EViLBERT: Learning Task-Agnostic Multimodal Sense Embeddings

Agostina Calabrese , **Michele Bevilacqua** and **Roberto Navigli**
Sapienza NLP Group, Department of Computer Science, Sapienza University of Rome
{calabrese.a, bevilacqua, navigli}@di.uniroma1.it

Abstract

The problem of grounding language in vision is increasingly attracting scholarly efforts. As of now, however, most of the approaches have been limited to word embeddings, which are not capable of handling polysemous words. This is mainly due to the limited coverage of the available semantically-annotated datasets, hence forcing research to rely on alternative technologies (i.e., image search engines). To address this issue, we introduce EViLBERT, an approach which is able to perform image classification over an open set of concepts, both concrete and non-concrete. Our approach is based on the recently introduced Vision-Language Pretraining (VLP) model, and builds upon a manually-annotated dataset of concept-image pairs. We use our technique to clean up the image-to-concept mapping that is provided within a multilingual knowledge base, resulting in over $258{,}000$ images associated with $42{,}500$ concepts. We show that our VLP-based model can be used to create multimodal sense embeddings starting from our automatically-created dataset. In turn, we also show that these multimodal embeddings improve the performance of a Word Sense Disambiguation architecture over a strong unimodal baseline. We release code, dataset and embeddings at http://babelpic.org.

1 Introduction

In recent years, there has been a surge of interest in learning representations of natural language that are grounded in visual perception, i.e., that are mapped to the external reality through image data. This research direction represents a promising step towards realising human-like language learning, and has its main focus in the integration of visual and language knowledge, grounding words and sentences into their visual representation [Kiros *et al.*, 2018]. This growing interest is also due to evidence that many concepts[1] are grounded in perception [Barsalou and Wiemer-Hastings, 2005].

The usefulness of perceptual information has already been demonstrated in several works (e.g., [Silberer and Lapata, 2014; Kiela *et al.*, 2014]), showing that multimodal systems outperform unimodal language-based models in many semantic Natural Language Processing (NLP) tasks, such as concept categorization and modelling similarity [Hill and Korhonen, 2014]. Multimodal embeddings can also be exploited to improve the performances of unimodal vision-based models. For instance, asking object detection systems to predict embeddings instead of classes enables zero-shot classification of unseen objects [Frome *et al.*, 2013]. Moreover, many artificial intelligence applications involve more than one modality, highlighting the importance of learning from diverse information sources.

Despite the growing interest, multimodal representation learning still remains a challenging problem. Most works on the topic focus on word-level embeddings, mainly due to the lack of a wide-coverage dataset illustrating concepts with images. For instance, abstract concepts are rarely included in computer vision datasets. As a result, the most comprehensive source of grounding information is search engines, which can only provide a mapping from surface representations (i.e., terms), not concepts. By using search engines, then, it is not clear how to handle polysemous items such as HACK, which can be used in either a *computer tampering* or *cutting* sense.

In this paper, we introduce Embedding Vision and Language with BERT[2] (EViLBERT), a BERT-based approach for learning task-agnostic multimodal sense embeddings. Our work is built on top of the recently introduced Vision-Language Pre-training (VLP) model [Zhou *et al.*, 2020], a system which achieves state-of-the-art results on language-vision tasks such as image captioning and Visual Question Answering (VQA). To produce the embeddings, as a wide-coverage mapping from images to non-concrete concepts is not readily available, we build upon BabelPic [Calabrese *et al.*, 2020] using an automatic technique for the verification of synset-image associations. BabelPic is a recently released dataset which includes manually annotated concept-image pairs with a focus on non-concrete concepts, and it is also linked to popular Lexical Knowledge Bases (LKB) like WordNet [Miller, 1995] and BabelNet [Navigli and Ponzetto,

[1] Unless further specification is provided, we use the terms *concept*, *sense* and *synset* (i.e., the representation of concepts in the WordNet [Miller, 1995] computational lexicon) interchangeably.

[2] Pre-training of Deep Bidirectional Transformers [Devlin *et al.*, 2019]

2012]. As a result, our automatic expansion produces a large mapping of around $258,000$ images to over $42,500$ concepts.

We utilize the VLP model to create multimodal sense embeddings from this extended resource. Specifically, we compute sense embeddings as centroids of VLP hidden states, feeding the model with image-definition pairs from the automatically expanded resource.

