# Deep Learning for Image Caption Generation

Efsun SEZER
Department of Computer Engineering
Hacettepe University
efsunsezer@cs.hacettepe.edu.tr

Derman AKGÖL
Department of Computer Engineering
Hacettepe University
dermanakgol@gmail.com


**Figure 1:** *Image captions*

## Abstract

Image captioning refers to the generation of a sentence that describes what is shown in the image such as current objects and their properties, the actions performed and interaction between objects. Since it is a task that resides at the intersection of Computer Vision and Natural Language Processing (NLP), the problem setting requires both an understanding of the pixel context in the images that is, the analysis of which features represent which objects and the creation of a semantic structure linked to these objects.

In this paper, we investigate the image captioning problem with the use of deep learning based techniques. For that purpose, we are going to employ a vision based CNN and RNN architecture that was previously described in the literature. Another goal of this project is that trying to improve existing approach with the use of attention module or moving to the better deep models usage.

**Keywords:** image captioning, attention mechanism, word embedding, word representation, neural network language model, deep learning.

## 1 Introduction

Image captioning is the process of creating a textual definition of an image by using both NLP and Computer Vision algorithms. It is an important area of research used in many fields. For example, it improves access to desired information by allowing us to sort and request images or image-based content in new ways. It can also be used to obtain real-time annotations that define the environment in order to improve the quality of life of visually impaired people.

Image captioning is performed by creating networks that can perceive contextual subtleties in images, associating observations with both the scene and the real world and extracting concise and accurate textual image definitions.

The structure of this paper is as follows: Recently proposed approaches are discussed in Section 2. In Section 3, a method is expressed. In Section 4, datasets and evaluation metrics are described. In Section 5, a schedule for the course project is presented. Finally, the conclusion is given in Section 6.

## 2 Related Works

The research for image captioning has existed for several years, but the effectiveness of the techniques was limited and they were generally not strong enough to cope with the real world. But with the emergence of AI breakthrough, the amazing performances are achieved by the use of deep learning-based methods in image captioning [Chen and Zitnick 2014] [Bahdanau et al. 2014].
For example, [Vinyals et al. 2015] proposed a method which is based on a vision based CNN and LSTM models. In their work, a vision based CNN is utilized to represent images and LSTM is used to generate textual description of images using CNN features. In another work, [Xu et al. 2015] introduce attention mechanism in the concept of image captioning to improve model performance. Their attention module has two variants: a hard attention mechanism and a soft attention mechanism.

## 3 Method

As mentioned before, image captioning is examined in two main branches. In the first step, computer vision techniques are used to extract image features that describe the image structure. After that, these features are fed into the language based model which translates these features and objects given by image based model to a natural sentence.

In this project, a CNN (Convolutional Neural Network) based model is going to be used to extract image features and RNN (Recurrent Neural Network) or its variants such as LSTM (Long-Short Term Memory), GRU (Gated Recurrent Units) will be used for the sentence generation part.
As a starting point, the CNN + RNN combination proposed in the literature will be implemented and results are going to be observed.

In addition to that, in the light of the research on the image captioning, a few ideas such as attention models usage and effective trial of other deep models combination that can help us to get a better results will be examined.
The main motivation to choose attention models is that it could help us in fine tuning our model performance. The other option is moving on to bigger and better techniques that is an effective and advanced trial of CNN or RNN models that recently have been proposed by the researchers.

## 4 Experimental Settings

### 4.1 Datasets

In this section, we explore the image captioning datasets which are listed below:

1. **Flickr8k [Hodosh et al. 2013] :** It consists of animals and human images. There are 8000 images in this dataset. While describing images, it has been used five sentences for each of them.

2. **Flickr30k [Young et al. 2014] :** This dataset is an extension of the previous dataset with 31738 images. However, this dataset consists of daily activities or events of human. Similarly, for each image, it has been used five sentences to describe the image.

3. **MSCOCO [Lin et al. 2014] :** It contains 123287 images which consist of daily scenes. The objects in images are labelled and like previous two datasets, five sentences has been used for each image for training set which formed 2/3 of all images. The rest of them has been used for validation set.

### 4.2 Evaluation Metrics

Evaluation metrics which are used to measure image captioning performance come from the machine translation. The evaluation process is carried out by examining the word overlap between generated and ground truth sentences.
Commonly used metrics in the literature are BLEU (Bilingual Evaluation Understudy) [Papineni et al. 2002], ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation) [Lin and Och 2004], METEOR (Metric for Evaluation of Translation with Explicit ORdering) [Lavie and Agarwal 2007], CIDEr (Consensus-based Image Description Evaluation) [Vedantam et al. 2015].

## 5 Schedule

The schedule for project progress is given by Table 1.

| Date | Plan |
|---|---|
| March 14 | Project proposal submission |
| March 21 | Literature review |
| March 28 | Get preliminary result |
| April 4 | Code review |
| April 11 | Progress report preparation |
| April 18 | Progress report submission |
| April 25 | Code review |
| May 2 | Experiments |
| May 9 | Optimization |
| May 16 | Experiments |
| May 23 | Project report preparation |
| May 30 | Project submission |

**Table 1:** *Project Schedule Table*

## 6 Conclusion

In this paper, the topic to be studied in NLP course is disccused. In this respect, we start with giving information about the motivation that led us to pick up image captioning topic. After that, we provide relevant background and describe datasets and evaluation metrics.

## References


BAHDANAU, D., CHO, K., AND BENGIO, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

CHEN, X., AND ZITNICK, C. L. 2014. Learning a recurrent visual representation for image caption generation. *arXiv preprint arXiv:1411.5654*.

HODOSH, M., YOUNG, P., AND HOCKENMAIER, J. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research 47*, 853–899.

KIROS, R., SALAKHUTDINOV, R., AND ZEMEL, R. 2014. Multimodal neural language models. In *International Conference on Machine Learning*, 595–603.

LAVIE, A., AND AGARWAL, A. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, Association for Computational Linguistics, 228–231.

LIN, C.-Y., AND OCH, F. J. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, 605.

LIN, T.-Y., MAIRE, M., BELONGIE, S., HAYS, J., PERONA, P., RAMANAN, D., DOLLÁR, P., AND ZITNICK, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, Springer, 740–755.

PAPINENI, K., ROUKOS, S., WARD, T., AND ZHU, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for*

*computational linguistics*, Association for Computational Linguistics, 311–318.

VEDANTAM, R., LAWRENCE ZITNICK, C., AND PARIKH, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

VINYALS, O., TOSHEV, A., BENGIO, S., AND ERHAN, D. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3156–3164.

XU, K., BA, J., KIROS, R., CHO, K., COURVILLE, A., SALAKHUDINOV, R., ZEMEL, R., AND BENGIO, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, 2048–2057.

YOUNG, P., LAI, A., HODOSH, M., AND HOCKENMAIER, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics 2*, 67–78.