International Journal of Engineering Research Excellence and Applied Science (IJEREAS)

Volume-II (Issue 4) – December 2024 ISSN: 3048-5355

TRANSFORMATIVE METHODS IN IMAGE CAPTIONING TECHNOLOGY

Minakshi Tomer [1†], Tripti Rathee[2*†]

[1]Information Technology, Maharaja Surajmal Institute of Technology, Janakpuri, New Delhi, 110058, New Delhi, India.

[2*]Information Technology, Maharaja Surajmal Institute of Technology, Janakpuri, New Delhi, 110058, New Delhi, India.

.

*Corresponding author(s). E-mail(s): tomer.minakshi@gmail.com; Contributing authors: rathee.tripti@gmail.com;

[†]These authors contributed equally to this work.

ABSTRACT

AI image captioning involves generating descriptive text for images using machine learning, often combining image processing with CNNs and NLP techniques like transformers. This research paper explores AI image captioning, focusing on integrating VGGNet and Transformers to generate descriptive captions. The study includes Image Processing, where CNNs like VGGNet and ResNet extract features, and NLP, where Transformers create captions using attention mechanisms. The experimental setup uses the Flickr30k dataset and evaluates model performance with BLEU and ROUGE metrics. The paper thoroughly analyzes efficacy, supported by insightful visualizations. The integration of VGGNet with Transformers enriches AI image captioning, offering insights for future advancements. This research bridges innovative methodologies with practical implementations, paving the way for further exploration in this burgeoning field.

Keywords:

Image captioning; Natural Language Processing; LSTM; VGGNet; Transformers

1. INTRODUCTION

In our daily routines, we come across countless images on different digital platforms like the internet, news articles, and advertisements. While humans can naturally understand these visuals without detailed descriptions, machines require explicit image captions to interpret them automatically. This highlights the importance of image captioning, which plays a vital role in various applications, such as automatic image indexing and content-based image retrieval, biomedicine, commerce, and education. The significance of image captions extends to social media, where they improve user experiences by providing contextual details like location, attire, and depicted activities. In artificial intelligence (AI), image captioning connects image understanding and linguistic description, allowing machines to identify and recognize objects within images, understand diverse perspectives, and object properties, and generate coherent sentences. To achieve these results, two complementary approaches are employed: traditional machine learning and deep machine learning. Traditional methods rely on manual processing and classifications, while deep learning algorithms, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have gained attention due to their ability to learn features from expansive datasets across various domains. Thus, image captioning is an essential tool in the world of digital media, and further research will continue to advance this technology, leading to new and exciting possibilities.

The field of artificial intelligence has seen significant advancements in recent years, particularly in the area of image understanding and captioning. Deep learning algorithms such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been pivotal in this progress. By autonomously learning and extracting features from large datasets, these algorithms have improved the accuracy of image descriptions. As research continues to explore the relationship between image comprehension and linguistic interpretation, diverse domains benefit from innovative application potential. The integration of traditional and deep learning methods in image captioning reflects the evolution of AI technologies and highlights the ongoing pursuit of sophisticated and efficient ways to interact with visual data in the digital age.

The field of AI-powered graphics has been gaining immense popularity lately, owing to its remarkable ability to bridge the gap between visual understanding and natural language graphics. This technology has revolutionized the way humans and machines communicate, making it seamless and effortless. Machine learning methods, especially Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), play a crucial role in extracting key features from images. The paper [1] provides valuable insights into this technology. It identifies common architectural models in AI image annotation models and connects traditional feature extraction techniques with modern deep learning techniques. The paper discusses the progress made in image feature extraction and highlights the shift from traditional methods like local binary models (LBPs) and scale-invariant feature transformation (SIFT) to deep learning models. This transformation is significant as deep learning enables machines to automatically extract complex features directly from raw image data, eliminating the need for artificial features and predefined algorithms.

This paper delves into the process of AI image captioning through two essential phases aimed at achieving the overarching objective of generating descriptive captions for images. The initial phase revolves around Image Processing, where various feature-extracting Machine Learning models are employed, notably including the implementation of VGGNet, a unique approach not commonly found in existing literature. This phase heavily relies on Convolutional Neural Networks (CNNs) to extract features from images, enhancing the model's understanding of visual content. Additionally, architectures like ResNet are utilized, offering simplicity, effectiveness, and innovative solutions to challenges like the vanishing gradient problem. In the subsequent phase, the focus shifts to Natural Language Processing (NLP), where advanced models such as Transformers are employed for caption generation, another distinctive aspect of this paper. Transformers further enhance this process by effectively transforming input sequences into output sequences, utilizing attention mechanisms to track sequential data relationships. The experimental setup leverages the Flickr30k dataset and offers quantitative insights into the model's performance through evaluation metrics such as BLEU and ROUGE. The integration of VGGNET with Transformers is a unique implementation that enriches the field of AI image captioning and offers fresh perspectives for further exploration and refinement. By seamlessly integrating these phases, the research endeavors to develop a robust system capable of generating accurate and contextually rich captions for images, thereby advancing the realm of AI-driven image captioning. The remaining part of the paper is divided as literature survey in section 2 followed by methodology in section 3. The section 4 explains

