# Image caption generation using transfer learning

R. KOPIŃSKI, K. ANTCZAK

radoslaw.kopinski@student.wat.edu.pl
karol.antczak@wat.edu.pl

Military University of Technology, Faculty of Cybernetics
Institute of Computer and Information Systems
Kaliskiego St. 2, 00-908 Warsaw, Poland

This paper describes an image caption generation system using deep neural networks. The model is trained to maximize the probability of generated sentence, given the image. The model utilizes transfer learning in the form of pretrained convolutional neural networks to preprocess the image data. The datasets are composed of a still photographs and associated with it, five captions in English language. Constructed model is compared to other similarly constructed models using BLEU score system and ways to further improve its performance are proposed.

**Keywords:** Neural networks, NLP, caption generation, machine learning, computer vision, deep learning, transfer learning.

**DOI:** 10.5604/01.3001.0053.9697

## 1. Introduction

Image captioning task is assigning to an image, a concise description of its contents. While describing our environment comes quite naturally to humans, doing so in an automated manner requires use of sophisticated algorithms processing both visual and textual data. Modern machine learning techniques of image allow constructions of captioning systems that provide on its output a reasonably accurate description. By combining methods from areas of Natural Language Processing (NLP) and object detection, we are able to extract the relevant information from the image to generate a meaningful human-readable caption describing in short words the content of the image.

While caption generation is a combination of two problems of object detection and text generation, it goes beyond those problems as not only information about individual objects has to be inferred from the image, but also their distinct features and relations between them. Other than that, a way to translate those features and relations to a natural language is of a different kind than regular translation. These additional sub-problems make the captioning of an image a harder problem than it is to be expected.

A most common usage of automatic image captioning might be found in search engines for finding the correct image given a text query and then performing a closest-match text search in collection of generated image descriptions. But as the captioning problem does not have to be constrained to images, one can imagine a situation where similar techniques can be used by some systems to supply self-diagnostic information to its users in natural language based on its abstract state vector

## 2. Related works

Image caption generation is a well-known task in both computer vision and natural language processing. Early approaches used various algorithms and techniques. One group of models, known as “retrieval-based” ones, generated sentences by composing preexisting sentences based on their adequacy score, as demonstrated by Farhadi et al. [1]. Alternative approaches utilized more or less complex templates representing syntactic structure of sentence that were “filled in” by various visual models. An example of this method can be seen in [2]. These early approaches mainly suffered from a rigid structure that limited possible generated sentences.

The newer approaches are almost exclusively based on neural networks. Their flexible nature provides more expressive power. There are, however many various models. A survey from 2018 [3] lists and categorizes many of them. A notable ones include "multimodal" learning models as one shown by Karpathy et al [10]. These approves work by extracting visual features first and then predicting consecutive words using language model, like word2vec [3].

The idea of using language models is extended in "encoder-decoder" group of methods, where caption generation is treaded as a machine translation problem. This led to improved models such as "Show and Tell" architecture by Vinyals et al [11].

Recently, a number of neural models were proposed utilizing so-called "attention" layers. They are inspired by mechanism of visual attention in animals and humans. They work by applying specific kind of filtering on visual features that allow models to focus on specific aspects of it while ignoring other. It has been shown that it can improve performance of many image vision models for various problems, including caption generation. An example of this is "Show, Attend and Tell" architecture [12].

## 3. Proposed model

An approach utilized in this paper uses neural model to generate consecutive words composing a concise description of image given at its input.

Textual data is interpreted as a sequence of tokens where each token uniquely represents one word and each sequence has guard tokens as first and last elements, being beginning and end of a sentence. A sequence comprising of just one token (a beginning guard token) is interpreted as empty sentence and is assigned to input image at the start of captioning process.

Input image data is downscaled and transformed by a CNN to get the 1-dimentional feature vector of fixed length. The architecture of used network is not relevant but it is important that it produces a feature vector that incorporates semantic information of the input image, by projecting it onto latent feature space. For that reason, pretrained convolutional object detection networks, with output layers removed, are used to produce such a vector, with assumption that classifiers that perform better, internally represent image with feature vector of better "quality".

