Image Classification And Captioning Using AI

S.Rakesh[1], K.Prakashraj[2], Dr. M. Babu[3]
[1, 2] Dept of Computer Science and Engineering
[3]Professor, Dept of Computer Science and Engineering
[1, 2, 3] GKMCET, Chennai, Tamilnadu, India.

Abstract- *In the past few years, the problem of generating descriptive sentences automatically for images has gained interest in natural language processing and computer vision research. Image captioning, in which the main aim is to generate a caption description for an image automatically, has attracted a lot of research attention in the field of cognitive computing. The meaningful description generation process of high level image semantics requires not only the understanding of objects or scene recognition in the image, but also the ability to analyze their states, understand the relationship among them and generate a semantically and syntactically correct sentence. Image caption generator is a popular research area of AI that deals with image understanding and a language description for that image. Simply it is the process of recognizing the context of an image and annotating it with relevant captions using deep learning, and computer vision. It includes the labeling of an image with English keywords by using the datasets provided during model training.*

I. INTRODUCTION

Nowadays, We apply AI in building powerful performance and highly intelligent machines and programs. Machine learning has a subset called deep learning , it provides high accuracy with its results so its performance is high through its output. Image description provides the process of describing the content from an image. Caption generating is an important task that is relevant to both computer vision and natural language processing. Computer vision module is for detecting features from objects or extracting features of images and NLP module is for generating correct syntactic and semantic image captions.

II. OBJECTIVES

By seeing an image human beings can say what is happening in the picture and explain the connection between the objects in the image. But for a machine classifying and generating the related descriptions for the images is a difficult task. This difficult task is to be done using AI algorithms. This captioning model is used for recommendations in editing applications, usage in virtual assistants, for image indexing, for visually impaired persons, for social media, and several other natural language processing applications.

III. LITERATURE REVIEW

A literature review surveys books, scholarly articles, and any other sources relevant to a particular issue, area of research, or theory, and by so doing, provides a description, summary, and critical evaluation of these works in relation to the research problem being investigated.

A. NEURAL ATTENTION FOR IMAGE CAPTIONING.

Image captioning is the task of automatically generating sentences that describe an input image in the best way possible. Some of the techniques used for generating caption used attentive deep learning models. In this survey they provide a review of literature related to attentive deep learning models for image captioning. We aim at finding the most successful types of attention mechanisms in deep models for image captioning. Soft attention, bottom-up attention, and multi-head attention are the types of attention mechanisms widely used in state-of-the-art attentive deep learning models for image captioning.

B. AUTOMATIC GENERATION ON FOOD IMAGE AESTHETIC CAPTIONING.

In this paper, they propose a novel model to generate aesthetic captions for food images. This model redefines food image aesthetic captioning as a compositional task that consists of two separated modules, i.e., a single-aspect captioning and an unsupervised text compression. The first module is guaranteed to generate the captions and learn feature representations of each aesthetic attribute. Then, the second module is supposed to study the associations among all feature representations and automatically aggregate captions of all aesthetic attributes to a final sentence. Experiments on the dataset demonstrate the effectiveness of the proposed model.

C. TOPIC-ORIENTED IMAGE CAPTIONING BASED ON ORDER-EMBEDDING.

We present an image captioning framework that generates captions under a given topic. The topic candidates are extracted from the caption corpus. A given image's topics are then selected from these candidates by a CNN-based multi-label classifier. The input to the caption generation model is an image-topic pair, and the output is a caption of the image. For this purpose, a cross-modal embedding method is learned for the images, topics, and captions. In the proposed framework, the topic, caption, and image are organized in a hierarchical structure, which is preserved in the embedding space by using the order-embedding method.

D. NEWS IMAGE CAPTIONING BASED ON TEXT SUMMARIZATION USING IMAGE AS QUERY.

News image captioning aims to generate captions or descriptions for news images automatically, serving as draft captions for creating news image captions manually. News image captions contain more detailed information such as entity names and events than generic image captions do. Detailed information is usually contained in news text but not in news images. Generic image captioning does not make full use of the accompanying news text to generate image captions. This paper proposes a news image captioning method based on the attentional encoder-decoder model through summarizing the news text according to the query image. The multi-modal attentional mechanism is proposed to compute the context vector. The proposed model is trained on the DailyMail news image captioning corpora which are created by collecting images, caption, news texts through parsing the html-formatted documents.

