Deep Learning-Based Image Caption Generator Using CNN and LSTM

Yash Aggarwal 2K22/CO/508, Yash Shival 2K22/CO/513
Department of Computer Science and Engineering
Delhi Technological University, India
Emails: yashaggarwal$_co22a8_23@dtu.ac.in, yashshival_co22a8_28@dtu.ac.in$

***Abstract*—In recent years, the task of automatically generating textual descriptions for images—known as image captioning—has gained significant attention in the field of artificial intelligence. This project presents a deep learning-based image caption generator that combines Convolutional Neural Networks (CNNs) for visual feature extraction with Long Short-Term Memory (LSTM) networks for sequence generation. We utilize a pre-trained CNN (VGG16) to extract image features and an LSTM decoder to generate grammatically meaningful and contextually relevant captions. The model is trained and evaluated on the Flickr8k dataset. Experimental results demonstrate the effectiveness of our approach in producing accurate and descriptive captions, showcasing the potential of neural architectures in bridging the gap between visual and textual modalities.**

***Index Terms*—Image Captioning, Convolutional Neural Network, Long Short-Term Memory, Deep Learning, Flickr8k**

I. Introduction

The ability to describe images in natural language is a fundamental problem that lies at the intersection of computer vision and natural language processing. Image captioning involves understanding the contents of an image and generating a coherent textual description that reflects the visual semantics. This has numerous applications, including aiding visually impaired individuals, improving image search engines, and enhancing human-computer interaction.

Traditional approaches to image captioning relied on template-based methods or retrieval techniques, which often lacked generalization and failed to capture complex visual relationships. With the advent of deep learning, especially the use of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), significant progress has been made in generating descriptive and grammatically correct captions.

In this project, we propose an image captioning model that leverages a pre-trained CNN (VGG16) to extract high-level visual features from images and an LSTM-based decoder to generate natural language descriptions. The model is trained and tested on the Flickr8k dataset, which consists of 8,000 images each annotated with five different captions. Our approach demonstrates how the fusion of visual and textual modalities using deep learning can result in meaningful and descriptive captions.

II. Related Work

Image captioning has seen rapid progress with the emergence of deep learning. Early approaches relied on template-based methods or retrieval-based systems which lacked flexibility and generalization.

The breakthrough came with the use of Convolutional Neural Networks (CNNs) for feature extraction and Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM), for sequence generation [6]. Vinyals et al. introduced the "Show and Tell" model, a pioneering end-to-end neural network framework that learns to generate captions from images.

Xu et al. further improved caption quality by incorporating attention mechanisms, allowing the model to focus on different parts of an image while generating each word [7]. More recent approaches have explored transformers, reinforcement learning, and large vision-language models like CLIP and BLIP to enhance accuracy and fluency.

Our work builds upon these foundational models by implementing a CNN-LSTM-based architecture using the Flickr8k dataset, demonstrating the feasibility of a compact yet effective image captioning pipeline.

III. Dataset Description

The Flickr8k dataset is a benchmark dataset for image captioning that contains 8,000 natural images collected from the Flickr website. Each image is manually annotated with five different captions that describe the visual content of the image. The captions are written in natural, grammatically correct English, making the dataset suitable for training language generation models.

The dataset is divided into three subsets: training (6,000 images), validation (1,000 images), and testing (1,000 images). The diversity in captions and the relatively small size of the dataset make it an ideal starting point for image captioning models that can be trained on limited resources.

IV. System Architecture

Our image captioning system follows a two-stage encoder-decoder architecture that combines visual feature extraction with sequence modeling:

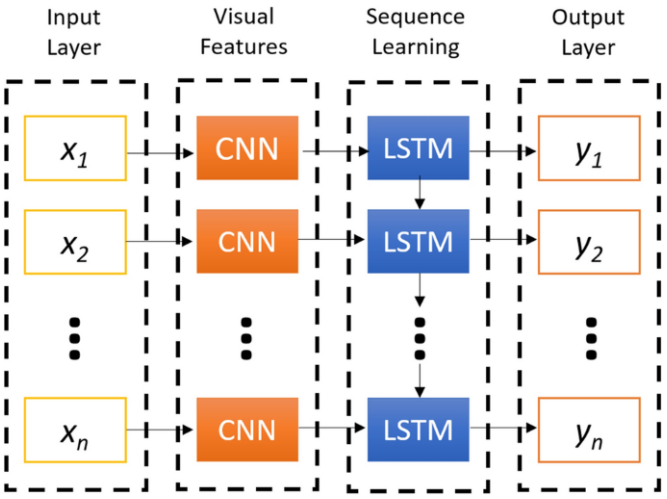

Fig. 1: Flowchart depicting the image caption generation architecture using CNN for feature extraction and LSTM for sequence learning.

