# *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 fine-tuning 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 (MS).

## 1 Introduction (MS)

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.

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. pretrained 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.

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.

Another way of employing attention in a similar model is presented in (Lu et al., 2016), 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.

The topic of domain adaptation in caption generation itself does not appear to be widely studied. However, (Hessel et al., 2015) touch 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.

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.

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.

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.

## 2 Materials and Methods (MS and DK)

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 (MS)

### 2.1 The image captioning model (DK)

The goal of this project was to analyze how the generation of captions changes when an already pre-trained model is fine-tuned on new data from a different domain. For this, we utilized code written by (Ilinykh) that implemented a model described by (Xu et al., 2015). We didn't adapt the model at all, but only fine-tuned it with a different dataset. In the following paragraph, the architecture model is summarized.

The model consists of an *Encoder* and a *Decoder*. The *Encoder* utilizes the ResNet-101 pre-trained model to extract features of the images (He et al., 2015). The last two layers are removed to only include the parts important for the feature extraction. The *Decoder* then initializes the hidden state of an LSTM with the image representation and generates the caption word by word. Hereby, the LSTM also makes use of an attention mechanism that weights different parts of the image higher and lower, depending on how important they are to generate a word at a certain step. This attention can then be visualized as a heat map on the image. Dropout is used to prevent overfitting to training data.

### 2.2 The code (MS and DK)

Loading the data was done in a *PyTorch Dataset* class. Here, it is possible to define multiple parameters for loading the data. First of all, only captions are loded that contain relations. Here, we implemented two different methods. One is filtering the captions based on the POS tag, and the other is a rule-based approach taken from (Ghanimifard and Dobnik, 2018). Furthermore, a total number of images that should be loaded can be defined. This loaded samples are then divided into training (0.8), validation (0.1) and testing (0.1) sets. Since the imageCLEF captions consists of multiple connected captions, we included an option to either just load the first caption or concatenate all with the conjunction *and*. Lastly, we noted that changing the dataset presents us with a lot of words that are not included in the original vocabulary. To deal with this, we implemented a variable filter threshold to exclude captions that contain too many unknown words (*unknown filter*).

The training loop was mostly taken from (Ilinykh). The pre-trained model is hereby fine-tuned on an unseen shuffled train dataset. After each epoch the model is validated with a validation dataset. This is done using BLEU-4 score. Since the imageCLEF dataset only consists of one independent caption instead of five as the Flickr8k dataset, the BLEU score is calculated only with one reference caption. The model is trained as long as the BLEU score doesn't improve for 20 epochs; the model with the best BLEU score is saved and used for testing and evaluating (DK).

The testing and evaluation part was conducted in a variety of Jupyter Notebook files. While the initial plan was to contain all the information in one file; unfortunately, due to GitHub limitations, that Notebook would have been too large to upload due to the number of generated and printed image and caption pairs. Therefore, all training data sized but 2k and 5k were tested in separate notebooks using the same methods, adapted from (Ilinykh), stored in a separate file, and imported into the notebook. The captions were generated using beam search and then presented together with the mappings of the attention on the image.

Other than relatively minor fixes that were required to make the code run, we decided to manually exclude the UNK token from the generation process by enforcing the lowest possible probability on it in the beam search. While this meant that we could not evaluate to what extent the models favored generating that token, it did allow us to obtain captions that sounded more natural, which was paramount for the subsequent human evaluation (MS).

### 2.3 Flickr8k (DK)

The model was trained on the Flickr8k dataset (Hodosh et al., 2013). It consists of 8,000 images that were hand-selected to depict actions and events of people or animals rather than scenery and mood. These images each are paired with five captions describing the image generally. The captions are not related to each other and were created by different annotators.

### 2.4 imageCLEF (MS)

The second dataset used in the project was the IAPR TC-12 Benchmark, called imageCLEF in this paper. This dataset consists of 20,000 images from a variety of real life situations. Each image is paired with a description consisting of a number of consecutive utterances specifying what the image is depicting (Grubinger et al., 2006). The assumption made in this project is that the first fragment of the description contains the most salient elements. Unlike Flickr8k, imageCLEF contains many more images that do not include people or animals. Additionally, the way the descriptions are formulated is also different from Flickr8k.

