International Journal of Creative and Open Research in Engineering and Management
ISSN: **3108-1754 (Online)**
Volume 02 Issue 05 May-2026 | Impact Factor: 3.5

An Automated Image Captioning Framework Based on Vision Transformers and LSTM Networks

Dr M Praneesh[1] Dr.D. Napoleon[2]

[1]Assistant Professor, Department of Computer Science with Data Analytics / Sri Ramakrishna College of Arts & Science / Bharathiar University, Coimbatore, India

[1]Associate Professor, Department of Computer Science, Bharathiar University, Coimbatore, India

Corresponding Author Email: raja.praneesh@gmail.com | ORCID: https://orcid.org/0000-0003-3691-1343

How to Cite this Article:

Praneesh, D. M. & Napoleon, D. (2026). An Automated Image Captioning Framework Based on Vision Transformers and LSTM Networks. International Journal of Creative and Open Research in Engineering and Management, 02(05). https://doi.org/10.55041/ijcope.v2i5.850

https://doi.org/10.55041/ijcope.v2i5.850

Abstract—

Image captioning is an important research area in artificial intelligence that integrates computer vision and natural language processing (NLP) to automatically generate descriptive textual interpretations of images. Conventional image captioning systems typically employ Convolutional Neural Networks (CNNs) for visual feature extraction and Long Short-Term Memory (LSTM) networks for generating sequential text descriptions. Although effective, CNN-based approaches may have limitations in capturing global contextual relationships within images.

This research introduces an improved image captioning framework that utilizes Vision Transformers (ViTs) as the feature extraction backbone instead of traditional CNN architectures. By leveraging self-attention mechanisms, Vision Transformers can effectively model long-range dependencies and capture comprehensive contextual information from visual data. The extracted image representations are subsequently provided to an LSTM network, which generates coherent and meaningful captions in a sequential manner.

The proposed model is evaluated using widely accepted image captioning performance metrics, including BLEU and METEOR scores. Experimental findings indicate that the Vision Transformer-based approach produces more accurate, descriptive, and context-aware captions compared to conventional CNN-LSTM models. The enhanced caption generation capability of the proposed framework makes it suitable for various real-world applications, including assistive technologies for visually impaired individuals, automated image annotation, content management systems, and intelligent multimedia retrieval.

Keywords— Image Captioning, Vision Transformer (ViT), Long Short-Term Memory (LSTM), Natural Language Processing (NLP), Computer Vision, Text Generation, BLEU Score

I. INTRODUCTION

Image captioning is a prominent task in artificial intelligence that aims to automatically generate descriptive text for images by combining techniques from computer vision and natural language processing (NLP). This technology has numerous practical applications, including assistive systems for visually impaired users, automated content creation, image retrieval, and multimedia management. The objective is not only to recognize objects within an image but also to understand their relationships and generate meaningful textual descriptions.

Conventional image captioning frameworks typically employ Convolutional Neural Networks (CNNs) to extract visual features from images and Long Short-Term Memory (LSTM) networks to generate captions sequentially. While these methods have achieved considerable success, CNN-based feature extractors often face challenges in capturing global contextual information and long-range dependencies within complex visual scenes. As a result, the generated captions may lack detailed contextual understanding and semantic richness.

To address these limitations, this study proposes an image captioning framework that utilizes Vision Transformers (ViTs) for visual feature extraction. Unlike CNNs, Vision Transformers process images as sequences of image patches and leverage self-attention mechanisms to model global relationships across the entire image. This capability enables the extraction of more informative and context-aware visual representations. The extracted features are subsequently provided to an LSTM network, which generates coherent and semantically meaningful captions.

The proposed model is trained and evaluated using the Flickr30k dataset, which contains 30,000 images accompanied by multiple human-generated captions. The availability of diverse image-caption pairs allows the model to learn a wide range of visual concepts and linguistic patterns. Performance evaluation is conducted using established image captioning metrics, including BLEU, METEOR, and CIDEr, which assess the quality, relevance, and fluency of the generated captions.

Experimental results indicate that the Vision Transformer-based approach significantly improves caption generation performance compared to traditional CNN-LSTM architectures. By effectively capturing global contextual information and complex visual relationships, the proposed framework produces more accurate, descriptive, and contextually relevant captions, making it suitable for advanced image understanding and real-world multimedia applications.