To demonstrate the usefulness of our multimodal sense embeddings, we show how better representations improve the subsequent learning process in the Word Sense Disambiguation (WSD) task [Navigli, 2009]. Our experiments in WSD show that multimodal sense embeddings, computed through our task-agnostic approach, achieve better performances than task-specific unimodal (i.e., textual) embeddings.

To summarise, our contributions are threefold:

- We create a large dataset with a multimodal system for the verification of candidate image-concept pairs. We show that our methodology is reliable in both concrete and abstract domains.
- We introduce EViLBERT, a VLP-based approach for learning task-agnostic multimodal sense embeddings.
- We demonstrate the effectiveness of EViLBERT sense embeddings by employing them within a WSD architecture, improving over a competitive baseline.

2 Related Work

To the best of our knowledge, no multimodal sense embeddings exist which cover a wide range of nominal and verbal concepts. There are, however, several works that tie images to words (which are often polysemous) instead of concepts. For example, in the Deep Visual-Semantic Embedding model [Frome *et al.*, 2013, DeViSE] word embeddings and image embeddings are mapped to a common space by maximising the similarity between them, enabling zero-shot object recognition. However, because of the use of word embeddings, the representations on the text side conflate multiple meanings of polysemous items. Hill *et al.* [2014] proposed a generalisation of the Skipgram model [Mikolov *et al.*, 2013] in order to learn multimodal representations of both abstract and concrete concepts. Nonetheless, concept representations are learned through cooccurrence of lexicalizations, i.e., of words, and are not linked to any LKB. In another approach, Picturebook [Kiros *et al.*, 2018], concepts are absent altogether. Image embeddings are created by extracting features from the top-k images that are returned by Google image search for each query word. Picturebook is restricted to the vocabulary of the word embeddings.

In addition to the problem of using words instead of concepts, it is often the case that vectors are not really learned in a multimodal way, but are the result of a mapping or shallow combination of information from single modalities. This is the case for Picturebook embeddings, in which the feature extractor exclusively takes images as input. Word embeddings are only integrated with a task-dependent gating mechanism, which enables the downstream model to select whether the image or language modality is more important. In the work of Hill *et al.* [2014] only information from object cooccurrences in visual data, but no actual image, is used.

Previous works have often not tackled joint multimodality and, as pointed out by Collell and Moens [2018], this represents a major weakness because they have to resort to mapping independently learned unimodal vectors to either another modality, or a shared embedding space. This results in the fact that, contrary to what is desirable, the neighborhood structure of the mapped vectors does not resemble that of the target vectors.

One promising way to produce deeply multimodal representations is that of extending language model pre-training, a technique which has proven to be immensely successful in NLP tasks. In fact, many multimodal extensions of BERT, currently the most popular language modeling architecture, have been released recently (e.g., VisualBERT [Li *et al.*, 2019], ViLBERT [Lu *et al.*, 2019], LXMERT [Tan and Bansal, 2019], VLP [Zhou *et al.*, 2020]) which achieve state-of-the-art results in many language-vision tasks.

In our approach, we address both the issues highlighted above, i.e., that of 1) grounding in visual perception concepts as encoded in an LKB, instead of words, and that of 2) embedding language and vision jointly. To reach our goals, we exploit the transfer learning capabilities of one of the aforementioned multimodal systems, i.e., VLP, by reducing the task of multimodal concept embedding to VQA. In what follows, we elaborate on the details of our contribution.

3 Our Approach

The core idea of this work is to create sense-level embeddings starting from both textual (i.e., the gloss of the target concept) and image data (i.e., images illustrating the target concept). To reach this goal, we need a dataset including wide-coverage concept-image associations which is also linked to an LKB (i.e., WordNet). Unfortunately, existing wide-coverage repositories like ImageNet [Deng *et al.*, 2009], COCO [Lin *et al.*, 2014], Flickr30kEntities [Plummer *et al.*, 2015] and Open Images [Kuznetsova *et al.*, 2020] are either limited to concepts denoting concrete, tangible things, or not linked to an LKB. To address this issue we start from BabelPic [Calabrese *et al.*, 2020], a hand-labeled dataset of concept-image associations, and use an automatic technique to extend its coverage to previously unseen concepts. Once the dataset has been created, we learn our multimodal representations by exploiting VLP, a recent BERT-based system for language-vision pre-training. Specifically, multimodal sense embeddings are obtained by aggregating the vectors resulting from the definition and the top-k images for each concept. All these steps are detailed below.

