I.J. Intelligent Systems and Applications, **2024, 2, 53-61**
Published Online on April 8, 2024 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2024.02.05

Optimized Image Captioning: Hybrid Transformers Vision Transformers and Convolutional Neural Networks: Enhanced with Beam Search

Sushma Jaiswal*
Guru Ghasidas Central University, Bilaspur (C.G.) and Post-Doctoral Research Fellow, Manipur International University, Imphal, Manipur, India
E-mail: jaiswal1302@gmail.com
ORCID iD: https://orcid.org/0000-0002-6253-7327
*Corresponding author

Harikumar Pallthadka
Manipur International University, Imphal, Manipur, India
E-mail: harikumar@miu.edu.in
ORCID iD: https://orcid.org/0000-0002-0705-9035

Rajesh P. Chinchewadi
Manipur International University, Imphal, Manipur, India
E-mail: rajesh.cto@miu.edu.in

Tarun Jaiswal
National Institute of Technology (NIT), Raipur (C.G.), India
E-mail: tjaiswal_1207@yahoo.com
ORCID iD: https://orcid.org/0000-0003-3963-4548

Received: 11 December 2023; Revised: 11 January 2024; Accepted: 20 February 2024; Published: 08 April 2024

Abstract: Deep learning has improved image captioning. Transformer, a neural network architecture built for natural language processing, excels at image captioning and other computer vision applications. This paper reviews Transformer-based image captioning methods in detail. Convolutional neural networks (CNNs) extracted image features and RNNs or LSTM networks generated captions in traditional image captioning. This method often has information bottlenecks and trouble capturing long-range dependencies. Transformer architecture revolutionized natural language processing with its attention strategy and parallel processing. Researchers used Transformers' language success to solve image captioning problems. Transformer-based image captioning systems outperform previous methods in accuracy and efficiency by integrating visual and textual information into a single model. This paper discusses how the Transformer architecture's self-attention mechanisms and positional encodings are adapted for image captioning. Vision Transformers (ViTs) and CNN-Transformer hybrid models are discussed. We also discuss pre-training, fine-tuning, and reinforcement learning to improve caption quality. Transformer-based image captioning difficulties, trends, and future approaches are also examined. Multimodal fusion, visual-text alignment, and caption interpretability are challenges. We expect research to address these issues and apply Transformer-based image captioning to medical imaging and distant sensing. This paper covers how Transformer-based approaches have changed image captioning and their potential to revolutionize multimodal interpretation and generation, advancing artificial intelligence and human-computer interactions.

Index Terms: ResNet101, Self-attention, Image Caption, ViT and CNN, Beam Search.

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. The Transformer is a ground-breaking design that was first shown for natural language processing and has since achieved enormous popularity. There has been interest in using the Transformer for image-related tasks because to its efficiency in capturing long-range dependencies and modelling intricate interactions within sequences. Instead of using the conventional 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. 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. 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 image captioning.

The problem statement for "Optimized Image Captioning: Hybrid Transformers, Vision Transformers, and Convolutional Neural Networks Enhanced with Beam Search" is to augment image captioning accuracy and efficiency. The primary challenges and goals are:

- Current image captioning systems may not provide accurate and contextually relevant captions. The goal is to improve caption accuracy by incorporating cutting-edge models like Hybrid Transformers, ViTs, and CNN.
- An innovative approach is hybrid model integration, which combines Transformers, Vision Transformers, and CNNs. The aim is to smoothly integrate these models to maximize their strengths and develop a hybrid system that captions images well.
- Vision Transformers could improve visual understanding. The problem statement optimizes Vision Transformers in the hybrid framework to extract significant image information and improve captioning.
- Image-related activities have generally been successful for CNNs. The hybrid approach uses CNNs to record hierarchical visual features for better interpretation and contextual captions.
- Advanced decoding can improve caption production. Beam Search is added to the decoding stage to provide accurate, diversified, and contextually suitable captions.
- Determining hybrid model training methodologies and performance improvement metrics is the difficulty. Selecting loss functions, optimizing hyperparameter, and rigorously testing against benchmark datasets are required.
- Scalability and efficiency become important with complicated models. Solving scalability and computational efficiency difficulties in the hybrid image captioning system is the problem.

