Software Project

Tatjana Chernenko, Siting Liang, Bente Nittka, Kim Almasan
Prof. Dr. Anette Frank
Institute for Computational Linguistics of University Heidelberg
Im Neuenheimer Feld 325, 69120 Heidelberg, Germany

1 Problem Definition

The recent significant improvements in generating natural language descriptions of images and easily available image-sentence data sets make the computer vision and natural language processing communities to engage in inventing more efficient joint models, which require both computer vision and natural language processing techniques[2]. Automatic image description generation is a problem which involves vision recognition in order to detect objects, substances, and locations. As the textual descriptions from data sets usually supply the spatial relations between the objects and provide the attributes for describing them, we aim to explore the impact of the features extracted from the textual descriptions and attributes and to create a model that ultimately generates a textual description that verbalizes the detected aspects of the image.

2 Approach and Implementation

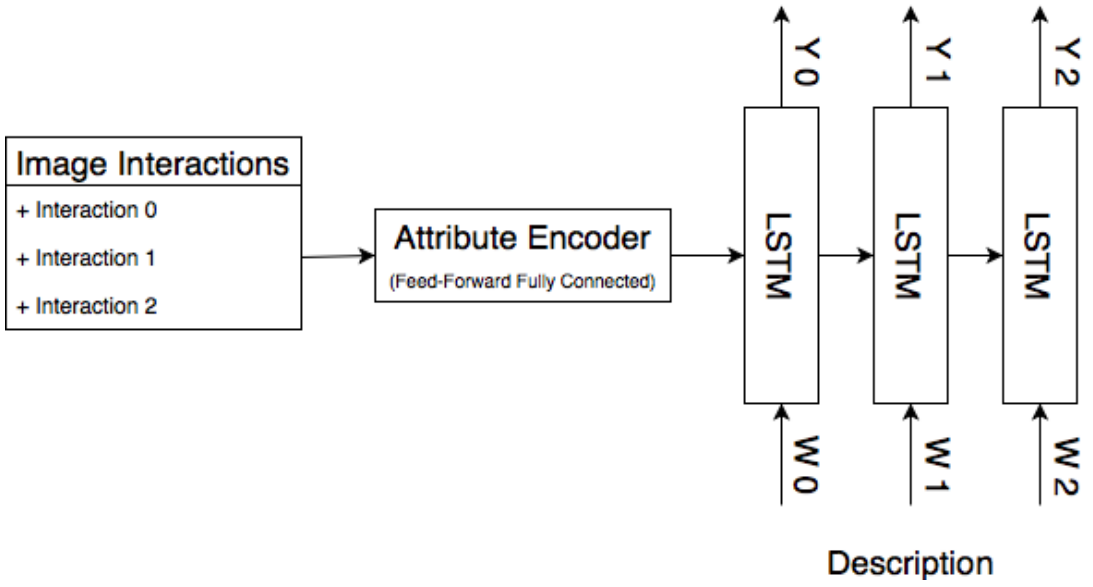

Figure 1: Model Diagram

2.1 General Approach and Architecture

This project aims to generate descriptions of images focusing on spatial relationships between the objects and sufficient attributes for the objects. We consider only the images with 2-5 objects in one image (see Dataset for more details). As an orientation for our architecture we take Dong et al. (2017) [1] who generate product reviews from attributes, a task that is similar in structure to ours.

The whole system has an encoder-decoder architecture. The feed-forward neural network encoder reads the normalized vector representations of all the attributes as input. The model transforms this into a rich fixed-length vector representation which in turn is used as the initial hidden state of a decoder that consists of several LSTMs. LSTMs take the words from the sequences from the description sentences as input and generate the target sentence.By inference we are going to evaluate the performance of the usage of the BeamSearch and Sampling methods.

2.2 System Development

The development of the system starts with exploring the impact of different input attributes on the quality of the generated descriptions. We consider the overlap of the V-COCO, MS COCO and COCO-a data sets (see Data set for more information) to get the attributes and image descriptions.
The first simplified version of the system takes only the information available in the MS COCO data set itself, that is objects and descriptions for an image. We expect that from such little information the model will learn descriptions that are common rather than appropriate descriptions for the situation in the image.
The second version includes attributes from the V-COCO data set (action and various semantic roles for this action (the object, the instrument, the agent)) and image descriptions from MS COCO for every considered image as input information.
The next variants of the system utilize additional attributes from the COCO-a data set. Finally, to make our approach multimodal, we could include in our model direct information about the images in form of image vectors, or try to produce our own attributes with an image recognition system.
We observe the performance of the system at every development step, using a set of automatic evaluation metrics (see Evaluation for more information). Human evaluations at every development step help to compare possible advantages and errors in the output.

