Building Image Caption Based on InceptionV3 Using Convolutional Neural Network

Vansh Ahlawat
Department of Computer Science
Bennett University
Greater Noida,Uttar Pradesh
e22cseu0403@bennet.edu.in

Utsav
Department of Computer Science
Bennett University
Greater Noida,Uttar Pradesh
e22cseu0396@bennet.edu.in

Ashwani Kumar
School of Computer Science
Engineering and Technology
Bennett University
Greater Noida,Uttar Pradesh
ashwani.kumar@bennett.edu.in

Avinash Kumar Sharma
Department of CSE, SSET
Sharda University
Greater Noida,Uttar Pradesh
avinash.sharma1@sharda.ac.in

***Abstract*—Image captioning is the generation of contextually relevant textual data from the visual content of an image. Advancements in computer vision and natural language processing in the field of deep learning have made great improvements. This work demonstrates a deep learning-based approach that effectively integrates Long Short-term Memory(LSTM) with Convolutional Neural Network(CNN) to generate meaningful captions. The CNN component of the model utilizes InceptionV3 architecture which is used to extract complex-level features from the images. These features are then passed to the LSTM network, a type of Recurrent Neural Network(RNN), which excels at predicting and learning long-term dependencies-making it suitable for generating contextually accurate captions from images. Flickr8k dataset is used for both training and testing, which consists of 8000 images, each annotated with five different captions. One of the most important part of the process is data preprocessing, which includes cleaning the text, tokenizing, and converting images into suitable feature vectors. The model is trained over multiple epochs and performance is calculated using standard evaluation measures like METEOR. The findings indicate that the model generates accurate and context-aware captions. The evaluation scores improve significantly with additional training, which indicates effective learning over time. This study shows the potential of combining CNNs and LSTMs for solving the image captioning problem and contributes in the ongoing research in multi-model deep learning.**

***Index Terms*—Natural Language Processing, Convolutional Neural Network, Recurrent Neural Network, Long ShortTerm Memory (LSTM) Network**

I. Introduction

The artificial intelligence technology known as image captioning works by uniting artificial vision capabilities with natural language processing (NLP). Through text caption generation systems, computers can interpret visual material in ways that resemble human understanding. Image captioning technology serves various essential functions which help both the visually impaired and provides automation in enhancing image searches and social media content creation. The current generation of image captioning systems relies on Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) with a specific implementation called Long Short-Term Memory (LSTM) networks as their fundamental design component. Through CNNs users can obtain layered graphical elements from images which detects items while performing feature recognition on picture patterns and spatial connections. The created features establish a solid base for determining image content meaning. .Sequential data processing serves RNNs especially LSTMs well for generating word-by-word captions which maintain both coherence and grammatical correctness. An image captioning system achieves efficient visual content-to-language translation through the integration of these two architectural types. This field has recently witnessed several developments in methods that boost the quality of generated captions. Through attention mechanisms the algorithm selects image components that are relevant to generate each word thus producing better quality captions. Transformer-based models along with multimodal frameworks continue to extend limitations through integration of large language models and external knowledge bases which generates sophisticated and professional descriptions. CIDEr and BLEU and METEOR serve as standard evaluation metrics to evaluate automated caption quality until they reach human-verified(floor) results. Certain innovative developments exist although future work needs to address multiple object scene processing along with detailed description maintenance and specialized captions for medical domains. Research activities focus on developing hybrid systems and bigger data resources and innovative training approaches to correct current system weaknesses.

II. Related Work

The task of image captioning involves generating textual content using visual data. This has gained a substantial amount of improvement due to the advances in natural language processing and computer vision. Earlier approaches failed to generalize unseen data but the deep learning techniques have enabled to learn directly from the image-caption pairs.

Chalcheema Sasidhar et al used a CNN model Exception which is used for extracting features from an image. These extracted features are fed to the Long Short Term Vector Machine (LSTM) model which creates image captioning possible. Normally RNNs are used which can process sequential data in the field of NLP, however they struggle with long term dependencies which is why LSTMs are preferred. CNN for image extraction and LSTM for caption generation make a powerful combination and approach which can be used for image captioning.

ResNet-50 is a popular architecture used for image feature extraction. It extracts the important and meaningful features of an image. Satish Kumar Satti et al used the ResNet-50 convolutional neural network and LSTM recurrent neural network as it serves as a robust feature extractor and with through fine-tuning it helps capture the semantic essence of the image.

K.Priya et al developed a system with an extension where it can generate descriptive captions and then convert them into audio output. It uses CNN to extract visual features from images which are then fed into a GRU to generate captions. These captions are then converted into audio output using text-to-speech techniques.

Dessy Santi et al focused and aimed to understand how the depth of the network affects the visual extraction and in improving textual descriptions. The study focuses on CNNs with residual network architectures ResNet-50,ResNet-101 and ResNet-152 tested on the Flickr 8k dataset where ResNet-101 model achieved the highest BlEU score among the tested models.

