Thesis Progress Overview

Wout Vekemans & Thijs Dieltjens

November 2, 2015

This report documents our progress so far. It has an overview of the papers we found most interesting. We also describe what we have been doing the last couple of weeks.

We started from the implementation from [Karpathy and Fei-Fei,] and were able to run it on the Paris server. We tried to make a couple of extensions to the code, investigating possible improvements. For now we still use a RNN as language model, but in the future we might consider using a maximum entropy model as found in [Mitchell et al., 2015]. We also looked at three recently published papers. The first looks at a new way of modelling a network (FSMN) [Zhang et al., 2015]. The second one proposes a new dataset containing alignments between snippets of texts and corresponding Flickr30k images [Plummer et al., 2015]. The last one is similar to [Xu et al., 2015] but it adds scene factorization. [Jin et al., 2015]

What we did so far

1. Understanding and running the code from Karpathy, both on our own machines and Paris.

2. Extracting the entities proposed in [Plummer et al., 2015] to a useable format. Right now we don't see how this can easily be integrated with the system to improve it.

3. Rough implementation of FSMN. This was mainly done to see if there was a time improvement compared to LSTM and RNN. The implementation is not perfect, but it gives us an estimate of execution time. It did not seem to improve much over RNN.

4. Following [Jin et al., 2015] we looked at the scene factorization. More concrete, we implemented LDA and ran it to extract topic distributions for each image in the Flickr30k training set. We also created a simple feed-forward neural network to extract a topic distribution from unseen images. This seems to work quite well.

What we will do in the near future

1. Use the topic distribution as extra input for the RNN and see how this affects the results.

2. Further understanding and possible implementation of an attention based model.

References

[Jin et al., 2015] Jin, J., Fu, K., Cui, R., Sha, F., and Zhang, C. (2015). Aligning where to see and what to tell: image caption with region-based attention and scene factorization. pages 1–20.

[Karpathy and Fei-Fei,] Karpathy, A. and Fei-Fei, L. Deep Visual-Semantic Alignments for Generating Image Descriptions.

[Mitchell et al., 2015] Mitchell, M., Doll, P., Iandola, F., Gao, J., and Zitnick, C. L. (2015). From Captions to Visual Concepts and Back. *Cvpr*.

[Plummer et al., 2015] Plummer, B. a., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. (2015). Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models.

[Xu et al., 2015] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., and Bengio, Y. (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention.

[Zhang et al., 2015] Zhang, S., Jiang, H., Wei, S., and Dai, L. (2015). Feed-forward Sequential Memory Neural Networks without Recurrent Feedback. pages 3–6.