S. No.	Year	Name of Paper	Author	Technique	Data Source	Description

the experimental setting used for the model followed by results in section 5. Finally, the paper concludes in

1	2019	A comprehensive Survey of Deep Learning for Image captioning	MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin and Hamid Laga	Encoder-decoder architectures, attention mechanisms, and reinforcement learning	Flickr8k, Flickr30k, MS COCO	The paper presents a comprehensive review of various deep learning based image captioning techniques and their comparison on the basis of performance, strength, limitation and datasets used.
2	2024	Exploring better image captioning with grid features	Jie Yan, Yuxiang Xie, Yanming Guo, Yingmei Wei & Xidao Luan	Transformer-based image captioning model - FeiM	MS COCO	The paper discusses how to use grid visuals' expressive features. It uses feature queries to capture specific visual information and a transformer architecture for multi-modal fusion.
3	2024	PICS: Pipeline for Image Captioning and Search	Grant Rosario, David Noever	LLaVa and Mistral-7B	Open Images	The paper represents PICS (Pipeline for Image Captioning and Search) that enhances the searching and accessibility capabilities in large databases automatically.
4	2024	Optimized Image Captioning: Hybrid Transformers Vision Transformers and Convolutional Neural Networks: Enhanced with Beam Search	Sushma Jaiswal, Harikumar Pallthadka, Rajesh P. Chinchewadi, Tarun Jaiswal	ResNet101, Self-attention, Image Caption, ViT and CNN, Beam Search	MS COCO	The paper presents a Transformer based image captioning techniques to address the issues of Multimodal fusion, visual text alignment and caption interpretablity for better medical imaging and distant sensing.
5	2024	CIC: A framework for Culturally-aware Image Captioning	Youngsik Yun, Jihie Kim	visual modality and Large Language Models	GD-VCR	The paper addresses the issue of cultural element identification in the images. Visual Modality and LLMs are combined to extract cultural visual elements and generate culturally aware captions using LLM.

6	2024	Image Captioning in News Report Scenario	Tianrui Liu, Qi Cai, Changxin Xu, Bo Hong, Jize Xiong, Yuxin Qiao, Tsungwei Yang	Computer Vision and Natural Language Processing	Flickr3K, Flickr8K	The paper explores the captioning of celebrity photographs used by news channels. Augmented automated news content generation to enhance the dissemination of information.
7	2023	Deep Learning Approaches on Image Captioning: A Review	Taraneh Ghandi, Hamidreza Pourreza, Hamidreza	Vision-language Pre-training	MS COCO, Flickr30k	The paper aims to provide a structured review of deep learning methods in image captioning by presenting a comprehensive taxonomy and discussions.
8	2023	Improving image captioning methods using machine learning approaches	Atliha, Viktar	Model Compression	MS COCO, Flickr8K, Flickr30K	The paper concentrates on reducing the model size of pre-existing image captioning models based on Computer Vision and Natural Language processing. The methods propose reduction that also aims to improve quality of captions generated.
9	2023	Automated testing of Image Captioning Systems	Boxi Yu, Zhiqing Zhong, Xinran Qin, Jiayi Yao, Yuancheng Wang, Pinjia He	CNN-RNN, Vision-Language Pre-Training (VLP), SOTA VLP IC model VIVO	MS COCO	The paper presents MetaIC, one of the metamorphic testing approach to validate Image Captioning. The process tries to eliminate errors - misclassification, omission and incorrect quantity.
10	2023	Controllable Image Captioning via Prompting	Ning Wang, Jiahao Xie, Jihao Wu, Mingbo Jia, Linlin Li	Encoder and Multi-modal fusion model	MS COCO, TextCaps	This paper presents a new image captioning approach that outperforms existing models in generating captions of desired length and style. It incorporates prompt-based captioning, validated manual prompts, and auto-prompt learning for better performance.

