

## Image Caption Generation Using Deep Learning

Sailee P. Pawaskar<sup>1\*</sup>, J. A. Laxminarayana<sup>2</sup>

<sup>1</sup>Computer Dept., Goa College of Engineering, Goa University, Farmagudi, Ponda, Goa, India

<sup>2</sup>Computer Dept., Goa College of Engineering, Goa University, Farmagudi, Ponda, Goa, India

\*Corresponding Author: sailee.spawaskar516@gmail.com,

Available online at: [www.ijcsonline.org](http://www.ijcsonline.org)

**Abstract**— From the perspective of humans and computers, a picture can be interpreted in distinct manner. In the case of humans, a picture will be clearly a few description or scene of a motion or environment and so forth, whilst with respect to computers, it is just a few aggregates of pixels or digital numbers. The system of photo captioning offers with assigning inner facts in the shape of captions with the aid of extracting the applicable functions from an input picture.

This venture aims at producing meaningful captions for a given picture. The proposed work is based on deep neural networks. The proposed work has three fundamental units. The first is the picture module that is given as input to the function extractor unit. The next unit is a feature extractor unit based on CNN (Convolutional Neural Network) which extracts the applicable characteristic. The final unit is the language generator. It generates sentences that describe the input image.

To assess the quality of the generated textual content, BLEU(Bi-Lingual Evaluation Understudy) rating is used. Suitable captions will help the users to search snapshots with lengthy queries. Such systems may also be beneficial for visually impaired humans in understanding pictures.

**Keywords**—BLEU rating, captions, CNN, deep neural network

### I. INTRODUCTION

People are excellent at processing pictures and collecting the information. They are able to study or infer a lot from a picture. However, the computers are inefficient to make proper sense of the given picture.

The main purpose of automated photo captioning consists of taking a photo, analysing its visual content, and producing a textual description that determines the capabilities of the image. A good description should be complete but concise whilst being officially correct. As an example, there are millions of photos uploaded over the net each day. Rapidly growing amount of visible data gives a mission to build smarter computer algorithms to recognize and summarize the information. Understanding of visible statistics is hence a critical issue in many elements of machine vision and AI. The device calls for to perceive the objects inside the image, apprehend their characteristic and extract the relationships among those objects. The extracted facts can then be used to generate a natural language caption. Suitable captions will help the users to search snapshots with lengthy queries. Such systems may also be beneficial for visually impaired humans in understanding pictures using automatic text to voice convertors.

The organization of this document is as follows. Section II covers the literature survey, which explains various methods

that have been adapted in the past for image captioning. Section III states the problem definition along with the proposed work. Section IV explains the detailed design methodology to achieve the objectives. Section V explains the implementation details and the experimental results of the proposed work. Finally, Section VI states the conclusion and the future scope of the project.

### II. LITERATURE SURVEY

#### A. Neural Network Approaches

Xinlei Chen et al presented two-way mapping between photos and their sentence primarily based entirely on descriptions [1]. The version projected was capable of generating novel captions given a picture, and reconstructing visible capabilities given a photograph description. The assessment was applied on many obligations just like the sentence generation, sentence retrieval, and image retrieval. As compared to human-generated captions, captions generated by means of this version were favoured via citizenry 21.0% of the time.

Jeff Donahue et al advanced a unique recurrent convolutional design and validated the value of this model on benchmark video reputation obligations, picture description, and retrieval troubles, and video narration demanding situations [2]. They made use of recurrent convolutional fashions that are “doubly deep”. These recurrent lengthy-term models

have been without delay connected to convent models after which have been skilled to simultaneously analyze temporal dynamics and convolutional perceptual representations.

Vinayshekhar Bannihatti Kumar et al improvised present technology used in famous social networking web sites like Twitter, to encompass a number of the brand new states of the art technology in system studying to construct features [3]. The version advanced changed into just like twitter along with extra features that consist of image captioning, auto-tagging of Tweets, sentiment detection, unsolicited mail filtering, and a progressive information feed generator.

#### B. Hidden Markov Approaches

Arnab Ghoshal et al presented a unique technique for computerized annotation of images with key phrases from a generic vocabulary of standards or objects for the reason of content-based photo retrieval [4]. A picture, represented as a chain of function vectors characterizing visual capabilities inclusive of shade, texture or oriented-edges, was modelled as having been stochastically generated via a hidden Markov model, whose states constitute ideas. The parameters of the version were expected from a set of manually annotated images. Each picture in a massive take a look at collection became then routinely annotated with the a posteriori chance of ideas present in it.

