Context-Based Enriching Image Captioning
For News Photos

Stephan Halbritter
Hamburg University of Applied Sciences, Dept. of Computer Science
Berliner Tor 7, 20099 Hamburg, Germany
stephan.halbritter@haw-hamburg.de

***Abstract*—Most image captioning algorithms are only descriptive and do not include contextual information. For some use cases, this is not sufficient. We look into different approaches on how to generate captions enriched with context, on how to extract that contextual information from external sources and if these strategies are suitable for the use case of enriching image captions of press photos with context from accompanying news articles.**

I. Introduction and Research Question

Generating captions of images combines computer vision and NLP: automatic detection of objects, calculation of their relationship with each other and the generation of a description in natural language. It's one of the many fields in which deep learning methods lead to great improvents in the last years. Today, common models can recognize hundreds of categories and create quite accurate descriptions of what is shown in the images.

On the other hand, there's still large room for further improvements. Of course, there is still room to further improve the error rate and generated captions can be just wrong or do not make any sense. Even if they do, the output often consists of just plain factual descriptions with a very basic sentence structure, for example "a woman holding a clock" or "a giraffe standing in the forest". Depending on the use case, this may or may not be a satisfactory result.

Another shortcoming lies in the available information used to create the caption as in many cases, the information is only based on the detected objects in the image. But images have a context and most human observers see more than the simple eye can meet. A photo is shot at a specific place and a specific date. People, streets and cities have names. As this (*meta*)-information cannot be extracted from the pixel value of the image alone, it has to come from additional sources.

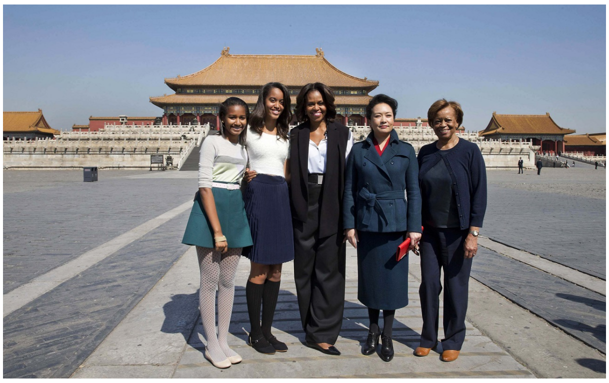


Figure 1: *"Sasha Obama, Malia Obama, Michelle Obama, Peng Liyuan et al. posing for a picture in front of the Forbidden City . . . "* (Tran et al. 2016)

In this paper, we evaluate different approaches for usage with a data set provided by *Deutsche Presse-Agentur (dpa)*. As a result, this will narrow the context to news texts and their accompanying images (see sec. IV). We provide insight into what useful context in news texts and their metadata can be in regard to the related images, what strategies exist to extract them and what possibilites to inject this data into the captioning process seem promising and worth pursuing.

The main goal is to answer the question, how image captions enriched with context-based information provided from corresponding text could be generated.

II. Related Work

In this section we provide background on related and useful work in regard of creating enriched captions. A more detailed view on some of the theoretical concepts used in image captioning can be found in Section III.

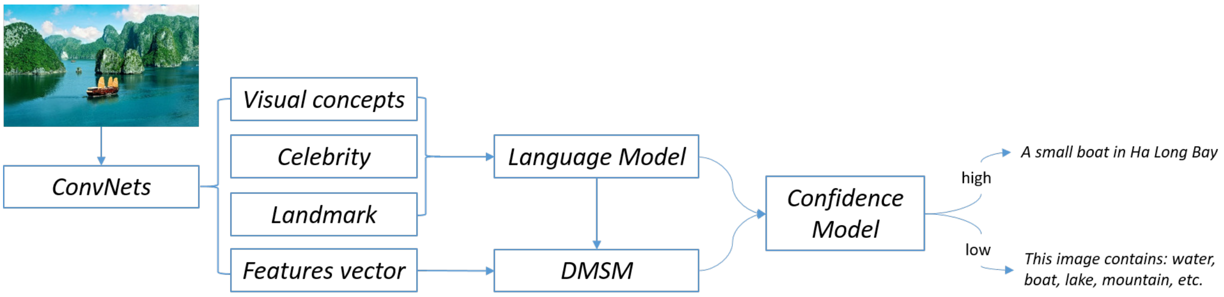


Figure 2: Image Caption Pipeline used in Tran et al. (2016)

