Image Caption Generation Using CNN-LSTM Based Approach

Bineeshia J [1]
{jb.cse@psgtech.ac.in[1]}

Department of CSE, PSG College of Technology, Coimbatore, India [1]

Abstract. The Picture Caption Generator automatically represents the content of an image, which is a key problem in artificial intelligence that links computer vision with natural language processing (NLP). There is a growing necessity for context-based natural language image descriptions. Recent advances in domains such as neural networks, natural language processing and computer vision, have paved the road for better description of images. It needs both computer vision approaches to interpret the content of the image and a language algorithm from the NLP sector to transform the image's interpretation into words in right order. To accomplish this, state-of-the-art algorithms like as Convolutional Neural Network (CNN) and sufficient image datasets with human-judged descriptions are used. It produces a regenerative neural model that relies on machine translation and computer vision. The proposed model generates natural statements that describe the image. This model combines Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). The former is employed for feature extraction of images and the latter is employed for generating sentences. The training of the model is done in such a way that it produces captions that almost define the image when the input source image is offered to the model. 6000 images have been used for training purpose and is trained over 20 epochs to finally obtain a loss value of 2.6380. The loss has been reduced exponentially through the span of 20 epochs. BLEU score metric is finally calculated to measure the model's performance. Unigram, bigram, trigram and 4-gram precision is calculated.

Keywords: Image, caption, Long Short Term memory (LSTM), Recurrent Neural Network, Convolutional Neural Networks.

1 Introduction

Humans communicate with each other through language either spoken or written. To explain the visual world around them, humans also use language. Another means of communication and comprehension for the physically disabled persons are pictures and signs. It is a very challenging and a difficult job to produce explanations automatically from an image as a correct statement [6].It can also assist and show a significant impact on a visually disabled person to understand the definition of images found in web. To imagine a picture, a good definition of the picture is required. In sentence generation, the development of mind image has a crucial influence. Upon getting a fast look at it, humans can also describe the picture. After reviewing current natural image definitions, progress will be made in achieving complicated human recognition objectives. The challenge of creating image caption and defining it is substantially more difficult than classification of images and identification of

ICCAP 2021, December 07-08, Chennai, India
Copyright © 2021 EAI
DOI 10.4108/eai.7-12-2021.2314958

objects. The image definition must include the entities of the image, the relationship among the entities and the behavior's reflected [7].

Majority of the study in understanding an image has focused on marking images with groups or categories that are already fixed, leading to significant advances in this area. Eventually, the vocabulary of closed visual concepts provides an effective and clear assumption model. After contrasting them with the immense amount of thinking ability that human beings possess, these concepts become widely restricted. However, natural language such as English should be used to convey semantic information above, i.e. it is important for visual comprehension of the language model. Several early efforts proposed combine the available solution to solve this challenge and produces descriptions from an image. In comparison, we will build a model that accepts an input image and trains itself to generate series of output words, each of which accurately represents the image. The relationship between visual meaning and explanations in natural language processing shifts to the issue of text summarization. Selecting or creating an abstract for the paper is the essential purpose of text summarization.

In image captioning, we would like to create a caption for any image that will identify the different characteristics of that image. This research proposes a model for generating correct interpretation of images. The dataset from Flickr has been used for this study. We are using CNN as well as RNN in this study. For the image classification task, the model uses Pre-trained CNN. This network serves as an encoder of the image and is fed as an input to the Recurrent Neural Network (RNN) as the last hidden layer. This network is a sentence-generating decoder. Often it seems that the produced sentence loses track or predicts an incorrect text instead of actual information of the picture. This phrase is derived from a dataset-wide definition, and it is only vaguely related to the image input.

2 Related Work

Sapkal describes a method for generation of image caption that aims to include two variants of attention. In this case, CNN functions as encoder, extracting features from the input image known as convolution features. This model uses CNN with few parameters than fully connected networks with the same number of hidden units and hence training the model is a lot easier. The generated sentences produce sensible qualitative predictions because they perform the same task for every element of a sequence, with the output being dependent on the previous computations. A realistic difficulty lies in having a large dataset of picture captions accessible on the web, however the captions are inconsistent and the origins of the photos are unclear. The models' versatility is limited since they frequently depend on visual notions and text patterns.

