International Journal of Research
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 07 Issue 05
May 2020

Image Captioning Using Deep Learning

Mr. P Rajasekhar Reddy
Department of Computer Science and Engineering
Anurag Group of Institutions, Hyderabad, Telangana, India
rajasekharreddycse@cvsr.ac.in

Sainath Omdas, Venkatesh Nangi and Neelam Chavada
Anurag Group of Institutions, Hyderabad, Telangana, India
sainathomdas@gmail.com, venkateshnangi18@gmail.com, chavadaneelam999@gmail.com

Abstract– In Artificial Intelligence, Caption generation is a challenging task where a textual description must be generated for a given image. It requires both methods from computer vision to understand the objects and actions involved in the image and a natural language processing model to generate a caption. Image captioning has various applications such as usage in virtual assistants, recommendations in editing applications, for image indexing, for social media, for visually impaired persons, and many other natural language processing applications. We propose a hybrid system using multilayer Convolutional Neural Network (CNN) to generate vocabulary describing the images and a Long Short Term Memory (LSTM) to accurately structure meaningful sentences using the generated keywords. The CNN compares the given image to a large dataset of training images, then generates an accurate description using the trained captions.

I. INTRODUCTION

The purpose of image captioning is to automatically describe an image with one or more natural language sentences. Every day, we encounter numerous images from various sources such as the news articles, internet, advertisements and document diagrams. Most of the images in these sources do not have a caption describing them, but the human can understand them by seeing the objects and the actions involved in it. However, a machine needs to interpret some kind of image captions if a human wants to generate automatic image captions from it. The web is filled with millions of images, helping to entertain and inform the world on an infinite variety of subjects. However, much of that image information is not accessible for visually challenged, or with slow internet speeds that prohibit the loading of images. Image captions that are manually added by website authors using Alt-text HTML, is a way to make this content more accessible, so that a natural language description for images that can be presented using text-to-speech systems. However, a small fraction of existing web images contain Alt-text HTML fields. Automatic image captioning can be used to solve this problem. Automatic image captioning has various applications such as usage in virtual assistants, recommendations in editing applications, for image indexing, for social media, for visually impaired persons, and many other natural language processing applications.

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 07 Issue 05
May 2020

The rest of the paper is organized as follows. Proposed algorithm is explained in section II. Related work is presented in section III. Experimental results are presented in section IV. Concluding remarks are given in section V.

II. PROPOSED ALGORITHM

The proposed system is composed of two parts: a) an image feature extractor model based on CNN to extract image features in the form of a fixed-length vector, and b) a language model that takes the outputted fixed vectors from the previous models as input and makes a final prediction.

Neural network models for image captioning involve two main elements:

a) Feature Extraction.

b) Language Model.

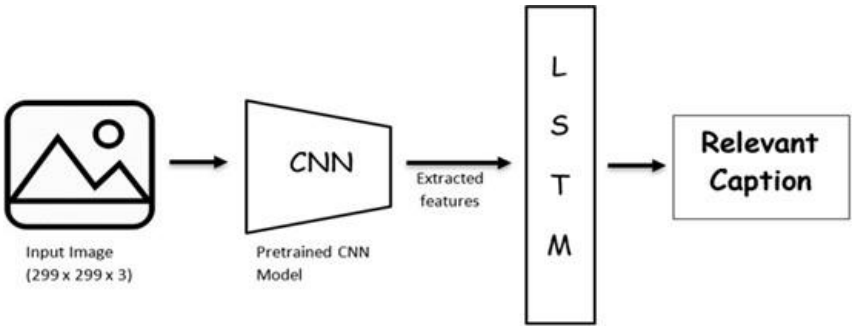

Figure 1. Architecture of our proposed model

a) Feature Extraction: The feature extraction model is a neural network used to extract the features of the given image, often in the form of a fixed-length vector. A deep convolutional neural network (CNN) is used as the feature extraction submodel. This network can be trained directly on the images in our dataset. Alternatively, we can use a pre-trained convolutional model. We are using the pretrained Inception-V3 model to extract image features created by Google Research for image classification. Inception-V3 is a widely-used image recognition model that has been shown to attain greater than 78.1% accuracy on the ImageNet dataset. The model is the culmination of many ideas developed by multiple researchers over the years. However, our purpose is not to classify images but to extract features and convert them into a fixed size vector. This method is called Automatic feature engineering. Hence, we remove the last softmax layer from the pretrained model and extract a 2048 length vector for the image.

