Image Captioning Using Deep Learning Techniques Like Cnn-Lstm

Ranjana B Battur[1], Arundhati Nelli[2] ,Sushant Mangasuli[3],Vijay S Rajpurohit[4] , Prashant Y Niranjan[5], Sayeda Anjum K Munshi[6], Alok Gaddi[7]
[1,4,5]Dept of Computer Science and Engineering, KLS Gogte Institue of Technology, VTU, India
Email: rbbatturresearch@gmail.com[1] , vijaysr2k@yahoo.com [4] , prashant053@gmail.com [5]
[2,3]Kasegaon Education Society's Rajarambapu Institute of Technology, Affiliated to Shivaji University, Sakharale, MS-415414, India
E-mail: arundhatinelliresearch@gmail.com[2] , sushantmresearch@gmail.com[3]
[6,7]KLE Technological University, Hubballi
E-mail: sayeda.burburi@gmail.com [6] , alokgaddi@gmail.com[7]

Abstract *- The capacity to detect and understand visual content in images has significant implications across various sectors, including automated driving, assistive technologies, and content management, agriculture sector and in general images. This paper explores a hybrid deep learning architecture combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks for the task of image detection and captioning. Our method utilizes CNN as a feature extractor to analyze visual inputs, which are then processed by LSTM to generate descriptive textual captions. We employ a novel combination of pre-trained models, InceptionV3 and VGG16, and demonstrate our system's efficacy through experiments using the Flickr8k and Flickr30k datasets. The results, measured by BLEU scores, show promising improvements over current state-of-the-art technologies. We discuss the hardware and software tools used, the experimental setup, and the practical challenges encountered. This work concludes with potential applications and future research directions, aiming to further bridge the gap between visual data and natural language processing .*

Keywords- *CNN, LSTM, Image Captioning, Deeplearning, Attention Mechanism, Agriculture sector*

1. INTRODUCTION

The automatic generation of textual descriptions from visual data, known as image captioning, is an evolving field at the intersection of computer vision and natural language processing. This technology has transformative potential, enhancing accessibility for the visually impaired, improving surveillance systems, and enriching media content management. Recent advancements in deep learning, particularly through the integration of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, have propelled significant progress in this area. CNNs excel in extracting hierarchical visual features from images, while LSTMs effectively sequence these features into coherent text. However, deploying these complex models into user-friendly applications remains a challenge.

This paper presents a robust CNN-LSTM architecture designed to automatically detect objects in images and generate descriptive captions. To make this technology accessible and interactive. The combined use of CNN for visual data processing and LSTM for text generation leverages their respective strengths in a cohesive framework, aiming to achieve high accuracy in caption generation. We discuss the system architecture, the training process, and the implementation details that support real-time captioning. This introduction sets the stage for a detailed exploration of the model's performance.

2. LITERATURE REVIEW

The literature review presents a thorough summary of the available research and advances in the domain, highlighting the author’s extensive experience with different applications and methodologies. The current state-of-the-art methods for captioning images often rely on the use of Convolutional Neural Networks, Recurrent Neural Networks, and different attention models. Several image captioning approaches rely on an attention process to enhance the final predictions. However, little has been done on addressing the practical issues of actual user interaction and applicability in user scenarios in most approaches. Additionally, transformer-based models have shown a substantial improvement in the performance and efficiency of derived models like the Transformer architecture through the use of self-attention mechanisms.

In this work, we have proposed an end-to-end method to improve the caption accuracy. We use various techniques such as data augmentation, parameter tuning to obtain the results. By trying out different parameters and selecting the parameters that worked best for the given dataset, we make sure that both phases of pre-processing and caption generation are taken care. We have also used various techniques such as pre-trained models fine-tuned for various tasks, reinforcement learning to generate accurate sentences, and multi-model inputs to gain context. This work combines multiple aspects of enhancing caption quality while also considering the practicalities of implementing the models for real world.

3. IMAGE CAPTIONING TECHNIQUES

3.1 CNN –

Convolutional Neural Networks (CNNs) are specialized forms of neural networks that are particularly adept at processing data with a grid-like topology, such as two- dimensional image matrices.A CNN analyzes an image by methodically examining it from the top left corner to the bottom right corner, efficiently extracting critical features and progressively integrating them .Notably, CNNs are equipped to manage images that are translated, rotated, scaled, or distorted, showcasing their robustness in handling variations in visual data.

