IMAGE CAPTIONING USING CNN AND LSTM

Nandam Sai Saketh (242IT021)
Department of Information Technology
National Institute of Technology Karnataka, Surathkal

Dasari Charan Srinivas Kumar Reddy (242IT007)
Department of Information Technology
National Institute of Technology Karnataka,Surathkal

Kunduru Phaneendra Reddy (242IT016)
Department of Information Technology
National Institute of Technology Karnataka, Surathkal

Mitesh Kumar Mandal (242IT020)
Department of Information Technology
National Institute of Technology Karnataka, Surathkal

Konduru Sai Kiran (242IT014)
Department of Information Technology
National Institute of Technology Karnataka, Surathkal

***Abstract*—This project focuses on the development of an image captioning system using deep learning techniques. The primary goal is to generate descriptive captions for images using a combination of Convolutional Neural Networks (CNN) for feature extraction and Recurrent Neural Networks (RNN), specifically LSTM (Long Short-Term Memory), for sequence prediction. The system uses the ResNet50 pre-trained model to extract image features[1], and the GloVe word embeddings to represent words in a continuous vector space[2]. The image and text features are combined in an encoder-decoder architecture, where the encoder processes the image features, and the decoder generates captions based on the learned image-text associations[3]. The model is trained on the Flickr8k dataset[4], and the captions are evaluated using the BLEU metric to measure the accuracy and fluency of the generated descriptions[5]. The system's performance demonstrates the effectiveness of combining CNN and RNN architectures for image captioning.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

Image captioning is a challenging task that involves generating a natural language description of an image. This task requires both visual understanding and language modeling, making it an important application of artificial intelligence and deep learning[6]. In recent years, significant progress has been made in this field using deep learning models that combine Convolutional Neural Networks (CNNs) for image feature extraction and Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, for generating sequences of words. The goal of this project is to create an image captioning model that can generate meaningful captions for images. We use the Flickr8k dataset, which contains 8,000 images, each associated with five human-generated captions[4]. The project involves two main components: image feature extraction and caption generation. For the image feature extraction, we employ the ResNet50 model, a deep CNN pre-trained on the ImageNet dataset. For caption generation, we use an LSTM-based decoder, which takes the extracted features and generates a sequence of words[1][3] that form a meaningful caption. The model is trained using the GloVe word embeddings to represent words in a vector space, allowing the model to better understand semantic relationships between words. The training process involves optimizing the model using the categorical cross-entropy loss function, and the generated captions are evaluated using the BLEU score, which measures the similarity between predicted captions and the ground truth captions. In the following sections, we describe the architecture of the image captioning model, the data preprocessing steps, the training procedure, and the evaluation metrics used to assess the model's performance. The results demonstrate the potential of deep learning models in bridging the gap between computer vision and natural language processing.

II. Related Work

A. CNN-LSTM Image Captioning Models

The CNN-LSTM hybrid architecture utilizes CNNs for extracting spatial features from images, which are then fed into LSTMs to generate sequential captions[2]. This combination benefits from CNNs' ability to capture local patterns and LSTMs' capacity to model dependencies between words in a sentence. The main advantage of this approach is that it effectively bridges the gap between visual data and natural language. However, the challenges lie in the computational demands and the need for large-scale image-caption paired datasets[3].

B. Attention Mechanism with CNN-LSTM

Attention mechanisms have been integrated with CNN-LSTM models to improve captioning accuracy by allowing the model to focus on specific regions of the image while generating descriptions[5][6]. In this architecture, the attention mechanism directs the LSTM to focus on important areas of the image, enhancing the relevance and detail of the generated captions. Although this method significantly improves the performance of captioning models, it can be complex to implement and requires fine-tuning of attention parameters[7].

C. Pre-trained CNNs with LSTM for Transfer Learning

In this approach, pre-trained CNNs (such as ResNet or Inception) are used as feature extractors, providing a powerful foundation for image understanding. These features are then passed into LSTM networks to generate captions. This transfer learning approach allows for the leveraging of large-scale image datasets, even when the caption dataset is limited[7][8], making it more efficient and scalable. However, this method may still struggle with domain-specific captions and require fine-tuning to match specific task requirements.

D. GAN-based Enhancement of Image Captioning

Generative Adversarial Networks (GANs) have been used to generate realistic images and enhance image captioning models. In this hybrid setup, a Generator produces images from text descriptions, and a Discriminator assesses the generated image's realism. The use of GANs in image captioning can improve the diversity and creativity of captions by generating visually coherent images that match the generated captions[9]. However, these models often face

Model	Contributions	Key Features
CNN-LSTM Hybrid	Image Feature Extraction, Caption Generation	Combines CNNs for Image Features, LSTMs for Text Generation
Attention-CNN-LSTM	Focused Captioning, Improved Accuracy	Attention Mechanism for Focused Learning
Pre-trained CNN + LSTM	Transfer Learning for Captioning	Utilizes Pre-trained CNNs for Feature Extraction
GAN-enhanced Image Captioning	Image Synthesis for Caption Diversity	Uses GANs for Creative and Diverse Captions