An attention-based approach for detecting items in pictures was proposed by Jimmy et al [2]. The presented method is a RNN that works with vital areas of the source images. Although being trained solely with labels of class, the model is trained to find and identify a variety of items. It performs better even when there is Gaussian noise. The model learns to localize several objects as well as identify them. Computational complexity is high in this method.

Pranay et al. [3] proposed a predictive model built on recurrent network which integrates latest breakthroughs in machine translation and computer vision to produce texts explaining a picture. The network has been trained to maximize the likelihood of the desired summary of text given the image. The network that employs InceptionV4 and encoder surpasses GoogLeNet, and is fast. Descriptions are created in a highly efficient manner, which

speeds up inference and eliminates the costs of traversing the entire beam search tree. Only PNG and JPEG formats of images are supported by this model.

A neural network architecture called RNN Encoder-Decoder consisting of two RNN was proposed by Kyunghyun Cho et al[4]. Among the two RNN, one of them encodes a sequence of symbol into a representation of fixed length vector, while the other RNN decodes into another sequence of symbol. To maximize the conditional likelihood of the sequence of target given a source sequence, the encoder and decoder of the proposed model are jointly trained. A meaningful representation of linguistic phrases is learned qualitatively by the proposed model to increase the conditional likelihood of the sequence of target given a source sequence.

General pipeline architecture does not require a large data set for training and does not constrain the generated descriptions to a predefined set of semantic classes of scenes, objects, attributes, and actions. Retrieval in multimodal space allows bidirectional models. It is difficult to refer the objects that are not depicted. It is difficult in providing background knowledge that cannot be derived directly from the image. Lun Huang [5] proposed an attention model, namely Adaptive Attention Time (AAT).This model aligned sources and targets for captioning of images. AAT helps the structure to gain knowledge on how many steps of attention must be taken at each decoding stage to output a caption phrase. With AAT, a region in the image can be linked to arbitrary number of image descriptions, whereas an arbitrary number of image regions can also be protected by a caption word. AAT prevents adding noise over the gradients of the parameters. AAT is also general and can be used for any learning task that is sequence-to-sequence. We empirically demonstrate on the image captioning task that AAT improves over state-of-the-art methods. NLP and video captioning which includes computer translation and text description are not allowed.

3 Approach

In this paper, the deep learning concepts CNN and RNN are used for generating captions of an image. Both CNN and LSTM have been combined for extracting features from an image and caption generation of the input image.

A. Convolutional Neural Network (CNN)

CNN, or Convolutional Neural Networks, is a kind of neural network that uses convolutional layers to analyze spatial information. A convolution layer contains a variety of kernels that learns to draw out various feature types from the input image. The kernel is a two-dimensional filter that is slid across the input to perform convolution. CNN's are efficient for caption generation tasks as the convolution layers can extract features horizontally from left to right and vertically from top to bottom. For classification tasks, the features seem to be important because it is difficult to find clues about class memberships, especially when they occur in different orders in input.

CNN is like a 2D matrix, which analyzes the information. The representation of the images is done as a 2D matrix and CNN is very helpful in dealing with images. The images are analyzed, transformed, resized, and given perspective modifications. A CNN blends the learnt features with data given as input and utilizes convolutional 2D layers. Multiple kernels that traverse the image are used to calculate a dot product. From the image, every filter extracts multiple features. With learnt features from the input, convolution maintains a link connecting pixels. The layer which carries out maximum pooling aids to minimize the size of transformed characteristics and also aids to minimize over-fitting by giving an abstract

representation. Max-pooling operation extracts the most important feature such as object, colour for each convolution and helps reduce noise by discarding noisy activations. ReLU function is a widely used linear function. The function returns zero if negative and returns positive if it gives the image.

