Image Captioning

Arrabelli Varsha Reddy
Guttha Bhanu Prakash Chowdary
Varikuti Sudhir Reddy

Under the guidence of
Pournami P N
Associate Professor- CSED
National Institute of Technology Calicut

Abstract

- Image captioning aims to detect information by describing the image content through image and text processing techniques. Recent progress in artificial intelligence (AI) has greatly improved the performance of models. However, the results are still not sufficiently satisfying. Machines cannot imitate human brains and the way they communicate, so it remains an ongoing task. This project focuses on the development of an image captioning system, a novel approach merging computer vision and natural language processing. Leveraging deep learning techniques, the system aims to automatically generate descriptive and contextually relevant captions for input images. By employing convolutional neural networks (CNNs) for image feature extraction and recurrent neural networks (RNNs) or long short-term memory (LSTM) for language modeling and adapting attention mechanism and GloVe word embeddings to enhance the model, the model learns intricate relationships between visual content and textual descriptions.

Index

I INTRODUCTION

The captioning of images is most widely used technology all over the world and still it is in improving position using advance technology. In today's rapidly evolving technological landscape, the art of image captioning stands as a testament to the remarkable strides made in the field of artificial intelligence and computer vision. Image captioning, a pivotal technology, transcends the boundaries of language, enabling the conversion of visual content into plain, comprehensible English. Its applications span a myriad of domains, including video generation, accessibility for the visually impaired, immersive gaming experiences, and the ever-expanding realm of multimedia technology.

The essence of crafting an image caption lies in the intricate process of extracting vital image features and weaving them into a coherent textual narrative. This process unfolds through a model that gracefully navigates the nuances of image boundaries, constructing captions phrase by phrase. Recent advancements in image captioning models have yielded impressive outcomes, surpassing traditional approaches in terms of accuracy and efficiency. Quantitative and qualitative assessments further underscore the quality and correctness of the generated captions.

Beyond the realm of recreational technology, image captioning finds utility in addressing the vast repositories of visual data encountered on social networking sites, where celestial objects and an overwhelming volume of images are shared daily. The magnitude of annotating such vast collections poses a formidable challenge, rife with the potential for human errors. Deep learning models, equipped with their ability to accurately interpret and compile images, have emerged as a solution, effectively eliminating the need for manual corrections.

The development of image annotation holds far-reaching implications, from enhancing accessibility for individuals with disabilities to streamlining the auto-

mated processing of images across diverse applications. It becomes a cornerstone for smart devices, image encoding, and facilitates the engagement of visually impaired individuals on social networking platforms.

This report embarks on a comprehensive exploration of image captioning, delving into the mechanisms of Convolutional Neural Networks (CNNs) employed for feature extraction and the role of Long Short-Term Memory (LSTM) or Recurrent Neural Networks (RNNs) in decoding image descriptions. It seeks to elucidate the remarkable journey of image captioning, unveiling the amalgamation of cutting-edge technology and its transformative impact on our digital landscape.

Ground truth: a large brick clock tower in the middle of a town

Ground truth: two men that are standing in a kitchen

Figure 1: Captions to the Images

II MOTIVATION

The motivation behind image captioning lies in the advancement of human-like technologies, offering a multitude of opportunities for meaningful applications in society. This technology enables computers to interact with humans and find specific use cases in child education, health assistance for the elderly and visually impaired, among many others. Studies have been dedicated to obtaining more accurate descriptions and making machines think like humans.

Image captioning, closely related to computer vision and natural language processing, plays a pivotal role in automating tasks involving image interpretation, simplifying processes in various domains, and enhancing accessibility. It improves user experiences on websites and applications by providing image descriptions. This technology aids in content indexing, making image searches efficient, and simplifying content management. Furthermore, it contributes significantly to the development of artificial intelligence, pushing the boundaries of computer vision and natural language understanding.

Image captioning finds applications in education, communication assistance, content summarization, social media engagement, machine translation, autonomous vehicles, medical image analysis, and content tagging. Its ongoing advancements introduce innovations that benefit society by facilitating better communication between humans and machines, enabling cross-cultural interactions, preserving cultural heritage, and aiding in environmental monitoring. In essence, image captioning has a transformative impact on technology and society, driven by its potential to automate and enhance various aspects of our daily lives.