A. *Rich Image Captioning in the Wild*

Tran et al. (2016) propose a framework with the ability to detect additional entities and as a result generate more detailed image captions, recognizing celebrities as well as landmarks. An example of a generated image caption is shown in figure 1.

These results are made possible by using multiple specialized models. Their basic pipeline can be seen in figure 2. Most models for image captioning use pretrained *Convolutional Neural Network (CNN)* for detecting objects. Here, a compositional approach is used. Huge collections of high-quality training data were created beforehand to train extra domain-specific models for the recognition of celebrities and landmarks. Especially landmarks pose an interesting problem if they are not displayed from the usual point of view. Additional models to evaluate and ensure the quality of the datasets were build.

The presented results are impressive, but that concept can be broken down to the idea of highly specialized CNNs and a huge effort on the preprocessing side and creation of the datasets. This makes this approach not really suitable to enrich captions with external contextual information, but shows the importance of a solid data collection.

B. *Image Captioning at Will*

In this paper, You et al. present an approach for injecting sentiment into captions (see figure 3). Their strategy is based on the findings of a single binary unit in text generating *Recurrent Neural Networks (RNN)*. This so called *sentiment unit* directly affects the sentiment in the text output (Radford et al. 2017). Two different models for injecting sentiment are presented.

Direct injection adds the sentiment value in the embedding layer of the text generating RNN. This way, it influences the outcome of each word in every step. Although this way there's a higher probability that the resulting outcome is affected by the sentiment unit, only a few parts of a sentence do actually contribute to sentimental meaning and the chance of unwanted side effects due to the injection rises.

Injection by sentiment flow takes this into account by adding an additional *sentiment cell*. This structure is interlinked with the *Long Short Term Memory (LSTM)* cells and responsible for injecting sentiment. This enables the model to learn on which words to activate the sentiment unit and avoid useless and faulty injections.

This paper shows how to injecting sentiment information during the text generation process in the RNN. This shows the possibility of adding information, which in contrast to Tran et al. (2016) is not extracted from the image itself.

C. *Globally Coherent Text Generation With Checklist Models*

The models of Kiddon et al. (2016) keep checklists of words that should be mentioned in the image caption.

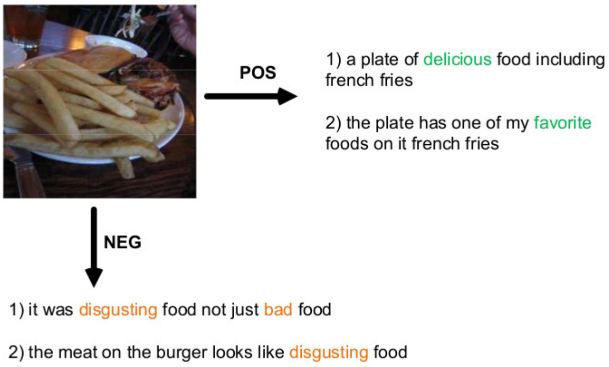

POS
1) a plate of delicious food including french fries
2) the plate has one of my favorite foods on it french fries
NEG
1) it was disgusting food not just bad food
2) the meat on the burger looks like disgusting food

Figure 3: Enriching with sentiment (You et al. 2018)

It is based on an idea similar to the sentiment cell described in Section II-B. An attention mechanism learns to predict the relevance of the checklisted words (for more details about attention see section III-B). This probability is later used in each step of the text generation process to decide if a word should be injected.

This is an interesting approach if you want to make sure that some contextual information has its impact on the image caption and is part of the outcome.

