Learning To Align Visual And Language Data

Xuewen Yang

May 31 2018

Our alignment model assumes an input dataset of images and their sentence descriptions. Our key insight is that sentences written by people make frequent references to some particular, but unknown location in the image. For example, in Figure 1,

the words Tabby cat is leaning refer to the cat, the words wooden table refer to the table, etc. We would like to infer these latent correspondences, with the eventual goal of later learning to generate these snippets from image regions. We build on the approach of Karpathy *et al.*, who learn to ground dependency tree relations to image regions with a ranking objective. Our contribution is in the use of bidirectional recurrent neural network to compute word representations in the sentence, dispensing of the need to compute dependency trees and allowing unbounded interactions of words and their context in the sentence. We also substantially simplify their objective and show that both modifications improve ranking performance.

We first describe neural networks that map words and image regions into a common, multimodal embedding. Then we introduce our novel objective, which learns the embedding representations so that semantically similar concepts across the two modalities occupy nearby regions of the space.

1. Representing images

Following prior work , we observe that sentence descriptions make frequent references to objects and their attributes. Thus, we follow the method of Girshick *et al.* to detect objects in every image with a Region Convolutional Neural Network (RCNN). The CNN is pre-trained on ImageNet [1] and finetuned on the 200 classes of the ImageNet Detection Challenge [3].

2. Representing sentences

To establish the inter-modal relationships, we would like to represent the words in the sentence in the same hdimensional embedding space that the image regions occupy. The simplest approach might be to project every individual word directly into this embedding. However, this approach does not consider any ordering and word context information in the sentence. An extension to this idea is to use word bigrams, or dependency tree relations as previously proposed [2]. However, this still imposes an arbitrary maximum size of the context window and requires the use of Dependency Tree Parsers that might be trained on unrelated text corpora.

3. Alignment objective

Since the supervision is at the level of entire images and sentences, our strategy is to formulate an image-sentence score as a function of the individual region word scores. Intuitively, a sentence-image pair should have a high matching score if its words have a confident support in the image.

4. Decoding text segment alignments to images

Consider an image from the training set and its corresponding sentence. We can interpret the quantity as the unnormalized log probability of the word describing any of the bounding boxes in the image. However, since we are ultimately interested in generating snippets of text instead of single words, we would like to align extended, contiguous sequences of words to a single bounding box. Note that the solution that assigns each word independently to the highest-scoring region is insufficient because it leads to words getting scattered inconsistently to different regions. To address this issue, we treat the true alignments as latent variables in a Markov Random Field (MRF) where the binary interactions between neighboring words encourage an alignment to the same region.

References

[1] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and F. F. Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[2] A. Karpathy, A. Joulin, and F. F. Li. Deep fragment embeddings for bidirectional image sentence mapping. In *NIPS*, 2014. 1

[3] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein. Ima-

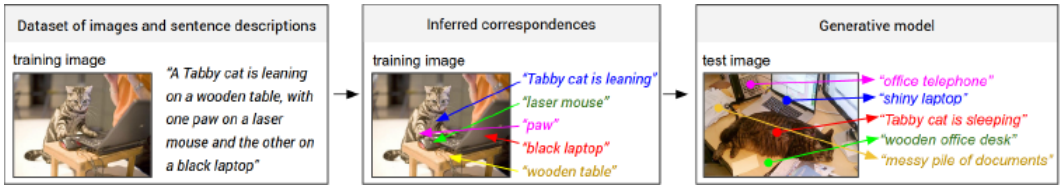

Figure 1. Overview of our approach. A dataset of images and their sentence descriptions is the input to our model (left).Our model first infers the correspondences (middle) and then learns to generate novel descriptions (right).

genet large scale visual recognition challenge. *IJCV*, 2014.
1