CLIP-Based Image Caption Generation Using Transformer Decoder

Mr. V. Jeevan Kumar,M.Tech
Assistant Professor
Dept. of CSE (Data Science)
Rajeev Gandhi Memorial College of Engineering and Technology
Nandyal, India
jeevanvarda@gmail.com

Anil Kumar Reddy
Dept. of CSE (Data Science)
Rajeev Gandhi Memorial College of Engineering and Technology
Nandyal, India
23095A3203@rgmcet.edu.in

Korlamadugu Ravi Shankar Sai
Dept. of CSE (Data Science)
Rajeev Gandhi Memorial College of Engineering and Technology
Nandyal, India
23095A3216@rgmcet.edu.in

Surapureddy Pushpa
Dept. of CSE (Data Science)
Rajeev Gandhi Memorial College of Engineering and Technology
Nandyal, India
23095A3214@rgmcet.edu.in

***Abstract*—Generating image captions is an essential process in computer vision and natural language processing, where an automatic description of the image is generated through a computer algorithm. Image captioning has attracted considerable attention owing to its usage in various domains such as assistive technologies, content management, and human-computer interaction. This paper presents a smart image caption generator that integrates deep learning algorithms and effective feature extraction methods to create precise and coherent captions.**

The proposed intelligent image caption generator adopts the CLIP model, which enables the extraction of high-quality semantic features from the image. The model uses a Transformer decoder to create the captions. The system is trained using the TextCaps dataset, which consists of images and their corresponding captions, thus enabling the learning of the relationship between the visual aspect and text. The beam search decoding technique is implemented to enhance the quality of the generated captions by choosing the appropriate sequence of words.

In order to ensure that the system is easy to use and deployable in the real world, a web application is built using Streamlit, allowing users to either upload images or take pictures through their device cameras. Moreover, features like activity detection and text-to-speech improve the ease of use and accessibility of the model. The proposed method proves that using CLIP for feature extraction and Transformer for caption generation is effective.

***Index Terms*—Image Captioning, CLIP, Transformer Decoder, Beam Search, Deep Learning, TextCaps, Streamlit, Action Detection, Text-to-Speech**

I. Introduction

The area of image captioning belongs to the realms of computer vision and natural language processing. The idea here is to generate meaningful text from pictures. Applications of such technology include assistive devices for the visually impaired people, picture indexing, and image retrieval systems among others. Given the vast amount of multimedia information that exists today, there is an ever-growing demand for intelligent systems that have the capability to understand visual data and provide textual descriptions of them.

Captioning algorithms based on conventional image captioning methods made use of CNNs paired with RNNs, specifically LSTMs. While these models produced decent results, they faced restrictions in terms of capturing dependencies and contextual relations in sequences. Deep learning research in recent times has seen significant progress in the realm of sequence generation due to the emergence of Transformer models with attention mechanisms.

In this project, we suggest the implementation of an image captioning system that uses CLIP (Contrastive Language-Image Pretraining) for feature extraction from the images, and the transformer-based decoder network to generate captions for the input images. In particular, the suggested system was trained on the TextCaps dataset and utilizes beam search as its decoding method. Also, a Streamlit-based interface is provided as part of our project that enables image captioning, as well as allows to detect actions in the images and produce captions in speech form.

II. Literature Review

Captioning images has been a research domain that involved both computer vision and NLP technologies. In the early days, templates were mostly used to produce captions, which included using fixed sentence structures to create descriptions of the objects detected in the image. However, even though this was an easy task to achieve, it did not offer any flexibility. The emergence of deep learning models changed everything about image captioning, and now encoder-decoder models were used for the purpose. The “Show and Tell” model used CNNs to extract features from the image and RNNs, particularly LSTMs, to generate captions. The model marked the beginning

of end-to-end learning in image captioning, and soon after, attention mechanisms began to be used to make the model capable of focusing on different parts of the picture while creating captions.