2.3 Special problems

2.3.1 Encoding the attributes:

The attributes from the COCO-a data set have the structure of a directed graph. There are several interaction groups for every image that have involved objects and attributes as subsections. The first interesting question is to find a good representation of these input attributes that are then taken as inputs by the encoder. We want to try a multidimensional structure of the input that saves the information about the hierarchical structure of the interaction groups.

The second question is the architecture of the encoder. We start with the architecture described in the subsection 2.1, using the feed-forward neural network.

The next question is the converting of the encoder output into a suitable for the decoder form. We will try the concatenation of the vectors.

If the performance will not be suitable, we can experiment with different architectures, using the one-dimensional representation of the input, BiRNN as the encoder and averaging the output vectors.

Subjects	Objects	Interactions	Actions	Adverbs
2.2	5.2	5.8	11.1	9.6

[hbt]

Table 1: Per Image statistics COCO-a

2.3.2 The limitations of the given descriptions from the MS COCO data set:

The provided description sentences usually don't cover all the interactions from the image. For example, the MS COCO descriptions of the example image (see subsection 3 for more details) often provide only the descriptions for one subject, ignoring other subjects in the image and the interactions between them. Since we are going to focus on the spacial relationships in the image, the MS COCO descriptions are not always the ideal training data for the task if we want to get more descriptions for more than one interactions. We want to explore, how can we combine the information from the attributes, how much can the attributes input contribute to the decoder and succeed to generate a new description for another subject, which doesn't have any descriptions in MS COCO data set.

3 Data sets

We base our work on the **MS COCO** data set [14], a big data set commonly used for the task of automatic image captioning. The intention of the data set is to depict objects in their natural context rather than in iconic images. The whole COCO data set contains 328k images with objects labeled as one of 91 object types. In average the data set has 3.5 categories and 7.7 object instances per image. Most importantly for our task, for each image the data set provides five different single-sentence descriptions that were produced by Amazon Mechanical Turk. Apart from the fact that it is a big data set that has previously been used for description generation, MS COCO has the advantage that there exist several external extensions of the data set that offer further useful information on the images.
We have chosen two of those in particular that we want to use in our work: V-COCO [15] and COCO-a [16]. **V-COCO** contains 10346 images from the MS COCO data set. It has been created for the task of visual semantic role labeling and thus adds action labels for each instance of a person, along with the semantic roles associated with the action and the objects that fill those roles, to the information from MS COCO.
In total, the data set contains 16199 people instances labeled with 26 different actions.
The second data set we want to include in our work is **COCO-a**. The data set focuses on human actions and interactions and wants to improve their recognition and understanding by providing a wide range of annotations. The annotations in COCO-a include the actions performed by each person in the data set, people and objects involved in the action, as well as interesting additional information like subject's posture and emotion, or prepositional attributes like mutual position and distance. The data set contains 10000 images and 140 actions that were obtained by analysis of VerbNet and from the MS COCO descriptions. See table 1 for more statistics.

We think that in combination, the three datasets are very useful for our task because together they contain a wide range of attributes that we can incrementally add to our model. Thus, they probably allow us to gain interesting insight about the change in quality of our generated descriptions, especially with a focus on interactions between humans and other objects and their spatial relations.
Links:

MS COCO
V-COCO
COCO-a
Let's consider an example image (id=360452) from the data set.

