Image Descriptor Generator Using Encoder-Decoder Model

Snigdha Biswas
Department of Computer Science and Engineering
Graphic Era Deemed to be University
Dehradun, India
snigdhaabiswas@gmail.com

Sachin Sharma
Department of Computer Science and Engineering
Graphic Era Deemed to be University
Dehradun, India
sachin.cse@geu.ac.in

***Abstract*— A visual representation is an image of a situation, object, or person, whereas a description is a description of the picture action. The generation of an image description is critical in many industries, including medicine, advertising, and robotics. In this paper, we considered the well-known deep learning problem of generating a textual description of an image. We used InceptionV3 and Beam search to run our picture descriptor generator.**

Keywords— Caption-generation, InceptionV3, Encoder-Decoder, Attention Model

I. INTRODUCTION

Image descriptor generator means the process of automatically generating an image description. The textual description of the visual is essentially a lesson. Since this field of research has recently come to the fore, the researchers are focusing two main advanced fields such as natural language processing computer vision. The generation of a descriptive grammatically correct phrase needs both semantic and syntactic language knowledge. Applications from the analysis of X-ray abnormalities to grouping photos by type of similar mountain, food, concert, etc. An algorithm for image descriptor is a function that uses an image and outputs descriptors/vector features. Feature descriptors encrypt interesting details into a series of numbers and act as a kind of digital "fingerprint" for differentiating one feature from another. Ideally, this data would be invariant in image transformation, so that the feature can be found again even if the image is transformed. Deep learning is an artificial intelligence subset of machine learning (ML), on the other hand. The term 'artificial intelligence' refers to computer techniques that imitate human behaviour. ML is a series of data-trained and all-in-one algorithms. Deep learning, by contrast, is just a type of ML based on a structure of the human brain. Deep learning algorithms attempt, through an ongoing analysis of data with the specific logical structure, to draw similar conclusions to people. Deep learning employs a multilayered algorithm structure, known as neural networks, to achieve this. In this paper, we have used the encoder-decoder attention model along with the pre-trained InceptionV3 to train our model to generate grammatically corrected visual description for the input image. The encoder-decoder model helps us to solve the famous sequence to sequence prediction problems by generating a pattern for using recurrent neural network problem statements. Attention model is a further add on to it as it improves the performance of the model on encountering longer sequences. InceptionV3 on the other hand is a pertained model, which was trained on the ImageNet dataset [1]. Instead of VGG and ResNet, we have chosen Incept V3 to save the Ram because weights are less than VGG and ResNet for Inception v3 [2]. For producing our results, we have also used a greedy research and beam search. The world that most likely comes in the sequence is selected in greedy search [13]. Whereas in beam search it expands all following words possible and maintains the K as likely to be any number specified by the user instead of greedily selecting the next word.

II. LITERATURE REVIEW

To fulfil the need of captioning a given image, we have to gather the semantic information of the image and express the same with the help of natural language [6], [9]. There are thus two subfields of computer vision and processing of natural languages in this research area [10], [11], [12]. Different approaches to this problem have been suggested. Many scientists have tried to build and automate the image-captioning process [14], [15]. In [3] the authors have given us an insight about the numerous approaches and datasets available to tackle the famous image-captioning problem in artificial intelligence. In [4] a convolutional neural network (CNN) that is the ResNet and VGG-16 architecture have been used for identifying the images and long short term memory and gated recurrent unit is used for the language processing. An automated process called SPICE was introduced in [5], it basically analyses the semantic structure of the caption and help us to judge the quality of generated captions for different available datasets like Flickr 8K, Flickr 30K and MS COCO. The author Tanti (et al) [8] has proposed the availability of two kinds of architectures-inject and merge. In Inject, a recurrent neural network (RNN) block takes the tokenized captions and image vectors as input. On the other hand, the merge architecture allows us to pass only the tokenized captions as input into the RNN block. The output obtained from this is then further merged with the image.

III. METHODOLOGY

a. *Encoder-Decoder:* The encoder refers to the reading the input sequence word by word and then stores it in the internal state vectors [Fig. 1]. It then produces a context vector as output. This output is the context vector further acts as the input to the decoder unit. Two special tokens -**<start>** and **<end>** are are added to the sequence before we pass the input to based on the context vector, the output is generated by the decoder. The major

disadavntage of this architecture is that the encoder is unable to produce a single context vector for a long sequence. This makes memorising the encoder difficult. By using the attention model, this problem can be overcome. It increases the importance of particular parts that result in the target sequence of the source sequence. Help us to better encode the long sequences. It allows the model to prioritise words while encoding them into a context vector in the source sequence.

