ulm university universität
uulm

May 6, 2021

Bachelor Thesis Proposal

Automatic Image Captioning

Johannes Scherer - johannes.scherer@uni-ulm.de

Supervisor: Prof. Dr. Ansgar Scherp, Ulm University
Collaboration: Dr. Deepayan Bhowmik, University of Stirling

1 Notes

As the main focus of the thesis has not yet been determined, this document is not yet in its final form. After a brief introduction to the subject and a compilation of scientific work, own ideas for the possible content of the bachelor thesis are given. The actual content is yet to be determined in joint discussions and this proposal will be reissued afterwards.

2 Motivation

Deep learning is used in computer vision to solve a wide variety of problems. Image classification and object detection may be the first tasks that come to mind. When it comes to describing the content of an image, the usual methods of Computer Vision are not enough to solve this problem. For an automatically created image description the objects contained in an image not only need to be detected and classified, but also a good description must

express how the objects relate to each other [8]. This makes the usage of natural language processing necessary. Therefore, generating natural language descriptions of images and their regions combines two large areas of machine learning: Computer vision and natural language processing. This makes this task a complex and a very interesting topic at the same time.

3 Related Work

The task of **Image captioning** is to *provide a brief and concise description of an image in natural language* [5]. Deep neural network based image captioning methods can be categorized depending on which main framework they use [1]. In the following, three of these subcategories together with the characteristics of their models are introduced. An example for each framework is named and briefly explained.

In **multimodal** based methods first image features are extracted, usually using a Convolutional Neural Network (CNN) [7]. Afterwards, a neural language model maps the image features into the common space with word features and finally generates a description by performing word prediction conditioned on the image features and previously generated words for this description [1]. The approach of Karpathy and Fei-Fei [4] is a multimodal approach, aiming to introduce a model that generates descriptions of image regions based on a dataset of images and related sentences. It consists of two separate models. A Region Convolutional Neural Network (RCNN) [2] is used to detect objects in an image and a Bidirectional Recurrent Neural Network (BRNN) [6] is used to split the image description into sentence segments. A structured objective then maps the two modalities into a common space so that each region feature is described by one textual feature. Afterwards, the so accomplished inter-modal correspondences between language and visual data can be used by a second model, a Multimodal Recurrent Neural Network, to learn to generate descriptions of image regions. Their Multimodal RNN can be used as both, full frame and region-level description generator. Since their approach consists of two separate models, end-to-end training is not possible.

Typically, Recurrent Neural Networks (RNN) are used in machine translation where the task is to transform a sentence written in a source language into its translation in the target language. Here an encoder RNN reads and transforms the source sentence into a fixed-length vector representation which is the input of the decoder RNN that generates the translated sentence. The

image captioning problem can as well be formulated as a translation problem where the input is an image and the output is a sentence describing the image. Image captioning methods using the **encoder-decoder** framework try to perform this translation by first encoding an image into a vector representation which is then decoded to generate a sentence describing the image [1]. Vinyals et al. [8] follow this approach. Replacing the encoder RNN, they use a CNN to encode the images and to obtain their features. Its last hidden layer is then used as the input to the decoder, which is a Long Short-Term Memory (LSTM) [3] Recurrent Neural Network that decodes the obtained image features into sentences.

It is clear that images can be very rich in information they contain while, depending on the formulated task, it is unnecessary to describe all details of a given image. **Attention guided** image captioning follows the idea that for a short and concise description only the most salient regions of the image are relevant and therefore utilizing the attention mechanism may lead to proper image descriptions [1]. Of course, this has the potential drawback of losing information and therefore prevent rich descriptions. Generally, in attention guided image captioning methods an attention mechanism is incorporated into the encoder–decoder framework to make the decoder focusing on certain aspects of the input image when it generates descriptions for the input image. Just like the already mentioned enoder-decoder model of Vinyals et al. the model of Xu et al. [9] use a CNN as encoder and a LSTM network as decoder. In addition to that they incorporate two attention mechanism variants, a stochastic hard and a deterministic soft attention mechanism. An image forwarded through the CNN results in not one, but L feature vectors, each representing a special part of the image. In each time step, the hard stochastic attention mechanism selects a visual feature from one of the L locations as the context vector to generate a word, while the deterministic soft attention mechanism combines visual features from all L locations to obtain the context vector to generate a word.

4 Proposals for research questions

- Introduce, explain and compare the architectures of the models from the different approaches mentioned above. More precisely: Introduce the included neural networks of a model in detail, explain how the model can be trained and which training data is used. Afterwards, evaluate the models based on the results from different scientific work and weight the advantages and disadvantages of the models.

- Creation, implementation and evaluation of an own version of a neural network, based on one of the models mentioned above.
- What are the use cases of automatic image captioning? Are there model architectures that serve well for special use cases?

References

[1] S. Bai and S. An. A survey on automatic image caption generation. *Neurocomputing*, 311, 05 2018.

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation, 2014.

[3] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[4] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. Technical report, Department of Computer Science, Stanford University, 2015.

[5] X. Liu, Q. Xu, and N. Wang. A survey on deep neural network-based image captioning. *The Visual Computer*, 35, 03 2019.

[6] M. Schuster and K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[7] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[8] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555, 2014.

[9] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044, 2015.