3.1 Gold Dataset

As more fully discussed in Hill *et al.* [2014], differences between processing of abstract and concrete concepts suggest that models trained for concrete concept learning may not necessarily work in the general case. However, since knowledge about tangible objects is already contained within pretrained object detection systems, we assume that the foregoing objection does not hold when, vice versa, a system that is trained for abstract concept learning is used on concrete concepts. Driven by this hypothesis, we employ as seed BabelPic [Calabrese *et al.*, 2020], which has an explicit focus

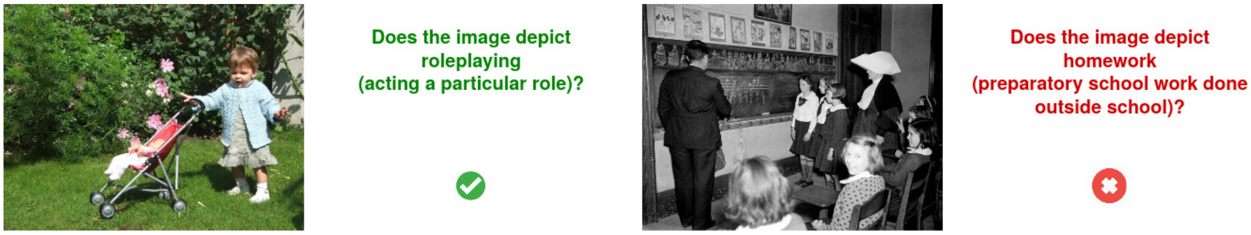

Figure 1: We enable image classification over an open set of concepts by reducing the task to VQA. Examples are taken from our gold dataset.

on non-concrete nominal and verbal concepts. BabelPic was built by selecting a subset of the abstract concepts contained in WordNet (i.e., any nominal synset descending from *feeling.n.01* or *event.n.01* and any verbal synset belonging to the *verb.competition*, *verb.motion* and *verb.social* lexicographer files). Note that, in order to guarantee the non-concreteness of such concepts, synsets descending from *physical_entity.n.01*, *shape.n.02* or *color.n.01* were discarded.

For each selected concept, 15 candidate images were gathered from Wikipedia. This was possible thanks to the automatic linking between WordNet and Wikipedia available through BabelNet [Navigli and Ponzetto, 2012], a large multilingual Lexical Knowledge Base. The quality of the dataset was guaranteed by filtering out any image where transparency was used, or where half of the pixels were white, as these were not likely to be relevant. Moreover, remaining noisy images were discarded during the manual validation phase, when each synset-image pair was hand-checked through an *ad hoc* graphical interface. As a result, the gold dataset includes 2,733 synsets and 14,931 images.

3.2 Silver Dataset

The creation of the embeddings requires the availability of a wide-coverage multimodal dataset. However, the time demands of the manual validation process put a serious limit on the feasibility of the task. To address this issue and create a larger dataset, we develop an approach for the automatic verification of synset-image associations by defining the problem as a VQA task with yes/no answers. In particular, we specify a question template as follows:

“Does the image depict l (g)?”

where l is the lemma of the first sense in WordNet of the target synset and g is the synset gloss, i.e., its textual definition. We instantiate our template for each synset-image pair of interest, thus obtaining a textual question for each instance. Figure 1 shows two examples of the template, one with a positive and one with a negative concept-image pair.

The system that we use to address the above VQA problem is the fine-tuned VLP model. Despite the fact that LXMERT achieves a slightly higher score on yes/no questions on the VQA 2.0 dataset [Goyal *et al.*, 2017], our preference goes to the VLP system since it is pre-trained on Conceptual Captions (CC), a wider and more general dataset including more than 3M image-caption pairs. More specifically, VLP is pre-trained using two unsupervised vision-language tasks: bidirectional and sequence-to-sequence masked language prediction. Input images are preprocessed using Faster R-CNN [Ren *et al.*, 2015] pre-trained on Visual Genome [Krishna *et al.*, 2017; Anderson *et al.*, 2018], hence extracting 100 object regions per image. In order to obtain class-aware region embeddings, region-level features are combined with the corresponding probability of each object label and with other region geometric information. The BERT-based architecture is thus fed the concatenation of gloss subword embeddings and class-aware region embeddings. During the fine-tuning phase, the hidden state of the encoder is given as input to a Multi-Layer Perceptron (MLP) in order to predict the corresponding answer. Here, we use VLP fine-tuned on VQA 2.0 and our gold dataset.

