

SAPIENZA UNIVERSITY OF ROME

DOCTORAL THESIS

**Latent Communication
in Artificial Neural Networks**

Author:
Luca MOSCHELLA

Supervisor:
Prof. Emanuele RODOLÀ
Prof. Francesco LOCATELLO

*A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy*

in the

GLADIA research lab
Department of Computer Science
Faculty of Information Engineering, Informatics, and Statistics

June 16, 2024

“Be less curious about people and more curious about ideas.”

Marie Curie

SAPIENZA UNIVERSITY OF ROME

Abstract

Faculty of Information Engineering, Informatics, and Statistics
Department of Computer Science

Doctor of Philosophy

Latent Communication in Artificial Neural Networks

by Luca MOSCHELLA

As NNs (Neural Networks) permeate various scientific and industrial domains, understanding the universality and reusability of their representations becomes crucial. At their core, these networks create intermediate neural representations, indicated as latent spaces, of the input data and subsequently leverage them to perform specific downstream tasks. This dissertation focuses on the *universality and reusability of neural representations*. Do the latent representations crafted by a NN remain exclusive to a particular trained instance, or can they generalize across models, adapting to factors such as randomness during training, model architecture, or even data domain? This adaptive quality introduces the notion of *Latent Communication* – a phenomenon that describes when representations can be unified or reused across neural spaces.

A salient observation from our research is the emergence of similarities in latent representations, even when these originate from distinct or seemingly unrelated NNs. By exploiting a partial correspondence between the two data distributions that establishes a semantic link, we found that these representations can either be projected into a universal representation (Moschella*, Maiorca*, et al., 2023), coined as Relative Representation, or be directly translated from one space to another (Maiorca* et al., 2023). Intriguingly, this holds even when the transformation relating the spaces is unknown (Cannistraci, Moschella, Fumero, et al., 2024) and when the semantic bridge between them is minimal (Cannistraci, Moschella, Maiorca, et al., 2023). Latent Communication allows for a bridge between independently trained NN, irrespective of their training regimen, architecture, or the data modality they were trained on – as long as the data semantic content stays the same (e.g., images and their captions). This holds true for both generation, classification and retrieval downstream tasks; in supervised, weakly supervised, and unsupervised settings; and spans various data modalities including images, text, audio, and graphs – showcasing the universality of the Latent Communication phenomenon. From a practical standpoint, our research offers the potential to repurpose and reuse models, circumventing the need for resource-intensive retraining; enables the transfer of knowledge across them; and allows for downstream performance evaluation directly in the latent space.

Indeed, several works leveraged the insights from our Latent Communication research (Kiefer and Buckley, 2024; Z. Wu, Y. Wu, and Mou, 2024; Jian et al., 2023; Norelli, Fumero, et al., 2023; G. Wang et al., 2023). For example, relative representations have been instrumental in attaining state-of-the-art results in Weakly Supervised Vision-and-Language Pretraining (C. Chen et al., 2023). Reflecting its significance, (Moschella*, Maiorca*, et al., 2023) has been presented orally at ICLR 2023 and Latent Communication has been a central theme in the UniReps: Unifying Representations in Neural Models Workshop at NeurIPS 2023, co-organized by our team.

Acknowledgements

First and foremost, I would like to express my gratitude to all my colleagues who worked in this line of research: Cannistraci Irene, Crisostomi Donato, Fumero Marco, Maiorca Valentino, Norelli Antonio, and Ricciardi Antonio. In particular, I am deeply thankful to Irene Cannistraci for her unwavering support, insightful discussions, joint work, and meticulous proofreading of this manuscript. Special thanks also go to Marco Fumero and Donato Crisostomi for their immense effort in organizing the first edition of the UniReps Workshop at NeurIPS 2023, which was a resounding success due to their dedication. A special mention goes to Valentino Maiorca, with whom I embarked on this journey. Working alongside Valentino has been an exceptional experience, and this dissertation owes much to our collaborative efforts.

I am immensely grateful to Prof. Emanuele Rodolà for fostering a free and creative environment. Initiating an entire research direction from playful observations at the laboratory has been truly inspiring. I am equally thankful to Prof. Francesco Locatello for his insightful discussions, supervision, and the opportunity to visit and collaborate during my fantastic period at ISTA. My heartfelt thanks also go to Dr. Maria Shugrina and Dr. Vojtěch Míčka for the fantastic experiences and valuable time spent during my internships at NVIDIA and NNAISENSE. I extend my sincere gratitude to the external reviewers, Prof. Nina Miolane and Prof. Marco Baroni, whose insightful remarks and suggestions have greatly improved this work.

My sincere thanks to all the other people with whom I had the honor to collaborate. The work we have accomplished together and the moments we have shared have been incredibly rewarding: Andrea Santilli, Cosimo Fiorini, Emanuele Frascaroli, Filippo Maggioli, Giambattista Parascandolo, Giorgio Mariani, Giovanni Trappolini, Leonidas Guibas, Luca Cosmo, Maks Ovsjanikov, Marco Ciccone, Matteo Boschini, Michele Bevilaqua, Nishkrit Desai, Or Litany, Or Perel, Pietro Barbiero, Pietro Liò, Riccardo Benaglia, Riccardo Marin, Roberto Dessi, Simone Antonelli, Simone Calderara, Simone Melzi, and Steve Azzolin. Moreover, I am thankful for the wonderful time and the enriching experiences shared with all the people at Sapienza, ISTA, NNAISENSE, and NVIDIA.

I am grateful to all the other colleagues from the GLADIA group at Sapienza and the Causal Learning and Artificial Intelligence group at ISTA for the stimulating discussions and the time spent together: Adrian R. Minut, Arianna Rampini, Berker Demirel, Daniele Baieri, Dingling Yao, Emilian Postolache, Irene Tallini, Lorenzo Basile, Marco Pegoraro, Michele Mancusi, Riccardo Cadei, and Silvio Severino.

My appreciation also extends to all the other people with whom I had stimulating discussions that have enriched and helped me along my journey: Alessandro Raganato, Alex Bronstein, Andrea Dittadi, Ari Morcos, Bogdan Gaza, Christos Tsirigotis, Clémentine Dominé, Emanuele Marconato, Emanuele Rossi, Even Oldridge, Fabrizio Frasca, Federico Scozzafava, Filip Szatkowski, Francesco Visin, Giovanni Zappella, Jonathan Masci, Luigi Gresele, Mateusz Pyla, Matthew Leavitt, Michael Bronstein, Patrik Reizinger, Pau Rodríguez López, Simone Azeglio, Simone Scardapane, Stefan Bejgu, Valentina Zantedeschi, Xavier Suau, and Zorah Lähner.

Finally, I want to thank everyone who shared their insights and encouragement with me, whether we worked together directly or simply had inspiring conversations. Your contributions have been truly valuable.

Contents

Abstract	iii
Acknowledgements	v
List of Figures	xi
List of Tables	xiii
Glossary	xv
List of Symbols	xix
Authored Publications	1
Latent Communication	1
Other Research Directions	1
I Introduction	3
1 Introduction to Latent Communication	5
1.1 Structure of the Thesis	8
2 Related Work	9
2.1 Representation similarity	9
2.2 Representation similarity measures	10
2.3 Manifold alignment	10
2.4 Model stitching	10
2.5 Relative information	11
2.6 Invariance and Equivariance in Representations	11
2.7 Theoretical Understanding	12
II Latent Communication	13
3 Problem Formalization	15
3.1 Framework	15
3.2 Problem Statement	17
3.2.1 Assumptions.	17
3.2.2 Problem.	18
3.3 Corollary problems	18
3.3.1 Zero-Shot Stitching	18
3.3.2 Latent Model Evaluation	18
3.3.3 Retrieval	19

4	Universal Representations	21
4.1	Introduction	21
4.2	Relative Representations	23
4.3	Latent Evaluation	25
4.3.1	Word Embeddings	25
4.3.2	Latent distance as a performance proxy	26
4.3.3	Training with Absolute vs. Relative representations	27
4.4	Zero-Shot Model Stitching	27
4.4.1	Image Reconstruction	28
4.4.2	Text Classification	29
4.4.3	Image Classification	30
5	Direct Translation	31
5.1	Introduction	31
5.2	Latent Space Translation	32
5.2.1	Assumptions	32
5.2.2	Method	33
5.3	Latent Communication via Translation	33
5.3.1	Cross-Architecture	34
5.3.2	Cross-Modality	37
5.3.3	Autoencoding	38
III	Overcoming Limitations in Latent Communication	41
6	Current limitations	43
7	Unknown Latent Transformation	45
7.1	Introduction	45
7.2	Infusing invariances	46
7.2.1	Method	47
7.2.2	Aggregation functions.	48
7.3	Experiments	48
7.3.1	Latent space analysis	48
7.3.2	Zero-Shot Stitching	49
7.3.3	Subspace selection	51
8	Limited Semantic Correspondence	53
8.1	Introduction	53
8.2	Method	53
8.3	Experiments	54
IV	Applying Latent Communication	57
9	Case Studies	59
9.1	ASIF: Coupled Data Turns Unimodal Models to Multimodal Without Training	59
9.2	From Charts to Atlas: Merging Latent Spaces into One	61
9.3	Zero-Shot Stitching in Reinforcement Learning	63

V	Conclusions	65
10	Conclusions	67
11	Contributions to the field	69
11.1	UniReps Workshop: Unifying Representations in Neural Models . . .	69
11.2	Works by other researchers	69
12	Limitations and Future Directions	71
VI	Appendices	75
A	Universal Representations	77
A.1	Anchors analysis	77
A.2	Dataset Information	78
A.3	Implementation Details	78
A.3.1	Word Embeddings	79
A.3.2	Relative representation space correlations	79
A.3.3	Latent distance as a performance proxy	80
A.3.4	Training with Absolute vs. Relative Representations	80
A.3.5	Image Reconstruction	80
A.3.6	Text Classification	81
A.3.7	Image Classification	81
A.4	Additional results	81
B	Direct Translation	89
B.1	Additional results	89
B.1.1	Scale invariance	90
B.1.2	Implementation Details	94
	Bibliography	97

List of Figures

1.1	Latent spaces learned by distinct trainings of the same AE	5
3.1	The Latent Communication Problem.	16
4.1	Latent spaces learned by distinct trainings of an high-dimensional AE	21
4.2	Relative Representation	24
4.3	Graph node classification task on Cora	26
4.4	Zero-Shot Stitching Reconstruction examples	28
5.1	Zero-shot stitching between absolute spaces	31
5.2	Direct translation illustration on a synthetic example.	32
5.3	Performance comparison of affine, linear, l-ortho, and ortho	34
5.4	Scale distribution in encodings of different pre-trained encoders	36
5.5	Performance comparison between different encoders and data modalities	37
5.6	Translation reconstruction examples grouped by dataset	38
7.1	CKA similarity of pretrained models on Fashion MNIST	46
7.2	Latent Spaces Cross-Architecture Similarity	49
7.3	Comparison of attention weights before and after fine-tuning	51
9.1	ASIF aligns latent spaces of frozen pre-trained encoders	60
9.2	Relative Latent Space Aggregation description	61
9.3	Environment variations in Car Racing	63
A.1	Accuracy vs Number of anchors	78
A.2	Self similarities correlations between each space	80
A.3	Correlations between performance and latent similarity	82
A.4	Alternative visualization of Table 4.1 with t-SNE	82
A.5	CIFAR-10 embeddings similarity across different models	88
B.1	Cross-domain stitching on CIFAR-10 and grayscale CIFAR-10	89
B.2	Performance comparison of affine, linear, l-ortho and ortho	90
B.3	Performance comparison between different encoders and data modalities	90
B.4	Translation reconstruction examples grouped by dataset.	91
B.5	Additional reconstruction examples grouped by dataset	91
B.6	Scale invariance of RoBERTa	93
B.7	Performance comparison of three MLPs	94

List of Tables

4.1	Similarity across word embeddings in absolute and relative spaces . . .	26
4.2	Performance comparison between relative and absolute representations . . .	27
4.3	Zero-Shot Stitching performance	28
4.4	Cross-lingual Zero-Shot Stitching performance comparison	29
4.5	Cross-architecture Zero-Shot Stitching performance comparison	29
4.6	Zero-Shot Stitching performance with different encoding techniques . .	30
5.1	Cross-architecture stitching with various \hat{T} and standard scaling . . .	35
5.2	Cross-architecture stitching with various \hat{T} and L2 normalization . . .	35
5.3	Zero-shot stitching for generation with various \hat{T}	39
7.1	Invariances summary	47
7.2	Graph and Text Stitching Performance	49
7.3	Image Stitching Performance Cross-Architecture and Cross-Seed . . .	50
7.4	Stitching Index Across Architectures and Seeds on Cora	50
7.5	Classification accuracy with pretrained stitched models	52
8.1	Qualitative and quantitative comparisons optimizing the Word2Vec space	54
8.2	Evaluation of the AO method in the vision domain	55
8.3	Cross-lingual Zero-Shot Stitching performance evaluation	55
9.1	Zero shot classification accuracy of different multimodal designs . . .	60
9.2	Relative Latent Space Aggregation classification accuracy comparison	62
9.3	Episode maximum return comparing in stitching	63
A.1	Additional results with different anchor selection strategies	83
A.2	Generalization of Section 4.3.1 to a different data modality	84
A.3	All the datasets utilized in Chapter 4 with their number of classes. . . .	84
A.4	Hyperparameter grid search performed in Section 4.3.2	85
A.5	The pretrained transformers used for the <i>Cross-lingual</i> setting	85
A.6	The pretrained transformers used for the <i>Cross-architecture</i> setting . . .	85
A.7	WikiMatrix analysis further details	85
A.8	Timm transformers used in Section 4.4.3.	86
A.9	Further results on Zero-Shot Stitching on Amazon Reviews coarse-grained	86
A.10	Further results on Zero-Shot Stitching on Amazon Reviews fine-grained	86
A.11	Zero-shot stitching performance comparison with XLM-R multilingual	87
A.12	Further results on Zero-Shot Stitching on CIFAR-100 fine-grained . . .	87
B.1	Zero-shot stitching for generation.	91
B.2	Cross-architecture stitching with various \hat{T} and standard sacling . . .	92
B.3	Cross-architecture stitching with various \hat{T} and L2 normalization . . .	92
B.4	HuggingFace models used as encoders (feature extractors)	95
B.5	Cross-architecture stitching for reconstruction tasks.	96

Glossary

AE AutoEncoder 5, 6, 21, 28, 38, 39, 80

AG News The AG News dataset (X. Zhang, Zhao, and LeCun, 2015), a collection of news articles for use in Natural Language Processing tasks such as text classification and sentiment analysis. 35, 92

ALBERT (albert-base-v2) The ALBERT model (Lan et al., 2020) with a base version 2 configuration, a lighter and more efficient version of BERT, designed to reduce model size while maintaining performance. Available pre-trained on HuggingFace 35, 37, 49, 90, 95

Amazon Reviews The Amazon Reviews dataset (Keung et al., 2020), a large collection of customer reviews, useful for sentiment analysis and other forms of Natural Language Processing. xiii, 29, 30, 55, 81, 84, 86, 87

AO Anchor Optimization xiii, 53–55

AT Affine Transformation 47

BERT-C (bert-base-cased) The BERT model (Devlin et al., 2019) with a base configuration and cased vocabulary, a breakthrough in the field of Natural Language Processing. Available pre-trained on HuggingFace. 29, 35, 37, 85, 90, 95

BERT-U (bert-base-uncased) The BERT model (Devlin et al., 2019) with a base configuration and uncased vocabulary, providing a foundational architecture for developing advanced Natural Language Processing systems. Available pre-trained on HuggingFace 29, 35, 37, 85, 90, 95

CCA Canonical Correlation Analysis 10

CIFAR-100 The CIFAR-100 dataset (Krizhevsky, 2009), similar to CIFAR-10 but with 100 classes, providing a more challenging task for image classification models. xiii, 22, 27, 28, 30, 34, 35, 38, 39, 48, 50–52, 77, 78, 81, 84, 87, 90–93, 96

CIFAR-10 The CIFAR-10 dataset (Krizhevsky, 2009), composed of 60,000 32x32 color images in 10 classes, with 6,000 images per class. xi, 26–28, 35, 36, 38, 39, 48, 50, 55, 84, 88, 89, 91, 92, 96

CiteSeer The CiteSeer dataset (Giles, Bollacker, and Lawrence, 1998), an academic literature digital library and search engine that focuses primarily on the literature in computer and information science. 27, 84

CKA Centered Kernel Alignment 10, 48, 49

CLIP (openai/clip-vit-base-patch32) The CLIP model (Radford et al., 2021) with a ViT-base and patch32 configuration, a state-of-the-art model for connecting visual and textual data, facilitating robust image-text understanding. Available pre-trained on HuggingFace 35, 50, 95

CNN Convolutional Neural Network 7, 9, 11, 22, 38, 63, 80

Cora The Cora dataset (Sen et al., 2008), a collection of scientific publications categorized into different classes, used for document classification and citation prediction. xi, xiii, 26, 27, 49, 50, 77, 78, 84

Cos. Cosine 47

DarkNet (cspdarknet53) The CSPDarkNet53 architecture, represents a significant advancement in the field of computer vision, offering a robust backbone for object detection models. Available pre-trained on HuggingFace 37, 90, 95

CV Computer Vision 35, 67

DBpedia The DBpedia dataset (Auer et al., 2007), derived from Wikipedia, provides a structured form of Wikipedia's content for various knowledge extraction and semantic search tasks. 29, 35, 48, 49, 84, 92

ELECTRA (google/electra-base-discriminator) The ELECTRA model (K. Clark et al., 2020) with a base discriminator configuration, introducing a novel pre-training methodology for language representations that is both efficient and effective. Available pre-trained on HuggingFace 29, 35, 37, 85, 90, 95

Eucl. Euclidean 47

FastText The FastText word embeddings, a powerful model designed for text representation and classification (Bojanowski et al., 2017). 10, 25, 26, 54, 80, 82, 83

F-MNIST (Fashion MNIST) The Fashion MNIST dataset (Xiao, Rasul, and Vollgraf, 2017), designed as a more complex alternative to the original MNIST, contains grayscale images of various fashion products. xi, 27, 28, 35, 38, 39, 46, 48, 50, 84, 91, 92, 96

GNN Graph Neural Network 22, 49, 50

ICA Independent Component Analysis 12

IMA Independent Mechanism Analysis 12

ImageNet1k The ImageNet1k dataset (J. Deng et al., 2009), a large-scale dataset designed for use in visual object recognition software research. 22, 30, 84

IMDB The IMDB dataset (Maas et al., 2011), a set of movie reviews for binary sentiment classification, widely used in Natural Language Processing research. 35, 92

IS Isotropic Scaling 47

LCP (Latent Communication Problem) The Latent Communication Problem or Latent Space Communication Problem is a novel formalization of a problem arising in machine learning, specifically within the context of neural networks. The problem focuses on unifying semantically related manifolds embedded in different data spaces. xi, 7–9, 15, 16, 18, 19, 21, 22, 24, 31, 32, 43, 45, 46, 48, 51, 53, 59, 67, 71–73

LLM Large Language Model 7

- LT** Linear Transformation 47
- MAE** Mean Absolute Error 29, 55, 86, 87
- MLP** Multi-Layer Perceptron xi, 35, 37, 48, 51, 52, 89, 90, 92–94
- MNIST** The MNIST dataset (L. Deng, 2012), a classic collection of handwritten digits widely used for training image processing systems. 5, 6, 21, 27, 28, 35, 38, 39, 48, 50, 84, 91, 92, 96
- N24News** The N24News dataset (Z. Wang et al., 2022), a collection of news articles with associated images. 35–37, 48, 90
- NLP** Natural Language Processing 6, 10, 35, 37, 54, 67
- NN** Neural Network iii, 5–7, 9–12, 15, 16, 18, 22, 24, 32, 39, 45, 49, 59, 71–73, 90, 91, 93
- OT** Orthogonal Transformation 47
- PCA** Principal Component Analysis 22
- PT** Permutation 47
- PubMed** The PubMed dataset (Sen et al., 2008), comprising abstracts from biomedical literature, serves as a resource for tasks like biomedical entity recognition and relation extraction. 27, 84
- PWCCA** Projection Weighted Canonical Correlation Analysis 10
- QKV** Query, Key, Value 51, 52
- RR** Relative Representation iii, xi, 7, 8, 22–32, 36, 43, 45–47, 50, 59, 61, 63, 72
- RexNet (rexnet_100)** The RexNet model (D. Han et al., 2020) with 100 layers, exemplifies the advances in neural network architecture for efficient and effective image processing. Available pre-trained on HuggingFace 30, 35, 37, 51, 52, 86, 87, 90, 95
- RL** Reinforcement Learning 59, 63, 67
- RLSA** Relative Latent Space Aggregation xi, xiii, 61, 62
- RoBERTa (roberta-base)** The RoBERTa model (Liu et al., 2019) with a base configuration, an optimized version of BERT that improves language understanding by carefully tuning hyperparameters and training with more data. Available pre-trained on HuggingFace xi, 29, 35, 37, 85, 90, 93, 95
- RSA** Representational Similarity Analysis 10
- SVCCA** Singular Value Canonical Correlation Analysis 10
- SVD** Singular Value Decomposition 10
- SVM (Support Vector Machine)** A Support Vector Machine (SVM) is a supervised learning model used for classification and regression tasks. It effectively creates a hyperplane or set of hyperplanes in a high-dimensional space, which can be used for classification, regression, or other tasks like outliers detection. 35, 37, 89, 94

TR Translation 47

TREC The TREC dataset (X. Li and Roth, 2002; Hovy et al., 2001), designed for research in text retrieval and information extraction, includes a wide range of question types. 29, 35, 48, 84, 92

VAE Variational AutoEncoder 9, 28, 80

ViT-B/16 (vit_base_patch16_224) The ViT model (Dosovitskiy et al., 2021) with a base configuration, patch size 16, and image size 224, introduces a novel approach to image classification, leveraging transformer architecture for visual tasks. Available pre-trained on HuggingFace 30, 35, 37, 50–52, 55, 79, 84, 86–88, 90, 95

ViT-B/16L (vit_base_patch16_384) The ViT model (Dosovitskiy et al., 2021) with a base configuration, patch size 16, and image size 384, introduces a novel approach to image classification, leveraging transformer architecture for visual tasks. Available pre-trained on HuggingFace 35, 37, 90, 95

RViT-B/16 (vit_base_resnet50_384) The ViT model (Dosovitskiy et al., 2021) with a base configuration, the ResNet50 backbone, and image size 384, combines the strengths of convolutional neural networks and transformer models for superior image analysis capabilities. Available pre-trained on HuggingFace 30, 35, 37, 50, 86, 87, 90, 95

ViT-S/16 (vit_small_patch16_224) The ViT model (Dosovitskiy et al., 2021) with a small configuration, patch size 16, and image size 224, introduces a novel approach to image classification, leveraging transformer architecture for visual tasks. Available pre-trained on HuggingFace 30, 35, 37, 55, 79, 84, 86–88, 90, 95

Word2Vec The Word2Vec word embeddings, renowned for transforming words into high-dimensional vector spaces, facilitating semantic analysis (Mikolov, K. Chen, et al., 2013). xiii, 25, 26, 54, 80, 82, 83

WVLP Weakly Supervised Vision-and-Language Pretraining iii, 8, 69, 70

XLM-R (xlm-roberta-base) The XLM-RoBERTa model (Conneau et al., 2020) with a base configuration, extends the RoBERTa architecture to support multilingual language processing, enabling improved cross-lingual performance. Available pre-trained on HuggingFace 35, 37, 90, 95

List of Symbols

- A Represents a subset of the training data X or Y , denoted as anchor samples. xix, xx, 17, 23, 24, 31–34, 43, 47, 53–55, 61, 77
- a Denotes a particular anchor point within A . 23, 53
- \hat{A} Out-of-domain anchors, i.e., anchors chosen from a domain different from the training distribution. Used to construct relative representations in domain adaptation tasks with limited correspondence and labeled data. 24, 55
- A_{XY} Anchor samples drawn simultaneously from both domains X and Y , denoted as parallel anchors, forming a subset of π . xix, 17, 23, 24, 32–34, 43, 53, 54
- Λ_{XY} A small subset of parallel anchors A_{XY} , i.e. $|\Lambda_{XY}| \ll |A_{XY}|$ xix, 53, 54
- Λ A small subset of anchors A , i.e. $|\Lambda| \ll |A|$ xix, 53, 54
- $\tilde{\Lambda}$ A small subset of anchors A embedded in a latent space through some encoding process 53, 54
- \tilde{A} The anchors embeddings, derived from the anchors A through an encoding process. xx, 23, 24, 31, 47, 53, 54
- \tilde{a} A specific anchor point from A , encoded into a latent space by E . 23, 24
- C An abstract semantic link between abstract data manifolds \mathcal{M}_X and \mathcal{M}_Y , critical for understanding semantic alignment between different data representations, though not directly observable. xx, 16–18, 33, 53
- \mathbb{D} The set of all possible decoder architectures and their parameterizations that can be employed to solve a specific task, representing the space of solutions in the context of neural network design. 17
- D The neural network’s decoding function, responsible to T , such as data classification or reconstruction. xx, 16–18, 31
- E The neural network’s encoding function, converting input data into a latent representation. xix, xx, 16–18, 23, 24, 46, 47, 53
- L_1 Represents the Manhattan distance metric. 47
- L_∞ Denotes the Chebyshev distance metric. 47
- \mathcal{L} The function measuring a neural network’s performance on a specified task T . 17, 18
- \mathcal{M} An abstract manifold representing the semantic of some data. xx, 15, 17
- \mathcal{M}_X The manifold for input space X , encapsulating the data underlying meaning in this domain. xix–xxi, 15–18

- $\widetilde{\mathcal{M}}_X$ The latent manifold associated with the latent space \widetilde{X} , representing the underlying meaning of encoded data from X . xx, xxi, 15, 16, 18, 21, 22, 25, 31, 33, 43, 45–47, 53
- \mathcal{M}_Y The manifold for input space Y , analogous to \mathcal{M}_X but for domain Y . xix, xx, xxii, 15–18
- $\widetilde{\mathcal{M}}_Y$ The latent manifold for space \widetilde{Y} , similar to $\widetilde{\mathcal{M}}_X$ but for encoded data from Y . xx, xxii, 15, 16, 18, 21, 22, 25, 33, 43, 45–47, 53
- N A general neural network architecture, described as combining encoder and decoder functions: $D \circ E$, capable of representing complex functions through layers of interconnected neurons. 16, 18
- ρ This function aggregates the relative spaces formed by different similarity functions d , for instance, through concatenation. 47, 48
- φ A embedding function, in the mathematical sense, that maps data points from a manifold \mathcal{M} to a space S , encapsulating the concept of manifold learning and data representation. 15, 17
- ϕ A collection of factors, including weight initialization, data shuffling, hyperparameters, and other stochastic factors, that collectively affect the training dynamics of neural networks, influencing the learned latent spaces. xxi, 17, 18, 21, 22, 25, 28, 43, 45, 46, 49, 61
- φ_X The embedding function mapping data points from the data manifold \mathcal{M}_X to the input space X , representing how abstract, low-dimensional structures are realized within the high-dimensional space. xx, 16, 18
- φ_Y The embedding function mapping data points from the data manifold \mathcal{M}_Y to the input space Y , analogous to φ_X but specific to the domain Y . 16, 18
- $\varphi_{\widetilde{X}}$ The embedding function mapping data points from the latent data manifold $\widetilde{\mathcal{M}}_X$ to the latent space \widetilde{X} , representing how encoded, abstract structures are realized within the latent high-dimensional space \widetilde{X} . xx, 16, 18, 21, 22, 25, 31, 33, 43, 45–47, 53
- $\varphi_{\widetilde{Y}}$ The embedding function mapping data points from the latent data manifold $\widetilde{\mathcal{M}}_Y$ to the latent space \widetilde{Y} , analogous to $\varphi_{\widetilde{X}}$ but specific to the domain Y . 16, 18, 21, 22, 25, 33, 43, 45–47, 53
- π Observable pairs from X and Y that are semantically connected, derived from the theoretical correspondence \mathcal{C} . xix, 16–18, 23, 32, 33, 43, 45, 53
- \mathbb{R} Denotes the set of all real numbers. 23, 33
- r Represents the relative representation of x with respect to certain anchors A . 23
- \mathbb{R}^d A Euclidean space with d dimensions. 22, 23
- R_p The relative projection function calculates the representations of a point \tilde{x} relative to a set of anchor points \tilde{A} , utilizing the similarity function d . 23, 24, 31, 47, 54
- S A generic space, potentially an input, latent, or any conceptual space where data or representations are defined. xx, xxi, 15, 17, 79

- s An element within the generic space S . 79
- d Aa generic function that assesses similarity between two latent representations. xx, 23, 24, 31, 45–47, 54, 61, 72
- T A specific problem addressed by neural networks, including classification, regression, and data generation, among others. xix, 17, 18
- θ The neural network's parameters, such as weights and biases, adjusted during training to minimize the loss function. 16, 17
- \mathcal{T} A transformation induced by changing the factors ϕ over the latent space. xxi, 16, 17, 21–24, 31–33, 43, 47, 53
- $\hat{\mathcal{T}}$ An estimate of the transformation between latent spaces, induced by changes in ϕ . xiii, 31–36, 39, 43, 91, 92
- \mathbb{T} An unknown class of transformations that the induced transformation \mathcal{T} belongs to. It depends on the data distribution, the task, additional constraints on the network, and possibly other factors. 16, 17, 21–24, 31–33, 38, 43, 45–49, 51, 72
- T_X A transformation applied to data in the latent space $T_X : \tilde{X} \rightarrow U_X$, used for unifying or translating between latent spaces. xxi, 16, 18, 21, 24, 25, 31, 32, 45, 47, 67
- T_Y A transformation applied to data in space $T_Y : \tilde{Y} \rightarrow U_Y$, similar to T_X but for domain Y . xxi, 16, 18, 21, 24, 25, 31, 32, 45, 47, 67
- U A universal space that includes the spaces U_X and U_Y , where latent manifold embeddings coincide for both X and Y . xxi, 16, 18, 21, 25, 31, 47, 48, 53
- \hat{U} A product space of invariant components approximating U , a universal space including U_X and U_Y with identical latent manifold embeddings for both domains. 47, 48
- U_X A universal space into which X (and their latent representations) can be mapped. xxi, 16, 18
- U_Y A universal space similar to U_X , but dedicated to the domain Y . xxi, 16, 18
- X The input space associated with domain X , encompassing all possible inputs that can be processed or generated within this context. xix–xxii, 15–18, 21–25, 31–34, 43, 45–47, 53, 54, 67, 79
- x A data point or sample from the input space X , representing an instance in this domain. xx, xxii, 15, 16, 23
- x A point on the manifold \mathcal{M}_X . xxii, 15–18
- \tilde{X} The latent space derived from input space X , representing encoded data. xx–xxii, 15–18, 21–25, 31–33, 43, 45–47, 53, 54, 61, 79
- \tilde{x} A sample in the latent space \tilde{X} , encoded from an input in X . xx, 15, 16, 23, 24, 47
- \tilde{x} A point on the latent manifold $\tilde{\mathcal{M}}_X$. xxii, 15, 47

Y The input space associated with domain Y , similar to X but for a potentially different data modality or domain. xix–xxii, 15–18, 21–25, 31–34, 43, 45–47, 53, 54, 67, 79

y A data point or sample from the input space Y , akin to x but from domain Y . 15

\mathcal{y} A point on the manifold \mathcal{M}_Y , akin to x but for domain Y . 15, 16, 18

\tilde{Y} The latent space derived from input space Y , analogous to \tilde{X} but for domain Y . xx–xxii, 15–18, 21, 22, 24, 25, 31–33, 43, 45–47, 53, 54, 61, 79

\tilde{y} A sample in the latent space \tilde{Y} , encoded from an input in Y . 15, 24, 54

$\tilde{\mathcal{y}}$ A point on the latent manifold $\tilde{\mathcal{M}}_Y$, similar to \tilde{x} . 15

Dedicated to my mother...

Authored Publications

Latent Communication

- Cannistraci, Irene, **Luca Moschella**, Marco Fumero, Valentino Maiorca, and Emanuele Rodolà (2024). “From Bricks to Bridges: Product of Invariances to Enhance Latent Space Communication”. In: *The Twelfth International Conference on Learning Representations (ICLR 2024, spotlight, top 5%)*. URL: <https://openreview.net/forum?id=vngVydDWft>.
- Cannistraci, Irene, **Luca Moschella**, Valentino Maiorca, Marco Fumero, Antonio Norelli, and Emanuele Rodolà (2023). “Bootstrapping Parallel Anchors for Relative Representations”. In: *The First Tiny Papers Track at ICLR 2023, Tiny Papers at ICLR 2023*. URL: <https://openreview.net/pdf?id=VBuUL2IWlq>.
- Crisostomi, Donato, Irene Cannistraci, **Luca Moschella**, Pietro Barbiero, Marco Ciccone, Pietro Lio, and Emanuele Rodolà (2023). “From Charts to Atlas: Merging Latent Spaces into One”. In: *NeurIPS 2023 Workshop on Symmetry and Geometry in Neural Representations*. URL: <https://openreview.net/forum?id=ZFu7CPTznY>.
- Maiorca*, Valentino, **Luca Moschella***, Antonio Norelli, Marco Fumero, Francesco Locatello, and Emanuele Rodolà (2023). “Latent Space Translation via Semantic Alignment”. In: *Thirty-seventh Conference on Neural Information Processing Systems*. URL: <https://openreview.net/forum?id=pBa70rGHlr>.
- Moschella***, **Luca**, Valentino Maiorca*, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà (2023). “Relative representations enable zero-shot latent space communication”. In: *The Eleventh International Conference on Learning Representations (ICLR 2023, oral, notable top 5%)*. URL: <https://openreview.net/forum?id=SrC-nwieGJ>.
- Norelli, Antonio, Marco Fumero, Valentino Maiorca, **Luca Moschella**, Emanuele Rodolà, and Francesco Locatello (2023). “ASIF: Coupled Data Turns Unimodal Models to Multimodal without Training”. In: *Thirty-seventh Conference on Neural Information Processing Systems*. URL: <https://openreview.net/forum?id=Xj0j3ZmWE1>.
- Ricciardi, Antonio Pio, Valentino Maiorca, **Luca Moschella**, and Emanuele Rodolà (2023). “Zero-shot stitching in Reinforcement Learning using Relative Representations”. In: *Sixteenth European Workshop on Reinforcement Learning*. URL: <https://openreview.net/forum?id=4tcXsImfsS1>.

Other Research Directions

- Frascaroli, Emanuele, Riccardo Benaglia, Matteo Boschini, **Luca Moschella**, Cosimo Fiorini, Emanuele Rodolà, and Simone Calderara (2023). “CaSpeR: Latent Spectral Regularization for Continual Learning”. In: *CoRR abs/2301.03345*. URL: <https://doi.org/10.48550/arXiv.2301.03345>.

*Equal contribution.

- Srivastava, Aarohi et al. (2023). “Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models”. In: *Transactions on Machine Learning Research*. ISSN: 2835-8856. URL: <https://openreview.net/forum?id=uyTL5Bvosj>.
- Crisostomi, Donato, Simone Antonelli, Valentino Maiorca, **Luca Moschella**, Riccardo Marin, and Emanuele Rodolà (2022). “Metric Based Few-Shot Graph Classification”. In: *The First Learning on Graphs Conference*. URL: <https://openreview.net/forum?id=VBXRmRBfRF>.
- Moschella, Luca**, Simone Melzi, Luca Cosmo, Filippo Maggioli, Or Litany, Maks Ovsjanikov, Leonidas Guibas, and Emanuele Rodolà (2022). “Learning Spectral Unions of Partial Deformable 3D Shapes”. In: *Computer Graphics Forum* 41.2, pp. 407–417. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.14483>.
- Norelli, Antonio, Giorgio Mariani, **Luca Moschella**, Andrea Santilli, Giambattista Parascandolo, Simone Melzi, and Emanuele Rodolà (2022). “Explanatory Learning: Beyond Empiricism in Neural Networks”. In: *CoRR* abs/2201.10222. URL: <https://arxiv.org/abs/2201.10222>.
- Trappolini, Giovanni, Luca Cosmo, **Luca Moschella**, Riccardo Marin, Simone Melzi, and Emanuele Rodolà (2021). “Shape Registration in the Time of Transformers”. In: *Advances in Neural Information Processing Systems*. Ed. by A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan. URL: <https://openreview.net/forum?id=ui4xChWcA4R>.