Overall, image captioning is a promising technology with the potential to benefit society in numerous ways. By automating tasks involving image interpretation, simplifying processes in various domains, and enhancing accessibility, image captioning can make our lives easier and more productive.

III LITERATURE SURVEY

A. Image Captioning using Neural Network Model

Waghmare et al.[1] focuses on developing a system for generating accurate image captions. It emphasizes the importance of enhancing feature extraction accuracy, giving priority to objects within images to improve caption quality. The dataset employed is the Flickr 8k dataset, comprising 8,000 images, each associated with five different text captions. Feature extraction is performed using a CNN model, and caption generation relies on an RNN with LSTM. The VGG 16 model effectively processes images by utilizing 16 deep layers to eliminate noise and ensure proper feature extraction. LSTM plays a crucial role in arranging words in a meaningful sequence, providing accurate and contextually relevant image captions

B. Image Caption using CNN in Computer Vision

Gaurav et al.[2] focuses on generating accurate image captions. In the context of generating image captions, they have explored various combinations of Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM). Their objective is to develop accurate image descriptions. These combinations involve CNN extracting image features, while RNN or LSTM generates the language model. A well-known approach uses CNN to recognize image content and objects, while RNN or LSTM is employed for language generation. Evaluations are done using metrics like BLEU, METEOR, ROUGE-L. The results revealed that combination of CNN and LSTM, yielding the best outcomes. The research leverages a dataset called MSCOCO, containing 330,000 images, each associated with five relevant image captions. The primary aim is to explore different deep learning techniques and models for image captioning, focusing on CNN for

image content recognition and RNN or LSTM for to establish a memory of past events and use it to process current input.

C. Hybrid Image Captioning Model

Raghab et al.[3] proposed a hybrid model is developed for automatic image captioning, combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) with an attention network. The architecture follows an encoder-decoder paradigm, where the CNN serves as the encoder, transforming images into vector representations, while the LSTM acts as the decoder, generating coherent sentences. To enhance performance, datasets are loaded progressively into the model, optimizing data processing. The introduction of an "attention model" improves accuracy and enables the system to handle video input, producing descriptive captions. The primary objectives of this study encompass the creation of a VGG16-based CNN-LSTM RNN encoder-decoder model and the application of attention mechanisms for prediction and feedback, addressing the multifaceted challenges of image captioning.

D. Analysis of Different Feature Extractors for Image Captioning Using Deep Learning

Dhaval et al.[5] presented various combinations of encoders and decoders for image captioning. The dataset employed is the Flickr 8k dataset, comprising 8,000 images, each associated with five different text captions. Models were trained and tested on the Flickr 8k dataset. BLEU score for each model was calculated and compared. Among different models, Attention resnet 101 with Glove 6B performed well as compared to other models.

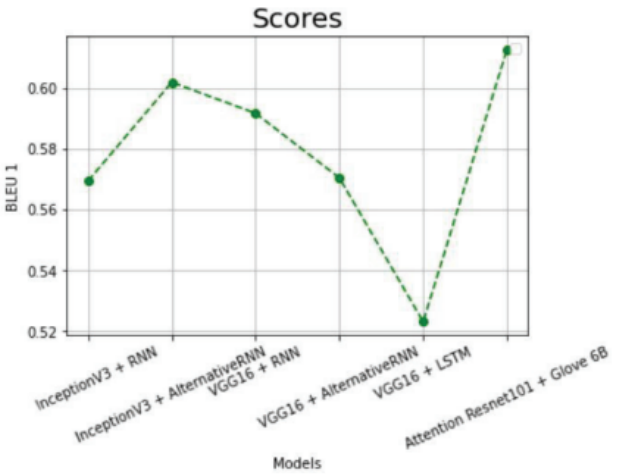

Figure 2: Graph of evaluation metric for different models used

E. Automatic Indonesian Image Captioning using CNN and Transformer-Based Model Approach

Rifqi et al.[7] focuses on the implementation of a Transformer-based model for Indonesian image captioning. The model uses a combination of CNN (Convolutional Neural Network) with ResNet as the encoder and Transformer with self-attention mechanism as the decoder. The study aims to generate Indonesian captions for images and the translation process for Indonesian image captioning involved using Google Translate and evaluate the performance of the captioning approach. The dataset employed is the Flickr 8k dataset, comprising 8,000 images, each associated with five different text captions. The results show that the Transformer-based strategy outperforms the previous approach, achieving high scores in metrics such as BLEU, METEOR, $\text{ROUGE}_L, and CIDEr$.