In layer where it is completely connected, the nodes of the input are associated to each node of the next layer. At the end of CNN, one or more completely linked layers are used. By addition of this layer it allows learning of combinations of non-linearity of the convolutional layers' higher level characteristics. The characteristics derived are transformed to a vector. Usually, the activation feature and dropout layer are used to minimize over-fitting between two layers and non-linearity. Dropout is used to minimize over-fitting in neural networks and it randomly makes few of the connections as zero. The image's features are extracted using CNN, and the knowledge from CNN is used by LSTM to help build an image representation.

B. LSTM

LSTM is artificial recurrent neural network architecture in the field of deep learning. RNN consists of neuron-like nodes which are connected to each other. Unlike traditional feed forward networks, LSTM incorporates feedback links. It can handle not only single information like pictures; it can also handle complete sequence of data such as video. Throughout the processing of inputs, LSTM may introduce relevant information and discard non-relevant information with a forgotten gate. The hidden unit of an RNN is its most significant characteristic.

LSTM networks, a type of RNN have the ability to remember information for a longer duration. The LSTM has the power, carefully controlled by structures called gates, which can delete information or add information to the cell. Gates are an optional way of allowing data to move through. They consist of a sigmoid layer and a point wise multiplicative process. The sigmoid layer produces values ranging from 0 to 1, indicating how much of every product should be allowed to pass. A zero value allows nothing to pass through, while one allows all of it to pass through. For monitoring and protecting the cell state, an LSTM has three of these gates. RNNs work very well with sequential data. By combining CNN and LSTM, a model that can understand pictures and using that knowledge to assist constructing a representation of these pictures can be developed. The LSTM architecture is depicted in Fig 1.

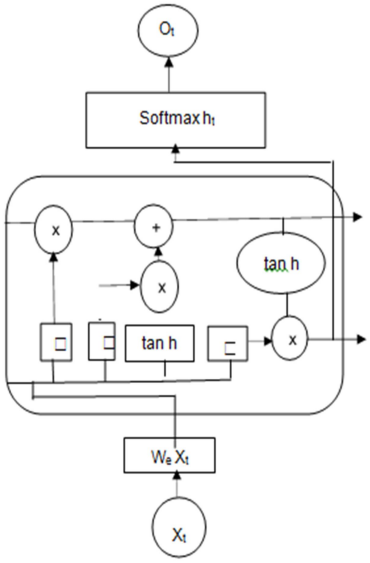

Fig. 1. LSTM Architecture

4 Result

A. Datasets

The dataset is Flikr8k_dataset and it consists of 8091 images. In the training data, there are a total of 6000 pictures. For training purposes, there are 6000 descriptions. The total vocabulary size of the unique words is 6889. The length of descriptions from the dictionary created using the tokenizer is 32..

B. Results

For a total of 20 epochs, the model was trained. Over the course of 20 epochs, the accuracy of the trained CNN-LSTM model improves. The graph as shown in Figure 2 represents the loss of CNN-LSTM model along y axis and number of epochs along x axis over the twenty epochs of training. The loss keeps decreasing over time. It decreases exponentially from 4.5348 to 2.6380 until the entire span of 20 epochs. This signifies that the images are more accurately classified and the captions generated are more accurate and precise as the number of epochs increase. For estimating the error of the model, the network is trained to use an optimization method which involves a function loss. Maximum likelihood is achieved by optimizing a likelihood function produced from the training phase to discover the optimal values for the model parameters.

.