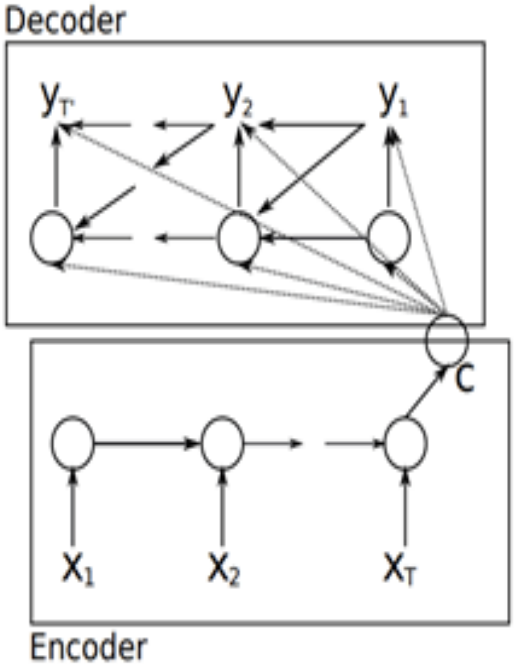

Fig. 1. Process of Encoder – Decoder.

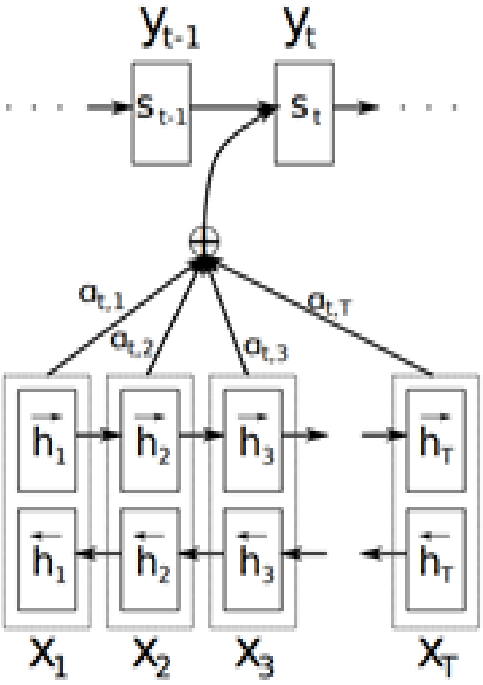

Fig. 2. Process of Transfer Learning-InceptionV3.

b. *Transfer Learning-InceptionV3:* Computer vision transfer refers to the neural network model, which was permitted to train in a series of images in different categories before actually using them on current training data [Fig. 2]. It enables us to load a pre-trained model that is trained in the ImageNet database using a million images of several classes [1]. InceptionV3 was used for visual recognition and analysis. The "Inception" micro-architecture was first presented when Szegedy et al. talked about it in 2014. [*Going Deeper with Convolutions*]. It served the purpose of a multi-level feature extractor by calculating *1×1*, *3×3*, and *5×5* convolutions within a single module of the network. It was referred to as *GoogLeNet*, but new demonstrations and versions have been given the term *Inception vN* in which N signifies the version number specified by Google. The keras core has the InceptionV3 architecture that was discussed in the subsequent year by Szegedy et al. in [Rethinking the Inception Architecture for Computer Vision (2015)]. We have chosen InceptionV3 because the weights smaller than both VGG and ResNet.

IV. Implementation and Result Analysis

a. *System Requirements and Dataset:* We implemented proposed model using a laptop with a stable Internet connection. We did it with Google Colab. It is a free cloud service for learning and research in machinery. The Jupyter notebook interface is similar. The runtime for deep learning and free access to a robust GPU and TPU is fully configured. All configurations were set to standard except for the hardware accelerator. From the drop-down menu we selected the TPU option. For implementing our idea, we used the FLICKR 8K dataset comprising of eight thousand images with 4 captions per each image stored in a text file.

b. *Dataset:* There are different datasets such as FLICKR 8K, FLICKR 30K, and MS-COCO available across the internet for image captioning. We downloaded the Flickr 8K dataset from ‘Kaggle’ as the other datasets were too expensive in terms of hardware requirements because of their size. Our dataset consisted of two sub-folders. One folder consisted of eight thousand images of different actions while the other folder named ‘caption.txt’ comprised of 4 captions each for every image. The caption described the major feature or action of the visual [Fig. 3].

c. Training: First, we imported all the necessary libraries like system libraries, data analysis libraries and tenser flow libraries. After importing the FLICKR 8K dataset for our model from ‘Kaggle’, we visualized the images and captions using different data analysis techniques and created mappings to understand our data. On doing this, we realized that each image is linked with 4 unique captions. Then we distributed the data into data frames to make the access and manipulations of data easy. Using matplotlib, we visualized the top thirty repetitive words from text file [Fig 4].