II. LITERATURE REVIEW

Existing image captioning research has largely been based on encoder–decoder architectures that combine Convolutional Neural Networks (CNNs) for visual feature extraction with Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks for caption generation. Notable models have demonstrated considerable success in producing image descriptions by learning visual representations and incorporating attention mechanisms to focus on relevant image regions. These approaches have significantly advanced the field of automatic image captioning.

Despite their effectiveness, CNN-based feature extraction methods exhibit certain limitations. Their localized receptive fields may restrict the ability to capture long-range relationships between objects and regions within an image. Additionally, the hierarchical nature of CNNs can result in the loss of important global contextual information, which is essential for accurately describing complex scenes. Consequently, generated captions may sometimes lack semantic richness and fail to fully represent intricate visual details.

To address these challenges, the proposed research adopts Vision Transformers (ViTs) as the image feature extraction component. Unlike CNNs, Vision Transformers divide an image into patches and utilize self-attention mechanisms to model interactions among all image regions simultaneously. This enables the network to capture both local and global contextual information more effectively, leading to a deeper understanding of visual content.

The extracted visual features are subsequently processed by an LSTM-based decoder to generate natural language descriptions. By combining the powerful representation learning capability of Vision Transformers with the sequence modeling strength of LSTMs, the proposed framework aims

to produce captions that are more accurate, contextually relevant, and linguistically coherent. This hybrid approach is expected to overcome several limitations of conventional CNN-based image captioning systems and improve the overall quality and diversity of generated image descriptions.

III. METHODOLOGY

To enhance the accuracy and contextual understanding of image captioning, we propose a Vision Transformer (ViT) + LSTM-based model that overcomes the limitations of traditional CNN-based approaches. Unlike CNNs, which extract local features through convolutional operations, ViT employs a self-attention mechanism to capture global dependencies across an image, leading to a more comprehensive representation of visual information. This enables the model to better understand complex scenes, object relationships, and fine-grained details, resulting in more descriptive and contextually rich captions.

In our approach, ViT extracts high-level visual features from images, which are then fed into an LSTM network to generate meaningful captions in natural language. The sequential nature of LSTM maintains coherence and fluency in generated sentences. Additionally, we incorporate attention mechanisms to enhance the alignment between visual features and textual output, ensuring that captions focus on the most relevant aspects of an image.

By combining the global feature extraction capabilities of ViT with the sequential text generation of LSTM, our model aims to achieve higher accuracy, fluency, and contextual relevance compared to traditional CNN-LSTM frameworks.

3.1 Vision Transformer (ViT)

Vision Transformer (ViT) is a deep learning model that applies transformer-based self-attention mechanisms to image analysis. Originally inspired by transformer architectures developed for natural language processing tasks, ViT adapts these concepts to effectively process visual information. Rather than analyzing an image as a continuous grid of pixels, the model divides the image into a sequence of smaller, fixed-size patches, such as 16 × 16 pixels.

Each image patch is transformed into a vector representation through a linear embedding process. Positional information is then incorporated into these embeddings to preserve the spatial arrangement of the patches within the image. The resulting sequence of embedded patches is passed through a transformer encoder composed of multiple self-attention and feedforward layers.

Through the self-attention mechanism, the model learns relationships among different image regions and captures both local and global visual dependencies. This enables the network to simultaneously focus on multiple areas of an image and extract rich contextual information. The final encoded representation provides a comprehensive understanding of the visual content, making Vision Transformers particularly effective for tasks such as image captioning, where accurate interpretation of complex scenes is essential for generating meaningful textual descriptions.

3.2 Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) networks play a vital role in the caption generation stage by effectively modeling sequential text data. The process begins with the extraction of visual features using the Vision Transformer (ViT), which provides a rich representation of the image content. The corresponding image captions are then pre-processed and converted into sequences of words or subword units through tokenization.

These tokens are transformed into dense numerical embeddings that capture semantic and syntactic relationships between words. Using the extracted visual features as contextual input, the LSTM decoder generates captions one word at a time. The sequence generation process is initiated with a special <start> token, and at each step, the model predicts the most probable next word through a probability distribution generated by a SoftMax layer.

