Image Description using Attention Mechanism

Abhilash Pandurangan, Vignesh Prabhakar, Poovammal E

Abstract— Image Description involves generating a textual description of images which is essential for the problem of image understanding. The variable and ambiguous nature of possible image descriptions make this task challenging. There are different approaches for automated image captioning which explain the image contents along with a complete understanding of the image, rather than just simply classifying it into a particular object type. However, learning image contexts from the text and generating image descriptions similar to human's description requires to focus on important features of the image using attention mechanism. We provide an outline of the various recent works in image description models employing various attention mechanism. We present an analysis of the various approaches, datasets and evaluation metrics that are utilized for image description. We showcase a model using the encoder-decoder attention mechanism based on Flickr dataset and evaluate the performance using BLEU metrics.

Keywords— Image Caption Generation; Attention Mechanism; Encoder and Decoder; Deep neural networks

I. INTRODUCTION

Recent advances in deep learning to handle tasks like object detection, image classification, and object segmentation have enhanced the feasibility of more complex problems like that of image captioning which can largely help us for better visual understanding in vision-based systems. Image captioning problem involves describing the given image with a natural sentence. With the advent of the exponentially increasing volume of unstructured data, which is largely available as images and videos, it is essential to have exact image description to leverage their value but it would be challenging task to manually process them. Automated description of images can be useful for image retrieval task by sorting image-based content leading to much efficient access of images. There are also plenty of other applications like helping the visually-impaired people by using some real-time annotations or utilizing it for visual question answering systems or to aid in robotics. Apart from these, we can also use image captioning to annotate the medical images which can help us to detect the diseases.

Image description generation system involves both the advances in natural language translation and computer vision due to which the main challenge faced is the need for bridging between these two different tasks. Image description can be used to build systems which are capable of perceiving contextual parts, to result in appropriate and accurate descriptions. Designing a captioning system which can generate captions which are equivalent to those generated by humans is a challenging task. An image can be described with the help of multiple sentences but we require only a single sentence which can be given as a caption for training the model. This leads to a problem of natural language processing for generation of natural sentences that describe the images accurately with a special focus on the important features of the image. Image description often requires focusing on particular portions of the image to generate an accurate natural sentence that provides insights about what the image is essentially portraying. The challenging part of this task is to not just extract the object features and its relations in the image but to express them in proper language which provides an understanding of the entire image.

II. LITERATURE REVIEW

Early approaches in image captioning depended on the usage of specific rules of grammar and the sentences were formed according to the results of detection of scene objects [1]. But these primitive systems do not form proper natural sentences and it is more dependent on the classification of elements in the image.

Vinyals et al. [2] have utilized a combination of a deep convolutional network with re-current networks for sequence modeling, a single network that produces descriptions of images. They have presented Neural Image Caption generator model which is an end-to-end neural network on a CNN which encodes an image into a vector representation which is followed by RNN that generates sentence related to the image.

Authors in [3] propose a similar system, using a combination of CNN and LSTM for encoder-decoder architecture and propose the utilization of LSTM in place of recurrent neural networks combined with powerful CNN architecture which was the primary factor for enhancement of performance. Additionally, they have added the feature representation from the previous LSTM of the stack which proved to be an effective way for increasing performance.

Xu et al. [4] were the pioneers to utilize the attention mechanism for image captioning. They present two types of systems, soft and hard attention mechanism which are extensions to the encoder-decoder framework. They have used a lower convolutional layer for extracting the image features using as annotation vectors which contain a summary of spatial locations of the image. Each of these vectors is given weight which is calculated by the attention model. These weights are just the possibilities for attending

Revised Version Manuscript Received on 10, September 2019.

Abhilash Pandurangan, Student, Computer Science and Engineering, SRM Institute of Science and Technology, SRM Nagar, Kattankulathur, Pin-603203, Chennai, Tamilnadu, India

Vignesh Prabhakar, Student, Computer Science and Engineering, SRM Institute of Science and Technology, SRM Nagar, Kattankulathur, Pin-603203, Chennai, Tamilnadu, India

Poovammal E, Professor, Computer Science and Engineering, SRM Institute of Science and Technology SRM Nagar, Kattankulathur, Pin-603203, Chennai, Tamilnadu, India

by the decoder during the next word generation or as the relative importance provided to a particular location. These models also help us to visualize the attention weights for each word from the output of the decoder.