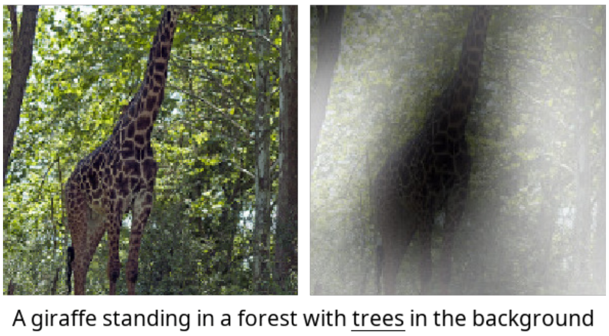


Figure 4: Attention visualization for generating the word *trees* in the image caption. Adapted from Xu et al. (2015)

III. Theoretical Background

Most current algorithms for image captioning are directly based or at least heavily inspired by *Show and Tell: A Neural Image Caption Generator* (Vinyals et al. 2015) and its immediate successor *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention* (Xu et al. 2015). The latter added an Attention mechanism for further improvements.

A. Encoder-Decoder Framework

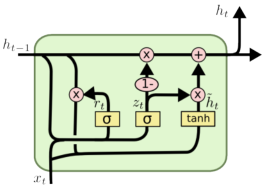
In general, the overall framework is based on an encoder-decoder framework with the encoder consisting of a CNN and the decoder of a RNN.

1) Encoder: a CNN takes over the part of the encoder. An image is fed into the encoder and through multiple convolutional layers is encoded into an intermediate representation. Vinyals et al. (2015) use the image representation from one of the CNNs' last, fully-connected layers. That means that the data contains probabilities about the categories the CNN could identify in the image. There is no spatial information of the original image left, we do not know which parts of the image are responsible for the categorizations.

To further enhance and accelerate the learning process, pretrained CNNs such as *Oxford VGGnet*[1] are used as a starting point.

The improved approach of Xu et al. (2015) uses the output of one of the convolutional layers further at the beginning of the CNN. Even after some convolutions, this multi-dimensional layer still contains spatial information about the input image. This is in contrast to using the fully-connected layer and allows the use of an attention mechanism in the decoder as we will see in Section III-B.

2) Decoder: an RNN generates the resulting image caption with the intermediate image represention from the encoder as input. In general, an RNN consists of LSTM cells or one of its variants like *Gated Recurrent Units (GRU)* (see figure 5). They are able to keep state about previous input features and allow the incorporation of that state into the evaluation of the current features and as a result in the generation of the next word. In the first models for image captioning, the input consisted of the end result of the CNN. This result of a fully-connected layer is a large vector and contains the probabilites of categories. This was fed into the RNN just once as an initial input. A big step forward in regard of the output quality was the introduction of an *Attention mechanism*.



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

Figure 5: Gated Recurrent Unit. (Olah 2015)

B. Attention Mechanism

The basic idea of *Attention* (sometimes called *attentional interface*) is to focus on just a subset of the information input, hopefully the most relevant for the current task.

This is especially useful when using and/or producing sequential data like text, for example in translation models. Here, attention helps the output RNN to focus on the relevant parts of the input sentence in regard

[1] http://www.robots.ox.ac.uk/~vgg/research/very_deep/

to the next output word and creates necessary context. Among other things, this produces better results where the sentence structure of input and output language differ a lot. A good overview is given by Olah and Carter (2016).

Attention is also very useful when generating image captions. In the training process, it learns the relationship between the single words of a caption and subsets of the input, in this case different areas of the image. This is why the intermediate representation from the encoder has to contain spatial information and therefore comes from one of the convolutional layers.

Figure 6 shows an overview on how attention can be used in the decoder for image captioning.

The black boxes in the upper part represent a LSTM or GRU cell like in figure 5. In each time step t, a new word y_t is generated, using the output word y_{t-1} and the hidden state h_{t-1} of the previous cell. In addition to previous models, there's an additional input, the context vector c_t. It is calculated by the attention mechanism shown in the orange block.

