Introducing Concept And Syntax Transition Networks for Image Captioning

Philipp Blandfort
University of Kaiserslautern
Kaiserslautern, Germany.
philipp.blandfort@dfki.de

Tushar Karayil
University of Kaiserslautern
Kaiserslautern, Germany.
tushar.karayil@dfki.de

Damian Borth
German Research Center for
Artificial Intelligence (DFKI)
Kaiserslautern, Germany.
damian.borth@dfki.de

Andreas Dengel
University of Kaiserslautern
German Research Center for
Artificial Intelligence (DFKI)
Kaiserslautern, Germany.
andreas.dengel@dfki.de

ABSTRACT

The area of image captioning i.e. the automatic generation of short textual descriptions of images has experienced much progress recently. However, image captioning approaches often only focus on describing the content of the image without any emotional or sentimental dimension which is common in human captions. This paper presents an approach for image captioning designed specifically to incorporate emotions and feelings into the caption generation process. The presented approach consists of a Deep Convolutional Neural Network (CNN) for detecting Adjective Noun Pairs in the image and a novel graphical network architecture called "Concept And Syntax Transition (CAST)" network for generating sentences from these detected concepts.

Keywords

Auto Caption, Image Captioning

1. INTRODUCTION

With its exponential growth in the last two decades, the Internet has become a major source of information exchange across the world. Powered by new technologies and increased computational resources, web sites have become more visual and animated. Moreover, the world wide web is flooded with new images everyday, e.g. Instagram has reported an average of 80 million photo uploads a day.[1] Most of these images come with titles which can be generic (e.g. IMG_123), descriptive or give additional information that can not be directly seen in the image.

Recently, there have been significant advances in generating descriptive image captions ([7],[8],[9]) but so far, the focus was on generating factual descriptive image captions like Microsoft COCO [3]. These datasets provide a rich textual description of images. However, these descriptions might not be representative for natural image captioning since human subjects were given clear instructions on writing the captions [3]. Such descriptions, although informative, constitute only a small subset of the different styles of captioning.

In contrast, large real-world datasets such as "Yahoo Flickr Creative Commom 100 Million" (YFCC100M [12]) displays a huge variety of captioning styles but these titles can only be considered to be weak labels (cf. [13]): Here, captions can be descriptive, emotional or mention information that is not visible in the image.

Humans often tend to associate a sentiment with an image and express that in the caption. One such method is by using an emoji.[2] A richer use of text would be another way to express the associated sentiment with the caption.

It was shown by [1] that adjectives can add an emotional component to nouns and the resulting Adjective Noun Pairs can express the visual contents in the image. Hence we assume that incorporating adjectives into machine generated captions is one feasible way of adding an emotional component to the caption. To this end, we describe a model that is capable of generating subjective image captions and thereby going beyond factual image descriptions. We train and test our model on the YFCC100M database.

2. RELATED WORK

The available methods in linking images to text can be classified broadly into three categories.

The first set of methods is used to detect a triplet (eg. $<object, action, scene>$) of scene elements in the image and convert them into sentences. Triplets provide a holistic idea of what is most important in the image and they are combined using various techniques to generate captions. [4], [8]

[1] https://www.instagram.com/press/

[2] http://instagram-engineering.tumblr.com/post/117889701472/emojineering-part-1-machine-learning-for-emoji

ICMR'16, June 06 - 09, 2016, New York, NY, USA

ISBN 978-1-4503-4359-6/16/06...15.00

DOI: http://dx.doi.org/10.1145/2911996.2930060

use this approach and a template based system to generate the sentences after identifying the objects.

The second set of methods bring the images and sentences into a single multi dimensional space by converting each of them into vectors. Thereafter a set of distance measures are used to find the closest matching description of a given image ([5], [11]). [11] uses neural networks to map images and sentences into the same vector space. Although the above mentioned methods have shown promising results they cannot be used for generating novel descriptions. Hence the performance of these methods drop when there are new compositions of objects in a given image (even though individual objects have been observed during training).

The third set of methods which have shown the most promising results use a combination of Deep Convolutional Neural Networks (DCNN) for feature extraction and a Multimodal Recurrent Neural Networks (RNN) on top of it for text generation from the extracted features [14].

However, despite the promising results, approaches of the third kind suffer from several shortcomings:

- **Robustness**: RNNs heavily rely on suitable ground truth information. We argue that training an RNN on the YFCC100M would require expensive preprocessing since there are too many images with titles that can not be learned with current methods (e.g. because the relation to the visual content is not straight-forward) and could therefore disturb the training.

- **Transparency**: It is hardly possible to interpret what exactly is happening inside the RNN. The problem we see with this lack of transparency is that it forces you to treat the whole sentence generation part as single task that can not readily be broken down into distinct parts. This makes it hard to gain new insights about human language processing from the performance of the system or to incorporate new insights into the model. As a result, the performance depends heavily on the training data and modifications often have to be done by trial and error.