Part I

Introduction

Chapter 1

Introduction to Latent Communication

The intrinsic meaning of data often resides in complex, lower-dimensional structures, we define as “abstract manifolds”. These manifolds, grounded in the Manifold Hypothesis (Fefferman, Mitter, and Narayanan, 2016), represent the compact, underlying essence of high-dimensional data. Indeed, both human and machine capabilities are limited to observing representations of meaning that manifest within high-dimensional spaces. These representations serve as proxies for deeper, underlying conceptual entities, which are not directly observable. To illustrate this concept, consider the entity “cat”. It does not reside within our immediate perceptual field but within an abstract manifold of meaning. The “cat” discerned and interpreted is not the conceptual entity per se, but its representation within a higher-dimensional space, such as images or textual descriptions of cats. Furthermore, it is crucial to recognize that different spaces may exhibit varying degrees of expressive power, potentially leading to differences in the underlying abstract manifold. Nonetheless, abstract manifolds that denote the same conceptual entities bear semantic connections and are intrinsically similar. When we examine these related manifolds embedded into high-dimensional spaces, we directly observe a correspondence between these representations – for instance, between the textual descriptions of cats and their visual images. This alignment, in essence, establishes a connection between the manifold meanings, bridging the abstract conception of “cat” as described in textual captions with its visual representation in images.

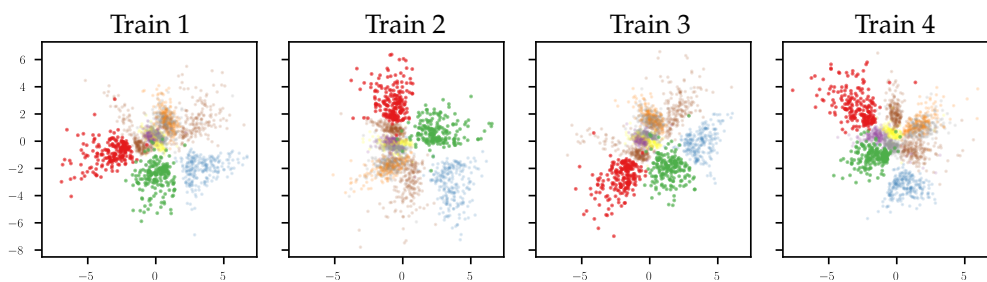


FIGURE 1.1: Latent spaces learned by distinct trainings of the same AE on MNIST. The bottleneck has size 2; thus, there is no dimensionality reduction in the latent space visualizations. The stochasticity in the training phase induces *intrinsically similar representations*.

NNs play a central role in this context, learning to transform high-dimensional data (e.g., images) into meaningful representations that are helpful for solving downstream tasks. Typically, these representations are seen as elements of a vector space, denoted as latent space, which corresponds to the constrained output (explicitly or implicitly) of a key component of the NN, e.g., the bottleneck in an AE (AutoEncoder),

or the word embedding space in NLP (Natural Language Processing) tasks. The foundational assumption is that learned latent spaces should encode the essential data characteristics and its abstract meaning required for task resolution. Thus, they should be an optimal encoding given the data distribution, the downstream task, and the network constraints.

In practice, however, learned latent spaces may vary, even when the initial assumptions are held constant. This phenomenon is illustrated in Figure 1.1, demonstrating the variation in latent spaces generated by training multiple times, from scratch, an AE with a two-dimensional bottleneck on the MNIST dataset. Perhaps unsurprisingly, these spaces differ from one another, breaking the fundamental assumptions made above. Indeed, the distribution of the latent embeddings is affected by various factors, such as the random initialization of the network weights, the data shuffling, hyperparameters, and other stochastic processes in the training phase. Although the resulting models may perform equally well on the task, this situation introduces several practical challenges. For example, it is notoriously challenging to compare latent spaces across different trainings or across different NNs; perhaps more importantly, re-using neural components trained on different embeddings of the same data becomes impossible, since they are incompatible.

Interestingly, although different, the learned representations in Figure 1.1 exhibit *intrinsic similarities*: the distances between the embedded representations are approximately the same across all spaces, even if their absolute coordinates differ. Indeed, in this case, the learned latent spaces are the same up to a nearly isometric transformation.¹ This symmetry arises from two main causes: (i) the existence of a semantic correspondence between their associated abstract manifolds, which are exactly the same in this instance since the same data and task is being considered; and (ii) the implicit biases underlying the optimization process, as noted by (Soudry et al., 2018), which compel the model to generalize. Consequently, this generalization ensures that similar samples – with respect to the task – are represented similarly. Discovering such symmetries and conserved quantities is a core step for extracting meaningful representations from raw data in biological and artificial systems (Higgins, Racanière, and Rezende, 2022; Benton et al., 2020; Lyle et al., 2020). Moreover, note that these emerging transformations, which relate intrinsically similar spaces, are limited to the embedded manifolds; thus, we cannot expect the same relationship between the entire ambient spaces, e.g., when considering out of distribution samples. This is a crucial point, as it implies that only a subset of the latent spaces is similar and *easily alignable*.

Leveraging this key observation, this dissertation investigates the *universality and reusability of these representations*. Do the latent representations crafted by a NN remain exclusive to a particular trained instance, or can they generalize across models, adapting to factors such as randomness during training, model architecture, or even data domain? This adaptive quality forms the basis of our exploration into *Latent Communication* – a novel paradigm that enables the unification or reuse of representations across disparate neural spaces, formally defined in Chapter 3. We investigate two principal strategies to harness this phenomenon, exploiting the partial semantic correspondence between the two spaces:

Universal Representation projects the latent spaces into a universal representation where the embedded manifolds are extrinsically equal. In practice, this universal space must be independent of the specific training regimen, architecture,

¹To the best of our knowledge, the first to acknowledge this behavior was (Olah, 2015) in a blog post.

data modality, and other stochastic factors; encoding only the intrinsic information underlying the data. We show an example of universal representation in Chapter 4, denoted as *RR (Relative Representation)* adopting a local coordinate system defined by the data itself.

Direct Translation translates directly between two specific latent spaces, explicitly approximating an ambient space transformation that induces an alignment between the manifolds embedded within them. We show an example of direct translation in Chapter 5, assuming the transformation relating the spaces is at most affine.

Through these strategies, we facilitate effective communication between latent spaces of different NNs, bridging the divide between various domains, models, architectures, and modalities.

From a practical standpoint, our research paves the way for model repurposing and *reuse*, eliminating the need for resource-intensive retraining and fostering a more sustainable AI development cycle; enables the *transfer of knowledge* across them, and allows for performance *evaluation directly in the latent space*. This holds true for generation, classification, and retrieval downstream tasks; in supervised, weakly supervised, and unsupervised settings; and spans various data modalities including images, text, audio, and graphs – showcasing the universality of the Latent Communication phenomenon. For example, it becomes feasible to classify images with a text classifier, or vice versa (Section 5.3.2).

This research makes multiple contributions to the field, summarized as follows:

- We empirically demonstrate that while representations learned by NNs can change due to various influencing factors, often the transformation that relates them is simple (e.g., the angles between latent embeddings often remain consistent).
- We introduce the novel concept of *Latent Communication* and formalize the LCP (Latent Communication Problem) (Chapter 3), providing a paradigm for understanding and leveraging the inherent connections between independently trained NNs; irrespective of their training regimen, architecture, or the data modality they were trained on – as long as the data semantic content stays the same (e.g., images and their captions).
- For the first time, we successfully showcase *Zero-Shot Stitching* (Section 3.3.1) of neural components produced by distinct training regimens, e.g., due to different seeds, neural architectures or data domains;
- Provide a *quantitative* latent measure of performance while training neural models, which is differentiable, does not need any labeled data, and is correlated with standard downstream performance measures such as accuracy.
- With an extensive set of experiments, we validate the performance of the proposed methods in multiple settings, tasks (classification, generation, retrieval), architectures (e.g., Transformers, CNNs (Convolutional Neural Networks)), and modalities (e.g., images, text, graphs, audio); showing that is possible to achieve Latent Communication across different architectural and modality changes.

Indeed, several works leveraged the insights from our Latent Communication research, and in particular the concept of Relative Representations. For example, they have proven fundamental in enabling continuous prompt transfers in LLMs

(Large Language Models) (Z. Wu, Y. Wu, and Mou, 2024); zero-shot image captioning without requiring any multimodal model training (Norelli, Fumero, et al., 2023); understanding shared speech-text representations (G. Wang et al., 2023); analyzing cognitive graphs (Kiefer and Buckley, 2024) and in the stitching of reinforcement learning agents in novel environments (Jian et al., 2023; Ricciardi et al., 2023); thus achieving zero-shot policy reuse. Our approach has also been instrumental in attaining state-of-the-art results in Weakly Supervised Vision-and-Language Pretraining (C. Chen et al., 2023). Reflecting its significance, (**Moschella**^{*}, Maiorca^{*}, et al., 2023) has been presented orally at ICLR 2023 and Latent Communication has been a central theme in the UniReps: Unifying Representations in Neural Models Workshop at NeurIPS 2023, co-organized by our team.

1.1 Structure of the Thesis

In this dissertation, we present a novel unified perspective on the LCP (Latent Communication Problem) research, reinterpreting several of our recent works (Cannistraci, **Moschella**, Fumero, et al., 2024; Cannistraci, **Moschella**, Maiorca, et al., 2023; Crisostomi, Cannistraci, et al., 2023; Maiorca^{*} et al., 2023; **Moschella**^{*}, Maiorca^{*}, et al., 2023; Norelli, Fumero, et al., 2023; Ricciardi et al., 2023). For the first time, we provide a formalization of the LCP (Latent Communication Problem) in Chapter 3.

Subsequently, in Chapters 4 and 5, we present two distinct methodologies to solve the LCP. The first method, discussed in Chapter 4, revolves around the concept of RR, which seeks to unify latent spaces into a universal space. The second approach, outlined in Chapter 5, focuses on establishing direct mappings between source and target spaces.

In Chapter 6 we discuss the limitations of the current approaches to solve the LCP, and in Chapters 7 and 8 we present two methods to overcome these limitations. In Chapter 7, we introduce a novel approach to tackle the LCP without any specific assumption on the transformation class relating the latent spaces. Meanwhile, in Chapter 8 we delineate a methodology to expand a small semantic correspondence between two latent spaces into a larger one, enabling the solution of the LCP even when it was not possible before.

To further illustrate the practical implications of LCP, we present three case studies in Sections 9.1 to 9.3. In Section 9.1 we show how to perform zero-shot captioning employing only unimodal models; in Section 9.2 we show how to merge distinct latent spaces; and in Section 9.3 we show how solving LCP enables zero-shot policy reuse in reinforcement learning.

The dissertation concludes in Chapter 10 with a summary of the key findings. In Chapter 11, we outline our main contributions to the field and discuss how other researchers have utilized our work. Finally, Chapter 12 presents potential avenues for future research.

Chapter 2

Related Work

2.1 Representation similarity

Recent years have witnessed a growing consensus among researchers in the deep learning community that “good” NNs tend to learn similar representations for semantically similar data, regardless of the architecture, training procedure, or domain in which they are applied.

This idea is supported by a plethora of empirical studies. For example, Morcos, Raghu, and S. Bengio, 2018 demonstrates that networks that generalize converge to more similar representations than networks that memorize; Y. Li et al., 2016 shows that some features are learned reliably in multiple networks; Kornblith et al., 2019 verifies that wider networks learn more similar representations; Bonheme and Grzes, 2022 shows that the VAE (Variational AutoEncoder) encoders representations in all but the mean and variance layers are similar across hyperparameters and learning objectives; Tsitsulin et al., 2020 develops an intrinsic method to characterize unaligned data manifolds of different dimensionality; Lenc and Vedaldi, 2015 shows the shallow representations in the first layers of CNNs are interchangeable across different networks; Lample et al., 2018 shows it is possible to build a bilingual dictionary by aligning monolingual word embeddings spaces in an unsupervised way; Rakotonirina et al., 2023 shows that automatically generated prompts can be learned on a language model and used to retrieve information from another; and many others (Barannikov et al., 2022; Chang, Tu, and Bergen, 2022; Antonello et al., 2021; Vulić, Ruder, and Søgaard, 2020; Movshovitz-Attias et al., 2017; Y. Bengio, Courville, and Vincent, 2014; Mikolov, Le, and Sutskever, 2013), recognizing that the phenomenon is particularly pronounced for large and wide models (Mehta et al., 2022; Somepalli et al., 2022). Furthermore, similar observations have been made in the context of biological models (Acosta et al., 2023; Raizada and Connolly, 2012; Kriegeskorte, Mur, and P. Bandettini, 2008; Laakso and Cottrell, 2000) and between artificial and biological representations (Sucholutsky and Griffiths, 2023; Sucholutsky, Muttenthaler, et al., 2023), suggesting foundational principles on information representation.

Although this is still not unanimously recognized (L. Wang et al., 2018) and missing strong theoretical justifications (Section 2.7), our framework is supported by the empirical evidence widely reported in these works. The LCP assumes that well-performing NNs trained on similar tasks and data produce intrinsically similar latent spaces, as formalized in Chapter 3, which allows us to unify them.

2.2 Representation similarity measures

Several metrics have been proposed to compare latent spaces generated by independent NNs (Klabunde et al., 2023), capturing their inherent similarity up to transformations that correlate the spaces. A classical statistical method is CCA (Canonical Correlation Analysis) (Hotelling, 1992), which is invariant to linear transformations. While variations of CCA seek to improve robustness through techniques like SVD (Singular Value Decomposition) and SVCCA (Singular Value Canonical Correlation Analysis) (Raghu et al., 2017) or to reduce sensitivity to perturbations using methods such as PWCCA (Projection Weighted Canonical Correlation Analysis) (Morcos, Raghu, and S. Bengio, 2018). Closely related to these metrics, the CKA (Centered Kernel Alignment) metric (Kornblith et al., 2019) measures the similarity between latent spaces while disregarding orthogonal transformations. However, recent research (Davari et al., 2022) demonstrates its sensitivity to transformations that shift a subset of data points in the representation space. Furthermore, highly relevant in the biological domain, RSA (Representational Similarity Analysis) is a method (Nili et al., 2014; Kriegeskorte, Mur, and P. A. Bandettini, 2008) used to compare and analyze the similarity of neural representations across different conditions, stimuli, or brain regions by correlating their respective similarity matrices.

2.3 Manifold alignment

Procrustes analysis has been instrumental in the alignment of latent spaces in deep NNs (C. Wang and Mahadevan, 2009; C. Wang and Mahadevan, 2008), particularly in NLP, where it is well-known that latent spaces of different languages are isomorphic (Vulić, Ruder, and Søgaard, 2020) and can be effectively aligned (Xing et al., 2015; Mikolov, Le, and Sutskever, 2013). Rooted in shape analysis, this method efficiently uncovers correspondences between latent spaces of different models through the estimation of an optimal orthogonal transformation (Gower, 1975). Previous works largely exploit Procrustes analysis to align latent spaces produced by the same architecture in different contexts (Csiszárík et al., 2021), such as multilingual FastText embeddings (Bojanowski et al., 2017; Smith et al., 2017). Procrustes analysis is termed “manifold alignment” because it aligns sets of keypoints that represent samples from lower-dimensional manifolds in a higher-dimensional space, thus aligning the intrinsic geometric structures of these manifolds.

Our research shares the common objective of aligning latent spaces, leveraging this alignment to enable or improve performance on various downstream tasks. For example, in Chapter 5 we broaden the scope of Procrustes analysis, applying it to Zero-Shot Stitching across disparate latent space dimensionalities, architectures, and data modalities.

2.4 Model stitching

Building on the observation of emergent intrinsic similarities between latent spaces, model stitching – combining different NNs to create a new model – has become an active research topic in the field of representation learning. For example, Lenc and Vedaldi, 2015 introduces *trainable* stitching layers that allow swapping parts of different networks; Csiszárík et al., 2021 demonstrates that the inner representations emerging in deep convolutional NNs with the same architecture, but different initializations can be matched with a surprisingly high degree of accuracy even with

a single, affine *trainable* stitching layer; while (Bansal, Nakkiran, and Barak, 2021; Csiszárík et al., 2021) employ stitching to quantitatively verify statements such as “good networks learn similar representations” and “more data, width, or time is better”. Other works, such as (Yaman et al., 2022; Biondi et al., 2021; Gygli, Uijlings, and Ferrari, 2021; Bianchi et al., 2020), try to directly produce compatible and reusable network components without stitching layers. In general, stitching has been mostly adopted in the literature to analyze NNs and verify statements regarding latent space similarity. An exception is (Lähner and Moeller, 2023) that, concurrently to the work presented in Chapter 5, targets the direct alignment of representational spaces, focusing on the compatibility of models trained end-to-end.

In our framework, we (i) sidestep the need for trainable stitching layers and propose for the first time *Zero-Shot* Model Stitching (Section 3.3.1); and (ii) propose to employ stitching to effectively reuse neural components, enabling many practical applications, some of which presented in Chapters 4, 5 and 7 to 9 and sections 9.1 to 9.3.

2.5 Relative information

Recognizing the importance of the relationships between data points, several methods have been proposed to exploit the relative information in the data. For example, the attention mechanism (Vaswani et al., 2017) and its variants (Kossen et al., 2021) exploit the relationship between features to extract meaningful representations; (Snell, Swersky, and Zemel, 2017) learn a metric space where the classification can be performed by measuring the distances with respect to prototype representations; You, Ying, and Leskovec, 2019 introduces Position-aware Graph Neural Networks (P-GNNs) to exploit position-aware node embeddings, Shalam and Korman, 2022 suggested the Self Optimal Transport feature transform to enrich the sample representations with higher order relations between the instance features, while Alvarez-Melis, Jegelka, and Jaakkola, 2019 suggested a general formulation of the optimal transport that accounts for global invariances in the underlying feature spaces.

Mathematically, the method presented in Chapter 4 bears resemblance to a kernel method (Hofmann, Schölkopf, and Smola, 2008); employing similarities of embedded features as a core ingredient. However, differently from kernel methods, we do not introduce learnable parameters and, crucially, we compute the representations explicitly without resorting to a kernel trick.

2.6 Invariance and Equivariance in Representations

Invariances in NN models can be enforced through various techniques operating at different levels, including adjustments to model architecture, training constraints, or input manipulation (Lyle et al., 2020). For example, (Benton et al., 2020) proposes a method to learn invariances and equivariances introducing augmentations in the training process; (Immer et al., 2022) introduces a gradient-based approach that effectively captures inherent invariances in the data. Meanwhile, (Ouderaa and Wilk, 2022) enables training of NNs with invariance to specific transformations by learning weight-space equivalents instead of modifying the input data. Other works directly incorporate invariances into the model through specific constraints, e.g., (Rath and Condurache, 2023) enforces a multi-stream architecture to exhibit invariance to various symmetry transformations without relying on data-driven learning; (Kandi et al., 2019) suggests an improved CNN architecture for better rotation invariance;

and (Gandikota et al., 2021) introduces a method for designing network architectures that are invariant or equivariant to structured transformations.

In contrast, the methodology presented in Chapter 4 proposes an alternative representation of the latent space that guarantees invariance to angle preserving transformation *of the latent space itself*, without requiring additional training but only a subset of the data. Building on this, Chapter 7 presents a method that directly *incorporates a set of invariances* into the learned latent space, creating a product space of invariant components which, combined, can capture complex transformations between the latent spaces.

2.7 Theoretical Understanding

While the empirical findings discussed throughout this manuscript provide substantial evidence for the similarity of representations in NNs, an exhaustive theoretical foundation is essential for fully understanding and leveraging these phenomena. Recent theoretical advancements have begun to shed light on the mechanisms behind the emerging representation similarity, offering a more solid ground for the empirical observations and methodologies employed in representation learning.

One direction of theoretical progress is the study of harmonics in NNs weights (Marchetti and Hillar, 2023), providing a mathematical framework for understanding the universality of neural representations. Furthermore, the intrinsic similarity of latent spaces, a core assumption of our framework, finds theoretical support in the field of linear identifiability within deep neural models, particularly in the context of nonlinear ICA (Independent Component Analysis) (Roeder, Metz, and Kingma, 2021; Khemakhem et al., 2020; Hyvarinen, Sasaki, and Turner, 2019; Hyvarinen and Morioka, 2016) and IMA (Independent Mechanism Analysis) (Ghosh et al., 2023; Sliwa et al., 2022; Gresele et al., 2021). This body of work suggests that, despite the complexity and non-linearity of deep learning models, their learned representations may converge towards similar structures when they capture the same underlying generative factors of data.

Part II

Latent Communication

Chapter 3

Problem Formalization

As discussed in Chapter 1, in machine learning and specifically within the context of NNs, our observational capabilities are confined to the high-dimensional representations of underlying conceptual entities. Consider the notion of a “cat”, a conceptual entity that resides within an abstract manifold of meaning. What we perceive and process are not these abstract entities themselves, but their embedding within a higher-dimensional space – namely, images of cats. When we have semantic correspondences between two distinct data spaces, we are effectively observing an alignment between these high-dimensional spaces, e.g., between captions and images, and indirectly, the correspondence between the caption’s meaning and the image’s meaning. The crux of our exploration is anchored in the fact that two semantically related manifolds are similar (but not necessarily isomorphic, since different spaces may have different expressive power) and easily alignable, even by transformations that operate on the entire ambient spaces in which they are embedded.

In the following sections, we formalize the LCP (Latent Communication Problem), an illustration of which is shown in Figure 3.1.

3.1 Framework

Data notation. We denote input data spaces as X and Y , containing data points x and y respectively. We indicate with \mathcal{M}_X and \mathcal{M}_Y their underlying abstract data manifolds, that contains data points denoted as \mathbf{x} and \mathbf{y} . Similarly, the symbols used to denote the latent spaces and the associated latent abstract manifolds are $\tilde{x} \in \tilde{X}$, $\tilde{y} \in \tilde{Y}$ and $\tilde{\mathbf{x}} \in \tilde{\mathcal{M}}_X$, $\tilde{\mathbf{y}} \in \tilde{\mathcal{M}}_Y$, respectively. We use S to indicate a generic space, e.g., the input space or the latent space of NNs.

Manifold embedding. We consider data semantics to reside on unknown and unobservable low-dimensional abstract manifolds, denoted as \mathcal{M} , which are embedded¹ through the operation φ_S into observable high-dimensional spaces, denoted as S :

$$\varphi_S : \mathcal{M} \hookrightarrow S, \quad (3.1)$$

where φ maps data points from the manifold \mathcal{M} to the ambient space S . The notation $\varphi_S(\mathcal{M})$ indicates the entirety of the *manifold embedding* within S . Although multiple mappings φ can embed \mathcal{M} into a generic high-dimensional space, for a specific configuration of S where the embedding is already established (e.g., a dataset of images or a latent space), the mapping φ_S that realizes \mathcal{M} within S is unique. The manifold embedding $\varphi_S(\mathcal{M})$ is precisely what the Manifold Hypothesis (Fefferman, Mitter, and Narayanan, 2016) refers to, suggesting that high-dimensional data observed in S

¹In the mathematical sense.

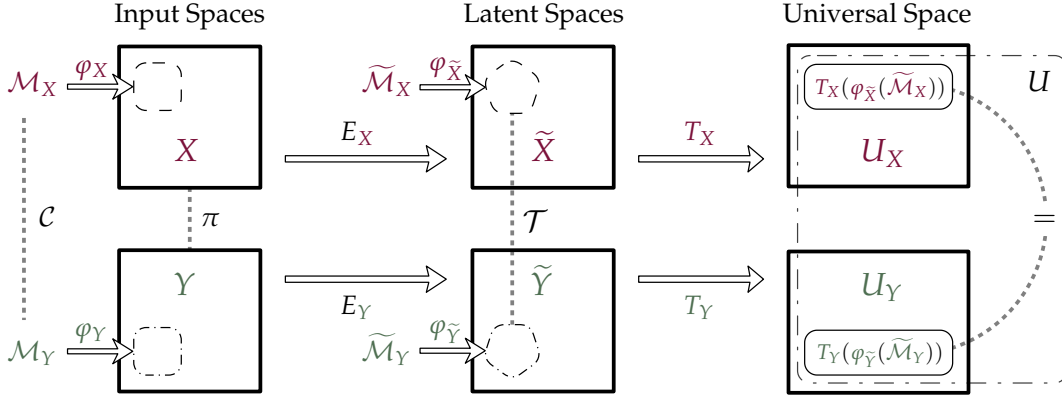


FIGURE 3.1: The LCP (Latent Communication Problem). The unobservable manifolds \mathcal{M}_X and \mathcal{M}_Y are embedded into the input spaces X and Y through φ_X and φ_Y . We can observe the semantic relationship between these manifolds, denoted as \mathcal{C} , through a partial correspondence π defined between the input spaces. The encoding functions E_X and E_Y map the input spaces to the respective latent spaces \tilde{X} and \tilde{Y} , modifying the embedded manifolds and inducing a correlation between them through some transformation $\mathcal{T} \in \mathbb{T}$. The objective is to discover two specific transformations, T_X and T_Y , that allow the latent spaces \tilde{X} and \tilde{Y} to be mapped into universal spaces U_X and U_Y . In the universal space U , the latent manifold embeddings must coincide:

$$T_X(\varphi_{\tilde{X}}(\tilde{\mathcal{M}}_X)) = T_Y(\varphi_{\tilde{Y}}(\tilde{\mathcal{M}}_Y)) \subseteq U.$$

actually lies on or near this embedded lower-dimensional manifold, capturing the intrinsic geometry and essential characteristics of the data.

Semantic correspondence. Given two data manifolds \mathcal{M}_X and \mathcal{M}_Y , they are semantically related if there exists a partial correspondence

$$\mathcal{C} \subseteq \mathcal{M}_X \times \mathcal{M}_Y \quad (3.2)$$

between the two manifolds, such that $\forall (\mathbf{x}, \mathbf{y}) \in \mathcal{C}$, \mathbf{x} and \mathbf{y} are related by the same semantic relationship. The correspondence \mathcal{C} is an abstract relation between the manifolds, and it is not directly observable. However, we can observe a partial correspondence π derived from \mathcal{C} and represented in the associated ambient spaces X and Y :

$$\pi \subseteq \{(\varphi_X(\mathbf{x}_i), \varphi_Y(\mathbf{y}_i)) \mid (\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{C}\}. \quad (3.3)$$

One example of such correspondence π involves images paired with one or more captions that describe them. Meanwhile, in \mathcal{C} , the corresponding elements associate their abstract meanings.

Latent Spaces. We consider NNs as parametric functions N^θ compositions of *encoding* and *decoding* functions, $N^\theta = D^{\theta_2} \circ E^{\theta_1}$, where the encoder E^{θ_1} is responsible for computing a latent representation $\tilde{\mathbf{x}} = E^{\theta_1}(\mathbf{x})$, $\mathbf{x} \in X$ for some domain X . This encoding function transforms the manifold embedded in the input space $\varphi_X(\mathcal{M}_X)$ into a latent manifold embedding $\varphi_{\tilde{X}}(\tilde{\mathcal{M}}_X)$, implicitly associated to some manifold $\tilde{\mathcal{M}}_X$. This latent representation is then exploited to solve downstream tasks, such as classification, reconstruction or generation, optimizing over some objective function. In the following, we will drop the dependence of parameters θ for notational convenience when not required, and indicate with \tilde{X} the latent space associated to X . We

will use the terms *latent representation* and *absolute representation* interchangeably, to refer to the output of the encoder. Moreover, for each module E (equivalently for D), we indicate with E_X if the module E was trained on the domain X .

Downstream task. We consider the decoder D to be responsible for solving a generic downstream task T at hand (e.g., classification, generation, etc.). We indicate with \mathcal{L}_D^T how well the decoder D is performing on the task T , i.e., the loss. Furthermore, we assume that the loss is computed on a test split, and the D is trained on a training split of the data.

Most importantly, we indicate with $\mathcal{L}^T(S)$ the lowest possible loss achievable by any decoder D trained from scratch on S to solve T :

$$\mathcal{L}^T(S) = \min_{D \in \mathbb{D}} \mathcal{L}_D^T(S), \quad (3.4)$$

where \mathbb{D} is the set of all possible decoders.

3.2 Problem Statement

3.2.1 Assumptions.

Semantic Correspondence. We assume that the data manifolds \mathcal{M}_X and \mathcal{M}_Y are related by a semantic correspondence \mathcal{C} , partially observable through π . Moreover, we assume such correspondence is partially provided as *parallel anchors* $A_{XY} \subseteq \pi$.

Good Encoders. Throughout our work, we assume that the encoders E_X and E_Y are good. Formally, we can express this assumption as follows:

$$\mathcal{L}^T(X) = \mathcal{L}^T(\tilde{X}) \quad \text{and} \quad \mathcal{L}^T(Y) = \mathcal{L}^T(\tilde{Y}), \quad (3.5)$$

that is, the task T can be solved with the same performance on the input spaces X and Y , as well as on the respective latent spaces \tilde{X} and \tilde{Y} .

In practice, this means that good encoders map data into the latent space without losing information useful for the task T (e.g., pre-trained universal feature extractors).

Emerging Similarities. As previously discussed, we argue that the learned latent spaces are not only a function of the data, the specific loss and the task; but in practice they are also affected by the optimization process used to train the network due to weight initialization, data shuffling, hyperparameters, data domain and other stochastic or non-semantic factors. We denote these factors collectively by ϕ .

In particular, as shown in Figures 1.1 and 4.1 and widely observed in the literature (Section 2.1), changing these factors induces some transformation \mathcal{T} over the latent manifold embedding:

$$\phi \rightarrow \phi' \quad \text{implies} \quad E^\theta(\varphi(x)) \rightarrow \mathcal{T}E^\theta(\varphi(x)), \quad \forall x \in \mathcal{M}, \quad (3.6)$$

where φ is the embedding operation described in Section 3.1. We assume that these transformations fall into some unknown class of transformation $\mathcal{T} \in \mathbb{T}$ (e.g., orthogonal transformations), when the variation factors are restricted to elements of ϕ .

3.2.2 Problem.

We are given the input spaces X and Y , their associated abstract manifolds \mathcal{M}_X and \mathcal{M}_Y in semantic correspondence $\mathcal{C} \subseteq \mathcal{M}_X \times \mathcal{M}_Y$ observable through π , and embedded in X and Y through the embedding functions φ_X and φ_Y ; and, two NNs $N_X = D_X \circ E_X$, $N_Y = D_Y \circ E_Y$ trained on X and Y , respectively, to solve the task T .

Our objective is to unify the latent manifold embeddings $\varphi_{\tilde{X}}(\tilde{\mathcal{M}}_X)$ and $\varphi_{\tilde{Y}}(\tilde{\mathcal{M}}_Y)$ into a universal space U , by finding $T_X : \tilde{X} \rightarrow U_X$ and $T_Y : \tilde{Y} \rightarrow U_Y$:

$$\begin{aligned} \forall (x, y) \in \mathcal{C}, \quad T_X(E_X(\varphi_X(x))) = T_Y(E_Y(\varphi_Y(y))) \subseteq U \\ \text{such that} \\ \mathcal{L}^\top(\tilde{X}) = \mathcal{L}^\top(U_X) \quad \text{and} \quad \mathcal{L}^\top(\tilde{Y}) = \mathcal{L}^\top(U_Y). \end{aligned} \tag{3.7}$$

Note that, the transformations are constrained to align only the latent manifold embeddings $\varphi_{\tilde{X}}(\tilde{\mathcal{M}}_X)$ and $\varphi_{\tilde{Y}}(\tilde{\mathcal{M}}_Y)$, not necessarily requiring alignment of the entire spaces \tilde{X} and \tilde{Y} . In practice, this implies we are trying to find transformations of the latent ambient spaces that align as best as possible the manifolds embedded in them, without losing information useful for the task T .

3.3 Corollary problems

Solving the general LCPs allows us to directly address several corollary tasks, which are of extreme practical interest. In the following sections, we describe how the LCP enables the reuse of neural components (Section 3.3.1), a downstream performance evaluation directly in the latent space (Section 3.3.2), and the development of advanced retrieval systems (Section 3.3.3).

3.3.1 Zero-Shot Stitching

Solving the LCP defined in Section 3.2 enables zero-shot interoperability of pre-trained neural components. In previous works, such as Bansal, Nakkiran, and Barak, 2021; Lenc and Vedaldi, 2015, stitching layers are *trainable* linear projections that allow comparing the representations of different networks. Instead, our framework unlocks the possibility of *Zero-Shot Stitching* different neural components, treating them as frozen black-box modules.

We define a generic *stitched model* as the composition of an encoder, that embeds data, plus an independent decoder specialized in a downstream task (e.g., classification, reconstruction):

$$N_{XY} = D_Y \circ E_X. \tag{3.8}$$

The stitching operation is always performed without training or fine-tuning, in a zero-shot fashion.

In Sections 4.4, 5.3, 7.3.2, 8.3 and 9.3, we show that the latent communication framework allows us to stitch together independent neural components, demonstrating empirically that re-using neural components is possible without the necessity for extensive retraining or fine-tuning.

3.3.2 Latent Model Evaluation

Unifying the latent spaces into universal spaces allows us to compare the latent spaces across variations of the factors ϕ . Interestingly, in Section 4.3.2 we show

that solving the LCP implies having a quantitative latent measure of downstream performance, provided that a reliable reference model is available. This measure does not require any labeled data and correlates with standard downstream performance measures. Consequently, we can assess the quality of a model directly, potentially during training, without the need to solve the downstream task explicitly.

3.3.3 Retrieval

The solution to the LCP facilitates the development of advanced retrieval systems that leverage independently computed representations. This allows for the retrieval of data points from one space using queries from another space, without the necessity for a shared training set, we showcase it in Sections 4.3.1 and 8.3. Finally, we demonstrate that solving the LCP also enables zero-shot captioning, by retrieving images using text queries, and vice versa, without any multimodal model training, as we show in Section 9.1.

Chapter 4

Universal Representations

*Relative representations enable zero-shot latent space communication*¹

In this Chapter, we tackle the LCP defined in Chapter 3 with the additional assumption that the latent manifold embeddings $\varphi_{\tilde{X}}(\tilde{\mathcal{M}}_X) \subseteq \tilde{X}$ and $\varphi_{\tilde{Y}}(\tilde{\mathcal{M}}_Y) \subseteq \tilde{Y}$ are always approximately related by a transformation $\mathcal{T} \in \mathbb{T}$, where \mathbb{T} is the class of transformations that preserve angle norms. Referring to Figure 3.1, we define analytically and independently T_X and T_Y as parameter-free *relative projections*, that implicitly unify the latent manifold embeddings in U .

4.1 Introduction

In Chapter 1, we discussed how the learned latent spaces are subject to changes even when the factors ϕ remain fixed. We illustrated this phenomenon in Figure 1.1 with a toy example on a bi-dimensional AE, and formalized the problem of unifying them in Chapter 3.

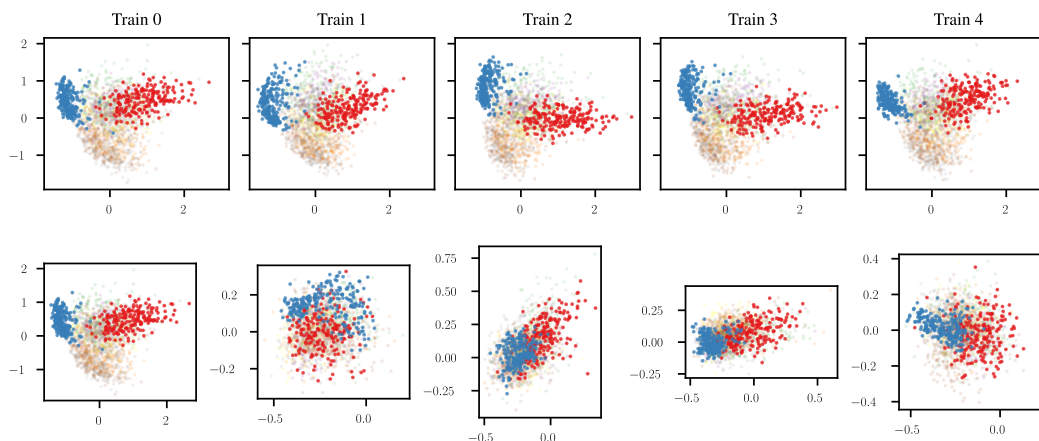


FIGURE 4.1: Latent spaces learned by distinct trainings of the same high-dimensional AE on the MNIST dataset. Each column is the latent space obtained by the AE with a different seed. On the first row, the dimensionality reduction is performed through PCAs fitted independently on each latent space, meanwhile, on the second row PCA is fitted on the leftmost latent space and then applied to all of them.

¹Luca Moschella*, Valentino Maiorca*, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà (2023). “Relative representations enable zero-shot latent space communication”. In: *The Eleventh International Conference on Learning Representations (ICLR 2023, oral, notable top 5%)*. URL: <https://openreview.net/forum?id=SrC-nwieGJ>

In Figure 4.1, we further illustrate the phenomenon, exploiting the properties of PCA (Principal Component Analysis) to demonstrate it also happens on high-dimensional latent spaces. Indeed, the *second row* of the figure proves that latent spaces learned by distinct trainings of the same high-dimensional AE are extrinsically different; since PCA fitted on one latent space and applied to the others does not align them (up to rotations and reflections). This extrinsic difference is a significant challenge in addressing any of the tasks outlined in Section 3.3; for instance, it hinders any form of reuse or comparison between neural components trained on different embeddings of the same data, since they are incompatible. Nevertheless, the *first row* in Figure 4.1 shows that the high-dimensional latent spaces, although extrinsically different, are *intrinsically similar*, as the PCA fitted independently on each latent space produces similar results.