### 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 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. What is important to mention is that sometimes the model could not generate a caption for a given image, or the captions were the same across multiple models. Luckily, there were always at least two different captions. Nevertheless, this does make the interpretation of the results somewhat more difficult.

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)

## 3.1 Hyperparameters (DK)

During the training, we tried different hyperparameters for loading the data. Changing the method for filtering captions that contain relations didn't provide big differences in first tests. Therefore, we utilized solely the simpler POS-based approach that we developed ourselves for all experiments.

Concatenating all captions of the imageCLEF dataset was a too big of a challenge for the model and yielded worse results in first test than only using the first caption for each image. We concluded that the difference in syntax (more complex and longer sentences) as well as in semantics (more detailed descriptions) were too big to the Flickr8k dataset, the model was pre-trained on, but this hypothesis would need to be tested further in the future. For our experiments, we only used the first section of the caption.

For the experiments, the variable parameters were the *number of samples* as well as the *unknown filter*. The model was fine-tuned on a sample size (training set size) of 100 (80), 200 (160), 500 (400), 1,000 (800), 2,000 (1,600) and 5,000 (4,000). For each of these sample sizes, four different unknown filter thresholds were used: 0.1, 0.15, 0.2 and 1 (DK).

As mentioned previously in the Materials and Methods section, the only models that consistently produced captions for almost all test images were the original and the "100" models. The "200" models still generated most of the captions, but many of them started being rather nonsensical or non-grammatical. This process continued in the "500" and "1000" models with fewer and fewer captions being generated successfully, and those generated ones being absolutely nonsensical. Finally, in the "2k" and "5k" models most captions could not be generated (MS).

While the human evaluation is the focus of this experiment, it is still insightful to also check, how the model fares during training and validation, more specifically, how the BLEU score changes. Interestingly, the models with few data samples achieve very low BLEU scores, while the models with more samples achieve higher BLEU scores. Table 1 shows the best achieved BLEU scores during validation with an unknown filter of 0.1. For other filter values, the results look similar. This is opposite to the observation mentioned before. A reason for this may be the calculation of the BLEU score. The BLEU score only evaluates the appearance of words, in this experiment 4-grams, but does not focus explicitly on grammaticality. A lower BLEU score, but grammatical correct sentences may mean that the model does not learn to relate the new captions to the image features, but that it still relies on the pre-trained generations. On the other hand, a higher BLEU score with low grammaticality may hint to the fact, that the model is now able to generate the correct words, but looses the pre-trained ability to understand syntax.

| number of samples | highest BLEU score | fine-tuned epoch |
|---|---|---|
| 100 | 0.065 | 13 |
| 200 | 0.091 | 9 |
| 500 | 0.146 | 19 |
| 1000 | 0.165 | 18 |
| 2000 | 0.201 | 13 |
| 5000 | 0.222 | 9 |

Table 1: BLEU scores for different sample sizes (unknown filter 0.1)

Looking at the unknown filter, higher thresholds produce in general better BLEU scores. Table 2 shows the highest BLEU score for different unknown filter with 100 samples. This applies especially to the models with a larger number of samples, but the effect can be seen in all models. A reason for this may be the higher number of unknown tokens in the training data. Since, the model is also trained on generating unknown tokens, a higher share of unknown tokens in the training data will result in a higher share in the generated caption. Both sentences are therefore getting more similar, which results in a higher BLEU score. A solution for this problem could be excluding the unknown token from the model's vocabulary (DK).

| unknown filter | highest BLEU score | fine-tuned epoch |
|---|---|---|
| 0.1 | 0.065 | 13 |
| 0.15 | 0.104 | 18 |
| 0.2 | 0.176 | 12 |
| 1 | 0.162 | 19 |

Table 2: BLEU scores for different unknown filters (number of samples: 100)

### 3.2 Caption Evaluation (MS)

The questionnaire was filled out by 6 people. While we would like a larger sample size, we believe that we can still draw some preliminary conclusions from their answers, and that it is sufficient taking the scope of the project into account. It is also likely that the length of the questionnaire could have put off other potential respondents. The people who filled the questionnaire out ranged between 25 and 31 years of age. Three of them identified as female, two as male, one selected "other" as a response in that option. Half of the respondents had an MA/MSc degree or equivalent, two a BA/BSc or equivalent, and one selected secondary school education as the highest completed level of education. None of the people who filled the questionnaire out were native speakers of English. More detailed data about their native language was not collected.