As we discuss in Section 4.1, the experiments demonstrate that our approach is resilient in the zero-shot classification of both abstract and concrete concepts, thus enabling verification across the whole sense inventory. We select all the WordNet synsets having at least one image in BabelNet, and, similarly to the gold dataset creation process, collect the first 15 corresponding images for each of them. Noisy images are discarded with the same heuristics used for the gold dataset. The 455,070 image-synset pairs obtained, corresponding to 44,868 different concepts, are then automatically validated using our VLP-based approach. As a result, our silver dataset includes 42,579 synsets and 257,499 images.

3.3 Multimodal Sense Embeddings

Having addressed the lack of a wide-coverage dataset, we aim to learn a dense representation for each concept by exploiting both textual (i.e., the gloss) and image data. To this end, we introduce EViLBERT, a new approach based on the general VLP architecture (i.e., without the MLP on top). Our approach involves the creation of static multimodal sense embeddings through the use of the context-dependent activations of the hidden layers of a pre-trained architecture. For the textual input, we feed the model again with a sentence including both the main lemma and the gloss of the target synset. As regards visual data, instead, we select the top-k images associated with the target concept (i.e., the first 5 images in the BabelNet ordering that are also accepted by our system) and we input them one at a time. By doing so we obtain at most 5 input pairs for each synset.

Given a concept-image pair, we generate a candidate embedding starting from the hidden states of the BERT-based VLP encoder. More specifically, we compute a weighted average of the hidden states at the various timesteps of the input sequence corresponding to the gloss. Note that this time step choice does not imply discarding all the visual information, since the corresponding hidden states have been calculated starting from the whole multimodal input sequence.

We compute an attention vector $a \in \mathbb{R}^T$ by summing the attention scores of all the heads of a Transformer layer for all the queries, and then averaging over the last N layers. After softmax-normalizing a, we obtain a vector associating each time step of the input sequence with an attention score. Specifically, a candidate embedding e is computed as:

$$e = \sum_{t=T_s}^{T_e} \left(\sum_{n=0}^{N} H_{n,t} \right) \cdot a_t$$

where T_s and T_e are, respectively, the first and the last time steps in the sequence corresponding to textual input and H is the three-dimensional tensor storing the hidden states of the last N (i.e., 4) layers of the BERT encoder. The multimodal embedding for a synset s is finally obtained as the centroid of the candidate embeddings $e^{(i)}$, $i \in [0, k)$, computed from the k instances associated with s.

We note that EViLBERT is similar to LMMS [Loureiro and Jorge, 2019], in which sense embeddings are built by feeding sense-tagged sequences and glosses to pre-trained BERT Large, but, crucially, differs in that we make joint use of textual and visual information, and do not need a sense-annotated corpus.

Coverage Extension

Given the limited number of concepts in the expanded silver dataset (45,312 out of a total of 117,659 WordNet synsets), we use the WordNet relational structure to assign to concepts with no image in our dataset, whenever possible, the embedding of the closest ancestor in our dataset. Thanks to this heuristic, we can compute embeddings for 80,414 concepts.

4 Experiments

Our experiments are organised in two main blocks. The first focuses on the evaluation of our proposed approach for the automatic verification of concept-image associations in both the concrete and non-concrete domains (Section 4.1). The second set of experiments, instead, assesses the effectiveness of our multimodal concept embeddings by evaluating them in the Word Sense Disambiguation task (Section 4.2).

4.1 Verification of Concept-Image Associations

As we recall, we start from BabelPic and use VLP to create a wide-coverage silver dataset. In order to do so, we refine the weights of the model fine-tuned on VQA 2.0. Training is performed by feeding VLP either a true example, i.e., a concept-image pair from the seed dataset, or a negative sample, which is created as described below.

Negative Samples

Our dataset can be formally characterised as the set $\mathcal{D}$, with each member being a pair $\langle c, i \rangle$, where c is a concept in WordNet and i is some image. For each concept c, we can produce negative samples by randomly selecting some other i' s.t. $\langle c', i' \rangle \in \mathcal{D}$ and $c \neq c'$. However, by defining the task as a validation of synset-image pairs images are allowed to be associated with multiple concepts and, thus, we need to ensure that i' does not depict c as well. A strategy to generate challenging negative instances is to look for **sibling** concepts, i.e., synsets that are connected to some concept c'' by the hypernymy relation (e.g., MARATHON and FUN RUN). Harder instances (namely, **polysemous**) can be obtained by selecting concepts containing at least one common lexicalization (e.g., the synsets of *swim.v.01* and *swim.v.02*). Finally, **unrelated** instances can be created by looking for synsets which are not connected to each other (e.g., GLADFULNESS and RACING).