Motivated by these empirical observations, in this Chapter, we address the LCP defined in Chapter 3 with an additional assumption that the latent manifolds embeddings $\varphi_{\tilde{X}}(\tilde{\mathcal{M}}_X) \subseteq \tilde{X}$ and $\varphi_{\tilde{Y}}(\tilde{\mathcal{M}}_Y) \subseteq \tilde{Y}$ are always approximately related by a transformation $\mathcal{T} \in \mathbb{T}$, where \mathbb{T} is the class of transformations that preserve angle norms. To tackle this simplified problem, we suggest adopting a local coordinate system defined by the data itself. Data points in the latent space becomes a set of coefficients that encode the point as a function of other data samples, instead of an independent point in \mathbb{R}^d . The proposed *RR (Relative Representation)* directly encodes the intrinsic information underlying the data, and with an appropriately chosen similarity function (e.g., cosine similarity), depends solely on the angles norms between embeddings by construction; de facto infusing an invariance to angle norm preserving transformations in the latent space that unifies them.

We show how neural architectures can leverage these RRs to guarantee, in practice, invariance to latent isometries and local rescalings, enabling a variety of applications from zero-shot model Section 3.3.1 stitching to latent space comparison Section 3.3.2 between diverse settings. Remarkably, this enables a form of compositionality between learning models; it allows, for instance, to stitch together an encoder trained on ImageNet1k with a decoder trained on CIFAR-100, as we showcase in our experiments. We extensively validate the generalization capability of our approach on different datasets, spanning various modalities (images, text, graphs), tasks (e.g., classification, reconstruction) and architectures (e.g., CNNs, GNNs, transformers).

The main contributions can be summarized as follows:

- We show that the representations learned by NNs are subject to change due to several training factors; nonetheless, the norm of the angles between latent embeddings are often preserved.
- We introduce a novel relative representation for latent embeddings, that is invariant by construction to the transformations induced by the factors ϕ .
- For the first time, we successfully demonstrate *Zero-Shot Stitching* (Section 3.3.1) of neural components produced by distinct training regimens, e.g., due to different seeds or different neural architectures; we validate our findings on different data modalities (e.g., images, text).
- Our framework also provides a *quantitative* latent measure of performance (Section 3.3.2) while training neural models, which is differentiable, does not need any labeled data, and is correlated with standard downstream performance measures such as accuracy.

4.2 Relative Representations

Assumption. In this Chapter, we make the core assumption that \mathbb{T} is the class of transformations that preserve the norm of the angles between elements of the latent space, namely $|\angle(\tilde{x}_i, \tilde{x}_j)| = |\angle(\mathcal{T}\tilde{x}_i, \mathcal{T}\tilde{x}_j)|$ for every $(x_i, x_j) \in X$. By “angle norm”, we mean the absolute value of the angle between two elements, which ensures that global reflections do not change the sign of the considered similarity. While this assumption might seem too restrictive, in practice it arises in several real scenarios, as we show in the following sections. Indeed, in classification tasks, this assumption is further supported by Figures B.6 and B.7 which show that the embeddings scale does not affect classification performance. Therefore, only the angle between embeddings is relevant. Additionally, in Chapter 7, we completely remove this assumption.

Method. To build our representation, we start by selecting a subset A of the training data X , which we denote as *anchors*. Every sample in the training distribution will be represented with respect to the embedded anchors $\tilde{a}_j = E(a_j)$ with $a_j \in A$. As a measure capturing the relation between the anchors and the other samples, we consider a generic similarity function $d : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$, yielding a scalar score between two absolute representations $d(\tilde{x}_i, \tilde{x}_j)$. Given the anchors A in an arbitrary ordering $a_1, \dots, a_{|A|}$, we define the RR of $x_i \in X$ as:

$$r_{x_i} = (d(\tilde{x}_i, \tilde{a}_1), d(\tilde{x}_i, \tilde{a}_2), \dots, d(\tilde{x}_i, \tilde{a}_{|A|})), \quad (4.1)$$

for convenience, we equivalently define the relative projection function R_p :

$$R_p(\tilde{x}; \tilde{A}, d) = \bigoplus_{\tilde{a}_i \in \tilde{A}} d(\tilde{x}, \tilde{a}_i) \quad (4.2)$$

where \bigoplus denotes row-wise concatenation and all embeddings are produced by the same encoding function E . For notational convenience, we denote the relative projection of a set of samples \tilde{X} with $R_p(\tilde{X}; \tilde{A}, d)$, which is defined as the collection of relative projections of individual samples $\tilde{x} \in \tilde{X}$. Figure 4.2 illustrates the key differences between absolute and RRs.

Choice of the anchors. Anchors directly affect the expressivity of the RR space, and are related to the task at hand. For example, in a classification task, we should sample anchors from each class in the training set, in order to well represent each data sample in X . We refer to Appendix A.1 for an analysis of different anchor selection strategies.

Parallel anchors. One case of interest arises when the data comes from different domains or modalities $X \neq Y$, and we are given a partial correspondence π , as defined in Section 3.1. In this case, we can obtain *parallel anchors*² by sampling simultaneously from both domains X and Y :

$$A_{XY} \subseteq \pi \quad (4.3)$$

We show an example of parallel anchors in Section 4.4.2, where X and Y are Amazon reviews in two different languages; illustrate a strategy to automatically expand this correspondence in Chapter 8; and further explore a multimodal application in Section 9.1.

²We use the term “parallel” to indicate they represent the same underlying meaning.

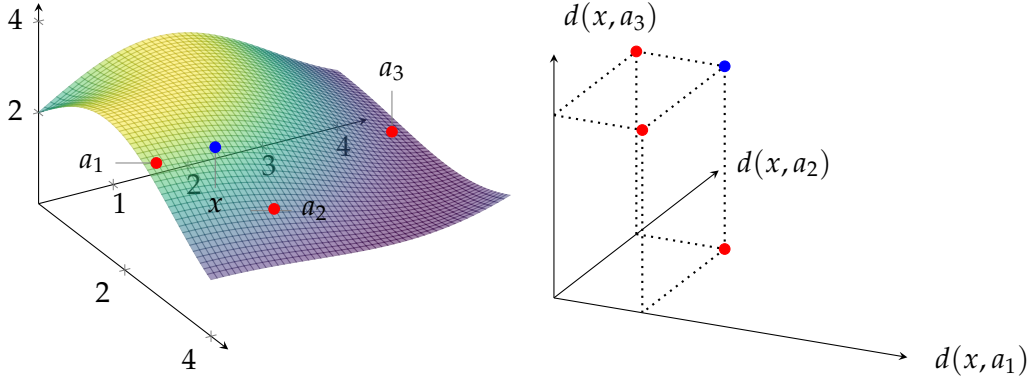


FIGURE 4.2: RR (Relative Representation). (left): a sample x and three anchor samples a_1, a_2, a_3 are embedded in a latent space and lie on the underlying embedded data manifold. (right): each dimension is treated as coefficients in a coordinate system defined by the anchors, the new representation of x is given by its similarities with respect to the anchors. Anchors are orthogonal in this example only for visualization purposes.

Out of domain anchors. Surprisingly, the choice of the anchors is not restricted to elements in the training distribution. Given an encoder pre-trained on a fixed training distribution, we can pick elements from a set \hat{A} that is out-of-domain w.r.t. \mathbf{X} , and build the RRs on top of \hat{A} . We refer to these as *OOD anchors* and exploit them, e.g., to solve domain adaptation tasks where we do not have access to a correspondence, and have scarce data labels. We refer to the Sections 4.4.2 and 4.4.3 for real-world examples.

Universal Representations. In this work, we choose the cosine similarity as the similarity function due to the properties it induces on the RR. The cosine similarity is the dot product of unit vectors, corresponding to the cosine of the angle $\cos \theta$ between the two. Importantly, $\cos \theta$ does not change if we apply the same angle-norm preserving transformation \mathcal{T} to them, i.e., the cosine similarity is invariant to rotations, reflections, and independent rescaling of each point. While this is not true for translations, NNs commonly employ normalization techniques (e.g., InstanceNorm (Ulyanov, Vedaldi, and Lempitsky, 2016)) to center the latent spaces. Under this assumption, cosine similarity guarantees RRs invariant also to translations.

This means we have the freedom to change the embedding function E with any other function E' that produces different representations with same angles, i.e.:

$$[d(\tilde{x}_i, \tilde{a}_1), \dots, d(\tilde{x}_i, \tilde{a}_{|A|})] = [d(\mathcal{T}\tilde{x}_i, \mathcal{T}\tilde{a}_1), \dots, d(\mathcal{T}\tilde{x}_i, \mathcal{T}\tilde{a}_{|A|})] \quad (4.4)$$

where d is the cosine similarity and \mathcal{T} , induced by E' , is an arbitrary angle-norm preserving transformation.

An implication of this invariance is that we can solve the LCP (defined in Chapter 3), simplified by the assumption that \mathbb{T} is the class of angle-norm preserving transformations. Indeed, we can define T_X and T_Y as independent relative projections over parallel anchors A_{XY} :

$$\begin{aligned} T_X(\tilde{x}) &= R_p(\tilde{x}; \tilde{A}_X, \text{cosine}) & \forall \tilde{x} \in \tilde{X} \\ T_Y(\tilde{y}) &= R_p(\tilde{y}; \tilde{A}_Y, \text{cosine}) & \forall \tilde{y} \in \tilde{Y}, \end{aligned} \quad (4.5)$$

these transformations are enough to unify the latent manifold embeddings:

$$T_X(\varphi_{\tilde{X}}(\tilde{\mathcal{M}}_X)) = T_Y(\varphi_{\tilde{Y}}(\tilde{\mathcal{M}}_Y)) \subseteq U, \quad (4.6)$$

as we demonstrate empirically in Sections 4.3 and 4.4. Additionally, as we demonstrate in Section 4.3.3, the original task can be solved in this universal space with a comparable performance.

We remark that other choices of similarity function can be made to enforce different invariances into the representation, refer to Chapter 7 for an extensive exploration of this aspect.

4.3 Latent Evaluation

In this Section, we demonstrate how RRs can effectively be used to produce latent spaces that are stable under a variety of factors ϕ as described in Section 3.3.2. To remark, our main question is the following: Given two different learning models that are trained independently, can we compare their latent embeddings? We answer in the positive, showing the gained invariance enables effective communication between different, but semantically equivalent latent spaces.

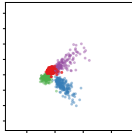
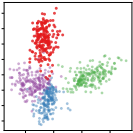
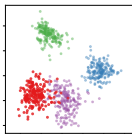
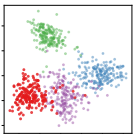
In particular, we analyze how different word embedding spaces, once projected onto RRs, are intrinsically the same (Section 4.3.1); we then show how the similarity between the relative counterparts of two or more embedding spaces is a surprisingly good predictor of model downstream performance (Section 4.3.2); finally, we confirm that RRs in the training phase are not detrimental to performance (Section 4.3.3).

4.3.1 Word Embeddings

Experimental setting. We select two different word embeddings on the English language, namely FastText and Word2Vec. Both models are pre-trained on different data, but partly share a vocabulary from which we extract $\approx 20\text{K}$ words. Using 300 randomly drawn parallel anchor, we convert each embedding space to a relative one. In Table 4.1 (left), we show the original and the relative embeddings. For each word w , we consider its corresponding encodings x and y in the source and target space. We apply three different metrics to measure their similarity (in a setting similar to Vulić, Ruder, and Søgaard, 2020): (i) *Jaccard*: the discrete Jaccard similarity between the set of word neighbors of x in source and target; (ii) *Mean Reciprocal Rank*: measures the (reciprocal) ranking of w among the top- k neighbors of x in the target space; (iii) *Cosine*: measures the cosine similarity between x and y . Additional details in Appendix A.3.1.

Result analysis. Table 4.1 (left) highlights clusters of semantically similar words and shows that the absolute representations are incoherent across the two latent spaces, while the relative embeddings are highly similar. The average Jaccard distance reported in Table 4.1 (right), says that the word neighborhoods of the RRs are matched exactly 34% of the time in one direction, and 39% of the time in the other one (the missing 61% is due to semantic differences, that are not taken into account by the discrete nature of the Jaccard metric). By contrast, the absolute embeddings are never matched exactly (Jaccard score equal to zero); for a match to happen, it would mean that the FastText and Word2Vec embeddings of a given English word are almost the same, which is highly unlikely. MRR, close to a perfect score for the RRs, shows that

TABLE 4.1: Qualitative (*left*) and quantitative (*right*) comparisons of English word embeddings using absolute and RRs. PCA is applied only for visualization. All metrics are calculated with $K = 10$ averaged over 20k words and across 10 different random seeds. See Figure A.4 for other dimensionality reductions, refer to Table A.2 and Figure A.5 for the same experiment on CIFAR-10, showcasing this result also holds on different data modalities.

		FastText	Word2Vec			
Absolute						
						
Relative	FT	FT	W2V	Jaccard \uparrow	MRR \uparrow	Cosine \uparrow
		W2V	W2V	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	W2V	FT	FT	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.00
		W2V	W2V	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.00
FT	FT	W2V	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	
	W2V	W2V	0.34 ± 0.01	0.94 ± 0.00	0.86 ± 0.00	
W2V	FT	FT	0.39 ± 0.00	0.98 ± 0.00	0.86 ± 0.00	
	W2V	W2V	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	

the most-similar word to a given one is usually itself, even if their cosine similarity doesn't reach 1.

Overall, these results show that RRs are preserved across different word embedding models, validating our assumptions.

4.3.2 Latent distance as a performance proxy

Experimental setting. In this experiment, we consider a node classification task on the Cora graph dataset. We first train a *reference* model that achieves good accuracy on a validation set. Then, we train ≈ 2000 models with various combinations of seed, number of epochs, number of layers, dropout probability, activation functions, optimizer type, learning rate or type of graph embedder (refer to Table A.4 for further details). All the models are trained using absolute representations, which are converted to relative post-training by projecting the embeddings onto 300 randomly drawn but fixed anchors. For each model, we measure its classification accuracy and compute the similarity of its space with the reference one. This similarity is computed as the average cosine similarity between the node embeddings produced by a given model and the corresponding embeddings in the reference one.

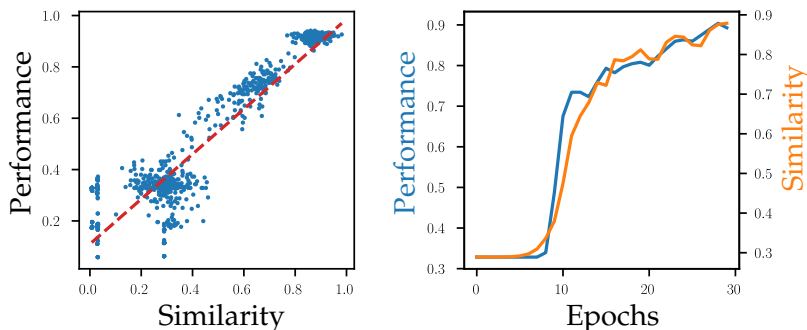


FIGURE 4.3: Graph node classification task on Cora. *Left*: Correlation between the performance of ≈ 2000 models and the similarity of their latent spaces with respect to a well-performing reference model. *Right*: The same correlation plotted over time. The mean Pearson correlation over all models is 0.955, after filtering out the models having the best validation accuracy below 0.5.

Result analysis. The scatter plot in Figure 4.3 (*left*) shows that better-performing models tend to be the ones with the latent spaces most similar to the reference model. The performance-similarity correlation also holds over time, as shown in Figure 4.3 (*right*). Additional correlation examples are in Figure A.3. Interestingly, this metric is differentiable, enabling an explicit supervision signal on the latent space, which does not require labeled data and could be readily exploited in a teacher-student framework.

Overall, these results suggest that the similarity between the RRs of latent spaces is a remarkably good proxy to evaluate model performance.

4.3.3 Training with Absolute vs. Relative representations

Experimental setting. Finally, we compare architectures that do or do not employ the RR while training. In these experiments, the models vary slightly according to the dataset; however, the relative and absolute versions are always comparable in terms of architecture, number of learnable parameters and hyperparameters. We refer to Appendix A.3 and the open-source code for further details on their implementation. In this Section, we consider classification tasks on several datasets, spanning the image domain (Xiao, Rasul, and Vollgraf, 2017; L. Deng, 2012; Krizhevsky, 2009) and the graph domain (Yang, Cohen, and Salakhutdinov, 2016).

TABLE 4.2: Performance comparison between relative and absolute representations on various image and graph datasets. The metric is the classification weighted F1 score (\pm std), over 6 seeds.

	Image Classification				Graph Node Classification		
	MNIST	F-MNIST	CIFAR-10	CIFAR-100	Cora	CiteSeer	PubMed
Relative	97.91 \pm 0.07	90.19 \pm 0.27	87.70 \pm 0.09	66.72 \pm 0.35	0.89 \pm 0.02	0.77 \pm 0.03	0.91 \pm 0.01
Absolute	97.95 \pm 0.10	90.32 \pm 0.21	87.85 \pm 0.06	68.88 \pm 0.14	0.90 \pm 0.01	0.78 \pm 0.03	0.91 \pm 0.01

Result analysis. The results, reported in Table 4.2, show that RRs, when used at training time, are not detrimental to performance in general. This is further shown in Tables 4.3 to 4.6 and A.9 to A.12, where a subset of the results compares the absolute and RRs on a variety of domains, datasets, and tasks. While the information relevant to the machine learning task seems to be preserved, an intriguing future research question is to determine what specific information is lost when infusing specific invariances to unify the representations.

Overall, these results show that RRs are effective when involved in end-to-end training, without significant performance drops in the downstream task.

4.4 Zero-Shot Model Stitching

Hereafter, we showcase the Zero-Shot Stitching, as defined in Section 3.3.1, capabilities of RR across combinations of different stochasticity sources (Figure 4.4 and table 4.3), neural architectures (Tables 4.4 and 4.5) or datasets (Table 4.6). Finally, we present strong real-world applications in NLP (Section 4.4.2) and CV (Section 4.4.3), e.g., zero-shot predictions on novel languages. Refer to Appendix A.3 for additional implementation details.



FIGURE 4.4: Zero-Shot Stitching Reconstruction examples. Each column is a different image, row pairs are different architectures. In each pair, we first report the non-stitched reconstructions, then the stitched ones.

4.4.1 Image Reconstruction

Experimental setting. We perform Zero-Shot Stitching with AEs and VAEs *trained on RRs* end-to-end on several datasets. For each combination of model and dataset, we perform 5 trainings with different seeds, and zero-shot stitch together the resulting encoders and decoders.

Result analysis. In Figure 4.4, the stitched models that employ absolute representations (*Abs.*) produce erroneous predictions, since the latent spaces obtained from distinct trainings are incompatible. Interestingly, although the absolute VAE does not produce compatible latent spaces, it is regularized. As a result, the embeddings produced by the encoders correspond to incorrect but semantically meaningful reconstructions. Instead, VAE based on RRs (*Rel.*) exhibit almost indistinguishable reconstructions between the models trained end-to-end and the stitched ones. Quantitative results are in Table 4.3.

These results support our claim that RRs are empirically invariant to the variation factors ϕ .

TABLE 4.3: Zero-Shot Stitching performance. The MSE (\pm std) between the ground truth \mathbb{X} and the reconstructions is computed over 5 different seeds. Stitching with our RRs yields an error up to two orders of magnitude less than the absolute counterpart.

			MNIST	F-MNIST	CIFAR-10	CIFAR-100	MSE \downarrow
AE	Abs.	Non-Stitch.	0.66 ± 0.02	1.57 ± 0.03	1.94 ± 0.08	2.13 ± 0.08	1.58 ± 0.05
		Stitch.	97.79 ± 2.48	120.54 ± 6.81	86.74 ± 4.37	97.17 ± 3.50	100.56 ± 4.29
	Rel.	Non-Stitch.	1.18 ± 0.02	3.59 ± 0.04	2.83 ± 0.13	3.50 ± 0.08	2.78 ± 0.07
		Stitch.	2.83 ± 0.20	6.37 ± 0.29	5.39 ± 1.18	18.03 ± 12.46	8.16 ± 3.53
VAE	Abs.	Non-Stitch.	1.31 ± 0.04	4.38 ± 0.03	2.68 ± 0.06	3.00 ± 0.03	2.84 ± 0.04
		Stitch.	98.51 ± 1.49	118.96 ± 2.96	69.02 ± 1.54	78.57 ± 1.88	91.27 ± 1.97
	Rel.	Non-Stitch.	2.97 ± 0.14	6.81 ± 0.06	5.18 ± 0.22	5.93 ± 0.14	5.22 ± 0.14
		Stitch.	13.43 ± 6.79	24.03 ± 13.15	11.20 ± 3.15	11.23 ± 2.38	14.97 ± 6.37

4.4.2 Text Classification

In this Section, we show practical examples of the use of parallel anchors (Section 4.2).

Experimental setting. We consider two different text classification settings.

Cross-lingual: given a review, predict the associated star rating, done on multi-lingual data from the Amazon Reviews dataset. Following the original paper, we work on a binarized version of the task, with FScore and MAE as metrics. In Table A.10, we report results on the fine-grained formulation. We adopt four different pre-trained language-specific RoBERTa transformers and evaluate their Zero-Shot Stitching performance on languages never seen by the classifier. We use parallel anchors in two modalities: (i) *Translated:* consider English reviews translated³ into the other languages; (ii) *Wikipedia:* adopt an external corpus, WikiMatrix (Schwenk et al., 2021), providing parallel sentences extracted from Wikipedia.

Cross-architecture: assessed on three different datasets: TREC (coarse), DBpedia, Amazon Reviews (English split). We adopt two different pre-trained BERT transformers, BERT-C and BERT-U, ELECTRA and RoBERTa.

TABLE 4.4: Cross-lingual Zero-Shot Stitching performance comparison. The table reports the mean weighted F1 (\pm std) and MAE on Amazon Reviews coarse-grained, across 5 different seeds.

Decoder	Encoder	Absolute		Relative			
		FScore	MAE	Translated		Wikipedia	
				FScore	MAE	FScore	MAE
en	en	91.54 \pm 0.58	0.08 \pm 0.01	90.06 \pm 0.60	0.10 \pm 0.01	90.45 \pm 0.52	0.10 \pm 0.01
	es	43.67 \pm 1.09	0.56 \pm 0.01	82.78 \pm 0.81	0.17 \pm 0.01	78.53 \pm 0.30	0.21 \pm 0.00
	fr	54.41 \pm 1.61	0.45 \pm 0.02	78.49 \pm 0.66	0.21 \pm 0.01	70.41 \pm 0.57	0.29 \pm 0.01
	ja	48.72 \pm 0.90	0.51 \pm 0.01	65.72 \pm 0.55	0.34 \pm 0.01	66.31 \pm 0.80	0.34 \pm 0.01

TABLE 4.5: Cross-architecture Zero-Shot Stitching performance comparison. The table reports the mean weighted F1 (\pm std) for each dataset, across 5 different seeds.

		TREC	DBpedia	Amazon Reviews	
				Coarse	Fine
				Abs.	Non-Stitch
	Stitch	21.49 \pm 3.64	6.96 \pm 1.46	49.58 \pm 2.95	19.01 \pm 2.04
Rel.	Non-Stitch	88.08 \pm 1.37	97.42 \pm 2.05	85.08 \pm 1.93	48.92 \pm 3.57
	Stitch	75.89 \pm 5.38	80.47 \pm 21.14	72.37 \pm 7.32	33.24 \pm 7.21

Result analysis. Tables 4.4 and 4.5 show for the first time that it is possible to learn to solve a downstream task on a specific language or transformer and perform predictions on another.

Stitching with absolute representations yields performances comparable to random guessing across the board, proving that RRs are a key element for the success of this kind of Zero-Shot Stitching. Moreover, Table 4.4 highlights the robustness that RRs have on the choice of anchors, even when they are noisy (*Translated* case), or their distribution differs from one of the downstream task (*Wikipedia* case), as long as

³We used the =GOOGLETRANSLATE function available in Google Sheets.

their encoding can be handled correctly by the encoder. In our case, the encoder is pre-trained to represent a variety of texts in a specific language, thus, even if Wiki-Matrix has a completely different domain from Amazon Reviews, the transformer still computes a meaningful representation, comparable with those of the reviews. We report in Tables A.9 and A.10 complete results on all languages combination, and in Table A.11 the performance obtained by a multi-lingual transformer; that, to the best of our knowledge, is the only alternative for obtaining compatible representations across languages.

According to these results, RRs show invariance to different architectures and data distribution shifts (e.g., different train languages).

4.4.3 Image Classification

In this Section, we show practical examples of the use of OOD anchors (Section 4.2).

Experimental setting. We consider a classification task on the datasets ImageNet1k and CIFAR-100 with coarse labels (20), and 4 different pre-trained image encoders: three variants of the ViT transformer (ViT-S/16, ViT-B/16 and RViT-B/16) and RexNet.

TABLE 4.6: Zero-Shot Stitching performance comparison with different encoding techniques. The table reports the mean weighted F1 (\pm std) on CIFAR-100 coarse-grained and ImageNet1k, across 5 seeds.

Decoder	Encoder	CIFAR-100		ImageNet1k	
		Absolute	Relative	Absolute	Relative
RexNet	RexNet	82.06 \pm 0.15	80.22 \pm 0.28	73.78 \pm 0.29	72.61 \pm 0.16
	ViT-B/16	-	54.98 \pm 0.44	-	37.39 \pm 0.36
	RViT-B/16	-	53.33 \pm 0.37	-	42.36 \pm 0.36
	ViT-S/16	-	59.82 \pm 0.32	-	43.75 \pm 0.27
ViT-B/16	RexNet	-	76.81 \pm 0.49	-	30.78 \pm 0.81
	ViT-B/16	93.15 \pm 0.05	91.94 \pm 0.10	80.91 \pm 0.29	78.86 \pm 0.33
	RViT-B/16	6.21 \pm 0.33	81.42 \pm 0.38	0.07 \pm 0.05	44.72 \pm 0.57
	ViT-S/16	-	84.29 \pm 0.86	-	48.31 \pm 0.72
RViT-B/16	RexNet	-	79.79 \pm 0.43	-	53.46 \pm 0.68
	ViT-B/16	4.69 \pm 0.07	84.46 \pm 0.19	0.08 \pm 0.04	62.21 \pm 0.54
	RViT-B/16	91.41 \pm 0.09	90.77 \pm 0.16	82.55 \pm 0.30	81.88 \pm 0.16
	ViT-S/16	-	84.66 \pm 0.16	-	61.32 \pm 0.36
ViT-S/16	RexNet	-	75.35 \pm 0.41	-	37.58 \pm 0.44
	ViT-B/16	-	81.23 \pm 0.31	-	50.08 \pm 0.63
	RViT-B/16	-	78.35 \pm 0.69	-	45.45 \pm 1.41
	ViT-S/16	90.07 \pm 0.19	88.85 \pm 0.44	77.73 \pm 0.41	76.36 \pm 0.40

Result analysis. The results in Table 4.6 highlight how the RRs allow stitching modules with different encoding dimensionality, since the decoder receives a RR with guaranteed equal size equal to the number of anchors. Furthermore, the results demonstrate the ability to generalize and perform Zero-Shot Stitching on CIFAR-100, although that data was never seen by the encoder since it is a frozen transformer trained on ImageNet1k. Interestingly, RexNet is the only transformer whose latent dimensionality is higher than the number of anchors, and the biggest drop in stitching performance happens when the decoder is trained on it. This suggests the number of anchors is an important hyperparameter; we refer to Figure A.1 for a more in-depth analysis.

Overall, these results prove that RRs can bridge general purpose encoders and pre-trained task-specific decoders.

Chapter 5

Direct Translation

*Latent Space Translation via Semantic Alignment*¹

In this Chapter, we address the LCP as defined in Chapter 3, incorporating an additional assumption: either T_X or T_Y is the identity function. Referring to Figure 3.1, this means that we assume the transformation $\mathcal{T} \in \mathbb{T}$, that maps the latent manifold embeddings from one space to another, can be directly approximated by some $\hat{\mathcal{T}}$. In the following, we show that good approximations $\hat{\mathcal{T}} \approx \mathcal{T}$ are simpler than previously thought, i.e., at most affine transformations, and can often be estimated using standard, well-understood algebraic procedures with closed-form solutions.

5.1 Introduction

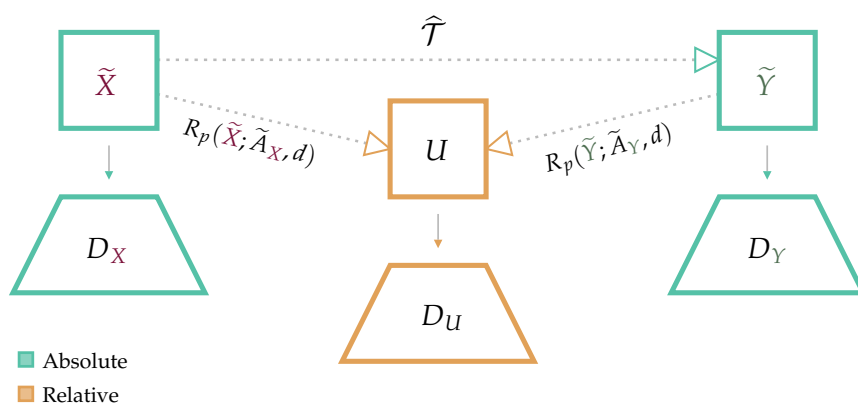


FIGURE 5.1: Zero-shot stitching between absolute spaces utilizing RRs and the method presented in this Chapter (the estimation of $\hat{\mathcal{T}} \approx \mathcal{T}$). The proposed approach does not require a decoder D_U specifically trained on RRs. Instead, we directly translate latent spaces, enabling the use of arbitrarily pre-trained decoders originally trained on absolute spaces, i.e., D_X and D_Y .

One of the key findings from Chapter 4 is the empirical evidence demonstrating that the signal encoded in the angle norms, with respect to a reduced set of data points A , suffices to represent the latent manifold $\varphi_{\tilde{X}}(\mathcal{M}_X)$ embedded within the latent space. This representation is accurate enough to allow downstream performance comparable to using absolute embeddings, in the specific tasks considered.

¹Valentino Maiorca*, Luca Moschella*, Antonio Norelli, Marco Fumero, Francesco Locatello, and Emanuele Rodolà (2023). “Latent Space Translation via Semantic Alignment”. In: *Thirty-seventh Conference on Neural Information Processing Systems*. URL: <https://openreview.net/forum?id=pBa70rGH1r>

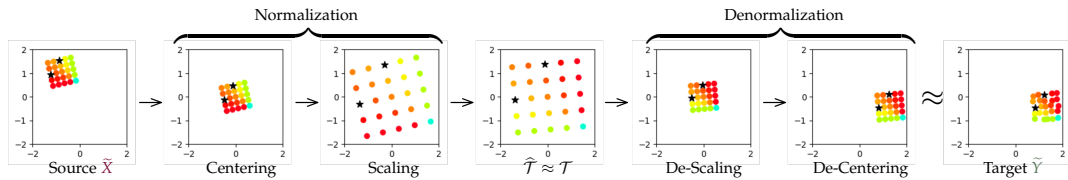


FIGURE 5.2: Method illustration on a synthetic example. Given a source space \tilde{X} , the steps to translate it to a target \tilde{Y} are sequentially applied as described in Section 5.2. Note that the translation is not perfect due to an arbitrary distortion of the data.

Building on this intuition of the existence of a relatively simple transformation relating the latent manifolds, we show the effectiveness and applications of *directly translating between different latent spaces*. Specifically, we show that it is feasible to directly approximate a transformation \mathcal{T} with some $\hat{\mathcal{T}}$, given that a partial (and possibly sparse) correspondence between data points $A_{XY} \subseteq \pi$ is established. Unexpectedly, the process of seamlessly combining different NNs – each pre-trained on different datasets, modalities, architectures, or domains – turns out to be surprisingly straightforward.

For instance, we show how it enables the ability to effectively integrate any pre-trained text encoder with any image classification head, and vice versa; without requiring any additional re-training or assumptions, e.g., without assuming the decoders are trained on RRs as in Chapter 4. The method difference is emphasized in Figure 5.1, Zero-Shot Stitching (Section 3.3.1) with RRs assumes the use of a single decoder specifically trained on a relative space; meanwhile, the method presented in this Chapter allows to zero-shot stitch and reuse decoders originally trained on the absolute spaces.

Our main contributions can be summarized as follows:

- We explore the direct translation between latent spaces of distinct NNs to solve the LCP, as defined in Chapter 3 and illustrated in Figure 3.1. In particular, leveraging a semantic correspondence between the input spaces $A_{XY} \subseteq \pi$, we directly approximate \mathcal{T} for the first time across different trainings, architectures, and modalities. We obtain excellent stitching performances even in cross-modal settings, where we *apply arbitrary text classifiers on top of pre-trained image encodings* (and vice versa).
- We show that different downstream tasks, namely classification and generation, require modeling different transformations to obtain the most out of the translation between their latent spaces.

5.2 Latent Space Translation

5.2.1 Assumptions

In this Chapter, we address the LCP described in Chapter 3 and Figure 3.1, with the additional assumption that either T_X or T_Y is the identity. This means that we are directly trying to approximate $\hat{\mathcal{T}} \approx \mathcal{T} \in \mathbb{T}$. Without loss of generality, we always assume that T_Y is the identity, thus $\hat{\mathcal{T}} \approx T_X = \mathcal{T}$. Furthermore, we assume that $\hat{\mathcal{T}}$ is at most an affine transformation. Please refer to Chapter 3 and Figure 3.1 for a formal definition of the LCP.

5.2.2 Method

Consider two latent spaces, $\tilde{X} \in \mathbb{R}^{n \times d_1}$ and $\tilde{Y} \in \mathbb{R}^{n \times d_2}$. Our objective is to estimate the transformation $\hat{\mathcal{T}} \approx \mathcal{T} \in \mathbb{T}$ that translates $\varphi_{\tilde{X}}(\tilde{\mathcal{M}}_X) \subseteq \tilde{X}$ into $\varphi_{\tilde{Y}}(\tilde{\mathcal{M}}_Y) \subseteq \tilde{Y}$, i.e.: $\varphi_{\tilde{Y}}(\tilde{\mathcal{M}}_Y) = \hat{\mathcal{T}}\varphi_{\tilde{X}}(\tilde{\mathcal{M}}_X)$, exploiting the *semantic alignment* \mathcal{C} observed through $A_{XY} \subseteq \pi$ between the input spaces X and Y .

Throughout this work, we identify two main steps in the translation process: pre-processing the spaces and estimating the transformation $\hat{\mathcal{T}}$, as outlined in Figure 5.2.

Pre-processing. Generally, the two spaces \tilde{X} and \tilde{Y} may have different dimensionalities – in those cases, we zero-pad the smaller one to match the dimension of the other without changing its underlying structure (Williams et al., 2021). Moreover, we standardize each feature to have zero mean and unit variance (standard scaling) if not otherwise specified, whose statistics are computed only on the anchor sets for both source and target space, to perform the necessary denormalization.

Estimating $\hat{\mathcal{T}} \approx \mathcal{T}$. In Chapter 4, it is empirically shown that the spaces often differ by an angle-norm preserving transformation. Nevertheless, we broaden our investigation by considering different ways of obtaining $\hat{\mathcal{T}}$ to evaluate the robustness of that assumption. Throughout our experiments, we primarily operate under the assumption that $\hat{\mathcal{T}}$ can be constrained to encode, at most, an affine transformation: $\hat{\mathcal{T}}(\tilde{X}) = \mathbf{R}\tilde{X} + \mathbf{b}$.

This general formulation, without additional constraints, corresponds to our affine method in the experiments, and it is optimized via gradient descent. The other transformations are trivially obtained by progressively adding constraints on this one:

- **linear.** To model a linear transformation, we can just set the bias term to zero $\mathbf{b} = \vec{0}$ and optimize via Least Square. Here, we are both simplifying the class of transformations and switching from a gradient descent optimization to a closed-form procedure.
- **1-ortho.** Additionally, we could require \mathbf{R} to be orthogonal to encode an isometry. In this case, we obtain this by applying Singular Value Decomposition (SVD) on the corresponding \mathbf{R} obtained by the `linear` solution. Through this, we aim to understand the implications of enforcing orthogonality on a transformation that was originally not constrained to be so, in a setting similar to Xing et al., 2015.
- **ortho.** To obtain the optimal orthogonal \mathbf{R} , we apply Procrustes analysis (Gower, 1975). Please refer to Section 2.3 for further details.

The transformation $\hat{\mathcal{T}}$ is estimated from samples in semantic correspondence $A_{XY} \subseteq \pi$, i.e., the parallel anchors defined in Section 3.2.

This methodology facilitates efficient and precise zero-shot translation between disparate latent spaces, providing a robust and versatile foundation for model reuse and interoperability in diverse machine learning contexts.

5.3 Latent Communication via Translation

In this Section, we evaluate the capabilities and effectiveness of our translation method through various scenarios, highlighting its applicability in diverse contexts.