Model used by the system, given a token sequence and feature vector, will generate a probability distribution of possible word continuations of input sentence. From this distribution a word with maximum probability is taken as the next word in the description of the image and unless terminating guard token is generated, its output is used to generate next word.

The neural network is constructed in a "Merge" manner described in Tanti et al [5]. Textual and visual inputs are processed concurrently and then both results are combined to produce final distribution. This approach allows to reduce the dimension of LSTM encoder layer in comparison to other architectures.

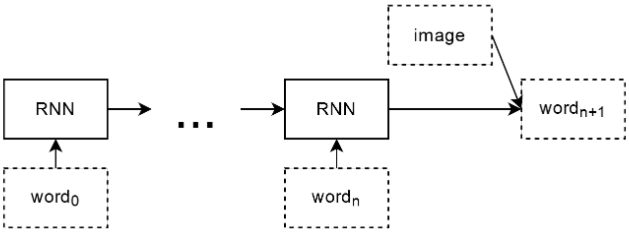


Fig. 1. General scheme of "merge" captioner architecture

Words in input sentence are encoded using word embedding vectors. This method has the advantage of small distance between representations of words with similar meanings. This mapping is trained along with the rest of the model. Embedded sequence is then fed to LSTM layer that transforms given sequence to fixed-size context vector.

Visual input is processed by CNN to a feature vector which is then linearly reduced to have its length equal to context vector to make its combination with it easier. The convolutional network used in this particular model is VGG-16 [6].

Feature vector and context vector are combined to make single vector that is then passed through dense layers to produce encoded word vector which is then transformed by dense layer with argmax activation to final probability distribution of the next word.

The exact method of combining feature and context vectors can be chosen freely, but it has been observed that the choice of this function has meaningful impact on results of the model.

In this model following functions were tested:

$$v_i = c_i + f_i \tag{1}$$

$$v_i = c_i f_i \tag{2}$$

$$v_i = (Wc)_i f_i + c_i \tag{3}$$

where:
$c_i$ – the element of context vector,
$f_i$ – the element of feature vector,
$W$ – a square matrix which weights are trained along with the model,
$v_i$ – the resultant vector.

Best BLEU scores were observed when using function **(3)** and function **(2)** generated numerical errors on larger dataset. Any further statistics regarding model performance are reported for model using function **(3)**.

The output layer has its dimension defined by $V$, which is the size of the vocabulary set, which in turn is dependent on the dataset used. To counteract overfitting, a dropout was used on the inputs of indicated layers during the training phase. The merge function is defined by one of equations shown earlier. The size of context vector was chosen arbitrary as a balance between model's performance and training time. Textual input layer does not depend on the size of the vocabulary, as for one-hot encoding only index is preserved.

## 4. Datasets

Flickr30k and Flickr8k (subset) were used to train and test the model's performance, standard dataset role divisions were used. Both datasets were divided into three subsets for training, evaluation and testing with proportions 8:1:1 respectively.

Tab. 1. Sizes of used dataset divisions

| Dataset name | Number of entries | | |
|---|---|---|---|
| | Train | Valid. | Test |
| Flickr8k | 6000 | 1000 | 1000 |
| Flickr30k | 25783 | 3000 | 3000 |

Each dataset comprises of number of photographic images in varying resolutions and contexts. Image descriptions are human generated through crowdsourcing method. Each image has 5 descriptions associated with it.

Before use, data is preprocessed in following way:

1. Punctuation symbols and single letter words are removed.
2. Descriptions which contain words that appear less than 5 times in whole dataset are removed.
3. Descriptions that are longer than 30 words are removed.
4. Images that do not have any descriptions left are removed.

The input-output pairs, are then generated from a word sequence $\boldsymbol{S}$ in a following manner:

$$S_i^x = S_0, \dots, S_{i-1} \tag{4}$$
$$S_i^y = S_i \tag{5}$$

where $S_i^x, S_i^y$ are input and output symbols respectively. Even though output symbol is a singular value, it is extended to unit probability distribution $\boldsymbol{y}$ that has its elements defined as:

$$\boldsymbol{y_i} = [i = d] \tag{6}$$

where *d* is the symbol index in a dictionary.