On the other hand, Transformer models have recently outperformed the classical recurrent network architectures due to their capability of learning long-term dependencies through attention mechanisms. In this way, Transformer models provide higher levels of parallelism, thereby leading to more accu- rate captions. Vision language models like CLIP (Contrastive Language-Image Pretraining) are also becoming more and more popular because of their ability to learn semantic features on a large scale of image-text pairs. This pre-trained model has also shown good performance on extracting features.

Moreover, many researchers have tried to improve the quality of the generated captions by employing decoding techniques, such as beam search, which considers several sequences before finding the best one. Finally, there has been an increased focus on real-time captioning models and their integration with user interfaces.

However, these developments do not come without their own set of issues, including bias in the data set, lack of fine granularity, and sometimes the production of generic captions. To tackle these issues, the proposed system uses a combination of feature extraction using CLIP and transformer decoder with beam search decoding and other factors like action detection and text-to-speech, creating an efficient image captioning tool.

III. PROPOSED SYSTEM

The proposed system is a deep learning–based image caption generator model that creates descriptive texts for the given image inputs. It uses an encoder-decoder approach, where CLIP is used for feature extraction, and the transformer decoder generates the output captions for the image.

A. *A. Image Input and Preprocessing*

There are two approaches for taking input of the images within this framework: uploading the image or capturing the image from the camera using Streamlit. The uploaded image goes through preprocessing which includes normalization and resizing based on the CLIP model requirements.

B. *B. Feature Extraction using CLIP*

The processed image is fed into the CLIP (ViT-B/32) encoder to generate a feature vector that captures the semantics of the image. The CLIP model was trained on huge data sets of image-text pairs and is able to recognize strong associations between the two. The output is used as input for generating the captions.

C. *C. Caption Generation using Transformer Decoder*

The generated feature vector will then be passed to a decoder that consists of a Transformer. This decoder generates the sentence caption for the image one word at a time. This Transformer model utilizes mechanisms such as self-attention and cross-attention to model dependencies and align word features.

D. *D. Beam Search Decoding*

The following decoding technique is used to enhance the captions generated by the neural network: beam search. In contrast to greedy decoding, which chooses the maximum probability of word selection at each stage, beam search considers multiple paths and uses the accumulated probability to select the best path.

E. *E. Action Detection Module*

Another module is added for detecting actions that occur within the caption generated. Actions within the caption are detected by analyzing the verbs in the caption and then mapping them to a corresponding action description, such as 'sitting', 'walking', and 'playing'.

F. *F. Text-to-Speech Integration*

To enhance accessibility, the generated caption is then spoken out loud through an application of the text-to-speech technology. This tool proves valuable for blind people as it enables them to listen to the image description.

G. *G. Streamlit-Based User Interface*

The whole process is performed using the Streamlit web app, which gives users an intuitive and interactive experience. They can input their images, produce captions, see confidence levels, detect actions, and listen to sound output directly from the interface.

IV. DATASET AND PREPROCESSING

The effectiveness of image captioning heavily relies on the quality and variety of the dataset used during the training process. For this particular experiment, TextCaps dataset is used. This dataset is specially developed for image captioning and includes images that contain text within their scenes. TextCaps has around 28,000 images, with more than 145,000 captions assigned to each of them. Texts appear in the form of captions describing several aspects of an image in different sentences.

A. *A. Dataset Characteristics*

TextCaps datasets are rich sources of images that have both complex visual and textual content. In this case, each image comes along with a series of captions that the model can use to generate descriptions for that visual scene in various ways. This will improve the generalization performance of the model.

B. *B. Image Preprocessing*

Prior to feeding the images into the model, some preprocessing operations are conducted. The first step includes resizing the input images into a common resolution that matches CLIP architecture. Normalization of the pixel values of the image in accordance with the pretraining distribution of CLIP is also done.

C. C. Caption Preprocessing

The text captions corresponding to each image are preprocessed in preparation for the training process. The text captions undergo tokenization to break down each sentence into individual words/tokens. Tokens including start-of-sequence token (<s>), end-of-sequence token (</s>), and padding token (<pad>) are included in order to preserve the structure of the sequence. This is followed by the creation of a vocabulary through assigning a numeric index to each distinct word.