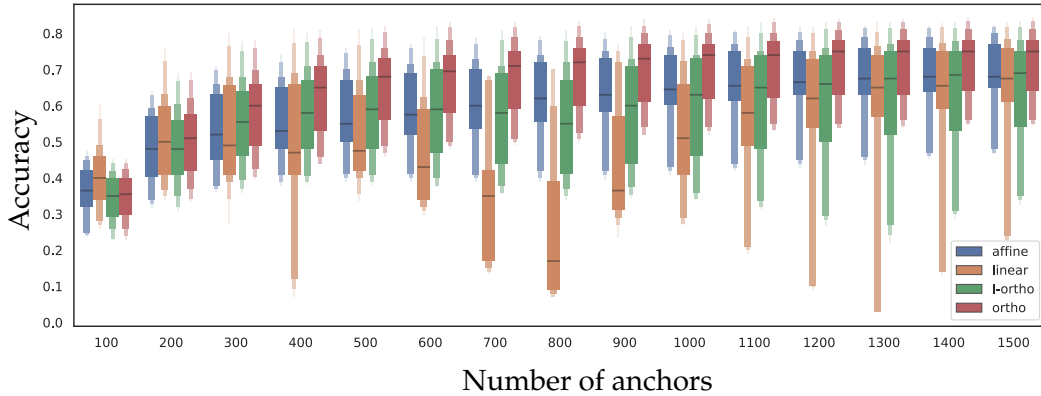


FIGURE 5.3: Performance comparison of affine, linear, 1-ortho, and ortho at varying number of anchors on classification accuracy. Results on CIFAR-100 fine-grained. The same analysis for the generation case is in Figure B.2 in the Appendix.

We present empirical results in three different settings: (i) cross-architecture; (ii) cross-modality; (iii) autoencoding. In each case, the translation performance of each method for obtaining the transformation $\hat{\mathcal{T}}$ is evaluated against two baselines, the naive absolute one and the relative one.

Stitching Procedure. In line with the *Zero-Shot Stitching* concept we introduced in Section 3.3.1, we combine independent encoders and decoders (e.g., classifiers, generators) without further training or fine-tuning. This study does not necessitate a decoder trained on relative representations; instead, we directly employ the original decoders trained on absolute spaces. Each one of the benchmarks we conduct follows the same procedure unless otherwise specified: we measure the mean performance over all the possible combinations of (encoder, decoder) for each test set in different settings:

- *no-stitch*. The end-to-end performance of the decoder applied to the original space it was trained on. This is useful to establish an upper-bound in performances.
- *absolute*. The result of using the encodings without any transformation, we consider this as a probe for any pre-existing compatibility among encodings and, therefore, a lower-bound.
- *translation*. These are the results of the application of our latent translation method, with the estimation of $\hat{\mathcal{T}}$ via affine, linear, 1-ortho and ortho.

In each instance, we use the same parallel anchors A_{XY} , that are uniformly chosen, in a quantity comparable with the dimensionality of the absolute representation.

5.3.1 Cross-Architecture

Firstly, we test our method in a cross-architecture setting, Zero-Shot Stitching together encodings coming from a variety of pre-trained networks and their associated absolute decoders (classifiers). This scenario provides an extensive testing ground for our method and demonstrates its robustness across different architectures. Please refer to Table B.5 in the Appendix for further results on cross-architecture stitching in generation tasks.

TABLE 5.1: Cross-architecture stitching with various methods for estimating $\hat{\mathcal{T}}$ and applying standard scaling. The stitched decoders are SVMs with a linear kernel. 5 runs for each encoder-decoder pair. (C) and (F) next to CIFAR-100 indicate, respectively, coarse-grained and fine-grained. Please refer to the Appendix in Table B.2 for additional results with MLPs as classification heads.

	Dataset	no-stitch	absolute	relative	affine	linear	l-ortho	ortho
Vision	CIFAR-10	0.95 ± 0.03	0.16 ± 0.22	0.80 ± 0.22	0.92 ± 0.05	0.88 ± 0.11	0.90 ± 0.09	0.93 ± 0.04
	CIFAR-100-C	0.85 ± 0.07	0.11 ± 0.21	0.54 ± 0.25	0.78 ± 0.09	0.73 ± 0.16	0.77 ± 0.11	0.81 ± 0.07
	CIFAR-100-F	0.76 ± 0.09	0.07 ± 0.21	0.30 ± 0.24	0.68 ± 0.11	0.62 ± 0.19	0.64 ± 0.16	0.71 ± 0.09
	F-MNIST	0.88 ± 0.01	0.15 ± 0.20	0.63 ± 0.23	0.86 ± 0.01	0.83 ± 0.06	0.82 ± 0.05	0.85 ± 0.02
	MNIST	0.96 ± 0.01	0.15 ± 0.21	0.50 ± 0.22	0.94 ± 0.01	0.89 ± 0.08	0.81 ± 0.11	0.91 ± 0.02
Text	TREC	0.87 ± 0.12	0.20 ± 0.06	0.36 ± 0.13	0.82 ± 0.12	0.74 ± 0.25	0.57 ± 0.25	0.79 ± 0.11
	AG News	0.73 ± 0.09	0.25 ± 0.02	0.39 ± 0.13	0.65 ± 0.08	0.62 ± 0.08	0.61 ± 0.10	0.66 ± 0.10
	DBpedia	0.78 ± 0.23	0.07 ± 0.01	0.16 ± 0.10	0.66 ± 0.24	0.62 ± 0.23	0.57 ± 0.23	0.66 ± 0.22
	IMDB	0.61 ± 0.04	0.50 ± 0.01	0.51 ± 0.02	0.59 ± 0.04	0.57 ± 0.04	0.56 ± 0.03	0.59 ± 0.04

Experimental setting. We consider a variety of Computer Vision (MNIST, F-MNIST (Fashion MNIST), N24News, CIFAR-10, CIFAR-100) and Natural Language Processing (TREC, DBpedia, N24News, AG News, IMDB) datasets. For the text domain we consider 6 different language models as encoders (BERT-C, BERT-U, ELECTRA, RoBERTa, ALBERT, XLM-R, and the text encoder of CLIP), and for the image domain 6 encoders (RexNet, ViT-S/16, ViT-B/16, ViT-B/16L, RViT-B/16, and the image encoder of CLIP), all pre-trained and frozen. The full encoder list can be found in Table B.4 in the Appendix. For each dataset and for each encoder, we train an SVM classification head (decoder) on top of their specific encodings. We then proceed with the standard stitching procedure outlined in Section 5.3 and collect the results. Please see Table B.5 in the Appendix for cross-architecture stitching in generation tasks, where we extend this analysis by verifying that our method works even across autoencoders of different bottleneck sizes.

TABLE 5.2: Cross-architecture stitching with various methods for estimating $\hat{\mathcal{T}}$ and applying L2 normalization. The stitched decoders are SVMs with linear kernel. 5 runs for each encoder-decoder pair. (C) and (F) next to CIFAR-100 indicate, respectively, coarse-grained and fine-grained. Please refer to Table B.3 in the Appendix for additional results with MLPs as classification heads.

	Dataset	no-stitch	absolute	relative	affine	linear	l-ortho	ortho
Vision	CIFAR-10	0.95 ± 0.03	0.16 ± 0.22	0.80 ± 0.22	0.93 ± 0.04	0.78 ± 0.27	0.88 ± 0.12	0.91 ± 0.09
	CIFAR-100-C	0.85 ± 0.07	0.11 ± 0.21	0.54 ± 0.25	0.79 ± 0.07	0.65 ± 0.25	0.73 ± 0.17	0.79 ± 0.10
	CIFAR-100-F	0.76 ± 0.09	0.07 ± 0.21	0.30 ± 0.24	0.69 ± 0.10	0.52 ± 0.25	0.62 ± 0.19	0.68 ± 0.13
	F-MNIST	0.88 ± 0.01	0.15 ± 0.20	0.63 ± 0.23	0.86 ± 0.01	0.65 ± 0.23	0.83 ± 0.06	0.84 ± 0.05
	MNIST	0.96 ± 0.01	0.15 ± 0.21	0.50 ± 0.22	0.94 ± 0.01	0.61 ± 0.23	0.90 ± 0.08	0.90 ± 0.04
Text	TREC	0.87 ± 0.12	0.20 ± 0.06	0.36 ± 0.13	0.82 ± 0.12	0.44 ± 0.20	0.74 ± 0.23	0.77 ± 0.12
	AG News	0.73 ± 0.09	0.25 ± 0.02	0.39 ± 0.13	0.66 ± 0.08	0.56 ± 0.10	0.62 ± 0.08	0.64 ± 0.10
	DBpedia	0.78 ± 0.23	0.07 ± 0.01	0.16 ± 0.10	0.66 ± 0.24	0.44 ± 0.20	0.62 ± 0.23	0.60 ± 0.22
	IMDB	0.61 ± 0.04	0.50 ± 0.01	0.51 ± 0.02	0.59 ± 0.04	0.55 ± 0.03	0.58 ± 0.04	0.59 ± 0.04

Result analysis. The stitching results are in Table 5.1. As expected, the *absolute* encodings obtain a score comparable to random guessing while also considering fewer encoder combinations out of the possible ones due to the dimensionality mismatch between some of them. These results show that the transformation relating to these

pre-trained encoders is indeed mostly orthogonal: (i) ortho and affine, the narrowest and the broadest transformation classes considered, are the better-performing translation methods. But while the former is obtained via a simple and efficient closed-form algorithm, the latter is SGD-optimized (Section 5.2.2). (ii) the 1-ortho version improves or has small drops in performances over the linear transformation it is obtained from, confirming that the least squares procedure converges to an \mathbf{R} which is almost orthogonal. Note that these results demonstrate the feasibility of combining pre-trained models without the need for retraining or fine-tuning, with negligible drops in performances across the board, and without any additional assumption on the decoders. Please refer to Tables B.2 and B.3 in the Appendix for results with different decoders. In the Appendix (Figure B.1), we extend the cross-architecture transfer to decoders trained on different domains (styles) of the same CIFAR-10 dataset: the original one and a grayscale one.

Sensibility to Anchor Quantity. The number of anchors is an essential parameter in our approach. In Figure 5.3, we evaluate how the quantity of these anchors impacts the residual error and the overall performance of our method for this experimental setting. This analysis offers insights into the optimal number of anchors necessary for efficient latent space translation.

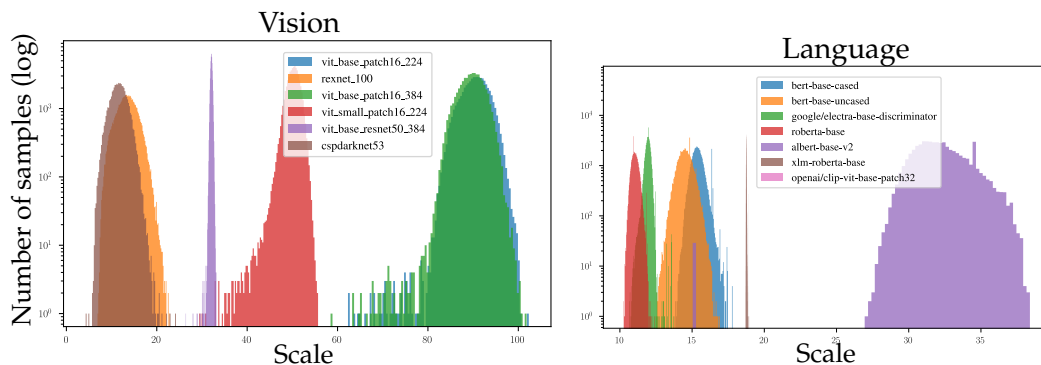


FIGURE 5.4: Scale distribution in encodings of different pre-trained encoders on the N24News dataset.

Role of Scaling. Our approach is designed to accommodate generic (re)scaling methods as pre-processing steps. We advocate for the use of standard scaling, as it shows reliable performance in our experiments, indicating that the scale of the data points is useful in estimating the latent transformation $\hat{\mathcal{T}}$.

However, for completeness, we also consider L2 normalization, which is the standard normalization in RRs. This normalization method generalizes the class of transformations handled by our method and introduces an element of complete scale invariance. It is important to note that when this level of generalization is introduced, a scale-invariant decoder is required, since the norm information is effectively removed. In Chapter 4, this is implicitly accomplished by training a decoder on RRs. In our setting, since we do not train the decoder, we just assume it is scale invariant; in Appendix B.1.1 we elaborate why this is a reasonable assumption that happens in practice.

This investigation exemplifies the flexibility of our approach, capable of adapting to different normalization and pre-processing strategies based on the specific requirements of the task at hand. The results presented in Table 5.2, when compared with Table 5.1, indicate a stronger reliance on the information encoded in the norm in the

text modality. This is aligned with existing literature in the NLP domain (Oyama, Yokoi, and Shimodaira, 2022), which suggests that the scale of the encodings contains information (e.g., it is correlated with the token frequency).

These results in diverse scenarios showcase the flexibility and adaptability of our method, especially its robustness in translating between latent spaces of different dimensionality and domains.

5.3.2 Cross-Modality

This scenario illustrates the applicability of our method in cross-modality settings, where we aim to translate between text and image latent spaces.

Experimental setting. We adopt N24News, a multimodal news classification dataset that contains both text and associated pictures. We apply the standard encoding procedure to these two features separately, using different pre-trained uni-modal encoders. Then, we train a classification head (an SVM, please refer to Appendix Figure B.3 for further results employing an MLP as classification head) on top of each one. Lastly, we zero-shot stitch each encoder with a classification head different from its corresponding one, measuring its classification accuracy, without further training or fine-tuning.

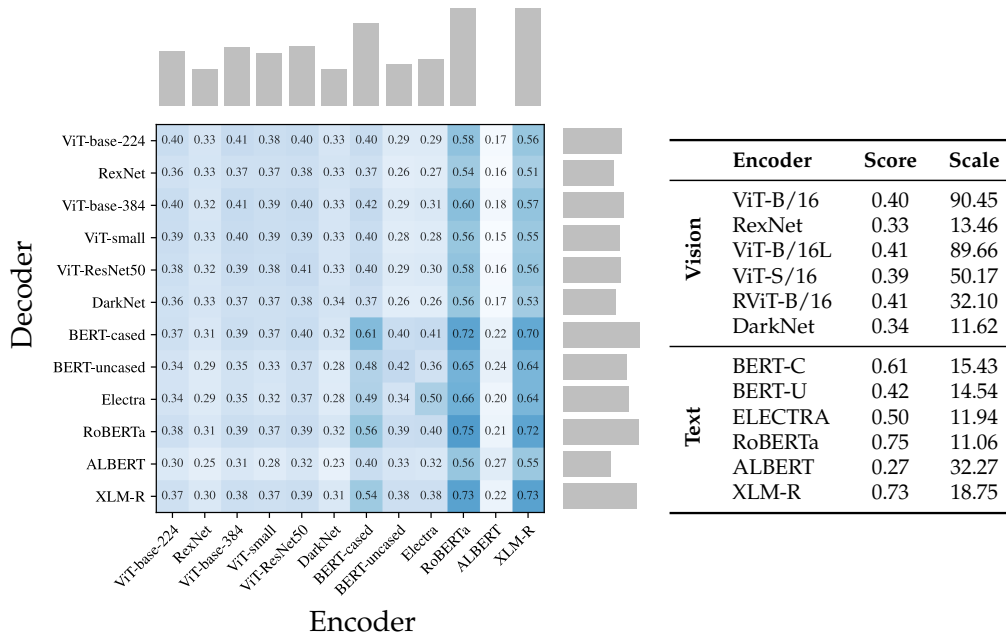


FIGURE 5.5: Performance comparison between different encoders and data modalities on the N24News multimodal dataset. On the right the accuracy of models trained end-to-end on a single data modality (Score) and their average norm (Scale). On the left the stitching performance between pairs of encoders and decoder. This shows the importance of translating from good encoders, that can even improve unimodal decoder performances. Results obtained with 2000 anchors and ortho, with an SVM as classification head. In the Appendix Figure B.3, additional results using MLPs as decoders.

Scale distributions. In Figure 5.4, we present the scale distribution of the embeddings produced by several encoders on the N24News dataset. This empirical analysis

shows a consistent pattern among encoders: the scale distribution of their embeddings follows a Gaussian one with a single mode and a well-defined mean, which are usually compatible with standard scaling. This consistent behavior across encoders is likely attributed to their architectural choices, such as the normalization techniques, regularizations and the optimization problems they are designed to solve.

Result analysis. The discrepancy in the mean accuracy represented by the marginal bar plots in Figure 5.5 is a signal that can be used to identify spaces more suited to be *decoded into* and the ones that are stronger in *encoding from*. In fact, the language models as source space for the translation exhibit stronger performance than the vision encoders. We relate this behavior to the higher generality of the text domain data used during pre-training with respect to the image domain one (Zhai et al., 2022a). A remarkable finding in this setting is the improvement in classification performance when a modality-specific classifier trained on images is fed zero-shot with corresponding text encodings translated to the image domain via our method. This result underlines the significance of a good encoder and demonstrates the broad applicability of our technique. In practice, this means we can seamlessly apply image classifiers on textual data, and vice versa.

These results show that our method: (i) obtains effective zero-shot translation over different modalities; (ii) improves unimodal decoders when translating from a better encoder than the one it was trained on.

5.3.3 Autoencoding

In this setting, our method is applied to align latent spaces of different trainings of the same AE. The novelty of this scenario lies in the generation setting itself, as most prior works (Section 2.3) primarily focus on classification tasks. One key observation explored in Chapter 7 is that the *task* at hand (e.g., classification, generation) defines a certain *class of transformations* \mathbb{T} (e.g., rotations) which act among the latent spaces. To ensure the best possible performance and efficiency, it is essential to limit the search for the transformation to the appropriate class.

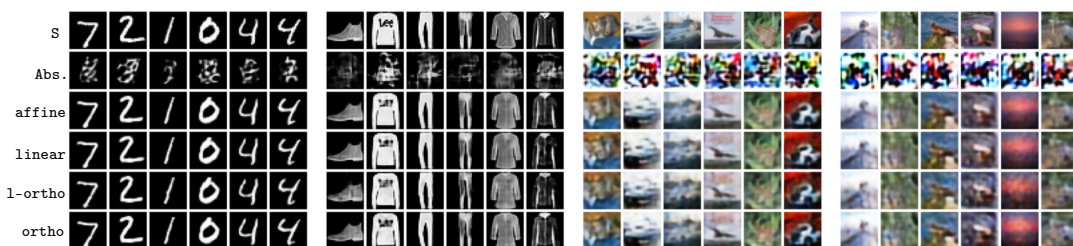


FIGURE 5.6: Reconstruction examples grouped by dataset. Each column is a different image, from top to bottom: original image, absolute stitching, affine stitching, linear stitching, 1-ortho stitching, and ortho stitching. No additional normalization applied on the decoder part. Please refer to Figures B.4 and B.5 in the Appendix for decoders trained with L2 normalization.

Experimental setting. We utilize four datasets for these experiments, namely MNIST, F-MNIST, CIFAR-10 and CIFAR-100. For each dataset, we train two standard CNN-based AE, with convolutions in the encoder and deconvolutions in the decoder, please refer to the Appendix for further implementation details. The two AEs are identical

TABLE 5.3: Zero-shot stitching for generation with various methods for estimating $\hat{\mathcal{T}}$. The representation is normalized using Standard Scaling, and no additional normalization is applied to the stitched decoders. We report the latent cosine similarity (*lcos*) and MSE (*lmse*) between the target encoding and the translated one, but also the reconstruction MSE (*rmse*) between the input and the output. The absolute space dimension is 500, and we used 1000 anchors. Please refer to Table B.1 for results on decoders scale-invariant by design (with L2 normalization on the encodings).

	MNIST			F-MNIST			CIFAR-10			CIFAR-100		
	<i>lcos</i>	<i>lmse</i>	<i>rmse</i>	<i>lcos</i>	<i>lmse</i>	<i>rmse</i>	<i>lcos</i>	<i>lmse</i>	<i>rmse</i>	<i>lcos</i>	<i>lmse</i>	<i>rmse</i>
absolute	0.09	0.27	0.14	0.17	0.23	0.23	0.30	0.29	0.34	0.34	0.53	0.40
affine	0.94	0.08	0.02	0.94	0.06	0.03	0.96	0.03	0.05	0.96	0.04	0.05
linear	0.92	0.09	0.02	0.93	0.07	0.04	0.94	0.03	0.05	0.94	0.04	0.06
1-ortho	0.79	0.14	0.02	0.78	0.12	0.05	0.85	0.05	0.06	0.84	0.07	0.07
ortho	0.90	0.10	0.02	0.90	0.08	0.04	0.94	0.03	0.06	0.93	0.04	0.06

in structure, differing only in the random seed used for weight initialization and data shuffling. To perform Zero-Shot Stitching, we first translate each data point from the latent space of the first encoder to the latent space of the second using 1000 parallel anchors. We then apply the second decoder to the translated data, without any additional training or fine-tuning.

Result analysis. This experiment analyzes the alignment of latent spaces in different training regimens of the same AE. The performance evaluation, as shown in Table 5.3, demonstrates that all methods *affine*, *linear*, *1-ortho*, and *ortho* yield satisfactory results. Moreover, qualitative results depicted in Figure 5.6 reveals minimal visual differences in the stitching outcomes across various datasets using different methods. Please refer to Figures B.4 and B.5 for other qualitative results. In fact, these results suggest that the latent spaces of image AEs are not exclusively correlated by orthogonal transformations. Consequently, in order to constrain and improve their approximation, more research is necessary to investigate and model the particular class of transformations that control the correlation between NNs during image autoencoding. For additional results pertaining to decoders with L2 normalization on their input, we refer to the Table B.1 in the Appendix.

Overall these results, combined with Cannistraci, **Moschella**, Fumero, et al., 2024 presented in Chapter 7 and Section 5.3.1, confirm that latent spaces in image AEs trained end-to-end are related by a class of transformations larger than orthogonal transformations.

Part III

Overcoming Limitations in Latent Communication

Chapter 6

Current limitations

The methodologies explored in Chapters 4 and 5 have demonstrated significant potential in addressing the LCP (Latent Communication Problem) illustrated in Figure 3.1 and detailed in Chapter 3. Despite these advancements, there exist major constraints within these approaches that merit further discussion.

Assumptions on the transformation class \mathbb{T} . The approaches delineated in Chapters 4 and 5 presuppose a *known* transformation class \mathbb{T} between latent manifold embeddings $\varphi_{\tilde{X}}(\tilde{\mathcal{M}}_X) \subseteq \tilde{X}$ and $\varphi_{\tilde{Y}}(\tilde{\mathcal{M}}_Y) \subseteq \tilde{Y}$. Specifically, Chapter 4 assumes that \mathbb{T} comprises transformations preserving angle norms, whereas Chapter 5 assumes it to be simple, exploring different possibilities (i.e., either affine, linear or orthogonal). However, this assumption does not always align with practical scenarios. Indeed, the latent manifolds embeddings are subject to changes due to several factors ϕ , as explained in Section 3.2.1, and the precise nature of \mathbb{T} connecting these embeddings often remains undetermined a priori.

Observable partial correspondence π . Both methodologies assume that a partial correspondence π between the input spaces exists and, most importantly, that it is at least partially observable through the parallel anchors $A_{XY} \subseteq \pi$. This premise, however, is not universally applicable, as the parallel anchors A_{XY} are typically not available in large quantities. The assumption that A_{XY} is sufficiently large to define RRs without losing information, in Chapter 4, and to accurately estimate the transformation $\hat{\mathcal{T}} \approx \mathcal{T}$, in Chapter 5, does not hold in many practical instances. This limitation is particularly evident in multimodal data contexts. Here, the available partial correspondence A_{XY} often falls short of the threshold necessary for the effective application of these methodologies, especially when considering domains different from images-text pairs.

In the following Chapters, we will explore methods to overcome these limitations. In Chapter 7, we introduce a novel approach to tackle the LCP without any specific assumption on the transformation class \mathbb{T} . Meanwhile, in Chapter 8 we delineate a methodology capable of discovering new parallel anchors from a limited known set, thereby expanding $A_{XY} \subseteq \pi$ and facilitating the communication between these spaces.

Chapter 7

Unknown Latent Transformation

*From Bricks to Bridges: Product of Invariances to Enhance Latent Space Communication*¹

In this Chapter, we address the LCP outlined in Chapter 3, without imposing any additional explicit assumptions on the transformation class \mathbb{T} that connects the latent manifold embeddings, $\varphi_{\tilde{X}}(\tilde{\mathcal{M}}_X) \subseteq \tilde{X}$ and $\varphi_{\tilde{Y}}(\tilde{\mathcal{M}}_Y) \subseteq \tilde{Y}$. Leveraging the RRs framework introduced in Chapter 4, we define both T_X and T_Y as multiple relative projections, each characterized by distinct similarity functions d . This approach enables the construction of a product space of invariant components, obviating the need for pre-existing knowledge about the specific invariances to be incorporated.

7.1 Introduction

Achieving invariance to specific groups of transformations within latent spaces is at the core of solving the LCP. In fact, the RRs framework presented in Chapter 4 enables communication between latent spaces by infusing an invariance to angle-norm preserving transformations in them. However, as shown in Figure 7.1, the transformations relating different latent manifold embeddings are not always consistently within a specific class of transformations \mathbb{T} . Determining a priori which \mathbb{T} relates distinct latent manifold embeddings is challenging due to complex interactions in the data, and multiple nuisance factors that are typically irrelevant but can nevertheless affect the representation, i.e., the factors ϕ outlined in Section 3.2.1.

To address this challenge, we expand upon the method of RR, presenting a framework to *efficiently incorporate a set of invariances into the learned latent space*. This is achieved by constructing a product space of invariant components on top of the latent representations of, possibly pretrained, neural models. Each component of this product space is a RR produced with a different similarity function d . Thus, we can infuse invariances to specific transformation classes into each component of the product space.

Our main contributions can be summarized as follows:

- We show that the transformation class \mathbb{T} that relates latent manifold embeddings learned by distinct NNs – trained on data semantically related by π – may vary and directly depends on the factors ϕ .

¹Irene Cannistraci, **Luca Moschella**, Marco Fumero, Valentino Maiorca, and Emanuele Rodolà (2024). “From Bricks to Bridges: Product of Invariances to Enhance Latent Space Communication”. In: *The Twelfth International Conference on Learning Representations (ICLR 2024, spotlight, top 5%)*. URL: <https://openreview.net/forum?id=vngVydDWft>

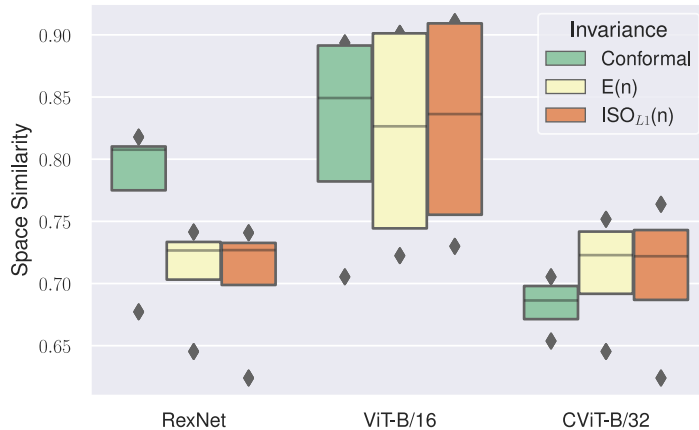


FIGURE 7.1: CKA similarity (Kornblith et al., 2019) of pretrained models on F-MNIST, measured on projections of the latent space onto subspaces invariant to specific classes of transformations (Conformal², Euclidean, Orthogonal). In each bar, we report the distribution of similarity to the other models while infusing a specific invariance. The score diversity highlights the absence of a universal transformation connecting all latent spaces.

- We introduce a framework to construct a product space of invariant components and improve the LCP solution proposed in Chapter 4; achieving the best performance without any prior knowledge of the transformation class \mathbb{T} or the factors ϕ that may affect it.
- We validate our findings on classification and reconstruction tasks, observing consistent latent similarity and downstream performance improvements in the Zero-Shot Stitching setting (Section 3.3.1). The experimental analysis comprises three modalities (vision, text, and graphs), twelve pretrained foundational models, eight benchmarks, and several architectures trained from scratch.

7.2 Infusing invariances

In this Chapter, our focus is to leverage different choices of the similarity function d in the RR framework to induce a *set of invariances* into the representations to capture complex transformations between latent spaces. Meanwhile, in Chapter 4, d was always the cosine similarity, inducing a RR invariant to angle-norm preserving transformations.

Overview. As illustrated in Figure 3.1 and Chapter 3, when considering different encoders E_X, E_Y , we are interested in modeling the class of transformations \mathbb{T} that relates their latent manifold embeddings $\varphi_{\tilde{X}}(\tilde{\mathcal{M}}_X) \subseteq \tilde{X}$ and $\varphi_{\tilde{Y}}(\tilde{\mathcal{M}}_Y) \subseteq \tilde{Y}$. In general, this transformation can be induced by any change in the factors ϕ which could affect $\varphi_{\tilde{X}}(\tilde{\mathcal{M}}_X)$ and $\varphi_{\tilde{Y}}(\tilde{\mathcal{M}}_Y)$ in unpredictable ways, as observed in Figure 7.1, e.g., their training dynamics, by architectural changes, or even domain changes. The resulting \mathbb{T} could be something known, e.g., rotations, or a nontrivial, complex class of transformations.

²More precisely, global orthogonal transformations composed with local rescalings.

7.2.1 Method

What we look for are the transformations T_X and T_Y , as per Figure 3.1, that independently projects the latent manifold embeddings $\varphi_{\tilde{X}}(\tilde{\mathcal{M}}_X)$ and $\varphi_{\tilde{Y}}(\tilde{\mathcal{M}}_Y)$ into a universal space U , where they become the same $T_X(\varphi_{\tilde{X}}(\tilde{\mathcal{M}}_X)) = T_Y(\varphi_{\tilde{Y}}(\tilde{\mathcal{M}}_Y)) \subseteq U$. This is achieved by independently enforcing an *invariance to \mathbb{T}* in each space, i.e.,

$$T_X(\tilde{x}) = T_X(\mathcal{T}\tilde{x}) \quad \forall \mathcal{T} \in \mathbb{T} \quad \text{and} \quad \forall \tilde{x} \in \varphi_{\tilde{X}}(\tilde{\mathcal{M}}_X), \quad (7.1)$$

and similarly for T_Y . Generalizing the RRs to arbitrary similarity functions d , or distance metrics, gives us a straightforward way to define representations invariant to specific classes of transformations.

However, \mathbb{T} is typically unknown a priori, and it is also improbable to accurately characterize it as a singular, well-defined class of transformations (as observed in Figure 7.1 and Section 7.3.1). To overcome this, we approximate U with a product space $\hat{U} := \prod_{i=1}^N U_i$, where each component is obtained by projecting samples of \tilde{X} and \tilde{Y} in a RR equipped with a different similarity function d_i . Each U_i will have properties induced by a similarity function d_i invariant to a specific, known, class of transformations \mathbb{T}_i (e.g., dilations). By combining this set of invariances, we want to construct T_X and T_Y such that they are invariant to \mathbb{T} . We define T_X , and equivalently T_Y , formally as the *product projection* from \tilde{X} to \hat{U} :

Definition (Product projection). *Given a latent space \tilde{X} produced by some encoding function E_X , a set of encoded anchors \tilde{A}_X , the definition of relative projection R_p in Equation (4.2), and a set of similarity functions \mathcal{D} each one invariant to a specific known class of transformations \mathbb{T}_i (e.g., rotations), i.e., $R_p(\tilde{x}; \tilde{A}_X, d_i) = R_p(\mathcal{T}\tilde{x}; \mathcal{T}\tilde{A}_X, d_i) \forall \mathcal{T} \in \mathbb{T}_i$. We define T_X as a product projection:*

$$T_X(\tilde{x}) = \rho \circ R_p(\tilde{x}; \tilde{A}_X, d_i) \quad \forall d_i \in \mathcal{D} \quad (7.2)$$

where ρ is an aggregation function (further details in Section 7.2.2) responsible for merging the relative spaces induced by each $d_i \in \mathcal{D}$.

Distance-induced invariances. We leverage the RR framework considering the following similarity functions d : Cos. (Cosine), Eucl. (Euclidean), L_1 , and L_∞ , each one inducing invariances to a specific, known class of transformations. In Table 7.1, we summarize the invariances guaranteed by different distance metrics concerning the following standard classes of transformations: IS (Isotropic Scaling), OT (Orthogonal Transformation), TR (Translation), PT (Permutation), AT (Affine Transformation) and LT (Linear Transformation).

TABLE 7.1: Invariances summary. Overview of the different distance-induced invariances.

Similarity	IS	OT	TR	PT	AT	LT
Absolute	×	×	×	×	×	×
Cos.	✓	✓	×	✓	×	×
Eucl.	×	✓	✓	✓	×	×
L_1	×	×	✓	✓	×	×
L_∞	×	×	✓	✓	×	×

Note how, in general, it is not straightforward to characterize the set of invariances induced by a similarity function. For example, the L_∞ distance does not enforce

isometry invariance (in the L_2 sense of rigidity) in the representation, but induces an invariance to perturbations in dimensions that are not the maximum one. Formalizing and analyzing such types of invariances presents challenges since these transformations cannot be neatly classified into a specific simple class of transformations.

7.2.2 Aggregation functions.

This section summarizes the different aggregation strategies ρ to construct the product space \hat{U} :

- *Concatenation* (Concat): the subspaces are independently normalized and concatenated.
- *Aggregation by sum* (MLP+Sum): similar to DeepSet (Y. Zhang, Hare, and Prugel-Bennett, 2019), the subspaces are independently normalized and non-linearly projected. The resulting components are summed.
- *Self-attention* (SelfAttention): the subspaces are independently normalized and aggregated via a self-attention layer.

When not specified, all the results are obtained using the *Aggregation by sum* strategy. For the implementation details of each strategy, please refer to Cannistraci, Moschella, Fumero, et al., 2024.

The product space \hat{U} is a *robust* latent representation, made of *invariant* components which are combined to capture *nontrivial, complex* transformations, improving LCP solutions and boosting the performance on downstream tasks.

7.3 Experiments

In this Section, we perform qualitative and quantitative experiments to analyze the effectiveness of our framework in constructing representations invariant to complex \mathbb{T} . Specifically, Section 7.3.1 provides empirical motivation, implicitly analyzing the transformations classes that emerge between different pretrained models on multiple datasets and modalities (i.e., vision and text); meanwhile, Section 7.3.2 evaluates the Zero-Shot Stitching performance of our framework across text, vision, and graph modalities; finally, Section 7.3.3 examines attention weights and their role in selecting the optimal relative subspace. Refer to Cannistraci, Moschella, Fumero, et al., 2024 for further experiments and details.

7.3.1 Latent space analysis

Experimental setting. In this Section, we analyze the similarity of latent spaces produced by pretrained foundational models in both the vision and text domains. For the vision domain, we evaluated five distinct foundational models (either convolutional or transformer-based) using the CIFAR-10, CIFAR-100, MNIST, and F-MNIST datasets. Meanwhile, in the text domain, we assessed seven different foundational models using the DBpedia, TREC (coarse), and N24News (Text) datasets.

Result Analysis. In Figure 7.2, we report the Linear CKA correlations for various models and datasets for vision (*left*) and text (*right*) modalities. This analysis highlights the absence of a universally shared transformation class that connects latent spaces of foundation models across distinct conditions. For example, on CIFAR-10

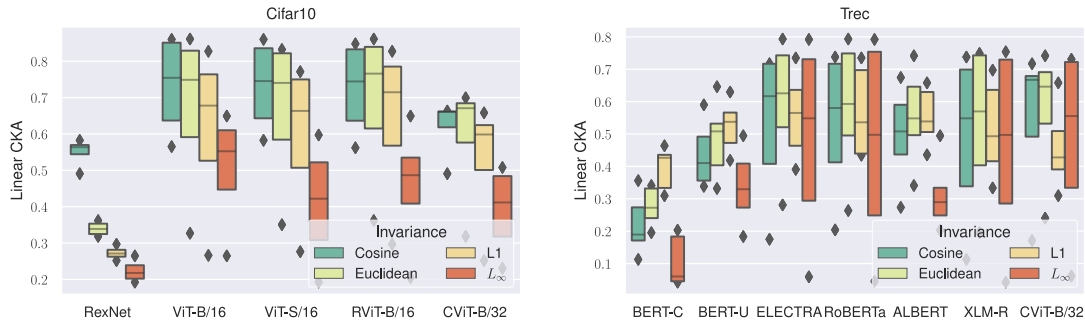


FIGURE 7.2: Latent Spaces Cross-Architecture Similarity. Linear CKA similarity of latent spaces across several pretrained models and datasets. In each bar, we report the space similarities distribution to the other models while infusing a specific invariance. There is no singular projection that consistently outperforms others across all configurations.

(left), the highest similarity is achieved with different projection functions when using different architectures. Interestingly, from Figure 7.2, it is possible to see that similar architectures (i.e., ViT-based models) exhibit similar trends; hinting at the possibility that the choice of architecture plays a major role in influencing the emerging transformation class \mathbb{T} .

Takeaway. The transformation class \mathbb{T} that correlates different latent manifold embeddings produced by different models depends on the dataset, architecture, modality, and possibly other factors ϕ .