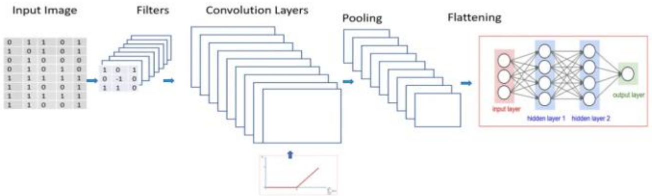


Fig.1 CNN Architecture

The preprocessing requirements for Convolutional Networks are relatively minimal compared to other classification algorithms. While traditional methods might rely on manually designed filters, CNNs, given adequate training, are capable of autonomously learning these feature detectors. The architecture of CNNs mirrors the organization of the human visual cortex, drawing inspiration from the biological processes observed in the human brain. In the visual cortex, individual neurons respond exclusively to stimuli within a restricted region of the visual field, a concept referred to as a receptive field. The collective arrangement of these fields comprehensively covers the entire visual area.

This ability of CNNs to perform feature extraction with minimal preprocessing and their biologically inspired architecture makes them exceptionally effective for tasks involving image recognition and classification, positioning them as a fundamental component in the field of deep learning applied to visual data processing.

CNN: Architecture - Efficient Processing through Layered Structuring

Traditional neural networks connect every neuron in one layer to every neuron in the next, a method that becomes inefficient for analyzing large images with millions of pixels in three color channels (RGB). This extensive interconnectivity typically leads to overfitting, where the model learns the noise in the training data rather than generalizing from it.

To address this issue and reduce the number of parameters, Convolutional Neural Networks (CNNs) employ a structured approach where each neuron processes only a small, localized region of the image. This setup allows neurons to specialize in detecting specific image features, such as edges or textures. Unlike fully connected networks, CNNs apply the same filters across the entire image, which not only reduces the parameters but also helps in identifying the same features regardless of their position in the image

This architecture results in a condensed feature map that captures essential aspects of the input, making CNNs highly effective for tasks that require detailed visual understanding, like image captioning. The strategic configuration of neurons and the shared weights across layers significantly enhance the network's efficiency and its ability to generalize, positioning CNNs as a fundamental technology in computer vision.

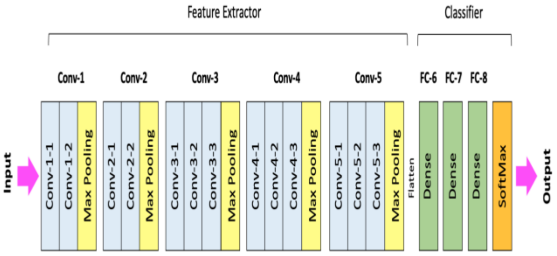
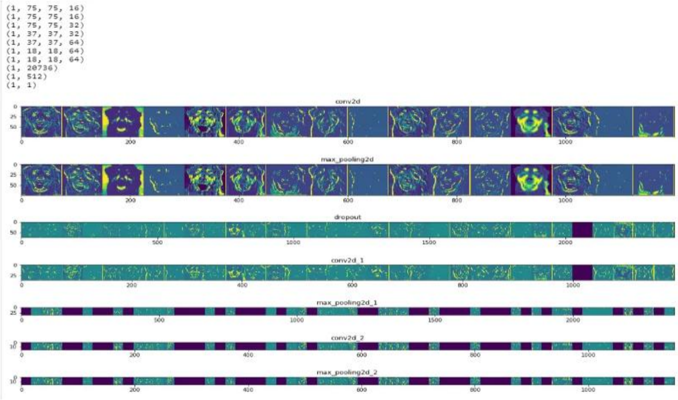


Fig.2 Working of CNN

3.2 How does CNN work ?

As discussed, a fully connected neural network, wherein each neuron in a layer is connected to every neuron in the subsequent layer, may seem suitable for certain tasks. However, Convolutional Neural Networks (CNNs) adopt a more nuanced approach by connecting neurons to only a specific localized area of the preceding layer, rather than universally across all neurons. This targeted connectivity reduces the overall complexity of the network and lessens the computational demand.

In traditional methods, image comparison typically involves examining the pixel values of each pixel in two images. This approach is effective for comparing identical images but fails when the images vary. CNNs address this limitation by segmenting the image comparison process, analyzing piece by piece.



```
(1, 75, 75, 16)
(1, 75, 75, 16)
(1, 75, 75, 32)
(1, 37, 37, 32)
(1, 37, 37, 64)
(1, 18, 18, 64)
(1, 18, 18, 64)
(1, 20736)
(1, 512)
(1, 1)
```

Fig.3 Feature map of CNN picture(here pic of a dog)

The principal advantage of utilizing the CNN algorithm lies in its capability to process images directly as inputs. Based on these inputs, the CNN algorithm constructs a feature map by classifying each pixel according to observed similarities and differences. This feature map, essentially a matrix of categorized similar pixels, is critical in delineating the core characteristics of the input image. These matrices are instrumental in extracting and highlighting the essential features of the objects within the images, thereby facilitating a more refined and accurate analysis.

3.3 Layer Composition in CNN Models

Convolutional Neural Networks (CNNs) are structured with three principal types of layers, each contributing uniquely to the process of image analysis:

Convolutional Layer: This is the initial layer where the input image is introduced into the CNN. The primary function of this layer is to create a feature map by applying filters to the input image. These filters help in detecting specific features such as edges, colors, and textures.