Split	S(%)	P(%)	U(%)
Training	10.20	1.95	37.85
Validation	10.18	1.98	37.84
Test	10.21	1.94	37.83
Zero-Shot	11.55	2.19	36.25

Table 1: Distribution of instances labelled as *sibling* (S), *polysemous* (P) and *unrelated* (U) in our dataset's splits.

Split	N	C	I
Training	23,891	2,618	13,311
Validation	2,986	1,442	2,740
Test	2,987	1,416	2,715
Zero-Shot	502	43	490

Table 2: Number of instances (N), concepts (C) and images (I) in our dataset's splits.

For each concept c, we define as many negative samples as the number of available hand-annotated associations for c. Consequently, we obtain a dataset which is perfectly balanced between the two output labels. We perform the splitting of the dataset according to the 80%/10%/10% rule, hence defining training, validation and test sets. The relations used to define the negative examples are proportionally distributed between the splits (see Table 1), and the same holds for the output classes. Moreover, we force both the validation and test sets to also contain instances involving concepts that are not present in the training set. This is done in order to evaluate the system's capability to handle new concepts, and we refer to the subset of the test set given by these instances as the zero-shot test. More statistics can be found in Table 2.

Hyperparameters

When training the VLP architecture on our gold dataset, we keep the same setting as in the original paper. That is, we set the number of both hidden layers and attention heads of the BERT encoder to 12. We train the model for 20 epochs with learning rate of $2 \cdot 10^{-5}$ and a dropout rate of 0.1, selecting the weights of the best epoch, i.e. the one achieving the highest F1 score on the validation set.

Model Selection

In the following we use the abbreviations P-VLP and F-VLP to refer, respectively, to the VLP model, first, pre-trained on CC only and, second, further fine-tuned for the VQA task on the VQA 2.0 dataset. Our experiments demonstrate that both systems are reliable on our task, achieving precision and

Model	Validation		Test		Zero-Shot	
	P	F1	P	F1	P	F1
P-VLP	71.93	78.97	72.48	79.33	71.43	77.90
F-VLP	76.14	77.50	75.94	75.99	77.67	71.67

Table 3: Precision and F1 scores (%) on the verification of concept-image associations in our dataset.

F1 scores that are over 70% on all the splits (see Table 3). However, in a common use case scenario it is more important to classify as correct *only* valid concept-image pairs rather than *all* valid concept-image pairs. In other words, we prefer precision over recall. Consequently, the F-VLP model proves to be the most suitable for the task.

Can We Learn What New Abstract Concepts Look Like?
With EViLBERT we aim to develop a system capable of classifying images over an open set of concepts. Therefore, it is of great interest to assess the performance of the model in a zero-shot scenario (i.e., where the target concept is new to the system). Results related to this experiment are reported in the last column of Table 3, and demonstrate that both the P-VLP and F-VLP models are robust to zero-shot classification. In fact, the scores achieved are comparable to the performances obtained on the other splits. Specifically, the F-VLP system is able to verify the associations between new concepts and images with 77.67% precision, hence enabling the automatic extension of our dataset to any other non-concrete concept.

Can We Learn Concreteness from Abstraction?
The building process of our gold dataset was based on the hypothesis that a system which is able to classify abstract concepts is also able to deal with concrete ones. This idea is supported by the fact that pre-trained object detection systems have already acquired knowledge about tangible things. In this paragraph we explore the reliability of our model on the annotation of concrete concepts. To this end, we create a new zero-shot test split starting from 500 random image-concept pairs in ImageNet. Given the nature of ImageNet, all the selected concepts are concrete. We define the negative instances (i.e., the irrelevant concept-image pairs) by associating concepts with random images from the rest of our ImageNet sample. The result is a test set which is perfectly balanced between the two output classes. Table 4 shows the performances achieved by the F-VLP model on this zero-shot test. Our system obtains a precision score of 76.03%, demonstrating that it is able to classify images over concrete concepts without any drop in performance. Interestingly, many false positives are due to concept-image pairs which, despite being associated at random, are still plausible and, hence, hard to classify (see, for instance, Figure 2). The F1 score is even higher than the one registered on BabelPic's zero-shot test, hence validating our initial hypothesis.