7.3.2 Zero-Shot Stitching

The Zero-Shot Stitching (Section 3.3.1) methodology allows combining components of different NNs to obtain a new model, where each element of the stitched model functions as an autonomous frozen module: the encoder handles data embedding, while the dedicated relative decoder manages the downstream task.

TABLE 7.2: Graph and Text Stitching Performance. Zero-shot accuracy scores across various decoders, seeds, and datasets. In the text domain, results are obtained from stitching across pretrained models, while in the graph domain, we train GNN (Graph Neural Network) models from scratch and evaluate the stitching across seeds. Using the *Aggregation by sum* (last row) we consistently achieve the best performance.

Projection	Text		Graph
	ALBERT		GNN
	DBpedia	Cora	Cora
Cosine	0.50 ± 0.02	0.54 ± 0.03	0.53 ± 0.06
Euclidean	0.50 ± 0.00	0.60 ± 0.03	0.27 ± 0.06
L_1	0.52 ± 0.01	0.65 ± 0.02	0.26 ± 0.06
L_∞	0.18 ± 0.02	0.29 ± 0.06	0.12 ± 0.03
Cosine, Euclidean, L_1 , L_∞	0.53 ± 0.01	0.65 ± 0.02	0.77 ± 0.01

Experimental setting. We perform Zero-Shot Stitching classification using text, vision, and graph modalities with various models and datasets. For the *Vision* and *Text* domains, we used the same datasets and pretrained models employed in Section 7.3.1. For the *Graph* domain, we employed the Cora dataset (Sen et al., 2008) and a GNN architecture trained from scratch. Relative decoders are trained with three different seed values, and the resulting representations are transformed into RRs by projecting the encodings onto 1280 randomly selected but fixed anchors.

TABLE 7.3: Image Stitching Performance Cross-Architecture and Cross-Seed. Zero-shot accuracy score across different pretrained models, seeds, and datasets. The proposed method with *Aggregation by sum* consistently achieves the highest accuracy score or comparable results, even without prior knowledge of the optimal projection to employ.

Encoder	Projection	Accuracy \uparrow			
		CIFAR-100	CIFAR-10	MNIST	F-MNIST
CLIP	Cosine	0.52 \pm 0.03	0.87 \pm 0.02	0.61 \pm 0.06	0.68 \pm 0.02
	Euclidean	0.53 \pm 0.02	0.87 \pm 0.02	0.66 \pm 0.05	0.70 \pm 0.03
	L_1	0.53 \pm 0.04	0.87 \pm 0.02	0.66 \pm 0.05	0.70 \pm 0.03
	L_∞	0.27 \pm 0.04	0.52 \pm 0.04	0.57 \pm 0.03	0.55 \pm 0.01
	Cosine, Euclidean, L_1 , L_∞	0.58 \pm 0.03	0.88 \pm 0.02	0.68 \pm 0.05	0.70 \pm 0.01
RViT-B/16	Cosine	0.79 \pm 0.03	0.94 \pm 0.01	0.69 \pm 0.04	0.76 \pm 0.03
	Euclidean	0.79 \pm 0.03	0.94 \pm 0.01	0.71 \pm 0.04	0.77 \pm 0.03
	L_1	0.77 \pm 0.04	0.95 \pm 0.01	0.71 \pm 0.04	0.79 \pm 0.03
	L_∞	0.31 \pm 0.03	0.75 \pm 0.04	0.61 \pm 0.05	0.60 \pm 0.03
	Cosine, Euclidean, L_1 , L_∞	0.81 \pm 0.04	0.95 \pm 0.01	0.72 \pm 0.04	0.76 \pm 0.04
ViT-B/16	Cosine	0.75 \pm 0.05	0.96 \pm 0.01	0.59 \pm 0.05	0.79 \pm 0.03
	Euclidean	0.76 \pm 0.05	0.96 \pm 0.01	0.65 \pm 0.06	0.81 \pm 0.02
	L_1	0.76 \pm 0.06	0.96 \pm 0.01	0.66 \pm 0.07	0.81 \pm 0.02
	L_∞	0.42 \pm 0.02	0.70 \pm 0.05	0.42 \pm 0.05	0.52 \pm 0.04
	Cosine, Euclidean, L_1 , L_∞	0.81 \pm 0.05	0.96 \pm 0.01	0.66 \pm 0.04	0.80 \pm 0.04

TABLE 7.4: Stitching Index Across Architectures and Seeds on Cora. Composing different projections using the *Aggregation by sum* (last row) enables Zero-Shot Stitching *without* any performance drop in this setting, ensuring competitive end-to-end performance.

Projection	Accuracy \uparrow	Stitching index \uparrow
Absolute	0.14 \pm 0.04	0.18
Cosine	0.53 \pm 0.06	0.71
Euclidean	0.27 \pm 0.06	0.58
L_1	0.26 \pm 0.06	0.58
L_∞	0.12 \pm 0.03	1.00
Cosine, Euclidean, L_1 , L_∞	0.77 \pm 0.01	1.00

Results Analysis. Tables 7.2 and 7.3 present the performance of various projection functions for different modalities. As previously observed in Section 7.3.1, the experiments reveal the absence of a single optimal projection function across architectures, modalities, and even within individual datasets. Our proposed method consistently

achieves superior accuracy across most scenarios. It is important to emphasize that the dimensionality of each independent projection and the aggregated product space remains constant, ensuring fair comparison.

To compare the performance with the end-to-end reference (reported in Cannistraci, Moschella, Fumero, et al., 2024), we also propose an additional evaluation metric named the *Stitching Index* computed as the ratio between the stitching score and the end-to-end score. It measures how closely the stitching accuracy aligns with the original score, i.e., a stitching score of one indicates there is no drop in performance when stitching modules. Results in Table 7.4 highlight that our method enables Zero-Shot Stitching *without* any performance drop in this setting, while still ensuring competitive end-to-end performance.

Takeaway. A product space with invariant components can improve the Zero-Shot Stitching performance, solving the LCP defined in Chapter 3, without any prior knowledge of the class of transformation \mathbb{T} that relates the manifold embeddings.

7.3.3 Subspace selection

In the preceding sections, we discussed integrating individual and multiple invariances into the representation through various projection functions and appropriate aggregation strategies. In this Section, we aim to analyze and understand if tuning only the aggregation strategy at stitching time is a reasonable cost for selecting the optimal subspace. We focus on the *Self-attention* aggregation, which is a single self-attention layer as described in Section 7.2.2, and fine-tune only the QKV (Query, Key, Value) parameters (i.e., the ones responsible for subspace blending). Each subspace is generated by its own projection function. We remark that stitching-time fine-tuning is exclusive to this experiment.

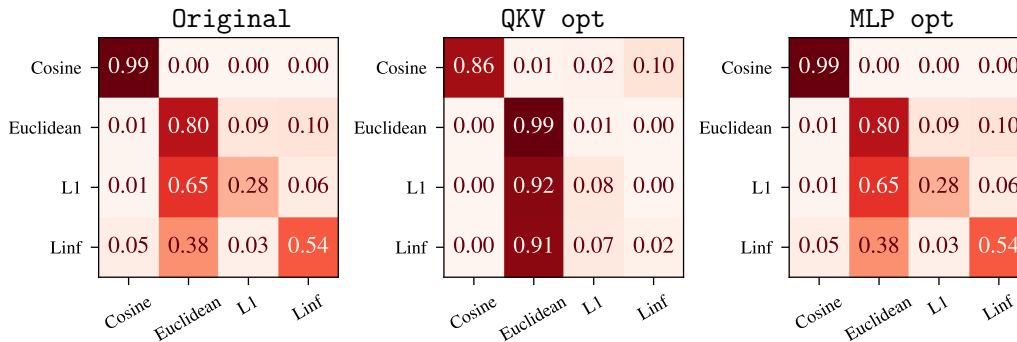


FIGURE 7.3: Comparison of attention weights for the stitched model with RexNet as encoder and ViT-B/16 as decoder on CIFAR-100, before and after fine-tuning. (*left*) the attention weights of the initial zero-shot stitched model, which remain unchanged (*right*) when fine-tuning the classifier (MLP opt). Conversely, fine-tuning the QKV projections (QKV opt) leads to a notable shift in attention weights (*center*), assigning lower importance to the subspace that performs worst individually.

Experimental setup. We identify two crucial components within the stitched model: (i) the linear projections associated with QKV in the attention mechanism, which is responsible for selecting and blending subspaces, and (ii) the MLP in the classification head following the attention mechanism, which classifies the aggregated embeddings. We examine two distinct approaches: the first approach fine-tunes only the first

component (QKV opt), while the second one fine-tunes the second component (MLP opt). All the experiments in this Section are conducted on the CIFAR-100 dataset, employing the RexNet as encoder and the ViT-B/16 as decoder.

TABLE 7.5: Classification accuracy for the stitched model with RexNet as encoder and ViT-B/16 as decoder on CIFAR-100, using different projection functions and aggregation strategies. Fine-tuning the subspace selection and blending part (QKV opt) has a more significant effect on performance than fine-tuning only the larger MLP (MLP opt).

Projection	Aggregation	Accuracy \uparrow
Cosine	-	0.50
Euclidean	-	0.38
L_1	-	0.24
L_∞	-	0.21
Cosine, Euclidean, L_1 , L_∞	SelfAttention	0.17
Cosine, Euclidean, L_1 , L_∞	MLP+Sum	0.45
Cosine, Euclidean, L_1 , L_∞	SelfAttention + QKV opt	0.75
Cosine, Euclidean, L_1 , L_∞	SelfAttention + MLP opt	0.52

Result Analysis. Table 7.5 summarizes downstream classification accuracy for the stitched model using various projection functions and aggregation strategies. Incorporating multiple invariances and aggregating them via *Self-attention* (fifth row) does not perform well; meanwhile, using the *MLP+Sum* or the Cosine projection alone is more effective. This is expected, considering the attention mechanism is primarily trained to improve end-to-end performance rather than to maximize compatibility between different spaces. Incorporating the adaptation strategies at stitching time significantly boosts performance, either focusing on the subspace selection and blending (QKV opt) or the classification head (MLP opt). We find that an informed fine-tuning of the parameters responsible for the subspace blending (i.e., only the QKV projections) significantly impacts performances, even more than tuning the whole classifier. Figure 7.3 illustrates the attention weights averaged over the test dataset: the *left* figure shows the attention weights of the zero-shot stitched model, that remain unchanged when fine-tuning only the classifier, as reported on the *right*. Meanwhile, the *center* figure shows that fine-tuning the QKV projection shifts the attention weights to allocate less importance to worse-performing projections (i.e., L_∞).

Takeaway. Appropriate subspace selection and aggregation are crucial to further enhance latent communication between neural models.

Chapter 8

Limited Semantic Correspondence

*Bootstrapping Parallel Anchors for Relative Representations*¹

The previous Chapters 4, 5 and 7 presented potential solutions to solve the LCP (Chapter 3). Nevertheless, they all rely on a certain amount of parallel anchors $A_{XY} \subseteq \pi$ to be given, which can be impractical to obtain in certain scenarios. To overcome this limitation, in this Chapter, we propose an optimization-based method to discover new parallel anchors from a limited known set Λ_{XY} , denoted as *seed*. Our approach expands the semantic correspondence between different domains, enabling the solution of the LCP in scenarios where it was previously not possible.

8.1 Introduction

The previous Chapters presented potential solutions to solve the LCP, either by projecting the latent manifold embeddings into a universal space U or by directly approximating a specific transformation \mathcal{T} that relates $\varphi_{\tilde{X}}(\tilde{\mathcal{M}}_X) \subseteq \tilde{X}$ and $\varphi_{\tilde{Y}}(\tilde{\mathcal{M}}_Y) \subseteq \tilde{Y}$. Nevertheless, they all rely on the existence of an abstract correspondence \mathcal{C} , observable directly in the input spaces through π and provided in input as a certain amount of parallel anchors $A_{XY} \subseteq \pi$. However, obtaining a sufficient number of parallel anchors in specific applications can be challenging or impossible, hindering the use of the aforementioned methods.

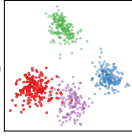
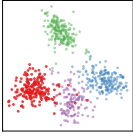
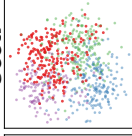
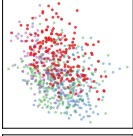
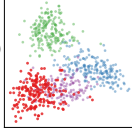
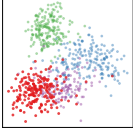
In this Chapter, we focus on the scenario where only a very limited number of parallel anchors is available, denoted as *seed* Λ , and we aim to expand this initial set through an AO (Anchor Optimization) process. Our method achieves competitive performance in NLP and Vision domains while significantly reducing the number of required parallel anchors by *one order of magnitude*.

8.2 Method

In this Section, we introduce an optimization procedure that reduces the required number of parallel anchors by one order of magnitude. This method does not require complete knowledge of A_{XY} but only of few initial *seed* anchors, denoted as $\Lambda_{XY} \subseteq A_{XY}$, where $|\Lambda| \ll |A_{XY}|$. With no prior knowledge of A_Y , we initialize the optimization process by approximating its embeddings $\tilde{\mathbf{A}}_Y$ with the known seed embeddings: $\tilde{\Lambda}_Y = \bigoplus_{a \in \Lambda_Y} E_Y(a)$, where \bigoplus denotes row-wise concatenation,

¹Irene Cannistraci, Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, and Emanuele Rodolà (2023). “Bootstrapping Parallel Anchors for Relative Representations”. In: *The First Tiny Papers Track at ICLR 2023, Tiny Papers at ICLR 2023*. URL: <https://openreview.net/pdf?id=VBuUL2IW1q>

TABLE 8.1: Qualitative (*left*) and quantitative (*right*) comparisons of the three methods when optimizing over the Word2Vec space, to discover the parallel anchors A_{XY} between Word2Vec and FastText. All metrics are calculated with $K = 10$ averaged over 20k words across 5 random seeds. Refer to Appendix A.3.1 for the metric definitions.

		FastText	Word2Vec	Source	Target	Jaccard \uparrow	MRR \uparrow	Cosine \uparrow	
GT				GT	FT	FT	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
					W2V	W2V	0.34 ± 0.01	0.94 ± 0.00	0.86 ± 0.00
Seed				Seed	FT	FT	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
					W2V	W2V	0.06 ± 0.01	0.11 ± 0.01	0.85 ± 0.01
AO				AO	FT	FT	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
					W2V	W2V	0.52 ± 0.00	0.99 ± 0.00	0.94 ± 0.00
					FT	W2V	0.50 ± 0.01	0.99 ± 0.00	0.94 ± 0.00
					W2V	W2V	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00

concatenated with $|A_{XY}| - |\Lambda_{XY}|$ random embeddings \mathbf{N} , i.e., $\tilde{\mathbf{A}}_Y = \tilde{\Lambda}_Y \oplus \mathbf{N}$ with $\mathbf{N} \sim \mathcal{N}(0, \mathbf{I})$. Thus, we define the following objective function optimizing over $\tilde{\mathbf{A}}_Y$:

$$\arg \min_{\tilde{\mathbf{A}}_Y \text{ s.t. } \|a\|_2=1 \forall a \in \tilde{\mathbf{A}}_Y} \sum_{\tilde{y} \in \tilde{Y}} \text{MSE}(R_p(\Pi(\tilde{y}), \tilde{\mathbf{A}}_X, d), R_p(\tilde{y}, \tilde{\mathbf{A}}_Y, d)) \quad (8.1)$$

where d is the cosine similarity and $\Pi: \tilde{Y} \rightarrow \tilde{X}$ is a correspondence estimated at each optimization step by the Sinkhorn algorithm (Cuturi, 2013) exploiting the initial seed and the current approximation of the remaining anchors:

$$\Pi = \text{sinkhorn}(R_p(\tilde{X}, \tilde{\mathbf{A}}_X, d), R_p(\tilde{Y}, \tilde{\mathbf{A}}_Y, d)). \quad (8.2)$$

After convergence, $\tilde{\mathbf{A}}_Y$ is discretized into $A_Y \subseteq Y$ considering the nearest embeddings in \tilde{Y} .

8.3 Experiments

This section assesses the effectiveness of the AO method in reducing the reliance on parallel anchors A_{XY} to the minimum necessary and automatically expanding the provided semantic correspondence between domains.

Experimental Setting We utilize 15 seed anchors to approximate 300 parallel anchors that serve as ground truth in all downstream tasks. Specifically, we compare the performance of our method against two different baselines: (i) *GT*, the Ground Truth employs all the anchors that our method aims to semantically approximate; and, (ii) *Seed*, exploits only the seed anchors. Refer to Appendix A.3.1 for the metric definitions, and to Cannistraci, Moschella, Maiorca, et al., 2023 for complete details.

Retrieval Task. AO effectively discovers new parallel anchors in the NLP and Vision domains, as demonstrated in Tables 8.1 and 8.2. Specifically, we explore different word embeddings and pre-trained foundational visual encoders, and assess

TABLE 8.2: Evaluation of the AO method in the vision domain on CIFAR-10. The table reports the mean results for each metric and its standard deviation across 5 different random seeds.

Mode	Type	Source	Target	Jaccard \uparrow	MRR \uparrow	Cosine \uparrow
GT	Absolute	ViT-B/16	ViT-B/16	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
			ViT-S/16	-	-	-
		ViT-S/16	ViT-B/16	-	-	-
			ViT-S/16	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	Relative	ViT-B/16	ViT-B/16	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
			ViT-S/16	0.11 ± 0.00	0.27 ± 0.01	0.97 ± 0.00
ViT-S/16		ViT-B/16	0.10 ± 0.00	0.28 ± 0.01	0.97 ± 0.00	
		ViT-S/16	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	
Seed	Absolute	ViT-B/16	ViT-B/16	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
			ViT-S/16	-	-	-
		ViT-S/16	ViT-B/16	-	-	-
			ViT-S/16	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	Relative	ViT-B/16	ViT-B/16	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
			ViT-S/16	0.03 ± 0.00	0.03 ± 0.01	0.97 ± 0.00
ViT-S/16		ViT-B/16	0.03 ± 0.00	0.04 ± 0.01	0.96 ± 0.00	
		ViT-S/16	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	
AO	Absolute	ViT-B/16	ViT-B/16	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
			ViT-S/16	-	-	-
		ViT-S/16	ViT-B/16	-	-	-
			ViT-S/16	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	Relative	ViT-B/16	ViT-B/16	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
			ViT-S/16	0.10 ± 0.01	0.23 ± 0.03	0.97 ± 0.00
ViT-S/16		ViT-B/16	0.10 ± 0.00	0.28 ± 0.01	0.97 ± 0.00	
		ViT-S/16	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	

the quality of the discovered anchors through a retrieval task. Results demonstrate that, when given the same number of starting anchors, our method outperforms the approach that relies solely on those without optimizing. Our results are *comparable* to those obtained employing all the ground truth parallel anchors.

TABLE 8.3: Cross-lingual Zero-Shot Stitching performance evaluation. The table reports the mean weighted F1 and MAE on Amazon Reviews fine-grained across 5 random seeds.

Decoder	Encoder	GT		Seed		AO	
		Fscore	MAE	Fscore	MAE	Fscore	MAE
en	en	0.64 ± 0.01	0.43 ± 0.01	0.50 ± 0.01	0.69 ± 0.01	0.62 ± 0.01	0.44 ± 0.01
	es	0.51 ± 0.01	0.67 ± 0.02	0.44 ± 0.01	0.80 ± 0.01	0.48 ± 0.01	0.70 ± 0.02
es	en	0.50 ± 0.02	0.72 ± 0.04	0.41 ± 0.01	0.92 ± 0.02	0.46 ± 0.01	0.76 ± 0.02
	es	0.60 ± 0.01	0.45 ± 0.01	0.48 ± 0.01	0.70 ± 0.01	0.61 ± 0.01	0.44 ± 0.01

Zero-Shot Stitching task. Furthermore, Table 8.3 demonstrates that our method can discover parallel anchors across different domains: the method finds aligned Amazon reviews in different languages with unavailable ground truth. Using only 15 out-of-domain anchors \hat{A} (refer to Section 4.2 for their definition), our method enables Zero-Shot Stitching (Section 3.3.1), allowing to train a classifier on one language and perform predictions on another one without any fine-tuning.

Part IV

Applying Latent Communication

Chapter 9

Case Studies

In this Chapter, we examine three case studies that highlight the potential impacts of solving the LCP. The first case, discussed in Section 9.1, shows that it is possible to create a multimodal model from unimodal models, solving CLIP-like tasks without ever training a multimodal model. The second case, discussed in Section 9.2, analyzes the possibility to merge latent spaces with differing sample and class compositions. This investigation provides insights into the theoretical and practical aspects of latent space manipulation, showcasing methods for the coherent integration of diverse datasets, which is crucial for the enhancement of model generalization and data utilization. Finally, the third case, outlined in Section 9.3, explores the Zero-Shot Stitching between policies and encoders trained on different variants of the Car Racing environment. This case study illustrates the application of LCP solutions in RL (Reinforcement Learning), particularly in the transfer and generalization of policies across varied environment, underscoring the potential for policy reuse in RL methodologies.

Through these case studies, this Chapter aims to showcase the broad applicability and significance of solving the LCP, ranging from multimodal data processing to RL. Please refer to Section 11.2 for additional applications of LCP solutions.

9.1 ASIF: Coupled Data Turns Unimodal Models to Multimodal Without Training¹

Large multimodal models such as CLIP (Radford et al., 2021) are rapidly becoming the standard for foundation models in computer vision. This is largely due to their zero-shot and open-world capabilities that enable diverse suites of downstream tasks, from classification to detection and visual search. Still, training NNs at such scale presents several challenges beside the obvious infrastructure and training costs. In fact, it requires collecting massive training sets, making it difficult to interpret the predictions of the model in light of their training data. Additionally, the training assets are often not owned by the institution training the model.

In Norelli, Fumero, et al., 2023, we present ASIF, building on the RR framework introduced in Chapter 4, to turn pre-trained uni-modal image and text encoders into a multi-modal model using a *relatively small*² collection of image-text pairs and no additional training, as depicted in Figure 9.1. The resulting model is functionally equivalent to CLIP, effectively producing aligned representations of images and their

¹Antonio Norelli, Marco Fumero, Valentino Maiorca, **Luca Moschella**, Emanuele Rodolà, and Francesco Locatello (2023). “ASIF: Coupled Data Turns Unimodal Models to Multimodal without Training”. In: *Thirty-seventh Conference on Neural Information Processing Systems*. URL: <https://openreview.net/forum?id=Xj0j3ZmWE1>

²CLIP (Radford et al., 2021) experiments used from 400M to 15M captioned images as training samples, LiT (Zhai et al., 2022b) from 901M to 10M. ASIF uses 1.6M.

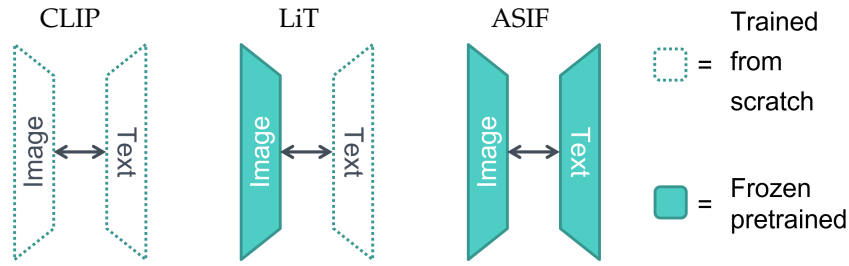


FIGURE 9.1: ASIF aligns latent spaces of frozen pre-trained encoders.

captions. The key insight is that captions of similar images are themselves similar, and therefore a representation crafted using just similarities to ground-truth multimodal pairs is quasi modality-invariant.

TABLE 9.1: Zero shot classification accuracy of different multimodal designs. CLIP and LiT implementations vary by dataset and the visual transformer used as image encoder.

Method	Dataset size	ImageNet	CIFAR100	Pets	ImageNet v2
CLIP Radford et al., 2021	400M (private)	68.6	68.7	88.9	-
CLIP Radford et al., 2021	15M (public)	31.3	-	-	-
LiT Zhai et al., 2022b	10M (public)	66.9	-	-	-
CLIP Zhai et al., 2022b	901M (private)	50.6	47.9	70.3	43.3
LiT Zhai et al., 2022b	901M (private)	70.1	70.9	88.1	61.7
ASIF (sup vis. encoder)	1.6M (public)	60.9	50.2	81.5	52.2
ASIF (unsup vis. encoder)	1.6M (public)	53.0	46.5	74.7	45.9

As shown in Table 9.1, despite the simplicity of the approach, a multimodal dataset that is up to 250 times smaller than in prior work, and the lack of actually training the model on multimodal data; ASIF achieves zero-shot classification accuracy on downstream datasets that is comparable to CLIP (Zhai et al., 2022b; Radford et al., 2021). For a comprehensive overview and discussion, please consult Norelli, Fumero, et al., 2023. The key points are summarized as follows:

- The introduction of ASIF, a method that transforms two pre-existing frozen unimodal encoders into an interpretable multimodal model.
- The demonstration of ASIF efficacy in zero-shot image classification tasks, exhibiting comparable performance to CLIP while requiring significantly fewer image-text pairs.

9.2 From Charts to Atlas: Merging Latent Spaces into One³

In Crisostomi, Cannistraci, et al., 2023, we investigate a natural follow-up question: when, and under what assumptions, *can two latent spaces be merged into one*? In principle, given two comparable representations that may partially overlap, or be entirely disjoint, it should be possible to generate a unified representation in which both coexist consistently. We refer to this problem as *Latent Space Aggregation*. Space aggregation raises several questions on (i) the representational power of the unified representation space, (ii) its ability to accommodate both spaces without collisions, and (iii) its robustness to complementary information present in only one of the two spaces. In fact, naively aggregating the sample representations by computing their mean *in absolute coordinates* would not account for the different latent configurations caused by the factors ϕ , resulting in an inconsistent aggregation of different entities based on spurious random factors.

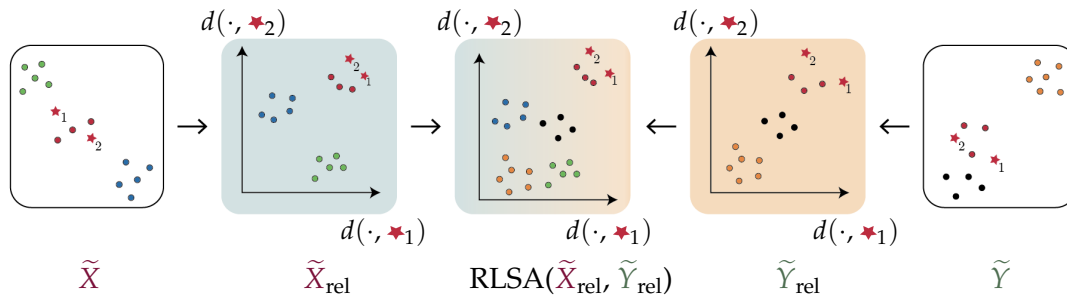


FIGURE 9.2: Relative Latent Space Aggregation description. Given two absolute spaces \tilde{X} and \tilde{Y} , we first project these spaces into two comparable RRs \tilde{X}_{rel} , \tilde{Y}_{rel} . Then, we combine these representations into a single, unified relative space $\text{RLSA}(\tilde{X}_{\text{rel}}, \tilde{Y}_{\text{rel}})$.

Motivated by the above challenges, we propose RLSA (Relative Latent Space Aggregation) illustrated in Figure 9.2. The approach involves two steps: we first switch to a RR (Chapter 4) where the latent spaces are represented with respect to a set of anchors A , and then aggregate the obtained representations of shared samples by computing their mean. The first step makes the spaces comparable, enabling a meaningful aggregation of samples that are common to multiple latent spaces, at the same time avoiding collisions.

To test the RLSA framework, we partition a classification dataset into multiple learning *tasks*. These tasks can vary in terms of class composition, such as covering disjoint subsets of classes, or in sample composition, such as being sampled with different class distributions. These diverse tasks enable us to train task-specific models, extract their latent spaces, and subsequently examine their aggregation. We consider three different cases: (i) tasks sharing a set of samples, (ii) tasks sharing the same classes but disjoint sample sets, and (iii) tasks disjoint both at the class and at the sample level. In the first case, we select the anchors from the shared samples, while in the disjoint scenarios they are sampled from unseen samples in the training dataset. We then analyze the quality of the aggregation by (i) comparing it to the space of an end-to-end model trained on all the tasks, (ii) assessing the performance

³Donato Crisostomi, Irene Cannistraci, Luca Moschella, Pietro Barbiero, Marco Ciccone, Pietro Lio, and Emanuele Rodolà (2023). “From Charts to Atlas: Merging Latent Spaces into One”. In: *NeurIPS 2023 Workshop on Symmetry and Geometry in Neural Representations*. URL: <https://openreview.net/forum?id=ZFu7CPtznY>

TABLE 9.2: Relative Latent Space Aggregation classification accuracy comparison. Each quarter shows a dataset-model combination, with end-to-end model accuracy on the right. For each S (shared classes), N (novel classes) combination, we report the accuracy of a classifier trained on the aggregated space over all the classes (*total*), along with accuracy when considering only *non-shared* and *shared* classes. *Improv* is the improvement over the end-to-end model, reported in the header, while *vanilla* the accuracy of naive merging.

	S	N	tasks	vanilla	non-shared	shared	total	improv	vanilla	non-shared	shared	total	improv
				VanillaCNN				0.39	EfficientNet				0.70
CIFAR100	80	10	2	0.36	0.60	0.39	0.43	+0.04	0.68	0.80	0.71	0.73	+0.02
	60	10	4	0.39	0.64	0.45	0.53	+0.14	0.72	0.82	0.76	0.79	+0.08
	40	10	6	0.42	0.64	0.50	0.58	+0.19	0.75	0.87	0.80	0.84	+0.14
	20	10	8	0.47	0.65	0.52	0.62	+0.23	0.80	0.88	0.84	0.87	+0.17
	80	5	4	0.37	0.77	0.41	0.49	+0.10	0.71	0.84	0.72	0.75	+0.05
	60	5	8	0.39	0.71	0.45	0.55	+0.16	0.76	0.85	0.78	0.81	+0.11
	40	5	12	0.44	0.74	0.49	0.64	+0.25	0.80	0.90	0.80	0.86	+0.16
	20	5	16	0.51	0.76	0.55	0.72	+0.33	0.83	0.93	0.83	0.90	+0.20
				VanillaCNN				0.22	EfficientNet				0.69
TINY	100	25	4	0.22	0.37	0.23	0.30	+0.08	0.68	0.75	0.71	0.73	+0.05
	50	25	6	0.24	0.36	0.36	0.36	+0.14	0.72	0.77	0.74	0.77	+0.08

of a classifier over the aggregated space, as reported in Table 9.2, and (iii) quantifying the separability of the classes within it.

For a complete description of the experiments and results, refer to Crisostomi, Cannistraci, et al., 2023. To summarize, the main contributions are three-fold:

1. The introduction of a novel framework for Latent Space Aggregation, which, for the first time, enables the merging of different latent spaces without requiring weight averaging, sharing, or any model-specific details.
2. The evaluation of the proposed framework on aggregating tasks sharing samples, classes, or neither, assessing representational power, class separability, and similarity to the global space.
3. The analysis of the improved performance on class-disjoint tasks, empirically demonstrating that it is a natural consequence of utilizing task-specific embedders.

9.3 Zero-Shot Stitching in Reinforcement Learning⁴

In the domain of RL, it is commonplace to train agents from scratch in an end-to-end manner, training both the feature extractor and the policy together. This methodology, while effective for singular tasks, presents scalability challenges both when agents need to adapt to multiple tasks within the same environments and when the same task must be tackled across environment variations.

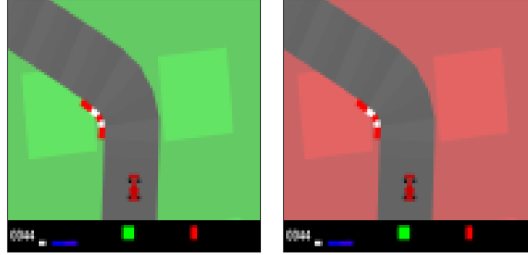


FIGURE 9.3: The modified version of the Car Racing environment, where we can change the color of the background.

An example of this challenge is illustrated in Figure 9.3, where we would like to train a policy that *drives the car* in one environment, and reuse it across variations of that environment without retraining. Ideally, a model trained on a specific task would maintain its policy while being able to substitute its encoder for another, thereby facilitating adaptation to environmental changes – such as varying weather conditions – without the need for retraining the policy. Driven by this insight, our investigation fully described in Ricciardi et al., 2023, and similar concurrent work (Jian et al., 2023), leverage RRs (Chapter 4) to prove the feasibility of Zero-Shot Stitching (Section 3.3.1) between encoders and policies trained on different environmental variations.

TABLE 9.3: Episode maximum return comparing in stitching. Encoder (rows) and policy (columns) colors represent the track background on which that module was trained on.

		Policy ↑							
		green		red		blue		yellow	
		Rel	Abs	Rel	Abs	Rel	Abs	Rel	Abs
Encoder	green	714 ± 288	863 ± 109	840 ± 94	7 ± 0.7	870 ± 84	26 ± 4	685 ± 237	12 ± 4
	red	774 ± 275	19 ± 3.8	692 ± 251	829 ± 116	288 ± 105	7 ± 0.7	638 ± 235	7 ± 0.7
	blue	256 ± 163	7 ± 1	307 ± 124	7 ± 0.6	690 ± 200	759 ± 288	556 ± 71	8 ± 3
	yellow	713 ± 204	7 ± 0.7	678 ± 81	33 ± 4	167 ± 66	26 ± 4	675 ± 233	874 ± 85

The preliminary exploration was carried out within a modified version of the CarRacing environment, featuring a discrete action space and the capability to alter the background color. We trained end-to-end agents, employing a conventional CNN as feature extractor, paired with a policy module. The empirical results, reported in Table 9.3, show that by employing RRs, a policy trained in conjunction with an encoder under a specific background color setting can be effortlessly used in a different background scenario with the corresponding encoder; resulting in minimal to no degradation in performance.

⁴Antonio Pio Ricciardi, Valentino Maiorca, **Luca Moschella**, and Emanuele Rodolà (2023). “Zero-shot stitching in Reinforcement Learning using Relative Representations”. In: *Sixteenth European Workshop on Reinforcement Learning*. URL: <https://openreview.net/forum?id=4tcXsImfsS1>

Part V

Conclusions

Chapter 10

Conclusions

In this dissertation, we introduced the *LCP (Latent Communication Problem)* framework, as formalized in Chapter 3 and illustrated in Figure 3.1. It is a novel, unifying approach that recognizes the presence of unobservable abstract manifolds, representing the underlying semantics of data. These manifolds become observable when embedded in high-dimensional spaces, such as images, texts, or latent spaces. The objective of the LCP is to identify two specific transformations, T_X and T_Y , that modify the entire latent spaces to align the manifolds within them. This framework has allowed us to *reinterpret several of our recent works*, Cannistraci, **Moschella**, Fumero, et al., 2024; Cannistraci, **Moschella**, Maiorca, et al., 2023; Crisostomi, Cannistraci, et al., 2023; Maiorca* et al., 2023; **Moschella***, Maiorca*, et al., 2023; Norelli, Fumero, et al., 2023; Ricciardi et al., 2023, from a new, unifying perspective.

Throughout this manuscript, we have showcased these methodologies to tackle the challenges presented by the LCP, building upon the foundational concepts introduced in the initial problem formalization. Our research spans from the exploration of *universal representations* (Chapter 4) and *direct translation* (Chapter 5), as well as overcoming inherent methodological limitations. These efforts include eliminating the need to know the specific transformation class relating different spaces (Chapter 7) and dealing with the limited available semantic correspondence between data domains (Chapter 8).

Indeed, beyond theoretical considerations, solving the LCP offers tangible benefits, as described in Section 3.3. One of the most salient outcomes is the concept of model compositionality through *Zero-Shot Model Stitching* (Section 3.3.1). This innovative methodology ensures that neural architectures can function as modular units, facilitating their reuse without the necessity for extensive retraining or fine-tuning. Furthermore, LCP solutions allow the direct comparisons between independently obtained latent spaces. Therefore, when an appropriate reference model is available, they provide a quantitative *latent measure of performance* (Section 3.3.2), which is often differentiable, and is correlated with standard performance measures such as accuracy on downstream tasks. Finally, it supports the development of *advanced retrieval systems* (Section 3.3.3) that leverage independently computed representations. This enables the retrieval of data points from one space using queries from another, eliminating the need for a shared training dataset. We illustrate these advantages with three detailed *case studies* in Chapter 9, covering diverse fields, including Computer Vision, Natural Language Processing and Reinforcement Learning.

This dissertation concludes with a brief overview of its impact on the broader field in Chapter 11, along with a discussion of the limitations, future research directions, and opportunities introduced by our works, detailed in Chapter 12.