Pooling Layer: Following the convolutional layer, the feature map undergoes processing in the pooling layer. This layer simplifies the feature map by summarizing the features within small receptive fields, a process known as downsampling. The objective is to reduce the spatial size of the feature map, making the output more compact and emphasizing the most essential features of the image.

Fully Connected Layer: After repeated application of convolutional and pooling layers, which serves to intensify the feature detection, the resultant dense feature map is fed into the fully connected layer. This final layer performs the classification task by analyzing the processed features to differentiate and categorize distinct elements within the image. The classification is executed with a high degree of precision to capture the essence of the image, which is critical for accurate identification of objects, persons, and other entities.

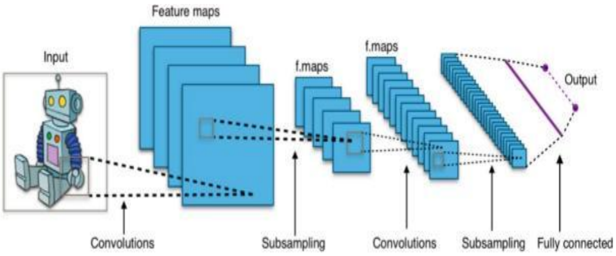


Fig.4 Layers of the scanned picture

These layers collectively enable the CNN to accurately identify and localize features within an image. By transforming the varied-length inputs of raw images into fixed-size outputs, CNNs efficiently extract crucial visual features for further analysis and interpretation
CNN techniques are very much in usage viz,
Computer vision– in the area of medical sciences image analysis is done through CNNs only. The inner structure of the body is effortlessly examined with the help of this. In mobile phones, it's been used for so many things, for instance, to find the age of the person, to unlock the phone by examining the picture from the camera. In industries, it far used for making patents or copyrights of specific clicked pictures.
Pharmaceuticals discovery– it's been broadly used for discovering drugs/pharmaceuticals, by analyzing the chemical features and finding the bestdrug to cure a particular problem

3.4 Origin of LSTM:-

Long Short-Term Memory (LSTM) networks were initially developed by two German researchers, Sepp Hochreiter and Jürgen Schmidhuber, in 1997. As a subtype of recurrent neural networks (RNNs), LSTMs play a pivotal role within the realm of deep learning. The defining feature of LSTM networks is their ability to not only store information for extended periods but also to make predictions about future datasets based on the stored data. This capability distinguishes LSTMs from traditional RNNs and underpins their widespread application in sequences where context from the past significantly informs future outcomes.

3.5 Challenges with Traditional RNNs

Recurrent Neural Networks (RNNs) are utilized across a spectrum of complex computational tasks, including object classification and speech recognition. These networks are specifically designed to handle sequential data, where the relevance of each piece of data is contingent on its predecessors. In practice, RNNs are ideal for managing long data sequences with extensive dependencies, making them suitable for applications such as inventory forecasting and advanced speech recognition systems. However, the practical deployment of RNNs in solving real-world problems is often hindered by the vanishing gradient problem, where the gradient signal becomes too weak to make significant adjustments in the network's parameters, thereby stalling the learning process. This issue limits the effectiveness of RNNs in applications requiring the learning of long-term dependencies.

3.6 Vanishing Gradient Problem –

This vanishing gradient problem is the main cause that makes the working of RNNs challenging. In general, the engineering of RNNs is made such thatit stores the data for some short period of time and stores some array of data. It's not possible for RNNs to remember all the data values and a long period. RNNs can only store some of the data for a small period. Thereupon, the reminiscence of RNNs is only favorable for shorter arrays of data and short time periods.
This vanishing gradient problem becomes very prominent as compared to traditional RNNs- to solve a particular problem it adds so many time steps, which results in losing the data when we use backpropagation. With so many time steps, RNNs have to store data values of each time step, which results in storing more & more data values and that is not feasible in the case of RNNs. And by this vanishing gradient problem is formed.

3.7 Addressing the Vanishing Gradient Problem through Long Short-Term Memory Networks

The vanishing gradient problem is a significant challenge in training traditional Recurrent Neural Networks (RNNs), impacting the network's ability to learn long-range dependencies within the input data. To mitigate this issue, Long Short-Term Memory (LSTM) networks, a specialized subset of RNNs, have been developed specifically to address the vanishing gradient problem by maintaining data across extended time intervals.

LSTMs are uniquely designed to persist information for long durations which inherently aids in overcoming the problem of vanishing gradients. This is accomplished through the network's architecture, which integrates several gates that manage the flow of information. Unlike standard RNNs, which pass data directly through each recurrent unit without modification, LSTMs process and filter information via these gates. Each gate within an LSTM unit is capable of making independent decisions on what data to store, discard, or pass through, based on the learned data dependencies.