4.2 Word Sense Disambiguation

To test whether the use of both visual and language modalities in EViLBERT results in better embeddings than the unimodal counterpart, we experiment on the use of our vectors in a competitive WSD architecture.

Model	Precision	F1
F-VLP	76.03	76.86

Table 4: Precision and F1 scores (%) of the zero-shot verification of concrete concept-image associations on the ImageNet sample.

Comparison Systems

We include as comparison systems the variants of the multimodal sense embeddings obtained through the use of EViLBERT at different stages of the fine-tuning procedure, namely EViLBERT-CC, EViLBERT-VQA and EViLBERT-FT. All these models have been pre-trained on CC, with the latter two being further fine-tuned on the VQA 2.0 dataset. In contrast to EViLBERT-VQA, however, EViLBERT-FT has also been trained on BabelPic gold. Results are reported both with and without the coverage extension. Additionally, we compare the performance of our systems with LMMS vectors [Loureiro and Jorge, 2019], which are produced with an approach that is similar to ours, but only uses textual information. In order to set a level playing field, we normalize the embeddings (both ours and LMMS's) and reduce the dimensions to $H = 512$ with a standard truncated SVD. The LMMS sense-level vectors are aggregated into synset embeddings by mean pooling. As regards LMMS, we also report results after reducing the coverage to the same level as EViLBERT's by discarding embeddings of concepts that are not in the latter. For completeness, we include the results of the current state of the art in WSD among models trained on SemCor, i.e., GlossBERT [Huang *et al.*, 2019].

Architecture

Our WSD system encodes words to be classified using a frozen BERT Large cased model. More specifically, for each subword that makes up the target, we compute a contextual vector by taking the sum of the last four corresponding hidden states. The vector for the target is just the centroid of the vectors of its subwords. The target contextual vector produced is given as input to a simple feed-forward classifier, which emits a probability distribution over all the concepts in the inventory, i.e., WordNet. When performing inference, for each target item, we take as prediction the synset with the highest probability only among those possible for the given target. Concept embeddings are used to initialize the corresponding rows in the output embeddings matrix $O \in \mathbb{R}^{|V| \times |H|}$ [Bevilacqua and Navigli, 2020]. In the baseline model, O is all randomly initialized. During training we update all the weights in O as usual. We train the system on the SemCor corpus for a maximum of 10 epochs, with the Adam optimizer and a learning rate of 10^{-4}, feeding the input in batches of 250 instances. We use SemEval-2015 [Moro and Navigli, 2015] as development set.

Results

Table 5 reports the results of the WSD experiments. We evaluate on the concatenation of all the standard evaluation datasets (ALL*) available in the framework of Raganato *et*

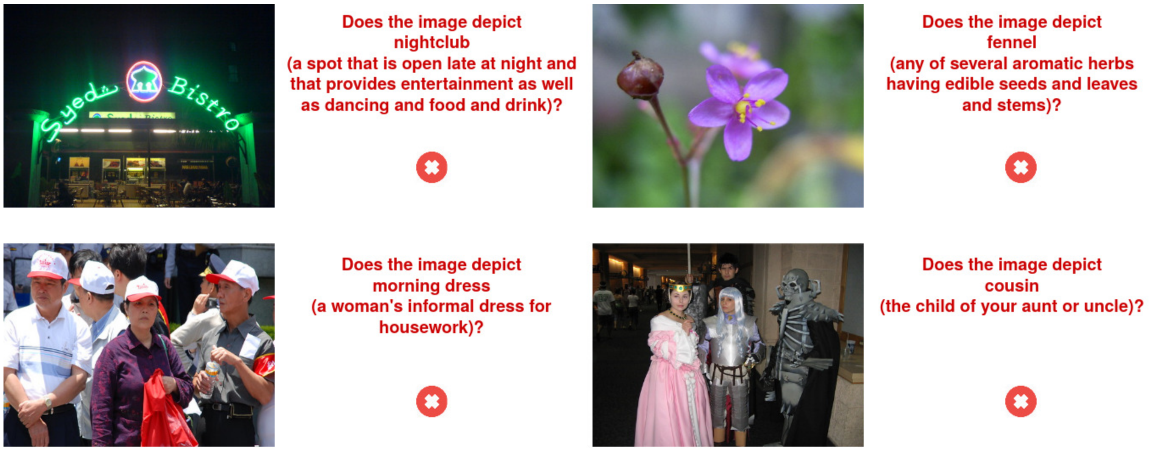
Does the image depict
nightclub
(a spot that is open late at night and that provides entertainment as well as dancing and food and drink)?
Does the image depict
fennel
(any of several aromatic herbs having edible seeds and leaves and stems)?
Does the image depict
morning dress
(a woman's informal dress for housework)?
Does the image depict
cousin
(the child of your aunt or uncle)?