Chapter 11

Contributions to the field

This Section delineates our contributions to the broader field, underscoring its adoption and impact across a diverse array of academic venues. The UniReps: Unifying Representations in Neural Models workshop at NeurIPS 2023, which our team co-organized, attests the significance and timeliness of our work (Section 11.1). Moreover, our research has been recognized and further developed in numerous preprints, peer-reviewed journals and articles presented at leading conferences (Section 11.2).

11.1 UniReps Workshop: Unifying Representations in Neural Models

The concept of *Latent Communication* has been a central theme at the NeurIPS 2023 workshop titled “UniReps: Unifying Representations in Neural Models”, which our team co-organized, and where I had the honor of serving as a Program Chair.

The workshop’s mission focused on core questions about when, how, and why different neural models converge on similar representations. This phenomenon has piqued the interest of researchers across Neuroscience, Artificial Intelligence, and Cognitive Science, indicating a thriving interdisciplinary field of study. It focused on three main themes: (i) *When*. Understanding the patterns by which these similarities emerge in different neural models and developing methods to measure them; (ii) *Why*. Investigating the underlying causes of these similarities in neural representations, considering both artificial and biological models; (iii) *What for*. Exploring and showcasing applications in modular deep learning, including model merging, reuse, stitching, efficient strategies for fine-tuning, and knowledge transfer between models and across modalities.

The workshop’s success is underscored by its substantial engagement and outcomes, demonstrating widespread interest in *Latent Communication*. It attracted 800+ attendees and received 90+ submissions, supported by a program committee of 150+ experts. Featuring invited talks from leading researchers in both industry (such as DeepMind, Anthropic) and academia (including UCSB, Princeton), the workshop was also sponsored by notable entities like Google DeepMind, Gatsby, and Translated.

11.2 Works by other researchers

This Section is dedicated to showcasing the influence and impact of our methodologies, briefly describing how they have been adopted, adapted, and extended by other researchers.

State of the art in WVLP. C. Chen et al., 2023 “proposes a *relative representation*-based WVLP (Weakly Supervised Vision-and-Language Pretraining) framework

that can both retrieve and generate weakly aligned image-text pairs for learning cross-modal representations” that “outperforms strong Weakly Supervised Vision-and-Language Pretraining baselines and further closes the performance gap between Weakly Supervised Vision-and-Language Pretraining and standard VLP”.

Generalizing Task Semantics Across Language Models. Z. Wu, Y. Wu, and Mou, 2024 “addresses a novel setting of zero-shot continuous prompt transfer, which allows for the reuse of continuous prompts across different language models”; suggesting “an encode-then-search strategy that maps a continuous prompt into a *relative space* for transfer between language models”.

Understanding Shared Speech-Text Representations. G. Wang et al., 2023 employs relative representations to reveal that the shared encoder learns a more compact and overlapping speech-text representation than the uni-modal encoders.

Policy Stitching: Learning Transferable Robot Policies. Jian et al., 2023 generalizes relative representations to enable “Policy Stitching, a model-free reinforcement learning framework for robot transfer learning among novel robot and task combinations” demonstrating clear advantages “in both zero-shot and few-shot transfer learning through simulated and physical 3D manipulation tasks”.

Relative Representations for Cognitive Graphs. Kiefer and Buckley, 2024 extends “*relative representations* to discrete state-space models, using Clone-Structured Cognitive Graphs (CSCGs)”; showing that “the probability vectors computed during message passing can be used to define relative representations on CSCGs” enabling effective Latent Communication across agents trained in different settings.

Model Stitching with Static Word Embeddings. Ye et al., 2024 introduces “MoSE-CroT, a novel and challenging task for (especially low-resource) languages where static word embeddings are available”, and proposes “a solution that leverages *relative representations* to construct a common space for source and target languages and that allows zero-shot transfer for the target languages”.

Knowledge Distillation with Relative Representations. Ramos, Alampay, and Abu, 2023 designs “a knowledge distillation scheme centered around matching the relative representations of a student to those of a teacher” and show that the proposed method “outperforms similar relation-based distillation approaches across a variety of benchmarks, with results extending to transfer learning”.

Direct Alignment of Latent Spaces. Löhner and Moeller, 2023 proposes, concurrently to the work presented in Chapter 5, “the theory that semantically related latent spaces even of very different network architectures are related by a linear transformation” and demonstrates that “aligning the latent space with a linear transformation performs best while not needing more prior knowledge”.

Boosting Visual-Language Models. H. Wang et al., 2023 employs relative representations to “propose a novel hard sample selection technique for the identification of hard negative samples” and “consistently improve CLIP model checkpoints by finetuning”.

Chapter 12

Limitations and Future Directions

In this Section, we summarize some open questions and potential avenues for future research based on the contributions presented in this dissertation. Our work has laid a strong foundation in the LCP framework, opening several pathways for further exploration and development. In the following paragraphs, we outline some of the most promising directions.

Modular Neural Components. Our exploration of *Latent Communication* paves the way for leveraging neural models in a modular, compositional fashion, potentially circumventing the exhaustive fine-tuning or retraining currently prevalent. Yet, bridging the performance gap across training modalities – especially between zero-shot and fine-tuning approaches – remains an elusive challenge. This disparity is particularly pronounced in industrial ML applications, where performance maximization often comes at the expense of computational efficiency. Our Zero-Shot Stitching methodology offers a partial solution; however, the quest for models that adapt dynamically to changes in feature representation with minimal retraining persists. Future research might focus on *self-adjusting mechanisms* akin to Test-Time Training (Sun et al., 2020; D. Wang et al., 2020), integrated with Latent Communication strategies, to address this gap.

Latent Communication without Semantic Correspondence. The dependency on initial seed anchors for parallel domains limits the current scope of Latent Communication. Removing this constraint could involve developing unsupervised methods for identifying parallel anchors or methods for learning the anchors from the data. The Gromov-Wasserstein distance (Mémoli, 2011) presents a promising theoretical underpinning for such methods, e.g., potentially revolutionizing cross-domain retrieval systems by eliminating the need for parallel data.

Latent Communication on sequences. Currently, we have devised solutions for the LCP when samples can be represented as a single embedding, i.e., a point in a high-dimensional space. An interesting direction could be to solve the LCP natively on sequences, e.g., by considering the latent space of a sequence of embeddings. This would allow a more natural handling of sequential data, such as textual embeddings, without the need to aggregate all the token embeddings into a single one (e.g., considering the final token, the CLS or the mean of the token embeddings). However, this would require the use of a similarity function that can compare embedding sequences, or the development of novel methods to achieve LCP in this context.

Neural Networks Inspection. Solutions to the LCP could be used to analyze the latent spaces of NNs, e.g., providing insights into the evolution of the representations

during training. They could be an interesting tool to understand when and how NNs develop or refine their ability to represent information. Indeed, it is yet to be understood how the latent spaces evolve during training, and there are a variety of phenomenon (e.g., emergent abilities (Wei et al., 2022), neural collapse (Papayan, X. Y. Han, and Donoho, 2020), double descent (Nakkiran et al., 2019)) that could be investigated exploiting the LCP framework and a known reference model, inspired by Section 3.3.2.

Partial Latent Communication. Requiring a full alignment of the manifold embeddings might be too restrictive in some scenarios, e.g., in cases where the two data distributions are only partially semantically overlapping. In these cases, it would be interesting to develop methods for *Partial Latent Communication* to align only a subset of the data manifolds. Similar techniques could be used to ensure the best alignment for a particular subset of the data of interest, such as particular classes or categories.

Representation Interpretability. The framework of RRs offers a novel lens for representation interpretability, associating specific meanings with each dimension through the anchor semantic. This could be further exploited by more tailored similarity functions, e.g., by performing a change of basis to obtain a more interpretable RR. This would allow interpreting the dimensions of the representation as the directions in the data space associated with specific semantic concepts, defined by the anchor, thus providing a more intuitive understanding of the latent space.

Learnable Similarity Functions. The framework described in Chapter 4 allows incorporating invariances into the latent representation, exploiting specific similarity functions. Moreover, in Chapter 7 we have shown how to extend it to infuse a set of invariances, instead of a single one. Nevertheless, this can still be limiting when the similarity function that induces an invariance to \mathbb{T} cannot be modeled analytically or expressed in closed form. In such cases, an interesting direction would be to *learn* the desired similarity function d .

Geodesic Relative Representations. Another fascinating line of research to improve the representation expressivity would be to estimate *geodesic* distances over the data manifold, instead of adopting distances in the ambient spaces. This could allow defining RRs that better capture the intrinsic structure of the data, especially when the manifold embedding is complex. However, this would require the development of innovative methods for efficient geodesic estimation in high-dimensional spaces.

Higher Order Relative Representations. The RR framework is currently limited to employ pairwise similarity functions, i.e., it can only capture first-order relationships between data points. In practice, this means that there can be only one anchor associated with each dimension. Extending it to higher-order n -way relationships, e.g., by considering triplets or quadruplets, could allow capturing more complex relationships between data points, and thus provide a more expressive representation.

Anchor selection methods. The interplay between anchor composition and the expressiveness of RRs warrants further investigation. Questions surrounding optimal anchor selection and the necessary number of anchors remain mostly unanswered. Developing methodologies for selecting anchors – guided by considerations of data

distribution, transformation classes, or specific tasks – stands as a significant frontier for future research.

Weights Similarity. Throughout this dissertation, we have explored the emerging similarities in the latent spaces of NNs. At the same time, a growing body of research has focused on emerging similarities between networks in the weight space, and how to exploit them (Ainsworth, Hayase, and Srinivasa, 2023; Ilharco et al., 2023; Ortiz-Jimenez, Favero, and Frossard, 2023; Ramé et al., 2023; Entezari et al., 2022; Matena and Raffel, 2022; Wortsman et al., 2022; Frankle et al., 2020; Singh and Jaggi, 2020; Tatro et al., 2020). An exciting research direction would be to investigate the relationship between the latent spaces and the weights of the models, answering the following question: “Are NNs with similar latent spaces also similar in the weight space?”

Automatic Data Curation. Data-centric AI and automatic data curation are experiencing rapid growth. This evolution underscores a realization: the “quantity” is not the sole determinant of AI performance. Rather, the “quality” of data plays a crucial, if not more significant, role in enhancing the training processes, boosting model performance, and optimizing model size. Within this context, the LCP methodology emerges as a powerful tool. It introduces a paradigm shift by employing trained models not just for predictions, but as a means to critically assess and ensure the quality of datasets – for instance, automating the process of identifying and eliminating noisy data alignments. This direction holds considerable promise for future research, particularly when dealing with multimodal data, where aligning diverse data modalities presents considerable challenges.

Part VI
Appendices

Appendix A

Universal Representations

A.1 Anchors analysis

The cardinality of the anchors set A and the choice of specific anchors is crucial to the quality of the relative representations. At the extreme, selecting one single anchor or the same repeated data points for all anchors, will produce collapsed relative representations. We believe that additional research is required to obtain a better understanding on the optimal choice for A . Questions like “Are anchors set composed only by stopwords worse than the ones composed by meaningful and diverse words?” require empirical evidence and could help revealing the semantics of the latent space. Indeed, each anchor is associated with a dimension in a relative representation; one could inspect the anchor data point to get a sense of the meaning of that latent dimension.

Anchor number. Below, we report a preliminary study on the performance sensitivity against the cardinality of the anchors set. In Figure A.1 we report the performance on the node classification task on Cora, with a model trained end-to-end adopting the relative representations while training, and on image classification tasks on CIFAR-100, with a frozen encoder. The performance improves monotonically as the number of anchors increase when the absolute representations are frozen (*right*). Differently, training models end-to-end proves to be more susceptible to model collapse and instabilities, as increasing the number of anchors does not always improve the performance (*left*). Further research on the relation between the absolute latent space dimensionality and the relative representation dimensionality (i.e., the number of anchors) is needed to clarify how the two quantities impact the performance, when training end-to-end or not.

Anchor selection. In Tables A.1 and A.2, we analyze different anchor selection strategies under an experimental setting analogous to the one described in Section 4.3.1:

- **uniform** The first selection strategy is the one adopted in Chapter 4. We randomly select the anchors with a uniform probability distribution over all the available samples.
- **fps** We select the anchors according to a farthest point sampling strategy.
- **kmeans** We select the anchors as the words more close to the centroids of K-means clustering with $K = \text{number of anchors}$.
- **top{k}** We select the anchors as the k most frequent words, after skipping the first 400 which are mostly stopwords.

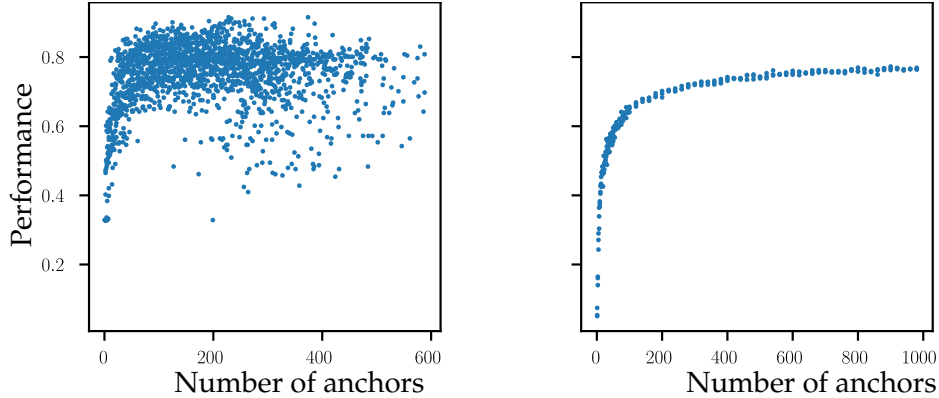


FIGURE A.1: Accuracy vs Number of anchors. Each point is a trained model. *Left*: Trained embedder on Cora, node classification. *Right*: Frozen transformer on CIFAR-100 coarse-grained, image classification. Left is less stable because the absolute embeddings are trained, and we are working on a domain that is less stable (graphs). Some collapsed examples are not visualized.

We expect strategies that better cover the absolute space with anchors to be the most effective ones. Indeed, the results are comparable across selection strategies, but fps reaches everywhere the best Jaccard and MRR scores while k-means the best Cosine ones. We attribute this behavior to their different nature: they both rely on the geometry of the latent spaces they are applied to, but k-means also favors high-density regions, and this can become a negative bias for the task at hand. In general, the uniform sampling is the most straightforward to apply, since it does not require additional computation for the selection process, and still achieves good performances.

A.2 Dataset Information

In Table A.3 we summarize the datasets utilized in Chapter 4, and for each one, we specify the number of classes, to give an idea about the classification difficulty.

A.3 Implementation Details

In this Section, following the corresponding sections in Chapter 4, we report implementation details for all the experimental settings considered.

Tools & Technologies. In all the experiments presented in this work, the following tools were used:

- *NN-Template* GrokAI, 2021, to easily bootstrap the project and enforce best practices.
- *PyTorch Lightning* (Falcon and The PyTorch Lightning team, 2019), to ensure reproducible results while also getting a clean and modular codebase.
- *Weights and Biases* (Biewald, 2020), to log experiments and compare runs across huge sweeps.

- *Transformers by HuggingFace* (Wolf et al., 2020), to get ready-to-use transformers for both text and images.
- *Datasets by HuggingFace* (Lhoest et al., 2021), to access most of the NLP datasets and ImageNet for CV.
- *DVC* (Kuprieiev et al., 2023), for data versioning.
- *PyTorch Geometric* (Fey and Lenssen, 2019), to handle graph datasets and get ready-to-use GNN architectures.

A.3.1 Word Embeddings

For both the Figure and the Table in Section 4.3.1, the number of anchors is set to 300 for a fair comparison with the dimensionality of the original spaces. For visualization purposes, we needed the figure to both show an easy clusterable and restricted set of word embeddings. They are obtained by subsampling the shared vocabulary with the following procedure: we select 4 random pivot words, and for each of them we consider the top-200 words in their neighborhood. This results in a total of 800 points divided in 4 clusters, the ones used only for the visualization part. For the quantitative part (table results), we select 20K random words from the shared vocabulary with a fixed seed for reproducibility purposes.

For the computer vision counterpart (Figure A.5 and table A.2), the procedure is similar but with the following differences: (i) the number of anchors is set to 500 to balance between the different encoding dimensions of the two transformers (384 for ViT-S/16 and 768 for ViT-B/16); (ii) the subsampling for visualization purposes is done by selecting 4 classes and randomly picking 200 samples for each of them;

Evaluation metrics. Consider the source space \tilde{X}' and target space \tilde{Y}' and a set of $\approx 20k$ samples $S \subseteq (X \cap Y)$ (words for the NLP test, images for the CV one); for any sample $s \in S$, we compute its representation in \tilde{X}' and \tilde{Y}' through the functions $f_{\tilde{X}'} : S \rightarrow \tilde{X}'$ and $f_{\tilde{Y}'} : S \rightarrow \tilde{Y}'$ (e.g. f can be the encoder composed with a relative projection) and define the metrics as follows:

$$\begin{aligned} \text{Jaccard}(s) &= \frac{|\text{KNN}_k^{\tilde{X}'}(f_{\tilde{X}'}(s)) \cap \text{KNN}_k^{\tilde{Y}'}(f_{\tilde{Y}'}(s))|}{|\text{KNN}_k^{\tilde{X}'}(f_{\tilde{X}'}(s)) \cup \text{KNN}_k^{\tilde{Y}'}(f_{\tilde{Y}'}(s))|} \\ \text{MRR}(s) &= \frac{1}{\text{Rank}_{\tilde{Y}'}(f_{\tilde{X}'}(s), f_{\tilde{Y}'}(s))} \\ \text{Cosine}(s) &= \frac{f_{\tilde{X}'}(s) \cdot f_{\tilde{Y}'}(s)}{\|f_{\tilde{X}'}(s)\| \|f_{\tilde{Y}'}(s)\|} \end{aligned}$$

where $\text{KNN}_k^S(v)$ is a function that returns the k -top similar samples (according to cosine similarity) to v in the space S , and $\text{Rank}_S(v, u)$ is a function that returns the index at which u is found in the ordered $\text{KNN}_k^S(v)$. The final score for each metric is the mean over each $s \in S$.

A.3.2 Relative representation space correlations

In this Section, we analyze how similarities in absolute and relative spaces are correlated. Let us consider two spaces alignable in the relative space. We denote elements

of the spaces with $\mathbb{A} \in \mathbb{R}^{m_1 \times n_1}$ and $\mathbb{B} \in \mathbb{R}^{m_2 \times n_2}$ and corresponding relative embeddings with $\mathbb{C} \in \mathbb{R}^{m_1 \times d}$, $\mathbb{D} \in \mathbb{R}^{m_2 \times d}$. Examples of \mathbb{A} and \mathbb{B} can be the FastText and Word2Vec word embedding spaces. We already observed in Table 4.1 how the spaces \mathbb{A} and \mathbb{B} are well aligned in the relative space. We can go further and analyze how self similarities in each space are preserved by the relative transform. In Figure A.2, we show that relative representations not only facilitate latent communication, but also preserve the underlying (absolute) latent space metric up to a certain degree.

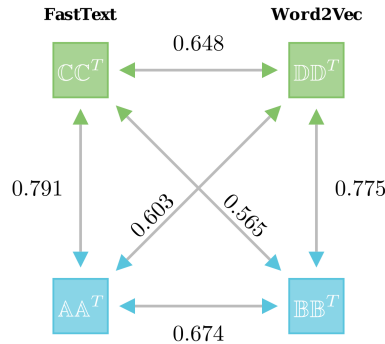


FIGURE A.2: Self similarities correlations between each space, measured with the Pearson correlation coefficient. In blue, we denote the self similarities in the absolute spaces \mathbb{A} , \mathbb{B} of FastText and Word2Vec; in green we depict the relative spaces \mathbb{C} , \mathbb{D} . The correlation in the vertical arrows indicate how much the underlying metric in the absolute space is preserved by the relative coordinate transformation.

A.3.3 Latent distance as a performance proxy

The hyperparameters used in Section 4.3.2 are summarized in Table A.4.

A.3.4 Training with Absolute vs. Relative Representations

The models trained on relative representations do not backpropagate through the anchors, which encourages a smoother optimization of the anchors’ representations.

Image Classification. The architecture is a standard deep CNN. We run a sweep for each dataset where we vary only the random seed (over 10 possible in total). We then aggregate by dataset and encoding type to obtain the final results with their standard deviation.

Graph Classification. We run a sweep identical to the one in Table A.4 for the reference model, except that we sweep on the “Number of layers” with two values: 32 and 64. Each configuration is repeated with 10 different seeds, then we aggregate by dataset and encoding type to obtain the final results with their standard deviation.

A.3.5 Image Reconstruction

The relative and absolute models appearing in Figure 4.4 are vanilla AEs and VAEs, the same for all the datasets, and have a comparable number of trainable parameters. Their architecture is composed by simple convolutions, deconvolutions and mean squared error as reconstruction loss. The number of anchors is 500 and the latent dimensionality of the absolute representations is 500.

A.3.6 Text Classification

We report in Tables A.5 to A.7 details on the transformers and anchors adopted in Section 4.4.2.

Preprocessing. Following the original work in which the Amazon Reviews dataset was proposed (Keung et al., 2020), we utilize both the *title* and *body* of each review. We differ in not using the category and in how we merge them; namely, we add the title as prefix for the body and add a full stop as separator when needed (avoiding duplicates). To obtain a single latent encoding for each sample, with fixed shape, we take the last hidden state and select the representation corresponding to the *[CLS]* token.

Wikipedia anchors. We use WikiMatrix, a corpus of sentences extracted from Wikipedia. The sentences are parallel between pairs of languages (i.e., same sentences translated in two languages), and since we are looking for a collection of parallel anchors between all 4 languages, we decided to use the English language as a pivot to compute the intersection. To get the final results, we considered only the sentences with margin score ≥ 1.06 , getting high-quality sentence alignments. In Table A.7 we show the total number of parallel sentences when computing the intersections. We randomly selected 768 samples to use as anchors.

A.3.7 Image Classification

The details of the transformers used in Section 4.4.3 are summarized in Table A.8.

A.4 Additional results

In this Section we report additional results on the correlation between latent similarity and performance in Figure A.3, results on the multilingual stitching both with Amazon coarse-grained in Table A.9 and fine-grained in Table A.10, results on the image classification stitching on CIFAR-100 fine-grained in Table A.12. Moreover, we evaluate the stitching performance of a multilingual transformer in Table A.11.



FIGURE A.3: Correlations between performance and latent similarity with the reference model for multiple different models, over time.

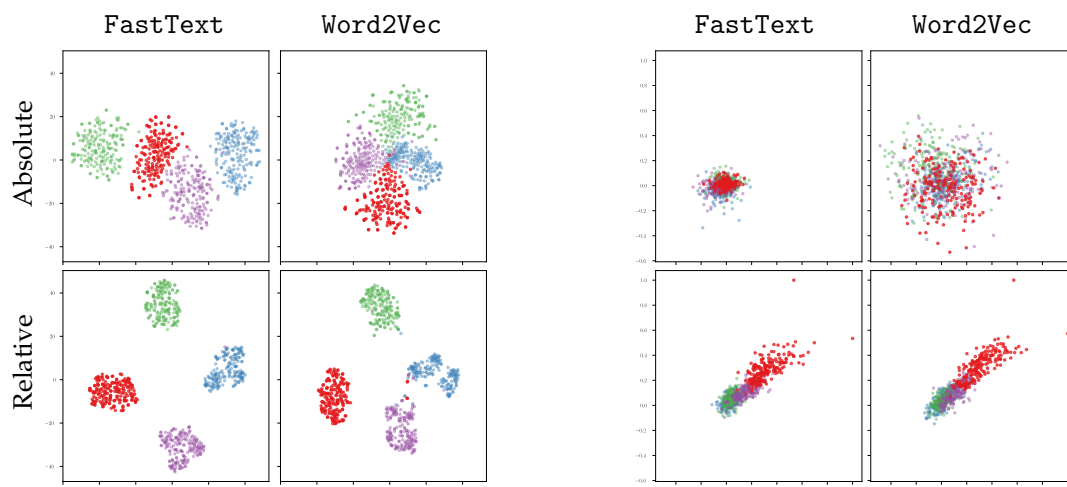


FIGURE A.4: Same encodings as in Table 4.1 (left) but with tSNE (left) dimensionality reduction or visualizing only their first two dimensions (right).

TABLE A.1: Extended results from Section 4.3.1 with different anchor selection strategies. The table reports the mean score for each metric and its std across 10 different seeds.

Mode	Type	Source	Target	Jaccard \uparrow	MRR \uparrow	Cosine \uparrow
uniform	Absolute	FastText	FastText	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
			Word2Vec	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.00
		Word2Vec	FastText	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.00
			Word2Vec	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	Relative	FastText	FastText	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
			Word2Vec	0.34 ± 0.01	0.94 ± 0.00	0.86 ± 0.00
fps	Absolute	FastText	FastText	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
			Word2Vec	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.00
		Word2Vec	FastText	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.00
			Word2Vec	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	Relative	FastText	FastText	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
			Word2Vec	0.34 ± 0.01	0.94 ± 0.00	0.81 ± 0.00
kmeans	Absolute	FastText	FastText	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
			Word2Vec	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.00
		Word2Vec	FastText	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.00
			Word2Vec	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	Relative	FastText	FastText	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
			Word2Vec	0.35 ± 0.00	0.94 ± 0.00	0.87 ± 0.00
top1000	Absolute	FastText	FastText	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
			Word2Vec	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.00
		Word2Vec	FastText	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.00
			Word2Vec	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	Relative	FastText	FastText	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
			Word2Vec	0.27 ± 0.01	0.84 ± 0.01	0.85 ± 0.00
top5000	Absolute	FastText	FastText	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
			Word2Vec	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.00
		Word2Vec	FastText	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.00
			Word2Vec	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	Relative	FastText	FastText	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
			Word2Vec	0.32 ± 0.00	0.92 ± 0.00	0.86 ± 0.00
top10000	Absolute	FastText	FastText	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
			Word2Vec	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.00
		Word2Vec	FastText	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.00
			Word2Vec	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	Relative	FastText	FastText	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
			Word2Vec	0.34 ± 0.00	0.93 ± 0.00	0.86 ± 0.00
Relative	FastText	FastText	0.39 ± 0.01	0.97 ± 0.00	0.86 ± 0.00	
		Word2Vec	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	
	Word2Vec	FastText	0.39 ± 0.01	0.97 ± 0.00	0.86 ± 0.00	
		Word2Vec	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	

TABLE A.2: Generalization of the results from Section 4.3.1 on word embeddings to a different data modality, with different anchor selection strategies (See Appendix A.1 for their description). The dataset considered is CIFAR-10, and the table reports the mean score for each metric and its std across 10 different seeds.

Mode	Type	Source	Target	Jaccard \uparrow	MRR \uparrow	Cosine \uparrow
uniform	Absolute	ViT-B/16	ViT-B/16	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
			ViT-S/16	-	-	-
		ViT-S/16	ViT-B/16	-	-	-
			ViT-S/16	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	Relative	ViT-B/16	ViT-B/16	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
			ViT-S/16	0.11 ± 0.00	0.27 ± 0.01	0.97 ± 0.00
	ViT-S/16	ViT-B/16	0.11 ± 0.00	0.30 ± 0.01	0.97 ± 0.00	
		ViT-S/16	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	
fps	Absolute	ViT-B/16	ViT-B/16	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
			ViT-S/16	-	-	-
		ViT-S/16	ViT-B/16	-	-	-
			ViT-S/16	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	Relative	ViT-B/16	ViT-B/16	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
			ViT-S/16	0.12 ± 0.00	0.37 ± 0.01	0.96 ± 0.00
	ViT-S/16	ViT-B/16	0.12 ± 0.00	0.39 ± 0.01	0.96 ± 0.00	
		ViT-S/16	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	
kmeans	Absolute	ViT-B/16	ViT-B/16	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
			ViT-S/16	-	-	-
		ViT-S/16	ViT-B/16	-	-	-
			ViT-S/16	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	Relative	ViT-B/16	ViT-B/16	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
			ViT-S/16	0.11 ± 0.00	0.25 ± 0.01	0.97 ± 0.00
	ViT-S/16	ViT-B/16	0.10 ± 0.00	0.27 ± 0.00	0.97 ± 0.00	
		ViT-S/16	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	

TABLE A.3: All the datasets utilized in Chapter 4 with their number of classes.

	Dataset	Number of Classes
Image	MNIST	10
	Fashion MNIST	10
	CIFAR-10	10
	CIFAR-100	20 (coarse) 100 (fine)
	ImageNet1k	1000
Graph	Cora	7
	CiteSeer	6
	PubMed	3
Text	TREC	6 (coarse) 50 (fine)
	DBpedia	14
	Amazon Reviews	2 (coarse) 5 (fine)

TABLE A.4: The reference model and exhaustive hyperparameter combinations pertaining Section 4.3.2.

Hyperparameter	Reference Model	Sweep
Seed	1	0, 1, 2, 3, 4
Epochs	500	10, 30, 50
Number of layers	32	32, 64
Dropout Probability	0.5	0.1, 0.5
Hidden Activations	ReLU	ReLU, Tanh
Convolution Activation	ReLU	ReLU, Tanh
Optimizer	Adam	Adam, SGD
Learning Rate	0.02	0.01, 0.02
Graph Embedder	GCNConv	GCNConv, GINConv

TABLE A.5: The HuggingFace transformers models employed in Section 4.4.2 to tackle the *Cross-lingual* setting.

Language	HuggingFace transformers name	Encoding Dim
English	roberta-base	768
Spanish	PlanTL-GOB-ES/roberta-base-bne	768
French	ClassCat/roberta-base-french	768
Japanese	nlp-waseda/roberta-base-japanese	768

TABLE A.6: The HuggingFace transformers models employed in Section 4.4.2 to tackle the *Cross-architecture* setting.

HuggingFace transformers name	Encoding Dim
bert-base-cased	768
bert-base-uncased	768
google/electra-base-discriminator	768
roberta-base	768

TABLE A.7: WikiMatrix analysis. Each row shows the number of parallel sentences having a translation available in all the languages of that row. Since we consider all four languages, we have 3338 parallel sentences available.

Languages	Number of Sentences
en, es	2302527
en, ja	264259
en, fr	1682477
en, es, fr	23200
en, es, ja	147665
en, fr, ja	20990
en, es, fr, ja	3338

TABLE A.8: Timm transformers used in Section 4.4.3.

Version	Timm model name	Encoding Dim	Training data
ViT	vit_base_patch16_224	768	JFT-300M, ImageNet
ViT	vit_small_patch16_224	384	ImageNet
ViT	vit_base_resnet50_384	768	ImageNet
RexNet	rexnet_100	1280	ImageNet

TABLE A.9: Stitching performance comparison with different encodings techniques. The table reports the mean weighted F1 (\pm std) and MAE classification performance on Amazon Reviews coarse-grained, across 5 different seeds. All the language pairs are shown.

Decoder	Encoder	Absolute		Relative			
		FScore	MAE	Translated		Wikipedia	
				FScore	MAE	FScore	MAE
en	en	91.54 \pm 0.58	0.08 \pm 0.01	90.06 \pm 0.60	0.10 \pm 0.01	90.45 \pm 0.52	0.10 \pm 0.01
	es	43.67 \pm 1.09	0.56 \pm 0.01	82.78 \pm 0.81	0.17 \pm 0.01	78.53 \pm 0.30	0.21 \pm 0.00
	fr	54.41 \pm 1.61	0.45 \pm 0.02	78.49 \pm 0.66	0.21 \pm 0.01	70.41 \pm 0.57	0.29 \pm 0.01
	ja	48.72 \pm 0.90	0.51 \pm 0.01	65.72 \pm 0.55	0.34 \pm 0.01	66.31 \pm 0.80	0.34 \pm 0.01
es	en	33.23 \pm 1.00	0.66 \pm 0.01	78.68 \pm 2.74	0.21 \pm 0.03	76.65 \pm 3.23	0.23 \pm 0.03
	es	91.64 \pm 1.02	0.08 \pm 0.01	89.96 \pm 0.77	0.10 \pm 0.01	89.62 \pm 0.94	0.10 \pm 0.01
	fr	47.66 \pm 0.70	0.52 \pm 0.01	78.57 \pm 1.80	0.21 \pm 0.02	75.25 \pm 0.76	0.25 \pm 0.01
	ja	53.10 \pm 2.27	0.46 \pm 0.02	67.69 \pm 0.24	0.32 \pm 0.00	61.84 \pm 0.61	0.38 \pm 0.01
fr	en	51.00 \pm 2.63	0.49 \pm 0.03	83.32 \pm 1.80	0.17 \pm 0.02	75.55 \pm 0.37	0.24 \pm 0.00
	es	51.96 \pm 2.81	0.48 \pm 0.03	82.50 \pm 0.83	0.17 \pm 0.01	77.12 \pm 0.88	0.23 \pm 0.01
	fr	88.22 \pm 0.75	0.12 \pm 0.01	85.68 \pm 1.37	0.14 \pm 0.01	86.45 \pm 0.96	0.13 \pm 0.01
	ja	50.32 \pm 4.16	0.50 \pm 0.04	69.38 \pm 0.73	0.31 \pm 0.01	62.79 \pm 0.27	0.37 \pm 0.00
ja	en	53.82 \pm 2.62	0.46 \pm 0.03	68.66 \pm 3.62	0.31 \pm 0.04	70.26 \pm 3.16	0.29 \pm 0.03
	es	44.91 \pm 2.21	0.55 \pm 0.02	70.37 \pm 6.94	0.29 \pm 0.06	58.54 \pm 1.21	0.41 \pm 0.01
	fr	66.46 \pm 1.30	0.34 \pm 0.01	76.49 \pm 1.13	0.23 \pm 0.01	63.94 \pm 2.70	0.36 \pm 0.02
	ja	83.30 \pm 0.67	0.17 \pm 0.01	81.04 \pm 0.82	0.19 \pm 0.01	80.80 \pm 1.25	0.19 \pm 0.01

TABLE A.10: Stitching performance comparison with different encodings techniques. The table reports the mean weighted F1 (\pm std) and MAE classification performance on Amazon Reviews fine-grained, across 5 different seeds. All the language pairs are shown.

Decoder	Encoder	Absolute		Relative			
		FScore	MAE	Translated		Wikipedia	
				FScore	MAE	FScore	MAE
en	en	65.46 \pm 2.89	0.38 \pm 0.02	61.18 \pm 1.92	0.44 \pm 0.02	62.36 \pm 2.23	0.43 \pm 0.02
	es	22.70 \pm 0.41	1.39 \pm 0.03	51.67 \pm 1.20	0.62 \pm 0.01	45.40 \pm 0.68	0.76 \pm 0.01
	fr	30.75 \pm 0.67	1.19 \pm 0.02	49.18 \pm 0.83	0.69 \pm 0.02	40.29 \pm 0.90	0.91 \pm 0.02
	ja	24.85 \pm 0.91	1.37 \pm 0.07	37.34 \pm 1.49	0.99 \pm 0.02	37.73 \pm 0.70	1.01 \pm 0.02
es	en	21.24 \pm 0.81	1.43 \pm 0.07	51.02 \pm 2.54	0.68 \pm 0.05	47.70 \pm 5.08	0.73 \pm 0.10
	es	61.29 \pm 3.04	0.43 \pm 0.02	57.89 \pm 3.80	0.48 \pm 0.03	57.96 \pm 4.40	0.48 \pm 0.03
	fr	29.02 \pm 0.85	1.26 \pm 0.05	48.40 \pm 1.02	0.71 \pm 0.02	44.92 \pm 1.83	0.77 \pm 0.01
	ja	29.23 \pm 1.32	1.22 \pm 0.02	37.22 \pm 1.56	1.03 \pm 0.04	34.56 \pm 0.87	1.08 \pm 0.04
fr	en	27.39 \pm 1.22	1.23 \pm 0.06	45.55 \pm 3.55	0.76 \pm 0.09	39.01 \pm 1.25	0.88 \pm 0.06
	es	29.47 \pm 3.68	1.18 \pm 0.07	40.29 \pm 1.72	0.90 \pm 0.04	41.29 \pm 2.01	0.83 \pm 0.04
	fr	56.40 \pm 1.89	0.51 \pm 0.01	53.58 \pm 0.70	0.57 \pm 0.01	54.23 \pm 0.95	0.56 \pm 0.01
	ja	25.92 \pm 1.31	1.25 \pm 0.05	38.60 \pm 1.03	0.96 \pm 0.02	35.22 \pm 0.56	1.08 \pm 0.02
ja	en	29.36 \pm 0.59	1.17 \pm 0.04	38.19 \pm 2.28	0.88 \pm 0.03	36.57 \pm 1.72	0.98 \pm 0.02
	es	25.64 \pm 1.77	1.28 \pm 0.04	34.23 \pm 2.62	1.00 \pm 0.05	33.16 \pm 2.28	1.06 \pm 0.03
	fr	31.79 \pm 1.91	1.06 \pm 0.02	38.50 \pm 2.46	0.89 \pm 0.02	36.68 \pm 3.14	1.00 \pm 0.05
	ja	54.09 \pm 1.35	0.60 \pm 0.02	50.89 \pm 1.70	0.65 \pm 0.02	51.64 \pm 1.47	0.65 \pm 0.02

TABLE A.11: Stitching performance comparison on XLM-R, a multilingual model by design. The table reports the mean weighted F1 (\pm std) and MAE classification performance on Amazon Reviews fine-grained, across 5 different seeds.