Rajwinder Kaur and Gurpreet Singh discuss the automatic image captioning field along with the essence of image captioning in the biomedical field. Extraction of such kind of information requires expert opinion. Automatic detection of the clinical information contained in these medical images can be possible using the image captioning techniques discussed in their paper.

Jianjun Xia and Xin Yang's research promotes the Tibetan caption generation method which integrates with LSTM encoding and decoding network to improve attention mechanisms. This method results to be superior than the traditional attention mechanism image caption generation method when experimented on the Flickr8k and Flickr30k-tic Tibetan caption datasets. VGG19,InceptionV3 and ResNet-101 are used as the backbone and group convolution is used to replace the attention mechanism. Shadika Afroze Ome et al utilizes a vision transformer replacing traditional CNN and a Meshed Memory Transformer for generating captions. The images

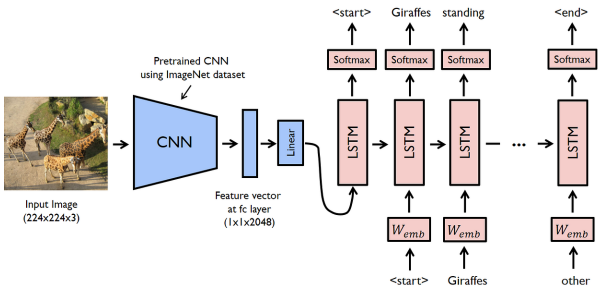

Fig. 1. Overview of Image Captioning Using CNN and LSTM

are spit into fixed-size patches by the vision transformer for extracting visual features which are further processed by memory-augmented encoder. The key features are preserved and reused and the decoder outputs the more accurate and rich captions. The model outperformed the conventional "CNN-LSTM" models and "CNN-Transformer" methods, achieving a 1.02 improvement in METEOR and 1.31 in SPICE on the MSCOCO dataset.

Tengfei Wan et al designed a framework which incorporates linear mapping strategies enhancing image recognition accuracy by adding fine-tuning matrices to a pretrained contrastive language-image mode. The linear mapping layer directly transforms image feature vectors into text feature vectors. It also facilitates a feedback between model outputs and knowledge base to improve performance. The model outperformed existing methods on major evaluation metrics such as BLEU-4 and CIDEr. Genc Hoxha et al used the CNN-RNN framework for Remote Sensing Image Captioning. It provides more semantic information with respect to the traditional tasks such as scene classification. Best caption is selected based on its lexical similarity with the reference captions after the CNN-RNN framework generates multiple captions for a target image.

III. Methodology

A. Data Description

Flickr8k dataset is used in this study containing 8000 images sourced from flickr.Each image is annotated with five captions written by different human annotators. The dataset includes variety of images making it suitable for training and testing the image captioning models.

B. Data Cleaning

To prepare the captions for training the models, the text was all lowered to lowercase and had all punctuation removed. The startseq and endseq special tokens were added to indicate the start and the end of each of these captions. A tokenizer was then used to transform the cleaned-up captions into a sequence of integer tokens according to the vocabulary created from the dataset.

C. Visual Feature Extraction

Visual features were extracted from the InceptionV3 pre-trained ImageNet convolutional neural network. The final

classification layer was removed from it and features were extracted from the average pooling layer generating a 2048-dim vector for each image. Images were resized to 299×299 pixels and normalized in accordance with InceptionV3's pre-processing requirements before feature extraction.

D. Model Architecture

The model has an encoder-decoder structure. InceptionV3 is utilized in the encoder to transform the image to feature vectors, which are then input into the decoder. The decoder has an embedding layer followed by an LSTM network in order to produce captions word for word. The embedding layer projects every word index to a dense representation, and the LSTM is trained on learning sequential relationships in the caption. The outputs of both the LSTM and image features are concatenated and fed through a dense softmax layer to make predictions for the upcoming word. The model is optimized using categorical cross-entropy loss via Adam. When generating captions, the model predicts word for word, and the output is fed back into the LSTM until it reaches the endseq token or reaches the maximum length for the caption.

IV. Training and Evaluation of Model

A. Training Process

The images that have been prepared go through the CNN-LSTM model for training. InceptionV3 is employed to process the image features, and then the last classification layer is not used. Features are extracted from the average pooling layer to create a 2048-dimensional image feature. Their captions are also transformed into a sequence of word indexes, and then both the features and the captions are combined. When training, the system uses each image feature vector and a sequence of words with the LSTM. Categorical cross-entropy was used to select the following word for the caption. Learning is performed using Adam with a learning rate of 0.001. An embedding is spread out over 256 dimensions, while the LSTM has 256 hidden units. Both the image feature projection and LSTM layer have an auxiliary dropout rate of 0.5 to deal with overfitting. Experiments are carried out by starting training with 64 records (batch size) and continuing for 25 to 50 epochs. After each batch, the model's weights are modified, while the validation loss is monitored to check the performance. The checkpoints are saved when the model performs best on the validation data. Using the GPU for training speeds up the process and leads to faster results.

