ORIGINAL RESEARCH PAPER
Volume-8 | Issue-3 | March-2019 | PRINT ISSN No 2277 - 8179
INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH

AUDIO DESCRIPTION OF IMAGE

Computer Science

Parth Zalavadia Student, Dept. of Computer Engineering, St. John College of Engineering and Management, India

Vrutti Patel* Student, Dept. of Computer Engineering, St. John College of Engineering and Management, India *Corresponding Author

Ronak Singh Sahni Student, Dept. of Computer Engineering, St. John College of Engineering and Management, India

Siddhesh Bhagat Assistant Professor, Dept. of Computer Engineering, St. John College of Engineering and Management, India

ABSTRACT

Audio Description of Image is a cornerstone in computer vision, a field which has seen much technical advancement in recent years. Describing the contents and audio of images is a challenge for the machine to achieve it. It requires not only accurate recognition of object and human, but also their attributes and relationship as well as scene information. Audio Description of Image makes initial attempts to deal with the above challenges to produce multi sentence natural language description of image contents and producing the audio of the same. It takes a local region based on the approach to extract regional image details and by combines multiple technique including attribute learning and deep learning through the use of machine learned features to create high level labels that can generate detailed audio description of real-world images.

Audio Description of Image will contain the function of scene classification, object detection and classification, attribute learning, relationship detection, sentence generation and audio generation. Currently caption generation systems do exist, however they are sharing resources to create the RoI (Region of Interest) layer. LSTM (Long Short Term Memory) provides a better approach to meaningful sentence creation using the different objects detected in the image using Convolutional Neural Network (CNN). By blending CNN and LSTM, it is possible to considerably reduce the sum of time and resources consumed by the system, also the caption generated by Audio Description of Image will be semantically more accurate.

KEYWORDS

CNN, RNN, LSTM, RoI, technology.

I. INTRODUCTION

Recent success in the field when applied by Deep Neural Networks to Computer Vision and Human Understandable Language Processing tasks have galvanised Artificial intelligence research are made to explore new research opportunity at the interchange of these previously separated domains. One such application of Deep Neural Networks is automatic description generation for the content of an image using properly formed English sentences and converting the same into audio. The task is significantly hard, for example, the well-studied image classification tasks has been main focusing in the computer vision community. Absolutely, a description must not only capture the objects containing in an image, but it should also consider how these objects are related to each other and their attributes and the activities they are involved. Moreover, the above semantic knowledge is to be expressed into human understandable language like English, which means that a language model is also needed along with the visual understanding.

II. Related Works

Our proposed idea inspired by Dong-Jin Kim. Sentence learning on deep CNN [1]. In this method, a transfer learning CNN is used for image representing with deep fisher kernels. CNN is used for image processing and guided LSTM are used as a sentence generator. Another inspiration is by Minsi Wang-Parallel fusion if RNN-LSTM [2]. In this system, CNN is used for object detection. The sentence are generated using RNN framework. They applies them to extract image features and align visual language data respective. The main inspiration is by Philip Kinghom-Deep learning image description generator [3]. In this method, a fusion of training and detection process is more accurate and for text generation LSTM is used.

From term of result, Dong's system uses guided LSTM which accurate result but it requires additional information takes more time for processing. Whereas Wang's system RNN-LSTM structure which also take more time and is complicated in case of working. So we are using a guided LSTM for generation of sentence as proposed by Philip. In terms object extraction from the image, while comparing the mentioned model RCNN gives a better result than CNN in term of accuracy and time. But still RCNN lags behind CNN in case of accuracy and time. So we are using CNN for object detection from the image and for sentence generation LSTM model is used and that sentence is converted into audio data using text to speech library.

III. ABBREVIATION AND ACRONYMS

LSTM is a module used along Neural Network in order to retain information after the text is generated and then converting it into audio data. LSTMs basically hold information as required. Guided LSTM uses input called “Guide” that contains the information of image. The term guide does not change through time. CNN is a neural network of handling image data. It consists of 3 layers, 1 convolution layer, 1 pooling layer, and 1 fully connected layer. Will look at how each of these work in further depth in the sections. CNN uses separate region proposal network to select region proposals of the input image. The CNN takes a feature map as input. Text to speech library are used to get the audio output.

APPROACH

We have created a layered architecture as shown below. It consists of two part one is CNN for object extraction from image and second part is LSTM for sentence formation.

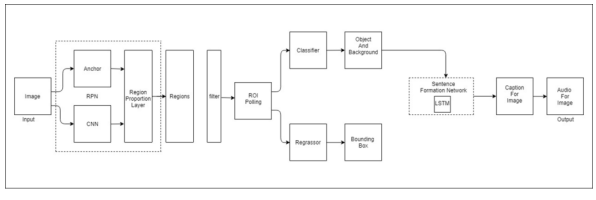

Figure 1. Network Architecture

A. Object Detection

Our object detection model consists of two sections, a deep convolutional network that proposes regions and a CNN spotter uses that proposed regions. The whole system is a network for object detection.

B. Region Proposal Network

As per Figure: 1, the image is feed into RPN (Region Proposal Network). It takes a picture as input and output as a set of selected object with a brightness score. The CNN tells the module where to see in an image and process this model with a CNN [11]. For the selected

region a small window type network is moved over a trait map which is given by the last layer of convolutional network. Each moving window is mapped 1N i^{th} lower dimensional feature (512d for VGG with Relu [13]). This components are fed into two connected layer i.e. box regression layer and box classification layer.