III. Dataset

The Flickr8k dataset is a popular benchmark used in the fields of computer vision and natural language processing, particularly for image captioning tasks. It consists of 8,000 images sourced from the Flickr photo-sharing platform, each accompanied by five descriptive captions that provide a comprehensive understanding of the image content. For this research, 6,000 images from the dataset are allocated for training, ensuring robust model learning and generalization. Additionally, 1,000 images are designated for validation, enabling the tuning of model hyperparameters and performance evaluation during development. The remaining 1,000 images are reserved for testing to objectively assess the model's final performance and its ability to generate accurate and contextually relevant captions

IV. Methodology

The methodology leverages both visual and textual data to achieve a cohesive and effective caption generation process. The image features, derived from a pre-trained CNN model, capture the critical visual elements of an image [1][2]. These features are passed through an InputLayer, followed by a Dropout layer, which reduces overfitting by preventing reliance on specific neurons during training [3]. To further optimize and compress the information, the features are projected into a smaller 256-dimensional feature vector using a Dense layer, ensuring that the visual representation is compact and suitable for integration with textual data [2][4].

In parallel, the textual input, which consists of sequential word data, undergoes its own processing pipeline. It starts with an InputLayer that handles raw text data, which is then transformed into dense numerical representations using an Embedding layer [5]. These embeddings map words into a 50-dimensional space, capturing semantic and syntactic information that aids in contextual understanding [6]. To ensure robustness and prevent overfitting, the embeddings are regularized using a Dropout layer before being fed into an LSTM layer [4]. The LSTM layer plays a critical role by capturing temporal dependencies and relationships within the sequence, ultimately producing a 256-dimensional vector that represents the text input [7].

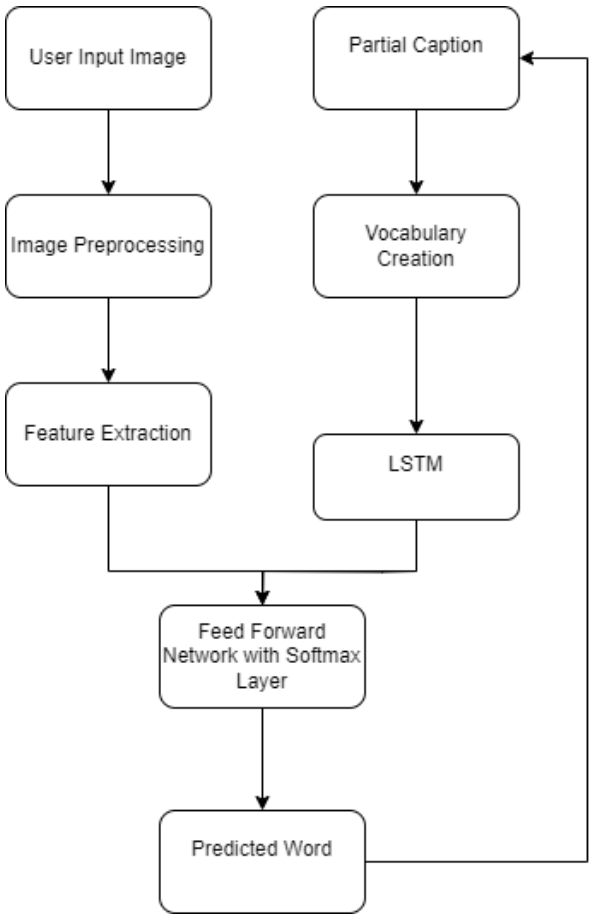

Fig. 1. Work Flow

A masking mechanism, such as a NotEqual layer, is employed to handle padding tokens in the sequence, ensuring that non-informative padding values do not interfere with the learning process. This step maintains the integrity of the textual features during LSTM processing [3]. Once both the image features and textual embeddings have been independently processed and reduced to 256 dimensions, they are combined through an Add layer [5]. This fusion effectively integrates the visual and linguistic modalities, enabling the model to consider both forms of information simultaneously [6].

The combined vector, which encapsulates the critical fea-

tures of the image and text, is further refined through two Dense layers. These layers enhance the representation, applying non-linear transformations to extract and distill meaningful patterns from the fused features [2][8]. The final Dense layer is tasked with generating predictions for the next word in the sequence, outputting probabilities across a vocabulary of size 1848. This vocabulary covers all possible words that the model can predict, enabling it to construct coherent and contextually relevant captions [4].

This approach harmonizes the strengths of CNN-based visual feature extraction and LSTM-based sequential modeling [2][7]. By integrating these two modalities, the model achieves a robust understanding of the visual content while maintaining the ability to generate accurate and fluent textual descriptions [9]. The use of dropout layers, dimensionality reduction, and multimodal fusion ensures a well-regularized and efficient architecture capable of producing high-quality image captions. This methodology exemplifies the power of combining advanced neural networks and thoughtful architectural design to solve complex generative tasks [10].

Fig. 2.

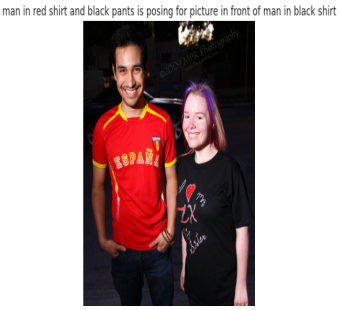

Fig. 3.