11	2023	Aesthetically Relevant Image Captioning	Zhipeng Zhong, Fei Zhou, Guoping Qiu	ARIC (Aesthetically Relevant Image Captioning) model	DPC2022	The research paper introduces the Aesthetically Relevant Image Captioning (ARIC) model, which addresses the challenge of generating image captions that capture aesthetic aspects effectively.
12	2023	Efficient Image Captioning for Edge Devices	Ning Wang, Jiangrong Xie, Hang Luo, Qinglin Cheng, Jihao Wu, Mingbo Jia, Linlin Li	CLIP (Contrastive Language-Image Pretraining) model	COCO and Visual Genome dataset (Krishna et al. 2017)	LightCap is a new image captioning model designed for mobile devices that efficiently extracts features without relying on object detectors. It achieves state-of-the-art performance with only 40M parameters and is highly efficient for real-world applications.
13	2023	Exploring Diverse In-Context Configurations for Image Captioning	Xu Yang, Yongliang Wu, Mingzhuo Yang, Haokun Chen, Xin Geng	Vision-Language Model (VLM), Natural Language Processing (NLP)	MS COCO	The paper discusses using few-shot learning in Vision-Language Models for image captioning and proposes practical strategies like Iterative Prompting .
14	2023	Image Captioning with Semantic Attention	Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, Jiebo Luo	Attention mechanism, Visual attribute prediction	Flickr30K, MS COCO	The paper presents the combination of top-down and bottom-up approaches of image captioning using semantic attention. The algorithm used is able to fuse semantic concepts with hidden states and outputs of RNNs.
15	2023	CaptionGenX: Advancements in Deep Learning for Automated Image Captioning	Shubham Bhaiyaji Derkar, Dipak Biranje, Laxman P Thakare, Swati Paraskar, Rahul Agrawal	CNN and RNN	MS COCO, Flickr8k, and Flickr30k	This paper proposes a novel approach for controllable image caption generation by embedding prompt learning into the model. It combines CNNs and RNNs with attention mechanisms to improve alignment.

16	2023	Enhancing Image Captioning with Neural Models	Pooja Bhatnagar, Sai Mrunaal, Sachin Kamnure	Merge and Inject Models	Yelp	This research delves into neural image captioning using deep learning, focusing on the "inject" model. It highlights the importance of data refinement and hyperparameter optimization for better performance and contributes to AI democratization.
17	2022	Explaining transformer-based image captioning models: An empirical analysis	Marcella Cornia ∗ , Lorenzo Baraldi and Rita Cucchiara	Transformer-based image captioning	COCO dataset and ACVR Robotic Vision Challenge	The paper investigates Transformer-based image captioning, proposing metrics for temporal alignment assessment between model predictions and attribution scores. It examines various image encoding methods and applies Reinforcement Learning for optimization..
18	2022	Image Captioning Encoder–Decoder Models Using CNN-RNN Architectures: A Comparative Study	K. Revati Suresh, Arun Jarapala1, P. V. Sudeep1	Neural Image Captioning (NIC), CNN, RNN	Flickr8K	The paper compares different models for image captioning using CNN-RNN architectures and finds that a ResNet-101 encoder with an LSTM decoder works best. They also highlight the effectiveness of parallel-inject concatenate models and beam search.
19	2022	An accurate generation of image captions for blind people using extended convolutional atom neural network	Tejal Tiwary & Rajendra Prasad Mahapatra	Automatic Image Captioning (AIC)	Freiburg Groceries Dataset and Grocery Store	The paper introduces an assistive technology using Automatic Image Captioning (AIC) to help visually impaired individuals recognize food items during online grocery shopping.
20	2022	Evaluating the effectiveness of automatic image captioning for web accessibility	Maurizio Leotta, Fabrizio Mori & Marina Ribaudo	Azure Computer Vision Engine, Amazon Rekognition, Cloudsight, Auto Alt-Text for Google Chrome	MS COCO, VizWiz	The paper presents four state of art tools - Azure Computer Vision Engine, Amazon Rekognition, Cloudsight, and Auto Alt-Text for Google Chrome that can be used to generate captions for web-based images automatically.