The goal is to improve image captioning by combining Transformers, Vision Transformers, and Convolutional Neural Networks. Beam Search enhances the integration by increasing caption diversity and conceptuality, making image captioning more accurate and efficient.

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 captioning that leverages the strengths of both ViT and CNN. The integration of ViT and CNN allows us to capture 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 architectures, 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 image descriptions. This optimization contributes to the overall performance of our proposed model by refining the captioning output and ensuring a more accurate depiction of the visual content.

2. Related Works

In 2017, Vaswani et al. [1] published "Attention Is All You Need" which revolutionized NLP and deep learning. The Transformer model, a pioneering neural network architecture, is used in many applications, especially sequence-to-sequence tasks. Self-attention layers replaced recurrent and convolutional layers in the Transformer, enabling efficient word relationship modelling. This groundbreaking architecture is now the foundation for many natural languages 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). 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 [3]. 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.

Dosovitskiy et al. [4] demonstrated that dividing an image into fixed-sized patches and treating them as "words" allows the application of the Transformer model. By lever-aging self-attention mechanisms and utilizing pre-training on large-scale image datasets, the Transformer-based model showcased remarkable performance in image recognition tasks. 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 [5], an innovative method for image 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 image captioning. Through self-attention mechanisms, the model analyses 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 [5]. The Transformer model, introduced in "Attention Is All You Need," emphasizes self-attention systems' ability to capture sequence relationships. When used for exemplar-based colorization, the Transformer's attention mechanism helps use reference examples accurately and contextually. The Transformer model's attention mechanism, which weighs distinct input sequence portions differently during processing, is its main innovation. This attention approach is parallelizable and efficient for long-range dependencies since it applies to each input sequence member independently [6].

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 embedding's, usually denoted as $x \in R^{(n\times d)}$, where d is the embedding dimension. Input embeddings are projected into Query (Q) obtained by projecting the input embedding using a learned weight matrix for queries. Key (K) obtained by projecting the input embedding's using a learned weight matrix for keys, and Value (V) Obtained by projecting the input embedding's using a learned weight matrix for values, matrices using the self-attention process.

We project the input embedding to obtain query.

$$x \in R^{(n\times d)} \tag{1}$$

Where d is the embedding dimension.

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)^V \tag{2}$$

Where d_k is the dimension of the key vectors. This parallelizable self-attention mechanism lets the model capture word associations at diverse sequence locations.

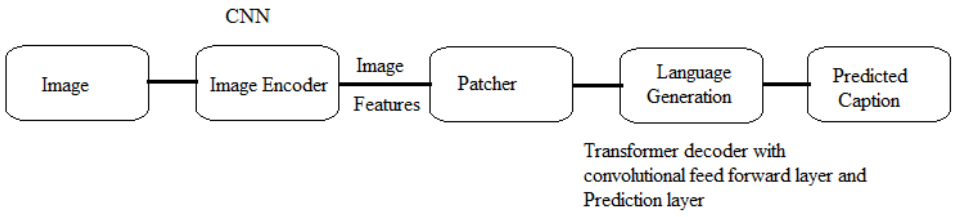

Fig.1. Model architecture

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,

Original Caption		**the**	**dog**	**is**	**running**	**on**	**the**	**grass**	
Input Label	<start>	the	dog	is	running	on	the	grass	<end>
	<start>	the	dog	is	running	on	the	grass	
	the	dog	is	running	on	the	grass	<start>	
Sequence Position	0	1	2	3	4	5	6	7	8

Fig.2. <start> and <end> tokens image caption

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. Training

Cross-entropy loss is a key mathematical formulation used in the training of the image caption generation model, according to K. Xu et al.'s [7] publication from April 2016. In this case, the cross-entropy loss equation is written as follows:

$$Cross - Entropy\ Loss = -\sum_{i=1}^{n}\sum_{j=1}^{V} y_{i,j} \cdot \log\left(p_{i,j}\right) \quad (3)$$

Where N is the number of training examples, V is the vocabulary size, $y_{i,j}$ is a binary indicator (1 if word j is the correct word for example i, 0 otherwise), and $p_{i,j}$ 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-entropy 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.