Decoder	Encoder	Absolute		Relative	
		FScore	MAE	FScore	MAE
en	en	65.27 ± 0.94	0.41 ± 0.01	58.24 ± 1.92	0.51 ± 0.03
	es	59.55 ± 0.76	0.48 ± 0.01	52.81 ± 1.57	0.62 ± 0.02
	fr	58.58 ± 1.04	0.49 ± 0.01	54.01 ± 1.34	0.59 ± 0.02
	ja	57.98 ± 0.77	0.52 ± 0.01	48.47 ± 2.67	0.71 ± 0.04
es	en	60.32 ± 1.50	0.47 ± 0.01	45.69 ± 2.19	0.87 ± 0.07
	es	61.25 ± 1.74	0.44 ± 0.01	57.61 ± 0.73	0.51 ± 0.01
	fr	59.50 ± 1.41	0.47 ± 0.01	45.16 ± 3.30	0.83 ± 0.09
	ja	58.24 ± 1.31	0.51 ± 0.02	41.14 ± 1.76	0.99 ± 0.05
fr	en	58.00 ± 4.21	0.49 ± 0.03	52.37 ± 1.66	0.66 ± 0.03
	es	56.87 ± 3.79	0.49 ± 0.03	54.99 ± 0.46	0.57 ± 0.01
	fr	57.99 ± 3.88	0.47 ± 0.02	57.00 ± 0.90	0.52 ± 0.01
	ja	55.83 ± 3.32	0.53 ± 0.03	39.15 ± 1.21	1.02 ± 0.03
ja	en	59.53 ± 1.73	0.48 ± 0.01	39.46 ± 2.34	1.04 ± 0.07
	es	57.02 ± 1.36	0.51 ± 0.00	40.74 ± 2.75	0.97 ± 0.09
	fr	57.48 ± 1.06	0.51 ± 0.01	43.36 ± 3.70	0.89 ± 0.11
	ja	61.43 ± 0.97	0.45 ± 0.01	57.67 ± 1.17	0.51 ± 0.01

TABLE A.12: Stitching performance comparison with different encodings techniques. The table reports the mean weighted F1 (\pm std) classification performance on CIFAR-100 fine-grained, across 5 different seeds.

Decoder	Encoder	Absolute	Relative
RexNet	RexNet	72.77 ± 0.19	71.39 ± 0.18
	ViT-B/16	-	40.68 ± 0.50
	RViT-B/16	-	38.18 ± 0.24
	ViT-S/16	-	44.11 ± 0.84
ViT-B/16	RexNet	-	57.81 ± 0.39
	ViT-B/16	88.69 ± 0.14	87.05 ± 0.34
	RViT-B/16	1.08 ± 0.19	66.65 ± 1.79
	ViT-S/16	-	73.73 ± 0.60
RViT-B/16	RexNet	-	66.91 ± 0.79
	ViT-B/16	1.10 ± 0.09	75.70 ± 0.68
	RViT-B/16	85.85 ± 0.18	85.04 ± 0.38
	ViT-S/16	-	75.52 ± 0.36
ViT-S/16	RexNet	-	56.60 ± 0.39
	ViT-B/16	-	70.14 ± 0.46
	RViT-B/16	-	62.85 ± 1.22
	ViT-S/16	84.11 ± 0.14	83.24 ± 0.13

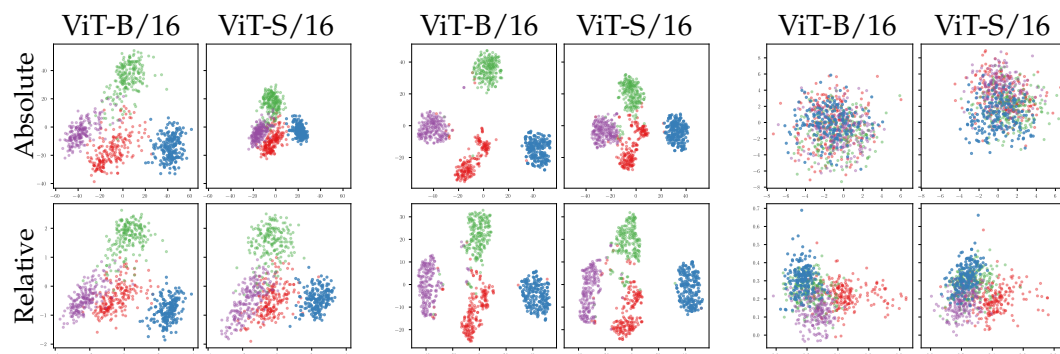


FIGURE A.5: Different dimensionality reduction techniques applied to absolute and relative spaces on CIFAR-10. From left to right: PCA (Principal Component Analysis), tSNE, and visualizing only their first two dimensions. Only 800 randomly sampled points are shown, belonging to the classes "bird", "ship", "cat", and "frog".

Appendix B

Direct Translation

B.1 Additional results

In Figure B.3, we present the outcomes of the multimodal experiment presented in Section 5.3.2 with an MLP employed as the classification head, instead of SVMs. The findings highlight the MLP’s capability to leverage cross-modal information, leading to improved performance. However, the underlying mechanisms responsible for this enhancement remain unclear and warrant further investigation.

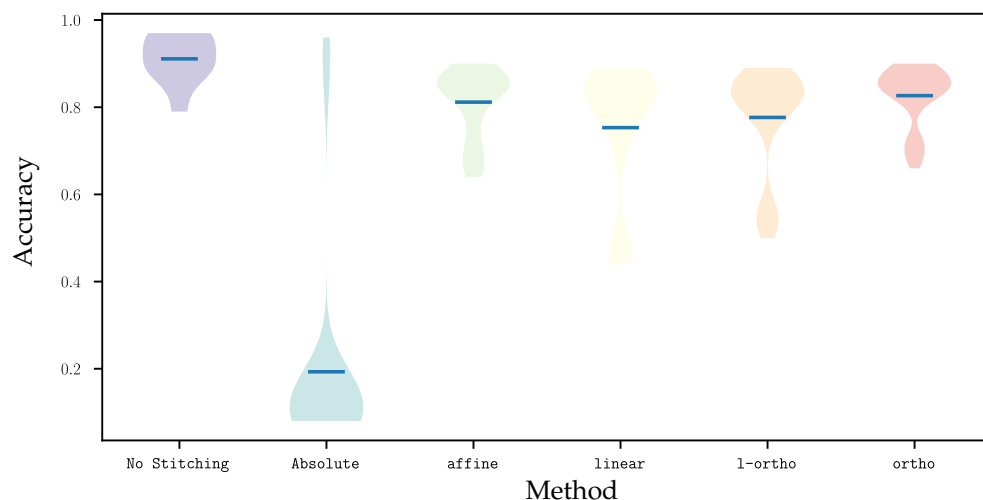


FIGURE B.1: Cross-domain stitching on CIFAR-10 and grayscale CIFAR-10. 84 stitched pairs (pre-trained encoder - SVM classifier) for 5 different seeds.

In Tables B.2 and B.3 quantitative results for stitching of MLP classifiers (again, differently from Tables 5.1 and 5.2 where SVMs are used) trained on top of pre-trained feature extractors, with and without additional L2 normalization, respectively.

In Figures B.4 and B.5, there are additional reconstruction examples with the same autoencoding setting as in Figure 5.6, and with additional L2 normalization, respectively.

In Table B.1 there are more quantitative results for stitching of autoencoders, with added L2 normalization (at training time) to the decoders of the reconstruction setting of Table 5.3.

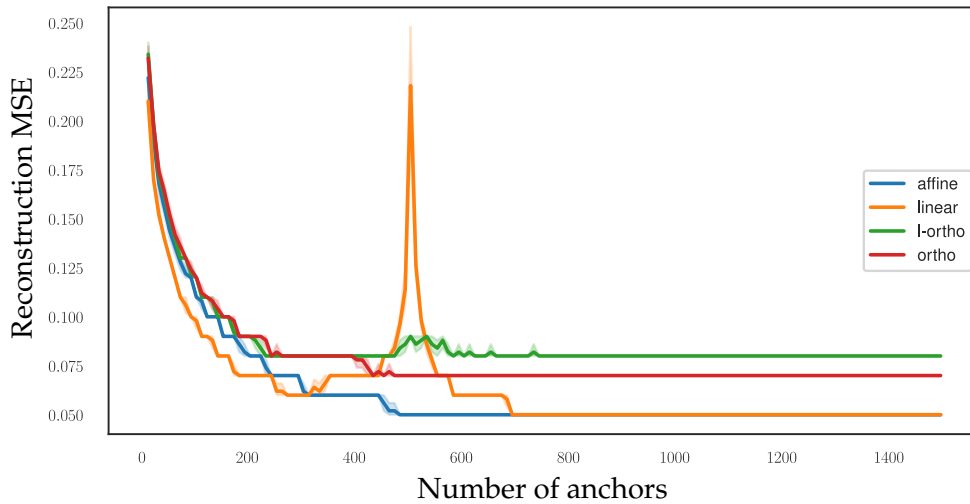


FIGURE B.2: Performance comparison (reconstruction error) of *affine*, *linear*, *l-ortho* and *ortho* at varying anchor number on reconstruction task. Results on stitching 2 different CIFAR-100-trained AEs with 5 samplings for each anchor quantity. The naive absolute baseline is flat on 0.38 as mean.

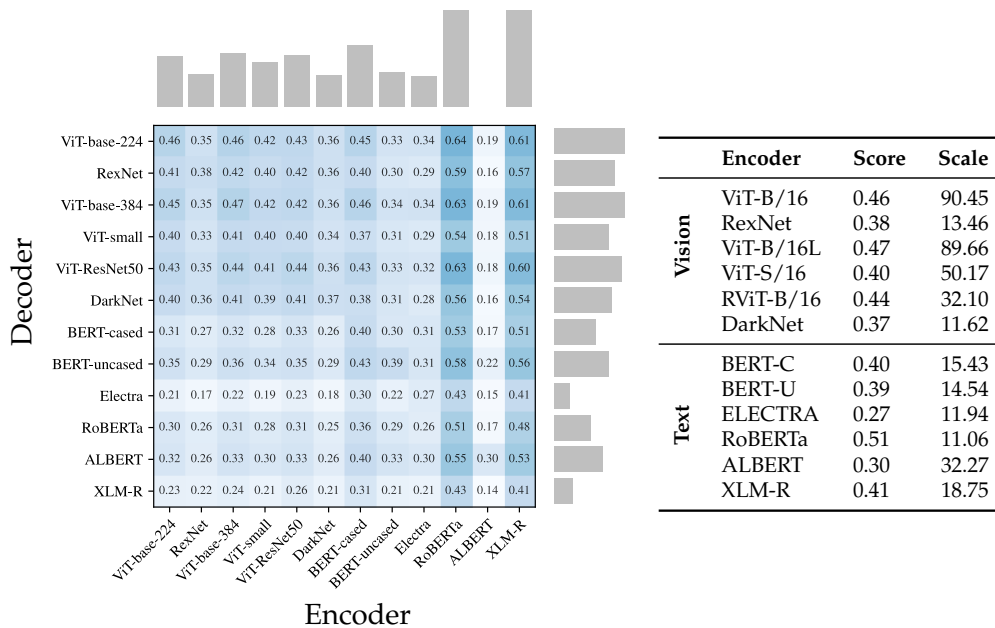


FIGURE B.3: Performance comparison between different encoders and data modalities on the N24News multimodal dataset. On the right, the accuracy of models trained end-to-end on a single data modality (Score) and their average norm (Scale). On the left the stitching performance between pairs of encoders and decoder. This shows the importance of translating from good encoders, that can even improve unimodal decoder performances. Results obtained with 2000 anchors and SVD, with a MLP as classification head.

B.1.1 Scale invariance

In this Section, we delve into the concept of scale invariance in NNs and its implications for model compositionality. We start by focusing on the effect of rescaling

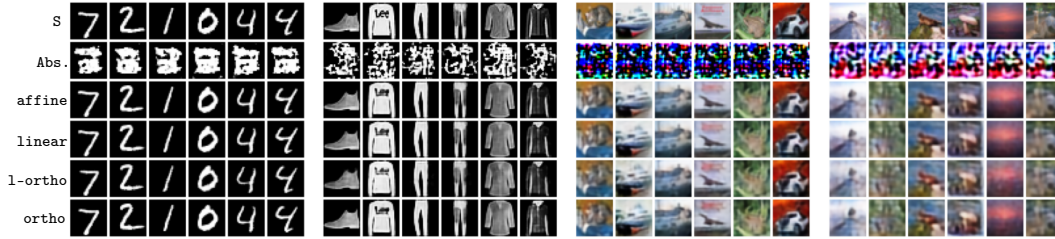


FIGURE B.4: Reconstruction examples grouped by dataset. Each column is a different image, from top to bottom: original image, absolute stitching, LSS stitching, OLSS stitching, and SVD stitching. An L2 normalization is applied to the decoder input.

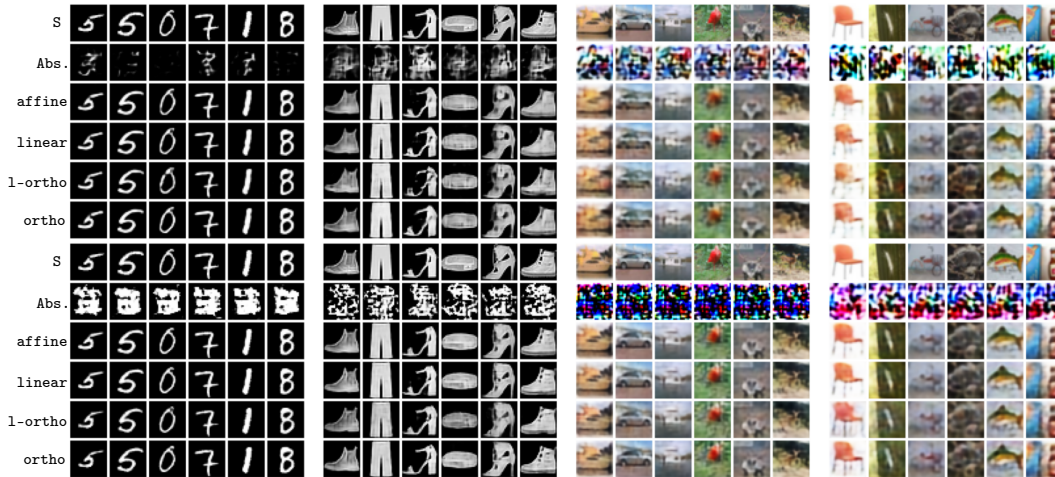


FIGURE B.5: Additional reconstruction examples grouped by dataset. Each column is a different image, from top to bottom: original image, absolute stitching, LSS stitching, OLSS stitching, and SVD stitching. In the first row, no additional normalization is applied on the decoder input; in the second row, an L2 normalization is applied instead.

TABLE B.1: Zero-shot stitching for generation. With SVD for estimating \hat{T} and standard scaling as pre-processing. An L2 normalization is applied to the decoder input. We report the latent cosine similarity ($lcos$) and MSE ($lmse$) between the target encoding and the translated one, but also the reconstruction MSE ($rmse$) between the input and the output.

	MNIST			F-MNIST			CIFAR-10			CIFAR-100		
	$lcos$	$lmse$	$rmse$	$lcos$	$lmse$	$rmse$	$lcos$	$lmse$	$rmse$	$lcos$	$lmse$	$rmse$
Abs.	0.39	0.98	0.28	0.53	0.97	0.33	0.62	1.23	0.46	0.59	1.17	0.38
affine	0.99	0.15	0.01	0.99	0.16	0.03	0.99	0.16	0.04	0.99	0.12	0.05
linear	0.98	0.17	0.01	0.98	0.18	0.03	0.99	0.16	0.04	0.99	0.13	0.05
1-ortho	0.89	0.41	0.02	0.91	0.41	0.04	0.96	0.39	0.05	0.93	0.30	0.08
ortho	0.97	0.21	0.02	0.97	0.23	0.03	0.99	0.21	0.05	0.96	0.22	0.07

operations on the latent input encodings and demonstrate that, by construction, certain classifiers exhibit scale-invariance properties without the need for additional priors. Then, by examining the behavior of networks when subjected to a specific type of input manipulation, *rescaling injection*, we aim to demonstrate the robustness and versatility of NNs in handling different scales of input data. As illustrated in Chapter 5, this is a key advantage in improving the adaptability of our method.

TABLE B.2: Cross-architecture stitching with various methods for estimating \hat{T} and employing standard scaling. The stitched decoders are simple MLPs. 5 runs for each encoder-decoder pair. (C) and (F) next to CIFAR-100 indicate, respectively, coarse-grained and fine-grained.

	Dataset	no-stitch	absolute	relative	affine	linear	l-ortho	ortho
Vision	CIFAR-10	0.95 ± 0.03	0.16 ± 0.22	0.73 ± 0.21	0.93 ± 0.05	0.89 ± 0.11	0.90 ± 0.09	0.93 ± 0.04
	CIFAR-100-C	0.82 ± 0.07	0.11 ± 0.21	0.39 ± 0.17	0.76 ± 0.08	0.71 ± 0.15	0.74 ± 0.11	0.78 ± 0.07
	CIFAR-100-F	0.68 ± 0.14	0.06 ± 0.20	0.13 ± 0.09	0.59 ± 0.13	0.55 ± 0.18	0.56 ± 0.17	0.62 ± 0.12
	F-MNIST	0.87 ± 0.02	0.14 ± 0.20	0.64 ± 0.12	0.85 ± 0.02	0.83 ± 0.05	0.80 ± 0.06	0.84 ± 0.02
	MNIST	0.92 ± 0.03	0.15 ± 0.20	0.36 ± 0.14	0.92 ± 0.03	0.87 ± 0.08	0.74 ± 0.12	0.88 ± 0.03
Text	TREC	0.41 ± 0.07	0.15 ± 0.04	0.27 ± 0.09	0.40 ± 0.08	0.37 ± 0.11	0.23 ± 0.08	0.41 ± 0.09
	AG News	0.76 ± 0.08	0.24 ± 0.02	0.36 ± 0.10	0.68 ± 0.08	0.65 ± 0.08	0.64 ± 0.10	0.68 ± 0.10
	DBpedia	0.64 ± 0.19	0.07 ± 0.02	0.15 ± 0.08	0.57 ± 0.19	0.53 ± 0.19	0.44 ± 0.21	0.56 ± 0.17
	IMDB	0.62 ± 0.04	0.50 ± 0.01	0.50 ± 0.01	0.59 ± 0.04	0.58 ± 0.04	0.57 ± 0.04	0.60 ± 0.04

TABLE B.3: Cross-architecture stitching with various methods for estimating \hat{T} and applying L2 as normalization. The stitched decoders are simple MLPs. 5 runs for each encoder-decoder pair. (C) and (F) next to CIFAR-100 indicate, respectively, coarse-grained and fine-grained.

	Dataset	no-stitch	absolute	relative	affine	linear	l-ortho	ortho
Vision	CIFAR-10	0.95 ± 0.03	0.16 ± 0.22	0.73 ± 0.21	0.93 ± 0.04	0.89 ± 0.11	0.89 ± 0.11	0.93 ± 0.04
	CIFAR-100-C	0.82 ± 0.07	0.11 ± 0.21	0.39 ± 0.17	0.77 ± 0.07	0.75 ± 0.13	0.71 ± 0.15	0.78 ± 0.06
	CIFAR-100-F	0.68 ± 0.14	0.06 ± 0.20	0.13 ± 0.09	0.60 ± 0.12	0.57 ± 0.18	0.54 ± 0.18	0.61 ± 0.12
	F-MNIST	0.87 ± 0.02	0.14 ± 0.20	0.64 ± 0.12	0.86 ± 0.02	0.79 ± 0.09	0.83 ± 0.05	0.84 ± 0.02
	MNIST	0.92 ± 0.03	0.15 ± 0.20	0.36 ± 0.14	0.91 ± 0.03	0.80 ± 0.17	0.86 ± 0.08	0.86 ± 0.04
Text	TREC	0.41 ± 0.07	0.15 ± 0.04	0.27 ± 0.09	0.51 ± 0.06	0.27 ± 0.10	0.47 ± 0.13	0.49 ± 0.06
	AG News	0.76 ± 0.08	0.24 ± 0.02	0.36 ± 0.10	0.68 ± 0.08	0.64 ± 0.10	0.65 ± 0.08	0.66 ± 0.10
	DBpedia	0.64 ± 0.19	0.07 ± 0.02	0.15 ± 0.08	0.55 ± 0.19	0.53 ± 0.21	0.51 ± 0.18	0.49 ± 0.15
	IMDB	0.62 ± 0.04	0.50 ± 0.01	0.50 ± 0.01	0.60 ± 0.04	0.58 ± 0.04	0.59 ± 0.04	0.59 ± 0.04

The softmax function, commonly used in neural classifiers, is known to be a temperature-controlled variant of the maximum function:

$$\text{softmax}(x)_i = \frac{e^{\frac{y_i}{T}}}{\sum_j^N e^{\frac{y_j}{T}}}. \quad (\text{B.1})$$

This means that the softmax temperature can be used to control the level of confidence of the classifier’s predictions. In this study, we show that a similar effect can also be achieved by rescaling the latent encodings given as input to a trained (and frozen) classifier.

In order to demonstrate this, we first note that the rescaling factor, α , can be factored out of the matrix multiplication in the Linear layers of the classifier. This can be represented mathematically as: $\mathbf{y} = \alpha \mathbf{W}\mathbf{x} + b$, where \mathbf{x} is the input latent encoding, \mathbf{W} is the weight matrix, b is the bias vector, α is the rescaling factor, and \mathbf{y} is the output of the linear layer. This implies that the rescaling operation can be “pushed through” the classifier without affecting its final prediction as it becomes equivalent to some temperature value applied at the softmax level.

Furthermore, we investigate the effect of rescaling when non-linear activation functions are involved and posit that as long as the function has a monotonic interval, if we rescale all the dimensions by an amount similar to the mean scale of the encodings on which the classifier was trained, we end up in the monotonic interval, without losing the scale-invariance property.

In summary, our study provides empirical evidence that neural classifiers that utilize the softmax activation function can, in practice, maintain their scale-invariance properties when the input latent encodings are rescaled. This property is essential

to our method, as it allows us to ignore the exact scale when decoding toward an L2-normalized absolute space.

Pre-trained models and scale-invariance. We observed that large pre-trained models, such as transformers and resnets, are robust to internal rescaling of the encodings. Although we do not have a strong theoretical explanation for this phenomenon, we hypothesize that normalization layers and the linear separability of the information encoded in the angles instead of the norms may play a significant role. In Figure B.6, we demonstrate the invariance a large transformer exhibits when the rescaling injection is applied at different layers: surprisingly, when the rescaling surpasses a certain threshold, the performance difference becomes negligible. These results further emphasize the robustness of these pre-trained models to the rescaling injection and suggest that the scale of the embedding is not a critical factor in their performance.

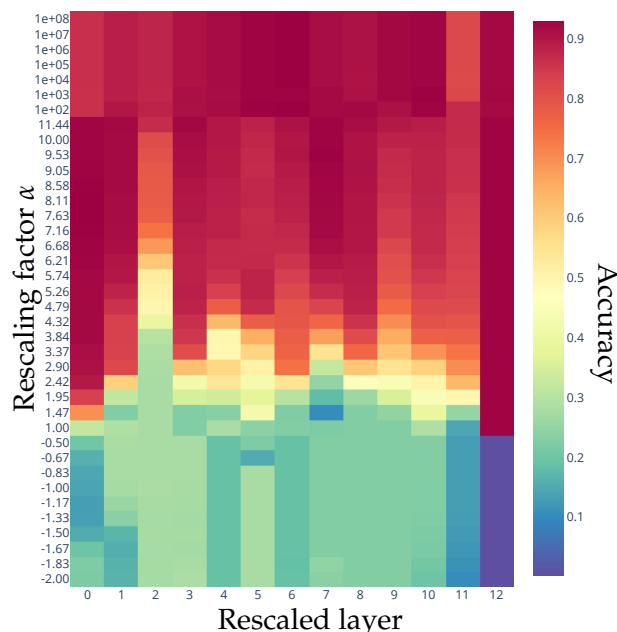


FIGURE B.6: Scale invariance of RoBERTa according to the performance of a downstream classifier trained on the encodings of the last attention layer. At each layer (with 0 being the embedding layer and 12 the output one), one for each run, we rescale the encodings by the specified α and measure its effect on the final accuracy. The performance without any rescaling is 0.92.

Rescale Injection. We define the *rescaling injection* as the operation of artificially altering the scale of the features produced at a specific layer of the network. This is achieved by normalizing the embeddings to unit norm and then rescaling them by a factor of α . By varying the value of α , we can observe how the network’s performance is affected at different scales. Through this empirical analysis, we aim to provide insight into the scale invariance properties of NNs and their potential for use in model compositionality.

In Figure B.7, we present experimental results investigating the scale invariance properties of NNs. We trained simple multi-layer perceptrons (MLPs) composed of two hidden layers, with no normalization layers, using encodings produced by the Clip Vision transformer (clip-vit-base-patch32) on the CIFAR-100 (fine) dataset.

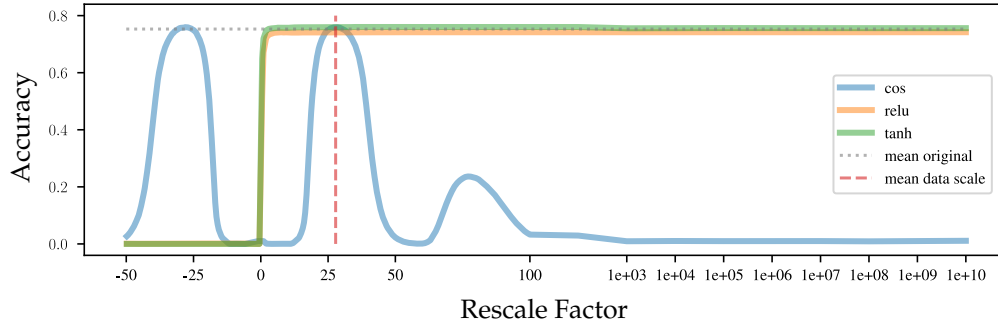


FIGURE B.7: Performance comparison of three MLPs with different activation functions, namely cosine (blue), ReLU (orange), and tanh (green) at different rescaling factors α . The ReLU and tanh MLPs exhibit scale invariance, while the cosine activation function is only invariant on the mean data scale and its periodic cycles.

The MLPs were evaluated using different activation functions: cosine (blue), tanh (orange), and ReLU (green). The rescaling injection technique was applied directly to the input embeddings, rescaling them by α .

We can observe that the scale of the embeddings does not significantly impact the MLPs' performance when using monotone activation functions that do not flip signs. This is a non-trivial result, as the nonlinearity of the activation function, the presence of bias terms b , and the absence of normalization layers make it difficult to predict the effect of an input rescaling on the performance of the network. It is particularly interesting to see that the cosine activation function shows an oscillatory performance, comparable to the original embeddings when rescaled by the mean embeddings scale (vertical red line) or its opposite since it is symmetric.

Our findings indicate that, surprisingly, even the internal layers of large deep learning models exhibit a *positive scale invariance*, as illustrated in Figure B.6. The underlying mechanism for this behavior is not straightforward, but we hypothesize that it may result from the interplay between various factors, such as the choice of activation function, the use of normalization layers, the optimization objective and regularization techniques employed during the training phase. Further research is needed to understand and explain this phenomenon fully.

B.1.2 Implementation Details

All the experiments were conducted using a machine equipped with an Intel Core i7-9700k CPU, 64 GB of RAM, and an NVIDIA 2080TI GPU.

Decoder structure. The full implementation details can be found in the attached code, the various experiments can be run by their corresponding notebook.

- *Autoencoding.* Since the autoencoders were used only on image data, the architecture was a simple sequence of convolutions (in the encoder part) and deconvolutions (in the decoder part). Each interleaved with nonlinear activations.
- *Classification.* Chapter 5 refers to "SVM" as the standard SVM implementation in scikit-learn (Pedregosa et al., 2011), with default parameters. The experiments with "MLP" as a classifier refer to a simple stack of 3 linear layers, interleaved by nonlinear activations.

Software and Technologies. The research of this study was facilitated by the use of various technologies and tools, which include:

- *NN-Template* (GrokAI, 2021), was used to kick-start the project while also ensuring best practices were adhered to.
- *DVC* (Kuprieiev et al., 2023), was implemented for data versioning.
- *PyTorch Lightning* (Falcon and The PyTorch Lightning team, 2019), contributed to maintaining the integrity of the results and promoting clean, modular code.
- *Weights and Biases* (Biewald, 2020), were employed for logging experiments, running comparisons over extensive sweeps, and sharing models.
- *Transformers by HuggingFace* (Wolf et al., 2020), provided pre-configured transformers for processing both image and text data.
- *Datasets by HuggingFace* (Lhoest et al., 2021), facilitated access to a majority of NLP datasets and ImageNet for computer vision purposes.

Pre-trained encoders. All the pre-trained encoders used come from HuggingFace and are listed in Table B.4. They are various both in terms of architecture and encoding size.

TABLE B.4: HuggingFace models used as encoders (feature extractors) in the various experiments, with their encoding dimensionality.

Modality	HuggingFace model name	Encoding Dim
Language	bert-base-cased	768
	bert-base-uncased	768
	google/electra-base-discriminator	768
	roberta-base	768
	albert-base-v2	768
	xlm-roberta-base	768
	openai/clip-vit-base-patch32	768
Vision	resnet101	1280
	cspdarknet53	768
	vit_small_patch16_224	384
	vit_base_patch16_224	768
	vit_base_patch16_384	768
	vit_base_resnet50_384	768
	openai/clip-vit-base-patch32	768

TABLE B.5: Cross-architecture stitching for reconstruction tasks. 5 different seeds, 2 different bottleneck sizes (250, 500) for the same architecture. Average over all combinations. 500 anchors used and standard scaling as normalization. The naive absolute baseline is impossible to compute due to the dimensionality mismatch.

	MNIST			F-MNIST			CIFAR-10			CIFAR-100		
	<i>lcos</i>	<i>lmse</i>	<i>rmse</i>	<i>lcos</i>	<i>lmse</i>	<i>rmse</i>	<i>lcos</i>	<i>lmse</i>	<i>rmse</i>	<i>lcos</i>	<i>lmse</i>	<i>rmse</i>
affine	0.95	0.09	0.02	0.95	0.09	0.04	0.98	0.06	0.05	0.98	0.07	0.06
linear	0.64	1.00	0.11	0.66	1.10	0.16	0.77	0.60	0.16	0.78	0.52	0.16
1-ortho	0.87	0.16	0.03	0.89	0.14	0.06	0.95	0.12	0.08	0.95	0.13	0.08
ortho	0.91	0.14	0.03	0.92	0.13	0.06	0.96	0.12	0.09	0.96	0.12	0.09

Bibliography

- Cannistraci, Irene, **Luca Moschella**, Marco Fumero, Valentino Maiorca, and Emanuele Rodolà (2024). “From Bricks to Bridges: Product of Invariances to Enhance Latent Space Communication”. In: *The Twelfth International Conference on Learning Representations (ICLR 2024, spotlight, top 5%)*. URL: <https://openreview.net/forum?id=vngVydDWft>.
- Kiefer, Alex B. and Christopher L. Buckley (2024). “Relative Representations for Cognitive Graphs”. In: *Active Inference*. Ed. by Christopher L. Buckley, Daniela Cialfi, Pablo Lanillos, Maxwell Ramstead, Noor Sajid, Hideaki Shimazaki, Tim Verbelen, and Martijn Wisse. Cham: Springer Nature Switzerland, pp. 218–236. URL: https://link.springer.com/chapter/10.1007/978-3-031-47958-8_14.
- Wu, Zijun, Yongkang Wu, and Lili Mou (2024). “Zero-Shot Continuous Prompt Transfer: Generalizing Task Semantics Across Language Models”. In: *The Twelfth International Conference on Learning Representations*. URL: <https://openreview.net/forum?id=26Xphug0cS>.
- Ye, Haotian, Yihong Liu, Chunlan Ma, and Hinrich Schütze (Jan. 9, 2024). *MoSECroT: Model Stitching with Static Word Embeddings for Crosslingual Zero-shot Transfer*. DOI: [10.48550/arXiv.2401.04821](https://doi.org/10.48550/arXiv.2401.04821). arXiv: 2401.04821 [cs]. URL: <http://arxiv.org/abs/2401.04821> (visited on 02/27/2024). preprint.
- Acosta, Francisco, Colin Conwell, Sophia Sanborn, David A. Klindt, and Nina Miolane (2023). “Evaluation of Representational Similarity Scores Across Human Visual Cortex”. In: *UniReps: the First Workshop on Unifying Representations in Neural Models*. URL: <https://openreview.net/forum?id=LhV3Ex8fky>.
- Ainsworth, Samuel K., Jonathan Hayase, and Siddhartha Srinivasa (Mar. 1, 2023). *Git Re-Basin: Merging Models modulo Permutation Symmetries*. DOI: [10.48550/arXiv.2209.04836](https://doi.org/10.48550/arXiv.2209.04836). arXiv: 2209.04836 [cs]. URL: <http://arxiv.org/abs/2209.04836> (visited on 04/28/2023). preprint.
- Cannistraci, Irene, **Luca Moschella**, Valentino Maiorca, Marco Fumero, Antonio Norelli, and Emanuele Rodolà (2023). “Bootstrapping Parallel Anchors for Relative Representations”. In: *The First Tiny Papers Track at ICLR 2023, Tiny Papers at ICLR 2023*. URL: <https://openreview.net/pdf?id=VBuUL2IWlq>.
- Chen, Chi, Peng Li, Maosong Sun, and Yang Liu (July 2023). “Weakly Supervised Vision-and-Language Pre-training with Relative Representations”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2023. Toronto, Canada: Association for Computational Linguistics, pp. 8341–8355. URL: <https://aclanthology.org/2023.acl-long.464>.
- Crisostomi, Donato, Irene Cannistraci, **Luca Moschella**, Pietro Barbiero, Marco Ciccone, Pietro Lio, and Emanuele Rodolà (2023). “From Charts to Atlas: Merging Latent Spaces into One”. In: *NeurIPS 2023 Workshop on Symmetry and Geometry in Neural Representations*. URL: <https://openreview.net/forum?id=ZFu7CPTznY>.
- Frascaroli, Emanuele, Riccardo Benaglia, Matteo Boschini, **Luca Moschella**, Cosimo Fiorini, Emanuele Rodolà, and Simone Calderara (2023). “CaSpeR: Latent Spectral Regularization for Continual Learning”. In: *CoRR* abs/2301.03345. URL: <https://doi.org/10.48550/arXiv.2301.03345>.

- Ghosh, Shubhangi, Luigi Gresele, Julius von Kügelgen, Michel Besserve, and Bernhard Schölkopf (2023). *Independent Mechanism Analysis and the Manifold Hypothesis*. arXiv: 2312.13438 [stat.ML].
- Illharco, Gabriel, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hananeh Hajishirzi, and Ali Farhadi (Feb. 1, 2023). "Editing Models with Task Arithmetic". In: URL: <https://openreview.net/forum?id=6t0Kwf8-jrj> (visited on 05/18/2023).
- Jian, Pingcheng, Easop Lee, Zachary Bell, Michael M. Zavlanos, and Boyuan Chen (2023). "Policy Stitching: Learning Transferable Robot Policies". In: *7th Annual Conference on Robot Learning*. URL: <https://openreview.net/forum?id=2qKBwyLnln>.
- Klabunde, Max, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich (2023). *Similarity of Neural Network Models: A Survey of Functional and Representational Measures*. arXiv: 2305.06329 [cs.LG].
- Kupriev, Ruslan, skshetry, Dmitry Petrov, Peter Rowlands, Paweł Redzyński, Casper da Costa-Luis, Alexander Schepanovski, Gao, David de la Iglesia Castro, Ivan Shcheklein, Batuhan Taskaya, Jorge Orpinel, Fábio Santos, Dave Berenbaum, daniele, Ronan Lamy, Aman Sharma, Zhanibek Kaimuldenov, Dani Hodovic, Nikita Kodenko, Andrew Grigorev, Earl, Nabanita Dash, George Vyshnya, maykulkarni, Max Hora, Vera, and Sanidhya Mangal (2023). *DVC: Data Version Control - Git for Data & Models*. Version 2.45.1. DOI: 10.5281/zenodo.7646429. URL: <https://doi.org/10.5281/zenodo.7646429>.
- Lähner, Zorah and Michael Moeller (2023). "On the Direct Alignment of Latent Spaces". In: *UniReps: the First Workshop on Unifying Representations in Neural Models*. URL: <https://openreview.net/forum?id=nro8tEfIfw>.
- Maiorca*, Valentino, **Luca Moschella***, Antonio Norelli, Marco Fumero, Francesco Locatello, and Emanuele Rodolà (2023). "Latent Space Translation via Semantic Alignment". In: *Thirty-seventh Conference on Neural Information Processing Systems*. URL: <https://openreview.net/forum?id=pBa70rGHlr>.
- Marchetti, Giovanni Luca and Christopher Hillar (Nov. 2023). *Harmonics of Learning: Universal Fourier Features Emerge in Invariant Networks*. <https://synthical.com/article/03032df5-f9c5-43a6-8253-d4af1d67e0d4>. arXiv: 2312.08550 [cs.AI].
- Moschella***, **Luca**, Valentino Maiorca*, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà (2023). "Relative representations enable zero-shot latent space communication". In: *The Eleventh International Conference on Learning Representations (ICLR 2023, oral, notable top 5%)*. URL: <https://openreview.net/forum?id=SrC-nwieGJ>.
- Norelli, Antonio, Marco Fumero, Valentino Maiorca, **Luca Moschella**, Emanuele Rodolà, and Francesco Locatello (2023). "ASIF: Coupled Data Turns Unimodal Models to Multimodal without Training". In: *Thirty-seventh Conference on Neural Information Processing Systems*. URL: <https://openreview.net/forum?id=Xj0j3ZmWE1>.
- Ortiz-Jimenez, Guillermo, Alessandro Favero, and Pascal Frossard (2023). "Task Arithmetic in the Tangent Space: Improved Editing of Pre-Trained Models". Version 2. In: DOI: 10.48550/ARXIV.2305.12827. URL: <https://arxiv.org/abs/2305.12827> (visited on 07/02/2023).
- Rakotonirina, Nathanaël Carraz, Roberto Dessi, Fabio Petroni, Sebastian Riedel, and Marco Baroni (2023). "Can discrete information extraction prompts generalize across language models?" In: *The Eleventh International Conference on Learning Representations*. URL: <https://openreview.net/forum?id=sbWVtxq8-zE>.