David Zajic & Bonnie Dorr proposed a singular utility of Hidden Markov models to the automated generation of informative headlines for English texts[5]. This model defined four interpreting parameters to make the headlines seem greater like Headlines, the language of informative newspaper headlines. It additionally allowed morphological variant in phrases among headline and tale English. Casual and formal reviews indicated that this method produced informative headlines.

Philo Sumi et al presented computerized caption technology for news photographs in affiliation with the related information article [6]. Here they gave one photograph and informative article as an input to the system. The device then generated the maximum vital key phrases that are associated with the photograph in association with the picture. To discover the photograph related key phrases, first, they determined out the input picture's capabilities the use of the SIFT method. Moreover, the usage of those capabilities it as compared the photograph with the snapshots which have been stored within the database. After locating the excellent-matched photograph, they extracted the keywords related to that photograph. Subsequently, after making use of grammatical rules to the key phrases, the perfect caption was generated.

#### C. Other Approaches

Krishnan Ramnath et al evolved a machine that helped a smartphone user to generate a caption for his or her

photographs [7]. It operated through importing the photo to a cloud carrier where some of the parallel modules were implemented to understand a diffusion of entities and relations. The outputs of the modules were then mixed to generate a huge set of candidate captions, which had been then again to the cellphone. The cellphone purchaser included a convenient person interface that could permit customers to choose their preferred caption, reorder, add, or delete phrases so that it will acquire the grammatical fashion they favored. The person can also pick out from more than one caption.

### III. PROBLEM DEFINITION

Image captioning aims to develop structures that generate textual descriptions approximately items in photos. Given the picture as an input, the system will extract the special objects, moves, and attributes, and subsequently, generate a meaningful sentence (caption) for the picture.

A caption should recognize the objects contained in a photograph, but it also ought to explore the connection between those objects. It needs to additionally recognize the attributes and therefore the activities they'll be concerned in. Eventually, the above information needs to also be presented in a natural language like English.

### IV. DETAILED DESIGN

The proposed model has three modules. First is the image module, second is the function extractor module and the last one, the language module. The photograph is considered as input to the system.

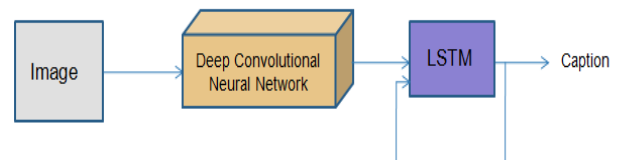


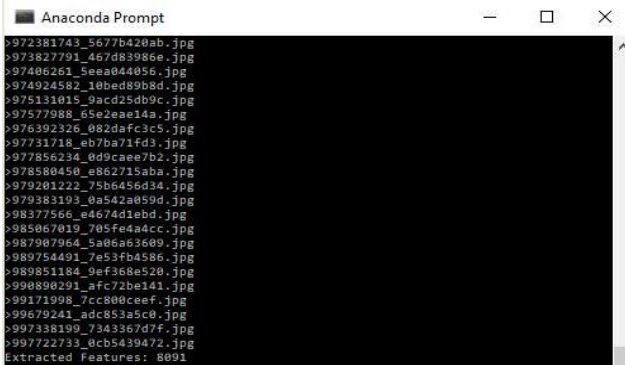
Figure 1 Architecture of the proposed model

The Deep CNN is used as a feature extractor. After extracting the features, the Deep CNN will produce output within the shape of a photograph vector. This vector is taken as input to the final module. The language module that we have chosen is the LSTM (Long Short Term Memory). The LSTM will recursively generate the phrases until it forms a sentence.

### V. IMPLEMENTATION AND EXPERIMENTAL RESULTS

This section describes the implementation and experimental results of the proposed model. The implementation was done using the features of Python programming language Flickr8k dataset containing 8092 photographs in JPEG format.