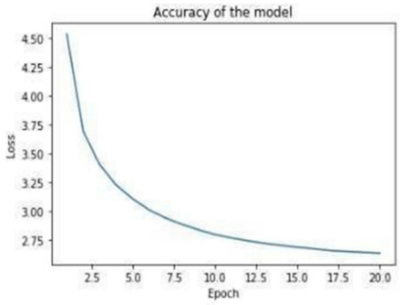

Fig. 2. Loss of CNN-LSTM model

start two dogs are playing with b

(a)

start man in red shirt is climbing rock end

(b)

start red jeep with a girl on it end

(c)

start green fish out of water end

(d)

start red apple is on the table end

(e)

Fig. 3. Generated Captions

Bilingual Evaluation Understudy or BLEU is a quality metric for evaluating a generated translation of text to a reference/human translation. The score is for comparing sentences and is based on an average of unigram, bigram, trigram, and 4-gram precision. A strong correlation gets a score of 1.0 and perfect mismatch results in score *0.0*. Figure 3 depicts some of the results obtained. BLEU-1 represents the unigram precision where each of the words in the caption generated is compared with the words in the original text dataset. BLEU-2 represents the bigram precision where neighboring pair of words are compared with the original text dataset. 0.5 weight is applied to both the unigram and bigram index which means that both are considered equally with equal importance. BLEU-3 represents the trigram precision where a set of three words are compared with the original text dataset. 0.3 weight is applied to unigram, bigram and trigram index which suggests that one-third of the results from each is considered and a cumulative is taken into account. BLEU-4 represents the 4-gram precision where a set of 4 words appearing next to each other are compared with the original text dataset. 0.25 weight is applied to all unigram, bigram, trigram and 4-gram indices which

means that a cumulative score in obtained. The scores that are obtained in BLEU-*1*=0.589, BLEU-2=0.335, BLEU-3=0.263 and BLEU-4=0.148.

5 Conclusion

Photo caption generator recognizes the image context and describes it in English language. It comprises of steps like data cleaning, extracting features, tokenizing the vocabulary, defining, training, validating and testing the model. In this paper the generation of caption for the image using CNN with LSTM is depicted. The advantage of training using CNN with LSTM are the ability to handle sequences of arbitrary length, and more importantly, the end-to-end maximization of the joint probability of the source and target sentence, have produced output in machine translation. The efficiency of the model is calculated using BLUE Score. This system is data-driven, therefore it can't anticipate phrases which are not in its vocabulary. The following tasks can be performed in the future. Text-to-speech technology, which automatically reads aloud to visually challenged people. Translating images directly into sentences, rather than creating image captions. Static images can only provide information about one particular instant in time to blind people, thus creating video captions may theoretically provide continuous real-time information to blind people. Using CNN and LSTM the developed system contains characteristics that predict an image's caption.

References

[1] D. D. Sapkal, Pratik Sethi, Rohan Ingle, Shantanu Kumar Vashishtha, Yash Bhan Professor, “A Survey on Auto Image Captioning”, International Journal of Innovative Research in Science,Engineering and Technology Vol. 5, Issue 2, February 2016.

[2] Jimmy Lei Ba, Volodymyr Minho, Koray Kavukcuoglu, “Survey on Feature Extraction of Images for Appropriate Caption Generation”, International Journal of Engineering Research and General Science Volume 4, Issue 1, January-February, 2016.

[3] Pranay Mathur, Aman Gill, Aayush Yadav, Anurag Mishra and Nand Kumar Bansode, “Camera2Caption:A Real-Time Image Caption Generator”, International Conference on Computational Intelligence in Data Science(ICCIDS), Chennai, 2017.

[4] Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, NazliIkizler-Cinbis, Frank Keller, Adrian Muscat, Barbara Plank, “Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures”, 26TH International Joint Conference on Artificial Intelligence (IJCAI-17), Melbourne, Australia, 2016.

[5] Lun Huang, Wenmin Wang,Yaxian Xia, JieChen, “Adaptively Aligned Image Captioning via Adaptive Attention Time”, 33rd Conference on Neural Information Processing Systems, Vancouver, Canada, 2019.

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

[7] 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.

[8] Ying Hua Tan, Chee Seng Chan,Phrase-based image caption generator with hierarchical LSTM network, Neurocomputing,Volume 333,2019.

[9] Jaing, W., Ma, L., Chen, X., Zhang, H., Liu, W.: Learning to guide decoding for image captioning. In: Thirty Second AAAI Conference on Artificial Intelligence (AAAI – 2018), pp. 6959–6966,2018.