5. Dataset Used

A key tool in the fields of computer vision and natural language processing is the Microsoft Common Objects in Context (MSCOCO) 2017 dataset for image captioning, dataset is the one we used [8]. 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 dividing 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 image captioning research, laying the groundwork for the creation of precise and contextually appropriate captioning systems.

6. Evaluation Metrcs

Bleu [9], METEOR [10], and Gleu [11] are the evaluation measures that are used. The Bleu (Bilingual Evaluation Understudy) [8] 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 contrasting machine translation systems by utilizing precision at different n-gram levels and solving problems like brevity penalty. This work promoted improvements in machine translation technology and provided a standardized 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) [10] 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. The improvements offered by METEOR over previous metrics Help Bridge the gap between automatic evaluation and human perception of translation quality. This paper significantly contributed to

advancing the field of machine translation evaluation, facilitating more accurate and nuanced assessments of translation systems.

- GLEU (Grammar and Length-based Evaluation Utility) [11], an automatic metric for assessing sentence-level fluency in machine-generated translations. GLEU evaluates fluency by considering both grammatical correctness and the length of matching phrases between the reference and candidate sentences. It addresses the limitations of other metrics by incorporating grammatical analysis, providing a more nuanced evaluation of translation quality at the sentence level. By focusing on fluency, GLEU offers valuable insights into the linguistic aspects of translations, aiding in the development and refinement of machine translation systems. This paper represents a significant contribution to the field of automatic evaluation metrics, enhancing our ability to assess the linguistic quality of translated text.

7. 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 validation split. The word embedding weights are initialized by the pre-trained Glove embedding's [12].

GloVe (Global Vectors) aims to represent words as vectors in a continuous vector space, capturing semantic relationships and meanings more accurately compared to traditional approaches. The method utilizes a global word co-occurrence matrix and employs optimization techniques to learn vector representations for words that preserve their semantic relationships. GloVe has demonstrated its effectiveness in various natural language processing tasks, making significant contributions to the field of word embedding's, 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.

Fig. 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 remained unchanged after 20.

Fig.3. Image caption generated by proposed method

The following may cause overfitting:

- The pre-trained ViT model initializes the CPTR encoder. The ViT study shows that the model performs well on ImageNet, a 21-million-image dataset [4]. Proposed model weights are randomly initialized, and we have fewer than 18.5 K images. According to Karpathy et al. [13], training, validation, and test datasets typically

have 1132875000, and 5000 images. We used the 80%, 20%, 20% split with far less photos in 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 features don't have to encapsulate an image that can be decomposed into a series of subgrids, this arrangement might not be the best one. ResNet101's features might be flattened 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.

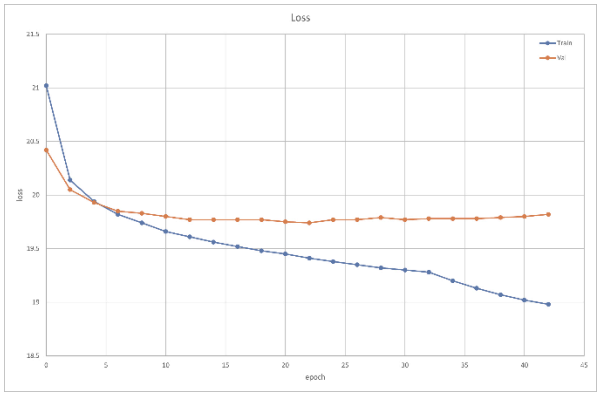

Fig.4. Loss curve

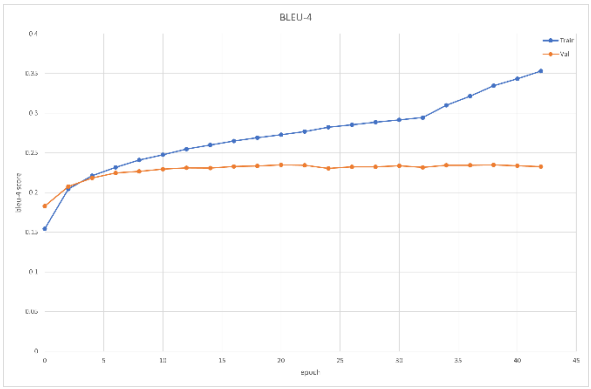

Fig.5. Bleu-4 score curve

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 previously 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.

