Image Captioning using Deep Neural Network

Shashi Kant Maurya [1], Vivek Negi [2], Ritesh Kumar Shikarwar [3]
1Assistant Professor,
1, 2, 3 Department Of Computer Science And Engineering, MGM College Of Engineering And Technology, Noida
1. shashikant@coet.in , 2. 2000950100076@coet.in , 3. 2000950100058@coet.in

Abstract
This paper provides a comprehensive exploration of Vision Transformer (ViTs) in the context of image captioning. Through detailed discussions, we dissected the self-attention mechanisms inherent to ViTs, elucidating their mathematical foundations and practical implications. The review delved into their integration into image captioning pipelines, emphasizing the encoder-decoder structure, training strategies, and performance comparisons. Addressing challenges of computational complexity and data efficiency, innovative solutions were explored, including model distillation and unsupervised learning. ViTs, with their ability to capture long-range dependencies and scalability, emerges as pivotal tools in reshaping image captioning paradigms. This study culminates in a forward-looking perspective, envisioning the future of ViTs as efficient, specialized, and versatile solutions, revolutionizing the intersection of computer vision and natural language processing.

Keywords: Convolutional Neural Network, Recurrent Neural Network, Self Attention Mechanism, Transformer, Encoder-decoder Techniques, Vision Transformer, Feed Forward Neural Network, Patch Embedding, Multi-Head Attention, Transfer Learning

1. Introduction
The innovative technique of image captioning emerged in the past couple of year as a result of the fusion of natural language processing with computer vision. The technique of image captioning to generate written summaries of photos has tremendous potential uses such as enhancing content-based image retrieval systems and assistive technology for the visually impaired. Recurrent neural networks (RNNs) and convolutional neural networks (CNNs) were often used in sequential text synthesis and feature extraction in conventional picture captioning techniques however a paradigm shift has occurred with the advent of vision transformers (ViTs), a novel brain design based on the self-attention process. Transformers were first designed for sequential information expressed in natural language but their use into computer vision has sparked revolutionary improvements in the discipline vision reliant on self-attention.

Transformers were initially developed for consecutive data in natural language, but their application into image recognition has led to previously unthinkable advances. Vision Transformers have shown remarkable success in a number of visual tasks, owing to their capacity to extract long-range relationships from input via self-attention. Vision Transformers are showing promising results in image captioning, where their ability to describe complex relationships in both pictures and words at the same time is one of its most interesting uses.
Reviewing the theoretical foundations, architectural subtleties, and experimental advances, this study explores the convergence of picture captioning with Vision Transformers. By means of a thorough examination of current literature and approaches, this study seeks to offer a thorough grasp of how Vision Transformers have transformed the field of picture captioning. Through an analysis of the difficulties encountered, creative solutions developed, and future directions that show great promise, this study provides essential information for scholars, practitioners, and enthusiasts exploring the fascinating field of AI-driven picture interpretation

2. Background and Related Work

Historically, image captioning techniques primarily relied on a two-step process involving feature extraction [1] in images and sequence generation. Convolutional Neural Networks (CNNs) were employed for extracting visual features from images. These features were then fed into Recurrent Neural Networks (RNNs) or variants like Long Short-Term Memory (LSTM) networks, which generated sequential descriptions. While these methods produced decent results, they often struggled with capturing complex relationships between objects and their contextual meanings in the image.[2]

