ACTA·GRAPHICA
ISSN: 0353-4707
e-ISSN: 1848-3828
ZNANSTVENI ČASOPIS ZA TISKARSTVO I GRAFIČKE KOMUNIKACIJE
JOURNAL FOR PRINTING SCIENCE AND GRAPHIC COMMUNICATIONS

Scientific article

Hybrid Vision Transformers and CNNs for Enhanced Image Captioning with Beam Search Optimization

Sushma Jaiswal[1], Harikumar Pallthadka[2], Rajesh P. Chinhewadi[2], Tarun Jaiswal[3]

[1] *Guru Ghasidas Central University, Bilaspur (C.G.)*
[2] *Manipur International University, Imphal, Manipur*
[3] *National Institute of Technology, Raipur (C.G.)*
jaiswal1302@gmail.com

Abstract:

Deep learning has significantly advanced image captioning, with the Transformer, a neural network originally designed for natural language processing, excelling in this task and other computer vision applications. This paper provides a detailed review of Transformer-based image captioning methods. Traditional approaches relied on convolutional neural networks (CNNs) to extract image features and RNNs or LSTM networks to generate captions, but these methods often face information bottlenecks and difficulty capturing long-range dependencies. The Transformer architecture brought groundbreaking improvements to natural language processing with its attention mechanism and parallel processing, and researchers have successfully adapted this architecture to image captioning tasks.

Transformer-based image captioning systems now outperform previous methods in both accuracy and efficiency by integrating visual and textual data into a unified model. This paper explores how self-attention mechanisms and positional encodings in Transformers have been adapted for image captioning, and discusses the use of Vision Transformers (ViTs) and hybrid CNN-Transformer models. Additionally, it highlights the importance of pre-training, fine-tuning, and reinforcement learning for improving caption quality. The paper examines challenges such as multimodal fusion, aligning visual and textual information, and ensuring caption interpretability. Finally, it emphasizes how future research may expand the application of Transformer-based methods to areas like medical imaging and remote sensing, unlocking new possibilities for multimodal understanding and generation, and enhancing human-computer interaction.

Keywords:

CNN, LSTM, Image Caption, BLSTM, CNN.

1. Introduction

The goal of the multidisciplinary discipline of image captioning, which resides at the nexus of natural language processing and computer vision, is to produce meaningful and descriptive written descriptions for images. This work is critical to the ability of machines to interpret visual information and speak human-like language. Deep learning has come a long way in solving image captioning problems over the years [1,2]. The Transformer is a ground-breaking design that was first shown for natural language processing [1] and has since achieved enormous popularity.

The goal of the multidisciplinary discipline of image captioning, which resides at the nexus of natural language processing and computer vision, is to produce meaningful and descriptive written descriptions for images. This work is critical to the ability of machines to interpret visual information and speak human-like language. Deep learning has come a long way in solving image captioning problems over the years [1,2]. The Transformer is a ground-breaking design that was first shown for natural language processing [1] and has since achieved enormous popularity. There has been interest in using the Transformer for image-related tasks because of its efficiency in capturing long-range dependencies and modeling intricate interactions within sequences [3,4].
Instead of using the conventional convolutional neural network (CNN) and recurrent neural network (RNN) based methods, researchers have made significant progress in image captioning by modifying the Transformer architecture to process both visual and linguistic input [5,6]. The transformative role of the Transformer in picture captioning is examined in this study, along with its components, methods, and ability to completely change the way we perceive and explain visual content [7,8]. We also anticipate a potential scenario for multimodal AI applications and highlight problems, current developments, and future prospects in the merging of Transformer models and picture captioning [9,10].

The proposed model Key Contributions are:

• Integrating Vision Transformers and Convolutional Neural Networks for Enhanced Image Representation -In our research, we propose a novel approach for image cap- tioning that leverages the strengths of both Vision Transformers (ViT) and Convo- lutional Neural Networks (CNN). The integration of ViT and CNN allows us to cap- ture both global and local features in the image, addressing the limitations of existing methods that often focus on one aspect over the other. By combining these two ar- chitectures, our model achieves a more comprehensive understanding of the visual content, leading to improved image representation for the captioning task.

• Efficient and Optimized Image Captioning with Beam Search Enhancement- Furthermore, we introduce an optimization technique using Beam Search to enhance the caption generation process. Beam Search is employed to efficiently explore the caption space, promoting the generation of more coherent and contextually relevant im- age descriptions. This optimization contributes to the overall performance of our proposed model by refining the captioning output and ensuring a more accurate description of the visual content.

