2018-9

# Entity-Grounded Image Captioning

Annika Lindh
*Technological University Dublin*, d14128847@mydit.ie

Robert Ross
*Dublin Institute of Technology*, robert.ross@dit.ie

John Kelleher
*Dublin Institute of Technology*, john.d.kelleher@dit.ie

# Entity-Grounded Image Captioning

Annika Lindh, Robert J. Ross, and John D. Kelleher

ADAPT Centre, Dublin Institute of Technology (DIT), Ireland

## 1 Image Captioning

The goal of an Image Captioning model is to generate a short, descriptive, natural language text when presented with an image. State-of-the-art models mainly employ neural encoder-decoder architectures with a Convolutional Neural Network (CNN) to handle the visual input and a Recurrent Neural Network (RNN) for language modelling. These models are typically trained with a cross-entropy loss to maximize the likelihood of replicating the human-written descriptions associated with the training images.

### 1.1 Generic Captions and Hallucinations

A noticeable limitation of RNN-based Image Captioning models is their tendency to output generic captions that avoid the interesting details which make each image unique [1]. There is also a tendency to "hallucinate" objects based on the context of previously generated words. RNN-based models have been shown to be particularly prone to producing exact copies of captions and parts of captions from the training set [2].

These limitations stem from multiple issues. On the training side, generic captions constitute a local minimum for the cross-entropy loss when averaged over all examples, particularly on biased datasets. On the evaluation side, the benchmark metrics reward models that learn to replicate common n-grams [3].

## 2 Sticking to the Facts

Methods of learning a more direct connection between text and visual elements have been proposed. Weighted attention maps and stochastically selected regions of focus have been shown to improve captioning results [4]; other models impose restrictions during word sampling based on visual detection results [5, 6].

We propose a model that takes the alignment between text and visual evidence a step further. Each segment of a caption is associated with a pair of regions along with their relative size and location. This allows the model to attend not only to specific objects but also their relationships. We hypothesize that training on individual segments will allow a stronger association between image content and relevant parts of the text.

### 2.1 Dataset

The proposed supervision is made possible by the Flickr30k dataset [7] with the additional Entities annotations [8]. These datasets contain images with human-annotated captions and bounding boxes for noun phrases. We preprocess the captions by dividing them into segments based on the entity annotations, as illustrated in Fig. 1.

### 2.2 Entity-Grounded Caption Generation

Our full model is composed of neural components for region-proposing, region ordering and caption generation. Since ground-truth data is available for each step of the pipeline, the components can be trained either end-to-end or individually on ground-truth data. This also allows us to establish upper-bounds for each trained component and to identify where improvement is most needed.

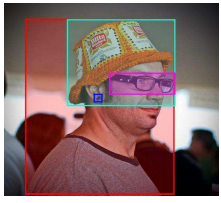


**Fig. 1.** A caption divided into segments relating to different image regions.

The key part of our model is the caption generation component based on an LSTM [9] architecture with the addition of a novel information gate. The information gate is implemented as a GRU unit [10] that updates the LSTM's cell state with information about the visual regions and their spatial relationship. Our model learns to output shorter segments of text followed by a "next"-token to request an update via the information gate with visual information relating to the next segment. By separating the input of visual information, we reduce the input dimensionality to the LSTM at each step while letting the GRU learn the mapping of the visual input.

Previous work [11] has shown correlations between visual grounding and a model's generalization capacity. Our hypothesis is that the stronger supervision along with the reduced input dimensionality of the language model will lead to improved visual grounding and better generalization to novel images. To test this hypothesis, the results will be evaluated both on the full test set and on sub-parts comprised of the most and least common types of images as in [2]. The captions will also be evaluated on the metrics for caption diversity in [1] to assess how well our proposed model addresses the issue of generic captions.

## 3 Conclusion

An urgent limitation of current Image Captioning models is their tendency to produce generic captions that do not always relate well to the content of the given image. We have proposed an approach to address this limitation by enforcing a stronger association between image regions and specific segments of text, including a novel information gate that allows separate frequencies and dimensionalities of the visual and textual input to LSTMs.

## References


1. Lindh, A., Ross, R.J., Mahalunkar, A., Salton, G., Kelleher, J.D.: Generating Diverse and Meaningful Captions: Unsupervised Specificity Optimization for Image Captioning. In: Artificial Neural Networks and Machine Learning - ICANN 2018. Springer (2018)
2. Devlin, J., Cheng, H., Fang, H., Gupta, S., Deng, L., He, X., Zweig, G., Mitchell, M.: Language Models for Image Captioning: The Quirks and What Works. 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) **2** (2015) 100–105
3. Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., Plank, B.: Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures. Journal of Artificial Intelligence Research **55**(1) (January 2016) 409–442
4. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In Bach, F., Blei, D., eds.: Proceedings of the 32nd International Conference on Machine Learning. Volume 37 of Proceedings of Machine Learning Research., Lille, France, PMLR (July 2015) 2048–2057
5. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Guided Open Vocabulary Image Captioning with Constrained Beam Search. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (2017) 936–945
6. Lu, J., Yang, J., Batra, D., Parikh, D.: Neural baby talk. CVPR (2018)
7. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics **2**(0) (February 2014) 67–78
8. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. International Journal of Computer Vision **123**(1) (May 2017) 74–93
9. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Comput. **9**(8) (November 1997) 1735–1780
10. Cho, K., Merrienboer, B.v., Bahdanau, D., Bengio, Y.: On the Properties of Neural Machine Translation: EncoderDecoder Approaches. Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (2014) 103–111
11. Tavakoliy, H.R., Shetty, R., Borji, A., Laaksonen, J.: Paying Attention to Descriptions Generated by Image Captioning Models. In: 2017 IEEE International Conference on Computer Vision (ICCV). (October 2017) 2506–2515