F. Image Caption Generation Using Encoder-Decoder Model

Abhishek et al.[8] proposed vision 360 is an innovative application designed to assist individuals with visual impairments through image captioning technology. It employs a sophisticated approach by utilizing a CNN and LSTM encoder-decoder model, coupled with two pretrained models, InceptionV3 for image feature extraction and GloVe for text encoding. Notably, the activation function (softmax) was eliminated from the pretrained models to retain only the essential feature vectors. The convergence of these models culminates in feeding data to an LSTM model, enabling the generation of descriptive captions

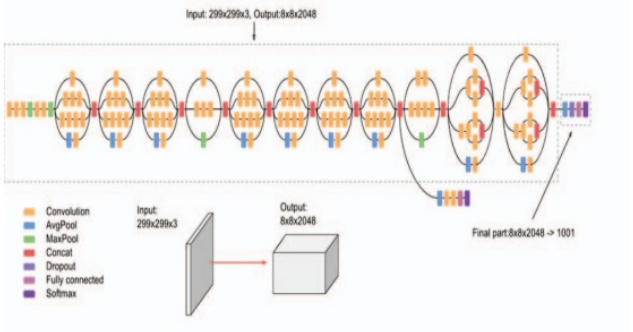

Figure 3: Inception V3 Architecture

IV METHODOLOGY

The encoder part receives the image as input and generates a feature vector of high dimensions. The

image goes through the layers of the Encoder model that converts the image to feature vector. The system does not understand anything but 0's and 1's and thus the image is converted to 0s and 1s. The image vector created is the form of binary format. The encoder is a group of convolutions that uses filters and uses pooling to help the system to focus on certain regions of the image. The encoded features from the encoder are given as input to the decoder. The decoder takes this high dimension features and generates a semantic segmentation mask. It is a process of recurrent units where it predicts an output at each stage. It accepts a hidden state and produces an output and its own hidden state. Various activation functions are used in the encoder-decoder model as per the application of the user.

Encoders
Convolutional and Recurrent networks are the head and tail of the encoder decoder model. The convolution neural network works on the two dimensional image data and can be used in other dimensional data too. Use ResNet-50 to extract visual features from images. These features will serve as the input to our model.

Attention Mechanism
Implement an attention mechanism that combines image features and word embeddings. This allows our model to give more attention to certain image regions when generating specific words in the caption.

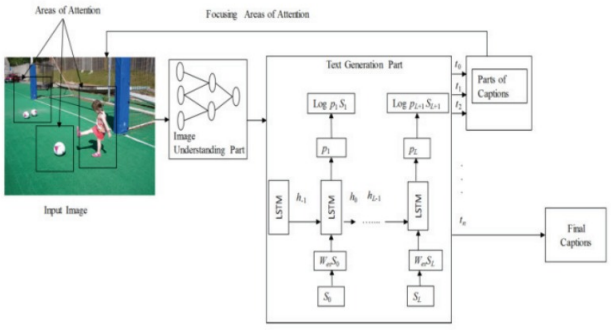

Figure 4: Attention mechanism

Decoders
Use LSTM (Long Short-Term Memory) as the recurrent neural network decoder. LSTM works by taking the image features and generating captions word by word. Maintain a memory of the previously generated words, enabling it to capture sequential dependencies in the text. Incorporate attention mechanisms to allow the model to focus on different parts of the image and provide more contextually relevant captions.

GloVe word embeddings
Incorporate gloVe word embeddings into the model. These embeddings enhance the model's ability to understand word semantics, generalize to diverse vocabulary, and improve overall performance.