The attention interface has two inputs, the hidden state h_{t-1} and the intermediate representation of the image, output by a convolutional layer of the encoding CNN. The image representation still conveys spatial information, it can be split into L different subimages a_1 to a_L.

For each of these subimages, a weight $\alpha_{t,i}$ is calculated (see equation 1). W_h is the respective trainable weight of the hidden state of the previous generated word, W_a for the image representation. As they are trained, they learn how relevant this image part is considered in regard to the next word in the caption. Calculating the hyperbolic tangent of their sum and normalizing the result with *softmax*, we map the image region's relevance to $[0,1]$ with a total sum of 1 for all regions. The resulting context c_t is the sum of the weighted image parts and fed into the RNN cells.

$$\begin{aligned} e_{t,i} &= tanh(W_h \cdot h_{t-1} + W_a \cdot a_{t,i}) \\ \alpha_{t,i} &= \text{softmax}(e_{t,i}) \\ c_t &= \sum_{i=1}^{L} a_{t,i} \cdot \alpha_{t,i} \end{aligned} \tag{1}$$

An interesting side effect of this is the ease to generate a visualization of this process. By multiplying the subimage weights $\alpha_{t,i}$ with 255, we can interpret them as a grey scale image. So according to (see eq. 1), the weight of unfocused areas tends to zero. For this reason, those areas are darker. Vice versa, focused areas are lighter. Overlaying the original image with this grey scale image, we can cleary see which parts were considered relevant. For an example, in figure 4, you see the that the trees in the background were focused when generating the word *trees* in the image caption »*A giraffe standing in a forest with trees in the background*«.

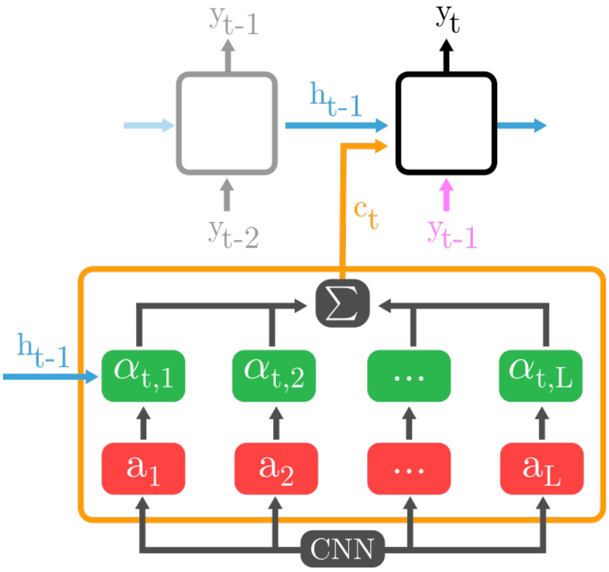


Figure 6: Example of attention used in image captioning

Attention enables the text generating network to recognize which parts of the input are relevant for each generated word. Hence it not only improves the quality of the text output, but seems the most promising part of the encoder-decoder framework for adding contextual information to enrich the caption.

IV. Data Set

The news corpus provided by *dpa* contains ~220 000 items in German, formatted in the »multimedia news exchange standard *NewsML-G2*«.[2] This will be the main source to identify contextual information to enrich the image captions of the accompanying images.

As the news texts are written by professional journalists, they are formal in contrast to e.g. tweets and contain only a negligible number of spelling mistakes and grammatical errors. Besides the actual news content put at disposal in plain text and HTML, a short

[2] https://iptc.org/standards/newsml-g2/

summary and a lot of meta data about authorship, dates, geolocations and events are contained. They are arranged in a hierarchical topic tree, the *IPTC media topics* with more than 1100 topics in 5 levels[3] and are sorted into 6 custom ressorts and tagged with 134 custom subjects.

With a high degree of certainty, news photos will be provided to us in the future as well. They come as JPEG files zipped with metadata in the same standard *NewsML-G2* as the news items and the same base metadata structure. In addition, they can also contain (extended) captions and information about entities like locations and depicted persons.