D. D. Feature Extraction

Once the images have been processed, they are fed into the CLIP model encoder to obtain the feature vectors. These feature vectors contain high-level semantic representations of the image which are then fed into the transformer decoder. The precomputation of these features makes training faster.

E. E. Data Splitting

The dataset is partitioned into a training set and a validation set to determine the efficiency of the model. An example partition would be where the training set comprises 80 percent of the data and the validation set comprises the remaining 20 percent.

F. F. Data Handling and Optimization

For better performance during training, features and captions are stored and retrieved through batch operations. This approach minimizes memory consumption and maximizes computational speed. Pre-computed CLIP features provide an additional advantage for training by eliminating the need to extract features repeatedly.

V. MODEL ARCHITECTURE

The design of the system is based on an encoder-decoder framework used to generate captions from images. The three major building blocks of the model include the CLIP image encoder, the caption generator that uses Transformers, and the beam search decoder.

A. A. CLIP Image Encoder

The first element of the framework is the CLIP model (Contrastive Language–Image Pretraining), which functions as the image encoder. For a particular input image I, the CLIP encoder produces a high-level semantic vector:

$$F = CLIP(I) \tag{1}$$

where $F \in \mathbb{R}^{512}$ is the image embedding vector. The CLIP model learns on large image-text datasets which enables it to learn a lot about the correlation between vision and language.

B. B. Transformer Decoder

The feature vector F is then fed into the decoder of the Transformer. The decoder creates caption by generating words one-by-one. The use of self-attention and cross-attention helps the decoder capture dependencies among words and align them with the image features.

The generation of a caption sequence $Y = (y_1, y_2, \ldots, y_T)$ is given by the following probability equation:

$$P(Y|F) = \prod_{t=1}^{T} P(y_t | y_1, y_2, \ldots, y_{t-1}, F) \tag{2}$$

The decoder module in the Transformer model is made up of many layers, which incorporate multi-head attention and feed-forward networks.

C. C. Beam Search Decoding

Beam search decoding technique is used in order to enhance the quality of captions generated by the model. As compared to greedy decoding method, which considers the most likely token each time, beam search algorithm maintains several candidates until completion.

The function used for the score will be expressed in the following form:

$$Score = \frac{1}{T^{\alpha}} \sum_{t=1} \log P(y_t) \tag{3}$$

Where represents the length normalization parameter

D. D. Overall Architecture

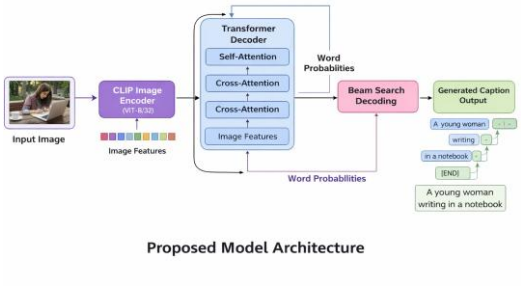

Fig. 1. Proposed Model Architecture

Pipeline of the whole system as such:

- CLIP encoder is employed to process the input image
- The feature vector is obtained by
- Transformer decoder produces sequence of captions
- The beam search algorithm chooses the optimal caption

The architectural design ensures efficiency in image comprehension and text generation, hence leading to contextual and sensible captions

VI. Results and Discussion

Evaluation of the performance of the image captioning method is done using the generated captions, where they have to be of high quality and relevance to the corresponding images. This method is trained on the TextCaps data set.

A. *A. Caption Generation Results*

The model is successful in producing captions that have meaning and relevance to the images fed into it. The employment of the CLIP encoder makes it easy for the model to capture high-level features, whereas the Transformer decoder facilitates fluency in the production of sentences.