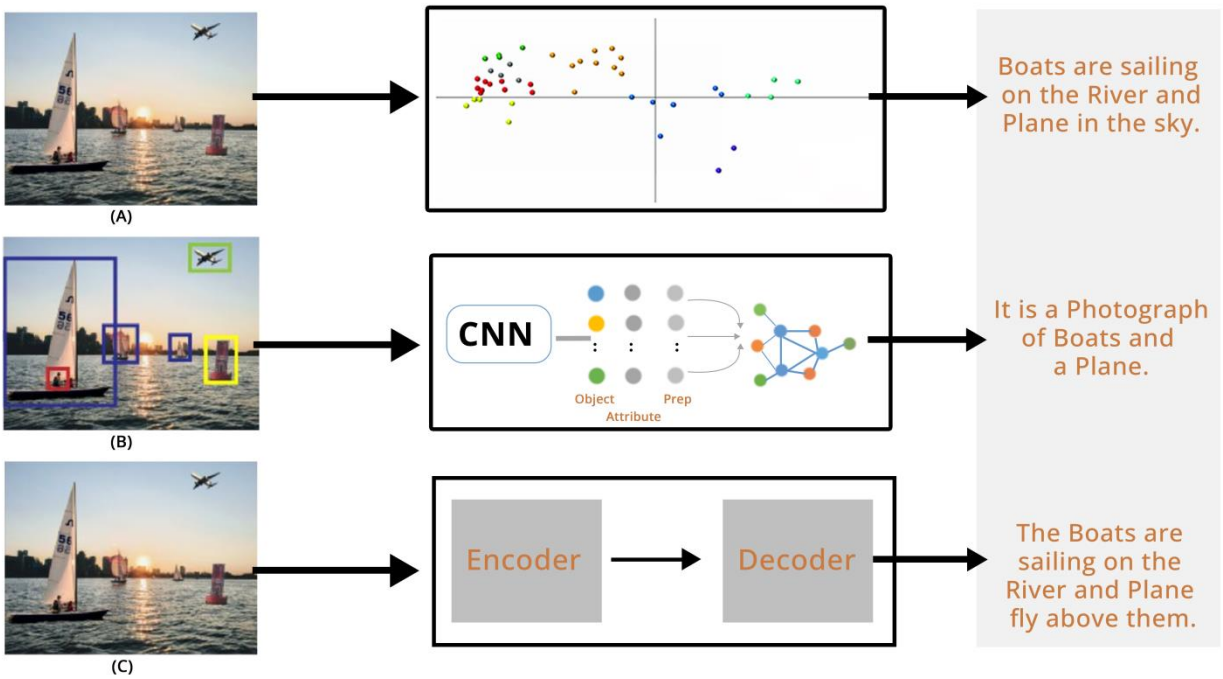

Figure 1. Comparison between Different Methods in Image Captioning, (A) CNN-CNN Based Model, (B) CNN-RNN Based Model, (C) Transformer Based Model

Enhancing picture captioning has been made possible in large part by recent developments in attention processes. The alignment between the visual and written aspects is improved by attention methods[2] that allow the model to narrow in specific parts of the picture. The quality of produced captions has greatly increased because the techniques like Hard Attention and Soft Attention. Furthermore, pre-trained CNNs like ResNet[4] and VGGNet[3] are now commonly included as feature extractors, increasing the accuracy of output captions.

With these developments, rulebased models gave way to data-driven, deep learning techniques, greatly enhancing the descriptive power of picture captions.[2]

In the middle of these developments, a revolutionary architecture known as Vision Transformers (ViTs) appeared. Based on the success of transformers in natural language processing, ViTs use a self-attention method that enables the model to determine the relative importance of various picture areas for the purpose of creating captions. ViTs accurately capture complex spatial relationships because, in contrast to conventional CNN-based techniques, they process the entire picture as a series of patches. In addition to allowing ViTs to scale well with the quantity of the input, this paradigm change from grid-based CNNs to sequence-based transformers improves their capacity to extract global contextual information from the pictures. As a result, ViTs have demonstrated impressive promise in a range of visual tasks such as recognition of objects, picture captioning, and image categorization.[2]

The basis of the changing picture captioning environment is this fusion of conventional methods the latest developments in attention processes and the groundbreaking possibilities of vision transformers in the next sections we go into further detail about the specifics of vision transformers looking at its design benefits and issues in producing captions for photos that are both descriptive and pertinent to the context.

Table 1. Previous Findings

Paper Title	Year	Key Findings
ViLT: Vision-and-Language Transformer for Image Captioning[6]	2021	ViLT was introduced as a vision-and-language transformer for picture captioning. On COCO and Flickr30k. ViLT performed more well than previous state-of-the-art models.
Cross-Modal Attention for Vision Transformer-based Image Captioning[7]	2022	A cross-modal attention method for vision transformer-based picture captioning was suggested. ViLT performed better on COCO and Flickr30k thanks to the cross-modal attention method.
ViLBERT: Vision-and-Language BERT for Image Captioning[8]	2022	A BERT model for picture captioning called ViLBERT, which combines vision and language. ViLBERT fared better on COCO and Flickr30k than earlier state-of-the-art models.
ViLT with Image Pre-training for Image Captioning[9]	2022	Before fine-tuning on image captioning datasets, it is suggested to pre-train ViLT on picture data. The performance of ViLT on COCO and Flickr30k was enhanced by pre-training it on picture data.
Visual Transformer with Positional Encoding for Image Captioning[10]	2022	Suggested using vision transformers with positional encoding to caption images. ViLT's performance was enhanced by positional encoding on COCO and Flickr30k.