Another interesting dataset we probably make use of is the *Common Objects in Context (COCO)* dataset.[4] It contains ~340 000 images with at least 5 english captions each, depicting »complex everyday scenes containing common objects in their natural context« (Lin et al. 2014). It is regulary used as the source data set to evaluate and compare image captioning models.

To collect further contextual details, databases like Wikipedia could be referenced. This could also happen in the form of *word embeddings* trained on the wikipedia corpus.

V. Identification and extraction of relevant contextual information

There are multiple possible approaches for the identification and extraction of relevant contextual information from the sources. We can rely on different methods of *Natural Language Processing (NLP)*.

A. Text Classification

Classification of the text could serve as a basic building block for further processing. It can be seen as a task similar to topic identification. Text classification as a supervised task works pretty well these days. Baseline algorithm implementations like *FastText* (Joulin et al. 2016) are fast and often reach error validation rates below 20% without any fine-tuning. One disadvantage is the need for a large training data set.

Recently, Howard and Ruder (2018) published *Universal Language Model Fine-tuning for Text Classification (ULMFiT)* with very promising results. Using the combined benefits of transfer learning and a universal language model, they need much smaller data sets to provide good results (see figure 7). The language model they used is based on the *WikiText 103* dataset[5], which contains over 100 million tokens extracted from wikipedia.org (Merity et al. 2016).

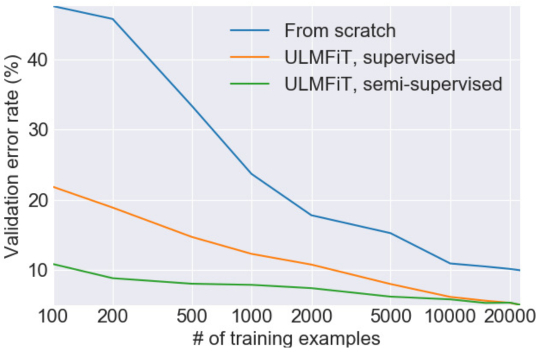


Figure 7: ULMFiT can reach a low error rate even with very few training examples (Howard and Ruder. 2018)

For our task, text classification could provide a basic building block. For example, its results can be used to find more relevant information from other sources by providing neccessary context narrowing the domain space.

B. Clustering and Topic Segmentation

While text classification concerns the whole text, clustering and topic segmentation try to break the text down into different parts. They can detect topics discussed in the text and segment the text according to them.

Clustering tries to find coherence and create groups which are separated from each other. This is a unsupervised task and it may not be clear which conditions led to a result. This is one difference to the supervised task of topic segmentation. Both concepts may reduce the context for further evaluation.

A naive but very cheap approach to extract the relevant part of an journalistic text is to ignore everything but the first few sentences or the lead paragraph. This is based on the experience that in many cases these parts already contain the essential facts of an article.

[3] https://iptc.org/standards/media-topics/

[4] http://cocodataset.org

[5] https://einstein.ai/research/the-wikitext-long-term-dependency-language-modeling-dataset

C. Summarization

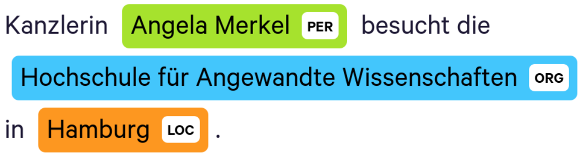
Summarization is in general either abstraction or extraction based. Abstract summarization rephrases the central points of the content in own words, while extraction based summarization creates the condensed output by extracting and reordering words and sentences already in the original text.

For our task, the means for creating a summary are not relevant. Although summarization is a very interesting task, the effort to add extra value through this method seems out of scope of this work. Both approaches are hard problems.

D. Named-Entity Recognition

Named-Entity Recognition (NER) is used to identify relevant named entities in text and map them to predefined categories, e.g. persons, locations or organizations.