Fig. 4.

Fig. 5.

V. Experiment

A. Dataset

The dataset comprises paired images and descriptive captions, formatted to align with the model input requirements.

B. Hardware

The experiments were conducted in a GPU-accelerated environment, facilitating parallel processing and faster model training.

C. Pre-trained Models

ResNet50 was used for extracting image features, while GloVe embeddings were employed for text representation.

D. Training parameters

Epochs : 50, Batch Size = 32, Loss Function : Categorical Cross Entropy, Optimizer : Adam

VI. Key Observations

A. Training efficiency

Total time for feature extraction and model training was measured to evaluate computational efficiency.

B. Model Performance

Performance was monitored through metrics such as training accuracy, validation loss, and sample outputs. Intermediate model weights were stored to facilitate performance comparison across different training epochs.

VII. Sample Image Outputs

Results

The use of transfer learning, particularly with ResNet50 and GloVe embeddings, proved effective for extracting robust features from image and text data. The combined architecture demonstrated improved predictive capability, generating fluent and contextually appropriate outputs. Performance metrics and qualitative assessments of output captions highlighted the advantages of leveraging pre-trained models for complex multimodal tasks.

VIII. Quantitative Analysis of Model Performance

To evaluate the performance of our image captioning model, we used BLEU scores at various n-gram levels (BLEU-1 to BLEU-4). These scores measure the overlap between the generated and reference captions, with higher scores indicating better performance.

A. BLEU-1: 0.468452

The model has a moderate agreement at the unigram level, capturing key words or concepts from the image.

B. BLEU-2: 0.279147

Performance drops slightly with bigrams, showing some difficulty in maintaining coherence in two-word sequences.

C. BLEU-3: 0.167669

The model struggles more with trigrams, indicating difficulty in generating contextually accurate multi-word combinations.

D. BLEU-4: 0.095209

The lowest score, highlighting challenges in generating longer, more coherent sequences of text.

IX. References

[1] Chu, Y., Yue, X., Yu, L., Sergei, M., Wang, Z. (2020) "Captioning images with proper descriptions automatically: AICRL model based on ResNet50 and LSTM with soft attention." Journal of Artificial Intelligence Research and Development, Harbin Engineering University, Harbin, China. Published under the Creative Commons Attribution License.

[2] Kamalam, S., Baby, S., Subiga, S., Yadhuvarshini, Y. (2024).

"Image caption generator using ResNet50 and LSTM." Proceedings of the 15th International Conference on Computing Communication and Networking Technologies (ICCCNT). DOI: 10.1109/ICCCNT61001.2024.10725128.

[3] Satti, S. K., Rajareddy, G. N. V., Maddula, P., Ravipati, N. V. V. (2023).

"Image caption generation using ResNet-50 and LSTM." Proceedings of the 2023 IEEE Silchar Subsection Conference (SILCON), Silchar, India, November 3-5. IEEE. DOI: 10.1109/SILCON59133.2023.10404600.

[4] Suresh, K. R., Jarapala, A., Sudeep, P. V. (2022) "Image captioning encoder–decoder models using CNN-RNN architectures: A comparative study." Circuits, Systems, and Signal Processing, 41, 5719–5742. DOI: 10.1007/s00034-022-02050-2.

[5] Santi, D., Ilham, A. A., Syafaruddin, Nurtanio, I. (2024) "Image caption generation through the integration of CNN-based residual network architectures and LSTM." Proceedings of the 2024 7th International Conference on Informatics and Computational Sciences (ICICoS), Semarang, Indonesia, July 17-18. IEEE. DOI: 10.1109/ICICoS62600.2024.10636926. [6] Singh, P., Kumar, C., Kumar, A. (2023) "Next-LSTM: A novel LSTM-based image captioning technique." Journal of Reliability: Theory and Applications, DOI: 10.1007/s13198-023-01956-7.

[7] Nehan, S., Chandrakanth, G., Lakshmi, A. V., Shambhavi, C. (2024) "Implementation of image caption generation using VGG16 and ResNet50." Proceedings of the 2024 Third International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), Hyderabad, India. IEEE. DOI: 10.1109/ICDCECE60827.2024.10548963.

[8] Liu, C., Zhao, R., Shi, Z. (2022) "Remote-sensing image captioning based on multilayer aggregated transformer." IEEE Geoscience and Remote Sensing Letters, 19, 6506605. DOI: 10.1109/LGRS.2022.3158123.

[9] Karthik, A. S., Karthik, M. H. S., Yashwanth, S., T, A. (2024) "Image captioning: Analyzing CNN-LSTM and Vision-GPT models." Proceedings of the 2024 IEEE 9th International Conference for Convergence in Technology (I2CT), Pune, India, April 5-7. IEEE. DOI: 10.1109/I2CT61223.2024.10543514.

[10] Rampal, H., Mohanty, A. (2020 "Efficient CNN-LSTM based image captioning using neural network compression." arXiv preprint arXiv:2012.09708v1, December 17.