The caption generation continues iteratively until a predefined <end> token is produced, indicating the completion of the sentence. This sequential prediction mechanism enables the model to generate grammatically coherent and semantically meaningful descriptions. By integrating the contextual feature extraction capability of Vision

Transformers with the sequence-learning strength of LSTM networks, the proposed framework effectively connects visual understanding with natural language generation, resulting in accurate, fluent, and contextually appropriate image captions.

3.3 Image Captioning Model Using ViT + LSTM

Proposed Algorithm: Image Caption Generation Using Vision Transformer (ViT) and LSTM
Input: Image Dataset (I) with corresponding captions (C)
Output: Generated image caption (G_C)

Step 1: Import Required Libraries

1. Import PyTorch and Torch vision libraries for image processing and Vision Transformer implementation.
2. Import Hugging Face Transformers library for loading the pre-trained ViT model.
3. Import Pillow (PIL) for image loading and preprocessing.
4. Import NumPy for numerical computations.
5. Import TensorFlow/Keras libraries for tokenization, sequence padding, and LSTM model development.

Step 2: Image Feature Extraction Using Vision Transformer

1. Load input image (I) using PIL.
2. Convert the image into RGB format.
3. Preprocess the image using the ViT Feature Extractor.
4. Pass the preprocessed image to the pre-trained Vision Transformer model.
5. Extract the hidden feature representations from the final transformer layer.
6. Compute the average of the extracted feature vectors.
7. Store the resulting fixed-length feature representation as (F_I).

Step 3: Caption Preprocessing

1. Collect all image captions from the dataset.
2. Apply text cleaning and preprocessing operations.
3. Initialize the Keras Tokenizer.
4. Build the vocabulary from the caption corpus.
5. Convert each caption into a sequence of integer tokens.
6. Determine the maximum caption length.
7. Apply sequence padding to ensure uniform sequence length.

Step 4: Construct LSTM-Based Caption Generator

1. Create a Sequential neural network model.
2. Add an Embedding layer to transform word indices into dense vector representations.
3. Add the first LSTM layer with sequence return enabled.
4. Add a second LSTM layer to learn deeper contextual dependencies.
5. Add a Dense layer with Softmax activation.
6. Configure the output layer to predict the next word from the vocabulary.

Step 5: Model Compilation

1. Set the loss function as Categorical Cross-Entropy.
2. Select Adam as the optimization algorithm.
3. Compile the model for training.

Step 6: Model Training

1. For each image-caption pair:
   - Extract image features (F_I) using the ViT model.
   - Convert captions into tokenized sequences.
   - Generate input-output training sequences.
2. Train the LSTM network using:
   - Image feature vectors (F_I)
   - Corresponding caption sequences
3. Update model parameters until convergence or the maximum number of epochs is reached.

Step 7: Caption Generation (Inference Phase)

1. Input a new image (I_{new}).
2. Extract image features (F_{new}) using the trained ViT model.
3. Initialize the caption sequence with the <start> token.
4. Repeat:
   - Feed the current sequence and image features into the trained LSTM model.
   - Predict the next word using Softmax probabilities.
   - Append the predicted word to the sequence.
5. Continue the process until:
   - The <end> token is generated, or
   - The maximum caption length is reached.

Step 8: Output

1. Remove special tokens (<start>, <end>).
2. Concatenate the generated words.
3. Return the final caption (G_C).

The proposed algorithm combines the powerful visual representation capability of Vision Transformers with the sequence-learning strength of LSTM networks to generate accurate, fluent, and context-aware image captions.

IV. RESULTS AND DISCUSSION

The effectiveness of an image captioning model is determined using evaluation metrics that measure the quality, accuracy, and linguistic coherence of the generated captions. Among the most widely adopted metrics are BLEU, METEOR, and CIDEr, each providing a distinct perspective on the model's caption generation performance. These metrics compare the generated captions with reference captions and help assess how well the model captures the semantic meaning and contextual relevance of the image content.