An example is shown in figure 8: the tokens in the string *Angela Merkel* are correctly identified as a single person, *Hochschule für Angewandte Wissenschaften* as an organization and *Hamburg* as a location.[6]



Kanzlerin Angela Merkel PER besucht die Hochschule für Angewandte Wissenschaften ORG in Hamburg LOC .

Figure 8: Named-Entity Recognition on a German sentence, created with *displaCy Named Entity Visualizer*.

Which and how many entity labels can be detected depends on the learning corpus. The same is true for the detection rate. The pre-trained models for German by *spaCy*[7] provide four labels for *person*, *organization*, *location* and *miscellaneous* entities like events or nationalities. In contrast, there are models for English available which are based on another corpus and are more fine-grained with seventeen labels, e.g. *product* or *country*.

StanfordNER also has a German NER model available with the same four labels, but it is trained on older data than the English version.[8]

We think named-entity recognition is the most practical and promising approach for detecting relevant context information in the source text, at least to get started.

VI. Conclusion and Outlook

In this paper we saw that there already exist several approaches to inject various extra information into the image captioning process. All of them can provide important insight on how to proceed in regard to transfer this task on the domain of press photos in the context of news texts. Especially the use of the attention mechanism seems to be promising.

There still remain a lot of unresolved questions. Is the provided data set sufficient for this task, despite its size and the wealth of metadata? What distance function works the best? How to connect all the different data sets and models and add additional sources? Most of these questions are of practical nature and are hopefully resolved in practice.

Our next step is to build a pipeline to evaluate different approaches. A first implementation will concentrate on using named entities (see Section V-D), following a simple strategy.

In a first step, we try to detect named entities in the news text. As we saw, pre-trained NER models for German only support four different labels. Learning a mapping between these labels and detected categories in the images and the training captions with consideration of the available metadata should provide acceptable results.

In the next step, we try to replace the mapped categories with their actual content (e.g. the NER label *Person* with *Angela Merkel*) and try to use Attention to inject these content values into the image captions. Again, the metadata provided with the *dpa* dataset makes it possible to use supervised training for this.

Acknowledgment

Thanks to Kai von Luck and the people at the //CSTI for all the support and Gerd Kamps for providing the incredible data set.

[6] Created with https://explosion.ai/demos/displacy-ent

[7] https://spacy.io/api/annotation#ner-wikipedia-scheme

[8] https://nlp.stanford.edu/software/CRF-NER.shtml

References

Howard, Jeremy, and Sebastian Ruder. 2018. "*Fine-tuned Language Models for Text Classification.*" *CoRR* abs/1801.06146.

Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. "*Bag of Tricks for Efficient Text Classification.*" *arXiv:1607.01759.*

Kiddon, Chloé, Luke S. Zettlemoyer, and Yejin Choi. 2016. "*Globally Coherent Text Generation with Neural Checklist Models.*" In *EMNLP*.

Lin, Tsung-Yi, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, et al. 2014. "*Microsoft COCO: Common Objects in Context.*" *CoRR* abs/1405.0312.

Merity, Stephen, Caiming Xiong, James Bradbury, and Richard Socher. 2016. "*Pointer Sentinel Mixture Models.*" *CoRR* abs/1609.07843.

Olah, Chris. 2015. "*Understanding LSTM Networks.*" *https://colah.github.io/posts/2015-08-Understanding-LSTMs/.*

Olah, Chris, and Shan Carter. 2016. "*Attention and Augmented Recurrent Neural Networks.*" *Distill*. doi:10.23915/distill.00001.

Radford, Alec, Rafal Józefowicz, and Ilya Sutskever. 2017. "*Learning to Generate Reviews and Discovering Sentiment.*" *arXiv:1704.01444.*

Tran, Kenneth, Xiaodong He, Lei Zhang, Jian Sun, et al. 2016. "*Rich Image Captioning in the Wild.*" *arXiv:1603.09016.*

Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. "*Show and Tell: A Neural Image Caption Generator.*" *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3156–64.

Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, et al. 2015. "*Show, Attend and Tell: Neural Image Caption Generation with Visual Attention.*" *arXiv:1502.03044.*