Exposé Final Project

Sempix
SoSe20

Alexander Koch, Meryem Karalioglu, Rodrigo Lopez Portillo Alcocer

July 16 2020

1 Introduction

At the intersection between the visual and textual modality, image captioning requires not only understanding of the image but also the ability to generate language. Being able to capture an image's most salient aspects is therefore a key characteristic of any image captioning model. MacLeod *et al.* found that visually impaired people for instance are very trusting when it comes to computer generated text [1], but are automatic image captions sufficiently accurate? Because it is not enough for a model to "understand" the image as a list of objects [2], the textual description of the image needs to be adequately specific.

Modern image captioning systems rely on neural network architectures. Commonly an encoder-decoder architecture is used consisting of two components: a convolutional neural network (CNN) for object detection and feature extraction, and a recurrent neural network (RNN) for language generation. These systems employ pre-trained CNN and RNN models. However, the generated image captions tend to become too generalized and apply to a multitude of images. We pose the assessment of the quality of image captions as a question of discriminativeness.

How Discriminative are Neural Image Captions?
Evaluation of Neural Image Captions Based on Caption-Image Retrieval.

Given a raw dataset of images, we will select one image captioning model to generate captions for all items in the set. We will split this dataset into multiple smaller sets. The task is to find the correct image in each of these smaller sets, given a textual description. Our system will retrieve a list of top images matching that query meant to describe the desired image. For this we will train a neural network to find a sentence representation that matches with the corresponding image representation. Matches can then be found using a k-nearest-neighbor approach (KNN) or similar. Additionally the input sets will contain a set of distractors to try to confuse the network. The final evaluation will be based on the success rate of correctly retrieved images over all the image sets.

2 Related Work

We base our work on *Show, attend and tell: Neural image caption generation with visual attention* [3] as it provides a captioning architecture to be used as a baseline. As an extension *Attacking Visual Language Grounding with Adversarial Examples:A Case Study on Neural Image Captioning* [4] proposes Show-and-Fool algorithm for crafting adversarial examples for neural image captioning. This can be useful when generating the set of distractors, such that the process can be automated instead of handpicking negative examples. The paper uses this approach to improve their model's robustness. Similar to this *Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data* [5] tries to improve the robustness and quality of captions. In this case however, this is done by performing self-retrieval. They use this self-retrieval as a guidance metric to encourage discriminative captions. The paper focuses especially on avoiding generalized caption and uses the REINFORCE algorithm with CiDEr as reward function. This shows an approach on how to evaluate our project's caption system.

3 Planned Work Packages

We will train our model on the COCO corpus. Possible pre-trained models for the encoder and decoder are VGG or ResNet and Word2Vec or GloVe respectively.

- Generate and test different image captioning models and find an appropriate dataset to test the model on.
- Create the smaller image batches on which to test the system.
- Implement language based models and convolutional neural nets to generate representations of captions and images.
- Train a model to find a weight matrix to map caption and image representations into the same space.
- Implement a system to retrieve images with the highest similarity to the given caption.
- Evaluate the system, report on different metrics.

References

[1] H. MacLeod, C. L. Bennett, M. R. Morris, and E. Cutrell, "Understanding blind people's experiences with computer-generated captions of social media images," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 5988–5999, 2017.

[2] R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, and B. Plank, "Automatic description generation from images: A survey of models, datasets, and evaluation measures," *Journal of Artificial Intelligence Research*, vol. 55, pp. 409–442, 2016.

[3] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, pp. 2048–2057, 2015.

[4] H. Chen, H. Zhang, P.-Y. Chen, J. Yi, and C.-J. Hsieh, "Attacking visual language grounding with adversarial examples: A case study on neural image captioning," *arXiv preprint arXiv:1712.02051*, 2017.

[5] X. Liu, H. Li, J. Shao, D. Chen, and X. Wang, "Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 338–354, 2018.