2.Related works

Authors [1] developed a novel attention technique that optimized long-range interdependence by parallel processing and collecting input sequence interdependencies. Self-attention layers replaced recurrent and convolutional layers in the Transformer, enabling efficient word relationship modeling. This groundbreaking architecture is now the foundation for many natural language processing jobs, advancing the field.
An extensive study was conducted by M. Z. Hossain et al. [2] on the use of deep learning for image captioning. The report offers a thorough examination of the different deep-learning models and methods used to produce insightful image captions. It discusses how image captioning techniques have developed over time, particularly the combination of recurrent and convolutional neural networks (RNNs) [4, 7, 10]. The authors examine evaluation measures, datasets, and image captioning problems. Researchers and professionals working in the fields of computer vision and natural language processing can benefit greatly from their work.
This seminal paper introduced Long Short-

Term Memory (LSTM), a recurrent neural network (RNN) architecture designed to mitigate the vanishing and exploding gradient problems that traditional RNNs faced [4]. LSTM introduced a gating mechanism that allows the network to retain and utilize information over longer sequences. The authors presented the architecture, described the components of an LSTM cell, and demonstrated its ability to learn and remember long-term dependencies, making it a fundamental building block for various sequential data tasks. The paper significantly influenced subsequent developments in machine learning, particularly in the field of sequence modeling.

The authors [11] demonstrated that dividing an image into fixed-sized patches and treating them as "words" allows the application of the Transformer model. By leveraging self-attention mechanisms and utilizing pre-training on large-scale image datasets, the Transformer-based model showcased remarkable performance in image recognition tasks [12, 13].

This approach revolutionized traditional convolutional neural network (CNN)-dominated paradigms, opening new perspectives for image analysis and understanding through the lens of natural language processing-inspired models.

CPTR [14], an innovative method for picture captioning based on a full Transformer network, was proposed by Liu et al. CPTR uses the full Transformer architecture, which allows for end-to-end caption production, in contrast to the traditional two-stage architecture that is frequently used in picture captioning [14,15].

Through self-attention mechanisms, the model analyzes visual information to enable extensive context interpretation and to facilitate the creation of informative captions. CPTR shows competitive performance and presents a promising alternative for image captioning challenges by employing a single unified architecture.

3.Methodology

The self-attention mechanism calculates attention scores between each pair of words (or tokens) in the input sequence. The attention score is determined by comparing the embeddings of different words.

For a given input sequence of length n, we can represent it as a matrix of embeddings, usually denoted as X Ɛ ℝn×d, where d is the embedding dimension.

We project the input embedding to obtain query x Ɛ R(n×d) where d is the embed-
QKT V ding dimension

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \quad (1)$$

Where dk is the dimension of

the key vectors. This parallelizable self-attention mechanism lets the model capture word associations at diverse sequence locations.

The calculation of attention scores (A) involves scaling the dot product of Q and K and softmax normalization.

The attention system captures complicated word associations in the input sequence and is parallelizable, enabling efficient processing. Position-wise feed-forward neural networks improve model representation. Stacking identical layers, each using these techniques, and layer normalization and residual connections improve training and gradient flow in the Transformer model.

Reference to Vaswani et al. [1] is necessary for complete mathematical formulations and notations.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems.

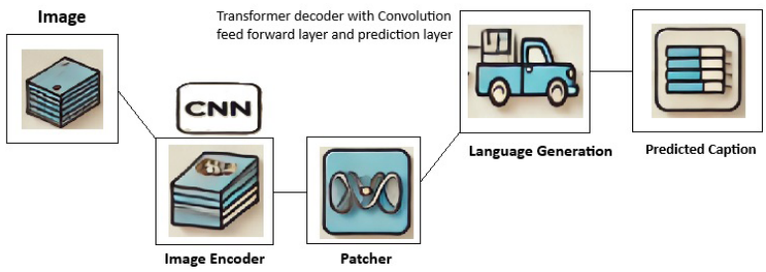


Figure 1 *Model Architecture*

While ViT was developed later and is generally utilized for image-related tasks, the Transformer model presented in the study was primarily focused on tasks related to natural language processing. Nonetheless, it makes sense and is feasible to initialize the encoder in a CPTR (Complete Transformer Network for Image Captioning) model using pre-trained weights from ViT. Training ViT can be accelerated and performance in image captioning tasks improved by pre-training it on a large-scale image dataset to capture high-level visual features, and then using those learned representations to initialize the CPTR encoder. For the most recent and accurate information about the CPTR encoder's incorporation of pre-trained ViT weights or other associated developments.