21	2021	Image Captioning through Cognitive IOT and Machine-Learning Approaches	Tarun Jaiswala, Manju Pandey b, and Priyanka Tripathi c	Deep learning architectures, Attention mechanisms, Transfer learning, Reinforcement learning and Ensemble methods	MS COCO, Flickr30K, Visual Genome and Open Images	The paper describes the comprehensive overview of prevalent deep-learning based text generation methods.
22	2020	Show, Interpret and Tell: Entity-Aware Contextualised Image Captioning	Khanh Nguyen, Ali Furkan Biten, Andres Mafla, Lluis Gomez, Dimosthenis Karatzas	Masked Named Entity Modelling (MNEM)	WIT	This work uses Wikipedia articles to generate contextualized captions for images. A technique called Masked Named Entity Modelling (MNEM) is proposed to improve entity recognition in the generated captions.
23	2020	Explainable AI (XAI) approach to image captioning	Seung-Ho Han, Min-Su Kwon, Ho-Jin Choi	Deep learning methods - CNN and RNN	Flickr30K, MS COCO	AI-powered image captioning can be improved by using explainable AI (XAI) that establishes visual links between image objects and words in captions. This approach was shown to be effective on MSCOCO and Flickr30K datasets.
24	2019	Uncertainty-Aware Image Captioning	Zhengcong Fei, Mingyuan Fan, Li Zhu, Junshi Huang* Xiaoming Wei, Xiaolin Wei	Dynamic programming, uncertainty-adaptive parallel beam search	MS COCO	This paper introduces an uncertainty-aware image captioning framework that addresses the challenge of generating captions with varying levels of uncertainty in words.
25	2019	A Systematic Literature Review on Image Captioning	Raimonda Staniut¯e and Dmitrij Šešok	Encoder-decoder, Attention mechanism, Novel objects, Sematics	Flickr30K, MS COCO	This study presents a Systematic Literature Review (SLR) summarizing advancements in image captioning over the past four years. It highlights common techniques, major challenges, and inconsistencies in result comparisons.

section 6.

Table 1: Summary of research papers

3. METHODOLOGY

The methodology employed in this research paper is designed to provide constructive insights into generating descriptive captions for images. The two integral phases, Image Processing and Natural Language Processing (NLP), are meticulously planned and executed to ensure that the research achieves its overarching objective. The first phase uses a variety of Machine Learning models to extract features from images, while the second phase uses advanced models such as Long Short-Term Memory networks (LSTMs) and transformers to process natural language. These phases are essential to the research and provide a constructive framework for generating descriptive captions for images.

Fig. 1 illustrates the workflow of AI Image Captioning Models. The models include three for computer vision and two for natural language processing, which can be combined to create six distinct AI image captioning models. By applying the Flickr_30k dataset to all of these AI image captioning models, comparative analysis can be carried out and a relatively exceptional model can be inferred from the given dataset.

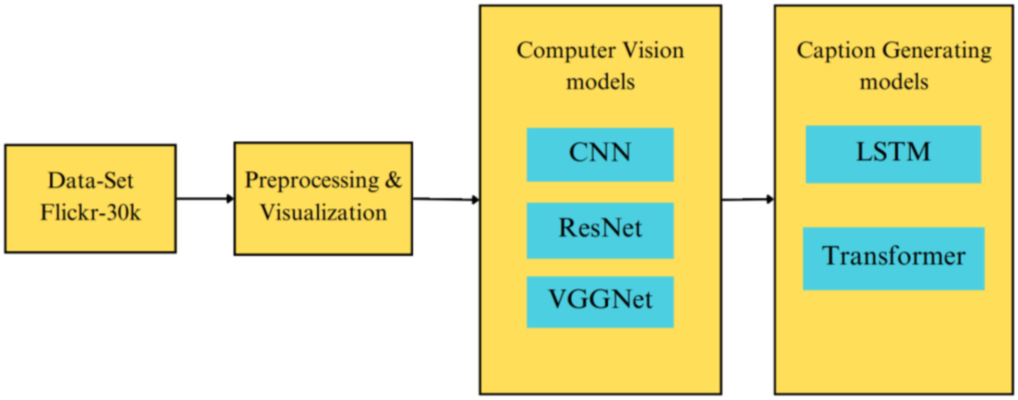

Fig 1: Workflow of AI Image Captioning

3.1 COMPUTER VISION PHASE

The Computer Vision Phase involves the extraction and processing of visual information from images. This phase utilizes advanced deep learning models to analyze and understand the visual content. By transforming raw pixels into meaningful representations, this phase serves as the foundational step in generating accurate and descriptive captions for images.

