# Augmenting a multimodal Recurrent Neural Network with textual context for automatically generating image descriptions

Flip van Rijn

June 3, 2015

**Abstract**

## 1 Introduction

People with a visual impairment are not able to read text or reliably interpret visual input. For this target group special types of books, such as braille books or audio books, are made to enable these people to still perform the same task but then via a different medium.

Similarly, user interfaces of programs on computers are often enhanced in such a way that screen readers, which are text-to-speech software, can easily read out what is being displayed on the screen. Examples of these enhancements are alternative texts for images or reorganizing the layout such that only relevant information is given to the screen reader. However, in a study about the frustrations of screen reader users on the web Lazar et al. [6] aggregated a list of causes of frustration from 100 blind users. These users have a screen reader. This list showed that the often used enhancements, such as alternative text for images, are not always enforced by websites. The top causes of frustrations include (among others) the layout that causes out of order auditory output from the screen reader, poorly designed forms and no alternative text for images.

The study about the frustrations of screen reader users on the web shows that one cannot rely on external sources to provide a user friendly experience, such as a well formatted layout or an additional description of the images. Not only on the web, but also in magazines or newspapers there are often images that are not or limited augmented with textual descriptions and thus are useless for people with a visual impairment. However, multiple fields within artificial intelligence, such as computer vision, machine learning and linguistics, can be used to help the end-user with providing a computer generated description of an image.

### 1.1 Related work

Literature shows many studies involving object recognition and classification [1, 3] using convolutional neural networks (CNNs). In these studies objects ranging from inanimate to animate in images are classified, which results in a label that describes

to which class each object belongs to. The most recent study [3] claims a performance that even surpasses human performance on classifying images. While this study focuses on the image classification by labeling images, other studies [8, 9, 13, 2] have focused on the generation of sentences from images by combining a computer vision approach with linguistic models. For description generation often two types of approaches are used in the literature. The first type is connecting the grammar of a sentence in a description to an object or a relation between objects [5]. Models that use this approach generate sentences for the description that are following the syntactically correctness of the language grammar. The description generation model that is mentioned in [9] follows this first approach. It is trained on the Flickr datasets which consists of images and descriptions. In order to generate meaningful sentences, the model uses co-occurrence statistics to compute the probability distribution within a noun phrase. Furthermore, the characteristics of visually descriptive text are inspected to determine what generally the structure is of this type of text. These statistics are then used in the model along with the computer vision input (number of objects, labels) to generate novel sentences.

The second type of approach is using probabilistic machine learning to learn the probability density over multimodal input such as text and images. These models also generate sentences for the description, but are not according to a grammar. This results into more expressive sentences, but may contain less sound grammatical structures. The models in [8, 4, 5] are according to this second approach and the authors describe the model which consists of a multimodal Recurrent Neural Network. What this network makes multimodal network is the multimodal layer. This layer connects the word representations layer with the image feature extraction network that is finally combined into a multimodal feature vector.

## 1.2 Existing model

The overall architecture of the model that will be used is based on the model by Karpathy and Fei-Fei where a CNN is combined with a Recurrent Neural Network [4] or by Vinyals et al. where a CNN is combined with a Long-Short Term Memory (LSTM) network [12]. Both approaches use a CNN by Simonyan and Zisserman [11] to extract features from an image and a model which can integrate the image features and a sequence of words.

When an LSTM model is compared to an RNN model an LSTM model tends to outperform the other model on sequence tasks such as generating sentences from images or translation [12].

In the next sections the CNN is explained in more detail as well as the LSTM approach for combining the image and text.

### 1.2.1 Image features

The CNN part of the multimodal model is designed by Simonyan and Zisserman. They participated in the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) 2014 as the Visual Geometry Group (VGG) with a very deep convolution network and finished with the first and second place. Prior to their work, CNNs consisted of only a few layers and convolution filters with large receptive fields were used. However, Simonyan and Zisserman took another approach by lowering the size of the convolution filters to 3×3 in all layers and by increasing the number of weight

layers to 16 or 19. These architectures were not only able to achieve state-of-the-art accuracy on the classification and localization tasks, but were also able to generalize to other datasets.

The last layer of the model has 1000 output units that represent the 1000 categories on which the model has been trained. The activation of each of those units are probabilities for each category to be present in the image.

In order to integrate this model in the overall multimodal model, the classification layers are stripped off to expose the rectified linear unit (ReLU) in order to get the more rich internal representation of 4096 abstract features.

### 1.2.2 Combining image and text

Vinyals et al. use an LSTM model to generate sentences from images. The main component of such a model is the memory cell which contains information at every time step about the inputs that have been given up until this step. Furthermore, three binary gates govern the behavior of the model by telling the cell to forget the current value, read its input and outputting its new value. A general overview of such a model is shown in Figure 1 where each part itself and the relation with other parts are depicted. The $\odot$ nodes indicate a multiplication with the gate values.