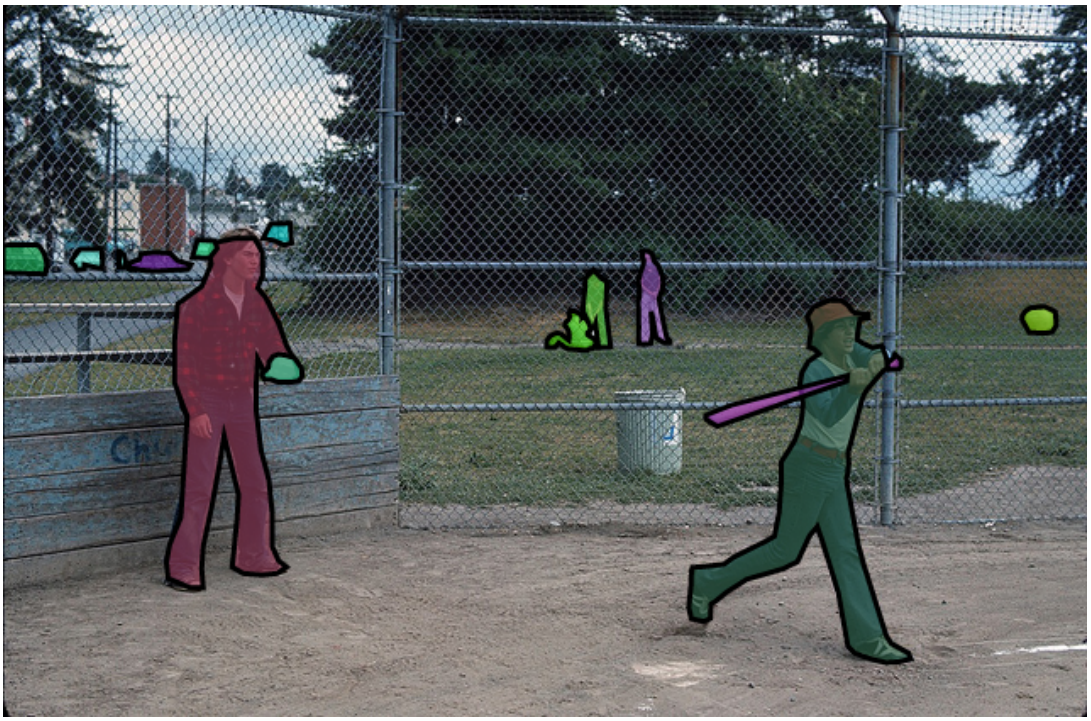

Figure 2: Example Picture

MS COCO provides the following information for this image:

- 5 objects (person, car, sports ball, baseball bat, baseball glove)

- 5 descriptions:

    1. child swinging bat playing baseball with adult watching and fence
    2. a young man holding a baseball bat in front of a ball.
    3. a boy hitting a ball with his bat on the baseball field.
    4. a young person swinging a baseball bat at a ball.
    5. a young boy takes a swing at a ball with his baseball bat.

COCO-a contains additional attributes for this image. It uses all the subjects and objects from the image and divides them into several interaction groups. Each interaction contains a subject, an object and some descriptive words (additional attributes: visual actions and visual attributes).

Visual adverbs are divided into three subcategories:

- location
- distance
- emotion

Visual actions have 8 subcategories:

contact	social	posture
perception	nutrition	communication
solo		objects

The above mentioned example image contains 5 interactions. Lets have a look at the first two interactions in tables 2 and 3:

Table 2: First Interaction

First Interaction	
Subject category:	person
Object category:	baseball bat
Visual actions:	hold, touch, wear
Visual adverbs:	full contact, in front

Table 3: Second Interaction

second Interaction	
Subject category:	person
Object category:	baseball glove
Visual actions:	hold, touch, wear
Visual adverbs:	full contact, in front

V_COCO data set provides triple attributes: object, subject (instrument) and action between these two elements.

If the model utilizes just the information from the MS COCO, then we have just five objects without any relations between them. It could be difficult to generate a sensible description. The first variant of our model will try to generate the descriptions, using the attributes from the V_COCO data set: object, subject (instrument) and an action between them. We expect that the model generates simple sentences, failing to capture the spacial relationship, emotions, distances between the objects, etc. The possible generated descriptions, containing this information, learned only from the descriptions, are expected to be wrong, because the relations on every image are always unique and need to be detected from the image (or its attributes).

Then we are going to explore the influence of the additional attributes from the COCO-a data set, focusing on the spacial relationships, using two subcategories of the visual adverbs (location and distance). This information is expected to generate more sensible sentences.

