IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Image Caption Generator Using CNN

,K. Praveen Kumar[1], V. Prakash Reddy[2] G. Indra Karan Reddy[3,] N.S. Ganesh[4]

SreeNidhi Institute of Science and Technology, Hyderabad

[1,2,3] UG Students,[4]Assistant Professor Hyderabad, Telangana

Abstract: In this paper we are presenting an Image caption generator using CNN. Automation is everybody's dream whether it is to ease their efforts or to make machines learn how to do something. In this paper we are presenting a way to generate captions based on the information in an image. This image caption generator can take input image and return the contents in the image as a caption. A CNN network along with LSTM is used to extract features and make words one by one and make a sensible sentences' model is trained in a way that the model will be able to accurately extract the contents of a image as captions. The model is trained on a dataset containing all kinds of images and it is observed that the model can detect the context of the image accurately.

Keywords: Image, caption, convolutional neural network, Description, Artificial intelligence.

I. INTRODUCTION:

In With the rise of users in internet the number of images and videos that are being shared on the internet is increasing a lot. This data is usually unstructured and raw data which does not carry much information. Whether to extract and use the right data from the internet or to organize the files captioning plays a huge role [1]. The Data that is available on the internet can be extracted to make a future prediction on any trends or to analyze the behavior of humans which can portray a wide variety of interests and demands. In the recent days with the rise of increasing users in internet everything can be accessed with text and images which can be searched using googles image search [1].

In this Project we will be predicting the contents of an image by words one by one and make a sensible sentence to be able to describe this image. A network of CNN and LSTM is used in this project. CNN is used to extract features and LSTM to store the words one by one and make a sentence. The caption should not only be able to describe the object but also make a sensible sentence which describe the action that's going on in that image [2].Most of the previous methods are based on indexing and labelling an image and categorizing them which can be a huge waste of human efforts when being compared to the automatic generation through deep learning .However simple words in any language can be feeded to the model to be able to predict the words of the objects In an image.

In order to get the caption of the image We will be designing a single model which takes an input image and then extract features and for the text preprocessing we will be adding start of series and an end of sequence to determine the starting and ending of caption. Then make a Dictionary of words in that image and use those words to describe the image and for sentence creation the first word that is predicted will be stored in LSTM and any further words predicted will be added one by one until the end of sequence appears.

The paper proposes a model which generates sensible captions for the images. The dataset used in this project contains 8091 images and 5 captions for a single input image. We are going to use ResNet 50 which consists of 50 hidden layers and can be able to classify 1million types of images and it is acting as an encoder. LSTM is used to store words one by one.

1.A child in a pink dress is climbing up a set of stairs in an entry way . 2. A girl going into a wooden building . 3. A little girl climbing into a wooden playhouse . 4.A little girl climbing the stairs to her playhouse . 5. A little girl in a pink dress going into a wooden cabin .

Fig.1 Example of input

II. APPROACH:

In this paper we are purposing a model which contains a neural network of CNN and LSTM together. CNN is used for image preprocessing also called as encoding and LSTM will be generating the sentences. All these neural networks can be mathematically represented fir better understanding and easier analysis.

A. *Convolutional Neural Network:*

Convolutional neural networks are best suited for image processing and visual analysis. The reason they are used is that the image can be easily manipulated by convoluting image data with a filter data for easier manipulations. CNN contains multiple layers, layers after CNN are connected as multilayer neural networks. The design of CNN allows it to be able to take a 2D image as an input. The output can be achieved by using multiple layers and weights just like in a neural network by convoluting one layer to another for manipulations and with various pooling techniques to get the feature vector.

In this Model we used ResNet50 which consists of 50 hidden layers for its model, which is a deep learning CNN model. This model is a pretrained model capable of categorizing 1million types of images. This model is used for feature extraction which can be further utilized for training.

Preprocessing using CNN:

Preprocessing of data implies that the input data is converted in a form that the computer can understand that easy training. Preprocessing is used to improve the performance by manipulating the image so that it can be easily understood by the machine. As ResNet50 accepts only 224 x 224 images it is important to first resize the images so it can be fed in to ResNet50.This reduction in size greatly enhances the performance as it has less data to work on.

Filtering:

In CNN, the main application of Convolution is done in filtering. In filtering the input image vectors are multiplied with the filter to get the modified output that we need. There are several types of filters such as sharpening, grayscale, blur and etc.

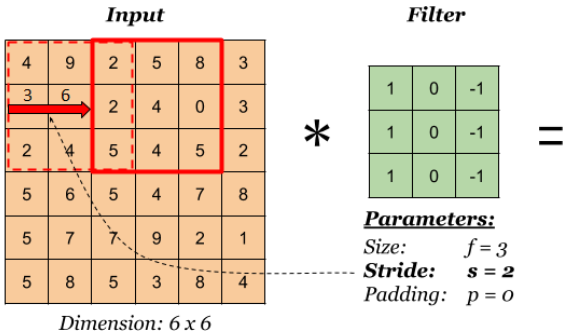

Fig 2: Filter

Feature extraction:

Just like the above filters CNN filters are used to extract features from an image, when compared to the usual filters CNN filters does not have any predefined values and these values are determined during the training period this helps the model to make filters by its own which can result in pretty amusing filters that humans can never think of manually. A 2D convolution filter is commonly used and it is referred as Conv2D.This filter adds up all the inputs and a single output is obtained from the image.