[10] Kinghorn, P., Zhang, L., Shao, L.: A hierarchical and regional deep learning architecture for image description generation. Pattern Recogin. Lett. 119, 1–9 ,2017.

[11] Seo, P. H., Sharma, P., Levinboim, T., Han, B., & Soricut, R. (2020). Reinforcing an Image Caption Generator Using Off-Line Human Feedback. Proceedings of the AAAI Conference on Artificial Intelligence, 34(03), 2693-2700,2019.
[12] Tariq, A., Foroosh, H.: A context - driven extractive framework for generating realistic image descriptions. IEEE Trans. Image Process. 26(2), 619–632,2017.
[13] Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models. IJCV 123(1), 74–93,2017.
[14] N. K. Kumar, D. Vigneswari, A. Mohan, K. Laxman and J. Yuvaraj, "Detection and Recognition of Objects in Image Caption Generator System: A Deep Learning Approach," 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), Coimbatore, India, 2019.
[15] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, et al, "Bottom-up and top-down attention for image captioning and visual question answering",Prodeedings of 2018 IEEE Conference on Computer Vision & Pattern Recognition, CVPR (2018), pp. 6077-6086.
[16] Xianrui Li, Zhiling Ye, Zhao Zhang, Mingbo Zhao,"Clothes image caption generation with attribute detection and visual attention model",Pattern Recognition Letters,Volume 141,2021,Pages 68-74,ISSN 0167-8655.
[17] Liu, D., Zhang, H., Zha, Z.-J., Wu, F."Learning to assemble neural module tree networks for visual grounding ",(2019) Proceedings of the IEEE International Conference on Computer Vision, 2019-October, art. no. 9009000, pp. 4672-4681. ISBN: 978-172814803-8,doi: 10.1109/ICCV.2019.00477.
[18] Yang, X., Tang, K., Zhang, H., Cai, J."Auto-encoding scene graphs for image captioning,(2019) Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition", 2019-June, art. no. 8953305, pp. 10677-10686. ISBN: 978-172813293-8,doi: 10.1109/CVPR.2019.01094.
[19] Sule Anjomshoae, Daniel Omeiza, Lili Jiang,"Context-based image explanations for deep neural networks",Image and Vision Computing,Volume 116,2021,104310,ISSN 0262-8856.
[20] R. Goplalakrishnan, A. Mohan, L. Ponrajsankar, D. S. Vijayan, "Characterisation On Toughness Property of Self- Compacting Fibre Reinforced Concrete", Journal of Environmental Protection and Ecology 21, No 6, 2153–2163 (2020).
[21] Khaing, Phyu & Yu, May. (2021). Two-Tier LSTM Model for Image Caption Generation. International Journal of Intelligent Engineering and Systems. 14. 22-34. 10.22266/ijies2021.0831.03.
[22] A. Verma, H. Saxena, M. Jaiswal and P. Tanwar, "Intelligence Embedded Image Caption Generator using LSTM based RNN Model," 2021 6th International Conference on Communication and Electronics Systems (ICCES), 2021, pp. 963-967, doi: 10.1109/ICCES51350.2021.9489253.
[23] L. Zhang and Q. Lu, "Image caption generation method based on an interaction mechanism and scene concept selection module," 2021 IEEE 32nd International Conference on Application-specific Systems, Architectures and Processors (ASAP), 2021, pp. 141-148, doi: 10.1109/ASAP52443.2021.00028.
[24] Tian, P.; Mo, H.; Jiang, L,"Image Caption Generation Using Multi-Level Semantic Context Information. Symmetry," 2021, 13,1184.https://doi.org/10.3390/sym13071184.
[25] Alam, Mohammad Shahnawaz & Narula, Vaishali & Haldia, Ruchika & Ganpatrao, Gitanjali. (2021). An Empirical Study of Image Captioning using Deep Learning. 1039-1044.