Automatic Image Captioning

Anand dhakne[1], Omkar Mangnale[2], Nihal Kazi[3], Dattray[4]
[1, 2, 3, 4] Dept of Information Technology Engineering
[1, 2, 3, 4] Genba Sopanrao Moze College of Engineering,,Balewadi.

Abstract- *The paper aims at generating automated captions by learning the contents of the image. At present images are annotated with human intervention and it becomes nearly impossible task for huge commercial databases. The image database is given as input to a deep neural network (Convolutional Neural Network (CNN)) encoder for generating "thought vector" which extracts the features andnuances out of our image and RNN (Recurrent Neural Network) decoder is used to translate the features and objects given by our image to obtain sequential, meaningful description of the image. In this paper, we systematically analyze different deep neural network-based image caption generation approaches and pretrained models to conclude on the most efficient model with fine-tuning. The analyzed models contain both with and without 'attention' concept to optimize the caption generating ability of the model. All the models are trained on the same dataset for concrete comparison.*

Keywords- automated captions, deep neural network, CNN, RNN, feature extraction, attention.

I. INTRODUCTION

A large amount of information is stored in an image. Everyday huge image data is generated on social media and observatories. Deep learning can be used to automatically annotate these images, thus replacing the manual annotations done. This will greatly reduce the human error as well as the efforts by removing the need for human intervention The generation of captions from images has various practical benefits, ranging from aiding the visually impaired, to enabling the automatic, cost-saving labelling of the millions of images uploaded to the Internet every day, recommendations in editing applications, beneficial in virtual assistants, for indexing of images, for visually challenged people, for social media, and several other natural language processing applications. The field brings together state-of-the-art models in Natural Language Processing and Computer Vision, two of the major fields in Artificial Intelligence. One of the challenges is availability of large number of images with their associated text ever expanding internet. However, most of this data is noisy and hence it cannot be directly used in image captioning model. For training an image caption generation model, a huge dataset with properly available annotated image is required. In this paper, we plan to demonstrate a system that generates contextual description about objects in images. Given an image, break it down to extract the different objects, actions, attributes and generate a meaningful sentence (caption/description) for the image.

II. SYSTEM ARCHITECTURE

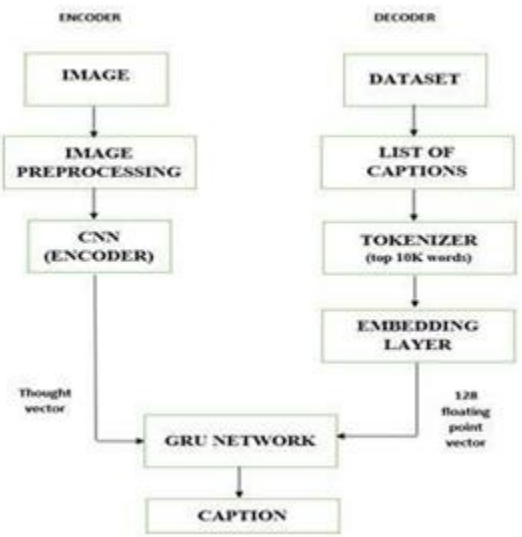

Fig: System Architecture

III. IMPLIMENTATION

To start with automatic image caption generation, image annotation was studied from Image Annotation via deep neural network [1] which proposes a novel framework of multimodal deep learning where the convolutional neural networks (CNN) with unlabeled data is utilized to pre-train the multimodal deep neural network to learn intermediate representations and provide a good initialization for the network then use backpropagation to optimize the distance metric functions on individual modality. This was followed by Automatic image annotation using DL representation [2] in which the last layer of CaffeNet of the CNN based model is replaced with a projection layer to perform regression and the resulting network is trained for mapping images to semantically meaningful word embedding vectors. Advantage of this modelling is: firstly, it does not require dozens of handcrafted features and secondly, the approach is simpler to formulate than any other generative or discriminative models.

A single network is created for generating captions of images in Show and Tell: A Neural Image Caption Generator

[3]. In this network, deep convolutional network is used for image classification and sentence generation is done by a powerful Recurrent Neural

Network which is trained with the visual input so that RNN can keep track of the objects explained by the text. A different approach to caption generation is incorporated in Show, Attend and Tell: Neural Image Caption Generation with Visual Attention [4] where, a form of attention, "hard" attention mechanism and "soft" attention mechanism are described. A deterministic "soft" attention mechanism is employed by standard back-propagation methods and a stochastic "hard" attention mechanism by maximizing an approximate variational lower bound. predefined number of unique words into integer tokens. Once the tokens are assigned, the embedding layer converts integer-tokens into vectors of 128 floating-point number since the RNN network works on vectors and not integers. The sequence vectors are padded to ensure that all the sequence vectors are of same length which is equal to the length of the maximum sequence. The GRU unit comprises of three gates: forget gate, output gate, input gate. The gates are defined as

IV. CONCLUSION

We have presented an end-to-end neural network system that can automatically view an image and generate a reasonable description in plain English. It is based on a convolution neural network that encodes an image into a compact representation, followed by a recurrent neural network that generates a corresponding sentence. The model is trained to maximize the likelihood of the sentence given the image. We also saw the effect of the encoder-decoder approach combined with attention and made analysis.

REFERENCES

[1] Sun Chengjian, Songhao Zhu, Zhe Shi, "Image Annotation Via Deep Neural Network", Published in: 2015 14th IAPR International Conference on Machine Vision Applications (MVA).
[2] Venkatesh N. Murthy, Subhransu Maji, R Manmatha, "Automatic Image Annotation using Deep learning representations", ICMR '15 Proceedings of the 5th ACM on International Conference on Multimedia Retrieva.
[3] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan, "Show and Tell: A Neural Image Caption Generator", published 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
[4] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, Yoshua Benjio, "Show, attend and tell: neural image caption generation with visual attention", ICML'15 Proceedings of the 32ndInternational Conference on Machine Learning – Volume 37, Pages 2048-2057.
[5] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation", Proceedings of the 40th Annual Meeting of theAssociation for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318.
[6] www.deeplearning.ai
[7] www.tensorflow.org