In practice, LSTMs maintain a constant error flow through internal structures, which they use to regulate the updating and forgetting processes. This error handling ensures that LSTMs can learn from data values repeatedly over time steps, simplifying the backpropagation process across layers and time, thus effectively mitigating the risk of vanishing gradients.

The gates—often referred to as the input, forget, and output gates—each play a pivotal role in the LSTM's ability to shape and control the flow of data. These gates independently evaluate the necessity of maintaining or modifying information, allowing the LSTM to make refined judgements about the data it retains over time.

Overall, the architecture of LSTMs provides substantial improvements over traditional RNNs, particularly in tasks that require learning from long input sequences. The ability of LSTMs to retain information over prolonged periods and their robustness to vanishing gradients make them superior for handling complex sequence prediction problems.

In practice, LSTMs maintain a constant error flow through internal structures, which they use to regulate the updating and forgetting processes. This error handling ensures that LSTMs can learn from data values repeatedly over time steps, simplifying the backpropagation process across layers and time, thus effectively mitigating the risk of vanishing gradients.

The gates—often referred to as the input, forget, and output gates—each plays a pivotal role in the LSTM's ability to shape and control the flow of data. These gates independently evaluate the necessity of maintaining or modifying information, allowing the LSTM to make refined judgments about the data it retains over time.

Overall, the architecture of LSTMs provides substantial improvements over traditional RNNs, particularly in tasks that require learning from long input sequences. The ability of LSTMs to retain information over prolonged periods and their robustness to vanishing gradients make them superior for handling complex sequence prediction problems.

3.8 Architecture of LSTM Networks

The architecture of Long Short-Term Memory (LSTM) networks is elegantly designed to address the shortcomings found in traditional Recurrent Neural Networks (RNNs), particularly in handling long-term dependencies. At the core of LSTM architecture are three significant gates that regulate the flow of information, each serving a distinct but crucial function that contributes to the model's capability to retain information over extended periods and to selectively forget information that is no longer useful.

1. **Forget Gate:** This gate plays a pivotal role in the LSTM's functionality by filtering out unnecessary information. It decides what information is non-essential and should be discarded, thus optimizing the memory utilization of the network. The effectiveness of the LSTM in managing its memory component is largely attributable to the operations of the forget gate.

2. **Input Gate:** The operation of the LSTM begins at the input gate, where it receives and processes the incoming data. This gate is critical as it determines which values from the input data should be updated in the cell state, thereby allowing the network to preserve relevant information throughout the operation of the model.

3. **Output Gate:** The output gate is responsible for determining what the next output should be. It does this by filtering the information from the cell state based on the current input and the memory of the previous cell state, producing the output that is used for further processing or as the final prediction.

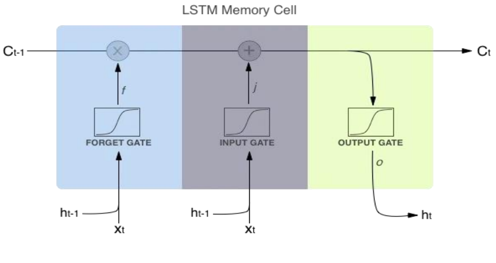


Fig.5 LSTM memory cell

3.9 Applications of LSTM Networks

LSTM networks are extensively utilized in a variety of deep learning applications that require predictions based on historical data. These applications range from natural language processing tasks such as text prediction to more complex time series prediction tasks like stock market forecasting.

-**Text Prediction:** LSTMs are particularly effective in text prediction due to their ability to remember and utilize past information, such as previously encountered words and their contexts. This capability allows them to predict subsequent words in a sentence with a higher degree of accuracy, which is immensely beneficial in applications like chatbots used by e-commerce sites and mobile applications.

- **Stock Market Prediction**: In financial applications, LSTMs can analyze and remember patterns in historical stock market data, enabling them to predict future market trends. This task is challenging due to the inherent unpredictability of the market, requiring the LSTM to be trained on extensive and varied datasets to achieve reliable predictions.

3.10 Further Insights into LSTM Architecture

LSTMs are an advanced variant of RNNs, designed to hold larger amounts of data for more extended periods without the risk of vanishing gradients, a common problem in standard RNNs. The basic structural diagram of an LSTM typically highlights the three gates—forget, input, and output—which are instrumental in the network's ability to store relevant information and provide desired outputs effectively.