Fig. 3. Viewgraph of used dataset for implementation.

Second, we performed the pre-processing by tokenizing the captions and generated a vocabulary of top five thousand words mentioned in 'captions.txt'. Then, we padded all the sequences as the longest length. After this, we loaded the pre-trained ImageNet weights of InceptionV3 and extracted the features of the images from the "images" folder using the last layer of pre-trained model. After this, we integrated both images and descriptions together to create the final dataset. After shuffling, with a random state of forty-two, the dataset was slit in an eighty – twenty ratio.

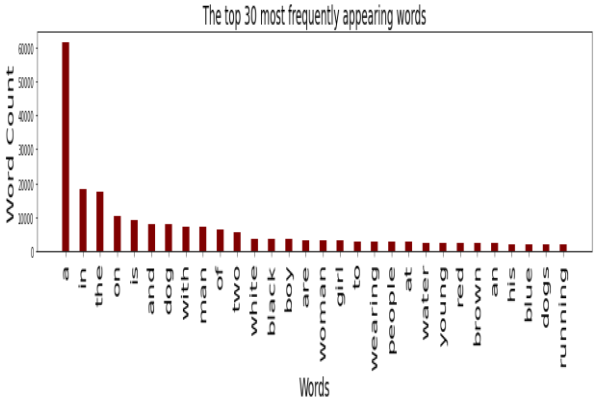

Fig. 4. View of Training dataset.

Third, using the last layer of pre-trained model-InceptionV3, we extracted the features of each image. The shape of each extracted image was 8x8x2048.After this, we built the encoder after calculating the attention weights. In the decoder, we set the optimizer and loss object. Then, we created the training and testing step functions along with the loss function for the test dataset. Lastly, we used greedy search and beam search to generate the captions for unseen images. In, Greedy Search 1 best word is selected as an input sequence every time. It can result into grammatically weak sentences while writing a long sentence. Whereas, the beam search allows us to choose more than one variation for an input sequence using conditional probability. This number of variations is specified using a parameter called the 'Beam Width- B'. It chooses B number of best alternate solutions with the highest probability.

Performance Measurement: On passing an unseen image [Fig. 5], beam search helped us to generate a caption for the image. It was then tested against the given data and Bilingual Evaluation Understudy (BLEU) score for the same was generated. It was first suggested by Kishore Papineni, et al [7]. BLEU is an algorithm for evaluating text quality machine-translations. BLEU was one of the first measures claiming to be highly correlated to human quality judgments and remains one of the most popular automated and cheap metrics. It refers to a metric for comparing a generated caption with its original references. 1 is considered to be an optimal case where as 0 refers to an imperfect mismatch. The actual caption of our image [Fig. 5] is 'a man walks down the road leading a cow with no rider and another cow with a rider' whereas the predicted caption is 'A man walks past a cow'. The BLEU score was 90%. Thus, we can say that our model almost predicts grammatically and visually correct sentences.

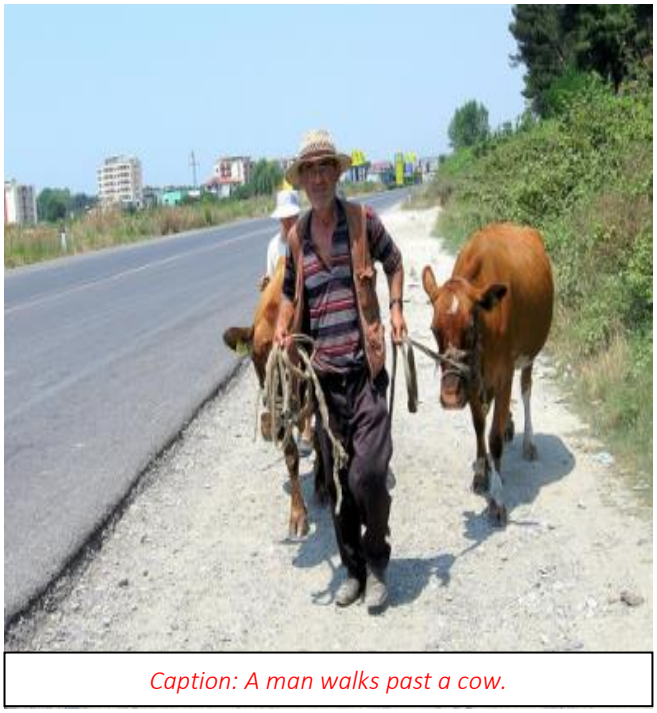

Fig. 5. Result of proposed model.