You et al. [5] note that traditional methods in image captioning are either top-down or bottom-up based on whether they start from the gist of an image and change it to words or vice-versa. However, they introduce a novel algorithm which combines both these methods and learns to attend selectively. This is possible by using a model of semantic attention, which is based on semantic concepts and the feature representation of the image encoding.

A. Karpathy et al. [6] introduce a novel approach of dense captioning where the model depicts certain portions of the image which is denoted by bounding boxes. This method produces results that might be more relevant when compared with captioning an entire image as such. It makes use of two independent different networks, one for text content and other for image regions, that represents inside the exact image-text space.

Authors of [7] claim that attending image at each step becomes unnecessary. They propose novel adaptive attention-based framework with visual sentinel that, while producing the next word at each step in the description, automatically tends to select whether to attend to the portion of the image. This adaptive context vector is formed with both the context vector in spatial attention model and the visual sentinel vector. This approach enables selective focus and it depends on each step to decide on which particular regions to attend and the output is generated accordingly.

A. Background

1) Convolutional Neural Network (CNN): These are neural networks comprised of multiple alternating convolutional layers, pooling layers and followed by fully connected layers and an optional softmax layer for classification. The architecture of CNN makes use of the two-dimensional structure of the input image to extract feature vectors from them. The convolutional layer neurons are not connected to all neurons from the previous layer like those of fully connected layers. This ensures that local features can be extracted from them. Low-level features are extracted from the first few layers and are combined with higher level features to give a vector representation as the output. Pooling layers are used to perform averaging to reduce the dimensions and to get better feature extraction. CNN's are used for image classification, object detection, and recognition problems.

2) Recurrent Neural Networks (RNN): It is a type of neural network which contains an internal memory and is useful for sequential tasks. It can remember important things about the input, which enables for precise next input predictions. RNN maintains a particular internal hidden state which is used to store context information that is calculated from past inputs. However, in the case of longer sequences, it becomes practically difficult to train them due to the problems of exploding and vanishing gradient.

3) Long-Short Term Memory (LSTM): These neural networks are typically just extensions of recurrent neural networks, which consists of an improvement in their memory. It is well tailored to learn from important experiences that have long-time gaps in between. LSTMs enable RNNs to remember the inputs over a long period of time as they contain their information in a memory which can be read, written and also deleted from its memory.

4) Encoder – Decoder Architecture: Most of the captioning systems are based on this framework which consists of two elements – encoder and decoder. The encoder reads the input image and transforms into a vector usually of fixed length representation. Decoder takes the encoded image as input and generates the description. Generally, the model consists of CNN as encoder and LSTM network for decoder as in Fig 1. These decoder layers can take input feature vector over different stacks of LSTMs and give the sentence as the output. These are trained together and can be used to extract the features from the input image and get the output a description of the image.

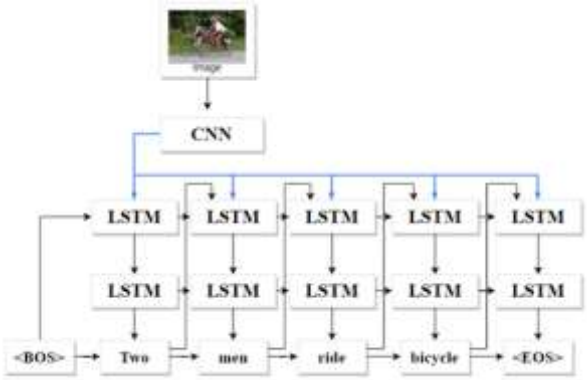

Fig. 1. Encoder-decoder architecture for image captioning system

5) Attention mechanism: An extension of encoder-decoder design for image caption generation is the utilization of attention mechanism which has enhanced these systems to perform better. When the RNN generates a word, this approach helps to focus on relevant portions of the image to ensure proper description. The attention model takes the arguments along with an additional vector for representing the context information and gives out vector which is the summary of features implying the weighted mean of the input feature and these weights are assigned according to the relevance with respect to context. The attention model allows, for each word, to concentrate on parts of the text that is relevant to the image.

There are different variations of this mechanism employed on top of the usual framework which helps us to visualize the output of the system and for improving the efficacy of the model to describe the images accurately.

III. DATASETS

The emerging innovations in image captioning systems require datasets with a large number of images having different captions for providing good training data. These datasets could be collected with either user-generated captions or by means of crowdsourcing the captions for these images. [12]

1) Flickr 8k [11] is a much huge and diverse dataset comprising of around 8,000 images gathered from Flickr and which is focused on animate entities performing some arbitrary action. Crowdsourcing had been used to gather five discrete captions for any given image describing the animated entity and the action being performed.

