Stacked Cross Attention for Image-Text Matching

Kuang-Huei Lee[1], Xi Chen[1], Gang Hua[1], Houdong Hu[1], Xiaodong He[2]*
[1] Microsoft AI and Research, [2] JD AI Research (* Work performed while working at Microsoft)

[Paper] [Code]

This is the project page of Stacked Cross Attention Network (SCAN) from Microsoft AI & Research. Stacked Cross Attention is an attention mechanism for image-text cross-modal matching by inferring the latent language-vision alignments. This work will appear in ECCV 2018.

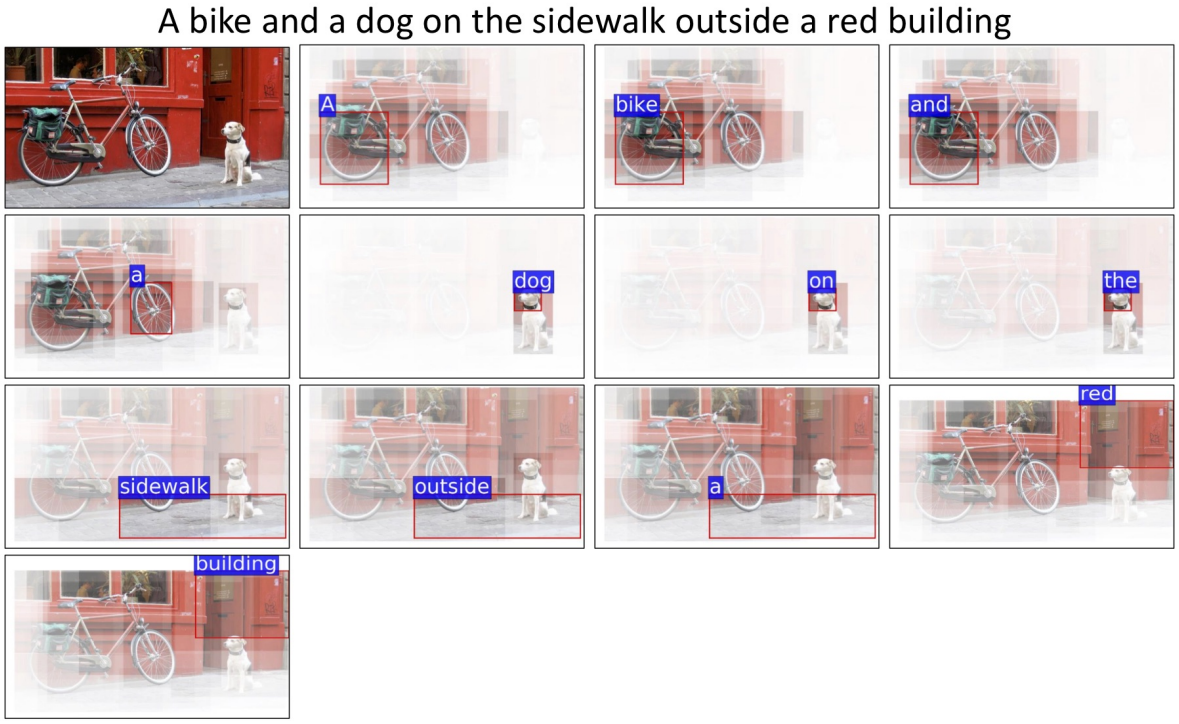
Abstract

In this paper, we study the problem of image-text matching. Inferring the latent semantic alignment between objects or other salient stuffs (e.g. snow, sky, lawn) and the corresponding words in sentences allows to capture fine-grained interplay between vision and language, and makes image-text matching more interpretable. Prior works either simply aggregate the similarity of all possible pairs of regions and words without attending differentially to more and less important words or regions, or use a multi-step attentional process to capture limited number of semantic alignments which is less interpretable. In this paper, we present Stacked Cross Attention to discover the full latent alignments using both image regions and words in sentence as context and infer the image-text similarity. Our approach achieves the state-of-the-art results on the MS-COCO and Flickr30K datasets. On Flickr30K, our approach outperforms the current best methods by 22.1% relatively in text retrieval from image query, and 18.2% relatively in image retrieval with text query (based on Recall@1). On MS-COCO, our approach improves sentence retrieval by 17.8% relatively and image retrieval by 16.6% relatively (based on Recall@1 using the 5K test set).

Infer the latent alignments between language and vision

An example of image-text matching showing attended image regions with respect to each word in the sentence. The brightness represents the attention strength, which considers the importance of both regions and words estimated by our model. This example shows that our model can infer the alignments between words and the corresponding objects/stuffs/attributes in the image (“bike” and

"dog" are objects; "sidewalk" and "building" are stuffs; "red" is an attribute.)