b) Language Model: We create an LSTM based model that is used to predict the sequences of words, called the caption, from the feature vectors obtained from the Inception V3 network. But the prediction of the entire caption to the given image does not happen at once. The caption will be predicted one word at a time. This model consists of numerous LSTM cells depending on the maximum length of any caption we want to generate. Each of these LSTM cells predicts the word by taking the

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 07 Issue 05
May 2020

sequence of previously generated words as input. Hence, we need a 'first word' to start the generation process and a 'last word' to end the generation process.

III. RELATED WORK

A significant amount of work has been done on generating captions for images. The first significant work in solving image captioning tasks was done by Ali Farhadi [8] where three spaces are defined namely the image space, meaning space and the sentence space where mapping is done from the respective image and sentence space to the meaning space.

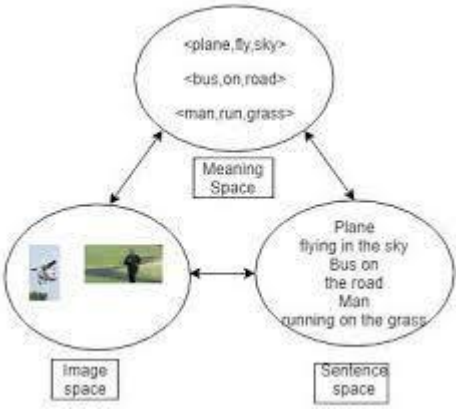

Figure 2. The three spaces defined by the work proposed by Ali Farhadi

With the help of mapping, similarity between the images and the sentence is evaluated, the meanings are stored as triplets of (image, action, object) and a score is evaluated by predicting the image and sentence triplets. If an image and sentence have a high level of similarity in terms of the predicted triplets then they will be highly compatible and have a high score. Thus, appropriate sentences can be generated. This model has many drawbacks such as the requirement of the middle meaning space and the results obtained from it are not at all highly accurate.

Pan et al. [2] proposed technique which work by annotating a particular part of image region, in this a word for each image region and then on combining we get the sentence, they discussed about considering the image regions as blobs-token which means an image regions based on its feature such as colour, texture, and position of object in image. But this technique has few limitations such as it is effective only for a small dataset, this work include much manual work as they have to provide annotated words and blob-tokens with an image, and this approach sometimes give results based on a training set, that means it is biased on a training set.

Jacob et al. [3], provided a technique that explores nearest neighbor images in training set to the query image and based on that appropriate k nearest neighbor images, their given captions are imported and based on that only caption is given for the query image. But there is a limitation of this approach as it performs better for highly similar images, but worse for highly dissimilar images.

In Ordonez, Vicente, GirishKulkarni, and Tamara L. Berg [4] , they used a similar approach

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 07 Issue 05
May 2020

as a candidate matching images with the query image are retrieved from the collection of captioned images, then after features are matched and based on the best rank obtained a caption is given to the query image. But this model has few limitations like re-ranking the given captions could create error for training images and related text, and also object and scene classifiers could give erroneous results, so the model could have given faulty results.

The work of D. Narayanswamy et al [5] which aims at generating labels that define the video frames, or that of D. Elliott and F. Keller [6], Image description using visual dependency representations, wherein the authors aim at identifying the different elements of an image. However the research stated above are generally concerned with and focused more on using image processing to detect and identify various objects in an image. They don't deal with identifying the context of these objects. These research papers have allowed us to greatly understand the concept of image processing and segmentation that plays a major role in our model. In addition to this, we try to obtain the contextual description of the image.