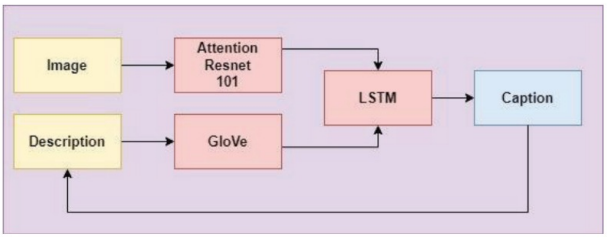

Fig. 3. Architecture for GloVe based mechanism

Figure 5: Glove word embeddings

By organizing these components into our model architecture, we can create a powerful system for generating image captions. The feature extraction step with ResNet-50 provides the visual context, the attention mechanism refines the focus, LSTM handles sequential word generation, and GloVe embeddings enhance the language understanding. This combination leverages both visual and semantic information to generate contextually relevant and coherent image captions.

V DESIGN

CNN Feature Extraction [Resnet - 50]

The input image typically undergoes preprocessing to meet the requirements of the ResNet-50 model. This may involve resizing the image to a specific input size that ResNet-50 expects, such as 224x224 pixels. Resizing ensures uniformity and compatibility with the network architecture.

Feature Extraction with ResNet: The resized image passes through a series of convolutional layers in the ResNet-50 architecture. These layers are designed to detect and extract hierarchical features from the image at different levels of abstraction. As the image progresses through these layers, it becomes encoded into a set of high-level features. The convolutional layers capture patterns, textures, and structures in the image, transforming it into a meaningful representation.
Convolutional Layers: The image then traverses through several convolutional layers, each consisting of residual blocks. ResNet-50 has 48 convolution layers