Cross-modal retrieval

Text retrieval for given image queries

(a)

1:Older women and younger girl are opening presents up . ✓
2:Two ladies and a little girl in her pajamas opening gifts ✓
3:A family opening up their Christmas presents . ✓
4:A mother and two children opening gifts on a Christmas morning . ✓
5:A little girl opening a Christmas present . ✓

(b)

1:Two men dressed in green are preparing food in a restaurant . ✓
2:A man , wearing a green shirt , is cooking food in a restaurant . ✓
3:A check with a green shirt uses a blowtorch on some food . ✓
4:An Asian man in a green uniform shirt with a white speckled headband is using a torch to cook food in a restaurant . ✓
5:An Asian man wearing gloves is working at a food stall . ×

(c)

1:A female runner dressed in blue athletic wear is running in a competition , while spectators line the street . ✓
2:A lady dressed in blue running a marathon . ✓
3:A young woman is running a marathon in a light blue tank top and spandex shorts . ✓
4:A lady standing at a crosswalk . ×
5:A woman who is running , with blue shorts . ✓

Image retrieval for given text queries

Query: *A man riding a motorcycle is performing a trick at a track .*

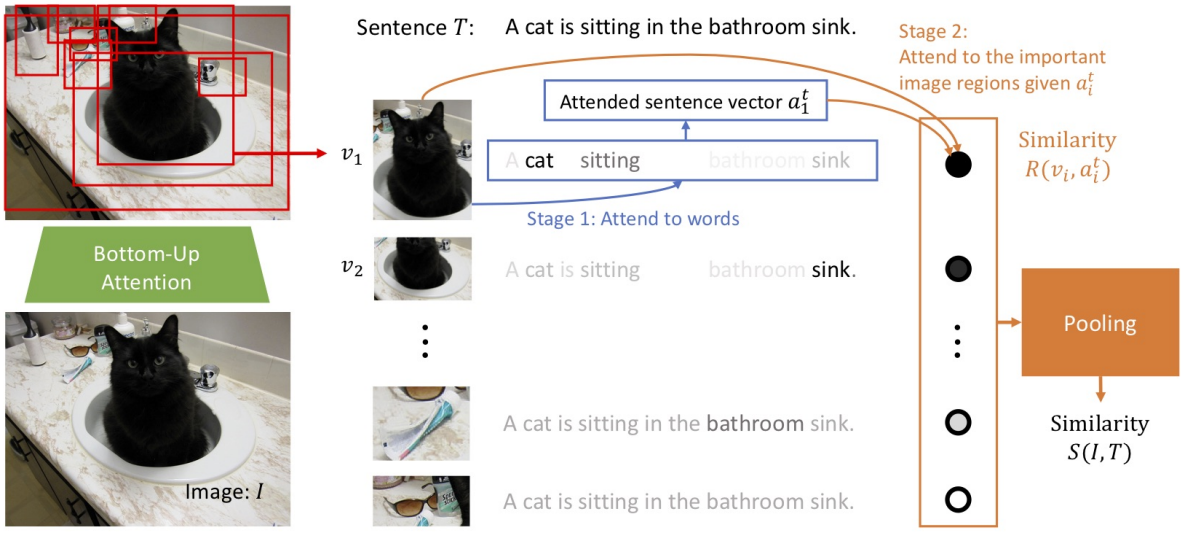
Query: *A baseball catcher trying to tag a base runner in a baseball game .*

Query: *Two dogs play by a tree .*

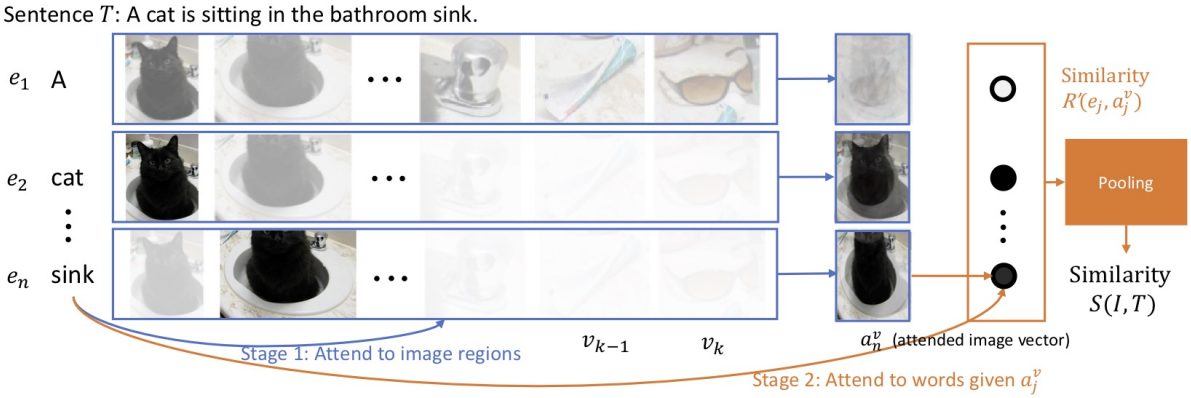
Query: *A construction worker is driving heavy equipment at a work site .*

Approach

Image-Text Stacked Cross Attention



Text-Image Stacked Cross Attention



Code

[Github repo]

Citation

```
@article{lee2018stacked,
  title={Stacked Cross Attention for Image-Text Matching},
  author={Lee, Kuang-Huei and Chen, Xi and Hua, Gang and Hu, Houdong and He, Xiaodong},
  journal={arXiv preprint arXiv:1803.08024},
  year={2018}
}
```

Acknowledgments

The authors would like to thank Po-Sen Huang and Yokesh Kumar for helping the manuscript, and Li Huang for helping with code release.