Learning language through pictures

Chrupala, Grzegorz; Kadar, Akos; Alishahi, Afra

Published in:
Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)

Publication date:
2015

Document Version
Publisher's PDF, also known as Version of record

A Image retrieval with single words

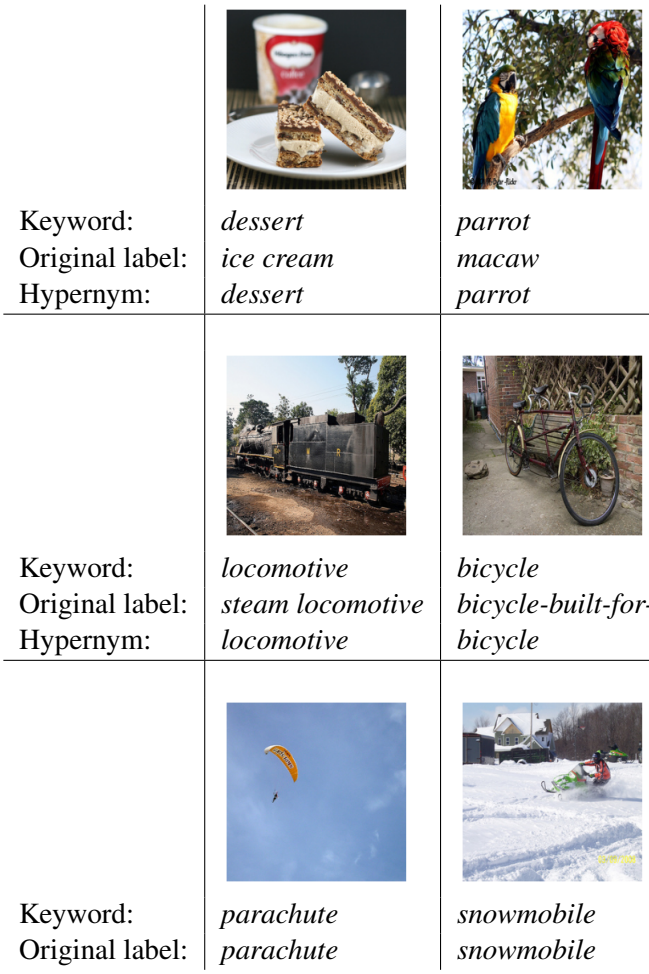


Figure 4: Sample images for single words. Under the images are the keywords that were used for the retrieval, the original label of the images and if it was not in our vocabulary its hypernym is included.

We visualize the acquired meaning of individual words using images from the ILSVRC2012 subset of ImageNet (Russakovsky et al., 2014). Labels of the images in ImageNet are synsets from WordNet, which identify a single concept in the image rather than providing descriptions of its full content. When the synset labels in ImageNet are too specific and cannot be found in our vocabulary, we replace them with their hypernyms from WordNet.

Figure 4 shows examples of images retrieved via projections of single words into the visual space using the MULTITASK model. As can be seen, the predicted images are intuitive. For those for which we use the hypernym as key, the more general term (e.g. *parrot*) is much more common in humans' daily descriptions of visual scenes than the original label used in ImageNet (e.g. *macaw*). The quantitative evaluation of this task is reported in the body of the paper.

B Effect of scrambling word order

In Figures 5–7 we show some illustrative cases of the effect for image retrieval of scrambling the input captions to the MULTITASK model trained on un-scrambled ones. These examples suggest that the model learns a number of facts about sentence structure. They range from very obvious, e.g. periods terminate sentences, to quite interesting, such as the distinction between modifiers and heads or the role of word order in encoding information structure (i.e. the distinction between topic and comment).

a pigeon with red feet perched on a wall .


feet on wall . pigeon a red with a perched


Figure 5: In the scrambled sentence, the presence of a full stop in the middle of a sentence causes all material following it to be ignored, so the model finds pictures with wall-like objects.

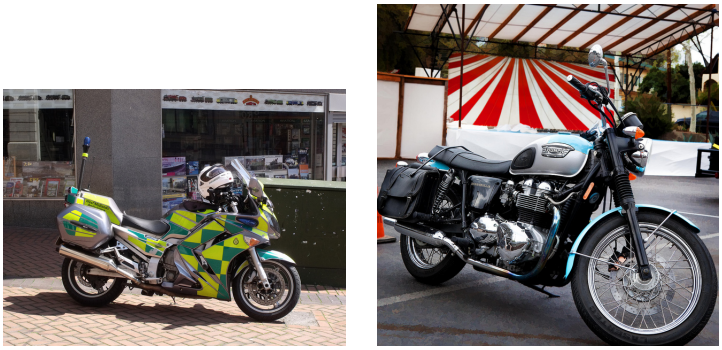
C Propagating distributional information through Multi-Task objective

Table 4 lists example word pairs for which the MULTITASK model matches human judgments closer than the VISUAL model. Some interesting cases are words which are closely related but which have the opposite meaning (*dawn, dusk*), or words which denote entities from the same broad class, but which are visually very dissimilar (*insect, lizard*). There are, however, also examples where there is no obvious prior expectation for the MULTITASK model to do better, e.g. (*maple, oak*).

Word 1	Word 2	Human	MULTITASK	VISUAL
construction	downtown	0.5	0.5	0.2
sexy	smile	0.4	0.4	0.2
dawn	dusk	0.8	0.7	0.4
insect	lizard	0.6	0.5	0.2
dawn	sunrise	0.9	0.7	0.4
collage	exhibition	0.6	0.4	0.2
bikini	swimsuit	0.9	0.7	0.4
outfit	skirt	0.7	0.5	0.2
sun	sunlight	1.0	0.7	0.4
maple	oak	0.9	0.5	0.2
shirt	skirt	0.9	0.4	0.1

Table 4: A sample of word pairs from the MEN 3K dataset for which the MULTITASK model matches human judgments better than VISUAL. All scores are scaled to the $[0, 1]$ range.

blue and silver motorcycle parked on pavement under plastic awning .



pavement silver awning and motorcycle blue on under plastic . parked

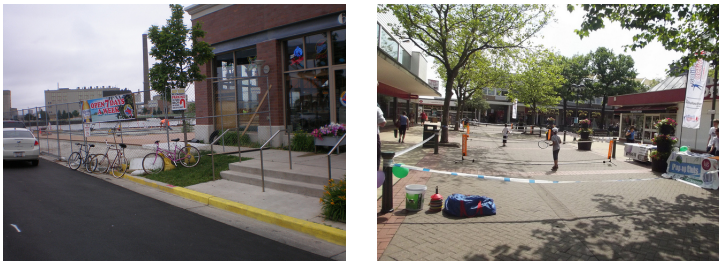


Figure 6: The model understands that *motorcycle* is the topic, even though it's not the very first word. In the scrambled sentence is treats *pavement* as the topic.

a brown teddy bear laying on top of a dry grass covered ground .

a a of covered laying bear on brown grass top teddy ground . dry

Figure 7: The model understands the compound *teddy bear*. In the scrambled sentence, it finds a picture of real bears instead.