TRANSFORMER ARCHITECTURE FOR IMAGE CAPTURING USING DEEP LEARNING APPROACH

Ajith PJ [1], Dr G.Kiruthiga [2]

[1] *student, Dept. of computer science and engineering, IES College of Engineering, Kerala, India*
[2] *Associate professor, Dept. of computer science and engineering, IES College of Engineering, Kerala, India*

ABSTRACT

Image Captioning is a tremendous job in both Natural Language Processing and Computer Vision. Most image captioning systems operate an encoder-decoder framework, where an input image is encoded into an intermediate representation of the information in the image, and then decoded into a descriptive text sequence. This paper bedrock a new deep learning approach based on transformer architecture method for Image Captioning purpose. A transformer is a deep learning model that adopts the mechanism of self-attention, differentially weighting the significance of each part of the input data. It is used primarily in the fields of natural language processing (NLP) and computer vision (CV).Here we are using two folders as dataset one contain the images and another contain the data. Here we are using EfficientNets which is a most powerful convolutional neural network. It uses Compound coefficient to scale up models in a very efficient manner. Compound Model Scaling helps to improve model performance, balancing the scale in all the three dimensions — width, depth, and image resolution , considering the variable available resources best improve the overall model performance.

***Keywords**: -Computer Vision, Convolution Neural Networks, Natural Language Processing.*

1. INTRODUCTION

.

Image Captioning is an interesting and quite confusing job for all. In computer vision image capturing is a branch to generate a correct description about the image or frame [1]. In this paper, we will take a look at an interesting multi modal topic where we will combine both image and text processing to build a useful Deep Learning application for Image Captioning. Image Captioning is the task of generating textual description from an image based on the objects and actions in the image. As per computer vision it is quite impossible a few years back, but now with the enhancement of Computer Vision and Deep learning algorithms, availability of relevant datasets, and Artificial Intelligence models, it becomes easier to build a relevant caption generator for an image. Even Caption generation is becoming a growing business in the world, by building such models many firms earning billions from this.

Our brain is capable of captioning of images appears in front of us. But in the case of computer vision it is quite easy task by the enhancement of deep learning algorithms. Basic knowledge of two techniques of deep learning including LSTM (a type of Recurrent Neural Network) and Convolutional Neural Networks (CNN) is required for the same. A long short-term memory (LSTM) network makes use of the previous output to generate a word at each time instant in dependence on a context vector, previously generated words, and the previous hidden state. Convolutional Neural Networks were designed to map image data to an output variable. They have proven so effective that they are the go-to method for any type of prediction problem involving image data as an input. Recurrent Neural Networks, or RNNs, were designed to work with sequence prediction problems. Some of these sequence prediction problems include one-to-many, many-to-one, and many-to-many.

There are a lot of works in deep learning to captioning the images. Traditional convolution neural networks ,ResNet,GPipe and ImageNet are modeled for image captioning.CNNs are commonly developed at a fixed resource cost, and then scaled up in order to achieve better accuracy when more resources are made available. For example, ResNet can be scaled up from ResNet-18 to ResNet-200 by increasing the number of layers, and recently, GPipe achieved 84.3% ImageNet top-1 accuracy by scaling up a baseline CNN by a factor of four. Here propose a new model "EfficientNet" a novel model scaling method that uses a simple yet highly effective compound coefficient to scale up CNNs in a more structured manner. Unlike conventional approaches that arbitrarily scale network dimensions, such as width, depth and resolution, our method uniformly scales each dimension with a fixed set of scaling coefficients. We have compared our EfficientNets with other existing CNNs on ImageNet. In general, the EfficientNet models achieve both higher accuracy and better efficiency over existing CNNs, reducing parameter size and FLOPS by an order of magnitude.

2. RELATED WORK

In order to substantiate the evolution within the area of image captioning, there are a lot of milestone works are there.In 2011 Y. Yang, C. Teo, H. Daumé, and Y. Aloimonos, were together researched about image captioning and they mark their findings as ''Corpus-guided sentence generation of natural images''[2], As their research uses empirical methods for natural language process. This work propose a sentence generation strategy that describes images by predicting the most likely nouns, verbs, scenes and prepositions that make up the core sentence structure. Use these estimates as parameters on a HMM that models the sentence generation process, with hidden nodes as sentence components and image detections as the emissions
.

In 2013, G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg were together worked for generating simple image description[3]. This system consists of two parts. The first part, content planning, smooths the output of computer vision-based detection and recognition algorithms with statistics mined from large pools of visually descriptive text to determine the best content words to use to describe an image. The second step, surface realization, chooses words to construct natural language sentences based on the predicted content and general statistics from natural language. We present multiple approaches for the surface realization step and evaluate each using automatic measures of similarity to human generated reference descriptions.

Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, propose a new algorithm in ''Image captioning with semantic attention''[4] , that combines both approaches through a model of semantic attention. Our algorithm learns to selectively attend to semantic concept proposals and fuse them into hidden states and outputs of recurrent neural networks. The selection and fusion form a feedback connecting the top-down and bottom-up computation. They evaluate algorithm on two public benchmarks: Microsoft COCO and Flickr30K. Experimental results show that our algorithm significantly outperforms the state-of-the-art approaches consistently across different evaluation metrics.