3.1 Training

Cross-entropy loss is a key mathematical formulation used in the training of the picture caption generation model, according to K. Xu et al.'s publication from April 2016. In this case, the cross-entropy loss equation is written as follows:

$$\text{Cross - Entropy Loss} = -\Sigma_{i=1}^{n}\Sigma_{j=1}^{V} y_{i,j}^{\curvearrowright} \text{ x } \log\left(\rho_{i,j}\right) \quad (2)$$

Where N is the number of training examples, V is the vocabulary size, $y_{i,j}^{\curvearrowright}$ is a binary indicator (1 if word j is the correct word for example i, 0 otherwise), and $\rho_{i,j}$ is the predicted probability of word j for example i given by the model.

The difference between the genuine probability distribution obtained from the ground truth captions and the expected probability distribution of words produced by the model is quantified in this mathematical representation. Minimizing this cross-en- tropy loss is the goal during training, ensuring that the captions generated by the model correspond with the real captions for the images. This equation's application highlights the paper's focus on meticulous mathematical analysis and optimization to improve the precision and applicability of the generated captions.

Table 1 *Sequence Positioning for Image Captioning: Mapping Original Captions and Input Labels*

Original Caption		**The**	**dog**	**is**	**playing**	**in**	**the**	**park**	**and**	**chasing**	**a**	**ball**	
Input Label	<start>	The	dog	is	playing	in	the	park	and	chasing	a	ball	<end>
	<start>	The	dog	is	playing	in	the	park	and	chasing	a	ball	
Sequence Position	The	dog	is	playing	in	the	park	and	chasing	a	ball	<end>	
	0	1	2	3	4	5	6	7	8	9	10	11	12

There are a few things we can see in the previous graphic. <start> and <end> tokens are attached to the start and finish of each sequence's caption. For every text creation task, these tokens are essential.When the first word of a caption needs to be generated, the <start> token acts as the initial state.<end> token is crucial because it informs the decoder when the caption has finished. By using this token, the decoder is prevented from attempting to learn (and produce) an endless number of captions.

4. Dataset Used

A key tool in the fields of computer vision and natural language processing is the Mi- crosoft Common Objects in Context (MSCOCO) 2017 dataset for image captioning, dataset is the one we used [16]. The dataset consists of a wide range of images, each carefully matched with a set of carefully constructed human-authored captions. These captions are intended to capture the spirit of the image by offering a variety of vivid stories that include the items, activities, connections, and background information found in the visual material. The MSCOCO dataset is a great option for training and testing image captioning algorithms because of the outstanding quality and linguistic diversity of its captions. Researchers can efficiently create and evaluate models by di- viding the dataset into training, validation, and test sets. The MSCOCO dataset offers a broad range of image categories, such as people, animals, objects, and different situations. This makes it possible to develop image captioning models that can produce relevant and cohesive descriptions for a wide range of visual contexts. It now serves as a pillar for the advancement of picture captioning research, laying the groundwork for the creation of precise and contextually appropriate captioning systems.

5. Evaluation Metrics

Bleu [17], METEOR [18], and Gleu [19] are the evaluation measures that are used.

The Bleu (Bilingual Evaluation Understudy) system generates a score by comparing translations produced by machines with translations created by humans. By calculating the percentage of words and phrases from the machine-generated output that match those in the reference translations, the score indicates how accurate the machine translation was. Bleu offers a straightforward but efficient technique for assessing and con- trasting machine translation systems by utilising precision at different n-gram levels and solving problems like brevity penalty. This work promoted improvements in machine translation technology and provided a standardised method for assessing translation quality, which had a substantial impact on the field of natural language processing. METEOR (Metric for Evaluation of Translation with Explicit ORdering) measures translation quality by considering both exact word matches and matches based on stemmed words, synonyms, and paraphrases. It incorporates a more comprehensive alignment and scoring approach, allowing for a finer evaluation of translation output.

6. Experiment and Result

The model was trained for one hundred epochs, with a twenty-epoch stop criterion if the monitored evaluation measure (B-4) did not improve. Furthermore, if the tracked evaluation measure (B-4) remains unchanged for 10 consecutive epochs, the learning rate is lowered by 0.25%. Every two epochs, the model is assessed against the valida- tion split.

