Image Caption Generation using Deep Learning Technique

Chetan Amritkar
Department of EnTC
Vishwakarma Institute of Technology
Pune, India
chetan.amritkar16@vit.edu

Vaishali Jabade
Department of EnTC
Vishwakarma Institute of Technology
Pune, India
vaishali.jabade@vit.edu

***Abstract*—In Artificial Intelligence (AI), the contents of an image are generated automatically which involves computer vision and NLP (Natural Language Processing). The neural model which is regenerative, is created. It depends on computer vision and machine translation. This model is used to generate natural sentences which eventually describes the image. This model consists of Convolutional Neural Network(CNN) as well as Recurrent Neural Network(RNN). The CNN is used for feature extraction from image and RNN is used for sentence generation. The model is trained in such a way that if input image is given to model it generates captions which nearly describes the image. The accuracy of model and smoothness or command of language model learns from image descriptions is tested on different datasets. These experiments show that model is frequently giving accurate descriptions for an input image.**

Keywords*—*Neural Network, Image, Caption, Description, Long Short Term memory(LSTM), Deep Learning.

I. Introduction

The people communicate through language, whether written or spoken. They often use this language to describe the visual world around them. Images, signs are another way of communication and understanding for the physically challenged people. The generation of descriptions from the image automatically in proper sentences is a very difficult and challenging task [1], but it can help and have a great impact on visually impaired people for better understanding of the description of images on the web. A good description of an image is often said for 'Visualizing a picture in the mind'. The creation of an image in mind can play a significant role in sentence generation. Also, human can describe the image after having a quick glance at it. The progress in achieving complex goals of human recognition will be done after studying existing natural image descriptions.

This task of automatically generating captions and describing the image is signifi-cantly harder than image classification and object recognition. The description of an image must involve not only the objects in the image, but also relation between the objects with their attributes and activities shown in images [20]. Most of the work done in visual recognition previously has concentrated to label images with already fixed classes or categories leading to the large progress in this field. Eventually, vocabularies of visual concepts which are closed, makes a suitable and simple model for assumption. These concepts appear widely limited after comparing them with the tremendous amount of thinking power which human possesses. However, the natural language like English should be used to express above semantic knowledge, that is, for visual understanding language model is necessary.

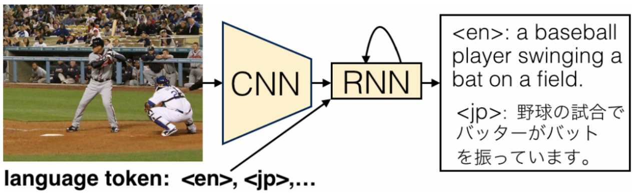
CNN
RNN
<en>: a baseball player swinging a bat on a field.
<jp>: 野球の試合でバッターがバットを振っています。
language token: <en>, <jp>,...

Fig. 1. Model based on Neural Networks

In order to generate description from an image, most of the previous attempts have suggested to combine all the current solutions of the above problem. Whereas, we will be designing a single model which takes an image as an input and is trained for producing a sequence of words where each word belongs to the dictionary that describes the image suitably as shown in Fig. 1.

The relation between visual importance and descriptions moves to the text summarization problem in natural language processing (NLP) [18]. The important goal of text summarization is selecting or generating an abstract for document. In problem of image captioning, for any image we would like to generate a caption which will describe various features of that image [21].

This paper proposes a model capable of generating novel descriptions from images. For this task, we have used Flickr 8k dataset consisting of 8000 images and five descriptions per image. The Fig. 2 illustrates the dataset structure in which an image is having five natural language captions. In this work, we are using CNN as well as RNN. Pre-trained Convolutional Neural Network (CNN) is used for the image classification task. This network acts as an image encoder. The last hidden layer is used as an input to Recurrent Neural Network (RNN). This network is a decoder which generates sentences. Sometimes, the generated sentence seems to lose track or predict wrong sentence than that of the original image content. This sentence is generated from description that is

978-1-5386-5257-2/18/$31.00 ©2018 IEEE

common in dataset and the sentence is weakly related to input image.