Figure 2: Unrelated concept-image pairs that have been wrongly classified by our system. Examples are taken from our ImageNet's sample.

	Embedding	ALL*	ALL	Cov.-WN	Cov.-ALL*
	Baseline	73.2	73.6	0.0	0.0
Limited	LMMS 1024	73.6	73.7	38.5	28.9
	LMMS 2048	73.6	73.7	38.5	28.9
	EViLBERT (CC)	73.4	73.8	38.5	28.9
	EViLBERT (VQA)	**73.7**	**73.9**	38.5	28.9
	EViLBERT (FT)	73.1†	73.4	38.5	28.9
Extended	LMMS 1024	74.1†	74.3	68.3	53.6
	LMMS 2048	74.0	74.2	68.3	53.6
	EViLBERT (CC)	74.0	74.2	68.3	53.6
	EViLBERT (VQA)	**75.0**	**75.1**	68.3	53.6
	EViLBERT (FT)	73.4	73.7	68.3	53.6
Full	LMMS 1024	75.0	75.2	100.0	100.0
	LMMS 2048	**75.5**	**75.7**	100.0	100.0
	GlossBERT	**76.2**	**77.0**	-	-

Table 5: F1 scores in the WSD evaluation. The row groups contain (top to bottom): baseline; results with coverage limited to EViLBERT; results with coverage limited to extended EViLBERT; results with full coverage; state of the art. Coverage (Cov.) as % of WordNet synsets and instances in ALL*. †: highest F1 that is statistically different from the row group best (McNemar's test, $p < 0.05$).

al. [2017], with the exception of SemEval-2015, our development set. To make the results easily comparable with other approaches in the literature, we also report results on the concatenation of all datasets in the framework (ALL). Experiments demonstrated that our EViLBERT-VQA embeddings are consistently better than the other multimodal variants. In particular, despite covering only 28.9% of the instances, our approach results in a gain of 0.5% over the baseline F1. When considering the extended EViLBERT the improvement is even more remarkable, outperforming the baseline by 1.8%. Our multimodal embeddings result in a substantial gain even when comparing with the unimodal LMMS-based architectures. More specifically, EViLBERT-VQA outperforms on ALL* the LMMS-based architectures by 0.1% and 1.0% in, respectively, the limited and extended settings, even though the latter are built with a larger BERT variant.

5 Conclusions

In this work, we introduced EViLBERT, a new approach for learning sense embeddings from both language and visual knowledge. EViLBERT is innovative in being the first multimodal learning technique which is able to learn task-agnostic representations for a wide range of concepts instead of words. Of course, learning multimodal embeddings requires the availability of a wide-coverage dataset illustrating concepts with images. Unfortunately, existing wide-coverage repositories are either limited to concrete concepts or not linked to a Lexical Knowledge Base. While most works in this field sidestep the issue by exploiting search engines, this is not an option for us since a methodology of this kind would not work in the case of polysemous words. To tackle this point, we created a large resource associating images with a wide-coverage set of concepts belonging to both the concrete and non-concrete domains. More specifically, we enabled image classification over an open set of concepts by reducing the verification of image-concept associations to Visual Question Answering. Our model relies on the recently introduced VLP architecture, and our experiments showed it to be reliable on zero-shot classification too. In addition, we described a strategy for creating effective embeddings from the hidden states of our VLP-based architecture, and we demonstrated their efficacy in the Word Sense Disambiguation task. Our experiments showed that multimodal information improves the performance of a simple WSD architecture, exceeding the scores obtained using competitive unimodal alternatives. EViLBERT is available at http://babelpic.org.

Acknowledgments

The authors gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE No. 726487 under the European Union's Horizon 2020 research and innovation programme.

References

[Anderson *et al.*, 2018] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proc. of CVPR*, pages 6077–6086, 2018.

