IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13469

Image Captioning using CNN and Transformers

K Lakshmipathi Raju[1], Venkat Rayidu[2], P Surendra[3], V Sai Satish[4], M. Sai Harsha[5]

Professor, Department OF Information Technology, SRKR Engineering College (Affiliated by JNTUK), Bhimavaram[1]

Student, Department OF Information Technology, SRKR Engineering College (Affiliated by JNTUK), Bhimavaram[2-5]

Abstract: Image captioning involves automatically describing images using words, attracting attention from researchers in natural language processing (NLP) and computer vision. Recent advancements primarily adopt an encoder-decoder framework, utilizing convolutional neural networks (CNNs) to extract image features and decoders to generate descriptions. Integration of attention mechanisms into this framework has notably improved performance. Leveraging the Transformer model, known for its effectiveness and efficiency in NLP tasks due to its attention mechanisms, we propose a novel approach combining CNNs and Transformers for image captioning. Our model utilizes a Transformer-Encoder to extract refined image feature representations, enabling the Transformer-Decoder to focus on pertinent image details when generating captions. Additionally, adaptive attention in the Transformer-Decoder determines the optimal utilization of image information during caption generation. Through extensive training on the Flickr8K_dataset, our model achieves an impressive 86.21% accuracy, demonstrating its efficancy and value in image captioning tasks.

Keywords: Image Caption, CNN, Deep Learning, Transformer, Attention mechanism, Flickr8k dataset.

I. INTRODUCTION

Machine translation, a pivotal domain within machine learning, is experiencing rapid evolution, catalyzing advancements across various technical sectors. The application of Artificial Intelligence (AI) and Neural Networks to intricate natural language processing tasks, like speech recognition and machine translation, is driving remarkable progress. Notably, within the realm of "Describing Images," significant strides have been made.

This task involves summarizing visual content, necessitating both comprehension of visual information and its translation into coherent sentences through natural language processing algorithms. Addressing this multimodal challenge requires hybrid models capable of seamlessly integrating visual and linguistic cues. Historically, prototype-based and retrieval-based approaches have been employed, yet they often falter in generalization and fail to produce novel descriptions for new images.

Recent innovations leveraging deep neural networks, particularly convolutional neural networks (CNNs), recurrent neural networks (RNNs) and long short term memory(LSTM), have demonstrated superior performance in image captioning tasks. These models implicitly learn common embeddings by encoding and decoding diverse modalities, yielding improved results across various caption generation datasets.

In recent years, numerous models have emerged in the domain of image captioning, sharing a common thread of integrating CNN and LSTM architectures. Popular datasets utilized for benchmark these approaches include Flickr8k and Flickr30k.

The Flickr8k dataset comprises 8091 images, each paired with five captions, while Flickr30k consists of 31783 images, each associated with five full-sentence level captions. These datasets serve as crucial testbeds for evaluating the efficacy of novel methods in image captioning, facilitating advancements in multimodal understanding and natural language processing tasks. In our model, we leverage the Transformer architecture and capitalize on the benefits of attention mechanisms by incorporating adaptive attention. Our contributions are outlined as follows:

- We harness the capabilities of the Transformer-Encoder to transform the image features extracted by CNN into a new representation. This aids the Transformer-Decoder in focusing on the most pertinent aspects of the image while generating subsequent words in the caption.

- By integrating the Transformer, renowned for its parallel training capacity and exceptional performance, we further enhance our model's capabilities. Additionally, we introduce adaptive attention within the Transformer-Decoder, enabling precise control over when and where the decoder utilizes the image features. This ensures that the model effectively incorporates relevant visual information at key points in the caption generation process

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13469

II. LITERATURE SURVEY

Sunil Varma,Nitika Kapoor et al. [1],discusses about The model, trained on image-text sets, demonstrates boundary sharing and unsupervised pre-training, accelerating downstream tasks like age-based or image captioning. It refines image-captioning models through two-stage training and utilizes past image-label sets for vision-language tasks, showcasing good and bad outputs, real-life photo caption predictions.