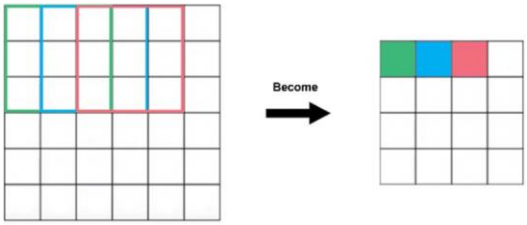

Fig 3. Feature Extraction

B. LSTM (Long -short term memory):

The transitory dynamics in a set of things are modelled by using a recurrent neural network [17]. It is exceedingly difficult for ordinary RNN to acquire long term dynamics as they get vanished and exploding weights or gradients [9]. The memory cell is the heart of LSTM. It is designed to store present value for a long amount of time. There are gates which are used for updating time Constants. Variants are represented by number of connections between the memory cells.

The LSTM used in this model is having the following memory and gates arranged in the manner of:

il = σ(Wixxl + Wimml-1) (1)

fl = σ(Wfxxl + Wfmml-1) (2)

ol = σ(Woxxl + Womml-1) (3)

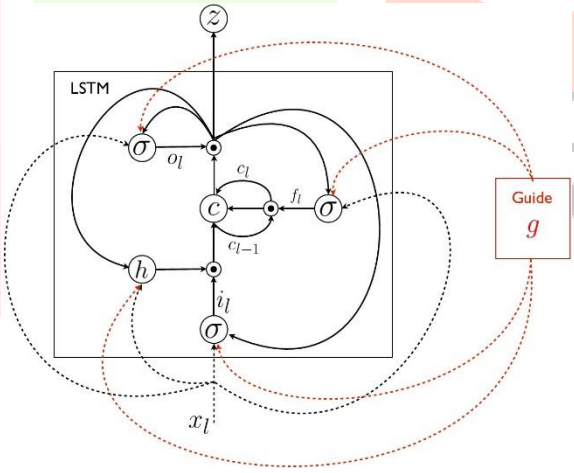

Fig 4. LSTM connections [9]

Where denotes multiplication performed elementwise, σ(.) denotes the sigmoid function and ~(.) denotes the hyperbolic tangent function. The variables il indicates the input gate, fl indicates the forget gate, ol indicates the output gate in the LSTM cell, cl indicates the state of the memory cell unit, and ml indicates the hidden state, and is the output of the block produced after processing in the LSTM, xl indicates a sequential parameter at time step l, and variable W[.][.] denotes model parameters. The loss function is represented by Eq. 6, with St denoting the created sentence at period t with regard to all LSTM and word embedding settings, this loss is always reduced.

III. GENERATION OF SENTENCE WITH CNN:

LSTM The idea of encoder decoder in network and machine translation modelling [1], [11]–[13], [16] is used to represent phrase creation in neural networks. The encoder is used to convert a variable sequence of words in natural language to a distributed vector in this paradigm. Then, based on the mapped vectors, a new sequence of words is created using a decoder in natural target language. The goal of the training procedure is to increase the likelihood of correct translation such that the phrase is in the native source language. When this approach is used to caption generation, the goal is to maximize the quantity of image captions created for a given image, namely.

$$\arg\theta \; X \; i \log(p(s1{:}Li \,|x\, i\, , \theta)) \quad (7)$$

where xi denotes an image, s1: Li represents group of words in properly formed sentence of length Li and θ represents model parameters. For ease of implementation, in the next step we ignore the superscript i whenever it is not significant or cleared from the context. As a sequence of words create each sentence, the Bayes chain rule is used to divide sentence which consists of words as its basic element.

$$\log (p\ (s1{:}\ L|x, \theta)) = \log (p(s1|x, \theta))+X\ l=2\ \log(p(s1|x, s1{:}l\text{-}1, \theta)) \quad (8)$$

where s1:L represents the block from sentence generated up to the l-th word. In whole training process, to maximize the purpose in Eq. 7, we have defined the log-likelihood.
log (p(s1:Li |x i , θ)) , it can be used with the hidden state in RNN.