4 Evaluation

We compare the end version of the system against the first development step as a baseline. In order to monitor the performance of the system on the trial data, we use a set of evaluation metrics at every development step. As no metric produce a correlation with human ratings [17], we use a mix of word-based and grammar-based metrics. The word-based group of evaluation

methods is represented by BLEU[18], METEOR [19] and CIDEr [20] scores, which measure the word-overlap. Semantic Similarity (SIM), based on distributional similarity and Latent Semantic Analysis and complemented with semantic relations extracted from WordNet, is another considered representative of the word-based metrics. The above mentioned scores are expected to show better correlations to human ratings of informativeness[17]. The grammar-based group is presented by Readability Flesch Reading Ease score (RE) and must show better correlations to quality and naturalness[17].
At every development step we also use human evaluation on a 6-point Likert scale for informativeness, naturalness and quality.

5 Organization

5.1 GitLab

We use GitLab as our main tool to organize, to parallel, and version our project. GitLab offers us a wide range of collaborative features, such as, Issue Managemnet with Milestones, Deadlines, Kanban Board, and so much more.

5.2 ShareLatex

In order to coordinate and facilitate the writing of latex documents for this project, we use ShareLatex, which offers an easy method to edit and maintain latex documents for multiple participants.

5.3 Possible Frameworks

Tensorflow is the most used machine learning library and offers a very active community and degree of customizability.

PlaidML is an alternative to Tensorflow which mostly functions as a backend for Keras. The advantage of PlaidML is that it supports GPUs with only OpenCL support, which makes it possible for us to use a GPU rig that we have access to.

Keras is only a layer on the top of popular libraries like Tensorflow and makes network/model development much less verbose. Furthermore, it makes it easier to exchange the "backend" library very easy.

NLTK will be mostly used as evaluation method to compare our results with the results of other trained models.

References

[1] Li Dong and Shaohan Huang and Furu Wei and Maria Lapata and Ming Zhou and Ke Xu:Learning to Generate Product Reviews from Attributes, Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, 2017.