## 5. Training

Model is trained to maximize the probability of next word, given an image data and an existing description.

The loss function used in training is *Categorical Cross Entropy* function, which

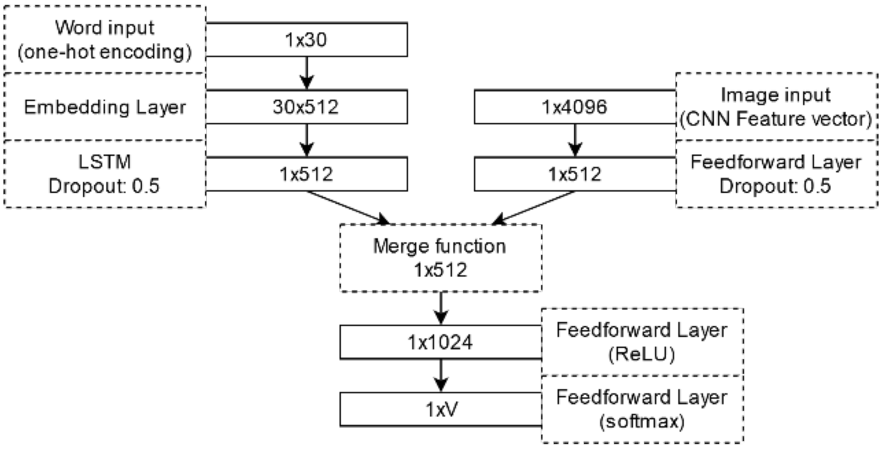


Fig. 2. Diagram of constructed model

computes the similarity between result and reference distribution as follows:

$$L(y,\hat{y}) = -\sum y_i ln(\hat{y_i}) \qquad (7)$$

It is minimized with relation to all parameters of dense, LSTM and embedding layers by the use of Adam [7] algorithm.

Training was performed until the value of the loss function showed no improvement. In practice it was observed that loss function stopped improving after around 6 epochs

The sentences were generated using the "Sampling" method where the word with maximum probability is chosen.

Training was performed with Tensorflow and Keras libraries [8], [9] on Mac mini with Apple M1 processor.

## 6. Evaluation

Model was evaluated and compared with other caption-generation models. To compute numerical score to assess the model performance the BLEU score was utilized, with ground-truth labels as references and generated sentence as a proposition. This scoring technique compares similarity of n-grams between reference labels and generated ones.

It has been shown that this score reflects well the human evaluations of translation systems, for this reason it is widely used to score the caption generation systems as we can treat the image at the input of the system as a sentence in "visual language".

A comparison has been made with other models that were described in recent years and are constructed with similar principles:

- **Karpathy et al. (2015)** – An "Init inject" type model without attention [10]
- **Show & Tell (2015)** – Similar "Init inject" model but with better results [11]
- **Show, Attend & Tell (2016)** – Model using an additional attention layer. [12]

Tab. 2. Model BLEU-n score comparison (Flickr8k dataset)

| Model | Flickr 8K | | | |
|---|---|---|---|---|
| | B-1 | B-2 | B-3 | B-4 |
| Karpathy et al. | 57,9 | 38,3 | 24,5 | 16 |
| S&T | 63 | 41 | 27 | – |
| S, A&T | 67 | 45,7 | 31,4 | 21,3 |
| **This model** | **44** | **26,2** | **18,6** | **8,5** |

Tab. 3. Model BLEU-n score comparison (Flickr30k dataset)

| Model | Flickr 30K | | | |
|---|---|---|---|---|
| | B-1 | B-2 | B-3 | B-4 |
| Karpathy et al. | 57,3 | 36,9 | 24 | 15,7 |
| S&T | 66,3 | 42,3 | 27,7 | – |
| S, A&T | 66,9 | 43,9 | 29,6 | 19,9 |
| **This model** | **45,9** | **26,4** | **18,4** | **8,7** |

