Automatic Image Captioning Using Convolution Neural Networks and LSTM

R. Subash, Yash Kamdar, and Nishit Bhatt
subash.r@ktr.srmuniv.edu.in, yashkamdar_ke@srmuniv.edu.in, nishitbhatt_vi@srmuniv.edu.in

Abstract—PC vision has turned out to be universal in our general public, with applications in a few fields. Given a lot of pictures, with its inscription, make a prescient model which produces regular, inventive, and intriguing subtitles for the concealed picture. A speedy look at a picture is adequate for a human to call attention to and portray a monstrous measure of insights regarding the visual scene. To rearrange the current issue of producing inscriptions for pictures by making a model which would give exact subtitles to these pictures which can be additionally utilized in other helpful applications and use cases. Be that as it may, this momentous capacity has ended up being a tricky errand for our visual acknowledgment models. Most of the past work in visual acknowledgment has concentrated on naming pictures with a fixed arrangement of visual classifica-tions and extraordinary advancement has been accomplished in these undertakings. For a question picture, the past strategies recover pertinent hopeful normal language states by outwardly contrasting the inquiry picture with database pictures. In any case, while shut vocabularies of visual ideas comprise a helpful demonstrating suspicion, they are boundlessly prohibitive when contrasted with the colossal measure of rich depictions that a human can form. These methodologies forced a breaking point on the assortment of inscriptions produced. The model ought to be free of suppositions about explicit hard-coded formats, standards or classes and rather depend on figuring out how to create sentences from the preparation information. The model proposed utilizes Convolution Neural Networks which help to separate highlights of the picture whose subtitle is to be created and afterward by utilizing a probabilistic methodology and Natural Language Processing Techniques reasonable sentences are framed and inscriptions are produced.

I. INTRODUCTION

Artificial Intelligence(AI) is currently at the core of development economy and in this manner the base for this task is additionally the equivalent. In the ongoing past a field of AI in particular Deep Learning has turned a ton of heads because of its noteworthy outcomes regarding exactness when contrasted with the effectively existing Machine learning calculations. The assignment of having the capacity to create a important sentence from a picture is a troublesome undertaking yet can have extraordinary effect, for example helping the outwardly impeded to have a superior comprehension of pictures. The undertaking of picture subtitling is essentially harder than that of picture characterization, which has been the fundamental concentration in the PC vision network. A portrayal for a picture must catch the connection between the items in the picture. Notwithstanding the visual comprehension of the picture, the above semantic learning must be communicated

in a characteristic language like English, which implies that a language demonstrate is required. The endeavors made in the past have all been to join the two models together. Utilizing Convolution Neural nets makes our errand a lot simpler in light of the fact that they are truly adept at discovering designs in pictures and arranging pictures. We use channels to discover the highlights of the picture and afterward apply pooling to lessen the size. Later on the tokens of the picture are passed to a Long Short Term Memory(LSTM) unit which at that point creates significant subtitles by utilizing probabilistic methodology.

II. RELATED WORK

Research scientists have been trying to generate human like caption for images even with the traditional and inefficient methods. The first paper studied in the survey from 2010 was able to generate sentences of just 3 words. The technique after it focused on labeling the images in categories and was unable to generate descriptive dense captions. Around 2015 started the era of deep learning and neural nets. Most research was done using CNNs. CNNs needed a large data set to arrive at accurate results. CNN paired along with RNNs provided a good method to do the task. Multimodal RNNs couldn't perform the task well because it was not able to use them with ReLU activation function. Other classifying techniques like mRNN plus nearest neighbors were able to simplify the task but generated errors on colored images. The last method in 2017 was feeding high frequency attributes into CNN and LSTM network which currently is a state-of-the-art method for generating captions but this method requires high amount of processing power and GPU and cannot possibly work well on a normal computer. Thus, a method needed to be found which could run on normal processing power and doesn't need huge data and it should be accurate.