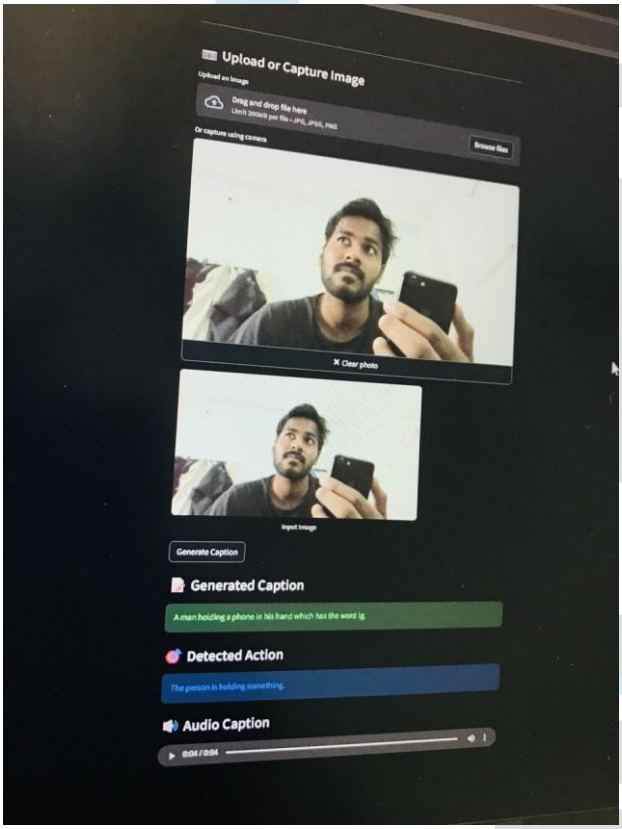

Fig. 2. Generated Caption Output

The output from the system can be seen in Fig. 2. The network is capable of providing an adequate description of the image. While certain captions may be a little vague, the sentence structure is still grammatically sound.

B. *B. Performance Analysis*

This model shows consistent behavior on various image samples. Adding the beam search technique results in better sentence coherence compared to using the greedy approach. Adding more features like action detection and text-to-speech makes the system easier to use.

C. *C. Limitations*

Though it performs well, it has certain weaknesses, which include:

- It might create general descriptions for intricate settings
- The accuracy of the output results relies on the training data
- Insufficient precise object recognition

D. *D. Discussion*

The outcome of the experiment shows that the integration of CLIP and Transformer decoder architecture is a successful method for creating captions to images. This technology ensures the proper combination of efficiency and accuracy, with real-time interaction enabled by Streamlit.

VII. CONCLUSION AND FUTURE WORK

VIII. Conclusion

The image captioning system in this project uses the CLIP model as the encoder and the Transformer as the decoder. In this case, the encoder provides rich semantic information about the images fed into the system, while the decoder produces semantically coherent and grammatically correct sentences. Beam search decoding contributes to the quality of the output sentences.

The model was trained on the TextCaps dataset and showed its capabilities in producing contextual captions for different images. Furthermore, the inclusion of live image capture, gesture recognition, and text-to-speech functions increases the interactivity of the proposed model.

In summary, the suggested methodology offers a compromise between precision and speed, thus proving that the fusion of vision-language models and Transformer structures is indeed a promising technique for generating image descriptions.

IX. Future Work

Even though the model works effectively, there are certain things that can be improved on:

- Fine-tuning the CLIP architecture by using dataset specific to domains for generating more accurate captions
- Using larger and more diverse datasets for training will help reduce biases
- Using state-of-the-art models like BLIP and GPT in decoding for more effective captioning
- Enhancing action detection through deep learning techniques for activity recognition
- Improving real-time speed to generate captions

References

[1] A. Radford et al., “Learning Transferable Visual Models From Natural Language Supervision,” Proc. ICML, 2021.
[2] Vaswani et al., “Attention is all you need,” in *Advances in Neural Information Processing Systems*, 2017.
[3] Vinyals O. et al., “Show and Tell: A Neural Image Caption Generator,” *Proc. CVPR*, 2015.
[4] K. Xu et al., ”Show, Attend and Tell: Neural Image Caption Generation with Visual Attention,” in *Proc. ICML*, 2015.
[5] Y. Wang et al., “TextCaps: A Dataset for Image Captioning with Reading Comprehension,” Proc. ECCV, 2020.