Sudhakar J,Viswesh Iyer V et al. [2], This paper addresses image captioning using ResNet50 and LSTM on the Flickr8k dataset, demonstrating ResNet50's superiority over VGG16, achieving 73% accuracy compared to 29%. It further explores converting captions to speech with gTTS and proposes future work to enhance model accuracy through larger datasets of images.

Kanimozhiselvi,Karthika V et al. [3], The proposed model utilizes CNN for feature extraction and LSTM for caption generation, achieving 72% accuracy with the Xception and LSTM combination over 50 epochs. LSTMs, crucial for sequence prediction tasks like speech recognition and machine translation, integrate bidirectional and sequence-to-sequence concepts, enhancing the ability to generate longer sentences.

Vaishnavi,Neha Tuniya et al. [4], The paper introduces an image caption generator employing an encoder and attention-based decoder, enhancing caption quality by focusing on relevant image information. However, limitations in training data and vocabulary may result in tags replacing unrecognized objects, potentially leading to nonsensical captions, particularly with multiple unknown objects in the input.

Soheyla Amirian et al. [5], This article discusses the integration of image captioning methodologies into video captioning, emphasizing algorithmic overlap rather than comprehensive review. Despite diverse approaches, it highlights the potential of deep learning-based captioning systems for enhancing accessibility and advancing search engine capabilities and recommendation systems. Yan Chu,Xiao Yue et al. [6], The paper introduces a joint ResNet50 and LSTM model with soft attention for automatic image captioning, enabling focus on specific image areas to enhance performance. Through stochastic gradient descent, the model is fully trainable and shows promising results in generating accurate image captions automatically. Ansar Hani,Najiba Tagougui et al. [7], The attention-based image captioning model encodes images with a convolutional neural network, then utilizes an attention layer and recurrent neural network to generate descriptions, yielding competitive results compared toexisting literature. The model's promising performance suggests its efficancy in generating image captions.

Pranay Mathur,Aman Gill et al. [8], The team develops a novel mobile application showcasing their model's capabilities and compares its performance with recent image captioning works. Leveraging Inception architecture and streamlining flow design, they optimize the model for real-time performance on mobile devices, achieving high-quality captioning results.

1.1 Existing System

Current automated image captioning systems predominantly rely on a combination of convolutional neural networks (CNNs) for image feature extraction and recurrent neural networks (RNNs) for sequential caption generation. This architectural choice has shown promise in various applications, leveraging the strengths of both CNNs and RNNs in processing visual information and generating coherent textual descriptions. The CNN component excels at extracting hierarchical features from images, while the RNN component handles the sequential nature of language generation, connecting visual features to linguistic structures.

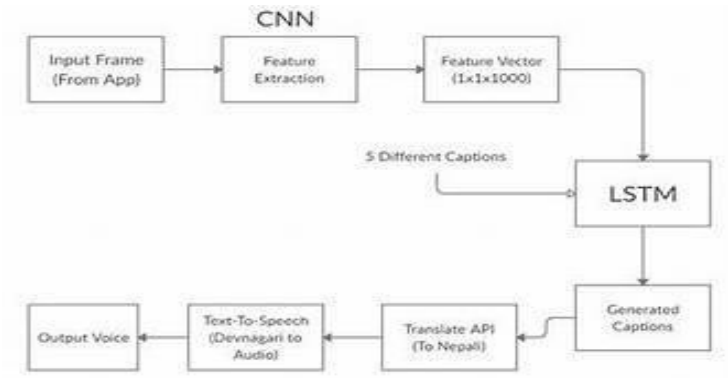


Fig. 1. Existing system Architecture

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13469

PROPOSED WORK

1.2 Proposed System

The proposed system of an image captioning model using a combination of convolutional neural networks (CNNs) and transformers. Overall, this system utilizes a combination of CNNs and transformers to generate descriptive captions for images. The CNN extracts visual features from images, which are then processed by the transformer encoder. The transformer decoder generates captions based on the encoded features, incorporating both visual and positional information.