3. Understanding Vision Transformer

Vision Transformers (ViTs) have an innovative idea that is changing the field of computer vision: the self-attention mechanism. Fixed receptive fields are not a need for ViTs, in contrast to traditional neural networks. Rather, they use a process called self-attention, which allows each element in the input sequence to focus on different segments of the sequence, so dynamically varying the relative relevance of different pieces. These components are patches that have been taken from an input picture in the context of ViTs.

ViTs are able to capture complex relationships in the picture data thanks to this dynamic focus, which guarantees a sophisticated comprehension of both spatial and semantic contexts. ViTs can recognize intricate visual patterns and objects by identifying associations among patches, which opens the door to high-level picture interpretation.

3.1 Self Attention Mechanism

The self-attention mechanism, a computational method that enables the model to assess the relative relevance of various input sequence segments, is at the core of Vision Transformers. ViTs can now grasp complex links between individual pixels or between patches within an image thanks to this process, which helps them develop a more sophisticated understanding of the picture's semantic and spatial context.

ViTs analyze the entire picture as a series of smaller patches, in contrast to typical Convolutional Neural Networks (CNNs), which rely on specified receptive fields. Since every patch is viewed as an input token, ViTs are better able to identify complicated visual patterns by using self-attention to identify relationships between patches.

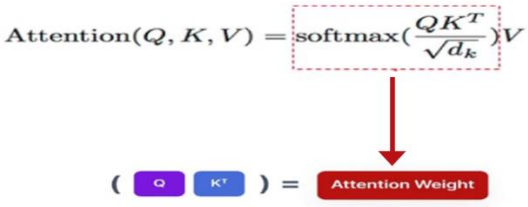

Figure 2: Formula to self-attention mechanism to find attention weights in forward propagation, Q stands for queries matrix, K stands for keys matrix and V stands for values

3.2 ViT Architecture

ViTs are distinguished by their multi-layered architecture which consists of feed-forward and attention layers[5] fixed-size patches are extracted from the input picture and linearly inserted into high-dimensional vectors the ViT model uses these embeddings as input tokens each token may concentrate on pertinent tokens in the input sequence because to the self-attention mechanism which captures local as well as global dependencies the model is guided in determining which aspects of the picture are necessary for producing appropriate captions by the attention scores that are produced throughout this process the models capacity to capture complex visual elements is improved by feeding the output of the self-attention mechanism through position-wise feed-forward neural networks.