We apply *DeepSentiBank* as fixed visual concept extraction and generate captions from the features with the novel CAST network architecture instead of RNN in order to address the aforementioned shortcomings. This combination is quite robust and you can follow and influence all the steps from detected concepts to final sentence, giving you much more control and allowing for easier future changes of the architecture.

3. PRESENTED APPROACH

The presented system follows a pipeline approach consisting of the following steps:

1. **Visual concept extraction**: We process the image with *DeepSentiBank* to extract concepts including emotional cues.

2. **CAST network**: Generate ranked sentences from the detected concepts.

3. **Templates**: If the rankings from the network are below a threshold, we use a template-based approach to create sentences.

3.1 Data Preparation

The YFCC100M dataset contains user captioned Flickr images. Users' captions/tags often do not provide the appropriate data required to train classifiers and generate graphical language models. For example, the images often contain camera generated captions, generic titles, single word captions, locations as reported in [6]. Therefore it was important to extract the relevant Image-Caption pairs useful for model training.

We filter and remove all images with titles that are generic (e.g. IMG_1234.jpg), have less than 2 words or contain one or more non-English words.

After applying the filter, we end up with a training set consisting of 9.6 million images and a validation set consisting of 1.2 million images where the split into train and validation set is based on the user identifier present in the image meta-data. The cleaned up set still contains captions that are not directly related to the image, often providing extra information. But as this is natural in human image captioning, we deliberately keep this kind of data.

We build our graph (including word2vec model) on this data only, showing that our method works well with noisy real-world data without any sophisticated preprocessing.[3]

3.2 Visual Concept Extraction

To generate human like caption, the first step is to capture the emotional and visual contents from the image. The work published in [1], introduced Adjective Noun Pair (ANP) concepts able to describe images beyond visual content (e.g. "dog") by capturing positive or negative polarity (e.g. "cute dog" or "scary dog"). The resulting set of ANPs as trained by a deep convolution neural network is called *DeepSentiBank* and was published by [2]. This pairing of adjectives and nouns does also provide an insight into the general emotion associated with an analyzed image.

Processing the image with *DeepSentiBank* gives us a feature vector where each element corresponds to one ANP from a list of 2089 ANPs.

These 2089 ANPs contain 231 distinct adjectives and 424 nouns. This size is quite small when we want to generate different styles of sentences. To increase the size of vocabulary we generated a word2vec [10] model from all the training titles. (See next point.)

3.3 Concept And Syntax Transition (CAST) Network

The CAST network is a multi-directed graph where each node in the network represents a concept (i.e. noun, adjective, verb or adverb) connected to other concepts. It is generated in the following steps:

1. **Nodes**: For each content word with occurrence count greater than 40 in the training titles, we create a node. This leads to a vocabulary of over 21000 words.

   Additionally, we add a START and an END node.

2. **Similarity edges**: We train a word2vec model on all training sentences. word2vec maps words to a vector space of a given dimension (200 in our case) where words that are used in similar contexts are mapped to vectors that are close together in the target space.

[3]In principle it would be possible to train the visual concept detector on the YFCC100M data as well.

Under the assumption that semantically similar words are used similarly, this makes it possible to compute semantic distances between words.

For each node we add similarity edges to all nodes that have a word2vec similarity above some fixed threshold. This accounts for the possibility of replacing words by semantically similar words in the sentence generation process.

3. **Syntax edge**: For each sentence in the training titles we check how content words are connected. For each such connection that does not use another content word, a directed edge is created between the corresponding concept nodes. The connecting string (usually consisting of propositions, articles, etc.) is used as edge label and the total number of connection occurrences is annotated as edge weight.

   These edges contain information about the syntax of the language and are used to connect different concepts.

The whole graph generation was done in less than 40h and the model can be used without any further training.

In CAST networks generating a sentence from a set of concepts is reduced to the problem of finding a path from the start to the end node through a set of activated nodes. Computing a list of such paths is done in a heuristic way and this list is then ranked by considering the weights of the included edges. The path with the highest score is then converted to a sentence in a straight-forward way, where similarity edges are used to substitute words. An illustration of a simple CAST network can be found in Figure 1.

In general, CAST networks provide a new possibility for the challenging task of generating sentences from an arbitrary sets of words. By using word similarity in the sentence generation process, they display a high degree of creativity, effectively extending the vocabulary. They do all this in a simple and transparent way which makes it easy to find the source of mistakes, allowing for systematic improvements or customization of the system in the future.

3.4 Template-based Approach

The idea of this approach is to use different templates of the kind "HUMAN with PROPERTY doing VERB on EVENT in LOCATION" to form sentences from a set of visual concepts that have been tagged by according category and are detected in the image.

For this we need:

- **Category tags**: We manually assigned category tags (e.g. "HUMAN" or "LOCATION") to all nouns that occur in any ANPs.
- **Templates**: A few (5) templates based on these category tags were created manually. From that we automatically generated different template variations by removing parts of the template. (E.g. the variations of the above template would include "HUMAN doing VERB in LOCATION" and "HUMAN with PROPERTY on EVENT in LOCATION".)