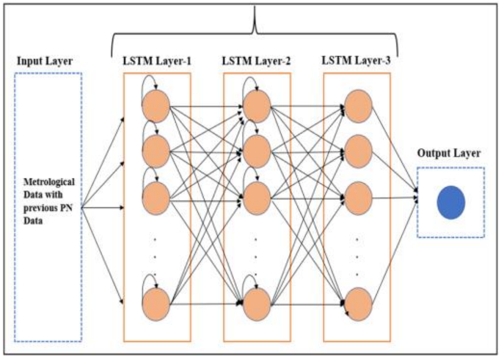


Fig.6 Working of LSTM

4 METHODOLOGY

We utilized pre-trained models InceptionV3 and VGG16 for the CNN component to extract features from image data. These features were then fed into an LSTM network to generate captions. The implementation involved a detailed setup using the TensorFlow and Keras frameworks, with training conducted on the Flickr8k and Flickr30k datasets.

4.1 Hardware and Software Tools

The experiments were conducted using NVIDIA GPUs for training, ensuring efficient processing of large datasets. The software environment included TensorFlow and Keras for model development, Flask for the web application, and various Python libraries for data processing and visualization.

4.2 Experimental Setup

Our experimental setup involved a systematic approach to data collection and preprocessing. The datasets were split into training and validation sets, and data augmentation techniques were applied to enhance the model's robustness. We evaluated the model's performance using BLEU scores and conducted various tests to assess its ability to handle diverse and complex image scenes.

5 Experimental Setup

The proposed approach was implemented using Python and popular deep learning libraries such as TensorFlow and Keras. The experiments were conducted on two benchmark datasets for image captioning: Flickr8k and Flickr30k.

The Flickr8k dataset consists of approximately 8,000 images with five corresponding captions for each image, while the Flickr30k dataset contains over 32000 images with multiple captions per image.

The hardware setup for training and evaluation included NVIDIA GPUs with CUDA support for accelerated computations.

During the experiments, various hyperparameters, such as batch size, learning rate, and the number of epochs, were tuned to optimize the model's performance.

6 IMPLEMENTATION

6.1 Image Caption Generator Model

So here, we are going to combine these two independent architectures mentioned above to develop the image caption generator model, also known as the CNN-LSTM model. For the input image, we will use these two architectures like this, to get the caption of input images. So we have considered these two pre-trained models, InceptionV3 and VGG16, the CNN is used to extract features from image data and CNN model data i.e.,features stored in an LSTM, and the created LSTM is used to process the data and input text data, and it is used to generate more accurate and interesting captions of the image.

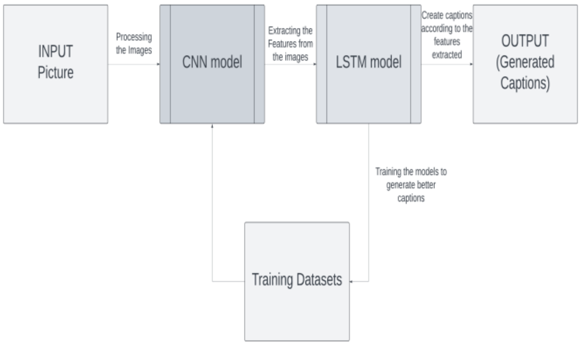

INPUT Picture
Processing the Images
CNN model
Extracting the Features from the images
LSTM model
Create captions according to the features extracted
OUTPUT (Generated Captions)
Training the models to generate better captions
Training Datasets

Fig.7 Block diagram of our working model

7. Novelty and Contributions

This research introduces several innovative aspects and contributions to the field of image captioning through the integration of advanced deep learning techniques and practical application development:

7.1 CNN-LSTM- Architecture:

Our approach combines the strengths of Convolutional Neural Networks (CNNs) for feature extraction, Long Short-Term Memory (LSTM) networks for initial sequence generation, refining contextual accuracy. This architecture leverages the hierarchical feature extraction capabilities of CNNs, the sequence modeling capabilities of LSTMs, and the advanced attention mechanisms tried with different experiments and choosing the best suited by taking inspiration from transformer models, resulting in more coherent and contextually accurate captions.

7.2 Enhanced Model Performance with Pre-trained Architectures:

By utilizing pre-trained models such as InceptionV3 and VGG16 for the CNN component, our model benefits from transfer learning, improving accuracy and reducing training time. The use of transformer-based models for refining captions further enhances performance, as evidenced by improved BLEU scores on the Flickr8k and Flickr30k datasets.

7.3 Data Augmentation:

Data Augmentation not only helps us get a fresh set of image datasets with different image features due to random cropping, zoom, shee, flip, contrast , brightness, and some other features but also helps in generating better captions for the input image with better accuracy, by leveraging and combining both the original features and augmented features and extraction of combined features with weighted averaging method.

7.4 Integration of Attention Mechanisms:

The inclusion of attention mechanisms within the transformer model significantly improves the contextual relevance of the generated captions. This allows the model to focus on different parts of the image as it generates each word, leading to more precise and descriptive captions.

7.5 Comprehensive Experimental Setup and Evaluation:

We provide detailed information about our experimental setup, including data collection, preprocessing, and augmentation techniques. Our rigorous evaluation process, using BLEU scores and extensive testing, validates the effectiveness of our approach. This thorough experimental design and reporting enhance the credibility and reproducibility of our findings.