The first step here is to extract the features from a given input image. Therefore we pre-compute the “photo features” and save them to file. Later we can load these features and feed them into our model as the interpretation of a given photo in the dataset. This will make our model to train faster and consume less memory.



```

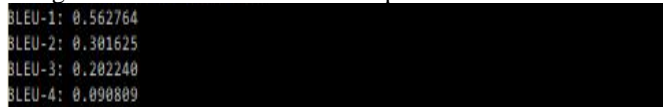
Anaconda Prompt
972381743_5677b420ab.jpg
973827791_467d83986e.jpg
97406261_5eea044056.jpg
974924582_10bed89b8d.jpg
975131015_9acd25db9c.jpg
97577988_65e2ee114a.jpg
976302326_082dafc3c5.jpg
97731718_eb7ba71fd3.jpg
977856234_0d9cae7b2.jpg
978580450_e862715aba.jpg
979201222_75b6456d34.jpg
979383193_0a542a059d.jpg
98377566_e4674d1ebd.jpg
985067019_705fe4a4cc.jpg
987907064_5a06a33809.jpg
989754491_7e53fb4586.jpg
989851184_9ef368e520.jpg
99080291_3fc72be141.jpg
99171998_7cc800ceef.jpg
99679241_adc853a5c0.jpg
997338199_7343367d7f.jpg
997722733_0cb5439472.jpg
Extracted Features: 8091

```

Figure 2 Output after extracting photo features

Every photograph has multiple descriptions related to it and every photo has a unique identifier. Next, we map the image identifier to the descriptions. We also want to smooth the outline with a view to reducing the size of the vocabulary of phrases. Then we ultimately print the scale of vocabulary.

After loading the data, we need to evaluate it. The actual and predicted descriptions are collected and evaluated using the BLEU score. It determines how close the generated text is to the expected text. We compare each generated description against all of the reference descriptions for the photograph using which BLEU scores are computed.



```

BLEU-1: 0.562764
BLEU-2: 0.301625
BLEU-3: 0.202240
BLEU-4: 0.090809

```

Figure 3 BLEU score

The last step is to validate the model. The sample photograph shown in Figure 4 is taken as input to the model. After training the model the language model i.e, LSTM generates the description as a caption.

Figure 5 shows the caption generated for the input image by the model.



Figure 4 Input to the model

little boy is playing in the pool

Figure 5 Generated caption

## VI. CONCLUSION AND FUTURE SCOPE

Image captioning is a completely tough venture than an easy category of classification of snapshots. This work supplied a deep learning model that automatically generates photograph captions. Our defined version is primarily based on a CNN that extracts the capabilities from a photo, followed by using an LSTM that generates corresponding sentences primarily based on the learned photo capabilities. The implementation of the photo caption technology in Python is efficient and user friendly. Moreover, BLEU assessment metric that was used to evaluate the quality of automatically generated texts works really well.

As a future work, multiple sentences could be generated with an exceptional content material. Further, a distinctive language model like GRU may also be used to test the performance of the generated sentences.

## REFERENCES

- [1] Chen and C. L. Zitnick, "Mind's eye: A recurrent visual representation for image caption generation," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 2422-2431.
- [2] J. Donahue et al., "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, Issue 4, pp. 677-691, April 1 2017.
- [3] V. B. Kumar, T. R. Baadkar, and V. Joshi, "CRYPTANITE: A New Look to the World of Social Networks Using Deep Learning," 2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), Naples, 2016, pp. 358-364.
- [4] Arnab Ghoshal, Pavel Ircing, Sanjeev Khudanpur "Hidden Markov Models for Automatic Annotation and Content-Based Retrieval of Images and Video".
- [5] Zajic R. Schwartz, D & Door, B & Schwartz, Richard "Automatic Headline Generation for Newspaper Stories", 2018.
- [6] PHILO SUMI , ANU.T.P " A Systematic Approach for News aption Generation", International Journal of Advanced Research in Computer Science & Technology (IJARCST 2014), Vol. 2, Issue 2, Ver. 1 (April -June 2014).
- [7] K. Ramnath et al., "AutoCaption: Automatic caption generation for personal photos," IEEE Winter Conference on Applications of Computer Vision, Steamboat Springs, CO, 2014, pp. 1050-1057.

## Authors Profile

Miss. Sailee P. Pawaskar pursued Bachelor of Engineering from Goa College of Engineering in Information Technology, Goa University, Farmagudi, Ponda-Goa in 2016 and Master of Engineering from Goa College of Engineering in Computer Science & Engineering, Goa University, Farmagudi, Ponda-Goa in 2018.