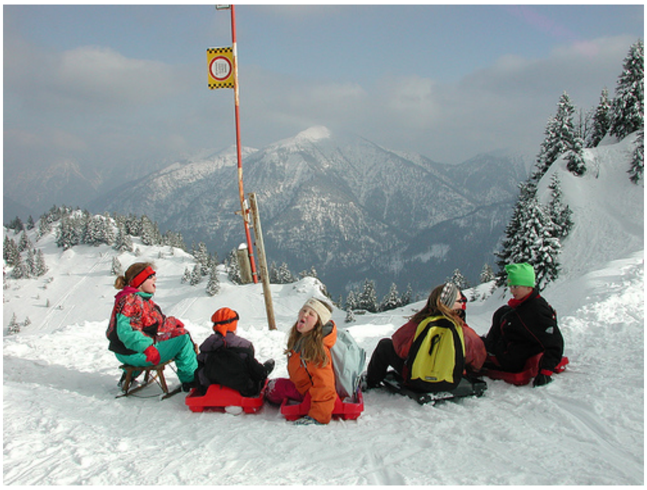

Fig. 2. Caption 1:A group is sitting around a snowy crevasse, Caption 2: A group of people sit atop a snowy mountain, Caption 3: A group of people sit in the snow overlooking a mountain scene, Caption 4: Five children getting ready to sled, Caption 5: Five people are sitting together in the snow.

II. Related Work

In computer vision, the problem of generating descriptions in natural language from visual data has long been studied [1]–[3]. The literature on image caption generation can be grouped into three categories. The first category consists of template based methods [4]–[7]. In this approach, the priority is given to detect objects, actions, scenes and attributes. The second category consists of transfer based caption generation methods [8]. In this approach, image retrieval is done. This approach fetches visually similar images and then the captions of these images are used for query image. Most of the researchers suggested that neural networks are useful in machine translation [10], use of neural language models for caption generation. The goal is to convert an image into sentence which explains it rather than translating a sentence from a source language into a required format.

This has made system more complex. They are made up of visual radical recognizers by using a formal language, eg. And-Or Graphs or logic systems, rule based systems are used for further conversion. Mao et al. [11] and Karpathy et al. [12] have suggested multimodal recurrent neural network model which is used for description generation of an image. Vinyals, Oriol, et al. [1] used NIC model (Neural Image Caption). In NIC model, the encoder used is CNN. The pretrained CNN is used for image classification and the last layer of network is used as input to RNN decoder. This RNN decoder further generate sentences. They have used LSTM [1], which is advanced type of RNN. Recently, Xu et al. [13] have suggested to summarize visual attention into the LSTM model for fixing its gaze on different objects during the process of generation of related words. Neural language models are useful in generating human-like image captions. Except for the very recent methods, most of them follows a similar encoding-decoding framework [13], which combines caption generation and visual attention.

This work related to the methods of third category of caption generation. In this approach, a neural model is designed which generates descriptions for image in natural language. CNN is used as image encoder. Firstly, pre-training is done for image classification task and then the RNN decoder uses this last hidden layer as input to generate the sentence.

III. Approach

In this work, neural framework is proposed for generating captions from images which are basically derived from probability theory. By using a powerful mathematical model, it is possible to achieve better results, which maximizes the probability of the correct translation for both inference and training.

A. *Convolutional Neural Network (CNN)*

The convolutional networks are currently used in visual recognition. There are number of convolutional layers in CNN. After these convolutional layers, next layers are fully connected layers as in multilayer neural network [14]. The CNN is designed in such a way that the benefit of 2D structure of input image can be taken. This target is accomplish with the help of number of local connections and tied weights along with various pooling techniques which result in translation invariant features. The main advantages of using CNN are ease of training and possessing less parameters as compared to other networks with equal number of hidden statesss.

For this work, we are using Visual Group Geometry(VGG) network, which is Deep CNN for large scale image recognition [15]. It is available in 16 layers as well as 19 layers. The classification error results for both 16 and 19 layers are almost same for validation set as well as test set, which is around 7.4% and 7.3%. This model gives the features of images which are used in further process of caption generation.