- **Encoder (CNN - VGG16):** We use the pre-trained VGG16 model as a feature extractor. The final classification layers are removed, and the output from the last fully connected layer before classification (a 4096-dimensional vector) is used to represent the image. This feature vector encapsulates high-level visual features and serves as the initial context for the caption generation process.
- **Decoder (LSTM):** The decoder is a Long Short-Term Memory (LSTM) network that generates captions word-by-word. The input to the decoder at each time step is the previous word (embedded into a dense vector) and the image feature vector. The LSTM outputs a probability distribution over the vocabulary for the next word in the sequence.
- **Training Objective:** The model is trained to minimize categorical cross-entropy loss between the predicted and actual words in the caption. Teacher forcing is used during training, where the true previous word is fed to the LSTM rather than the predicted one.
- **Inference:** During inference, a greedy or beam search strategy is used to generate captions by selecting the most probable words sequentially until an end token is produced.

V. Results

The model was trained on the Flickr8k dataset using a pre-trained VGG16 encoder and an LSTM-based decoder. The captions generated by the model were evaluated qualitatively by observing the relevance and grammatical correctness of the generated descriptions.

Examples of generated captions:

- **Ground Truth:** A child is playing with a dog in the grass.
  Generated: A child plays with a dog in a field.

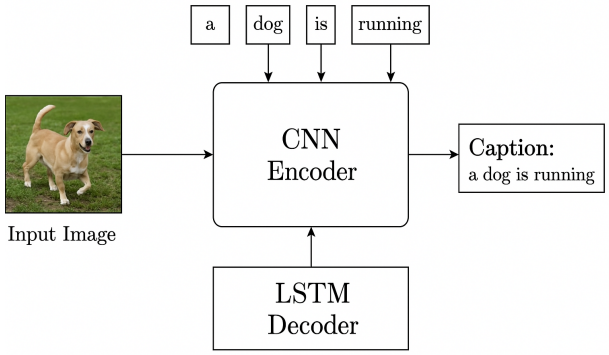

Fig. 2: Model Architecture for Image Captioning using CNN and LSTM.

(a) Predicted: *"man stands in front of very tall building"*

(b) Predicted: *"three dogs are standing in the grass and person is sitting next to them"*

Fig. 3: Generated captions by the model on test images from the Flickr8k dataset.

- **Ground Truth:** A group of people are hiking in the mountains.
  Generated: People are walking on a mountain trail.

To evaluate the performance of our image captioning model, we employed the BLEU (Bilingual Evaluation Understudy) metric, which is widely used for assessing the quality of machine-generated text against human-written reference sentences. BLEU measures the n-gram precision between the generated captions and the ground truth annotations.

We used a subset of 100 randomly selected test images from the Flickr8k dataset to calculate BLEU scores. The results are as follows:

TABLE I: BLEU Scores on Test Set (100 Images)

BLEU Metric	Score
BLEU-1	0.5672
BLEU-2	0.4117
BLEU-3	0.2852
BLEU-4	0.1861

These scores indicate that the model is able to accurately generate relevant words (BLEU-1), while also demonstrating reasonable fluency and grammatical correctness in longer n-

gram sequences (BLEU-2 to BLEU-4). A BLEU-1 score of over 0.56 suggests that the model is capturing key objects and actions effectively. As expected, the performance decreases for higher-order BLEU scores due to the inherent challenges in modeling complex grammar and sentence structure.

VI. Conclusion

This project demonstrates the capability of deep learning models to generate natural language descriptions of images by combining visual and language features. By using a CNN to extract visual features and an LSTM to generate sequences, our model effectively bridges the gap between computer vision and natural language processing.

While the model shows promising results on the Flickr8k dataset, future improvements could include:

- Incorporating attention mechanisms to focus on specific regions of the image while generating each word.
- Using larger datasets like MS COCO or Flickr30k to improve generalization.

This work lays a solid foundation for building more advanced multi-modal systems capable of understanding and describing visual content.

Acknowledgment

We would like to thank the faculty of Delhi Technological University for their continuous support and guidance throughout the development of this project. Special thanks to the Department of Computer Science and Engineering for providing the necessary resources and encouragement to pursue this research.

References

[1] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, "Show and Tell: A Neural Image Caption Generator," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3156–3164.

[2] A. Karpathy and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 664–676, April 2017.

[3] M. Hodosh, P. Young, and J. Hockenmaier, "Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics," in *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013.

[4] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[5] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," in *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[6] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. CVPR*, 2015, pp. 3156–3164.

[7] K. Xu et al., "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. ICML*, 2015, pp. 2048–2057.