C. Anchors

Each sliding window multiple selected region are predicted. Where the maximum possible proposals are given by k. The regression layer gives 4k output encoding co-ordinates of boxes and classification layer gives 2k output which estimates the probability of objects. The kth proposals are parameterized according to reference boxes. This is called as anchor. An anchor is placed at the center of the moving window with a scale and ratio. We are using 3 scale of 128*128,256*256 and 512*512 and ratio of 1:1, 1:2 and 2:1. For a feature map of region M*N, there are MNk anchors in total.

D. Training RPN

RPN can be trained end to end by back propagation and clastic gradient descent [14]. We are using image centric sapling strategy from [4] to train our network. Each batch creates a single image that contain positive and negative sample anchor. We randomly sample 256 anchors in a group to compute the loss function of a batch. The ratio of positive and negative anchor are 1:1. We randomly initialize entire new layer with a 0 mean Gaussian distribution standard deviation as 0.01. All other layers are initialized by training model on Image Net dataset.

E. Sentence Generation

The core of LSTM is a memory of cell. The behavior of cell are controlled by "gates". The value of gates will be either 1 or 0 particular, 3 gates are being used which controls weather forgot the current cell value (forget gate) if it should read its value (input gate) and weather to output the new cell value (output gate). The definition of gates is as follow in [2]. We have used a softmax for probability distribution of words.

F. Training LSTM

LSTM model is competent to predict each words for framing the sentence after it has seen the picture as well as the preceding words. For this purpose we are using LSTM. A copy of LSTM memory is built from the picture and each words for the sentence such that all LSTM share the same parameter. It represent each word as a vector and we have a start and stop word that we designate the start and end of sentence. The LSTM will give this stop word and signal that the complete.

G. Text to Speech

Text to speech are abbreviated as TTS. It's a form of voice amalgamation that convert sentence into voice output. Text to speech automations were developed to offer a computer generated voice that would read sentences to the user.

IV. Result

A. Accuracy

For accuracy of our system, we are going to use BLEU algorithm to check the precision of captions being generated. BLEU stand for Bilingual Evaluation Understudy. It's an algorithm for evaluation of quality text which machine has generated. The output is always between 0 and 1.

Table 1: Comparison of accuracy on Bleu score

Serial No.	BLEU Score	
	System Name	*Score*
1.	DONG MODEL	0.22
2.	WANG MODEL	0.27
3.	PHILIP MODEL	0.41
4.	AUDIO DESCRIPTION OF IMAGE	0.52

V. CONCLUSION AND FUTURE WORK

We have presented Audio Description of Image, an end-to-end neural network that takes image as an input and automatically generate on a reasonable caption in English and converted into audio data. Audio Description of Image is based on solution neural network that encodes and image in compact: sentation, followed by a LSTM that generates a corresponding caption and then text to speech converted converts it into audio format. The model is trained to maximize the likelihood of the sentence given the image. For qualitative and initiative evaluation BLEU is used to generate proper sentence. Furthermore, it be interesting to see how one can use unsupervised data, from images alone and text alone and audio alone, to improve image on approaches.

REFERENCES

[1] D. Kim, D. YOO, B. Sim, I. Kweon, "Sentence Learning on Deep Convolution Network for Image Caption Generation", 2016 13th International Conference on Ubiquitous Robots and Ambient Intlligence (URAI).

[2] M. Wang, L. Song, X. Yang, C. Luo, "A Parallel-Fusion RNN-LSTM Architecture for Image Caption Generation", 2016 IEEE International Conference on Image Processing(ICIP).

[3] P. Kinghorn, L. Zhang, L. Shao, "Deep Learning based Image Description Generation", 2017 International Joint Conference on Neural Networks(UCNN).

[4] R. Girshick, "CNN", 2015 IEEE International Conference on Computer Vision.

[5] S. Ren, K. He, R. Girshick, J.Sun, "CNN: Towards Real-Time Object Detection with Region Proposal Networks", IEEE Transactions on Pattern and Machine Intelligence, vol.39, issue.6, pp. 1137-1149, 2017.

[6] X.Chen, C.L.Zitnick, "Mind's Eye: A Recurrent Visual Representation for Image Caption Generation", 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2422-2431, 2015.

[7] Athelas.com 'A brief history of CNN's, 2017.[Online]. Available: https://blog.athelas.com/a-brief-history-of-cnns-in-image-segmentation-from-r-cnn-to-mask-r-cnn-34ea83205de4.

[8] Medium.com 'Machine Learning is Fun!' 2017. [Online]. Available:https://medium.com/@ageitgey/machine-learning-is- fun-part-3-deep-learning-and-convolutional-neural-networks-f40359318721.

[9] Tryolabs.com 'Object detection: an overview in the age of deep learning', 2017.[Online].Available:https: //trylabs.com/blog/2017/08/30/olect-detection-an-overview-in-the-age-of-deep-learning.

[10] Tensorflow.com 'Image Recognition | Tensorflow', 2017[Online]. Available:https://www.tensorflow.org/ tutorials/image_recognition.

[11] J. Long, E. Shelhamer, and Darrel "fully connected network for semantic segmentation", in IEEE Conference on computer vision, 2015.

[12] R.Girishiek, J.Donahue, T.Darrell, "Rich Feature for accurate Object Detection and semantic Segmentation in IEEE Conference on computer vision, 2014.

[13] V. Nair and G.E. Hinton, "Rectified linear units improves restricted Boltzmann machine", in International conference on machine Learning, 2010.

[14] Y. Leun, B. Boser, J.S. Denker and L.D. Jackel "Back propagation applied to handwritten zip code recognition", neural computation,1989.