| Image | Best model (% responses) |
|---|---|
| 15551 | 100_0.1 (50%) |
| 20144 | original (50%) |
| 20343 | 100_0.1 (33.3%) or 100_1.0 (33.3%) |
| 12761 | original (33.3%) or 100_1.0 (33.3%) |
| 22381 | 100_0.1 (83.3%) |
| 10821 | 100_0.1 (100%) |
| 18448 | original (100%) |
| 17173 | original (66.7%) |
| 13152 | original (66.7%) |
| 20272 | original (100%) |
| 11310 | 100_0.1 (66.7%) |
| 14183 | 100_1.0 (50%) |
| 10622 | original (50%) |
| 13123 | original (100%) |
| 19181 | 100_1.0 (83.3%) |
| 16920 | 100_0.1 (100%) |
| 11308 | original (50%) |
| 14016 | original (83.3%) |
| 21029 | original (100%) |
| 18355 | original (33.3%) or 100_0.15 (33.3%) or 100_0.2 (33.3%) |

Table 3: Best captions per image in the unfiltered imageCLEF test set.

While the captions themselves can be seen in the Jupyter Notebook files, the aim of the questionnaire was to reveal which model or models performed the best on which test set. Therefore, the results presented here are divided between the sets and include only the information on which model's caption was deemed the best, as well as how many people voted for it. Table 3 contains the information for the unfiltered imageCLEF set, Table 4 - for the filtered one, and Table 5 for Flickr8k. Image names are the file names without the .jpg extension. In some cases more than one model's caption received the same amount of votes, or the winning caption was generated by more than one model.

| Image | Best model (% responses) |
|---|---|
| 40416 | original (83.3%) |
| 39158 | original (50%) |
| 25053 | original (83.3%) |
| 30620 | 100_1.0 (66.7%) |
| 32397 | 100_0.2 (66.7%) |
| 38937 | original (50%) |
| 39005 | 100_0.1 (33.3%) or 100_0.15 (33.3%) or 100_0.2 (33.3%) |
| 40120 | original (66.7%) |
| 23588 | 100_1.0 (83.3%) |
| 39472 | 100_0.1 (66.7%) |
| 30138 | original (50%) |
| 40202 | 100_0.1 (83.3%) |
| 35895 | original (66.7%) |
| 32663 | 100_0.2 (83.3%) |
| 30705 | original (83.3%) |
| 38081 | 100_0.15 (66.7%) |
| 31571 | original (100%) |
| 35858 | original (66.7%) |
| 37836 | 100_0.1 (33.3%) or 100_0.15 (33.3%) |
| 39239 | 100_0.15, 100_0.2, and 100_1.0 (50%)[1] |

Table 4: Best captions per image in the filtered imageCLEF test set.

As can be noticed in Table 3, the majority of the "good" captions for the unfiltered imageCLEF set were generated by the original, non-fine-tuned model, or by the one with a 10% filter. Some captions from the model fine-tuned on non-filtered data turned out to have been rated high as well. Finally, other models' captions were rated as good when they overlapped with the original model's caption. Overall, while the judgements are split rather evenly, the original model seems to generate better captions for this test set. It is worth pointing out that relatively few captions were unanimously voted best.

When it comes to the filtered imageCLEF test set (Table 4), where a maximum of 10% tokens in the original caption could be unknown tokens, the spectrum of what models' captions were considered good is even wider. There was only one case where one caption was voted best by all the participants. Half of the best captions were generated by the original model, and half by various fine-tuned ones, with none of them being a clear front-runner in this case.