[Barsalou and Wiemer-Hastings, 2005] Lawrence W Barsalou and Katja Wiemer-Hastings. Situating abstract concepts. *Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thought*, pages 129–163, 2005.

[Bevilacqua and Navigli, 2020] Michele Bevilacqua and Roberto Navigli. Breaking through the 80% glass ceiling: Raising the state of the art in Word Sense Disambiguation by incorporating knowledge graph information. In *Proc. of ACL*, 2020.

[Calabrese *et al.*, 2020] Agostina Calabrese, Michele Bevilacqua, and Roberto Navigli. Fatality killed the cat or: BabelPic, a multimodal dataset for non-concrete concepts. In *Proc. of ACL*, 2020.

[Collell and Moens, 2018] Guillem Collell and Marie-Francine Moens. Do neural network cross-modal mappings really bridge modalities? In *Proc. of ACL*, pages 462–468, 2018.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. ImageNet: A large-scale hierarchical image database. In *Proc. of CVPR*, pages 248–255, 2009.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional Transformers for language understanding. In *Proc. of NAACL-HLT*, pages 4171–4186, 2019.

[Frome *et al.*, 2013] Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. DeViSE: A deep visual-semantic embedding model. In *Proc. of NeurIPS*, pages 2121–2129, 2013.

[Goyal *et al.*, 2017] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proc. of CVPR*, pages 6325–6334, 2017.

[Hill and Korhonen, 2014] Felix Hill and Anna Korhonen. Learning abstract concept embeddings from multi-modal data: Since you probably can't see what I mean. In *Proc. of EMNLP*, pages 255–265, 2014.

[Huang *et al.*, 2019] Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. GlossBERT: BERT for Word Sense Disambiguation with gloss knowledge. In *Proc. of EMNLP-IJCNLP*, pages 3507–3512, 2019.

[Kiela *et al.*, 2014] Douwe Kiela, Felix Hill, Anna Korhonen, and Stephen Clark. Improving multi-modal representations using image dispersion: Why less is sometimes more. In *Proc. of ACL*, pages 835–841, 2014.

[Kiros *et al.*, 2018] Jamie Ryan Kiros, William Chan, and Geoffrey E. Hinton. Illustrative language understanding: Large-scale visual grounding with image search. In *Proc. of ACL*, pages 922–933, 2018.

[Krishna *et al.*, 2017] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[Kuznetsova *et al.*, 2020] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The Open Images dataset V4. *International Journal of Computer Vision*, 2020.

[Li *et al.*, 2019] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557, 2019.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proc. of ECCV*, pages 740–755, 2014.

[Loureiro and Jorge, 2019] Daniel Loureiro and Alípio Jorge. Language modelling makes sense: Propagating representations through WordNet for full-coverage Word Sense Disambiguation. In *Proc. of ACL*, pages 5682–5691, 2019.

[Lu *et al.*, 2019] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proc. of NeurIPS*, pages 13–23, 2019.

[Mikolov *et al.*, 2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proc. of ICLR*, 2013.

[Miller, 1995] George A. Miller. WordNet: A lexical database for English. *Commun. ACM*, 38(11):39–41, 1995.

[Moro and Navigli, 2015] Andrea Moro and Roberto Navigli. SemEval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proc. of SemEval@NAACL-HLT*, pages 288–297, 2015.

[Navigli and Ponzetto, 2012] Roberto Navigli and Simone Paolo Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, 193:217–250, 2012.

[Navigli, 2009] Roberto Navigli. Word Sense Disambiguation: A Survey. *ACM Computing Surveys*, 41(2):1–69, 2009.

[Plummer *et al.*, 2015] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k Entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proc. of ICCV*, pages 2641–2649, 2015.

[Raganato *et al.*, 2017] Alessandro Raganato, José Camacho-Collados, and Roberto Navigli. Word Sense Disambiguation: A unified evaluation framework and empirical comparison. In *Proc. of EACL*, pages 99–110, 2017.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proc. of NeurIPS*, pages 91–99, 2015.

[Silberer and Lapata, 2014] Carina Silberer and Mirella Lapata. Learning grounded meaning representations with autoencoders. In *Proc. of ACL*, pages 721–732, 2014.

[Tan and Bansal, 2019] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from Transformers. In *Proc. of EMNLP-IJCNLP*, pages 5099–5110, 2019.

[Zhou *et al.*, 2020] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and VQA. In *Proc. of AAAI*, 2020.