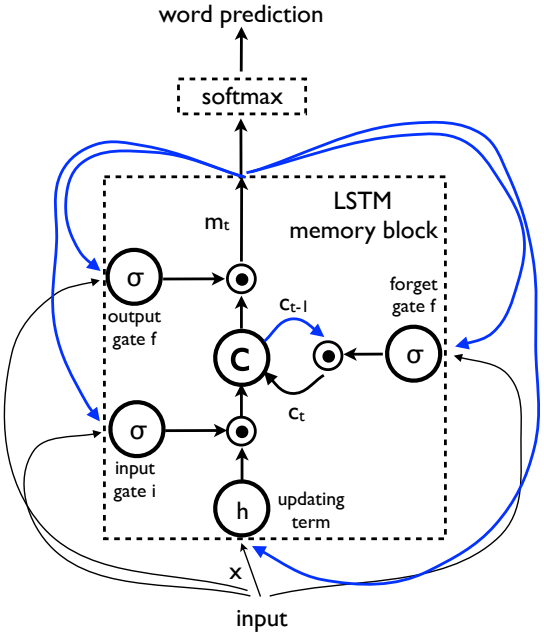


Figure 1: Illustration of an LSTM model adopted from [12]: Memory cell $C$ which is controlled by three gates $\sigma_{output}$, $\sigma_{input}$ and $\sigma_{forget}$. The blue lines indicate the recurrent connections from the output $m_{t-1}$ back to the gates at time $t$.

The image $I$ is fed into the model ($x$) only at the initial step. After this step the sequence of words in the ground truth sentence $S = (S_0, \ldots, S_N)$ is the input of the model. To mark the start and the ending of the sentence, $S_0$ and $S_N$ are special tokens. The words in the sentence are encoded as a 1-of-$K$ vector, where $K$ is the

number of words in the entire vocabulary (training set) and the 1 is the index of the word in the vocabulary.

$$L(I, S) = -\sum_{t=1}^{N} \log p_t(S_t) \tag{1}$$

The output of the model is a probability distribution $p_t$ over all words in the vocabulary. Finally, the loss function formalized in Equation 1 is minimized taking into account all the parameters of the model, the image features from the CNN and the word encodings.

# 2 Methods

## 2.1 Dataset

This consists of images, their descriptions and the textual context in which these images occur. Many datasets, such as ImageNET [10] and MSCOCO [7], exist of which many only consist of images and their description. Though, next to images and their annotations, in this approach the dataset also needs to include the context in which the image is in. In the literature the ImageCLEF dataset is used for various image retrieval tasks as well as for image annotation tasks. In order to compare the models in the literature with the model in the thesis and since this dataset contains the required components to develop and train a model for automatic image description generation, the ImageCLEF 2014 dataset is a suitable candidate. Furthermore, this dataset also contains the webpages on which the images are hosted. This can be used as the context for the images. However, the model of this thesis will be used by Dedicon and data that they have might not represent the data that is used by ImageCLEF. First of all the language (English vs. Dutch) is different which will have a big impact if this model were to be trained on the ImageCLEF dataset and then tested on the Dedicon dataset. Furthermore, the images in the Dedicon dataset might be of an entirely different category compared with the images in the ImageCLEF dataset. Therefore, only one of the datasets can be used to both train and test the model on. Once a working prototype of the model has been created and trained, the model is tested on a subset of the dataset that is used.

# Bibliography


[1] Peter Carbonetto, Nando de Freitas, Kobus Barnard, N Freitas, and Kobus Barnard. A statistical model for general contextual object recognition. In *Computer Vision-ECCV 2004*, pages 350–362. Springer, 2004.

[2] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 6314 LNCS, pages 15–29, 2010. ISBN 364215560X. doi: 10.1007/978-3-642-15561-1_2.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *arXiv preprint arXiv:1502.01852*, 2015.

[4] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *arXiv preprint arXiv:1412.2306*, 2014.

[5] Andrej Karpathy, Armand Joulin, and Li Fei-Fei. Deep Fragment Embeddings for Bidirectional Image Sentence Mapping. In *Advances in Neural Information Processing Systems*, pages 1–9, 2014.

[6] Jonathan Lazar, Aaron Allen, Jason Kleinman, and Chris Malarkey. What frustrates screen reader users on the web: A study of 100 blind users. *International Journal of human-computer interaction*, 22(3):247–269, 2007.

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. *arXiv preprint arXiv ...*, cs.CV, 2014. doi: 10.1007/978-3-319-10602-1_48.

[8] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014.

[9] Margaret Mitchell, Jesse Dodge, Amit Goyal, Kota Yamaguchi, Karl Stratos, Alyssa Mensch, Alex Berg, Xufeng Han, Tamara Berg, Oregon Health, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé III. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756. Association for Computational Linguistics, 2012. ISBN 978-1-937284-19-0.

[10] Olga Russakovsky and Li Fei-Fei. Attribute learning in large-scale datasets. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 6553 LNCS, pages 1–14, 2012. ISBN 978-3-642-35748-0. doi: 10.1007/978-3-642-35749-7_1.

[11] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[12] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555*, 2014.

[13] Yezhou Yang, Ching Lik Teo, Hal Daume, and Yiannis Aloimonos. Corpus-Guided Sentence Generation of Natural Images. In *Proceedings of EMNLP*, pages 444–454, 2011. ISBN 1937284115.