© 2017 IJSRST | Volume 3 | Issue 7 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology

Image Description Using Deep Neural Network

Akanksha P. Deshmukh[1], Dr. A. S. Ghotkar[2]
[1]PG student, Dept. of Computer Engineering, Pune Institute of Computer Technology, Pune, India
[2]Associate Professor, Dept. of Computer Engineering, Pune Institute of Computer Technology, Pune, India

ABSTRACT

Recent research in computer vision and machine learning has demonstrated some great abilities at detecting and recognizing objects in natural images. Image description is a good starting point for imparting artificial intelligence to machines by allowing them to analyze and describe complex visual scenes. Computer software recently become smart enough to recognize objects in pictures, but not finding exactly what activities happening inside pictures. So, there is a need to develop system that can generate natural language descriptions from images. Such system can be useful for childhood education, image retrieval and visually impaired people. Automatic description from image is a challenging problem that contains interest from the domain like computer vision and natural language processing. The vision based image description system uses deep learning Convolution Neural Network and Recurrent Neural Network for generating description of images. As a result, Neural Network shows better result for description of images with increasing Bilingual Evaluation Understudy (BLEU) score of 0.64, Consensus-based Image Description Evaluation (CIDEr) score of 0.72 and minimizes validation loss to 2.5.

Keywords : Natural Language Processing, Neural Network, Torch , Convolution Neural Network, Recurrent Neural Network.

I. INTRODUCTION

Image description is description of the visual features of the contents in images. They describe elementary characteristics such as the shape, the color, the texture or the motion. It is sufficient for a human to point out and describes large amount of details about visual description. It requires identifying and detecting objects, people, scenes etc., reasoning about spatial relationships and properties of objects, combining several sources of information into a equivalent sentence. Hence it is a complex task to define an image or a scene; which is an important problem in the field of computer vision. Even though it is a challenging one, a lot of research is going on which explores the capability of computer vision in the field of image processing and it helps to narrow the gap between the computer and the human beings on scene understanding.

Computer Vision task includes processing, acquiring, analysing and understanding a digital image which deal with extraction of high dimensional data from real world in order to produce symbolical information. Natural language generation constitutes one of the fundamental research problems in natural language processing (NLP) and is core to a wide range of NLP applications such as machine translation, summarizing, dialogue systems and machine assisted revision. Connecting visual imagery with visually descriptive language is a challenge for computer vision that is becoming more relevant as recognition and detection methods are beginning to work. Studying such language has the potential to provide: training data for understanding how people describe the world and general knowledge about the visual world implicitly encoded in human language.

Natural language generation still remains an open research problem. Most previous work in NLP on automatically generating captions or descriptions for images is based on retrieval and summarizing. Obtaining sentence level descriptions for images is becoming an important task and has many applications, such as early childhood education, image retrieval and navigation for the blind.

Recent research in deep learning have inspired works which discuss a deep learning based approach inspired by recent advances in the applications of Convolution deep neural networks and recurrent neural networks. To reduce the training time required for the Neural Image

Captioning as well as integrate the decoder part into the network, while applying the convolution part to adapt to the dataset. The encoder part of NIC consists of a Convolution Neural Network (CNN) called GoogLeNet. Thus, in order to cut down on the training time, we tried to adapt the size of the network to the dataset by evaluating its performance on the dataset with multiple size.

II. Related Work

A. Convolution Neural Network

Convolutional Neural Networks (CNN) are biologically-inspired variants of Multi Layered Perceptrons. It is comprised of one or more convolution layers (often with a subsampling step) and then followed by one or more fully connected layers as in a standard multilayer neural network. The architecture of a CNN is designed to take advantage of the 2D structure of an input image. This is achieved with local connections and tied weights followed by some form of pooling which results in translation invariant features. Another benefit of CNNs is that they are easier to train and have many fewer parameters than fully connected networks with the same number of hidden units. CNN have been widely used and studied for image tasks, and are currently state-of-the art for object recognition and detection[4]. Several statistical measures are used for performance evaluation - An image using the 1024 × 1 final layer of VGG , denoted as g(I) for an image I. We train a linear transformation of g(I) that maps it into the 512 × 1 input dimensions expected by our LSTM network.