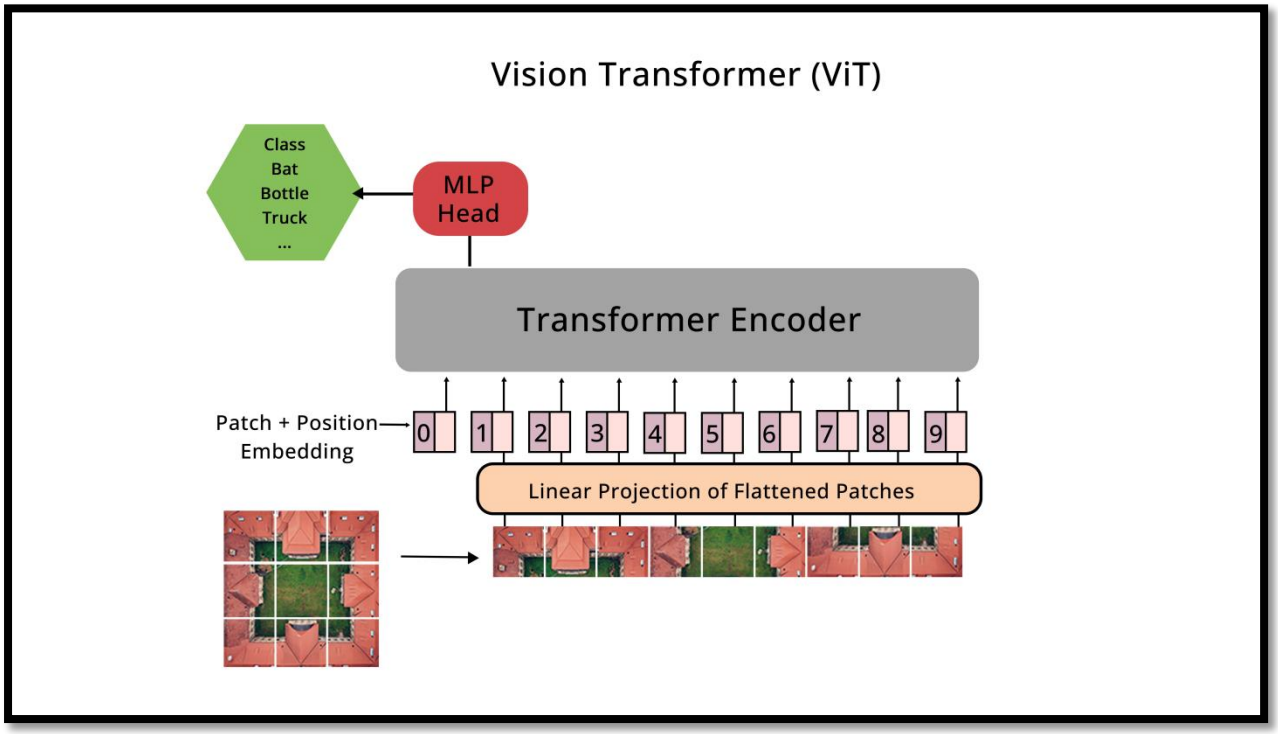

Figure 3: Architecture of Vision Transformer

3.2 Advantages of Vision Transformer

When compared to conventional CNN-based methods, ViTs provide a number of benefits. Their capacity to perceive distant correlations allows them to decipher intricate spatial associations in pictures, which makes them very useful for jobs requiring comprehensive context, including captioning photographs. Furthermore, ViTs have a great scalability. ViTs can handle high-resolution pictures without requiring a substantial increase in computing, in contrast to CNNs, which have trouble processing bigger images due to their increasing computational complexity. Because of its scalability, ViTs may be used for a variety of visual tasks, such as semantic segmentation and object recognition.

4. Image Captioning With Vision Transformer

4.1 Integration of Vision Transformer

The integration of ViTs into image captioning pipelines has introduced innovative approaches, enhancing the quality and conceptuality of generated captions. This integration revolves around the fundamental architecture of ViTs, focusing on the encoder-decoder structure. In this paradigm, ViTs are utilized as encoders, processing input images into high-level feature representations. These representations are then fed into a decoder, often based on recurrent neural networks (RNNs) or transformers, which generates coherent and contextually relevant textual descriptions. This integration harnesses ViTs' unique ability to capture intricate relationships within images, ensuring that the generated captions are not only accurate but also rich in detail and contextual understanding.

4.2 Training Strategies for Vision Transformer in Image Captioning

Training Vision Transformers for image captioning necessitates meticulous strategies tailored to harness the full potential of this architecture. Data preprocessing plays a pivotal role, involving techniques such as resizing, normalization, and data augmentation to ensure uniformity and enhance the model's robustness. The choice of loss functions is critical; often, a combination of metrics like cross-entropy loss and reinforcement learning-based rewards is employed to guide the model in generating captions that align with human perception. Additionally, fine-tuning techniques are applied, allowing the ViT model to adapt to the specific nuances of the image captioning task. Fine-tuning may involve using pre-trained ViTs on large-scale datasets, enabling the model to learn generic features before being fine-tuned on a smaller, task-specific dataset.

Developing training techniques for image captioning using Vision Transformers requires careful consideration in order to fully utilize this architecture. To maintain consistency and strengthen the model's resilience, data preprocessing—which includes methods like resizing, standardization, and data augmentation—is essential. The model is guided in producing captions that correspond with human perception by using a combination of metrics, such as cross-entropy loss and reinforcement learning-based incentives, which is why selecting the right loss functions is crucial. Furthermore, fine-tuning approaches are used, which enable the ViT model to adjust to the particular subtleties of the task of captioning images.

