2026 Volume: 02 Issue: 04 www.ijrpa.com ISSN 2456-9995 Case Study

International Journal Research Publication Analysis

Page: 01-06

A COMPARATIVE STUDY OF IMAGE CAPTION GENERATION USING CNN–LSTM AND TRANSFORMER-BASED MODELS

***Dr. Ramya B. N., Vijaya Krishna, Samarth**

Department of Computer Science and Engineering Jyothy Institute of Technology, Bengaluru, India.

Article Received: 21 March 2026
Article Revised: 11 April 2026
Published on: 01 May2026

***Corresponding Author: Dr. Ramya B. N.**
Department of Computer Science and Engineering Jyothy Institute of Technology, Bengaluru, India.

DOI: https://doi-doi.org/101555/ijrpa.5727

ABSTRACT

Image caption generation is a fundamental problem in artificial intelligence that combines computer vision and natural language processing to generate textual descriptions for images. The task requires a model to identify objects, understand their relationships, and express this information in coherent natural language. This paper presents a detailed experimental study of image caption generation using CNN–LSTM architectures and Transformer-based models, including Vision Transformer with GPT-2 and BLIP-based captioning systems. The models are evaluated on real-world images using quantitative metrics such as BLEU score and extensive qualitative analysis. Experimental results demonstrate that Transformer-based models, particularly BLIP-Large, produce more descriptive and context-aware captions compared to traditional CNN–LSTM approaches. The study highlights the strengths, limitations, and practical trade-offs of each model.

***INDEX TERMS*:** Image Caption Generation, CNN–LSTM, Trans-former, Vision Transformer, BLIP, Deep Learning.

INTRODUCTION

The rapid growth of visual content on the internet has created a strong demand for automated systems capable of un-derstanding and describing images. Image caption generation aims to bridge the semantic gap between visual perception and natural language understanding by producing grammatically correct and semantically meaningful descriptions for images. This task has significant applications in assistive technologies for visually impaired individuals, image retrieval systems, content moderation, and human–computer interaction.

Traditional computer vision systems focused primarily on image classification and object detection. However, generating captions requires a deeper level of understanding, including object attributes, spatial relationships, and contextual informa-tion. Early image captioning approaches relied on template-based or retrieval-based methods, which lacked generalization and scalability.

With the advent of deep learning, encoder–decoder archi-tectures became the dominant paradigm. In these systems, an encoder extracts visual features from an image, and a de-coder generates a natural language description based on these features. More recently, Transformer-based architectures have achieved state-of-the-art performance by leveraging attention mechanisms to model long-range dependencies. This paper presents a comprehensive comparison of CNN–LSTM and Transformer-based image captioning models.

RELATED WORK

Early image captioning research focused on rule-based systems that generated captions using predefined sentence structures. While computationally efficient, these approaches failed to generalize to complex real-world images. Retrieval-based methods later emerged, selecting captions from visually similar images; however, their performance was limited by the diversity of available captions.

The introduction of deep learning enabled end-to-end encoder–decoder frameworks. Vinyals et al. proposed the CNN–LSTM architecture, where convolutional neural net-works extract visual features and LSTMs generate captions sequentially. Xu et al. introduced attention mechanisms that allow the model to focus on relevant image regions during caption generation.

Transformer architectures eliminated recurrence entirely and relied on self-attention mechanisms. Vision Transformer (ViT) encoders model global image context by dividing images into patches, while Transformer-based decoders generate captions in an autoregressive manner. Recent vision–language pretrain-ing models such as BLIP have further improved caption quality by learning joint visual and textual representations.

PROBLEM FORMULATION

Given an input image I, the goal of image caption genera-tion is to produce a sequence of words $S = \{w_1, w_2, \ldots, w_T\}$ that maximizes the conditional probability:

$$P(S|I) = \prod_{t=1}^{T} P(w_t|I, w_1, \ldots, w_{t-1}) \quad (1)$$

In CNN–LSTM models, image features extracted by the CNN encoder initialize the hidden state of the LSTM decoder. In Transformer-based models, visual features are integrated using attention mechanisms, enabling the model to capture global dependencies between image regions and language tokens.

SYSTEM ARCHITECTURE

The overall architecture of the image caption generation system is shown in Fig. 1. The system follows an encoder–decoder framework supporting multiple captioning pipelines, including CNN–LSTM, ViT–GPT2, and BLIP-based models.

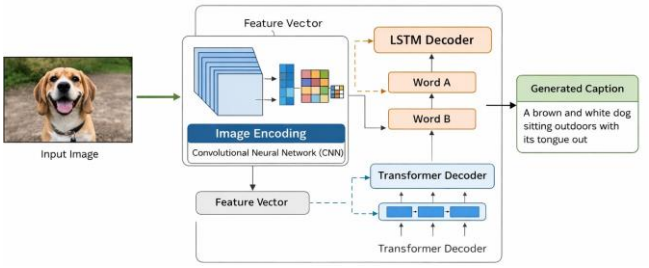

Fig. 1. System architecture of image caption generation using CNN/Trans-former encoder and LSTM/Transformer decoder.