2) Flickr 30k [13] dataset is an addition to the Flickr 8k dataset where in around 30,000 images of diurnal activities and happenings have been described by these images through crowdsourcing.

3) MS COCO [10] captions dataset extends Microsoft's Common Objects in Context dataset comprising of the images of intricate diurnal activities and general objects in their original context. Dataset provides a set of training and validation images and we can upload the generated captions to the COCO server for evaluation.

Flickr 30k and MS COCO are widely believed to be the standard datasets for the task of image captioning in most of the deep learning research as they contain most structurally complex sentences along with large image samples.

IV. EVALUATION METRICS & RESULT

Description of an image requires summarisation of the most relevant part in terms of objects and attributes along with the deduction of what is new and involving, expression of semantic content with grammatically correct sentences which they describe. Evaluation of such descriptions causes problems. It is evident that providing emphasis on one more aspect may still vary the resultant sentences considerably although they may be accurate.

Evaluation of captions can be done by experts through crowdsourcing. Human dependent evaluations are subject to additional costs and are usually hard to reproduce. The utilization of automatic metrics could be quicker, more accurate and inexpensive. The output score of the metrics compares the given sentence and reference sentence.

1) Bilingual Evaluation Understudy [8] or BLEU score is a popular algorithm for evaluating generated sentence and is one of the pioneering metrics used for image captioning. It matches the generated sentence with the reference sentence and score of 1.0 indicates a perfect match. It uses n-gram modified precision method for the comparison of sentences.

2) SPICE [14] is another algorithm employed for the evaluation of captions. The quality of image captions is measured using an F-measure which is based on a content of semantic propositional for test and reference sentences represented as scene graphs. This model is well known due to their endorsement by Microsoft evaluation server which allows regular comparison using an even implementation of specified metrics.

However, these automatic metrics do not often correlate with artificial decisions as they outperformed artificial upper bounds according to automatic metrics but human judges had a predilection towards humanly generated captions. Additional issues with the metrics are that the replacement of words with similar meanings decreases the scores of metrics.

V. IMPLEMENTATION

For the task of image description, we have implemented the Show, Attend and Tell [4] model with visual attention mechanism using Flickr 8k datasets. We have trained the model in Keras with Telsa K80 GPU with 12GB RAM machine using 6,000 images for the training set and rest 1000 split for testing and validation sets. The model consists of CNN encoder based on the VGG Net pre-trained model and LSTM decoder architecture with hard attention weights for context vectors. For data pre-processing, we tokenized the given captions in the dataset to build a vocabulary of all unique words which is required to generate proper descriptions as in Fig 2.

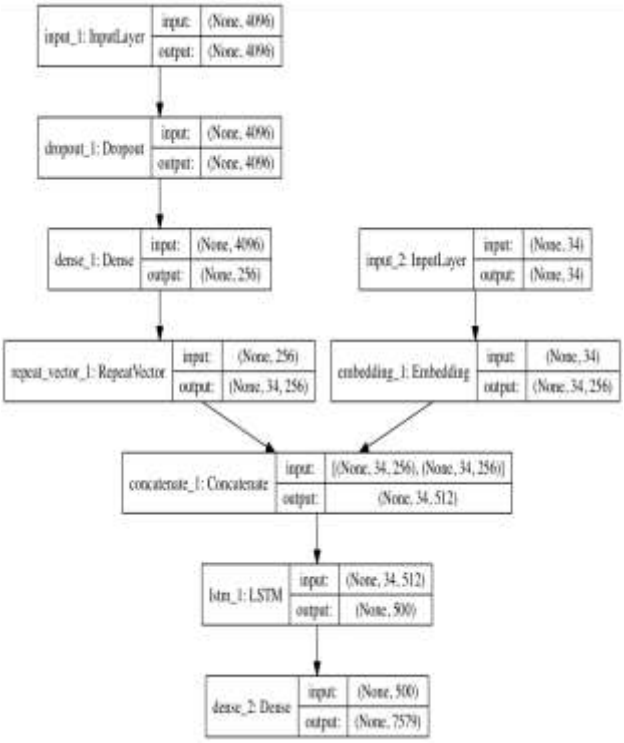

Fig 2. The model structure used for generating captions

During the training, we extract the features using VGG16 model and pass it to the encoder whose output is then sent to the decoder. The decoder gives the predictions which are passed back to model to calculate the loss and backpropagate it. While generating the captions, the input to the decoder at each step is given by previous predictions along with encoder output and attention weights. These attention weights help to generate better sentences which are based on the context vectors. The model is then evaluated with BLEU metrics which results in an effective score of 0.62.