Finally, as for the original Flickr8k-based test set, as seen in Table 5, the original model without fine-tuning performed best, although there were instances where the 10% or 20% fine-tuned models generated better captions. Many of the original model's captions were also chosen unanimously.

| Image | Best model (% responses) |
|---|---|
| 2116444946[...] | original (100%) |
| 2316097768[...] | original (100%) |
| 2439384468[...] | original (66.7%) |
| 2112921744[...] | 100_0.2 (50%) |
| 2392460773[...] | original (100%) |
| 2434006663[...] | original (100%) |
| 2308256827[...] | original (100%) |
| 2111360187[...] | original (66.7%) |
| 2271671533[...] | original (66.7%) |
| 2328616978[...] | original (100%) |
| 2456907314[...] | original (66.7%) |
| 2229179070[...] | 100_0.1 (100%) |
| 2279980395[...] | 100_0.1 (88.3%) |
| 2393971707[...] | original (83.3%) |
| 211277478[...] | 100_0.1 (50%) |
| 2337919839[...] | original (83.3%) |
| 2447035752[...] | original (50%) |
| 23445819[...] | 100_0.2 (66.7%) |
| 2448210587[...] | original (33.3%) |
| 2445654384[...] | 100_0.1 (50%) or 100_0.2 (50%) |

Table 5: Best captions per image in the Flickr8k test set.

When it comes to the comments left by the respondents after each section, they were not very positive. The captions in the first content section were judged as "unfitting," "strange," "ungrammatical," "inaccurate," "nonsensical," or "confusing," with some of the respondents admitting that the mismatch between the image and captions made them anxious or made their "brain hurt." For the second content section, some people considered the captions better, and some worse than previously, and the feeling of anxiety persisted here too. In addition, half of the respondents pointed out that the captions tend to include people in the descriptions even when there are no people in the images. It seems that the impressions of the captions in the last content section were better: while people said that they were still "strange," "weird," or "confusing," they also admitted that the verbs in the sentences tended to be accurate descriptions of the events in the images, and that they were overall "somewhat accurate." The person who expressed their feelings of anxiety previously now said that they "started to get used to them," which could also be a consequence of there being a smaller mismatch between the images and the captions.

In the final remarks many of the respondents expressed their dissatisfaction with the overall quality of the captions, leaving comments such as "r u b b i s h" or "my brain melted." Two people mentioned again how the captions would include people despite there not being any in the images. One person said that the simpler captions or images were clear, while the more complex ones were problematic or mismatched. Another person said that these "didn't sound like human-written captions," and yet another that they oftentimes "missed the essence of the picure."

Overall, the original model performed best on the original test set. The most captions generated by fine-tuned models were selected as best for the filtered imageCLEF test set. The unfiltered test set was also fairly balanced between the fine-tuned and original models. While for the unfiltered data the unfiltered and 10% fine-tuned models can be identified as the best out of the fine-tuned ones, the same is impossible to determine for the filtered test set. From these results it can be concluded that while fine-tuning does improve the quality of the captions somewhat (both for imageCLEF and Flickr8k), it does not have that effect for every image. Judging by the open responses from the participants, the captions generated for Flickr8k were the least wrong, but their quality was still low, and the ones for imageCLEF were even worse.

Clearly, while some degree of improvement can be obtained by this kind of fine-tuning, it is far from perfect and the captions rarely match the images. This is likely due to the disparity between the kinds of images in the two sets. Another clue to that being a major issue is the models' tendency to - as pointed out by the respondents - insert people into captions when there were none in the original image. A large portion of the captions in the Flickr8k dataset are constructed that way, and it must be one of the things that the model learned to try to identify in the image at all cost and feature in the caption, and that the fine-tuning did not help mitigate.

Finally, it is worth keeping in mind that some of the apparent weirdness of the captions may stem from the relatively small vocabulary that the models had, which was likely lacking the proper words to describe certain elements in the images, even when the ones whose captions contained many new words were filtered out.

## 4 Discussion (DK)

As seen in this project, domain adaption for a pre-trained image captioning model can be very challenging. It was surprising to what extent the pre-trained model was disrupted when fine-tuned on data from another domain. Even though the dataset for fine-tuning was very different than the dataset for pre-training, we hoped that the fine-tuned model would still perform better on the new domain than the original. This was not the case for multiple reasons, mentioned above. At the very least, we expected it to be able to generate sensible captions for most of the images and that it would only produce big mistakes for a smaller part. With the rather opposite results discussed before, two major challenges for domain adaption got highlighted. First, the model cannot adapt well to a different structure of captions. This results in ungrammatical sentences, when fine-tuned on many samples from the new domain. Secondly, the model is not able to adapt to a different topic of images. That results in 'reusing' old captions on new images and not learning its image features, when fine-tuned on few samples. For a deeper analysis, it may be helpful to first fine-tune the model on a domain that is closer in one of these aspects, for example a domain, that uses a similar way of describing the images, but focuses on topics. This could test the hypothesis made in this paper.