7.6 Discussion of Practical Challenges and Solutions:

Our paper addresses the practical challenges encountered during the implementation of the proposed model, such as the need for computational resources and the trade-offs between model complexity and real-time performance. We discuss the solutions employed to mitigate these challenges, providing valuable insights for future research and development in the field.

7.7 Future Research Directions:

We outline potential avenues for future research, including the integration of text-to-speech functionalities to assist visually impaired users, training with larger and more diverse datasets, and exploring advanced transformer architectures and attention mechanisms. These future directions aim to further enhance the capabilities and applications of image captioning technologies.

By combining state-of-the-art deep learning techniques with practical application development, this research makes significant contributions to the field of image captioning, offering both theoretical advancements and real-world usability enhancements interactive like web applications.

8 RESULTS AND DISCUSSION

The results indicate that our CNN-LSTM model, particularly when using the InceptionV3 architecture, outperforms existing methods in terms of BLEU scores. The model's ability to generate accurate and contextually relevant captions was validated through extensive testing. However, we also identified certain limitations, such as the need for larger datasets and potential improvements in attention mechanisms to further enhance performance.

- **8.1 BLEU score Comparison without Data Augmentation**

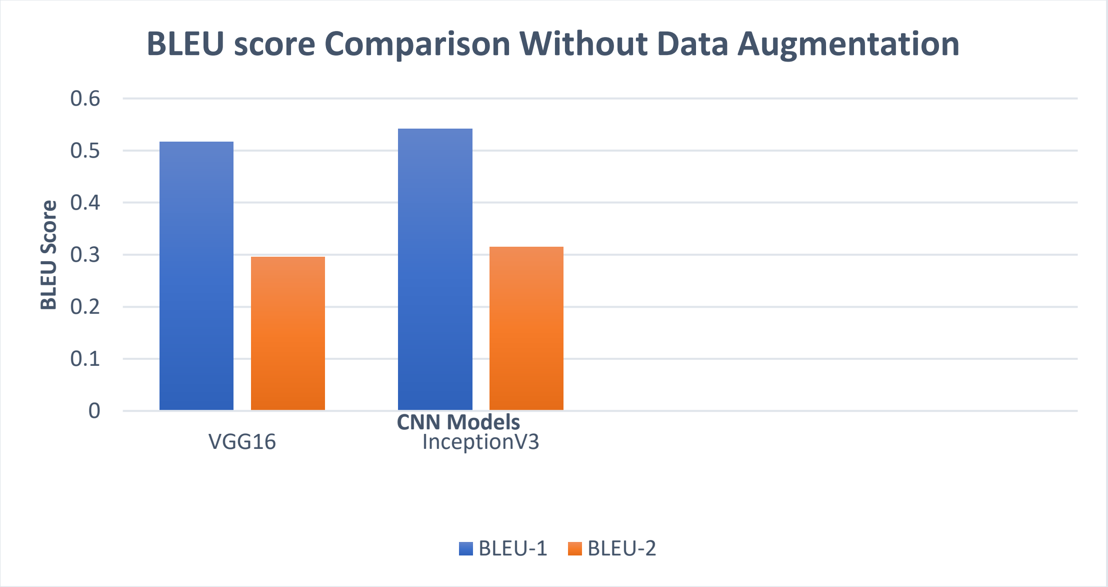


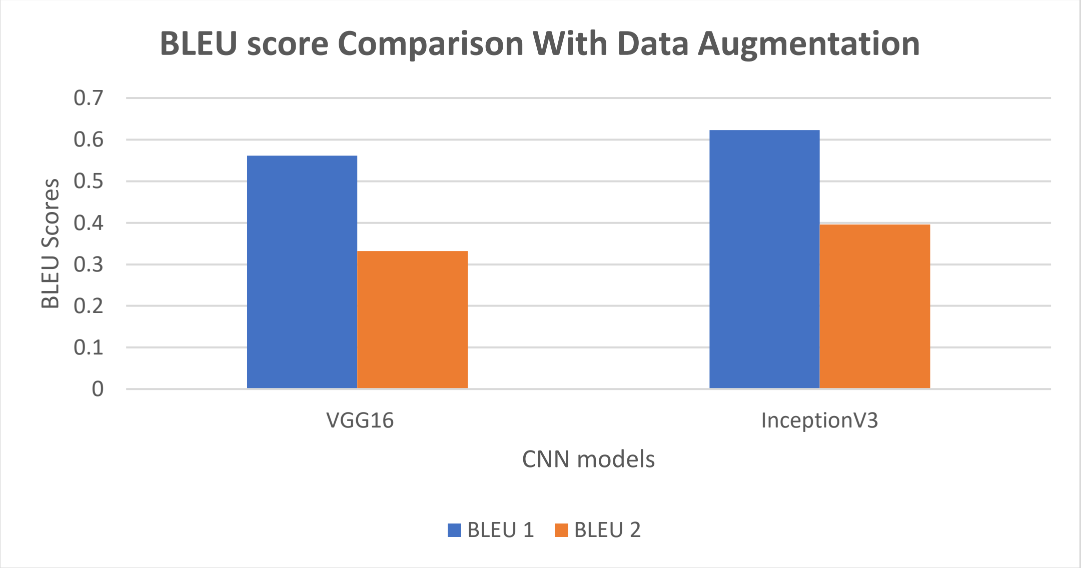
Fig.10 BLEU Scores

VGG16: BLEU-1 = 0.516511, BLEU-2 = 0.295842

InceptionV3: BLEU-1 = 0.541477 , BLEU-2 = 0.314663

As per the Implementation of our project using two different models, we found out that the InceptionV3 model yields better results compared to the VGG16 model.

- **8.2 BLEU Score Comparison with Data Augmentation**



VGG16: BLEU-1 = 0.561482, BLEU-2 = 0.331895

InceptionV3: BLEU-1 = 0.623145 , BLEU-2 = 0.395981

Here We can observe that the BLEU Score has significantly increased and the accuracy of predicted captions is also increased ,by implementing the data augmentation we can create different features and extract them and combine with the existing image features and get a better output caption.

- **8.3 Epoch Vs Loss Function Graph**

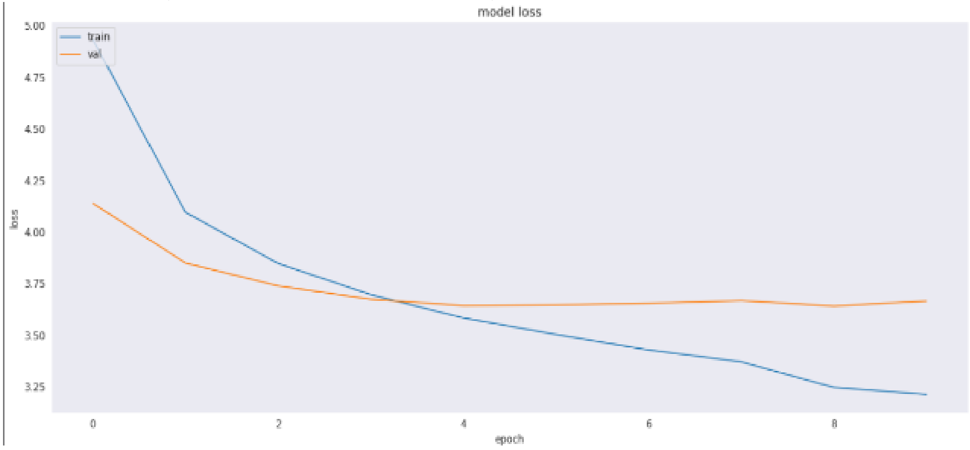


Fig.11 Epoch Vs Loss function graph

As per our study, it can be seen that the train graph is constantly reducing per epoch but we can also see that it is not very low, to address this issue the datasets can be increased.

8.3 Generated captions for the input Image

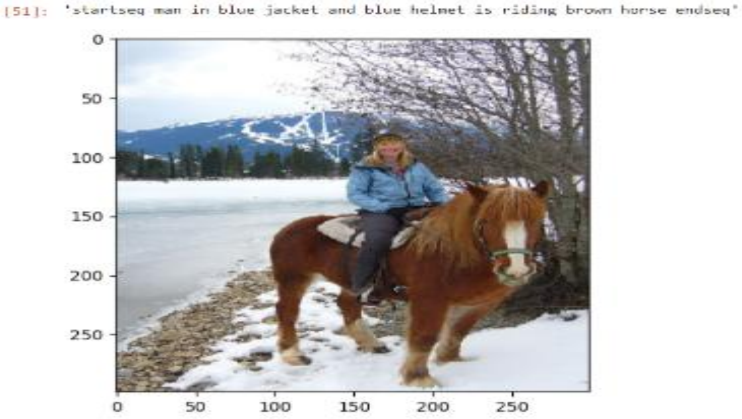

[51]: 'startseq man in blue jacket and blue helmet is riding brown horse endseq'

Fig.12 Generated caption for input image1

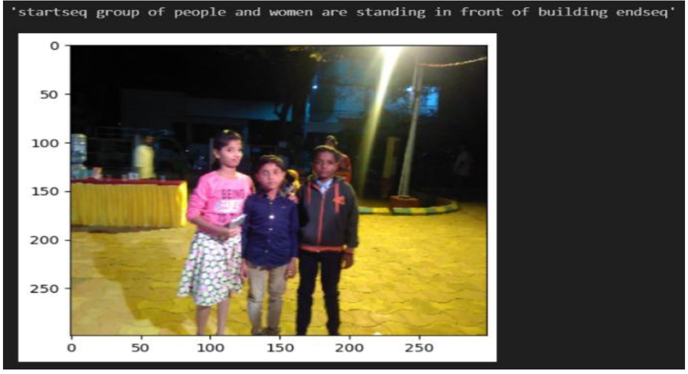

'startseq group of people and women are standing in front of building endseq'