Through our model we aim to produce semantically and visually grounded description of abstract images, the proposed description of which would be in natural language i.e. human perceived description. By leveraging the techniques like CNN, RNN, and data sets such as those of Flick8K, Flickr30K and MS-COCO, we strive to attain the human level perception of given images.

IV. EXPERIMENT AND RESULT

The proposed system is trained on the Flickr 8K dataset provided by the University of Illinois at Urbana-Champaign. This dataset contains 8000 images with 5 captions for each image. The whole setup has run on the Google Colab IDE for a faster training process. Keras 2.0 was used to implement the deep learning model because of the presence of the Inception V3 which was used for object identification. Tensorflow library is installed as a backend for the Keras framework for creating and training deep neural networks. TensorFlow is a deep learning library developed by Google. The results are shown below.

		
Woman in white shirt playing tennis	Group of people are standing in front of building	Man in green and blue shirt with helmet is riding bike

Table 1. Caption generated by our model for the respective image

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 07 Issue 05
May 2020

V. CONCLUSION

We have implemented a CNN-RNN model for generating the captions for images. We used a very small dataset consisting of 8K images. For production level, we need to train on larger datasets having more than one lakh images which can produce better accuracy models. We must understand that the images used for predicting the caption must be semantically related to those used for training the model. For example, if we train our model on the images of dogs, cats, etc. we must not test it on images of waterfalls, air planes, etc. This is an example where the distribution of the train and test sets will be different and in such cases no Machine Learning model in the world will give good performance.

REFERENCES

[1] Gurjar, Sonu& Gupta, Shivam&Srivastava, Rajeev. (2017). Automatic Image Annotation Model Using LSTM Approach. Signal & Image Processing : An International Journal. 8. 25-37. 10.5121/sipij.2017.8403.

[2] Pan, Jia-Yu, Hyung-Jeong Yang, Pinar Duygulu, and Christos Faloutsos. "Automatic image captioning."In Multimedia and Expo, 2004.ICME'04. 2004 IEEE International Conference on, vol. 3, pp. 1987-1990. IEEE, 2004.

[3] Devlin, Jacob, Saurabh Gupta, Ross Girshick, Margaret Mitchell, and C. Lawrence Zitnick. "Exploring nearest neighbor approaches for image captioning."arXiv preprint arXiv:1505.04467 (2015).

[4] Ordonez, Vicente, GirishKulkarni, and Tamara L. Berg. "Im2text: Describing images using 1 million captioned photographs." In Advances in Neural Information Processing Systems, pp. 1143-1151. 2011.

[5] Barbu, A. Bridge, Z. Burchill, D. Coroian, S. Dickinson, S. Fidler, A. Michaux, S. Mussman, S. Narayanaswamy, D. Salvi, et al. Video in sentences out. Uai, 2012

[6] D. Elliott and F. Keller. Image description using visual dependency representations.EMNLP, 2013.

[7] Pan, Jia-Yu, Hyung-Jeong Yang, Christos Faloutsos, and Pinar Duygulu. "Gcap: Graph-based automatic image captioning." In Computer Vision and Pattern Recognition Workshop, 2004.CVPRW'04. Conference on, pp. 146-146. IEEE, 2004.

[8] Farhadi, Ali, Mohsen Hejrati, Mohammad Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. "Every picture tells a story: Generating sentences from images."Computer vision–ECCV 2010 (2010): 15-29.

[9] https://www.ijedr.org/papers/IJEDR1804011.pdf

[10] https://cs.stanford.edu/people/karpathy/cvpr2015.pdf

[11] https://www.ijedr.org/papers/IJEDR1804011.pdf

[12] https://www.ijert.org/image-captioning-using-deep-learning

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 07 Issue 05
May 2020

[13] https://towardsdatascience.com/image-captioning-with-keras-teaching-computers-to-describe-pictures-c88a46a311b8

[14] https://www.ripublication.com/ijaer18/ijaerv13n9_102.pdf

[15] https://arxiv.org/abs/1512.00567