$$CNN(I) = W(I)\, g(I) + b(I)$$

B. Recurrent Neural Network

Recurrent neural networks (RNN) are quite popular for text generation, and so many researchers use them in this task, albeit in different settings Karpathy and Fei-Fei[3], Vinyals et al [4] are influenced by modern ANN based machine translation systems, and they employ a encoder decoder type architecture for their model. Recurrent Neural Networks (RNNs) are models that have shown great promise in many NLP tasks. The concept of RNNs is to make use of sequential information. In a traditional neural network we assume that all inputs (and outputs) are independent of each other. But for many tasks thats not effective. If you want to predict the next word in a sentence you have to know which words came before it. RNNs are called recurrent because they perform the same task for every element of a sequence, with the output being depended on the previous computations. Alternatively RNNs can be thought of as networks that have a memory which captures information about what has been calculated so far. In theory RNNs can make use of information in arbitrarily long sequences, but in practice they are limited to looking back only a few steps. RNN being unfolded into a full network. By unrolling we mean that we write out the network for the complete sequence. For example, if the sequence we care about is a sentence of 5 words, the network would be unrolled into a 5-layer neural network, one layer for each word[4]. LSTM defines a more complex memory cell at each time step. Each memory cell con- tains an internal state c t that stores information about inputs up to time LSTM also has three types of gates (input gate i t , forget gate f t , output gate o t) that control how information enters and leaves each cell. The input gate i t controls the degree to which LSTM will allow the current input x t to influence the hidden state h t . The forget gate f t modulates the influence of previous hidden state h t1 to current hidden state h t (i.e. how much to forget about previous hidden state). The output gate o t controls how much information is transferred from the memory cell to the hidden state at current time. Specifically, the hidden state h t in a LSTM model is computed as follows:

$$i_t = \sigma(W_{ix}\, x_t + W_{ih}\, h_{t-1})$$
$$f_t = \sigma(W_{fx}\, x_t + W_{fh}\, h_{t-1})$$
$$o_t = \sigma(W_{ox}\, x_t + W_{oh}\, h_{t-1})$$

In contrast, other recent work has focused more on the visual recognition aspect by detecting content elements (e.g. scenes,objects,attributes, actions, etc) and then composing descriptions from scratch Kulkarni et al.[2], Yang et al.[7], Li et al. [6], or by retrieving existing whole descriptions from visually similar images Ordonez et al.[16]. For the latter approaches, it is unrealistic to expect that there will always exist a single complete description for retrieval that is pertinent to a given query image. For the former approaches, visual recognition first generates an intermediate representation of image content using a set of English words, then language generation constructs a full description by adding function words an optionally applying simple re-ordering. Because the generation process sticks relatively closely to the recognized content, the resulting descriptions often lack the kind of coverage, creativity, and complexity typically found in human-written text.

Sentences are richer than lists of words, because they describe activities,properties of objects, and relations between entities (among other things). Such relations are revealing: Gupta and Davis show that respecting likely spatial relations between objects markedly improves the accuracy of both annotation and placing [7]. Li and Fei-Fei show that event recognition is improved by explicit inference on a generative model representing the scene in which the event occurs and also the objects in the image [8]. Using a different generative model, Li and Fei-Fei demonstrate that relations improve object labels, scene labels and segmentation [9]. Gupta and Davis show that respecting relations between objects and actions improve recognition of each [10, 11]. Yao and Fei-Fei use the fact that objects and human poses are coupled and show that recognizing one helps the recognition of the other [12]. Relations between words in annotating sentences can reveal image structure. Berg et al. show that word features suggest which names in a caption are depicted in the attached picture, and that this improves the accuracy of links between names and faces [13].

