A young boy with a tree made of trees: Domain Adaptation of an LSTM-based Image Caption Generation Model

Dominik Künkele
Master in Language Technology
University of Gothenburg
guskunkdo@student.gu.se

Maria Irena Szawerna
Master in Language Technology
University of Gothenburg
gusszawma@student.gu.se

Abstract

Within this project we aim to explore how fine-tuning a CNN and LSTM-based image captioning model on a new dataset influences its performance on both the old and new data, along with what role some of the parameters, such as finetuning data size or the prevalence of out-of-vocabulary tokens in captions play in the process. We discover that differences in vocabulary play a major role and that providing too much fine-tuning data can cause the whole system to fail. We conclude that domain adaptation can be difficult to carry out, but perhaps worth attempting in certain cases.[1]

1 Introduction

Here we will have the introduction: some general talk, a little summary of the background reading (do we need more reading?), our questions that we want answered.

Automated image caption generation is an intersectional topic, combining advances in the fields of computer vision and natural language processing. It is not only an interesting and challenging task, but also one with useful real-life implementations, such as assisting people with vision impairments in a variety of ways. It is therefore quite relevant to explore a variety of issues that can be encountered in the field, in order to improve the performance of potential models and increase their relevance in practical applications.[2]

One potential issue with image captioning systems is that of domain adaptation, meaning adapting a trained system to somewhat different data. While it is not uncommon to encounter e.g. pre-trained image processing components in image caption generation models, it would appear that the general consensus is to train a new captioning model for every new dataset. This, however, may lead to an unnecessary waste of resources, especially if large amounts of data are used for training; another potential issue is there not always being enough data to train a model from scratch. An alternative in that case would be domain adaptation.[3]

Some more lightweight, but well-performing image captioning architectures include ones with a CNN-based image processing component and LSTM-based language generation component; their performance can be further increased by incorporating some form of attention, which allows for a more precise identification of what is relevant to the caption generation at what point. The particular model used in this project is based on (Xu et al., 2015), with the implementation by (Ilinykh). In the model, the encoder is a convolutional network, from which representations of elements of the picture - feature vectors - are obtained. The decoder is an LSTM network with attention, determining what parts of the image are the most relevant for generating a given element of the caption (ADD WHICH KIND OF ATTENTION?).[4]

Another way of employing attention in a similar model is presented in (**?**), where a special "visual sentinel" helps determine when to attend to the data from the image and when to the information from the language model, as not all the words are equally represented in an image - for instance, articles or other determiners, particles, etc., stem more from the requirements of a given grammatical structure or content word rather than the scene that is being described. However, since this project did not delve into the contribution of each of these elements in caption generation, it was found to be superfluous in our particular case, even though it can lead to improved performance.[5]

The topic of domain adaptation in caption generation itself does not appear to be widely stud-

[1] MS
[2] MS
[3] MS
[4] MS
[5] MS

ied. However, (Hessel et al., 2015) touche upon it in their paper on the importance of the language model versus the image processing component, where they prove that a good language model can produce decent captions even with poor image input. However, what the authors actually discuss is only fine-tuning the image recognition element of the model. While that is naturally helpful, especially since models of that kind require massive amounts of training data, it does not fully address the issue of fine-tuning a whole image caption generation model, and, in fact, utilizing pre-trained CNN components has become rather common.[6]

While all of the aforementioned papers seem to approach this issue as "generation from visual input," (Bernardi et al., 2016) describe other approaches to the issue of image caption generation, showing that this is simply one of the possible routes. Some shortcomings of this approach that they name include the inability of the system to generate captions for visual input where the system cannot recognize any salient elements. They also tackle the issue of evaluation, listing the variety of measures that can be used to that end when it comes to caption generation (as, naturally, simpler ones like accuracy would not be a good match); they admit that human judgement or measures that mimic human judgement are key when it comes to evaluating automatically generated captions.[7]