B. *Long Short-Term Memory (LSTM)*

The transitory dynamics in a set of things are modelled by using a recurrent neural network [17]. It is very difficult for ordinary RNN to acquire long term dynamics as they get vanished and exploding weights or gradients [9]. The memory cell is main block of LSTM. It stores the present value for long period of time. Gates are there for controlling update time of state of cell. The number of connections between memory cell and gates represent variants.

Our model is based on the LSTM block which depends on the LSTM with no peephole architecture as shown in Fig. 3. The memory cell and gates of LSTM are having following relations:

$$i_{\mathrm{l}} = \sigma(W_{\mathrm{ix}} x_{\mathrm{l}} + W_{\mathrm{im}} m_{\mathrm{l\text{-}1}}) \tag{1}$$

$$f_{\mathrm{l}} = \sigma(W_{\mathrm{fx}} x_{\mathrm{l}} + W_{\mathrm{fm}} m_{\mathrm{l\text{-}1}}) \tag{2}$$

$$o_{\mathrm{l}} = \sigma(W_{\mathrm{ox}} x_{\mathrm{l}} + W_{\mathrm{om}} m_{\mathrm{l\text{-}1}}) \tag{3}$$

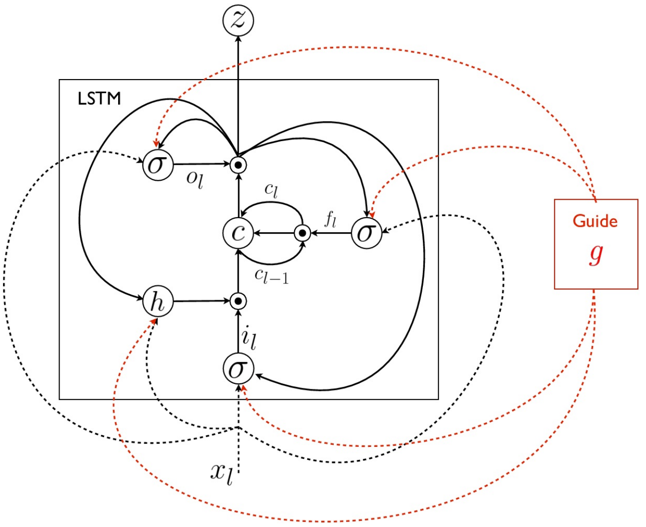

Fig. 3. Connection diagram of LSTM [9]

$$c_{\mathrm{l}} = f_{\mathrm{l}} \odot c_{\mathrm{l\text{-}1}} + i_{\mathrm{l}} \odot \hbar(W_{\mathrm{cx}} x_{\mathrm{l}} + W_{\mathrm{cm}} m_{\mathrm{l\text{-}1}}) \tag{4}$$

$$m_{\mathrm{l}} = o_{\mathrm{l}} \odot c_{\mathrm{l}} \tag{5}$$

$$L(I,S) = -\sum_{t=1}^{N} log(p_{\mathrm{t}}(S_{\mathrm{t}})) \tag{6}$$

where $\odot$ denotes multiplication peformed elementwise, $\sigma(.)$ denotes the sigmoid function and $\hbar(.)$ denotes the hyperbolic tangent function. The variable $\mathrm{i_l}$ represents the input gate, $\mathrm{f_l}$ represents the forget gate, $\mathrm{o_l}$ represents the output gate in the LSTM cell, $\mathrm{c_l}$ represents state of memory cell unit and $\mathrm{m_l}$ represents hidden state which is the output of the block generated after processing in LSTM, $\mathrm{x_l}$ represents a parameter of sequence at time step l and variable $\mathrm{W_{[.][.]}}$ denotes parameters of the model. Eq. 6 represents loss function with S_{t} representing generated sentence at time t. This loss is always minimized with respect to all parameters of LSTM and word embeddings.