Recent research in deep learning have inspired works which discuss a deep learning based approach inspired by recent advances in the applications of Convolutional deep neural networks and recurrent neural networks [5][6]. Another paper that uses a similar technique written at about the same time is Long-Term Recurrent Convolutional Network (LRCN)[7]. These two works have invoked our interest. To reduce the training time required for the Neural Image Captioning (NIC) algorithm proposed in [4] as well as integrate the decoder part specified in [7], into the network, J. Donahue and L. A. Hendricks[12], describes a new approach to the problem of image caption generation, casted into the framework of encoder-decoder models. For the encoder, we learn a joint image-sentence embedding where sentences are encoded using long short-term memory (LSTM) recurrent neural networks.

III. Proposed Methodology

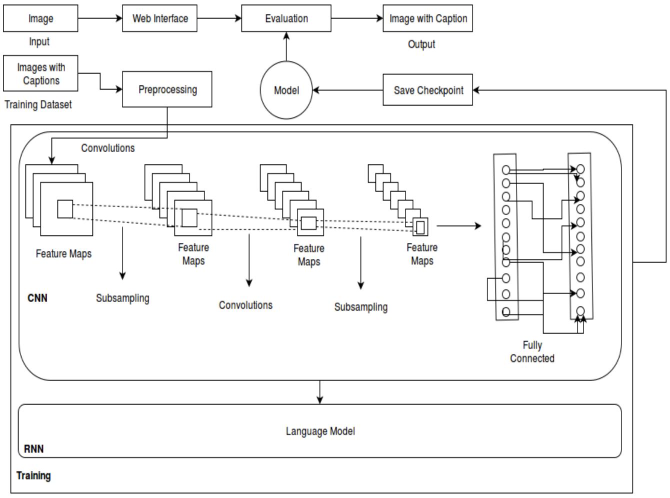


Figure 1 : Architecture Design Of Proposed System

Figure 1 shows architectural design of proposed system. Following are important components in the system :

Web Interface : User can upload an image through web interface to predict sentence description of upload image.

Pre-Processing : It takes JSON of the form of image with caption as an input. It does some basic Pre-Processing on the captions , creates a special UNK token, and encodes everything to arrays. It produces JSON and hdf5 file as an output. It has a dictionary that contains: an ’ix-to-word’ field storing the vocabulary in form ix:’word’, where ix is 1-indexed an ’images’ field that is a list holding auxiliary information for each image. It contains several fields: Images which are (N,3,256,256) unsigned int 8 array of raw image data in RGB format and labels is (M,maxlength) unsigned int 32 array of encoded labels, zero padded label-start-ix and label-end-ix are unsigned int 32 arrays of pointers to the first and last indices (in range 1..M) of labels for each image label-length stores the length of the sequence for each of the M sequences.

Training : JSON and hdf5 file generated by Pre-Processing are given as an input to training phase. This phase also requires VGG 16-layer network, Using the VGGNet , we transform the pixels inside an image to a 4096-dimensional vector. After getting the visual features, training an LSTM to obtain linguistic captions. At last we fine tune the pre-trained model to get a more suitable and save checkpoint to model for the natural language caption generation task.

Evaluation : Evaluation phase uses model which is generated after training phase and it predicts image description for each image which is uploaded by user.

IV. Experiment and Results

Experiment is performed such that in the training set, there are 414113 captions in total, for an average of 5.002 captions per image. We preprocess the caption dataset by replacing words that appear less than five times in the training dataset with an UNKs token, prepend each sentence with a SOS token, and append each sentence with a EOS token. The mean and median length of the post-processed captions is 12.55 and 12 respectively. The following Table 1 shows the output of preprocessing:

Table 1: Preprocessing Output

Total number of Words	6447836
Number of bad words	67.71%
Number of words in vocabulary	9566
Number of UNKs	0.54%
Max length sentence in raw data	49

C. *Dataset*

MS COCO is a large image dataset designed for object detection, segmentation and caption generation. The Microsoft COCO dataset contains 82,783 training images and 40,504 validation images, each With 5 human generated descriptions. We used the training set and validation set to train our model in our experiments and uploaded our generated captions on the testing set (40,775 images) to the COCO server for evaluation.