1. **BLEU (Bilingual Evaluation Understudy):** BLEU measures the n-gram overlap between generated and reference captions.

$$BLEU=BP\times\exp(n=1\sum Nwn\log Pn)$$

2. **METEOR (Metric for Evaluation of Translation with Explicit Ordering*):*** METEOR considers synonyms, stemming, and word order, making it more aligned with human judgment.

$$METEOR=Fmean\times(1-Penalty)$$

3. **CIDEr (Consensus-based Image Description Evaluation):** CIDEr emphasizes consensus by comparing generated captions with multiple human-annotated references.

$$CIDEr=N1n=1\sum Nwn\times TF\text{-}IDF\ Score$$

For the sample caption, *"A man is riding a horse on the beach,"* the evaluation metrics provide a comprehensive assessment of the quality and relevance of the generated description. The BLEU score falls within the range of 30–35, indicating a reasonable degree of similarity between the generated caption and the reference captions based on n-gram matching. This suggests that the model successfully captures key elements of the image description.

The METEOR score of 0.44 (44%) reflects a stronger alignment with human-generated captions by considering factors such as synonym matching, stemming, and word ordering. As a result, METEOR offers a more nuanced evaluation of semantic similarity and linguistic quality than BLEU alone.

Additionally, the CIDEr score of 0.90 (90%) indicates a high level of agreement between the generated caption and human annotations. By incorporating TF-IDF weighting, CIDEr emphasizes the importance of informative and distinctive words, thereby providing a robust measure of descriptive accuracy.

Collectively, these evaluation results demonstrate the capability of the proposed ViT-LSTM framework to generate captions that are both contextually meaningful and linguistically coherent, closely resembling descriptions produced by human annotators.

Figure -1 Sample Caption

Figure -2 Sample Caption

Overall, the experimental findings indicate that the proposed ViT-LSTM framework achieves superior performance compared to conventional CNN-based image captioning models across various evaluation scenarios. The enhanced feature extraction capability of Vision Transformers contributes to the generation of more accurate and contextually meaningful captions. Nevertheless, there remains potential for further improvement. Future research may focus on incorporating advanced attention mechanisms and performing domain-specific fine-tuning to enhance the model's understanding of visual content and improve caption quality, accuracy, and semantic relevance.

Figure – 3 Output six children with their hands on their chins

Figure – 4 Output a skater leaps in the air on a city

Figure – 4 An image of a cat

V. CONCLUSION

This study demonstrates the effectiveness of combining Vision Transformers (ViT) with Long Short-Term Memory (LSTM) networks for automated image caption generation. The use of ViT enables strong global attention-based visual feature extraction, while the LSTM component supports sequential language modeling for generating coherent textual descriptions. This integrated framework shows improved performance over traditional CNN-based approaches in terms of contextual understanding, caption accuracy, and linguistic fluency.

The model's evaluation using standard metrics such as BLEU, METEOR, and CIDEr confirms its ability to produce detailed and semantically relevant image captions. These results highlight the advantage of leveraging transformer-based visual encoders along with recurrent language models for improved image-to-text generation.

For future enhancements, several directions can be considered. Incorporating large pre-trained language models such as GPT or BERT may further improve the fluency and grammatical quality of generated captions. In addition, developing fully transformer-based hybrid architectures that combine Vision Transformers with advanced language models could enhance overall system performance. Further improvements may also be achieved by training on larger and more diverse datasets or by applying reinforcement learning techniques to refine caption quality. Finally, optimizing the model for real-time deployment using edge computing techniques could extend its applicability to practical domains such as assistive technologies, automated image annotation, and intelligent content generation.

REFERENCES

[1] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, "Show and tell: A neural image caption generator," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3156-3164, doi: 10.1109/CVPR.2015.7298935.

[2] P. Mathur, A. Gill, A. Yadav, A. Mishra and N. K. Bansode, "Camera2Caption: A real-time image caption generator," 2017 International Conference on Computational Intelligence in Data

Science(ICCIDS), 2017, pp. 1-6, doi:10.1109/ICCIDS.2017.8272660.

[3] Wang, Haoran, Yue Zhang, and Xiaosheng Yu. "An overview of image caption generation methods." Computational intelligence and neuroscience 2020 (2020).

[4] P. Anderson et al., "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 6077-6086, doi: 10.1109/CVPR.2018.00636.