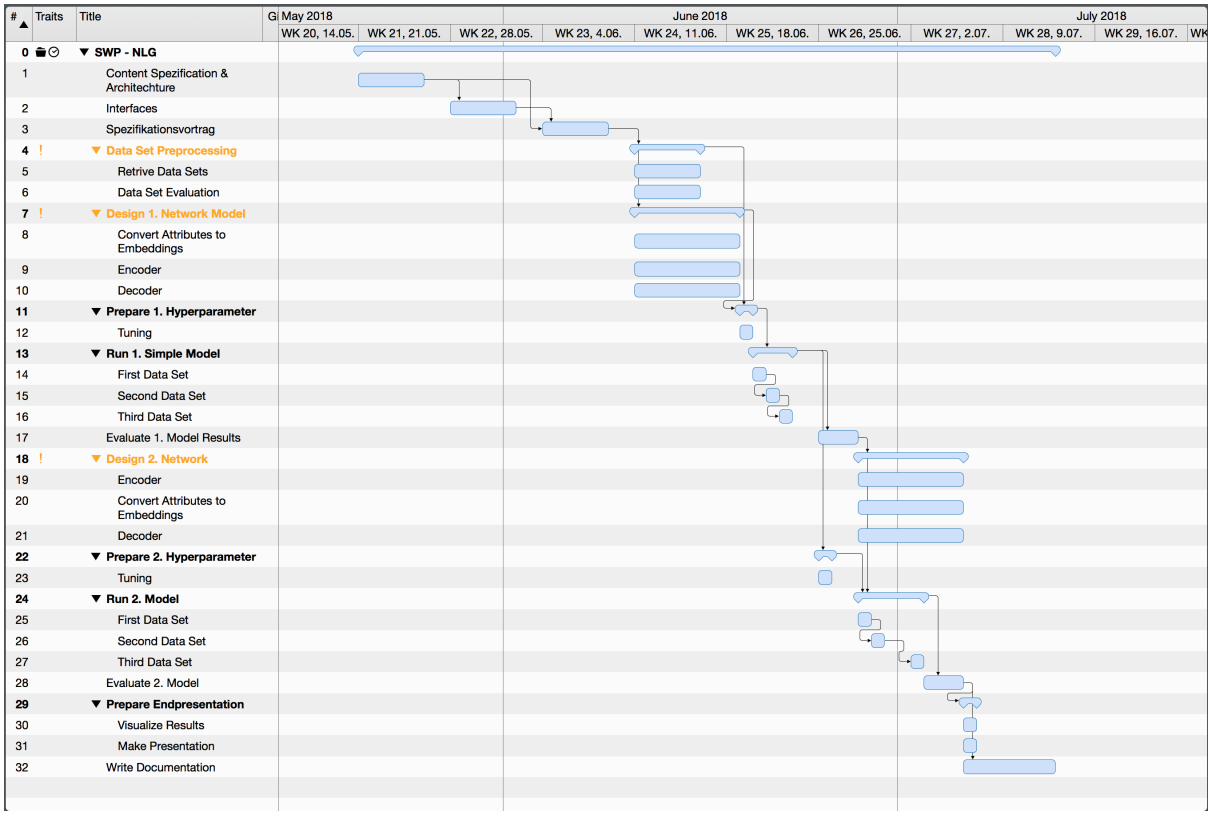

Figure 3: Gantt Diagramm

[2] Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, Barbara Plank: Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures. CoRR,abs/1601.03896, 2016.

[3] Yatskar, Mark and Zettlemoyer, Luke and Farhadi, Ali: Situation Recognition: Visual Semantic Role Labeling for Image Understanding. pages 5534-5542, CVPR, 2016.

[4] Mark Yatskar, Vicente Ordonez, Luke Zettlemoyer, Ali Farhadi: Commonly Uncommon: Semantic Sparsity in Situation Recognition.Conference on Computer Vision and Pattern Recognition.arXiv:1612.00901v1, 2016.

[5] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, Noam Shazeer(Google Search): Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks. CoRR, abs/1506.03099, 2015.

[6] Spandana Gella,Frank Keller: An Analysis of Action Recognition Datasets for Language and Vision Tasks. CoRR, abs/1704.07129, 2017.

[7] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan: Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge. CoRR,abs/1609.06647, 2016.

[8] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. pages 529-545, ECCV, 2014.

[9] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping.CoRR,abs/1406.5679, 2014.

[10] A. Karpathy, L. Fei-Fei: Deep Visual-Semantic Alignments for Generating Image Descriptions. IEEE Conference on Computer Vision and Pattern Recognition. pages 664 - 676, CVPR, 2015.

[11] Girish Kulkarni Visruth Premraj Sagnik Dhar Siming Li Yejin Choi Alexander C Berg Tamara L Berg: Baby Talk: Understanding and Generating Image Descriptions. pages 1601 - 1608, CVPR, 2011.

[12] Laura Perez-Beltrachini,Claire Gardent: Analysing Data-To-Text Generation Benchmarks. CoRR, abs/1705.03802, 2017.

[13] Marc Tanti, Albert Gatt, Kenneth P. Camilleri: Where to put the Image in an Image Caption Generator. CoRR, abs/1703.09137, 2017.

[14] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár and C. Lawrence Zitnick: Microsoft COCO: Common Objects in Context .CoRR, abs/1405.0312, 2014.

[15] Gupta, Saurabh and Malik, Jitendra: Visual Semantic Role Labeling. arXiv:1505.04474, 2015.

[16] Matteo Ruggero Ronchi and Pietro Perona. Describing common human visual actions inimages. In Mark W. Jones Xianghua Xie and Gary K. L. Tam, editors,Proceedings of theBritish Machine Vision Conference (BMVC 2015), pages 52.1–52.12. BMVA Press, Septem-ber 2015.

[17] Ekaterina Novikova, Ondrej Dusek, Amanda Cercas Curry, Verena Rieser: Why We Need New Evaluation Metrics for NLG. CoRR, abs/1707.06875, 2017.

[18] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu: BLEU: a Method for Automatic Evaluation of Machine Translation. ACL, pages 311-318, 2002.

[19] Alon Lavie,Abhaya Agarwal: Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. StatMT, pages 228-231 , 2007.

[20] Ramakrishna Vedantam, Virginia Tech, C. Lawrence Zitnick, Microsoft Research, Devi Parikh, Virgnia Tech: CIDEr: Consensus-based Image Description Evaluation.CoRR, abs/1411.5726, 2014.