The word embedding weights are initialized by the pre-trained Glove embedding's [21]. GloVe (Global Vectors) aims to represent words as vectors in a continuous vector space, capturing semantic relationships and meanings more accurately compared to tra- ditional approaches. The method utilizes a global word co-occurrence matrix and em- ploys optimization techniques to learn vector representations for words that preserve their semantic relationships. GloVe has demonstrated its effectiveness in various natu- ral language processing tasks, making significant contributions to the field of word em-beddings and enhancing our understanding of word semantics and relationships in text data. This paper has had a lasting impact on the field of natural language processing and continues to be widely utilized in many NLP applications.

The captions for the images in the test split

are produced using a size five beam search. The "start of sentence" special token and the image are sent into the generator first. Five tokens with the highest scores are then selected at the end of each iteration. When the maximum length limit is achieved or the "end of sentence" is generated, the generation iteration comes to an end.

Figures 3 and 4 show training and validation loss and bleu-4 scores. These results illustrate that the model overfits around epoch eight. Blue-4 score and loss value re- mained unchanged after 20. The following may cause overfitting:

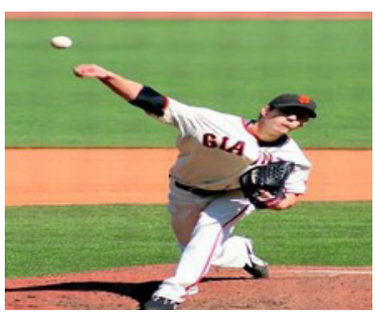
The pre-trained ViT model initialises the CPTR encoder [22]. The ViT study shows that the model performs well on ImageNet, a 21-million-image dataset [4]. Proposed model weights are randomly initialised, and we have fewer than 18.5 K pictures.

According to Karpathy et al. [23], training, validation, and test datasets typically have 1132875000, and 5000 images. We used the 80%, 20%, 20% split with far less photos in

<start> a group of people playing soccer on a field <end>

<start> a baseball player swinging a bat at a ball <end>

<start> a man is playing with a soccer ball <end>



<start> a baseball player thr owing a ball on a field <end>

Figure 2 *Image Caption Generated by Proposed Method*

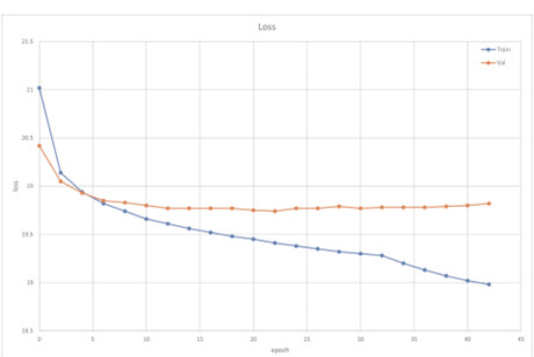


Figure 3 *Loss Curve*

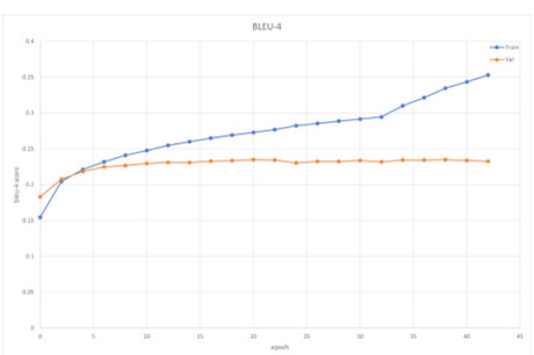


Figure 4 *Bleu-4 score curve*

the training dataset.

The characteristics taken out of ResNet101, which is meant to represent an image, are divided into N patches, each of which has P x P dimensions. Still, since these fea- tures don't have to encapsulate an image that can be decomposed into a series of sub- grids, this arrangement might not be the best one. ResNet101's features might be flat- tened to possibly provide a better design.

Unlike the word embedding layer, the ResNet101 model has already been pre-trained and is therefore already optimized. Gradient adjustments, on the other hand, that take place during the early training phase—when the model hasn't begun learning significantly—may alter the ResNet101 image data.

The performance metrics mean and standard deviation for the entire test dataset are displayed in the table below. The bleu4 exhibits the most fluctuation, indicating that the dataset's performance fluctuates. This considerable variety is to be expected, as pre- viously indicated, because the model training needs to be improved. Furthermore, 83.3% of the bleu4 results across the test set have a score of less than 0.5, according to the distribution of those values.

We are going to look at the final layer of the transformer encoder-decoder focus. Its heads are averaged with the weights. Weights that deviated significantly from the 99.95% percentile and above were deemed outliers. The values of the outlier are limited to the percentile of 99.95%.