[5] Pedersoli M, Lucas T, Schmid C, Verbeek J (2017) Areas of attention for image captioning. In: 2017 IEEE international conference on computer vision (ICCV), pp 12511259

[6] Preksha Khant, Vishal Deshmukh, Aishwarya Kude, Prachi Kiraula, Image Caption Generator using CNN-LSTM International Research Journal of Engineering and Technology (IRJET), 2021

[7] Tanti M, Gatt A, Camilleri KP. What is the role of recurrent neural networks (rnns) in an image caption generator?. arXiv preprint arXiv:1708.02043. 2017 Aug 7.

[8] Chunseong Park C, Kim B, Kim G. Attend to you: Personalized image captioning with context sequence memory networks. InProceedings of the IEEE conference on computer vision and pattern recognition 2017 (pp. 895-903).

[9] Yang, Yang & Zhou, Jie & Ai, Jiangbo & Bin, Yi & Hanjalic, Alan & Shen, Heng & Ji, Yanli. (2018). Video Captioning by Adversarial LSTM. IEEE Transactions on Image Processing. 27. 1-1.10.1109/TIP.2018.2855422.

[10] Sherstinsky A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena. 2020 Mar 1;404:132306.

[11] You, Quanzeng, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. "Image captioning with semantic attention." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4651-4659. 2016.

[12] Rennie, Steven J., Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. "Self-critical sequence training for image captioning." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7008-7024. 2017.

[13] Aneja, Jyoti, Aditya Deshpande, and Alexander G. Schwing. "Convolutional image captioning." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5561-5570. 2018.

[14] Feng, Yang, Lin Ma, Wei Liu, and Jiebo Luo. "Unsupervised image captioning." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4125-4134. 2019.

[15] Cui, Yin, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie. "Learning to evaluate image captioning." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5804-5812. 2018.

[16] Huang, L., Wang, W., Chen, J. and Wei, X.Y., 2019. Attention on attention for image captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 4634-4643).

[17] Yao, Ting, Yingwei Pan, Yehao Li, and Tao Mei. "Exploring visual relationship for image captioning." In Proceedings of the European conference on computer vision (ECCV), pp. 684-699. 2018.

[18] Xu Yang, Kaihua Tang, Hanwang Zhang, Jianfei Cai. Auto-Encoding Scene Graphs for Image Captioning. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pages 10685-10694, Computer Vision Foundation IEEE, 2019. [doi]

[19] Wang, C., Yang, H. and Meinel, C., 2018. Image Captioning with Deep Bidirectional LSTMs and Multi-Task Learning. ACM Transactions on Multimedia Computing, Communications, and Applications, 14(2s), pp.1-20.

[20] Ren Z, Wang X, Zhang N, Lv X, Li LJ. Deep reinforcement learning-based image captioning with embedding reward. InProceedings of the IEEE conference on computer vision and pattern recognition 2017 (pp.290-298).

[21] M. P. R, M. Anu and D. S, "Building A Voice Based Image Caption Generator with Deep Learning," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2021, pp. 943-948, doi:10.1109/ICICCS51141.2021.9432091.

[22] S. Han and H. Choi, "Domain-Specific Image Caption Generator with Semantic Ontology," 2020 IEEE International Conference on Big Data and Smart Computing (BigComp), Busan, Korea (South), 2020, pp. 526- 530, doi:10.1109/BigComp48618.2020.00-12.

[23] C. Amritkar and V. Jabade, "Image Caption Generation Using Deep Learning Technique," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-4, doi: 10.1109/ICCUBEA.2018.8697360.

[24] N. K. Kumar, D. Vigneswari, A. Mohan, K. Laxman and J. Yuvaraj, "Detection and Recognition of Objects in Image Caption Generator System: A Deep Learning Approach," 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), Coimbatore, India, 2019, pp. 107-109,doi: 10.1109/ICACCS.2019.8728516.

[25] V. Kesavan, V. Muley and M. Kolhekar, "Deep Learning based Automatic Image Caption Generation," 2019 Global Conference for Advancement in Technology (GCAT), BENGALURU, India, 2019, pp. 1-6, doi: 10.1109/GCAT47503.2019.8978293.