Machine Learning for Caption-Image Retrieval

Junyang Qian & Giacomo Lamberti {junyangq, giacomol}@stanford.edu

Stanford University - CS 229: Machine Learning

Motivation & Objective

Motivation: large amount of pictures are produced and stored every day, from medical images to personal photos; in this context, automated caption-image retrieval is becoming an increasingly attractive feature to search inside these enormous databases of images.

Goal: accurate retrieval of images given an input description.

Dataset

We train the models using the Microsoft COCO dataset [1]:

- 123,287 images: 113,287 for training, 5,000 each for validation and test;
- 5 human-annotated captions per image
- Example:

- Three teddy bears laying in bed under the covers.
- A group of stuffed animals sitting next to each other in bed.
- A white beige and brown baby bear under a beige white comforter.
- A trio of teddy bears bundled up on a bed.
- Three stuffed animals lay in a bed cuddled together.

References

[1] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, pp. 740–755, Springer, 2014.

[2] B. Klein, G. Lev, G. Sadeh, and L. Wolf, "Associating neural word embeddings with deep image representations using fisher vectors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4437–4446, 2015.

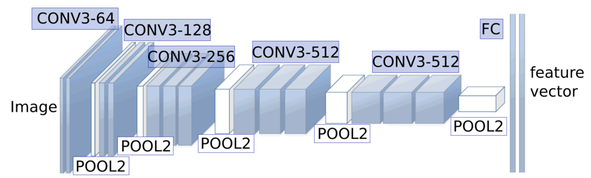
[3] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

[4] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun, "Order-embeddings of images and language," *arXiv preprint arXiv:1511.06361*, 2015.

Methods

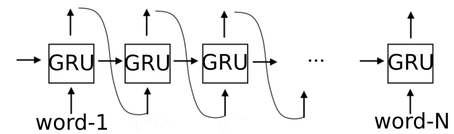
Image Model

The $fc7$ features (f_{VGG}) of the 19-layer VGG pre-trained network [2]



Language Model

- GloVe word vectors (f_{GloVe}) [3]
- Gated Recurrent Unit (GRU)



Baseline Model: $\hat{W}_{c,i} = \arg\min_W \sum_k \|f_{\text{VGG}}(i_k) - W \cdot f_{\text{GloVe}}(c_k)\|^2$.

Order Embeddin[4]: for each caption-image pair (c, i), we minimize an order violation:

$$f_i(i) = |W_i \cdot f_{\text{VGG}}(i)|, \qquad f_c(c) = |\text{GRU}(c)|, \qquad S(i,c) = -\left\|\max\{0, f_i(i) - f_c(c)\}\right\|^2.$$

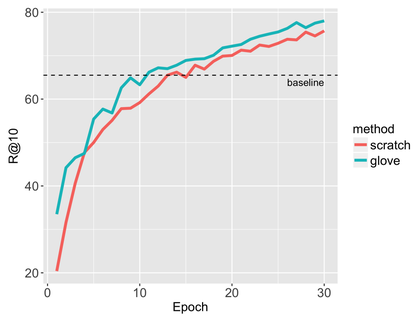
Dynamic Attention with Bidirectional RNN (in progress):

$$f_c(c) = |\sum_t w_{c,t} h_t|, \qquad h_t = \text{Bi-GRU/LSTM}(h_{t-1}, h_{t+1}, f_{\text{GloVe}}(c_t)).$$

Results

Evaluation metric on the test set:

$$\text{R@K} = \frac{\#\{\text{correct retrievals among top } K \text{ images}\}}{\#\{\text{text queries}\}}$$



Model	R@10
Baseline	65.5%
Order embedding	75.7%
Order embedding + GloVe	78.0%

Baseline

[dog]

[a guy riding a bike next to a train]

Order embedding + GloVe

[dog]

[a guy riding a bike next to a train]

Conclusions and future work

- By using pre-trained word vectors and fine-tuning on the present problem, we were able to achieve a R@10 of 78.0%.
- Future work will focus on finishing the dynamic attention mechanism and context-aware retrieval.