III. OBJECTIVE

The objective of this project is to ease the task of generating human captions for images. Humans can just by looking at images generate a very descriptive caption for the image. But machines despite having the processing power and learning ability fail to generate human like captions. We use a model having convolution neural network whose output is paired to Long Short Term Memory network which helps us generate descriptive captions for the image. For helping us achieve this on normal machines, we have used techniques such as spatial pooling, filters, strides, convolution operator, etc. We use the MSCOCO data set for this which is open source and
freely available. It has been contributed by various people worldwide providing captions for all images available there. The applications of this project can be used in every way to make life easier for a lot of people. The first application would be using it for blind people. The application can connect through the mobile phone camera and people can hear accurate captions through the earphones. The second application can be used by social media companies for generating captions for the images on their platform allowing partially blind users to listen to them. Lastly, if the same technology is applied for finding scenes in a video by description then we can save a huge amount of time and also it can be really accurate in finding the exact situation in a footage of a video where there is some mishap occurring.

IV. MODEL

In the model proposed in the paper we try to combine this into a single model which consists of a Convolution Neural Network (CNN) encoder which helps in creating image encoding. We use the VGG16 architecture with some modifi-cations. We could have used some of the recent and advanced classification architectures but that would have increased the training time significantly. These encoded images are then passed to a LSTM network which are a type of Recurrent Neural Network. The input to the network is an image which is first converted in a 224*224 dimension. We use the MSCOCO data set to train the model. The model outputs a generated caption based on the dictionary it forms from the tokens of caption in the training set. The generated caption is compared with the human given caption via BLEU score measure.

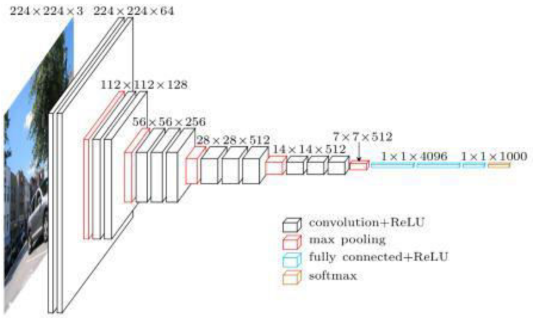

Fig. 1. VGG16 Architecture

These operations are the basic building blocks of every Convolution Neural Network.

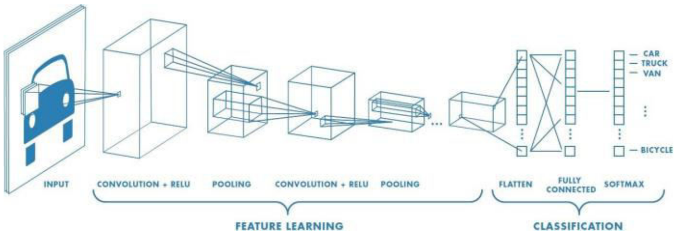

Fig. 2. Architecture of Convolution Neural Net

LSTMs are explicitly designed to avoid the long-term de-pendency problem. Remembering information for long periods of time is practically their default behavior, not something they struggle to learn! All recurrent neural networks have the form of a chain of repeating modules of neural network. In standard RNNs, this repeating module will have a very simple structure, such as a single tanh layer. LSTMs also have this chain like structure, but the repeating module has a different structure. Instead of having a single neural network layer, there are four, interacting in a very special way. The key behind the LSTM network is the horizontal line running on the top which is known as the cell state. The cell state runs through all the repeating modules and is modified at every module with the help of gates. This causes the information in a LSTM network to persist.

Convolution Neural Networks (ConvNets or CNNs) are a category of Artificial Neural Networks which have proven to be very effective in the field of image recognition and classification. They have been used extensively for the task of object detection, self driving cars, image captioning etc. The entire modules of a convolution neural net can be explained using four main operations namely,