V. Conclusion And Future Work

This paper investigates image recognition using transfer learning (InceptionV3), and we implemented our idea using

the encoder decoder architecture in conjunction with the attention model. InceptionV3 was chosen over VGG and ResNet because it has fewer weights and thus uses less RAM. Image descriptors in a computer view are descriptions of the content visual properties in pictures, videos, or algorithms or applications producing these descriptions. They describe, inter alia, elemental properties like shape, colour, texture or motion. The title generated by our model for the unseen images, as well as the corresponding BLEU score, demonstrated the success of our approach. There are numerous applications for caption generation in our daily lives. This project can be tailored to meet the needs of a specific industry, such as health care or robotics. The healthcare industry uses this to automatically detect abnormalities in X-ray or CT scans, and, like robotics, the generation of captions can be used to train the robots and familiarise them with the environment. At the same time, by building on this concept in the future, we can give back to society. We can connect a text to a translator for speech and a real-time camera to help the blind better understand their surroundings.

REFERENCES

[1]. Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[2]. Snigdha Biswas, Anirudh Ghildiyal, and Sachin Sharma. "Classification of Indian Dance Forms using Pre-Trained Model-VGG." In 2021 Sixth International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), pp. 278-282. IEEE, 2021.

[3]. Bernardi, Raffaella, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. "Automatic description generation from images: A survey of models, datasets, and evaluation measures." Journal of Artificial Intelligence Research 55 (2016): 409-442.

[4]. Sharma, Grishma, Priyanka Kalena, Nishi Malde, Aromal Nair, and Saurabh Parkar. "Visual image caption generator using deep learning." In 2nd International Conference on Advances in Science & Technology (ICAST). 2019.

[5]. Feng, Yansong, and Mirella Lapata. "Automatic caption generation for news images." IEEE transactions on pattern analysis and machine intelligence 35, no. 4 (2012): 797-812.

[6]. Piyush Juyal, and Sachin Sharma. "Locating people in Real-World for Assisting Crowd Behaviour Analysis Using SSD and Deep SORT Algorithm." In 2021 Sixth International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), pp. 350-353. IEEE, 2021.

[7]. Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. "Bleu: a method for automatic evaluation of machine translation." In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp. 311-318. 2002.

[8]. Blandfort, Philipp, Tushar Karayil, Damian Borth, and Andreas Dengel. "Image captioning in the wild: how people caption images on Flickr." In Proceedings of the Workshop on Multimodal Understanding of Social, Affective and Subjective Attributes, pp. 21-29. 2017.

[9]. Anirudh Ghildiyal, Sachin Sharma, and Ayush Kumar. "Street Cleanliness Monitoring System using Deep Learning." In 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), pp. 868-873. IEEE, 2021.

[10]. Amit Juyal, and Sachin Sharma. "A Study of Landslide Susceptibility Mapping using Machine Learning Approach." In 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), pp. 1523-1528. IEEE, 2021.

[11]. Anirudh Ghildiyal, Komal Singh, and Sachin Sharma. "Music Genre Classification using Machine Learning." In 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 1368-1372. IEEE, 2020.

[12]. Piyush Juyal, and Sachin Sharma. "Detecting the Infectious Area Along with Disease Using Deep Learning in Tomato Plant Leaves." In 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), pp. 328-332. IEEE, 2020.

[13]. Vinchoo, Mohit Manoj, and Rugved Vivek Deolekar. "Comparative analysis of different approaches to solve the job assignment problem." In 2017 International Conference on Trends in Electronics and Informatics (ICEI), pp. 129-134. IEEE, 2017.

[14]. Anirudh Ghildiyal, Sachin Sharma, Ishita Verma, and Urvi Marhatta. "Age and Gender Predictions using Artificial Intelligence Algorithm." In 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), pp. 371-375. IEEE, 2020.

[15]. Piyush Juyal, and Sachin Sharma. "Estimation of Tree Volume Using Mask R-CNN based Deep Learning." In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1-6. IEEE, 2020.