Table 1. The mean and standard deviation of the performance metrics across the test dataset

Method	**Evaluation Metrics (mean ± std)**					
	B-1	B-2	B-3	B-4	GLEU	METEOR
Proposed Method	0.7199 ± 0.16	0.601 ± 0.220	0.3698 ± 0.222	0.2896 ± 0.213	0.2756 ± 0.169	0.4889 ± 0.189

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 visual 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.

8. Limitations and Challenges

"Optimized Image Captioning: Hybrid Transformers, Vision Transformers, and CNN Enhanced with Beam Search" is a novel and advanced image captioning system, but it has limitations and challenges:

- Transformers, Vision Transformers, and Convolutional Neural Networks in the hybrid model may raise training and inference computational needs. This may restrict system scalability and deployment, especially on resource-constrained devices.
- The training dataset quality and diversity greatly affect the suggested model's performance. The model may generalize poorly if the dataset is biased or lacks domain representation, resulting in real-world inaccuracies and limits.
- Optimizing the hybrid model requires setting many hyperparameters, which may affect performance. Finding the best configuration may take time, and the model's resilience may vary between datasets or contexts.
- Hybrid models with multiple architectures may make decision-making interpretation difficult. Deciphering how each component affects caption production is difficult, limiting model interpretability.
- Overfitting is possible with the hybrid model's complexity, especially if the training dataset is small or undiversified. A major problem is generalizing the model to new data and image categories.
- It can be difficult to seamlessly integrate diverse architectures, which may affect the hybrid model's effectiveness. Making sure components work together without conflicts or redundancies is difficult.
- Beam Search increases caption diversity but may reduce computing efficiency. Consider the trade-off between caption quality and real-time performance.
- Model performance may vary by domain or dataset. It may perform well in certain images but poorly in others, restricting its uses.

In order to address these constraints, viable directions for further study must be carefully considered. The success of the suggested approach in actual image captioning situations depends on finding a balance between model complexity and realistic deployment constraints.

9. Conclusion and Discussion

Image captioning employing Transformer architecture's self-attention and positional encodings improves computer vision. Through self-attention, ViTs find image patch linkages. The global understandings allow ViTs to model long-range image dependencies and create relevant captions. Also effective are CNN-Transformer hybrid models. Transformer-based designs collect global contextual information and interactive aspects, while CNNs extract visual features. Self-attention and CNN-Transformer synergy improve captioning and visual representation. Beam search improves captions by extending candidate sequences during decoding. To generate more variety and high-quality captions, beam search selects the best sequences using established scoring criteria. Transformer architecture, ViTs, and hybrid CNN-Transformer models show that positional encodings and self-attention may caption images descriptively. Together with beam search, these tactics improve caption accuracy, diversity, and contextual relevance. The suggested strategy yields promising evaluation metrics across dimensions. The mean values for BLEU-1 (B-1), BLEU-2 (B-2), BLEU-3 (B-3), and BLEU-4 (B-4) are 0.7199, 0.601, 0.3698, and 0.2896, respectively, with accompanying standard deviations indicating model stability (± 0.16, ± 0.220, ± 0.222, and ± 0.213). Furthermore, GLEU (0.2756 ± 0.169) and METEOR (0.4889 ± 0.189) metrics provide a comprehensive evaluation of the suggested method's efficacy. The mean values' precision and standard deviations' consistency demonstrate the suggested method's reliability across many evaluation criteria. The suggested captioning approach captures language nuances and contextual significance competitively, according to these results. These metrics must be interpreted in light of the application domain's requirements and peculiarities. The method's positive mean values and comparable standard deviations across evaluation metrics demonstrate its reliability. These results make the proposed caption generation method promising for further study and real-world use.