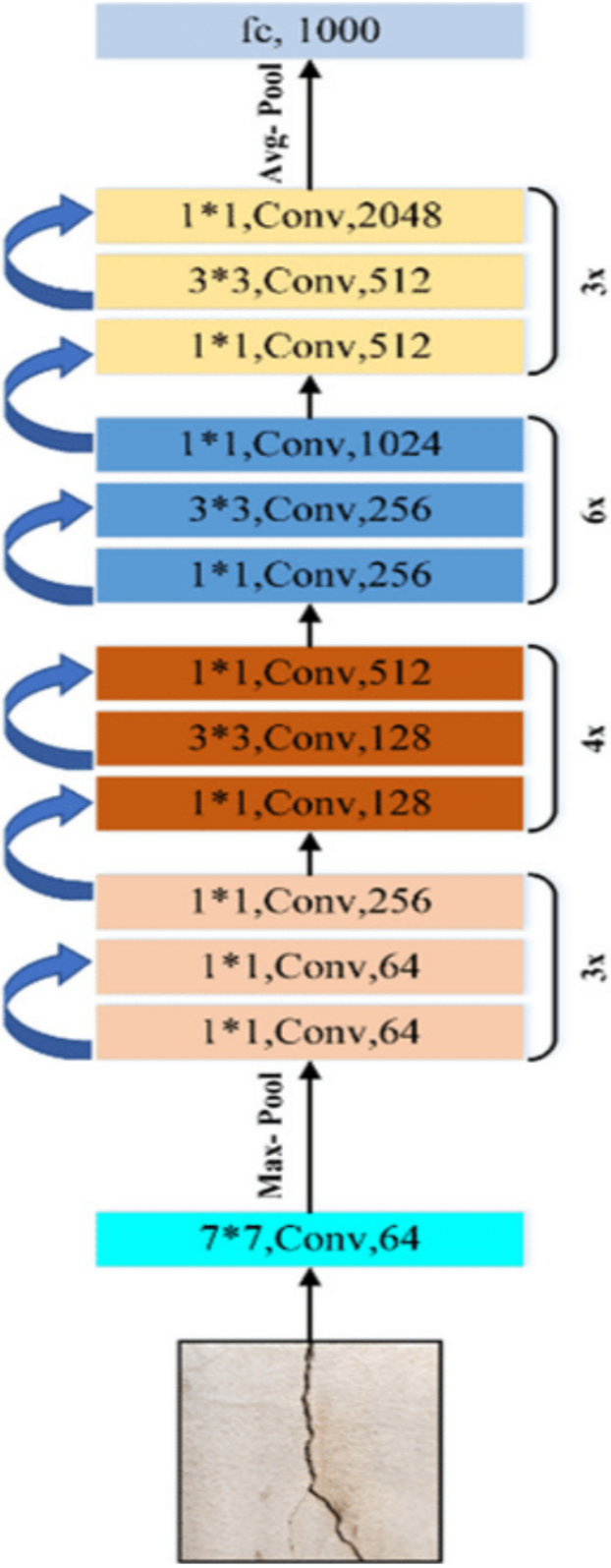

Figure 6: Architecture of Resnet-50

grouped into residual blocks, where the residual connections allow for more efficient training by mitigating the vanishing gradient problem. Within these blocks, convolution operations extract features at various levels of abstraction, identifying patterns, edges, and textures.

Pooling Layers: At specific intervals, max-pooling and average-pooling layers downsample feature maps, reducing spatial dimensions to capture the most relevant information. These pooling layers help in preserving important features while reducing computational load and preventing overfitting.

Fully Connected Layers: Post the convolutional layers, ResNet-50 employs fully connected layers, usually comprising an average pooling layer, a dropout layer for regularization, and a flatten layer to prepare the feature maps for the final classification or output. The features are aggregated, flattened into a vector, and passed through these layers for higher-level feature abstraction and classification.

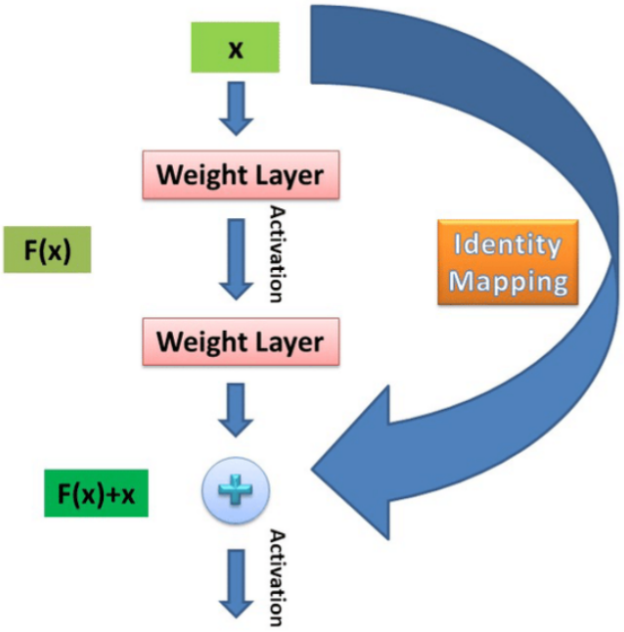

Figure 7: Skip Connections

Output Layer: The output of ResNet-50 is a condensed feature vector capturing the most significant information extracted from the image. This vector serves as a rich representation, often used for various downstream tasks such as image classification or, in the case of image captioning models, as input to generate descriptive captions.

CNN Feature Vector Creation: After traversing

through the convolutional layers, the processed image features are aggregated and condensed into a fixed-length vector known as the CNN feature vector. This vector encapsulates the most relevant and discriminative information extracted from the image by ResNet-50. This condensed representation serves as a rich and informative input to the subsequent stages, often used as the basis for generating captions using recurrent neural networks (RNNs).

LSTM

The feature vector is fed as an input to the LSTM model. This initiates the LSTM's sequence learning process. The LSTM is a type of recurrent neural network (RNN) capable of retaining information over sequences, making it suitable for processing sequential data, such as text.

Word Generation Process: The LSTM decodes the feature vector by learning to predict the next word in the sequence, effectively generating captions. Initially, the LSTM receives the feature vector and a start token, signaling the beginning of the caption generation process. Then, at each time step, the LSTM predicts the probabilities of the next word in the sequence based on the feature vector and the previous words generated.