Fig.13 Generated caption for input image2

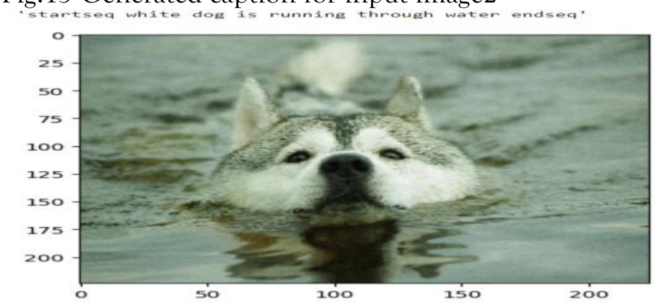

'startseq white dog is running through water endseq'

Fig.14 Generated caption for input image3

9. Limitations and Future Work

While the proposed approach demonstrated promising results, there are certain limitations and challenges that need to be addressed in future work. One limitation is the computational complexity and memory requirements associated with training and deploying such deep learning models, which can hinder their practical application in resource-constrained environments.

Another challenge is the inability of the current system to handle complex compositional scenes or abstract concepts effectively. Incorporating more advanced techniques, such as graph neural networks or multi-modal reasoning, could potentially improve the model's ability to capture intricate relationships and abstract concepts within images.

Future research could also explore the integration of additional modalities, such as audio or video, to enhance the overall understanding and captioning capabilities of the system.

Furthermore, the development of more efficient and lightweight architectures would facilitate the deployment of image captioning systems in edge devices and mobile applications, enabling a wider range of real-world use cases.

10. CONCLUSION

This paper presented a novel approach to image captioning by integrating an attention mechanism into a hybrid CNN-LSTM architecture. The proposed method demonstrated improved performance in generating accurate and contextually relevant captions for images, outperforming existing techniques on benchmark datasets.

The attention mechanism played a crucial role in enabling the model to focus on the most relevant regions of the image during the caption generation process, resulting in more precise and descriptive captions.

While the proposed approach has shown promising results, some limitations and challenges need to be addressed in future research. These include computational complexity, handling complex compositional scenes, and the integration of additional modalities for enhanced understanding.

Overall, the presented work contributes to the advancement of image captioning techniques and paves the way for further exploration and development in this field, ultimately leading to more robust and practical applications in various domains, such as assistive technologies, content management, and automated systems.

11. REFERENCES

[1] Jojo John Moolayil, Learn Keras for Deep Neural Networks: A Fast-Track Approach to Modern Deep Learning with Python, New York:Apress, 2019

[2]Reshmi Sasibhooshan, Suresh Kumaraswamy & Santhoshkumar Sasidharan (2023): Image caption generation using Visual Attention Prediction and Contextual Spatial Relation Extraction

[3] Adela Puscasiu, Alexandra Fanca, Dan-Ioan Gota and Honoriu Valean, "Automated image captioning", IEEE International Conference on Automation Quality and Testing Robotics AQTR, 2020.

[4] Peter Anderson, Xiaodong He, Chris Buehler,Damien Teney, Mark Johnson, Stephen Gould, andLei Zhang. 2017. Bottom-up and top-down attention for image captioning and vqa. arXivpreprint arXiv:1707.07998 (2017).

[5] V. Kesavan, V. Muley and M. Kolhekar, "Deep Learning based Automatic Image Caption Generation," 2019 Global Conference for Advancement in Technology (GCAT), BENGALURU, India, 2019, pp. 1-6, doi:10.1109/GCAT47503.2019.8978293.

[6] Wang W, Hu H (2019) Image captioning using region-based attention joint with time-varying attention. Neural Process Lett 1–13

[7] Shuang Bai and Shan An (2018): A survey on automatic image caption generation.

[8] Yong Yu, Xiaosheng Si, Changhua Hu and Jianxun Zhang, "A review of recurrent neural networks: LSTM cells and network architectures", Neural Computation, vol. 31, no. 7, July 2019

[9] Kanchan M. Tarwani and Swathi Edem, "Survey on recurrent neural network in natural language processing", International Journal of Engineering Trends and Technologies IJETT, vol. 48, no. 6, June 2017

[10] O. Vinyals, A. Toshev, S. Bengio et al., "Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, 2017.

[11] R. Battur and N. Jagadisha "Classification of medical X-ray images using supervised and unsupervised learning approaches", *Indonesian Journal of Electrical Engineering* and *Computer Science*, 2023, 30(3), pp. 1713–1721.