Context-based Enriched Image Captioning

Stephan Halbritter

May 8, 2018

Hamburg University of Applied Sciences

Motivation

Research question

Image Captioning

Enriching Captions

Text-Mining For Enrichment

Outlook

Motivation

dpa Data Set

Large

- corpus w/ about 220.000 german news *items*

Well structured

- NewsML-G2, a *»multimedia news exchange format standard«*

High quality

- formal texts, written by professional journalists

Metadata

- 6 custom ressorts, 134 custom subjects and more
- IPTC media topics: 1100 terms, 17 top level terms, 5 levels

Long-term updates

- GraphQL access w/ regular updates (soon)

dpa Data Set - Images

NewsML-G2

- Brandnew as of 23 January, 2018!
- Zipped XML- and JPEG files

Metadata

- 6 custom ressorts, 134 custom subjects and more
- IPTC media topics: 1100 terms, 17 top level terms, 5 levels
- More custom keywords
- Headline, author, caption and extended caption
- Infos about geolocation, event, depicted persons

Example Image with Caption

»Eine Ente schwimmt am 22.11.2017 auf dem Schliersee in Schliersee (Bayern) und spiegelt sich dabei im Wasser.«

https://pixabay.com/en/ducks-waterfowl-mallard-bird-3089530/
(Creative Commons CC0)

Example Image with Caption

»Eine Ente schwimmt am 22.11.2017 auf dem Schliersee in Schliersee (Bayern) und spiegelt sich dabei im Wasser.«

https://pixabay.com/en/ducks-waterfowl-mallard-bird-3089530/
(Creative Commons CC0)

Research question

Research question

How to generate image captions enriched with context-based information from corresponding text?

Image Captioning

Image Captioning

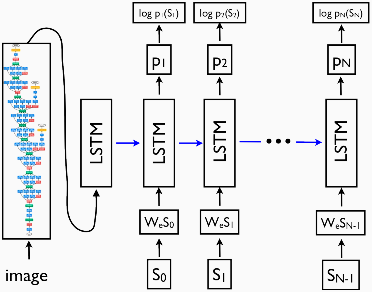
Generating a textual description of an image

- Subtask: Image Classification and Text Generation
- Encoder-Decoder framework
- Supervised Learning

CNN-RNN

Show and Tell: A neural image caption generator (Vinyals et al. 2015)

- CNN for image embedding
- LSTM-based text generating w/ word embedding vectors
- image feature from fully-connect layer of CNN
- static represenation of image is feed into RNN just once

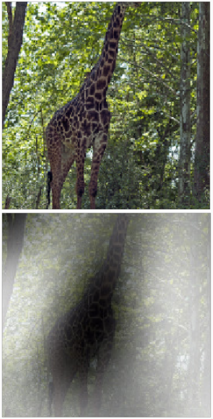


Vinyals et al. (2015)

Image Captioning with Attention 1/2

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention (Xu et al. 2015)

- basically the same CNN-RNN structure as Vinyals et al. (2015)
- attention mechanism uses weighted image features in each step
- extracts multiple features from convolutional layer



A giraffe standing in a forest with trees in the background.

Adapted from Xu et al. (2015)

Image Captioning with Attention 2/2

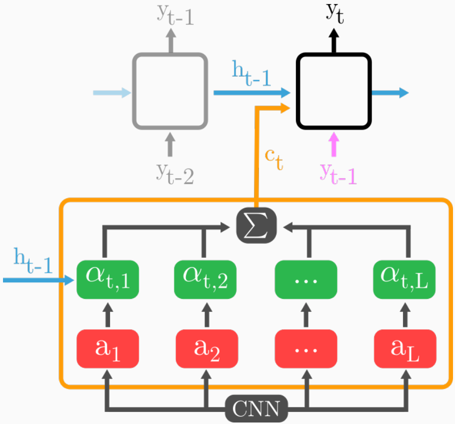
Current word y_t is generated using the previous word y_{t-1}, the hidden state h_{t-1} and the context c_t.

Soft Attention mechanism

$$c_t = \sum_{i=1}^{L} a_{t,i} \cdot \alpha_{t,i}$$

$$\alpha_{t,i} = \text{softmax}(e_{t,i})$$

$$e_{t,i} = tanh(W_h \cdot h_{t-1} + W_a \cdot a_{t,i})$$

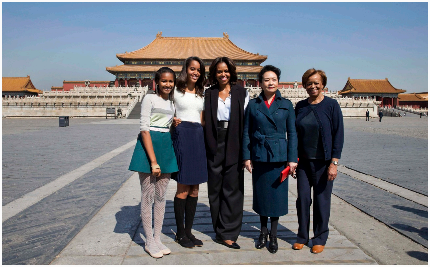


Enriching Captions

Train on More Specific Data

Rich Image Captioning in the Wild (Tran et al. 2016)

- *»data collection and visual model learning are two closely coupled problems«*
- Created large-scale databases of celebrity and landmark images and entity descriptions
- Trained multiple domain-specific CNNs



»Sasha Obama, Malia Obama, Michelle Obama, Peng Liyuan et al. posing for a picture with Forbidden City in the background« (Tran et a. 2016)

Direct and Indirect Injection

Image Captioning at Will: A Versatile Scheme for Effectively Injecting Sentiments into Image Descriptions (You et al. 2018)

- Sentiment Unit in RNN (Radford et al. 2017)
- Direct injection concats weighted sentiment value with the current value in the input gate
- Indirect injection uses *sentiment cells* which is interlinked with the LSTM memory cells

Using a Checklist of Words to Use

Globally Coherent Text Generation with Neural Checklist Models (Kiddon et al. 2016)

- keeps a checklist of words that have to be mentioned in the final text
- predicts in each step if a checklisted word is relevant
- uses this probability during generation

Text-Mining For Enrichment

Text-Mining For Enrichment

Classification

- Works pretty well, e.g. FastText (Joulin et al. 2016)
- Basic building block

Clustering and Topic Segmentation

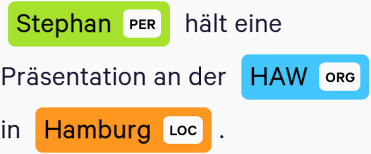
- Idea: restrict text to parts relevant for captioning

Summarization

- Abstraction- or extraction-based
- Hard problem

Named-Entity Recognition

- Number of entity labels depends on the learning corpus
- pre-trained models by spaCy and StanfordNER provide 4 labels PER, LOC, ORG, MISC
- spaCy achieves a F-score for German of 82.85



Stephan PER hält eine Präsentation an der HAW ORG in Hamburg LOC .

Visualized with *displaCy Named Entity Visualizer*

Outlook

Outlook

- Build pipeline for multiple models and different data sets, word embeddings, …
- Interplay of image - caption - text

References i

Joulin, Armand, Edouard Grave, et al. 2016. "*Bag of Tricks for Efficient Text Classification.*" *arXiv:1607.01759*. http://arxiv.org/abs/1607.01759.

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

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

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

Vinyals, Oriol, Alexander Toshev, et al. 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, et al. 2015. "*Show, Attend and Tell: Neural Image Caption Generation with Visual Attention.*" *arXiv:1502.03044*.

You, Quanzeng, Hailin Jin, and Jiebo Luo. 2018. "*Image Captioning at Will: A Versatile Scheme for Effectively Injecting Sentiments into Image Descriptions.*" *arXiv:1801.10121*.

Any questions or answers?