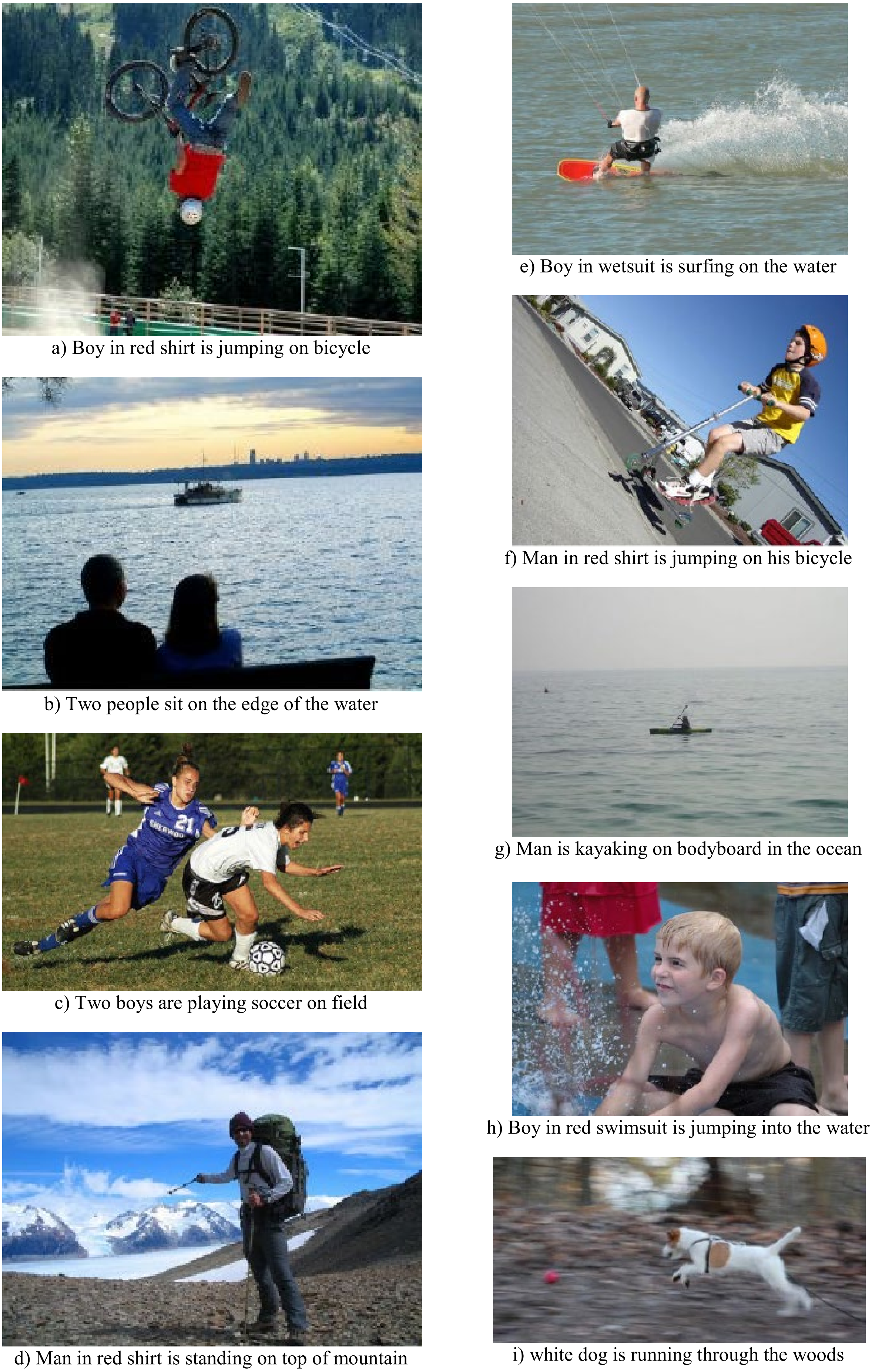
It can be seen that the presented model scores a bit lower than other models built on the similar principles. Most of this difference can be attributed to overall lower performance of "Merge" architectures versus "Init-inject" architectures used in the models. The tradeoff for lower scores is smaller number of trainable parameters for compute-heavy "LSTM" layers as they do not need to encode the image along with textual data, on the other hand the interaction between the two is also smaller. The score could also be increased by utilizing a more advanced CNN like ConvNeXt [13] instead of basic VGG-16[2] network.