It comprises three main components: a convolutional neural network (CNN) for feature extraction, a transformer encoder, and a transformer decoder. The CNN, based on the EfficientNetB0 model pre-trained on ImageNet, extracts image features. The transformer encoder block processes these features by applying multi-head self-attention mechanisms, layer normalization, and feed-forward networks. Similarly, the transformer decoder block employs self-attention and encoder-decoder attention mechanisms along with positional embeddings to generate captions sequentially.

The entire model is orchestrated within the imagecaptioningmodel class, which handles training and evaluation procedures. It implements custom training loops for optimizing captioning performance across multiple captions per image, incorporating loss calculation and accuracy tracking. This system offers a robust framework for generating descriptive captions for images using transformer-based architectures, allowing for flexibility in experimentation and fine-tuning

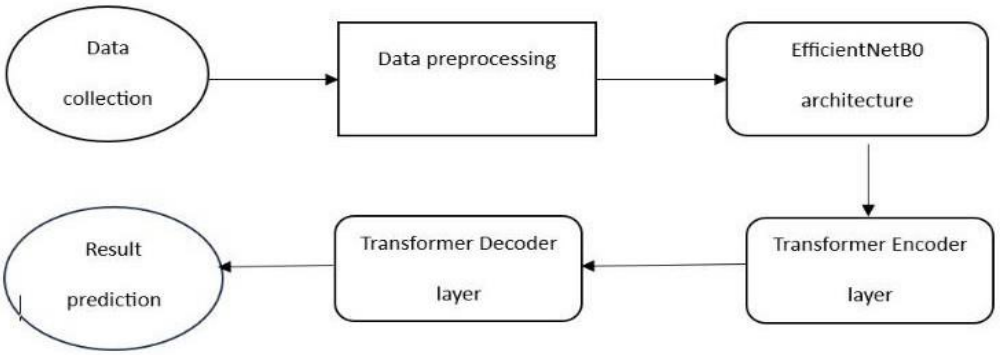


Fig. 2.Proposed Architecture

1.3 Problem Definition:

Generating the captions for the images which includes the time consuming for manual identification of objects in the images.

1.4 Data Collection:

For the image caption generator, we will utilize the Flickr_8K dataset. There is likewise another enormous Flickr 8K dataset. There are likewise other enormous datasets like Flickr_30K and MSCOCO dataset however it can require weeks just to prepare the organization so we will be utilizing a little Flickr8k dataset.