B. Evaluation Metrics

The performance of the model is predominantly assessed using the METEOR metric due to its ability to consider synonymy, stemming, and word order. METEOR is more balanced and semantically aware compared to BLEU. METEOR score in our experiments improved steadily from about 0.16 at epoch 5 to about 0.21 at epoch 25. The score shows the model's increasing capacity to learn contextually relevant and syntactically correct captions. Although METEOR is predominantly utilized in the current research, BLEU, CIDEr, and ROUGE-L can also be utilized to present complementary assessment of model performance.

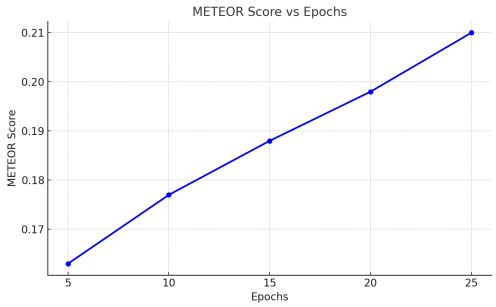

Fig. 2. meteor score vs epochs

Fig. 3. Generated Caption: "Black dog swimming in the water"

Epochs	METEOR Score
5	0.16
10	0.18
15	0.19
20	0.20
25	0.21

TABLE I
PERFORMANCE OF THE MODEL ACROSS DIFFERENT EPOCHS USING METEOR SCORE.

C. Impact of Image Captioning

The performance of the model is predominantly assessed using the METEOR metric due to its ability to consider synonymy, stemming, and word order. METEOR is more balanced and semantically aware compared to BLEU. METEOR score in our experiments improved steadily from about 0.16 at epoch 5 to about 0.21 at epoch 25. The score shows the model's increasing capacity to learn contextually relevant and syntactically correct captions. Although METEOR is predominantly utilized in the current research, BLEU, CIDEr, and ROUGE-L can also be utilized to present complementary assessment of model performance.

V. Results and Discussion

A. Performance Evaluation

The performance of the model is predominantly assessed using the METEOR metric due to its ability to consider synonymy, stemming, and word order. METEOR is more balanced

and semantically aware compared to BLEU. METEOR score in our experiments improved steadily from about 0.16 at epoch 5 to about 0.21 at epoch 25. The score shows the model's increasing capacity to learn contextually relevant and syntactically correct captions. Although METEOR is predominantly utilized in the current research, BLEU, CIDEr, and ROUGE-L can also be utilized to present complementary assessment of model performance.

B. Impact of Attention Mechanism

While our present model lacks an explicit attention mechanism in its implementation, the findings indicate the addition of attention can improve performance considerably. Attention has been shown to enable models to target specific locations of an image of interest, typically resulting in higher accuracy and more descriptive captions. An attention-based LSTM decoder will be included in future versions of the model for comparison and measuring its impact on performance and caption quality.

C. Effect of Training Data

Flickr8k dataset of 8000 images annotated with five captions was used to train the model. Although the dataset is adequate for proof of concept, its small size is a limitation on the model's generalization capacity. In experiments with other related research, it has been demonstrated that increasing to Flickr30k- or MSCOCO-sized datasets can improve caption quality significantly and boost evaluation measures. Having a larger and more varied training set would enable the model to learn more complex relationships between visual features and text descriptions.

D. Qualitative Analysis

The captions produced by our model tend to be coherent and pertinent to the visual elements. For instance, the model generated captions like "A man on a skateboard in the park" and "A black dog is jumping in the water" that accurately described the scenes. However, in some cases, the model produced ambiguous and inaccurate captions where there were cluttered background scenes or overlapped objects. In these cases, the captions either omitted important details or were repetitive. These errors identify gaps in the model's ability to capture in-depth spatial relationships among images. Table 1 presents the METEOR scores for various epochs.

E. Comparison with Different Architectures

Although transformer models such as Vision Transformers (ViTs) and transformer decoders have demonstrated excellent performance in image captioning tasks, we specifically worked on the conventional CNN-LSTM pipeline. Even though transformers have state-of-the-art performance on large datasets, these usually demand more computation and larger datasets for optimal performance. Our CNN-LSTM model performed stably on a small dataset such as Flickr8k and generated well-structured, coherent captions using less number of epochs. In low-resource environments, CNN-LSTM continues to be an efficient and viable way.