D. *Results and Discussion*

The BLUE score for each iterations and BLUE-1, BLUE-2, BLUE -3 and BLUE-4 gives the 4 reference sentence for each image given by for getting better result. BLEU (bilingual evaluation understudy) is an algorithm for evaluating the quality of text which has been machine translated from one natural language to another. It compute the geometric average of the modified n-gram precision, Pn using n-grams up to length N and positive weights Wn summing to one. Next, let c be the length of the candidate translation and r be the effective reference corpus length. It compute the brevity penalty BP as,

$$BP = \{ 1 \text{ if } (c > r) \text{ OR } e^{(1 - rc)} \text{ if } (c < r) \}$$

Then,

$$BLEU = BP * \exp\left(\sum_{n=1}^{N} w_n \, log p_n \right)$$

The graph in Figure 2 shows performance metric BLEU-1, BLEU-2, BLEU-3, BLEU-4 score which calculates score for 4 reference sentence at each iteration. Calculation for BLEU score is described in Section 1.6.2. The graph shows that BLEU-1 gives higher score of 0.64, therefore the first reference sentence matches with image description. While BLEU-2, BLEU-3, BLEU-4 also increases with each iteration but not higher than BLEU-1. The figure 1 shows graph of performance metric BLEU score which calculate score for 4 reference sentence at each iterations. It shows that BLEU-1 score which gives approximate matching with sentence generated by our system.

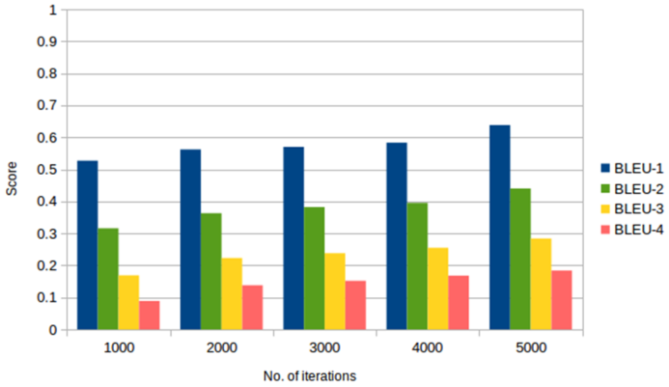
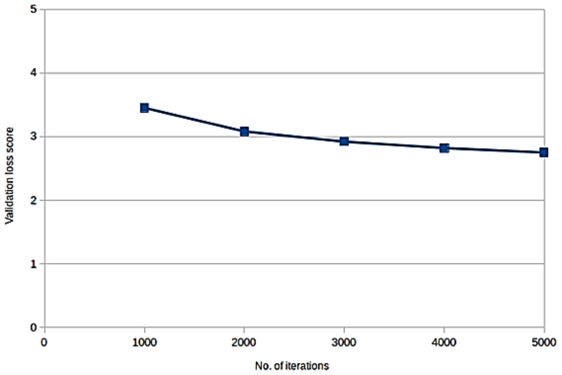


Figure 2 : BLEU_Score



Loss is a summation of the errors made for each example in training or validation sets. In the case of neural networks the loss is usually negative log-likelihood and residual sum of squares for classification and regression respectively. The main objective in a learning model is to reduce (minimize) the loss function's value with respect to the model's parameters by changing the weight vector values through different optimization methods.

The parameters of the model at each iteration calculate using the cross entropy loss of the predictions on each sentence. The loss function minimized as:

J(S|I; θ) = $\sum_{t=1}^{N} log p_t$ J(S_t|I; θ)where $p_t(S_t)$ is the probability of observing the correct word S_t at time t. This loss is minimized with regards to parameters in the set θ, which are all the parameters of the LSTM above, the parameters of the CNN and the word embeddings. The figure 10.3 shows graph of validation loss. It shows how the loss is reduced at each iteration. For better accuracy of system, validation loss should be minimum.