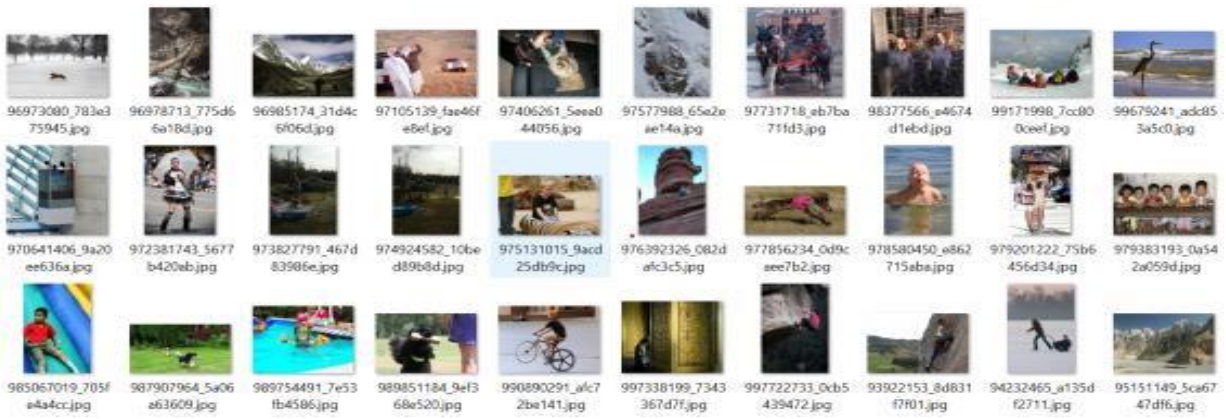


Fig. 3. Flickr_8k image dataset

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13469

1.5 Preprocessing:

Preprocessing pipeline involves loading and mapping caption data to corresponding images, followed by splitting the dataset into training and validation sets. Text data is preprocessed using custom standardization to lowercase and remove specified characters, then vectorized using Text Vectorization to convert text into integer sequences. For image data, data augmentation techniques including horizontal flipping, rotation, and contrast adjustment are applied using a Sequential model. Overall, this preprocessing pipeline ensures data readiness for subsequent training of models in image captioning tasks.

1.6 Feature Extraction:

An EfficientNetB0 model pre-trained on ImageNet is utilized as a feature extractor for images. This model extracts high-level features from images, which can be further processed and utilized by subsequent layers or models for various tasks such as image classification or captioning. Additionally, the TransformerEncoderBlock and PositionalEmbedding layers are employed for text processing in the context of image captioning. These layers are responsible for embedding and encoding textual information, allowing the model to understand and generate captions based on the extracted image features. Therefore, the features to be generated from the image can be considered as high-level semantic representations captured by the EfficientNetB0 model, which are then integrated with textual information through the TransformerEncoderBlock and PositionalEmbedding layers to generate coherent captions. presentations captured by the EfficientNetB0 model, which are then integrated with textual information through the TransformerEncoderBlock and PositionalEmbedding layers to generate coherent captions.

1.6.1 EfficientNetB0 Model

- **Role:** In the image processing pipeline, the CNN plays a vital role by extracting intricate features from input images. In particular, the utilization of the EfficientNetB0 model ensures efficient and effective feature extraction.

- **Importance:** CNNs excel in deciphering complex visual patterns within images, making them indispensable for tasks like image captioning. Leveraging a pre-trained model such as EfficientNetB0 capitalizes on the wealth of knowledge learned from a vast dataset like ImageNet. This strategy enables the extraction of rich and meaningful image features, enhancing the model's ability to comprehend image content accurately.

1.7 Methodology

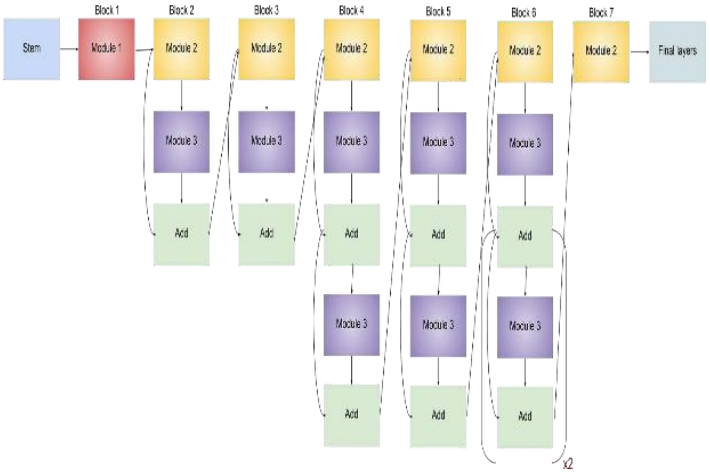


Fig. 4. Architecture of EfficientNetB0 Model

1.7.1 Transformer Encoder-Decoder

- **Role:** In the context of image captioning, the Transformer architecture assumes a crucial role as the backbone for sequence-to-sequence tasks. Here, the encoder module processes extracted image features, while the decoder module generates captions based on these features.

- **Importance:**

Encoder: Tasked with encoding image features, the encoder employs self-attention to discern and emphasize various aspects of the image, effectively capturing spatial relationships. This mechanism enables the model to comprehend and represent the visual content comprehensively.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13469

Decoder: Responsible for generating captions word by word, the decoder leverages the context representation from the encoder along with previously generated words. Through self-attention, the decoder module focuses on pertinent words at each step, ensuring the coherence and contextual relevance of the generated captions.