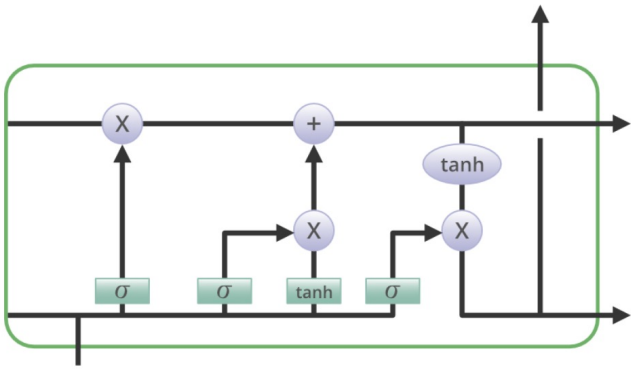

Figure 8: Word Generation Process

Input Gate: The input gate determines which values from the new input should be updated and added to the cell state. It uses a sigmoid activation function to generate values between 0 and 1, where 0 means "ignore" and 1 means "keep." These gate values are multiplied by the hyperbolic tangent layer output. Hyperbolic Tangent Layer: The hyperbolic tangent layer processes the new input, creating a vector of new candidate values. The layer squashes the values to be between -1 and 1, making it suitable for the LSTM cell state. Combining Input Gate and Tanh Output: The values obtained from the input gate (scaled by the sigmoid) are multiplied by the corresponding values from the hyperbolic tangent layer. This product represents the new information that will be added to the cell state.
Introducing New Information: The product of the input gate and hyperbolic tangent layer introduces new information to the cell state. The purpose is to update the cell state selectively, considering the importance of each element in the new input.
The multiplicative filters allow to effectively train LSTM, as they are good to prevent the exploding and vanishing gradients. Nonlinearity is provided by the sigmoid and the hyperbolic tangent. The last equation, is fed to the softmax function to calculate the probability distribution over all words. This function is calculated and optimized on the entire training dataset. The word with maximum probability is selected at each time step and fed into next time step input to generate a full sentence.

Attention added to LSTM

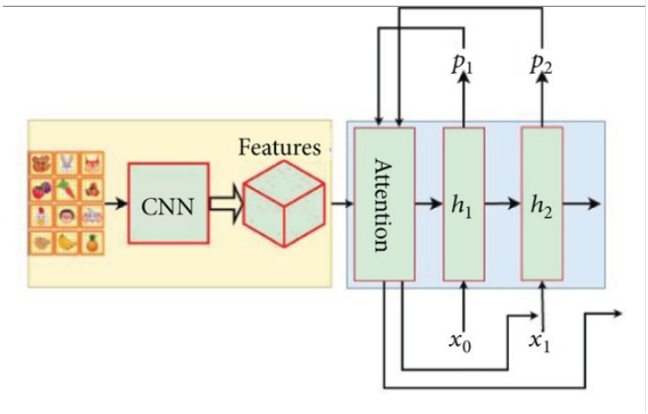

Figure 9: Model with added attention to LSTM [10]

Attention Implementation:

Use soft attention, such as Bahdanau or Luong attention mechanisms, which calculate attention weights over different parts of the image features during caption generation by the LSTM decoder. Calculation of Attention Scores:

Calculate attention scores based on the alignment between the current decoder LSTM hidden state and the encoded image features. Common methods include: Dot Product Attention: Computes scores by taking the dot product between the decoder hidden state and the encoder image features. Additive Attention: Calculates scores by passing both encoder and de-

coder hidden states through a neural network to obtain compatibility scores. Attention Mechanism Workflow:

During each decoding step, compute attention scores, normalize them using softmax to obtain attention weights, and create a context vector by weighting the image features with these weights. Caption Generation:

Use the context vector along with the LSTM hidden state to predict the next word in the caption sequence. The combined implementation of attention and LSTM can be seen in Figure 4.