V. Comparative Analysis

The following Table shows analysis with existing research by O. Vinyals[4] and J. H. Mao[7] uses Flickr30k dataset which contains 30,000 images and achieved BLEU score of 0.66 and 0.60 respectively. J. Donahue[12] and A. Karapathy [3] uses MSCOCO dataset which contains 82,783 images and achieved BLUE score of 0.62. In our work, we have used MSCOCO dataset and achieved BLUE score of 0.64.

Sr. No.	Author	Dataset	BLEU Measure
1.	O. Vinyals et al. [4]	Flickr30k	0.66
2.	J. Donahue et al. [12]	MSCOCO	0.62
3.	J. H. Mao et al. [7]	Flickr30k	0.60
4.	A.Karapathy et al. [3]	MSCOCO	0.62
5.	Our Work	MSCOCO	0.64

VI. Conclusion and Future Work

Proposed system uses Convolution Neural Network for extracting features from an image and encodes an image into a compact representation, followed by a Recurrent Neural Network that generates a corresponding sentence. The model is trained using MSCOCO dataset that contains 82,783 images to maximize the likelihood of the sentence. As a result, proposed system gives BLEU score of 0.64, CIDEr score of 0.72 and minimizes validation loss to 2.5. The score increases and validation loss decreases with each iteration. Vision based image description system generates only description for images. Thus, research work can be extend to explore the description of videos. The proposed system can also be extend for GIF images.

VII. REFERENCES

[1]. Anurag Kishore and Sanjay Singh, "Natural langauage image descriptor", *IEEE Recent Advances in Intelligent Computational Systems (RAICS),* pp. 10-12, December, 2015.

[2]. G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L.Berg,"Baby talk: Understanding and generating image descriptions", *IEEE Transaction On Pattern Analysis And Machine Intelligence*, Vol. 35, NO. 12, December, 2013.

[3]. Andrej Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions",*CVPR*, March, 2015

[4]. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan "Show and tell: A neural image caption generator", *arXiv preprint arXiv:1411.4555,* March, 2014.

[5]. Frome, Andrea, Greg S. Corrado, Jon Shlens, Samy Bengio, Jeff Dean, and Tomas Mikolov, "Devise: A deep visual-semantic embedding

model", *InAdvances in Neural Information Processing Systems*, pp. 2121-2129, 2013.

[6]. Andrej karapathy, Armand Joulin, and Li Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping", *arXiv preprnt arXiv:1406.5679,* 2014.

[7]. J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)", *ICLR,* 2015.

[8]. Polina Kuznetsova, Vicente Ordonez, Alexander C. Berg,Tamara L. Berg and Yejin Choi, "Collective Generation of Natural Image Descriptions".

[9]. X. Chen and C. L. Zitnick, "Learning a recurrent visual representation for image caption generation", *arXiv preprint arXiv:1411.5654*, 2014.

[10]. J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach,S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description", *CoRR,vol. abs/1411.4389*, 2014, [Online]. Available: http://arxiv.org/abs/1411.4389.

[11]. K. Cho, B. van Merrienboer, C. Gulcehre, F. Bougares,H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation", *In EMNLP*, 2014.

[12]. J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition", *In ICML*, 2014.

[13]. M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics", *Journal of Artificial Intelligence Research(JAIR)*, Vol. 47, pp. 853899, 2013.

[14]. Mitchell, Margaret, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daum 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*, pp. 747-756, 2012.

[15]. Kiela, Douwe, and Lon Bottou, "Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics", *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),* pp. 36-45, 2014.

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

[17]. K. Papineni, S. Roukos, T. Ward, and W. Jing Zhu, "Bleu: A Method for Automatic Evaluation of Machine Translation", *Proc. 40th Ann. Meeting of Assoc. for Computational*, pp. 311-318, 2012.

[18]. P. Kuznetsova, V. Ordonez, T. Berg, and Y. Choi,"Treetalk: Composition and compression of trees for image descriptions", *Trans. of the Association for Computational Linguistics*, pp. 351362, 2014.