K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio [5] propose an attention based approach that gives state of the art performance on three benchmark datasets using the BLEU and METEOR metric. They recommend how to learn attention can be exploited to give more interpretability into the models generation process, and demonstrate that the learned alignments correspond very well to human intuition.

J. Ba, V. Mnih, and K. Kavukcuoglu present an attention-based model [6] for recognizing multiple objects in images. Their model includes a deep recurrent neural network trained with reinforcement learning to attend to the most relevant regions of the input image. Their model model learns to both localize and recognize multiple objects despite being given only class labels during training.

Patel and A. Varier propose a concept " Hyper parameter analysis for image captioning " [7] perform a thorough sensitivity analysis on state-of-the-art image captioning approaches using two different architectures: CNN+LSTM and CNN+Transformer. Experiments were carried out using the Flickr8k dataset. The biggest takeaway from the experiments is that fine-tuning the CNN encoder outperforms the baseline and all other experiments carried out for both architectures.

H. Lu, R. Yang, Z. Deng, Y. Zhang, G. Gao, and R. Lan were together publish a paper ''Chinese image captioning via fuzzy attention-based DenseNet-BiLSTM'' [8] which goes through the challenges of Chinese image description generation. They propose a fuzzy attention-based DenseNet-BiLSTM Chinese image captioning method in this article. Here improve the densely connected network to extract features of the image at different scales and to enhance the model's ability to capture the weak features. At the same time, a bidirectional LSTM is used as the decoder to enhance the use of context information.

3. EXISTING SYSTEM

Existing system uses Residual network, which solves the solve the problem of the vanishing/exploding gradient; this architecture introduced the concept called Residual Blocks. Resnets are made by stacking these residual blocks together. The approach behind this network is instead of layers learning the underlying mapping; we allow the network to fit the residual mapping. So, instead of say H(x), initial mapping, let the network fit. But in the case of image captioning its quite time consuming.

4. PROPOSED SYSTEM

In this paper we prefer EfficientNet is convolutional neural network architecture and scaling method that uniformly scales all dimensions of depth/width/resolution using a compound coefficient. The compound scaling method is justified by the intuition that if the input image is bigger, then the network needs more layers to increase the receptive field and more channels to capture more fine-grained patterns on the bigger image. The compound scaling method is based on the idea of balancing dimensions of width, depth, and resolution by scaling with a constant ratio.

EfficientNet is based on the baseline network developed by the neural architecture search using the AutoML MNAS framework. The network is fine-tuned for obtaining maximum accuracy but is also penalized if the network is very computationally heavy. It is also penalized for slow inference time when the network takes a lot of time to make predictions. The architecture uses a mobile inverted bottleneck convolution similar to MobileNet V2 but is much larger due to the increase in FLOPS. This baseline model is scaled up to obtain the family of EfficientNets

4.1 DATASET

Here we use data sets as Input images and Encoded captions. The set of images obtained from MS COCO must have pixels values in the domain b ∈ {0,1} to be compatible with the pre-trained convolutional model used as the encoder block. For the effect, a normalization of the RGB channels. In order to be able to manipulate the descriptions associated with each image in the dataset, the model uses a mapping system supported by a dictionary. Within this file, each word used in the captioning of the entire dataset has an identification number. In this way, each ground-truth will be represented as a numerical array according to the equivalences defined by the mapping system.

5. CONCLUSIONS

We propose, Transformers based image captioning using EfficientNet model, which is very flexible model compare to other implementation. It takes less time to predict the caption of the image. Another attraction to the proposed model is very easy to implement. This model can be extended for captioning and it will be a very useful for every filed like medical and research.

6. REFERENCES

[1] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, ''Self-critical sequence training for image captioning,'' inProc .IEEEConf .Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 7008–7024.

[2] Y. Yang, C. Teo, H. Daumé, and Y. Aloimonos, ''Corpus-guided sentence generation of natural images,'' in Proc. Conf. Empirical Methods Natural Lang. Process., 2011, pp. 444–454.

[3] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, ''Babytalk: Understanding and generating simple image descriptions,'' IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 12, pp. 2891–2903, Oct. 2013.

[4] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, ''Image captioning with semantic attention,'' in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 4651–4659.

[5] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, ''Show, attend and tell: Neural image caption generation with visual attention,'' in Proc. Int. Conf. Mach. Learn., 2016, pp. 2048–2057.

[6] J. Ba, V. Mnih, and K. Kavukcuoglu, ''Multiple object recognition with visual attention,'' 2015, arXiv:1412.7755.

[7] A. Patel and A. Varier, ''Hyper parameter analysis for image captioning,'' 2020, arXiv:2006.10923.

[8] H. Lu, R. Yang, Z. Deng, Y. Zhang, G. Gao, and R. Lan, ''Chinese image captioning via fuzzy attention-based DenseNet-BiLSTM,'' ACM Trans. MultimediaComput.Commun.Appl.,vol.17,no.1,Mar.2021,Art. no. 48, doi: 10.1145/3422668.

IJARIIE