- Ramé, Alexandre, Kartik Ahuja, Jianyu Zhang, Matthieu Cord, Léon Bottou, and David Lopez-Paz (Jan. 27, 2023). *Model Ratatouille: Recycling Diverse Models for Out-of-Distribution Generalization*. arXiv: 2212.10445 [cs]. URL: <http://arxiv.org/abs/2212.10445> (visited on 04/28/2023). preprint.
- Ramos, Patrick, Raphael Alampay, and Patricia Abu (2023). “Knowledge Distillation with Relative Representations for Image Representation Learning”. In: *Progress on Pattern Classification, Image Processing and Communications*. Ed. by Robert Burduk, Michał Choraś, Rafał Kozik, Paweł Ksieniewicz, Tomasz Marciniak, and Paweł Trajdos. Lecture Notes in Networks and Systems. Cham: Springer Nature Switzerland, pp. 133–143. ISBN: 978-3-031-41630-9. DOI: 10.1007/978-3-031-41630-9_14.
- Rath, Matthias and Alexandru Paul Condurache (2023). “Deep neural networks with efficient guaranteed invariances”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 2460–2480.
- Ricciardi, Antonio Pio, Valentino Maiorca, **Luca Moschella**, and Emanuele Rodolà (2023). “Zero-shot stitching in Reinforcement Learning using Relative Representations”. In: *Sixteenth European Workshop on Reinforcement Learning*. URL: <https://openreview.net/forum?id=4tcXsImfsS1>.
- Srivastava, Aarohi et al. (2023). “Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models”. In: *Transactions on Machine Learning Research*. ISSN: 2835-8856. URL: <https://openreview.net/forum?id=uyTL5Bvosj>.
- Sucholutsky, Ilia and Thomas L. Griffiths (Nov. 2, 2023). “Alignment with Human Representations Supports Robust Few-Shot Learning”. In: *Thirty-Seventh Conference on Neural Information Processing Systems*. URL: <https://openreview.net/forum?id=HYGnmSLBCf> (visited on 02/27/2024).
- Sucholutsky, Ilia, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Erin Grant, Iris Groen, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine L. Hermann, Kerem Oktar, Klaus Greff, Martin N. Hebart, Nori Jacoby, Qiuyi Zhang, Raja Marjeh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas P. O’Connell, Thomas Unterthiner, Andrew K. Lampinen, Klaus-Robert Müller, Mariya Toneva, and Thomas L. Griffiths (Nov. 2, 2023). *Getting Aligned on Representational Alignment*. DOI: 10.48550/arXiv.2310.13018. arXiv: 2310.13018 [cs, q-bio]. URL: <http://arxiv.org/abs/2310.13018> (visited on 02/27/2024). preprint.
- Wang, Gary, Kyle Kastner, Ankur Bapna, Zhehuai Chen, Andrew Rosenberg, Bhuvana Ramabhadran, and Yu Zhang (2023). “Understanding Shared Speech-Text Representations”. In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10095508.
- Wang, Haonan, Minbin Huang, Runhui Huang, Lanqing Hong, Hang Xu, Tianyang Hu, Xiaodan Liang, and Zhenguo Li (May 9, 2023). “Boosting Visual-Language Models by Exploiting Hard Samples”. In: DOI: 10.48550/arXiv.2305.05208. arXiv: 2305.05208 [cs]. URL: <http://arxiv.org/abs/2305.05208> (visited on 02/27/2024). preprint.
- Barannikov, Serguei, Ilya Trofimov, Nikita Balabin, and Evgeny Burnaev (2022). “Representation Topology Divergence: A Method for Comparing Neural Network Representations.” In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato. Vol. 162. Proceedings of Machine Learning Research. PMLR, pp. 1607–1626. URL: <https://proceedings.mlr.press/v162/barannikov22a.html>.

- Bonheme, Lisa and Marek Grzes (2022). “How do Variational Autoencoders Learn? Insights from Representational Similarity”. In: *ArXiv preprint abs/2205.08399*. URL: <https://arxiv.org/abs/2205.08399>.
- Chang, Tyler A, Zhuowen Tu, and Benjamin K Bergen (2022). “The Geometry of Multilingual Language Model Representations”. In: *ArXiv preprint abs/2205.10964*. DOI: [10.18653/v1/2022.emnlp-main.9](https://doi.org/10.18653/v1/2022.emnlp-main.9). arXiv: [2205.10964](https://arxiv.org/abs/2205.10964) [cs.CL]. URL: <https://arxiv.org/abs/2205.10964>.
- Crisostomi, Donato, Simone Antonelli, Valentino Maiorca, **Luca Moschella**, Riccardo Marin, and Emanuele Rodolà (2022). “Metric Based Few-Shot Graph Classification”. In: *The First Learning on Graphs Conference*. URL: <https://openreview.net/forum?id=VBXRMnRBfRF>.
- Davari, MohammadReza, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky (2022). “Reliability of CKA as a Similarity Measure in Deep Learning”. In: *arXiv preprint arXiv:2210.16156*.
- Entezari, Rahim, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur (July 5, 2022). “The Role of Permutation Invariance in Linear Mode Connectivity of Neural Networks”. DOI: [10.48550/arXiv.2110.06296](https://doi.org/10.48550/arXiv.2110.06296). arXiv: [2110.06296](https://arxiv.org/abs/2110.06296) [cs]. URL: <http://arxiv.org/abs/2110.06296> (visited on 04/28/2023).
- Higgins, Irina, Sébastien Racanière, and Danilo Rezende (2022). *Symmetry-Based Representations for Artificial and Biological General Intelligence*. arXiv: [2203.09250](https://arxiv.org/abs/2203.09250) [q-bio.NC].
- Immer, Alexander, Tycho van der Ouderaa, Gunnar Rätsch, Vincent Fortuin, and Mark van der Wilk (2022). “Invariance learning in deep neural networks with differentiable Laplace approximations”. In: *Advances in Neural Information Processing Systems 35*, pp. 12449–12463.
- Matena, Michael S. and Colin A. Raffel (Dec. 6, 2022). “Merging Models with Fisher-Weighted Averaging”. In: *Advances in Neural Information Processing Systems 35*, pp. 17703–17716. URL: https://proceedings.neurips.cc/paper_files/paper/2022/hash/70c26937fbf3d4600b69a129031b66ec-Abstract-Conference.html (visited on 04/28/2023).
- Mehta, Raghav, Vítor Albiero, Li Chen, Ivan Evtimov, Tamar Glaser, Zhiheng Li, and Tal Hassner (2022). “You Only Need a Good Embeddings Extractor to Fix Spurious Correlations”. In: arXiv: [2212.06254](https://arxiv.org/abs/2212.06254) [cs.CV].
- Moschella, Luca**, Simone Melzi, Luca Cosmo, Filippo Maggioli, Or Litany, Maks Ovsjanikov, Leonidas Guibas, and Emanuele Rodolà (2022). “Learning Spectral Unions of Partial Deformable 3D Shapes”. In: *Computer Graphics Forum 41.2*, pp. 407–417. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.14483>.
- Norelli, Antonio, Giorgio Mariani, **Luca Moschella**, Andrea Santilli, Giambattista Parascandolo, Simone Melzi, and Emanuele Rodolà (2022). “Explanatory Learning: Beyond Empiricism in Neural Networks”. In: *CoRR abs/2201.10222*. URL: <https://arxiv.org/abs/2201.10222>.
- Ouderaa, Tycho FA van der and Mark van der Wilk (2022). “Learning invariant weights in neural networks”. In: *Uncertainty in Artificial Intelligence*. PMLR, pp. 1992–2001.
- Oyama, Momose, Sho Yokoi, and Hidetoshi Shimodaira (2022). “Norm of word embedding encodes information gain”. In: *ArXiv abs/2212.09663*. DOI: [10.18653/v1/2023.emnlp-main.131](https://doi.org/10.18653/v1/2023.emnlp-main.131). URL: <https://api.semanticscholar.org/CorpusID:254853643>.

- Shalam, Daniel and Simon Korman (2022). “The Self-Optimal-Transport Feature Transform”. In: *ArXiv preprint abs/2204.03065*. URL: <https://arxiv.org/abs/2204.03065>.
- Sliwa, Joanna, Shubhangi Ghosh, Vincent Stimper, Luigi Gresele, and Bernhard Schölkopf (2022). “Probing the Robustness of Independent Mechanism Analysis for Representation Learning”. In: *UAI 2022 Workshop on Causal Representation Learning*. URL: <https://openreview.net/forum?id=MnKSQVBVpBQ>.
- Somepalli, Gowthami, Liam Fowl, Arpit Bansal, Ping Yeh-Chiang, Yehuda Dar, Richard Baraniuk, Micah Goldblum, and Tom Goldstein (2022). “Can Neural Nets Learn the Same Model Twice? Investigating Reproducibility and Double Descent from the Decision Boundary Perspective”. In: *ArXiv preprint abs/2203.08124*. DOI: [10.1109/cvpr52688.2022.01333](https://doi.org/10.1109/cvpr52688.2022.01333). arXiv: [2203.08124](https://arxiv.org/abs/2203.08124) [cs.LG]. URL: <https://arxiv.org/abs/2203.08124>.
- Wang, Zhen, Xu Shan, Xiangxie Zhang, and Jie Yang (June 2022). “N24News: A New Dataset for Multimodal News Classification”. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis. Marseille, France: European Language Resources Association, pp. 6768–6775. URL: <https://aclanthology.org/2022.lrec-1.729>.
- Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus (June 26, 2022). “Emergent Abilities of Large Language Models”. In: *Transactions on Machine Learning Research*. ISSN: 2835-8856. URL: <https://openreview.net/forum?id=yzkSU5zdwD> (visited on 02/27/2024).
- Wortsman, Mitchell, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt (June 28, 2022). “Model Soups: Averaging Weights of Multiple Fine-Tuned Models Improves Accuracy without Increasing Inference Time”. In: *Proceedings of the 39th International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, pp. 23965–23998. URL: <https://proceedings.mlr.press/v162/wortsman22a.html> (visited on 04/28/2023).
- Yaman, Muammer Y., Sergei V. Kalinin, Kathryn N. Guye, David Ginger, and Maxim Ziatdinov (2022). *Learning and predicting photonic responses of plasmonic nanoparticle assemblies via dual variational autoencoders*. DOI: [10.1002/sml1.202205893](https://doi.org/10.1002/sml1.202205893). URL: <https://arxiv.org/abs/2208.03861>.
- Zhai, Xiaohua, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer (2022a). “LiT: Zero-Shot Transfer With Locked-Image Text Tuning”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18123–18133. DOI: [10.1109/cvpr52688.2022.01759](https://doi.org/10.1109/cvpr52688.2022.01759).
- (2022b). “Lit: Zero-shot transfer with locked-image text tuning”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18123–18133.
- Antonello, Richard, Javier S Turek, Vy Vo, and Alexander Huth (2021). “Low-dimensional Structure in the Space of Language Representations is Reflected in Brain Responses”. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34.

- Curran Associates, Inc., pp. 8332–8344. URL: <https://proceedings.neurips.cc/paper/2021/file/464074179972cbbd75a39abc6954cd12-Paper.pdf>.
- Bansal, Yamini, Preetum Nakkiran, and Boaz Barak (2021). “Revisiting Model Stitching to Compare Neural Representations”. In: *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*. Ed. by Marc’Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, pp. 225–236. URL: <https://proceedings.neurips.cc/paper/2021/hash/01ded4259d101feb739b06c399e9cd9c-Abstract.html>.
- Biondi, Niccolo, Federico Pernici, Matteo Bruni, and Alberto Del Bimbo (2021). “CoReS: Compatible Representations via Stationarity”. In: *ArXiv preprint*. DOI: 10.1109/tpami.2023.3259542. URL: <https://arxiv.org/abs/2111.07632>.
- Csiszárík, Adrián, Péter Kőrösi-Szabó, Ákos K. Matszangosz, Gergely Papp, and Dániel Varga (2021). *Similarity and Matching of Neural Network Representations*. URL: <https://arxiv.org/abs/2110.14633>.
- Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby (2021). “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. URL: <https://openreview.net/forum?id=YicbFdNTTy>.
- Gandikota, Kanchana Vaishnavi, Jonas Geiping, Zorah Löhner, Adam Czapliński, and Michael Moeller (2021). “Training or architecture? how to incorporate invariance in neural networks”. In: *arXiv preprint arXiv:2106.10044*.
- Gresele, Luigi, Julius Von Kügelgen, Vincent Stimper, Bernhard Schölkopf, and Michel Besserve (2021). “Independent mechanism analysis, a new concept?” In: *Advances in Neural Information Processing Systems*. Ed. by A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan. URL: <https://openreview.net/forum?id=Rnn8zoAkrwr>.
- GrokAI (2021). *nn-template bootstraps PyTorch projects by advocating reproducibility & best practices in deep learning*. Software available from <https://github.com/grok-ai/>. URL: <https://github.com/grok-ai/nn-template>.
- Gygli, Michael, Jasper Uijlings, and Vittorio Ferrari (2021). “Towards Reusable Network Components by Learning Compatible Representations”. en. In: *AAAI 35.9*, pp. 7620–7629. DOI: 10.1609/aaai.v35i9.16932.
- Kossen, Jannik, Neil Band, Clare Lyle, Aidan N. Gomez, Tom Rainforth, and Yarin Gal (2021). *Self-Attention Between Datapoints: Going Beyond Individual Input-Output Pairs in Deep Learning*. URL: <https://arxiv.org/abs/2106.02584>.
- Lhoest, Quentin, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf (2021). “Datasets: A Community Library for Natural Language Processing”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 175–184. DOI: 10.18653/v1/2021.emnlp-demo.21. URL: <https://aclanthology.org/2021.emnlp-demo.21>.

- Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. (2021). “Learning transferable visual models from natural language supervision”. In: *International conference on machine learning*. PMLR, pp. 8748–8763.
- Roeder, Geoffrey, Luke Metz, and Durk Kingma (July 1, 2021). “On Linear Identifiability of Learned Representations”. In: *Proceedings of the 38th International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, pp. 9030–9039. URL: <https://proceedings.mlr.press/v139/roeder21a.html> (visited on 02/25/2024).
- Schwenk, Holger, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán (2021). “WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 1351–1361. DOI: 10.18653/v1/2021.eacl-main.115. URL: <https://aclanthology.org/2021.eacl-main.115>.
- Trappolini, Giovanni, Luca Cosmo, **Luca Moschella**, Riccardo Marin, Simone Melzi, and Emanuele Rodolà (2021). “Shape Registration in the Time of Transformers”. In: *Advances in Neural Information Processing Systems*. Ed. by A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan. URL: <https://openreview.net/forum?id=ui4xChWcA4R>.
- Williams, Alex H., Erin Kunz, Simon Kornblith, and Scott W. Linderman (2021). “Generalized Shape Metrics on Neural Representations”. In: *Advances in Neural Information Processing Systems*. Ed. by Marc’Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan. Vol. 34, pp. 4738–4750. URL: <https://proceedings.neurips.cc/paper/2021/hash/252a3dbaeb32e7690242ad3b556e626b-Abstract.html>.
- Benton, Gregory, Marc Finzi, Pavel Izmailov, and Andrew G Wilson (2020). “Learning Invariances in Neural Networks from Training Data”. In: 33. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, pp. 17605–17616.
- Bianchi, Federico, Jacopo Tagliabue, Bingqing Yu, Luca Bigon, and Ciro Greco (2020). “Fantastic Embeddings and How to Align Them: Zero-Shot Inference in a Multi-Shop Scenario”. In: *ArXiv preprint abs/2007.14906*. URL: <https://arxiv.org/abs/2007.14906>.
- Biewald, Lukas (2020). *Experiment Tracking with Weights and Biases*. Software available from wandb.com. URL: <https://www.wandb.com/>.
- Clark, Kevin, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning (2020). “ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators”. In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. URL: <https://openreview.net/forum?id=r1xMH1BtvB>.
- Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov (2020). “Unsupervised Cross-lingual Representation Learning at Scale”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747. URL: <https://aclanthology.org/2020.acl-main.747>.
- Frankle, Jonathan, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin (July 18, 2020). “Linear Mode Connectivity and the Lottery Ticket Hypothesis”. In: *arXiv*. DOI: 10.48550/arXiv.1912.05671. arXiv: 1912.05671 [cs, stat]. URL: <http://arxiv.org/abs/1912.05671> (visited on 04/28/2023).

- Han, Dongyoon, Sangdoon Yun, Byeongho Heo, and YoungJoon Yoo (2020). "Rethinking Channel Dimensions for Efficient Model Design". In: *ArXiv preprint abs/2007.00992*. DOI: [10.1109/cvpr46437.2021.00079](https://doi.org/10.1109/cvpr46437.2021.00079). URL: <https://arxiv.org/abs/2007.00992>.
- Keung, Phillip, Yichao Lu, György Szarvas, and Noah A. Smith (2020). "The Multilingual Amazon Reviews Corpus". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 4563–4568. DOI: [10.18653/v1/2020.emnlp-main.369](https://doi.org/10.18653/v1/2020.emnlp-main.369). URL: <https://aclanthology.org/2020.emnlp-main.369>.
- Khemakhem, Ilyes, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen (June 3, 2020). "Variational Autoencoders and Nonlinear ICA: A Unifying Framework". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. International Conference on Artificial Intelligence and Statistics. PMLR, pp. 2207–2217. URL: <https://proceedings.mlr.press/v108/khemakhem20a.html> (visited on 02/25/2024).
- Lan, Zhenzhong, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut (2020). "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations". In: *International Conference on Learning Representations*. URL: <https://openreview.net/forum?id=H1eA7AEtvS>.
- Lyle, Clare, Mark van der Wilk, Marta Kwiatkowska, Yarin Gal, and Benjamin Bloem-Reddy (2020). "On the benefits of invariance in neural networks". In: *arXiv preprint arXiv:2005.00178*.
- Papayan, Vardan, X. Y. Han, and David L. Donoho (Oct. 6, 2020). "Prevalence of Neural Collapse during the Terminal Phase of Deep Learning Training". In: *Proceedings of the National Academy of Sciences* 117.40, pp. 24652–24663. DOI: [10.1073/pnas.2015509117](https://doi.org/10.1073/pnas.2015509117). URL: <https://www.pnas.org/doi/10.1073/pnas.2015509117> (visited on 02/27/2024).
- Singh, Sidak Pal and Martin Jaggi (2020). "Model Fusion via Optimal Transport". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., pp. 22045–22055. URL: <https://proceedings.neurips.cc/paper/2020/hash/fb2697869f56484404c8ceee2985b01d-Abstract.html> (visited on 06/29/2023).
- Sun, Yu, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt (Nov. 21, 2020). "Test-Time Training with Self-Supervision for Generalization under Distribution Shifts". In: *Proceedings of the 37th International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, pp. 9229–9248. URL: <https://proceedings.mlr.press/v119/sun20b.html> (visited on 10/06/2023).
- Tatro, Norman, Pin-Yu Chen, Payel Das, Igor Melnyk, Prasanna Sattigeri, and Rongjie Lai (2020). "Optimizing Mode Connectivity via Neuron Alignment". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., pp. 15300–15311. URL: <https://proceedings.neurips.cc/paper/2020/hash/aecad42329922dfc97eee948606e1f8e-Abstract.html> (visited on 04/28/2023).
- Tsitsulin, Anton, Marina Munkhoeva, Davide Mottin, Panagiotis Karras, Alexander M. Bronstein, Ivan V. Oseledets, and Emmanuel Müller (2020). "The Shape of Data: Intrinsic Distance for Data Distributions". In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. URL: <https://openreview.net/forum?id=HyebplHYwB>.
- Vulić, Ivan, Sebastian Ruder, and Anders Søgaard (Nov. 2020). "Are All Good Word Vector Spaces Isomorphic?" In: *Proceedings of the 2020 Conference on Empirical*

- Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 3178–3192. DOI: [10.18653/v1/2020.emnlp-main.257](https://doi.org/10.18653/v1/2020.emnlp-main.257). URL: <https://aclanthology.org/2020.emnlp-main.257>.
- Wang, Dequan, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell (Oct. 2, 2020). “Tent: Fully Test-Time Adaptation by Entropy Minimization”. In: International Conference on Learning Representations. URL: <https://openreview.net/forum?id=uX13bZLkr3c> (visited on 10/06/2023).
- Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush (2020). “Transformers: State-of-the-Art Natural Language Processing”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, pp. 38–45. DOI: [10.18653/v1/2020.emnlp-demos.6](https://doi.org/10.18653/v1/2020.emnlp-demos.6). URL: <https://aclanthology.org/2020.emnlp-demos.6>.
- Alvarez-Melis, David, Stefanie Jegelka, and Tommi S. Jaakkola (2019). “Towards Optimal Transport with Global Invariances”. In: *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*. Ed. by Kamalika Chaudhuri and Masashi Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, pp. 1870–1879. URL: <http://proceedings.mlr.press/v89/alvarez-melis19a.html>.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: [10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423). URL: <https://aclanthology.org/N19-1423>.
- Falcon, William and The PyTorch Lightning team (2019). *PyTorch Lightning*. Version 1.4. DOI: [10.5281/zenodo.3828935](https://doi.org/10.5281/zenodo.3828935). URL: <https://github.com/Lightning-AI/lightning>.
- Fey, Matthias and Jan E. Lenssen (2019). “Fast Graph Representation Learning with PyTorch Geometric”. In: *ICLR Workshop on Representation Learning on Graphs and Manifolds*.
- Hyvarinen, Aapo, Hiroaki Sasaki, and Richard Turner (Apr. 11, 2019). “Nonlinear ICA Using Auxiliary Variables and Generalized Contrastive Learning”. In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics. The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 859–868. URL: <https://proceedings.mlr.press/v89/hyvarinen19a.html> (visited on 02/25/2024).
- Kandi, Haribabu, Ayushi Jain, Swetha Velluva Chathoth, Deepak Mishra, and Gorthi RK Sai Subrahmanyam (2019). “Incorporating rotational invariance in convolutional neural network architecture”. In: *Pattern Analysis and Applications 22*, pp. 935–948.
- Kornblith, Simon, Mohammad Norouzi, Honglak Lee, and Geoffrey E. Hinton (2019). “Similarity of Neural Network Representations Revisited”. In: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 3519–3529. URL: <http://proceedings.mlr.press/v97/kornblith19a.html>.

- Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019). “RoBERTa: A Robustly Optimized BERT Pretraining Approach”. In: *ArXiv preprint abs/1907.11692*. URL: <https://arxiv.org/abs/1907.11692>.
- Nakkiran, Preetum, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever (Sept. 25, 2019). “Deep Double Descent: Where Bigger Models and More Data Hurt”. In: *International Conference on Learning Representations*. URL: <https://openreview.net/forum?id=B1g5sA4twr> (visited on 02/27/2024).
- You, Jiaxuan, Rex Ying, and Jure Leskovec (2019). “Position-aware Graph Neural Networks”. In: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. *Proceedings of Machine Learning Research*. PMLR, pp. 7134–7143. URL: <http://proceedings.mlr.press/v97/you19b.html>.
- Zhang, Yan, Jonathon Hare, and Adam Prugel-Bennett (2019). “Deep set prediction networks”. In: *Advances in Neural Information Processing Systems* 32.
- Lample, Guillaume, Alexis Conneau, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou (2018). “Word translation without parallel data”. In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. URL: <https://openreview.net/forum?id=H196sainb>.
- Morcos, Ari S., Maithra Raghu, and Samy Bengio (2018). “Insights on representational similarity in neural networks with canonical correlation”. In: *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*. Ed. by Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, pp. 5732–5741. URL: <https://proceedings.neurips.cc/paper/2018/hash/a7a3d70c6d17a73140918996d03c014f-Abstract.html>.
- Soudry, Daniel, Elad Hoffer, Mor Shpigel Nacson, and Nathan Srebro (2018). “The Implicit Bias of Gradient Descent on Separable Data”. In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. URL: <https://openreview.net/forum?id=r1q7n9gAb>.
- Wang, Liwei, Lunjia Hu, Jiayuan Gu, Zhiqiang Hu, Yue Wu, Kun He, and John E. Hopcroft (2018). “Towards Understanding Learning Representations: To What Extent Do Different Neural Networks Learn the Same Representation”. In: *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*. Ed. by Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, pp. 9607–9616. URL: <https://proceedings.neurips.cc/paper/2018/hash/5fc34ed307aac159a30d81181c99847e-Abstract.html>.
- Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov (2017). “Enriching Word Vectors with Subword Information”. In: *Transactions of the Association for Computational Linguistics* 5, pp. 135–146. DOI: [10.1162/tacl_a_00051](https://doi.org/10.1162/tacl_a_00051). URL: <https://aclanthology.org/Q17-1010>.
- Movshovitz-Attias, Yair, Alexander Toshev, Thomas K. Leung, Sergey Ioffe, and Saurabh Singh (2017). “No Fuss Distance Metric Learning Using Proxies”. In: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, pp. 360–368. DOI: [10.1109/ICCV.2017.47](https://doi.org/10.1109/ICCV.2017.47). URL: <https://doi.org/10.1109/ICCV.2017.47>.

- Raghu, Maithra, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein (2017). "Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability". In: *Advances in neural information processing systems* 30.
- Smith, Samuel L., David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla (2017). "Offline bilingual word vectors, orthogonal transformations and the inverted softmax". In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. URL: <https://openreview.net/forum?id=r1Aab85gg>.
- Snell, Jake, Kevin Swersky, and Richard S. Zemel (2017). "Prototypical Networks for Few-shot Learning". In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Ed. by Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, pp. 4077–4087. URL: <https://proceedings.neurips.cc/paper/2017/hash/cb8da6767461f2812ae4290eac7cbc42-Abstract.html>.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Ed. by Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, pp. 5998–6008. URL: <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>.
- Xiao, Han, Kashif Rasul, and Roland Vollgraf (2017). *Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms*. URL: <https://arxiv.org/abs/1708.07747>.
- Fefferman, Charles, Sanjoy Mitter, and Hariharan Narayanan (2016). "Testing the manifold hypothesis". In: *Journal of the American Mathematical Society* 29.4, pp. 983–1049.
- Hyvarinen, Aapo and Hiroshi Morioka (2016). "Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear ICA". In: *Advances in Neural Information Processing Systems*. Vol. 29. Curran Associates, Inc. URL: https://papers.nips.cc/paper_files/paper/2016/hash/d305281faf947ca7acade9ad5c8c818c-Abstract.html (visited on 02/25/2024).
- Li, Yixuan, Jason Yosinski, Jeff Clune, Hod Lipson, and John E. Hopcroft (2016). "Convergent Learning: Do different neural networks learn the same representations?" In: *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. URL: <http://arxiv.org/abs/1511.07543>.
- Ulyanov, Dmitry, Andrea Vedaldi, and Victor Lempitsky (2016). "Instance Normalization: The Missing Ingredient for Fast Stylization". In: *ArXiv preprint abs/1607.08022*. URL: <https://arxiv.org/abs/1607.08022>.
- Yang, Zhilin, William W. Cohen, and Ruslan Salakhutdinov (2016). "Revisiting Semi-Supervised Learning with Graph Embeddings". In: *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*. Ed. by Maria-Florina Balcan and Kilian Q. Weinberger. Vol. 48. JMLR Workshop and Conference Proceedings. JMLR.org, pp. 40–48. URL: <http://proceedings.mlr.press/v48/yanga16.html>.
- Lenc, Karel and Andrea Vedaldi (2015). "Understanding image representations by measuring their equivariance and equivalence". In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE

- Computer Society, pp. 991–999. DOI: [10.1109/CVPR.2015.7298701](https://doi.org/10.1109/CVPR.2015.7298701). URL: <https://doi.org/10.1109/CVPR.2015.7298701>.
- Olah, Christopher (2015). *Visualizing representations: Deep learning and human beings*. en. <http://colah.github.io/posts/2015-01-Visualizing-Representations/>. Accessed: 2022-8-2.
- Xing, Chao, Dong Wang, Chao Liu, and Yiye Lin (June 2015). “Normalized Word Embedding and Orthogonal Transform for Bilingual Word Translation”. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado: Association for Computational Linguistics, pp. 1006–1011. DOI: [10.3115/v1/N15-1104](https://doi.org/10.3115/v1/N15-1104). URL: <https://aclanthology.org/N15-1104>.
- Zhang, Xiang, Junbo Jake Zhao, and Yann LeCun (2015). “Character-level Convolutional Networks for Text Classification”. In: *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*. Ed. by Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, pp. 649–657. URL: <https://proceedings.neurips.cc/paper/2015/hash/250cf8b51c773f3f8dc8b4be867a9a02-Abstract.html>.
- Bengio, Yoshua, Aaron Courville, and Pascal Vincent (2014). *Representation Learning: A Review and New Perspectives*. arXiv: [1206.5538](https://arxiv.org/abs/1206.5538) [cs.LG].
- Nili, Hamed, Cai Wingfield, Alexander Walther, Li Su, William D. Marslen-Wilson, and Nikolaus Kriegeskorte (2014). “A Toolbox for Representational Similarity Analysis”. In: *PLoS Computational Biology* 10. URL: <https://api.semanticscholar.org/CorpusID:16823738>.
- Cuturi, Marco (2013). “Sinkhorn Distances: Lightspeed Computation of Optimal Transport”. In: *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*. Ed. by Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, pp. 2292–2300. URL: <https://proceedings.neurips.cc/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html>.
- Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). “Efficient Estimation of Word Representations in Vector Space”. In: *ArXiv preprint abs/1301.3781*. URL: <https://arxiv.org/abs/1301.3781>.
- Mikolov, Tomás, Quoc V. Le, and Ilya Sutskever (2013). “Exploiting Similarities among Languages for Machine Translation”. In: *CoRR abs/1309.4168*. arXiv: [1309.4168](https://arxiv.org/abs/1309.4168) [cs.CL]. URL: <http://arxiv.org/abs/1309.4168>.
- Deng, Li (2012). “The mnist database of handwritten digit images for machine learning research”. In: *IEEE Signal Processing Magazine* 29.6, pp. 141–142.
- Raizada, Rajeev DS and Andrew C Connolly (2012). “What makes different people’s representations alike: neural similarity space solves the problem of across-subject fMRI decoding”. In: *Journal of cognitive neuroscience* 24.4, pp. 868–877. DOI: [10.1162/jocn_a_00189](https://doi.org/10.1162/jocn_a_00189).
- Maas, Andrew L., Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts (June 2011). “Learning Word Vectors for Sentiment Analysis”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, pp. 142–150. URL: <http://www.aclweb.org/anthology/P11-1015>.
- Mémoli, Facundo (Aug. 1, 2011). “Gromov–Wasserstein Distances and the Metric Approach to Object Matching”. In: *Foundations of Computational Mathematics* 11.4,

- pp. 417–487. ISSN: 1615-3383. DOI: [10.1007/s10208-011-9093-5](https://doi.org/10.1007/s10208-011-9093-5). URL: <https://doi.org/10.1007/s10208-011-9093-5> (visited on 02/27/2024).
- Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12, pp. 2825–2830.
- Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li (2009). “ImageNet: A large-scale hierarchical image database”. In: *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. IEEE Computer Society, pp. 248–255. DOI: [10.1109/CVPR.2009.5206848](https://doi.org/10.1109/CVPR.2009.5206848). URL: <https://doi.org/10.1109/CVPR.2009.5206848>.
- Krizhevsky, Alex (2009). “Learning Multiple Layers of Features from Tiny Images”. In: pp. 32–33. URL: <https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf>.
- Wang, Chang and Sridhar Mahadevan (2009). “Manifold Alignment without Correspondence”. In: *IJCAI 2009, Proceedings of the 21st International Joint Conference on Artificial Intelligence, Pasadena, California, USA, July 11-17, 2009*. Ed. by Craig Boutilier. Vol. 2, pp. 1273–1278. URL: <http://ijcai.org/Proceedings/09/Papers/214.pdf>.
- Hofmann, Thomas, Bernhard Schölkopf, and Alexander J Smola (2008). “Kernel methods in machine learning”. In: *The annals of statistics* 36.3, pp. 1171–1220. DOI: [10.1214/009053607000000677](https://doi.org/10.1214/009053607000000677).
- Kriegeskorte, Nikolaus, Marieke Mur, and Peter Bandettini (2008). “Representational similarity analysis - connecting the branches of systems neuroscience”. In: *Frontiers in Systems Neuroscience* 2. ISSN: 1662-5137. DOI: [10.3389/neuro.06.004.2008](https://doi.org/10.3389/neuro.06.004.2008).
- Kriegeskorte, Nikolaus, Marieke Mur, and Peter A. Bandettini (2008). “Representational Similarity Analysis – Connecting the Branches of Systems Neuroscience”. In: *Frontiers in Systems Neuroscience* 2. URL: <https://api.semanticscholar.org/CorpusID:10873381>.
- Sen, Prithviraj, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad (2008). “Collective Classification in Network Data”. In: *AIMag*. 29.3, p. 93. ISSN: 2371-9621. DOI: [10.1609/aimag.v29i3.2157](https://doi.org/10.1609/aimag.v29i3.2157).
- Wang, Chang and Sridhar Mahadevan (2008). “Manifold alignment using Procrustes analysis”. In: *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*. Ed. by William W. Cohen, Andrew McCallum, and Sam T. Roweis. Vol. 307. ACM International Conference Proceeding Series. ACM, pp. 1120–1127. DOI: [10.1145/1390156.1390297](https://doi.org/10.1145/1390156.1390297). URL: <https://doi.org/10.1145/1390156.1390297>.
- Auer, Sören, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives (2007). “DBpedia: A Nucleus for a Web of Open Data”. In: *The Semantic Web*. Springer Berlin Heidelberg, pp. 722–735. DOI: [10.1007/978-3-540-76298-0_52](https://doi.org/10.1007/978-3-540-76298-0_52).
- Li, Xin and Dan Roth (2002). “Learning Question Classifiers”. In: *COLING 2002: The 19th International Conference on Computational Linguistics*. DOI: [10.3115/1072228.1072378](https://doi.org/10.3115/1072228.1072378). URL: <https://www.aclweb.org/anthology/C02-1150>.
- Hovy, Eduard, Laurie Gerber, Ulf Hermjakob, Chin-Yew Lin, and Deepak Ravichandran (2001). “Toward Semantics-Based Answer Pinpointing”. In: *Proceedings of the First International Conference on Human Language Technology Research*. DOI: [10.3115/1072133.1072221](https://doi.org/10.3115/1072133.1072221). URL: <https://aclanthology.org/H01-1069>.

- Laakso, Aarre and G. Cottrell (2000). "Content and cluster analysis: Assessing representational similarity in neural systems". In: *Philosophical Psychology* 13, pp. 47–76. DOI: [10.1080/09515080050002726](https://doi.org/10.1080/09515080050002726).
- Giles, C. Lee, Kurt D. Bollacker, and Steve Lawrence (1998). "CiteSeer: an automatic citation indexing system". In: *Digital library*. DOI: [10.1145/276675.276685](https://doi.org/10.1145/276675.276685). URL: <https://api.semanticscholar.org/CorpusID:514080>.
- Hotelling, Harold (1992). "Relations between two sets of variates". In: *Breakthroughs in statistics: methodology and distribution*, pp. 162–190.
- Gower, J. C. (Mar. 1975). "Generalized Procrustes Analysis". In: *Psychometrika* 40.1, pp. 33–51. ISSN: 1860-0980. DOI: [10.1007/BF02291478](https://doi.org/10.1007/BF02291478). (Visited on 05/17/2023).