At timestep l+1 the probability distribution of word for the whole vocabulary can be calculated with the help of SoftMax function z(.) which is based on output ml of the memory cell, pl+1 = z (ml) like [1]. Images and sentences are encoded as fixed-length vectors before using them as inputs to LSTM. First, for each image, CNN features are computed and then they are mapped to the embedding matrix. A new sequence is generated by concatenating sequence of words and an image in a sentence. In this new sequence, image is considered as beginning symbol of sequence and the sequence of words is treated as the remaining part of new sequence. This new sequence is used as an input to the LSTM network for training purpose by iterating the recurrence connection for l from 1 to Li. The transfer matrix which is linear in nature for image features, word embedding matrix and some arguments of LSTM are parameters of neural model. The image caption model has three sub models, first one is image model which repeats the image feature vector 28 times having dimension 28 x 4096 here 28 represents the maximum number of words in a caption. The second one is language model consisting of single LSTM unit and outputs the matrix having dimension 28 x 256, 256 is the output size of LSTM unit and the final model merge these two vectors and pass it to another LSTM unit having output dimension 28 x 915. For training we pass same encoded text vector as target vector but while testing we just encode" sol" to feature vector along with test image feature vector and we get matrix of dimension 28 x 915, and we decode that matrix into sequence words.

IV. RESULT:

A. *Dataset:*

The data set consists of 8091 images and for every individual image there are 5 different versions of captions that describe the image.

B. *Results:*

The model is trained for 500 epochs and was able to get a loss of 0.235 which is pretty good and was able to get 91% accuracy.

Fig. Results

ACKNOWLEDGEMENT:

The authors would like to express sincere gratitude to the management of SreeNidhi Institute of Science and Technologyfor their continuous support and encouragementin this work.

REFERENCES:

[1] Yao, Benjamin Z., et al. ”I2t: Image parsing to text description.” Proceedings of the IEEE 98.8 (2010): 1485-1508.

[2] Amritkar, C., & Jabade, V. (2018). Image Caption Generation Using Deep Learning Technique.

[3] Zahisham, Z., Lee, C. P., & Lim, K. M. (2020). Food Recognition with ResNet-50. 2020

[4] Vinyals, Oriol, et al. ”Show and tell: A neural image caption generator.” Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on. IEEE, 2015.

[5] Gerber, Ralf, and N-H. Nagel. ”Knowledge representation for the generation of quantified natural language descriptions of vehicle traffic in image sequences.” Image Processing, 1996. Proceedings., International Conference on. Vol. 2. IEEE, 1996.

[6] Yao, Benjamin Z., et al. ”I2t: Image parsing to text description.” Proceedings of the IEEE 98.8 (2010): 1485-1508

[7] Farhadi, Ali, et al. ”Every picture tells a story: Generating sentences from images.” Euro-pean conference on computer vision. Springer, Berlin, Heidelberg, 2010

[8] Yang, Yezhou, et al. ”Corpus-guided sentence generation of natural images.” Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011.

[9] Kulkarni, Girish, et al. ”Babytalk: Understanding and generating simple image descriptions.” IEEE Transactions on Pattern Analysis and Machine Intelligence 35.12 (2013): 2891-2903.

[10] Mitchell, Margaret, et al. ”Midge: Generating image descriptions from computer vision de-tections.” Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2012.

[11] Kuznetsova, Polina, et al. ”Collective generation of natural image descriptions.” Proceed-ings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. Association for Computational Linguistics, 2012.

[12] Jia, Xu, et al. ”Guiding long-short term memory for image caption generation.” arXiv pre-print arXiv:1509.04942 (2015).

[13] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. ”Neural machine translation by jointly learning to align and translate.” arXiv preprint arXiv:1409.0473 (2014).

[14] Mao, Junhua, et al. ”Deep captioning with multimodal recurrent neural networks (m-rnn).” arXiv preprint arXiv:1412.6632 (2014).

[15] Karpathy, Andrej, and Li Fei-Fei. ”Deep visual-semantic alignments for generating image descriptions.” Proceedings of the IEEE conference on computer vision and pattern recog-nition. 2015.

[16] Xu, Kelvin, et al. ”Show, attend and tell: Neural image caption generation with visual at-tention.” International Conference on Machine Learning. 2015.

[17] El Housseini, Ali, Abdelmalek Toumi, and Ali Khenchaf. ”Deep Learning for target recognition from SAR images.” Detection Systems Architectures and Technologies (DAT), Seminar on. IEEE, 2017.

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

[19] Donahue, Jeffrey, et al. ”Long-term recurrent convolutional networks for visual recogni-tion and description.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[20] Lu, Jiasen, et al. ”Knowing when to look: Adaptive attention via a visual sentinel for im-age captioning.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Vol. 6. 2017.

[21] Ordonez, Vicente, Girish Kulkarni, and Tamara L. Berg. ”Im2text: Describing images us-ing 1 million captioned photographs.” Advances in neural information processing systems. 2011.

[22] Chen, Xinlei, and C. Lawrence Zitnick. ”Mind’s eye: A recurrent visual representation for image caption generation.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[23] Feng, Yansong, and Mirella Lapata. ”How many words is a picture worth? automatic caption generation for news images.” Proceedings of the 48th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2010.

[24] Rashtchian, Cyrus, et al. ”Collecting image annotations using Amazon’s Mechanical Turk.” Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk. Association for Computational Linguistics, 2010.