Transformer-Encoder

In our model, the Transformer-Encoder constitutes the second part of our encoder, featuring multiple layers. Each layer contains two sublayers:

Multi-Head Self-Attention: This sublayer allows the model to concentrate on different parts of the input vectors, capturing dependencies and relationships effectively.

Feed-Forward Networks: These networks introduce non- linearities, enabling the model to capture complex patterns and relationships within the data.

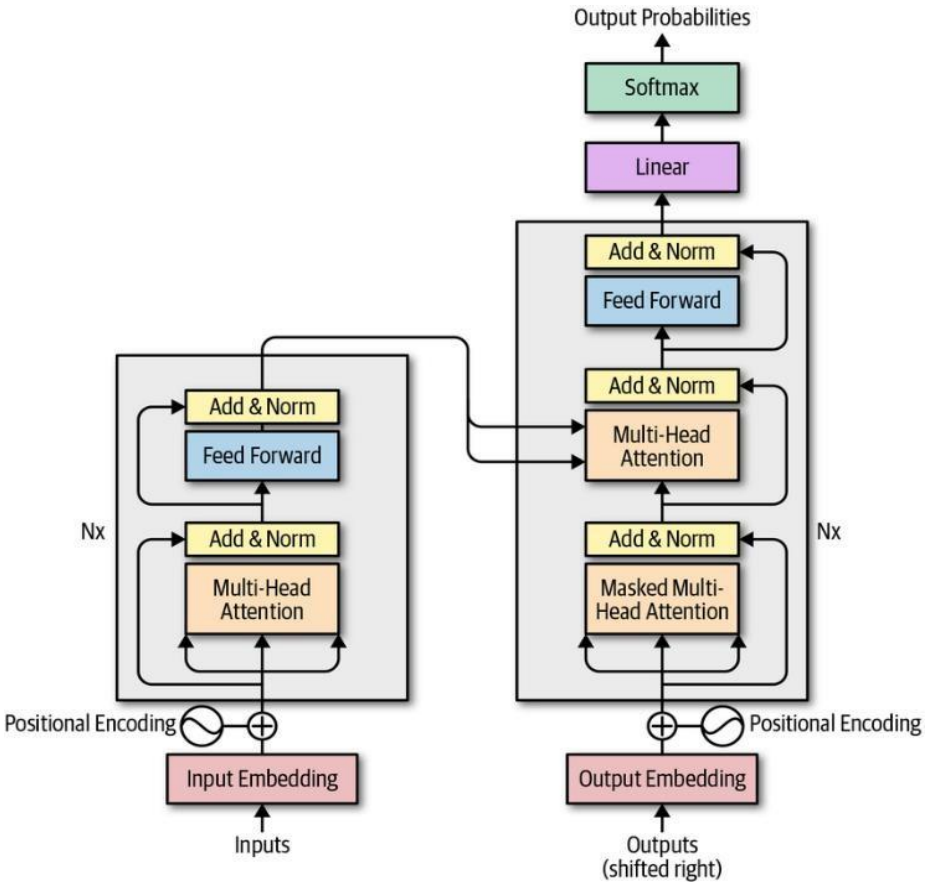


Fig. 5.Transformer Architecture

Before passing the output tensor (7x7x2048) from the CNN to the Transformer-Encoder, we preprocess it. Since the Transformer-Encoder lacks inherent knowledge of input order, we perform positional encoding to incorporate positional information. We employ sine and cosine functions to encode the positional information for each pixel. Specifically, for each of the 7x7 pixels, we generate two 1024-dimensional vectors representing the row and column positions. These vectors are then concatenated to form a final 2048-dimensional positional feature vector.

$$PE_{(pos,i)}=\sin\left(\frac{pos}{10000^{\frac{i}{1024}}}\right),\ PE_{(pos,i\ 1024)}=\sin\left(\frac{pos}{10000^{\frac{i}{1024}}}\right)\ 1$$

Subsequently, the image features, along with their positional encodings, are fed through the Transformer-Encoder layers. Here, the model undergoes self-attention and feed-forward operations, resulting in the generation of a new representation of the image features.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024
DOI: 10.17148/IJARCCE.2024.13469

This representation aids the Transformer-Decoder in focusing on the most relevant aspects of the image during the caption generation process.