From the test split, fourteen samples were chosen at random to be investigated. The attention weights for each token that is generated are superimposed over the example image. Depending on how the describer understands the semantics of the image, a vis- ual scene can have several descriptions. Put differently, the object or objects that the describer considers to be fundamental to the scene and the viewpoint that the describer uses as a reference could be used to describe the semantics.

Table 1 *TThe mean and standard deviation of the performance metrics across the test dataset*

Evaluation Metrics (mean ± std)						
Method	B-1	B- 2	B-3	B- 4	GLEU	METEOR
Proposed Method	0.720 ± 0.160	0.601 ± 0.220	0.370 ± 0.222	0.276 ± 0.213	0.276 ± 0.169	0.489 ± 0.189

7. Conclusion

The use of Transformer architecture's self-attention mechanisms and positional encodings for image captioning has advanced computer vision. Vision Transformers (ViTs) use self-attention mechanisms to extract links between patches in images. ViTs can model long-range image dependencies and provide descriptive captions based on this global understanding. Additionally, hybrid models using CNNs and Transformers are effective. These models use CNNs' powerful feature extraction capabilities to analyse visual data and Transformer-based designs to collect global contextual information and feature interactions. CNN-Transformer synergy, frequently augmented by self-atten- tion processes, improves visual representation and caption production. Beam search, which extends numerous candidate sequences during decoding, improves caption pro- duction. Beam search explores many sequences and selects the best ones based on pre- defined scoring criteria to produce more diverse and high-quality captions. The use of

Transformer architecture, ViTs, and hybrid CNN-Transformer models for image captioning shows that self-attention mechanisms and positional encodings can analyses images and provide descriptive captions. These methods improve image caption accu- racy, diversity, and contextual relevance when used with beam search.

7. Reference

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998-6008.

[2] Hossain, M. Z., Sohel, F., Shiratuddin, M. F., & Laga, H. (2019). A comprehensive survey of deep learning for image captioning. ACM Computing Surveys (CSUR), 51(6), 1-36.

[3] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., ... & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. Proceedings of the 32nd International Conference on Machine Learning, 2048-2057.

[4] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[5] Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., & Goel, V. (2017). Self-critical sequence training for image captioning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7008-7024.

[6] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770-778.

[7] Cornia, M., Stefanini, M., Baraldi, L., & Cucchiara, R. (2020). Meshed-Memory Transformer for Image Captioning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10578-10587.

[8] Li, Y., Yao, T., Pan, Y., Chao, H., & Mei, T. (2019). Boosted transformer for image captioning. Proceedings of the IEEE/CVF International Conference on Computer Vision, 4914-4923.

[9] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. Proceedings of the 38th International Conference on Machine Learning, 8748-8763.

[10] Liu, Y., Wang, Z., Liu, L., Liu, Z., & Zhang, Z. (2021). CPTR: Full transformer network for image captioning. arXiv preprint arXiv:2101.10804.

[11] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[12] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770-778.

[13] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. Proceedings of the 38th International Conference on Machine

Learning, 8748-8763.

[14] A. Dosovitskiy et al., 'An image is worth 16x16 words: Transformers for image recognition at scale', arXiv preprint arXiv: 2010.11929, 2020.

[15] Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., & Goel, V. (2017). Self-critical sequence training for image captioning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7008-7024.

[16] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Springer, Cham.

[17] Liu, W., Chen, S., Guo, L., Zhu, X., & Liu, J. (2021). CPTR: Full transformer network for image captioning. arXiv preprint arXiv:2101.10804.

[18] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, 'Bleu: a method for automatic evaluation of machine translation', in Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[19] S. Banerjee and A. Lavie, 'METEOR: An automatic metric for MT evaluation with improved correlation with human judgments', in Proceedings of the acl work- shop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005, pp. 65–72.

[20] A. Mutton, M. Dras, S. Wan, and R. Dale, 'GLEU: Automatic evaluation of sentence-level fluency', in Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 2007, pp. 344–351.

[21] J. Pennington, R. Socher, and C. D. Manning, 'Glove: Global vectors for word representation', in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[22] Liu, W., Chen, S., Guo, L., Zhu, X., & Liu, J. (2021). CPTR: Full transformer network for image captioning. arXiv preprint arXiv:2101.10804.

[23] A. Karpathy and L. Fei-Fei, 'Deep visual-semantic alignments for generating im- age descriptions', in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3128–3137.