References

[1] Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I., “Attention is all you need”, Advances in neural information processing systems 30, 2017.

[2] Hossain, M.Z., Sohel, F., Shiratuddin, M.F., & Laga, H.,“A Comprehensive Survey of Deep Learning for Image Captioning”, ACM Computing Surveys (CSUR) 51, ACM Computing Surveys, Volume 51, Issue 6, Article No.: 118, pp 1–36, 2018. https://doi.org/10.1145/3295748

[3] Hochreiter, S., & Schmidhuber, J., “Long Short-Term Memory”, Neural Computation, Volume 9, Issue 8, November 15, pp 1735–1780, 1997 https://doi.org/10.1162/neco.1997.9.8.1735.

[4] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N., “An image is worth 16x16 words: Transformers for image recognition at scale”, arXiv

preprint arXiv: 2010, 11929, 2020.
[5] Liu, W., Chen, S., Guo, L., Zhu, X., & Liu, J., "CPTR: Full transformer network for image captioning", arXiv preprint arXiv: 2101, 10804, 2021.
[6] Yin, W., Lu, P., Zhao, Z., & Peng, X., "Yes, "Attention Is All You Need", for Exemplar based Colorization", In Proceedings of the 29th ACM international conference on multimedia, pp. 2243-2251, 2021.
[7] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A.C., Salakhutdinov, R., Zemel, R.S., & Bengio, Y., "Show, attend and tell: Neural image caption generation with visual attention", In International conference on machine learning, pp. 2048-2057. PMLR, 2015.
[8] Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C.L., "Microsoft coco: Common objects in context", In Computer Vision–ECCV 2014, 13th European Conference, Zurich, Switzerland, September 6-12, Proceedings, Part V 13, pp. 740-755. Springer International Publishing, 2014.
[9] Papineni, K., Roukos, S., Ward, T., & Zhu, W., "Bleu: a method for automatic evaluation of machine translation", In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp. 311-318, 2002.
[10] SBanerjee, S., & Lavie, A., "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments", In Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, pp. 65-72, 2005.
[11] Mutton, A., Dras, M., Wan, S., & Dale, R., "GLEU: Automatic evaluation of sentence-level fluency", In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pp. 344-351, 2007.
[12] Pennington, J., Socher, R., & Manning, C.D., "Glove: Global vectors for word representation", In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543, 2014.
[13] Karpathy, A., & Fei-Fei, L., "Deep visual-semantic alignments for generating image descriptions", In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3128-3137, 2015.

Authors' Profiles

Sushma Jaiswal is an Assistant Professor in the department of CSIT of Guru Ghasidas Central University, Bilaspur, (C.G.). She has completed Ph.D. in the field of image processing and her fields of interest are computer graphics and machine vision. She has teaching experience of 18 years, and she has published many papers in various national and international journals as well as presented her work at various conferences.

Harikumar Pallthadka is a Professor and Vice Chancellor of Manipur International University, Imphal, Manipur. His research interests span both computer networking and network science. Much of his work has been on improving the understanding, design, and performance of parallel and networked computer systems, mainly through the application of data mining, statistics, and performance evaluation.

Rajesh P. Chinchewadi is CTO & Dean Innovation Manipur International University, Imphal, Manipur. His research focuses on the development of computational methods for scalable and responsible discovery science. He has to his credit of publishing number of research papers including in International and National Journals.

Tarun Jaiswal is a Research Scholar at the Department of Computer Applications, National Institute of Technology, Raipur. His research interests include Machine Learning, Artificial Intelligence and IoT. He has published several research papers in leading national and international journals, as well as presented His work at various conferences. He has also worked on several projects related to machine learning, which has given his practical experience in applying theoretical concepts to real-world problems.

How to cite this paper: Sushma Jaiswal, Harikumar Pallthadka, Rajesh P. Chinchewadi, Tarun Jaiswal, "Optimized Image Captioning: Hybrid Transformers Vision Transformers and Convolutional Neural Networks: Enhanced with Beam Search", International Journal of Intelligent Systems and Applications(IJISA), Vol.16, No.2, pp.53-61, 2024. DOI:10.5815/ijisa.2024.02.05