IV. Generation Of Sentence With LSTM

The process of sentence generation in neural network is taken from principle of encoder decoder in modelling of network and machine translation [1], [11]–[13], [16]. In this modelling, variable sequence of words in natural language is mapped to distributed vector by using encoder. Then, a new sequence of words is generated by using decoder in natural target language depending on mapped vectors. In training process, the aim is to maximize chances of perfect translation such that the sentence is in natural source language. Applying this principle while generating the captions, the target is to maximize the amount of the image caption generated given an image, namely

$$arg_{\theta} \sum\nolimits_{i} log(p(s_{1:\mathrm{L_i}}|x^{\mathrm{i}}, \theta)) \tag{7}$$

where $\mathrm{x^i}$ denotes an image, $\mathrm{s_{1:L_i}}$ represents group of words in properly formed sentence of length $\mathrm{L_i}$ 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.

$$log(p(s_{1:\mathrm{L}}|x,\theta)) = log(p(s_1|x,\theta)) + \sum\nolimits_{l=2} log(p(s_1|x, s_{1:l\text{-}1}, \theta)) \tag{8}$$

where $\mathrm{s_{1:L}}$ 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 $\mathrm{log(p(s_{1:L_i}|x^i}, \theta))$, 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, $p_{\mathrm{l+1}}$ = z ($\mathrm{m_l}$) similar to [1].

Images and sentences are encoded as fixed-length vectors before using them as inputs to LSTM. Frst of all for each images, 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 coonsidered 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 $\mathrm{L^i}$. 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.

V. Results

A. Datasets

These datasets consist of images and description of image in the form of senetences in natural language such as English. The statistics of datasets are as shown in Table I.

In these datasets, each image is described by observers with 5 different sentences that are relatively visible and impartial.

B. Results

The model has been trained for 50 epochs. As number of epochs used are more, it helps to lower the loss to 3.74. If we

TABLE I
DATASET STATISTICS

Dataset Name	Size		
	Train	***Valid***	***Test***
Flickr8k [1]	6000	1000	1000
Flickr30k [1]	28000	1000	1000
MSCOCO [1]	82783	40504	40775

consider the large dataset then we should use more epochs for accurate results.

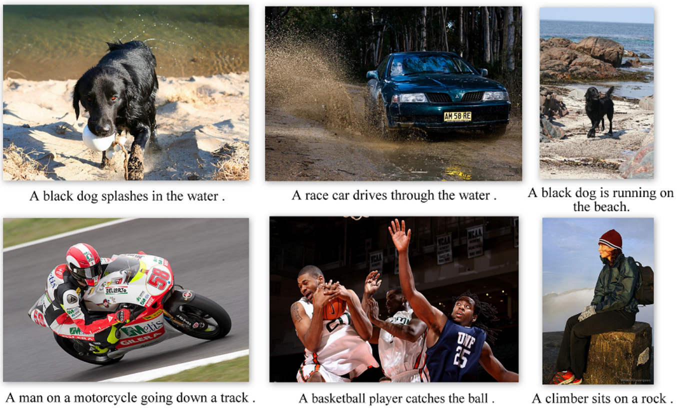
A black dog splashes in the water .
A race car drives through the water .
A black dog is running on the beach.
A man on a motorcycle going down a track .
A basketball player catches the ball .
A climber sits on a rock .

Fig. 4. Selection of Evaluation Results

Some results generated are as shown in Fig. 4. By using the Flickr8k dataset for training model and running test on the 1000 test images available in dataset results in BLEU = 0.53356. For Flickr30k dataset, running test on same number of test images available in dataset results in BLEU = 0.61433 and for MSCOCO dataset running test on images results in BLEU = 0.67257.

VI. CONCLUSION

This work presents a model, which is a neural network that can automatically view an image and generate appropriate captions in natural language like English. The model is trained to produce the sentence or description from given image. The descriptions or captions obtained from the model are categorized into:

- Description without errors
- Description with minor errors
- Description somewhat related to image
- Description unrelated to image

The categories in results are due to neighborhood of some particular words, i.e., for word like car it's neighborhood words like vehicle, van, cab etc. are also generated which might be incorrect. After so much of experiments, it is conclusive that use of larger datasets increases performance of the model. The larger dataset will increase accuracy as well as reduce losses. Also, it will be interesting that how unsupervised data for both images as well as text can be used for improving the image caption generation approaches.

REFERENCES

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

[2] 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.

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

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

[5] 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.

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

[7] 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.

[8] 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.

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

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

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

[12] 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.

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

[14] 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.

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

[16] 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.

[17] 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.

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

[19] 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.

[20] 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.

[21] 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.