Automatic Image Captioning Using Convolution Neural Networks and LSTM

To cite this article: R. Subash et al 2019 J. Phys.: Conf. Ser. 1362 012096

View the article online for updates and enhancements.
Automatic Image Captioning Using Convolution Neural Networks and LSTM

R. Subash¹, Dr.R.Jebakumar², Yash Kamdar³, Nishit Bhatt⁴

¹²³⁴SRM Institute of science and Technology, Chennai
Subash.r@ktr.srmuniv.ac.in¹, jebakumar.r@ktr.srmuniv.ac.in², yashkamdar_ke@srmuniv.edu.in³, nishitbhatt_vi@srmuniv.edu.in⁴

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 research in scene acknowledgment has concentrated on naming pictures with a predetermined arrangement of visual classifications and extraordinary advancement has been accom- plished 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 exempt of suppositions regarding explicit pre decided 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.

1. 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.
2. RELATED WORK

Research scientists have been trying to generate human-like captions 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.

3. 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.

4. MODEL

In the model proposed in the paper we attempt to consolidate this into a solitary model which comprises of a Convolution Neural Network (CNN) encoder which helps in making picture encoding. We utilize the VGG16 design with certain changes. We could have utilized a portion of the ongoing and propelled order structures yet that would have expanded the preparation time altogether. These encoded pictures are then passed to a LSTM arrange which are a sort of Recurrent Neural Network. The contribution to the system is a picture which is first changed over in a 224*224 measurement. We utilize the MSCOCO informational collection to prepare the model. The model yields a created inscription dependent on the lexicon it frames from the tokens of subtitle in the preparation set. The created inscription is contrasted and the human given subtitle by means of BLEU score measure.
Convolution Neural Networks (ConvNets or CNNs) are a class of Artificial Neural Networks which have ended up being viable in the field of picture acknowledgment and order. They have been utilized widely for the assignment of article recognition, self-driving vehicles, picture subtitling and so on. The whole modules of a convolution neural net can be clarified utilizing four fundamental tasks to be specific,

a. Convolution
b. Rectified Linear Unit
c. Pooling
d. Fully Connected Layer

These tasks are the essential structure of each Convolution Neural Network.

LSTMs are expressly intended to maintain a strategic distance from the long haul reliance issue. Recalling data for significant lots of time is for all intents and purposes their default conduct, not something they battle to learn! All repetitive neural systems have the type of a chain of rehashing modules of neural system. In standard RNNs, this rehashing module will have a basic structure, for example, a solitary tanh layer. LSTMs additionally have this chain like structure, however the rehashing module has an alternate structure. Rather than having a solitary neural system layer, there are four, associating in an exceptionally unique manner. The key behind the LSTM organize is the level line running on the top which is known as the cell state. The cell state goes through all the rehashing modules and is adjusted at each module with the assistance of doors. This causes the data in a LSTM system to persevere.
5. RESULTS

Bilingual evaluation understudy (BLEU) is a calculation that assessed the nature of content which has been deciphered by a machine. It was one of the principal measurements to accomplish high connection 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.

Fig. 3: Overall architecture of the network

Fig. 4: Example image 1 with generated caption
6. CONCLUSION

Our start to finish framework neural system framework is equipped for review a picture and creating a sensible depiction in English relying upon the words in its lexicon produced based on tokens in the inscriptions of train pictures. The model has a convolution neural system encoder and a LSTM decoder that helps in age of sentences. The reason for the model is to expand the probability of the sentence given the picture. Testing the model with MS COCO informational index show not too bad outcomes. We assess the exactness of the model based on BLEU score. The exactness can be expanded if a similar model is worked upon a greater informational collection. Moreover, it will intrigue perceive how one can utilize unsupervised information, both from pictures alone and message alone, to improve picture portrayal approaches. The undertaking of picture subtitling can be put to incredible use for the outwardly impeded. The model proposed can be incorporated with an android or ios application to fill in as a constant scene descriptor. The exactness of the model can be improved to accomplish cutting edge results by hyper tuning the parameters. The model’s exactness can be supported by conveying it on a bigger dataset with the goal that the words in the vocabulary of the model increment fundamentally. The utilization of moderately more up to date engineering, as ResNet and GoogleNet can likewise expand the exactness in the arrangement task subsequently decreasing the blunder rate in the language age. Aside from that the utilization of bidirectional LSTM system and Gated Recurrent Unit may help in improving the exactness of the model.

7. 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 retrieval: 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 evaluation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[5]. Wang, Cheng, et al. "Image captioning with deep bidirectional LSTMs." Proceedings of the 2016 ACM on Multimedia Conference. ACM, 2016.

[6]. 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.

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

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

[9]. S.Li,G.Kulkarni,T.L.Berg,A.C.Berg,andY.Choi. Composing simple image descriptions using web-scale n-grams. In CoNLL, 2011.

[10].Gupta, A., Davis, L.: Objects in action: An approach for combining action understanding and object perception. In: CVPR. (2007)

[11].NHK K. ISMAIL*,”Estimation Of Reliability Of D Flip-Flops Using Mc Analysis”, Journal of VLSI Circuits And Systems 1 (01), 10-12,2019.

[12].Sulyukova,”Analysis of Low power and reliable XOR-XNOR circuit for high Speed Applications”,Journal of VLSI Circuits And Systems 1 (01), 23-26,2019.