1. Convolution
2. Non- Linearity (ReLU)
3. Pooling or Sub Sampling
4. Classification (Fully Connected Layer)

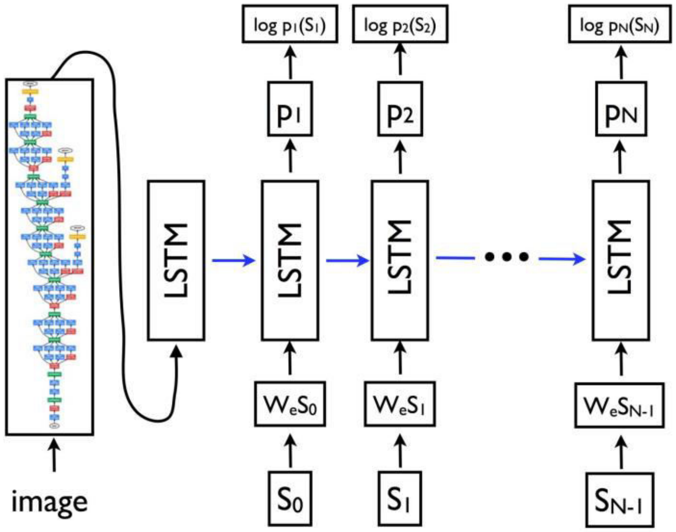

Fig. 3. Overall architecture of the network

V. RESULTS

Bilingual evaluation understudy (BLEU) is an algorithm that evaluated the quality of text which has been translated by a machine. It was one of the first metrics to achieve high correlation with human judgment. Our Model works well, and we have compared our model with the other models out there on the MSCOCO data set. Here are a few of the results of Test image along with the generated captions.

VI. CONCLUSION

Our end-to-end system neural network system is capable of viewing an image and generating a reasonable description in English depending on the words in its dictionary generated on the basis of tokens in the captions of train images. The model has a convolution neural network encoder and a LSTM decoder that helps in generation of sentences. The purpose of the model is to maximize the likelihood of the sentence given the image. Experimenting the model with MS COCO data set show decent results. We evaluate the accuracy of the model on the basis of BLEU score. The accuracy can be increased if the same model is worked upon a bigger data set. Furthermore, it will be interesting to see how one can use unsupervised data, both from images alone and text alone, to improve image description approaches. The task of image captioning can be put to great use for the visually impaired. The model proposed can be integrated with an android or ios application to work as a real-time scene descriptor. The accuracy of the model can be improved to achieve state-of-the-art results by hyper tuning the parameters. The model's accuracy can be boosted by deploying it on a larger dataset so that the words in the vocabulary of the model increase significantly. The use of relatively newer architecture, like ResNet and GoogleNet can also increase the accuracy in the classification task thus reducing the error rate in the language generation. Apart from that the use of bidirectional LSTM network and Gated Recurrent Unit may help in improving the accuracy of the model.

REFERENCES

1. Chen, Xinlei, and C. Lawrence Zitnick. "Learning a recurrent visual representation for image caption generation". In arXiv preprint arXiv:1411.5654, 2014.

2. Datta, R., Li, J., Wang, J.Z.: Content-based image re-trieval: approaches and trends of the new age. In: MIR '05. (2005) 253–262

3. Fang, Hao, et al. "From captions to visual concepts and back." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

4. Vedantam, Ramakrishna, C. Lawrence Zitnick, and Devi Parikh. "Cider: Consensus-based image description evalua-tion." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

5. Wang, Cheng, et al. "Image captioning with deep bidirec-tional LSTMs." Proceedings of the

6. 2016 ACM on Multimedia Conference. ACM, 2016.

7. L. Fei-Fei, C. Koch, A. Iyer, and P. Perona. What do we see when we glance at a scene. Journal of Vision, 4(8), 2004.

8. Li, L.J., Fei-Fei, L.: What, where and who? classifying event by scene and object recognition. In: ICCV. 2007

9. M. Yatskar, L. Vanderwende, and L. Zettlemoyer. See no evil, say no evil: Description generation from densely labeled images. Lexical and Computational Semantics, 2014.

10. S.Li,G.Kulkarni,T.L.Berg,A.C.Berg,andY.Choi. Compos-ing simple image descriptions using web-scale n-grams. In CoNLL, 2011.

11. Gupta, A., Davis, L.: Objects in action: An approach for combining action under-standing and object perception. In: CVPR. (2007)