Given how little information we could find on the fine-tuning of CNN+LSTM image captioning models, within this project we would like to explore the extent to which a trained model of such architecture can be adapted to another domain and how some of the hyperparameters of that process can influence the outcome. We expect this process to be effective to an extent, but it may also break the model at some point. We hope for the fine-tuned model to perform better on the images from the new domain than the non-fine-tuned one, which we aim to evaluate using human judgement. Throughout the project, Dominik focused on the influence of hyperparameters, while Maria on the evaluation, with both authors inventing ways to adapt the model to their needs and identify and fix potential issues.[8]

In Section 2 we describe the materials and methods used to address the aforementioned questions, such as the architecture of the model, the datasets, and the evaluation methods. In Section 3 we present the results of our inquiry. In Section 4 we discuss how they answer our questions and how they relate to previous work. Finally, in Section 5 we offer conclusions and suggestions for future projects on this topic.[9]

2 Materials and Methods

Here we need to describe the databases that we used and the code that we used, and what we added on our own. Then we need to describe how we went about evaluation.

Within this section we will describe the resources that we used in the project. We will specify the characteristics of the image captioning model, the two datasets that were used to train or fine-tune it, and the questionnaire we utilized to elicit human judgements of the best captions.[10]

2.1 The image captioning model DK

2.2 The code MS and DK

2.3 Flickr8k MS or DK

2.4 imageCLEF MS or DK

2.5 Human judgement questionnaire MS

Having produced results (captions) using the model and the data described in the previous sections, we decided to test which of the plausible-sounding captions were the best match for the images. The only models that were somewhat consistent producing coherent captions were the non-fine-tuned (original) model and the "100" training data size models (with 4 variations of UNK filter: 0.1, 0.15, 0.2, no filter). In order to collect judgements, 20 images from 3 test sets each were collected, together with the captions generated by the aforementioned models. The test sets were: unfiltered imageCLEF, filtered (max 10% UNK tokens) imageCLEF, and Flickr8k. The questionnaire was constructed and hosted using Google Forms.

The questionnaire itself consisted of five sections. In the first one, a modicum of personal information, such as age, gender, education, native language was collected. The following three sections were constructed the same way, and each of them corresponded to a different test set. In each of those sections, 20 images, along with the corresponding captions were presented. The order of the

[6] MS
[7] MS
[8] MS
[9] MS
[10] MS

This is	a table
a table	with things
and even	more things

Table 1: Example table.

captions for each image was randomized to avoid bias. The participants were instructed to select one caption for each image which they thought was the best fit or the best description of the image. At the end of every section the participants had an opportunity to voice their general thoughts on the captions. Finally, in the last section the participants had the chance to include their final thoughts and opinions before submitting their answers.

The form was subsequently distributed online, and the responses were later processed and a summary will be presented in the Results section.

3 Results MS and DK?

Here we should summarize our results, in terms of:

1. *where did the performance exceed the best prior performance (we got BEST checkpoints?)* ***DK***
2. *which models actually generated reasonable captions? where did they start falling apart?* ***MS or DK***
3. *subjectively, which captions were best? include the questionnaire results here or subsequently* ***MS***

Include figures and tables as much as we can/as much as is reasonable.

3.1 Hyperparameters DK

3.2 Caption Evaluation MS

4 Discussion MS and/or DK

Here we compare our results to our initial expectations and decide how they answered questions, we also contrast them with prior work (if there is any). Highlight why what we did was relevant to the field.

5 Conclusions and further work MS and/or DK

Here we should summarize what we have done and suggest what more could be done on this topic.

References

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures.

Jack Hessel, Nicolas Savva, and Michael J. Wilber. 2015. Image representations and new domains in neural image captioning. *CoRR*, abs/1508.02091.

Nikolai Ilinykh. Image captioning tutorial.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044.

A Example Appendix

Here we should link our repository and also make sure that it is well-organized and has proper READMEs.