Fig 2. Results of Image Description Model

As in Fig.2, these sentences generated "Man is riding the bike on the mountains" and "Black dog is running through the water" are natural and describe the scene in the image clearly. The descriptions from the model with an average of 0.62 BLEU score can be compared to the human-generated sentences. We have also deployed our model by saving it and running the model on web service using flask server as in Fig 3.

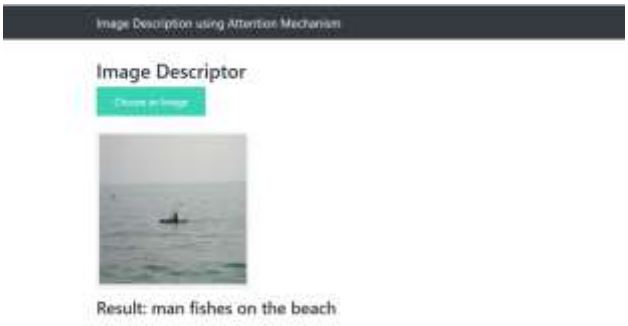

Fig 3. Model deployed as a web app

VI. CHALLENGES

The goal in such image description systems is to be able to fabricate image descriptions which could be judged as being not bad as models are still a long way from being classified as perfect. However, there are some problems which are needed to be focused on future research. Some of these issues are inherently internal whereas some of them are external such as utilization of a particular dataset. One such problem for this task is the assessment of such generated captions. Colouring has been a major problem with images and identification of texture has been problematic as well.

Authors in [2] have presented the problem using a different perspective and they found that transferring of the model to different dataset results in a decrease of BLEU scores. The generated text also depends on the dataset used for training. Lack of variety in the training data was observed as the captioning is being produced from the training set and such issues replicated in over more than half of the time. The major problem for this task is the assessment of such generated captions. While most of the evaluation metrics being used provide some sort of metrics, their usage will depend based on their accuracy and comparison to human level evaluations.

VII. CONCLUSION

We have presented an outline of various advances in the domain of image captioning, with the implementation of a model which employs encoder-decoder design with the attention mechanism. The benefit of using such models is that they can be trained together, mapping from images to natural sentences. The general encoder-decoder framework along with the attention mechanism enables to attend only the most salient portions of the image while producing the next sequence of the output sentence. We have described different variations of attention mechanism which help us to fine tune the performance and showcased the model based on visual attention mechanism.

Most of the work in this field deals with models which generate image descriptions, however, similar models can be applied to videos as well, which are just multiple frames of images. We can also employ reinforcement learning approaches in addition to the general model which usually need lesser training data. Further, such models can be used to visualize the focus parts of the image for the output sentences for better understanding the neural networks.

REFERENCES

1. G. Kulkarni et al., "Baby talk: Understanding and generating simple image descriptions," CVPR 2011, Colorado Springs, CO, USA, 2011, pp. 1601-1608.
2. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 3156-3164.
3. J. Donahue et al., "Long-term recurrent convolutional networks for visual recognition and description," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2625–2634.
4. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, Yoshua Bengio, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, arXiv:1502.03044
5. Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, Jiebo Luo, Image Captioning with Semantic Attention, arXiv:1603.03925
6. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3128–3137.
7. J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via A visual sentinel for image captioning," arXiv preprint arXiv:1612.01887, 2016.
8. K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in Proceedings of the 40th annual meeting on association for computational linguistics, 2002, pp. 311–318.
9. X. Chen et al., "Microsoft COCO captions: Data collection and evaluation server," arXiv preprint arXiv: 1504.00325, 2015
10. M. Hodosh, P. Young and J. Hockenmaier (2013) "Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics", Journal of Artificial Intelligence Research, Volume 47, pages 853-899
11. F. Ferraro, N. Mostafazadeh, L. Vanderwende, J. Devlin, M. Galley, and M. Mitchell, "A survey of current datasets for vision and language research," arXiv preprint arXiv:1506.06833, 2015.
12. P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," Transactions of the Association for Computational Linguistics, vol. 2, pp. 67–78, 2014
13. E. van Miltenburg and D. Elliot, "Room for improvement in automatic image description: an error analysis," arXiv:1704.04198, 2017
14. Peter Anderson, Basura Fernando, Mark Johnson, Stephen Gould, "SPICE: Semantic Propositional Image Caption Evaluation", In Proceedings of European Conference on Computer Vision, (ECCV) 2016