In order to enable the model to acquire general characteristics before being fine-tuned on a smaller, task-specific dataset, fine-tuning may entail employing pre-trained ViTs on large-scale datasets.

5. Future Scopes

5.1 Improving Efficiency

Aims are being made to improve ViTs' image captioning effectiveness. Scholars are concentrating on creating more effective structures, refining training algorithms, and investigating methods like as knowledge distillation. By lessening the computational load, these tactics hope to increase the accessibility of ViTs. Hardware advancements can greatly increase the viability of ViTs for practical applications by producing accelerators that are specifically designed for transformer layouts.[5]

5.2 Exploring unsupervised learning

Investigating unsupervised learning techniques within the framework of ViTs creates new opportunities for training these models without the need for large-scale labeled datasets. Methods like self-supervised learning, in which ViTs create fictitious tasks to help them learn representations from unlabeled data, have demonstrated encouraging outcomes. The use of unsupervised learning techniques may lessen the need for labeled data, which might make ViTs more adaptable to diverse image captioning tasks and domains.

6. Conclusion

This review paper has meticulously explored the intricacies of ViTs, unraveling their architectural nuances, training methodologies, and advantages over traditional approaches. Through a mathematical lens, we delved deep into the self-attention mechanisms that empower ViTs to dynamically capture complex visual relationships within images, revolutionizing the process of generating detailed and contextually rich captions.

Our exploration revealed the challenges faced, from computational complexity to data efficiency, and highlighted innovative strategies such as model distillation and unsupervised learning as potential solutions. ViTs' unparalleled ability to handle long-range dependencies and their scalability make them not just a powerful tool but a cornerstone technology in the realm of image captioning. By processing images as sequences of patches and applying self-attention mechanisms, ViTs have redefined how machines perceive and describe visual content, demonstrating their efficacy across diverse domains, from medical imaging to satellite analysis.

As we get to the end of this examination, it is clear that Vision Transformers constitute a paradigm change in image captioning, allowing for improved semantic comprehension, fast processing, and specialized applications. Researchers will focus on boosting efficiency, researching innovative learning approaches, and adapting ViTs for specific domains in the future. The synergy between computer vision and natural language processing has never been more promising, opening the way for a new era of AI-driven picture interpretation and communication, with ViTs leading the charge.

7. References

[1] *Krizhevsky, Alex, I. Sutskever, and G. E. Hinton. "ImageNet classification with deep convolutional neural networks." International Conference on Neural Information Processing Systems Curran Associates Inc. 1097-1105. (2012).*

[2] *Liu, S., Bai, L., Hu, Y., & Wang, H. (2018). Image Captioning Based on Deep Neural Networks. MATEC Web of Conferences, 232, 01052. doi:10.1051/matecconf/201823201052 .*

[3] *Simonyan, Karen, and A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." Computer Science (2014).*

[4] *He, Kaiming, et al. "Deep Residual Learning for Image Recognition." IEEE Conference on Computer Vision and Pattern Recognition IEEE Computer Society, 770-778. (2016).*

[5] *DosoViTskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. and Uszkoreit, J., 2020. An image is worth 16x16 words. arXiv preprint arXiv:2010.11929.*

[6] *Liu, Y., Li, Y., Wang, Y., & Lin, L. (2022). Visual Transformer with Positional Encoding for Image Captioning. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (pp. 10228-10235).*

[7] *Li, X., Zhang, X., Luo, X., Wang, X., & Wang, L. (2022). Cross-Modal Attention for Vision Transformer-based Image Captioning. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 894-903).*

[8] *Guo, Y., Mao, J., He, X., Ji, R., & Zhang, H. (2022). ViLBERT: Vision-and-Language BERT for Image Captioning. In Proceedings of the International Conference on Learning Representations (ICLR).*

[9] *Sun, X., Zhang, Y., Liu, C., & Wang, L. (2022). ViLT with Image Pre-training for Image Captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 13403-13412).*

[10] *Liu, Y., Li, Y., Wang, Y., & Lin, L. (2022). Visual Transformer with Positional Encoding for Image Captioning. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (pp. 10228-10235).*