Fig. 4. Generated Caption: "Two children on the beach at sunset"

F. Generated Captions by Our Models

The captions produced by our model proved to be syntactically sound and semantically accurate in most test cases. For example, in figure 3 the caption "Two children on the beach at sunset" accurately described the contents of the image. However, occlusions of objects and changing lighting conditions occasionally resulted in partial or inaccurate descriptions.

G. Limitations and Future Work

Despite promising results, our model has several limitations. It struggles with images that contain complex or crowded scenes and occasionally generates repetitive or overly simplistic captions. Furthermore, the reliance on a relatively small dataset like Flickr8k limits the model's ability to generalize to diverse visual inputs. Future work will involve integrating an attention mechanism, experimenting with transformer-based models, and scaling up to larger datasets such as MSCOCO. Enhancing the diversity of training data and incorporating external knowledge bases could also help the model generate richer and more informative captions.

VI. Conclusion

In this study, we developed an image captioning model that effectively combines convolutional neural networks for visual feature extraction with Long Short-Term Memory networks for sequential caption generation. Using the InceptionV3 model to extract high-level image features and an LSTM decoder to generate descriptive text, the model was trained and evaluated on the Flickr8k dataset. The METEOR score was used to assess the quality of the generated captions, with results showing consistent improvements across training epochs. While the model was able to produce coherent and relevant captions for a variety of images, it showed limitations when handling complex or crowded scenes. The study demonstrates that even with a relatively small dataset, a well-structured CNN-LSTM architecture can achieve strong results in image captioning tasks. Future work will focus on integrating attention mechanisms, exploring transformer-based architectures, and scaling to larger, more diverse datasets to further enhance caption quality and generalization.

REFERENCES

1) Chalcheema Sasidhar, Dr. Madan Lal Saini, Medarametla Charan, Avula Venkata Shivanand, Dr. Vijay Mohan Shrimal. *Image Caption Generator Using LSTM*. https://ieeexplore.ieee.org/document/10841294
2) Satish Kumar Satti, Goluguri N V Rajareddy, Prasad Maddula, N V Vishnumurthy Ravipati. *Image Caption Generation using ResNET-50 and LSTM*. https://ieeexplore.ieee.org/document/10404600
3) K. Priya, R. Vijay Babu, M. Muralidhar Reddy, T. H. Mohan Reddy, M. Maanesh. *Image Caption Generation using CNN and Audio Conversion*. https://ieeexplore.ieee.org/document/10760457
4) Dessy Santi, Syafaruddin, Amil Ahmad Ilham, Ingrid Nurtanio. *Image Caption Generation Through the Integration of CNN-Based Residual Network Architectures and LSTM*. https://ieeexplore.ieee.org/document/10636926
5) Rajwinder Kaur, Gurpreet Singh. *Automatic Image Captioning for Medical Diagnosis Report Generation*. https://ieeexplore.ieee.org/document/10444545
6) Jianjun Xia, Xin Yang, Qiong Ni, Dingguo Gao. *Research on Image Tibetan Caption Generation Method Fusion Attention Mechanism*.https://ieeexplore.ieee.org/document/10348351
7) Shadika Afroze Ome, Tanvir Azhar, Asaduzzaman. *A Transformer-based Model for Image Caption Generation with Memory Enhancement*. https://ieeexplore.ieee.org/document/10915030
8) Tengfei Wan, Huiyi Liu, Lijie Geng. *A New Multimodal Large Model Framework for Knowledge-enhanced Image Caption Generation* https://ieeexplore.ieee.org/document/10898381
9) Genc Hoxha, Farid Melgani, Jacopo Slaghenauffi. *A New CNN-RNN Framework For Remote Sensing Image Captioning*. https://ieeexplore.ieee.org/document/9105191
10) J Sudhakar, Vishwesh V Iyer, Sree T Sharmila. *Image Caption Generation using Deep Neural Networks*. https://ieeexplore.ieee.org/document/9726074
11) Zainab Umair Kamangar, Ghulam Mutjaba Shaikh, Saif Hassan, Nimra Mughai, Umair Ayaz Kamangar. *Image Caption Generation Related to Object Detection and Colour Recognition Using Transformer-Decoders*. https://ieeexplore.ieee.org/document/10099161
12) L Abisha Anto Ignatious, S Jeevitha, M Madhurambigai, M Hemalatha. *A Semantic Driven CNN – LSTM Architecture for Personalised Image Caption Generation*. https://ieeexplore.ieee.org/document/9087299
13) Swati Srivastava. *Review of Recent Datasets Used in Image Captioning Models*. https://ieeexplore.ieee.org/document/10864233
14) Rohit Kumar, Gaurav Goel. *Image Caption using CNN in Computer Vision*. https://ieeexplore.ieee.org/document/10085162
15) Rongrong Yuan, Haisheng Li. *An Image Captioning Model Based on SE-ResNest and EMSA*. https://ieeexplore.ieee.org/document/10332008