1.7.1.1 Transformer-Decoder

Our decoder, known as the Transformer-Decoder, comprises multiple layers, each with distinct sublayers:

- Masked Multi-Head Self-Attention: This sublayer focuses on capturing correlations within the sentence while generating captions. It employs masking techniques to prevent attending to future words during training, ensuring that each word only attends to previous words.

- Adaptive Multi-Head Attention: Here, the decoder dynamically focuses on relevant image regions while generating captions. The keys and values receive input from the encoder's output, while the queries come from the output of the Masked Multi- Head Attention sublayer. This mechanism allows the decoder to selectively utilize image features based on their relevance to the caption being generated.

- Feed-Forward Networks: These networks introduce non-linearities, aiding in capturing intricate patterns and relationships within the data, enhancing the decoder's ability to generate coherent and contextually relevant captions.

In our decoder, we also handle word embeddings. We either customize the dimensions of initial word embeddings or utilize pre-trained word embeddings like GloVe, which encode semantic relationships between words through unsupervised learning. Furthermore, we perform positional encoding on the captions to preserve the word order when entering the Transformer-Decoder.

To enable the decoder to focus on the most relevant image regions, we employ spatial information. By calculating the correlation between keys and queries, we obtain a weight vector that determines the importance of each image feature. This weighted vector is then used to adjust the values associated with each image region, allowing the decoder to attend to relevant visual information during caption generation.

The decoder incorporates two attention mechanisms:

- Self-Attention Mechanism: This mechanism attends to correlations within the sentence, facilitating the generation of coherent captions.

- Adaptive Attention Mechanism: This mechanism enables the decoder to determine when and where to utilize image features. An adaptive gate ensures that the decoder selectively incorporates image features based on their relevance

Multi-Head self Attention Mechanism

- **Role:** Multi-head self-attention serves as a fundamental element integrated into both encoder and decoder blocks within the Transformer architecture.

- **Importance:** Self-attention empowers the model to assign different levels of importance to different elements within the input sequence, facilitating the capturing of dependencies among diverse elements. With multi-head attention, the model gains the capability to attend to multiple positions within the input sequence concurrently. This functionality enriches the model's comprehension of intricate relationships, thereby playing a pivotal role in tasks like image captioning. By enabling the model to concentrate on pertinent visual and contextual details, multi-head self-attention significantly enhances its ability to generate accurate and meaningful captions for images.

1.7.2 Layer Normalization and Feed-Forward Networks

- **Role:** Layer normalization and FNN play integral roles within both encoder and decoder blocks, serving to augment the model's learning capabilities.

- **Importance:**

Layer Normalization: This technique standardizes the input to each layer, contributing to the stabilization of training processes by mitigating internal covariate shift. By maintaining activations within a reasonable range throughout training, layer normalization facilitates quicker convergence, thus expediting the learning process.