Sentences are now generated in the following steps:

1. **Input**: Given ANP scores from *DeepSentiBank*, we consider all ANPs that have a score above a fixed threshold to create sentences from all suitable template variations. (If no score exceeds the threshold we take the ANP with highest confidence and return it as caption.)
2. **Ranking**: We rank all resulting sentences based on a scalar rating score that is computed for each sentence individually, using for the computation the *DeepSentiBank* scores of all ANPs that are present in the sentence.
3. **Output**: The sentence with the highest score is given as caption.

4. RESULTS

In order to evaluate the *humanness* factor of the generated caption, we selected 200 random images from our test set. These images were assigned two captions: The original caption present in the YFCC100M dataset and the caption generated by our method. Without informing the individual about the source of the two captions, we asked human subjects to choose one among the two captions which they thought were generated by a human. To compensate for the subjective bias in human evaluation, each image was shown to three different individuals and the opinion of the majority was decided as the final result for that image.

We report that 31.5% of the captions generated by our method were reported as more human-like in comparison to the original caption by at least two subjects. In 62.5% of images at least one subject chose our caption over the original one. These results are encouraging. The generated captions often read naturally and convey emotions. The creativity and subjectivity that is displayed in some of the captions is very entertaining. Figure 2 shows a small selection of titles generated by our method.

5. CONCLUSIONS

We presented an approach of combining the top Adjective Noun Pairs detected in an image with a graphical model to form captions that are not only descriptive but also carry subjective meaning. A human evaluation of our method on a subset of the YFCC100M dataset we often obtain natural image captions.

To improve the existing model and to get the caption quality closer to human levels we are planning to extend our work by incorporating the following points: The grammar of the whole sentence needs to be given more weight. We are currently working on an additional ranking mechanism to take that into account.

Also, so far the confidences of the detector are only respected in the thresholding and then discarded. We plan to either modify the network traversing algorithm such that it also respects the concept scores or respect the scores in the final ranking of the proposed sentences. We also want to use additional concept detectors to get more different sentences from the network and optimize the whole network on more data.

6. ACKNOWLEDGMENTS

This work was partially funded by the BMBF project Multimedia Opinion Mining (MOM: 01WI15002).

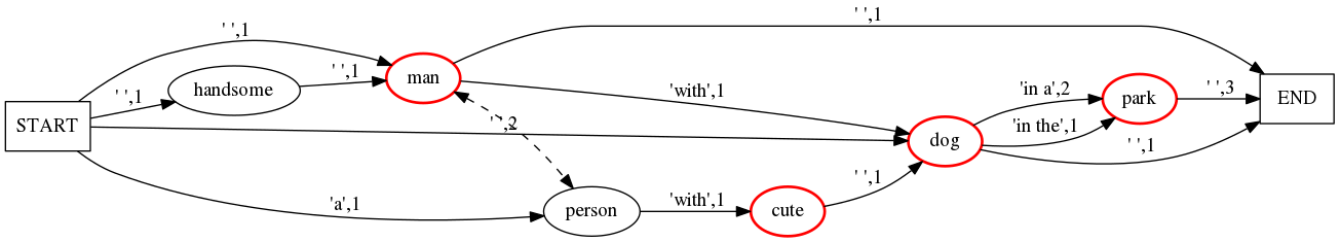

Figure 1: Example of a CAST network generated from the titles "handsome man", "a person with cute dog", "dog in a park", "dog in the park" and "man with dog in a park". The dashed line indicates a similarity edge (and is in this toy example not generated from the given sentences). If the red nodes denote the activated concepts, the resulting sentence would be "person with cute dog in a park". (Substituting "man" by "person" because a similarity edge was traversed.)

Figure 2: Qualitative results of our approach for images of the YFCC100M dataset. The captions in black are the ground truth titles, in blue we have captions produced by the combination of *DeepSentiBank* and CAST. The "X" marks indicate which caption the majority of people in our evaluation experiment believed to be created by a human.

7. REFERENCES

[1] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale Visual Sentiment Ontology and Detectors Using Adjective Noun Pairs. In *ACM Int. Conf. on Multimedia (ACM MM)*, 2013.

[2] T. Chen, D. Borth, T. Darrell, and S.-F. Chang. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586*, 2014.

[3] X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv:1504.00325*, 2015.

[4] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *Computer Vision–ECCV 2010*, pages 15–29. Springer, 2010.

[5] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, pages 853–899, 2013.

[6] S. Kalkowski, C. Schulze, A. Dengel, and D. Borth. Real-time analysis and visualization of the yfcc100m dataset. In *MM COMMOMS Workshop*, 2015.

[7] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015.

[8] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 220–228. ACL, 2011.

[9] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014.

[10] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546, 2013.

[11] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the ACL*, 2:207–218, 2014.

[12] B. Thomee, B. Elizalde, D. A. Shamma, K. Ni, G. Friedland, D. Poland, D. Borth, and L.-J. Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[13] A. Ulges, D. Borth, and T. M. Breuel. Visual concept learning from weakly labeled web videos. In *Video Search and Mining*, pages 203–232. Springer, 2010.

[14] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555*, 2014.