## 7. Conclusion

This paper presented an architecture and construction of an end-to-end neural network system that given an image, will generate a reasonably accurate description of an image in English language. The model is based on joining together, separately encoded vectors from convolutional and recurrent networks to generate a most probable sentence based on training data.

Even though CNN that was used to produce feature vector was trained for object detection task, separately from the rest of the model, its output provides meaningful information to the network that performs captioning task, positively impacting its score. One can wonder if other parts of the neural network could also be transferred from networks that solve particular subproblems, to achieve better results.

A comparison was made to other models with similar construction. Although the BLEU scores were comparable to the reference models, there is still much room for improvement. Ways to improve achieved results were proposed and possibly pursued in the future.



a) Boy in red shirt is jumping on bicycle

b) Two people sit on the edge of the water

c) Two boys are playing soccer on field

d) Man in red shirt is standing on top of mountain

e) Boy in wetsuit is surfing on the water

f) Man in red shirt is jumping on his bicycle

g) Man is kayaking on bodyboard in the ocean

h) Boy in red swimsuit is jumping into the water

i) white dog is running through the woods

Fig. 3(a–i). Examples of images from Flickr8k dataset, labeled by the model. Sample of erroneous labels can be seen in examples: d, f, h; where model would generate a wrong adjective in an otherwise correct label

## 8. Bibliography

[1] Farhadi A., et al., “Every picture tells a Ssory: Generating sentences from images”, *Computer Vision – ECCV 2010*, LNCS 6314, pp. 15–29, Springer 2010.

[2] Mitchell M., et al., “Midge: Generating image descriptions from computer vision detections”, in: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 747–756, April 2012.

[3] Bai S., An S., “A survey on automatic image caption generation”, *Neurocomputing*, Vol. 311, 291–304 (2018).

[4] Mikolov T., Chen K., Corrado G., Dean J., “Efficient estimation of word representations in vector space”, arXiv preprint arXiv : 1301.3781, September 2013.

[5] Tanti M., et al., “Where to put the image in an image caption generator”, *Natural Language Engineering*, Vol. 24(3), 467–489 (2018).

[6] Simonyan K. et al., “Very deep convolutional networks for large-scale image recognition”, *CoRR*, abs/1409.1556v6 (2014).

[7] Kingma D.P., Ba J., “Adam: A Method for stochastic optimization”, *CoRR*, abs/1412.6980v9 (2017).

[8] Abadi M., et al., “TensorFlow: Large-scale machine learning on heterogeneous systems”, November 2015, Software available from www.tensorflow.org.

[9] Chollet F., et. al., Keras, 2015, https://keras.io.

[10] Karpathy A., Fei-Fei L., “Deep Visual-Semantic Alignments for Generating Image Descriptions”, *CoRR*, abs/1412.2306 (2014).

[11] Vinyals O. et al., “Show and Tell: A Neural Image Caption Generator”, *CoRR* abs/1411.4555 (2014).

[12] Xu K. et al., “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”, *CoRR*, abs/1502.03044 (2015).

[13] Liu Z. et al., “A ConvNet for the 2020s”. *CoRR*, abs/2201.03545 (2022).

## Generowanie podpisów na podstawie zdjęć z użyciem uczenia transferowego

R. KOPIŃSKI K. ANTCZAK

W tym artykule opisano system generujący podpisy do zdjęć z wykorzystaniem głębokich sieci neuronowych. Model jest trenowany pod kątem maksymalizacji prawdopodobieństwa wygenerowanego zdania, dla zadanego obrazu. Model wykorzystuje uczenie transferowe w postaci wytrenowanych wstępnie neuronowych sieci konwolucyjnych. Zbiory danych wykorzystane do trenowania modelu składają się z fotografii, oraz przypisanych do niej pięciu zdań w języku angielskim. Skonstruowany model jest potem porównany z innymi modelami o podobnej konstrukcji z wykorzystaniem punktacji BLEU.

**Słowa kluczowe**: Sieci neuronowe, generowanie podpisów, maszynowe uczenie, widzenie komputerowe, głębokie uczenie, uczenie transferowe.