DATASET DESCRIPTION

Experiments were conducted using a collection of real-world images obtained from publicly available sources. The dataset includes images containing animals, humans, indoor scenes, outdoor scenes, and complex backgrounds. Images were resized and normalized before being passed to the encoder. Captions were manually verified for qualitative eval-uation.

METHODOLOGY

A. CNN–LSTM Model

The CNN–LSTM model uses a convolutional neural net-work to extract visual features, which are then passed to an LSTM decoder. The decoder generates captions sequentially, predicting the next word based on the visual context and previously generated words. Although effective, this approach struggles with long-range dependencies.

B. Transformer-Based Models

Transformer-based models replace recurrent connections with self-attention mechanisms. Vision Transformer encoders divide images into patches and learn global representations, while decoders such as GPT-2 generate captions autoregres-sively. BLIP models further

improve performance through vision–language pretraining.

EVALUATION METRICS

Model performance was evaluated using BLEU score, which measures n-gram overlap between generated captions and reference captions. In addition to quantitative metrics, qual-itative analysis was performed to assess descriptive richness, grammatical correctness, and contextual accuracy.

RESULTS AND ANALYSIS

A. Quantitative Results

B. Qualitative Analysis

Transformer-based models generate longer and more de-scriptive captions compared to CNN–LSTM models. BLIP-Large demonstrates improved understanding of object at-tributes and actions, while CNN–LSTM models often produce generic descriptions.

TABLE I BLEU SCORE COMPARISON.

Model	BLEU Score
CNN–LSTM	0.31
ViT–GPT2	0.38
BLIP–Base	0.40
BLIP–Large	**0.42**

LIMITATIONS

Despite improved performance, the models occasionally misinterpret object counts or colors in visually ambiguous images. Computational complexity and memory requirements are higher for Transformer-based models.

IMPLEMENTATION DETAILS

The following code snippet demonstrates the implemen-tation of the BLIP-Large model used for image caption generation. The model is loaded using the Hugging Face Transformers library and generates captions for input images.

CONCLUSION

This paper presented a detailed comparative study of image caption generation using CNN–LSTM and Transformer-based models. Experimental results show that Transformer-based architectures, particularly BLIP-Large, outperform traditional CNN–LSTM approaches in terms of caption quality and contextual understanding. Future work may explore domain-

specific fine-tuning and multilingual caption generation.

REFERENCES

1. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: A Neural Image Caption Generator," in Proc. CVPR, 2015.
2. K. Xu et al., "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," IEEE TPAMI, 2015.
3. Karpathy and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," in Proc. CVPR, 2015.
4. Vaswani et al., "Attention Is All You Need," in Proc. NeurIPS, 2017.
5. J. Donahue et al., "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," in Proc. CVPR, 2015.
6. P. Anderson et al., "Bottom-Up and Top-Down Attention for Image Captioning and VQA," in Proc. CVPR, 2018.
7. M. Cornia et al., "Meshed-Memory Transformer for Image Captioning," in Proc. CVPR, 2020.
8. J. Lu et al., "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Rep-resentations," in Proc. NeurIPS, 2019.
9. H. Tan and M. Bansal, "LXMERT: Learning Cross-Modality Encoder Representations," in Proc. EMNLP, 2019.
10. Y. Chen et al., "UNITER: Learning Universal Image-Text Representa-tions," in Proc. ECCV, 2020.
11. J. Li et al., "BLIP: Bootstrapping Language-Image Pre-training," in Proc. ICML, 2022.
12. R. Bernardi et al., "Automatic Description Generation from Images: A Survey," JAIR, 2016.
13. P. Young et al., "From Image Descriptions to Visual Denotations," IEEE Signal Processing Magazine, 2017.
14. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," in Proc. ICML, 2021.
15. X. Huang et al., "Attention on Attention for Image Captioning," in Proc. ICCV, 2019.
16. M. Johnson et al., "Google's Multilingual Neural Machine Translation System," in Proc. ACL, 2017.
17. L. Zhou et al., "Unified Vision-Language Pre-training for Image Cap-tioning and VQA," arXiv:1909.11059, 2019.

18. J. Devlin et al., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proc. NAACL, 2019.
19. Elliott and F. Keller, “Image Description Using Visual Dependency Representations,” in Proc. ACL, 2014.
20. T. Yao et al., “Boosting Image Captioning with Attributes,” in Proc. ICCV, 2017.
21. S. Rennie et al., “Self-Critical Sequence Training for Image Captioning,” in Proc. CVPR, 2017.
22. J. Aneja et al., “Convolutional Image Captioning,” in Proc. CVPR, 2018.
23. M. Hossain et al., “A Comprehensive Survey of Deep Learning for Image Captioning,” ACM Computing Surveys, 2019.
24. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” in Proc. ICLR, 2021.
25. T. Chen et al., “Exploring Simple Siamese Representation Learning,” in Proc. CVPR, 2021.
26. Z. Dai et al., “Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context,” in Proc. ACL, 2019.
27. Y. Liu et al., “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” arXiv:1907.11692, 2019.
28. Bahdanau et al., “Neural Machine Translation by Jointly Learning to Align and Translate,” in Proc. ICLR, 2015.
29. S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” Neural Computation, 1997.
30. Sutskever et al., “Sequence to Sequence Learning with Neural Net-works,” in Proc. NeurIPS, 2014.