E. GENERATING IMAGE CAPTIONS IN ARABIC USING ROOT-WORD BASED RECURRENT NEURAL NETWORKS AND DEEP NEURAL NETWORKS.

Despite advanced research in English caption generation, research on generating Arabic descriptions of an image is extremely limited. Semitic languages like Arabic are heavily influenced by root-words. We leverage this critical dependency of Arabic and in this paper are the first to generate captions of an image directly in Arabic using root-word based Recurrent Neural Networks and Deep Neural Networks. This report the first BLEU score for direct Arabic caption generation. Experimental results confirm that generating image captions using root-words directly in Arabic significantly outperforms the English-Arabic translated captions using state-of-the-art methods.

III. PROPOSED SYSTEM

In this proposed model, we are proposing recognition an image and classify image based on the structured dimensional Convolutional Neural Networks (CNNs) type of VGGNet to classify the image and improve the accuracy of workflow. The proposed method for this project is to train a Deep Learning algorithm capable of classifying the images and data preprocessing and visualizing the image then feature extracting through CNN using given image dataset. We can classify the whole project and the task can be divided into two modules: Image based module-which uses CNN for extracting features from image and Language based model-which uses LSTM for generating sentences. it such as Covid-19 or Normal this system using CNN model.It is predicted that the success of the obtained result’s accuracy and related sentence will increase if the CNN and LSTM method is trained by adding more training dataset and increases the period of training in the system.

IV. SYSTEM ARCHITECTURE

The system architecture gets an image as input to the trained deep learning model. This deep learning model is made with the algorithms of cnn-lstm for image captioning and a image dataset is used to create a trained model.

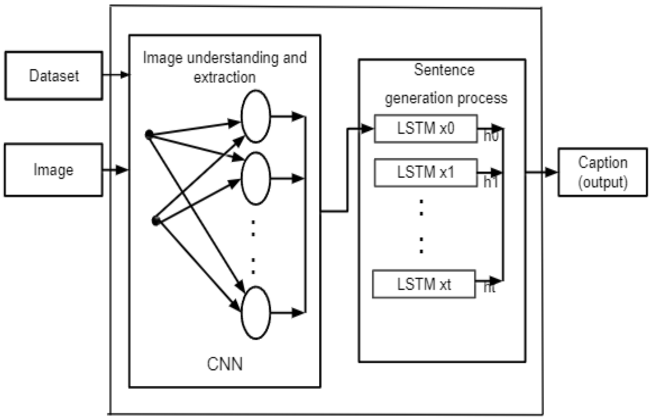
Dataset
Image
Image understanding and extraction
CNN
Sentence generation process
LSTM x0
h0
LSTM x1
h1
LSTM xt
ht
Caption (output)

Figure 4.1: System architecture

V. USE CASE DIAGRAM

This diagram shows various use cases and different types of users that the system has.The main purpose of the use-case diagram is to help development teams visualize the functional requirements of a system, including the relationship of “actors”, who will interact with the system to essential processes, as well as the relationships among different use cases.

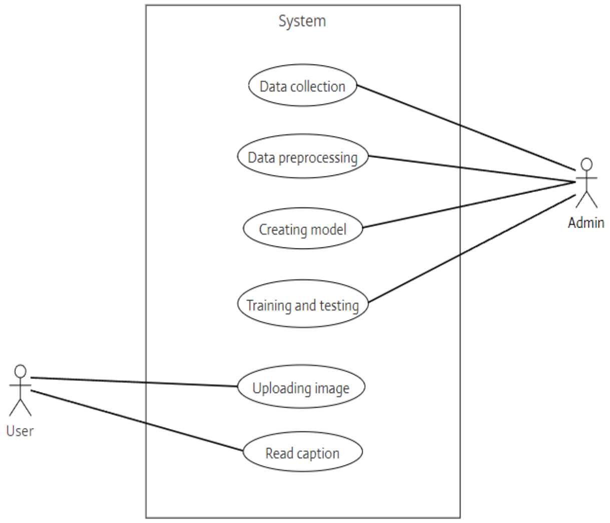

Figure 4.2: Use case diagram

VI. MODULE DESCRIPTION

DATA COLLECTION AND PREPROCESSING:

We have to import our dataset and clean it to create an efficient trained model. There are many open source datasets available for this problem, like Flickr 8k (containing 8k images), Flickr 30k (containing 30k images), MS-COCO (containing 180k images), etc.Data collection is the process of gathering and measuring information on targeted variables in an established system, which then enables one to answer relevant questions and evaluate outcomes.

Data cleaning is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted. But, as we mentioned above, it isn’t as simple as organizing some rows or erasing information to make space for new data.

TRAINING THE MODEL:

To train our model the required algorithms for the CNN and LSTM should be implemented. We do training of model by increasing the epoch value which may increases the efficiency. We can use the deep CNN architecture to extract features from the image which are then fed into the LSTM architecture to generate the caption. The feature vector is linearly transformed to the LSTM network. This network is trained as a language model on our feature vector.

Output Layer:

First this system asks to upload an image for which captions should be generated. Then it will load the image into the memory and extract the image details. It will later enter into the LSTM system to produce meaningful relevant sentences related to the action in the image.

VII. FUTURE WORK

Generating captions for images is achieved through this model. But for captioning a video frame by frame is a challenging and complex task as compared to image captioning. The image captioning model should be developed for many languages that can be used for different purposes around the world.

VIII. CONCLUSION

We have implemented an image caption generator with a trained model of CNN-LSTM. From the generated output it shows that the captions are more accurate than the captions which are generated by existing systems. By observing the pattern we can say that this system will generate more accurate result when trained using larger dataset.

REFERENCES

[1] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth (2010), “Every picture tells a story: Generating sentences from images”. European Conference on Computer Vision: Part IV, ECCV’10, pages 15–29, Berlin, Heidelberg.

[2] Andrej Karpathy and Fei-Fei Li (2014) “Deep visual-semantic alignments for generating image descriptions”. CoRR, abs/1412.2306.

[3] Chen Fang, Quanzeng You, Hailin Jin, Zhaowen Wang and Jiebo Luo (2016) “Image captioning with semantic attention”. CoRR, abs/1603.03925.

[4] Dumitru Erhan, Oriol Vinyals, Alexander Toshev and Samy Bengio (2014) “Show and tell: A neural image caption generator”. CoRR, abs/1411.4555.

[5] Girish Kulkarni,Simin Li, Tamara L. Berg, Alexander C. Berg, and Yejin Choi (2011) “Composing simple image descriptions using web-scale n-grams”. Computational Natural Language Learning, CoNLL ’11, pages 220–228, Stroudsburg, PA, USA.

[6] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille (2014) “Deep captioning with multimodal recurrent neural networks (m-rnn)”. CoRR, abs/1412.6632.

[7] Jurgen Schmidhuber and Sepp Hochreiter (1997) “Long short-term memory. ¨ Neural Comput.”, 9(8):1735–1780, November.

[8] Kate Saenko, Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan and Trevor Darrell (2014) “Long-term recurrent convolutional networks for visual recognition and description”. CoRR, abs/1411.4389.

[9] Lawrence. C Zitnick and Xinlei Chen (2014) “Learning a recurrent visual representation for image caption generation”. CoRR, abs/1411.5654.

[10] Polina Kuznetsova, Vicente Ordonez, Alexander C. Berg, Tamara L. Berg, and Yejin Choi (2012) “Collective generation of natural image descriptions”. Long Papers - Volume 1, ACL ’12, pages 359–368, Stroudsburg, PA, USA.

[11] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel (2014) “Unifying visual-semantic embeddings with multimodal neural language models”. CoRR, abs/1411.2539.