GloVe Word Embeddings

To integrate GloVe word embeddings into an LSTM model, follow these steps:
Download GloVe Embeddings: Obtain pre-trained GloVe word embeddings from sources like Stanford NLP. These embeddings come in various dimensions (e.g., 50D, 100D, etc.).
Load GloVe Embeddings: Use Python to load the GloVe embeddings into the script. Iterate through the GloVe file and create a dictionary mapping words to their respective embedding vectors.
Map Words to GloVe Vectors: Map the words in the dataset to their GloVe vectors using the dictionary created in the previous step. This mapping allows words in the dataset to have corresponding GloVe embeddings.
Use GloVe Embeddings in LSTM Model: In the LSTM model's embedding layer, set the weights to the GloVe embedding matrix. Ensure the embedding layer is non-trainable to retain the pre-trained GloVe weights.
Train the Model: Train the LSTM model on the specific task, whether it's sentiment analysis, language translation, or any text-based task. During training, the LSTM learns to leverage the contextual information encoded in GloVe embeddings for improved performance.
This integration helps enhance the model's understanding of language nuances by leveraging pre-trained word embeddings.

VI CONCLUSION

The development of image captioning represents a significant leap in the intersection of artificial intelligence and computer vision. It enables the transformation of visual content into understandable text, thereby facilitating a wide range of applications, including accessibility for the visually impaired, enhanced gaming experiences, and improved multimedia technology. However, despite impressive advancements in image captioning, replicating the nuanced communication of human brains remains a challenge.

This project focuses on bridging the gap by merging computer vision and natural language processing through deep learning techniques. It employs Convolutional Neural Networks (CNNs) for extracting image features and Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) for language modeling. These components work together to understand the intricate relationships between visual content and textual descriptions.

The motivation behind image captioning lies in its potential to improve human-computer interactions, benefiting fields like child education and health assistance for the elderly and visually impaired. Image captioning simplifies image interpretation tasks and enhances accessibility across various applications.

The literature survey highlights various approaches to image captioning, from combining CNNs and LSTMs for accurate caption generation to the use of attention mechanisms for improved accuracy. The choice of feature extractors, such as ResNet-50, and the incorporation of GloVe word embeddings play a pivotal role in achieving contextually relevant and coherent image captions.

In conclusion, image captioning is a promising technology that has the potential to benefit society in numerous ways, from automating image interpretation to enhancing user experiences and pushing the boundaries of artificial intelligence. By bridging the gap between visual content and language, image captioning offers transformative solutions for our digital landscape and daily lives.plications in our increasingly visual and data-driven world.

REFERENCES

[1] Prachi Waghmare, Swati Shinde, Jayshree Katti, "Image Captioning using Neural Network Model," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2022.

[2] Rohit Kumar, Gaurav Goel, "Image Caption using CNN in Computer Vision," in IEEE Transactions on Image Processing, 2023.

[3] Lipismita Panigrahi, Raghab Ranjan Panigrahi, Saroj Kumar Chandra, "Hybrid Image Captioning Model," in IEEE Computer Vision and Pattern Recognition (CVPR), 2022.

[4] Omkar Sargar, Shakti Kinger, "Image Captioning using Neural Network Model," Journal of Artificial Intelligence Research, 2022.

[5] Srusti Shinde, Dhaval Hatzade, Shubhankar Unhale, Gaurav Marwal,"Analysis of Different Fea-

ture Extractors for Image Captioning Using Deep Learning," 2022.

[6] Vishal Singh, Ajay Shankar Singh, K Anandhan, "Image Captioning Using Machine/Deep Learning," 2022.

[7] Rifqi Mulyawan, Andi Sunyoto, Alva Hendi Muhammad, "Automatic Indonesian Image Captioning using CNN and Transformer-Based Model Approach," 2022.

[8] Ankita Kumari, Abhishek Chauhan, Abhishek Singhal, "Image Caption Generation Using Encoder-Decoder Model," 2022.

[9] Rifqi Mulyawan, Andi Sunyoto, Alva Hendi Muhammad, "Automatic Indonesian Image Captioning using CNN and Transformer-Based Model Approach," 2022.

[10] Yan Chu, Xiao Yue, Lei Yu, Mikhailov Sergei, Zhengkui Wang, "Automatic Image Captioning Based on ResNet50 and LSTM with Soft Attention", Wireless Communications and Mobile Computing, vol. 2020