## 5 Conclusions and further work (MS)

Throughout this project we have explored and tested how a CNN and LSTM-based image caption generation model could be fine-tuned, or adapted to a new domain, as represented by a somewhat different image-caption dataset. We have identified points that can be especially problematic, such as differences in the vocabulary and sentence structure, and the subsequent generation of UNK tokens in the captions as well as ungrammatical sentences. We have determined that if the domains are too different and too much fine-tuning data is fed to the model, the generation will start to completely fail or focus only on a very narrow range of words. We have also collected human judgements and determined that while fine-tuning does slightly increase the quality of the captions on the new dataset, all of them, regardless of the model, were still considered weird, mismatched, or unnatural, and that the structure of the sentences in the original dataset may have a large influence on how they are generated even after fine-tuning (e.g. trying to feature people in every caption).

While we have addressed the issues that we raised in the introduction, we were not able to explore all the possible variables in domain adaptation of models of this kind, as it was beyond the scope of this project. However, we are certain that it would be very interesting to see how other hyperparameters than just the training data size can influence the outcome, and what the interplay between them could be. It would also be fascinating to see similar research pertaining to other model architectures, especially trying to identify the optimal hyperparameters for fine-tuning transformer-based image caption generation models. Of course, following up on the idea from the Discussion section, to verify our conclusions from the results by fine-tuning on a more similar dataset is another topic worth investigating. Finally, it would be interesting to explore other potential ways of dealing with the vocabulary mismatch between different datasets.

## 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.

Mehdi Ghanimifard and Simon Dobnik. 2018. Language model perplexities as multi-word distributional vectors of spatial relations. In *Swedish Language Technology Conference 2018*. SLTC2018.

Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. 2006. The iapr tc12 benchmark: A new evaluation resource for visual information systems. *Workshop Ontoimage*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

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

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Intell. Res.*, 47:853–899.

Nikolai Ilinykh. Image captioning tutorial.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2016. Knowing when to look: Adaptive attention via a visual sentinel for image captioning.

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 Appendix

1. The project repository with instructions: `https://github.com/Turtilla/aics-project`

## B Statement of Contribution

While really detailed information on who developed what is available in the code, the code is not the only thing that we shared work on. Below one can find the details of who contributed what to the project.

1. **Developing project ideas:** both Dominik and Maria. Selecting the initial ideas from the ones suggested in the course page mostly by Dominik, ideas about working with comicbooks and contacting researchers related to that by Maria. Final idea selected under the advice from Nikolai and Simon by both project authors.

2. **Setting up resources:** figuring out how to access the datasets used in the project by Dominik, using the existing code by Nikolai both by Maria and Dominik.

3. **Code:** detailed information on code contribution can be found in code documentation.

   (a) Evaluation Jupyter Notebooks: Maria

   (b) Dataset classes: Dominik

   (c) Relation filters: Maria and code shared by Simon

   (d) Fine-tuning loop: adapted by Maria and Dominik from Nikolai's code.

   (e) Models and some preprocessing: Nikolai's code.

   (f) Caption generation and attention visualization: Nikolai's code, with changes by Maria and Dominik.

   (g) Division of the code into separate files, turning it into classes: Dominik.

   (h) Documentation: Maria with Dominik's help.

4. **Questionnaire:** Maria.

5. **Evaluation:** evaluation of the outputs of the notebooks by Dominik and Maria (for the questionnaire), evaluation of questionnaire answers by Maria.

6. **Writing up:** both Dominik and Maria. While the initials signify the person who wrote most of the section, small changes by the other person are possible - and are marked by (initials) at the end of a paragraph. This is done especially if a certain paragraph was moved from one section to the other. Otherwise, for sections where authorship is more evenly shared, the same method of marking contribution is employed.

7. **Repository management:** Maria with Dominik's help.

8. **Presentation:** Maria with Dominik's help.