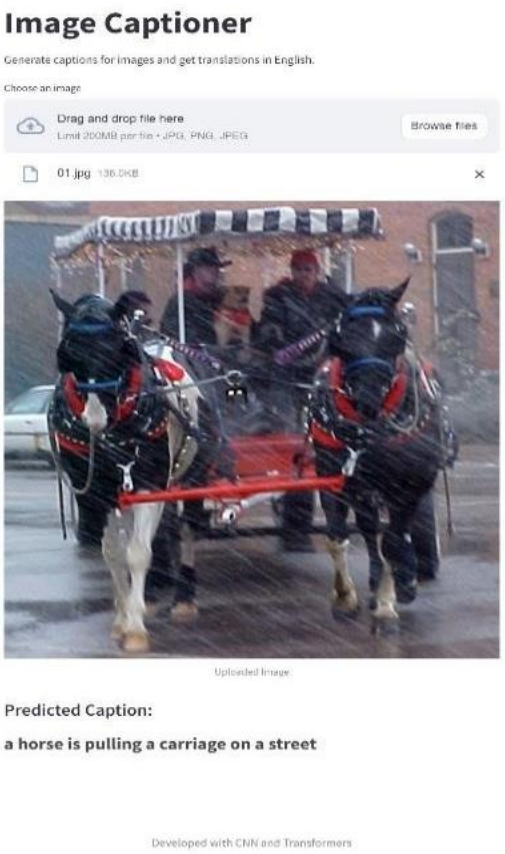
IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13469

Feed-Forward Networks: Employed to introduce non-linearities into the model, feed- forward networks are instrumental in enabling the capture of intricate patterns and relationships within the data. Through their transformative nature, these networks enhance the learned representations, rendering them more expressive and facilitating the extraction of more insightful features from the input data.



