Impact Factor Value 4.046 **e-ISSN: 2456-3463**
National Conference on "Recent Advances in Engineering and Technology" SAMMANTRANA 19
Organized by Government College of Engineering, Nagpur
International Journal of Innovations in Engineering and Science, Vol 4 No.8, 2019
www.ijies.net

Development of an Approach for Image Captioning and Context Writing

Anugrah Chimanekar [1] ,Veer Kalantri [2] ,Mohit Adlakha [3],Aditi Lokhande [4] ,Ayushi Mishra [5]

[1,2,3,4,5] *Student, S.B. Jain Institute of Technology, Management and Research, Nagpur, India - 441501*

Abstract -*The project is based on image captioning with context generation for the generated caption using the features extracted from the input image by the convolutional neural networks. The problem is interesting not only because it has important practical applications, such as helping visually impaired people see, but also because it is regarded as a grand challenge for image understanding which is a core problem in computer vision. Generating a meaningful natural language description of an image requires a level of image understanding that goes well beyond image classification and object detection.*

Keyword: image captioning, context writing, machine learning, convolutional neural networks..

INTRODUCTION

Automatically generating a natural language description of an image, a problem known as image captioning, has recently received a lot of attention in Computer Vision. The problem is interesting not only because it has important practical applications, such as helping visually impaired people see, but also because it is regarded as a grand challenge for image understanding which is a core problem in Computer Vision. Generating a meaningful natural language description of an image requires a level of image understanding that goes well beyond image classification and object detection. What is most impressive about this problem is that it connects Computer Vision with Natural Language Processing which are two major fields in Artificial Intelligence.

OBJECTIVES

- To develop a system that generate caption for an image provided as input.
- To provide a short context for the reason behind the caption generated.

LITERATURE SURVEY

"Survey on Feature Extraction of Images for Appropriate Caption Generation" - Such systems help visually impaired people in understanding pictures. They can be used for providing alternate text for images in parts of the world where mobile connections are slow. [1]

"Show, Attend and Tell: Neural Image Caption Generation with Visual Attention" - Attention based model that automatically learns to describe the content of images & through visualization the model is able to learn to fix its gaze on salient objects while generating the corresponding words in the output sequence. [2]

"A Survey on Auto Image Captioning" - Demonstrate that alignment model produces results in retrieval experiments on datasets such as MS-COCO. [3]

"Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures" - Classify the known approaches based on how they conceptualise this problem and provide a review of existing models, highlighting their advantages and disadvantages. [4]

"Learning CNN-LSTM Architectures for Image Caption Generation" - Implements a generative CNN-LSTM model that beats human baselines by 2.7 BLEU-4 points and is close to matching (3.8 CIDEr points lower) the current state of the art. [5]

METHODOLOGY

The task of image captioning can be divided into two modules logically – one is an image based model –which extracts the features out of our image, and the other is a language based model – which translates the features and objects given by our image based model to a natural

sentence. For our image based model (viz encoder) – we usually rely on a Convolutional Neural Network model. And for our language based model (viz decoder) – we rely on a Recurrent Neural Network.

A. Flow of the system

1) The user will pass image as input to the front end (web app) and the image will be passed to convolutional neural network (CNN) to extract the features from it. These extracted features will form a 2D linear feature vector to pass to the Long Short Term Memory (LSTM) network and then further to the Softmax function for feature extracted sentence generation with Natural Language Toolkit (NLTK). This will generate the approximate caption for the input image which will be printed on the user side output screen. The following flowchart explains this flow for caption generation phase:

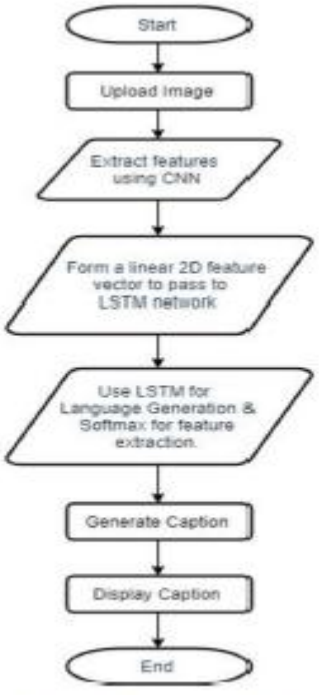

Fig. 1- Image Caption Generation

2) Next, in the inference generation phase, the existing pretrained model for caption generation will also be trained for the image context generation with inference dataset. This dataset is containing 10000 pre-generated inferences for the images with respect to the keywords found in the caption as well as the extracted features with respect to the VGG16 model trained on ImageNet dataset of 1000 objects. After the training is completed, the pre-generated caption will be passed along with the feature vector generated from CNN as input to the model. After analyzing the feature vector and the generated caption, the model will make the most relevant approximate context or inference for the given input image and this context will be printed on user side output screen. The flow of context writing phase is a follows:

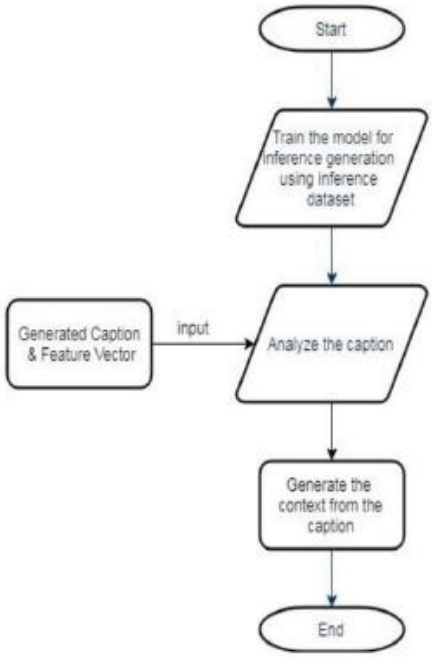

Fig. 2- Context Generation

B. Functional Modules

1) Feature Extraction: The image passed as input will be passed to CNN with various filters and fixed eight for feature extraction. These extracted features will form a 2D Linear feature vector after passing from the CNN i.e. feature extraction phase.

2) Caption Generation: The feature vector will be passed to the LSTM network and then Softmax function for approximate caption generation. This will use Python NLTK library for grammar generation for the sentence along with LSTM and Softmax. As the feature vector is passed through LSTM it goes to Softmax function and the most probable feature is used as the actual feature in the image.

3) Context Generation: After caption generation, the feature vector along with generated caption is passed to the model trained on inference dataset. The model accordingly generated the inference and prints it as

context along with the generated caption with respect to the extracted features and keywords from caption. The sentence generation is done using LSTM and NLTK libraries for Python.

C. Comparison with existing models:

Name Of Model	BLEU1 Score	BLEU2 Score	BLEU3 Score	BLEU4 Score
Google: Show, Attend and Tell	67	45.7	31.8	21.3
Our Project	43.3	23.6	16.5	7

The BLEU metric ranges from 0 to 1. Few translations will attain a score of 1 unless they are identical to a reference translation. For this reason, even a human translator will not necessarily score 1. It is important to note that the more reference translations per sentence there are, the higher the score is. Thus, one must be cautious making even "rough" comparisons on evaluations with different numbers of reference translations: on a test corpus of about 500 sentences (40 general news stories), a human translator scored 0.3468 against four references and scored 0.2571 against two references. **[6]**

CONCLUSION

Hence, we have till now successfully implemented GUI of our project which is going to accept the image from the user and will display the caption. And two modules, Feature Extraction and Caption Generation has been successfully implemented.

REFERENCES

[1] *Survey on Feature Extraction of Images for Appropriate Caption Generation by Aswathy K S, Prof. (Dr.) Gnana Sheela K.*
[2] *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention by Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel,Yoshua Bengio.*
[3] *A Survey on Auto Image Captioning by D. D. Sapkal, Pratik Sethi, Rohan Ingle, Shantanu Kumar Vashishtha, Yash Bhan.*
[4] *Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures by Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, Barbara Plank.*
[5] *Learning CNN-LSTM Architectures for Image Caption Generation by Moses Soh.*
[6] *BLEU: a Method for Automatic Evaluation of Machine Translation Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu*