Solving Visual Madlibs with Multiple Cues

Tatiana Tommasi[1]
ttommasi@cs.unc.edu
Arun Mallya[2]
amallya2@illinois.edu
Bryan Plummer[2]
bplumme2@illinois.edu
Svetlana Lazebnik[2]
slazebni@illinois.edu
Alexander C. Berg[1]
aberg@cs.unc.edu
Tamara L. Berg[1]
tlberg@cs.unc.edu

[1] University of North Carolina at Chapel Hill, (NC) USA

[2] University of Illinois at Urbana-Champaign, (IL) USA

This paper focuses on answering multiple choice questions from the Visual Madlibs dataset [2] which was created by asking people to write fill-in-the-blank descriptions about persons (action, attribute, location), objects (affordance, attribute, location), and high-level concepts as future and past events.

We posit that in order to truly understand an image and answer questions about it, it is necessary to leverage rich and detailed global and local information. To explore this assertion, we represent the images by using CNN architectures trained on task-specific sources to recognize more than 200 scenes, 900 actions and 300 attributes (see Fig. 1). We extract the features both from the whole image and from regions selected to best match people and objects mentioned in the answers. We project both the visual and textual information in a joint CCA-embedding space [1] and at test time, we select the putative answer which obtains the highest cosine similarity with the image features. Finally we integrate multiple cues, through low-level visual feature stacking and high-level CCA score combinations. Our results show a significant improvement over the previous state of the art (see Tab. 1), and indicate that answering different question types benefits from examining a variety of image cues and carefully choosing informative image sub-regions.

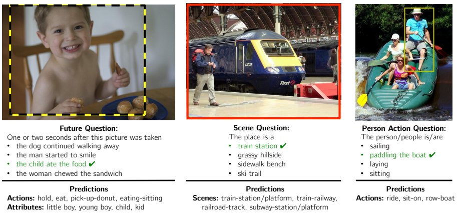
Future Question:
One or two seconds after this picture was taken
• the dog continued walking away
• the man started to smile
• the child ate the food ✔
• the woman chewed the sandwich
Predictions
Actions: hold, eat, pick-up-donut, eating-sitting
Attributes: little boy, young boy, child, kid
Scene Question:
The place is a
• train station ✔
• grassy hillside
• sidewalk bench
• ski trail
Predictions
Scenes: train-station/platform, train-railway, railroad-track, subway-station/platform
Person Action Question:
The person/people is/are
• sailing
• paddling the boat ✔
• laying
• sitting
Predictions
Actions: ride, sit-on, row-boat

Figure 1: Our method uses multiple deep networks trained on external knowledge sources to predict action, attribute, scene, and other diverse features from specific regions in the image. A CCA model trained on these features allows to score the putative answers and select the correct one for different different types of questions.

	Question Type		Baseline VGG	CCA Ensemble
a)	Interesting	Easy	79.53	83.20
		Hard	55.05	57.70
	Past	Easy	80.24	86.36
		Hard	54.35	60.00
	Future	Easy	80.22	86.88
		Hard	55.49	62.39
b)	Person Attribute	Easy	53.56	68.50
		Hard	42.58	55.90
	Person Action	Easy	84.71	88.34
		Hard	68.04	71.65
	Person Location	Easy	84.95	85.70
		Hard	64.67	63.92
	Person Object Relationship	Easy	73.63	78.93
		Hard	56.19	58.63
c)	Object Attribute	Easy	50.35	58.94
		Hard	45.41	54.50
	Object Affordance	Easy	82.49	87.29
		Hard	64.46	68.37
	Object Location	Easy	67.91	70.03
		Hard	56.71	58.01

Table 1: Improvement in accuracy by combining CCA scores from multiple cues.

[1] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV*, 2014.

[2] L. Yu, E. Park, A. C. Berg, and T. L. Berg. Visual Madlibs: Fill in the blank Image Generation and Question Answering. In *ICCV*, 2015.