4.2 Predicted Captions

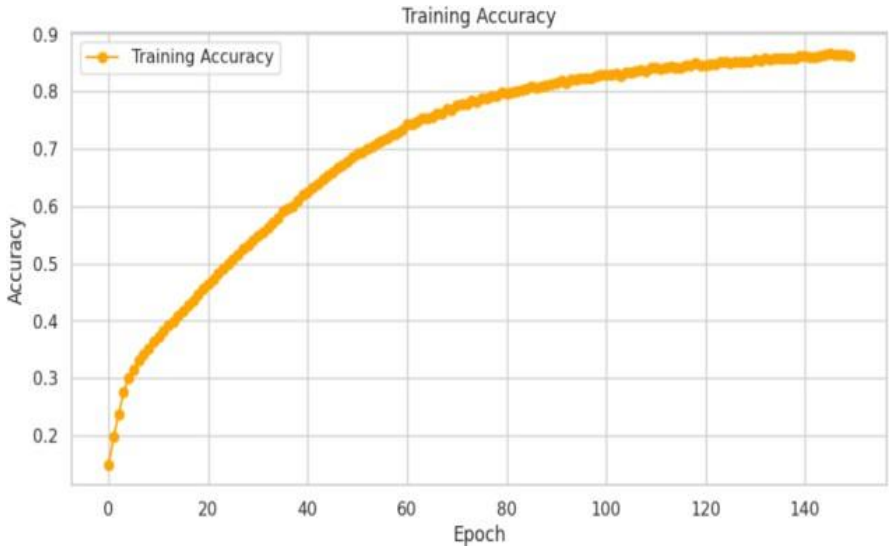


Fig. 6. Training accuracy

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13469

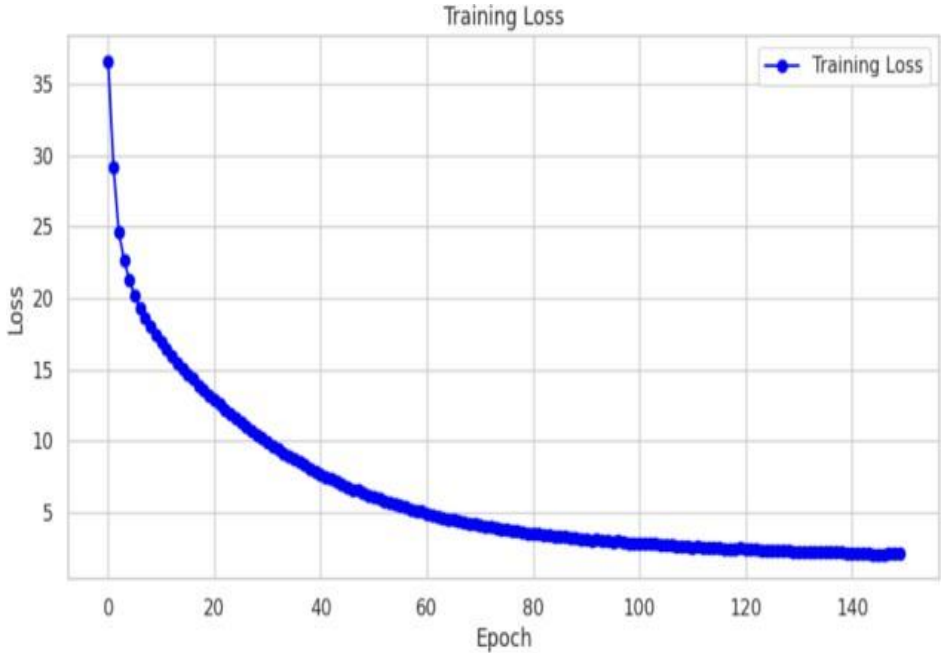


Fig. 7. Training Loss

III. RESULTS

1.8 Accuracy and Loss

Based on our model by generating the captions of the images,the accuracy of the proposed system achieves 86.21% which enhances the efficiency of the model in generating captions according to the previous existing models.

Fig. 8.Test Image-1

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024
DOI: 10.17148/IJARCCE.2024.13469

IV. CONCLUSION AND FUTURE WORK

In our research paper, we introduce a novel encoder-decoder framework that combines Convolutional Neural Networks (CNN) and Transformers to generate precise and contextually relevant captions for images within our dataset. Through rigorous training and validation processes, our model demonstrates remarkable accuracy in caption generation for images from the test set. Our proposed approach utilizes an EfficientNetB0 CNN for feature extraction, with a Transformer-Encoder serving as the encoder and a Transformer-Decoder as the sequence generator which the accuracy of 86.21%.Notably, our proposed ap surpasses alternative methods such as CNN-RNN, CNN- GUR, and CNN-LSTM, as well as those incorporating attention mechanisms.

The integration of our approach into various image captioning applications is extensive, ranging from automating visual interpretation and image indexing to generating traffic analysis reports using street-mounted cameras to guide drivers efficiently. Furthermore, our model holds potential for application recommendation systems, showcasing its versatilityecy across diverse domains.Future works will be center on preparing for a bigger number of pictures and datasets to make strides the model's overall accuracy and gives way better creating demonstrate for caption era.

REFERENCES

[1]. X. Li et al., “Oscar: Object-Semantics Aligned Pre- training for vision language tasks,” arXiv [cs.CV], 2020.
[2]. Kuznetsova, Alina; “Unified image classification, object detection, and visual relationship detection at scale”, 2018.
[3]. Caesar, Holger ; “nuScenes: A multimodal dataset for autonomous driving”, 2019.
[4]. Cai, Han ;Zhu, Ligeng ;Han, Song; ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware,2018.
[5]. Smith, Leslie N.; A disciplined approach to neural network hyper-parameters: Part 1 -- learning rate, batch size, momentum, and weight decay,2018.
[6]. A.Karpathy,Li Fei-Fei; Deep visual-semantic alignments for generating image descriptions,2014.
[7]. Polina Kuznetsova, Vicente Ordonez,Collective Generation of Natural Image descriptions,2012.
[8]. A.Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal,H. Larochelle, A. Courville, and B. Schiele, ‘‘Movie description,’’ Int. J. Comput. Vis., vol. 123, no. 1, pp. 94– 120, 2017.
[9]. Marcus Rohrbach, Wei Qiu,Translating video content into natural language descriptions,2013.
[10]. Michaela Regneri, Marcus Rohrbach GroundingAction descriptions in videos,2013.
[11]. Borneel Bikash Phukan, Amiya Ranjan Panda; An Efficient Technique for Image Captioning using Deep Neural Network,2020
[12]. A. Karpathy, Li Fei-Fei; Deep visual-semantic alignments for generating image descriptions,2014.