3.1.1 CNN - As discussed in [18], CNN is an essential as well as one of the most widely used tools for captioning images because it acts as a conduit between the creation of captions and raw images. It comprises a classifier and a feature extractor with pooling and convolutional layers as shown in Fig 2. While deeper layers extract intricate forms and patterns, the Conv layers catch fundamental elements like color and borders. These characteristics improve the network's categorization skills by enabling precise visual interpretation and captioning

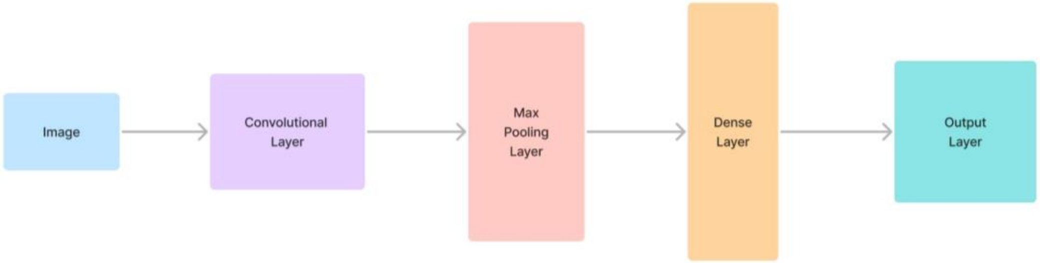

Fig 2: Architecture of CNN

3.1.2 RESNET- It is one of the crucial models used in the computer vision phase as stated in [4]. With skip connections and ReLU activations between layers (as shown in Fig 3), ResNet (Residual Network) tackles the vanishing gradient issue in deep neural networks. It is available in multiple variants with varying amounts of layers, ranging from ResNet-18 to ResNet-152. The top-5 error rate often drops with an increase in layer count, with ResNet-152 having the lowest mistake rate. However, because ResNet-152 is more sophisticated and has a higher parameter count, ResNet-50 and ResNet-101 are frequently used for testing. [4]

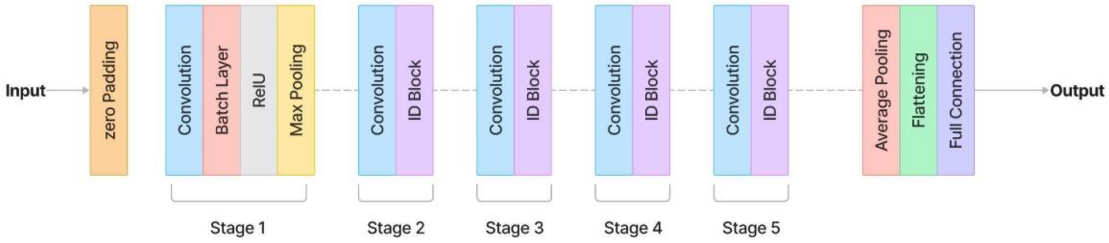

Fig 3: Architecture of RESNET

3.1.3 VGGNET - As discussed in [35], VGGNet (Visual Geometry Group Network) is well-known for its potent performance in computer vision tasks. In order to extract features and learn complicated features, it stacks 3x3 kernels in convolutional layers. Max-pooling layers help maintain features and strengthen the model by reducing the size of the input volume as shown in Fig 4. Three fully connected layers make up the network for high-level representation. There are two versions of VGGNet: VGG-16 and VGG-19. These versions vary in depth, which affects model correctness and complexity.

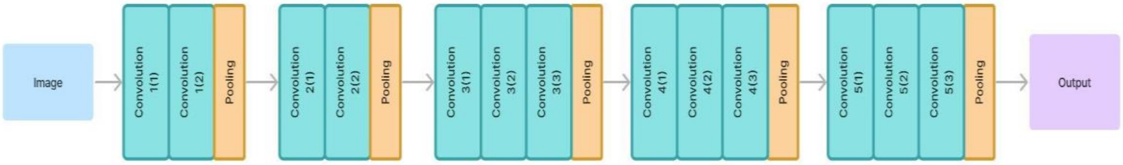

Fig 4: Architecture of VGGNET

3.2 CAPTION GENERATION PHASE

The Caption Generation Phase focuses on translating the visual representations obtained from the Computer Vision Phase into coherent and contextually relevant textual descriptions. This phase uses advanced natural language processing models to synthesize visual information and linguistic structures to produce accurate and descriptive captions.

3.2.1 LSTM - LSTM is an outstanding NLP model as per the paper [33]. One kind of RNN that is particularly good at identifying long-term dependencies in sequential data are LSTM (Long Short-Term Memory) networks. In order to store long-term data, they keep their cell state constant, controlled by input, output, and forget gates (as shown in Fig 5). An output gate regulates information that is exposed, an input gate incorporates fresh information, and a forget gate eliminates unnecessary information. Because of this careful control, long-term dependencies may be captured by LSTM networks, making them perfect for applications such as time series analysis and natural language processing.

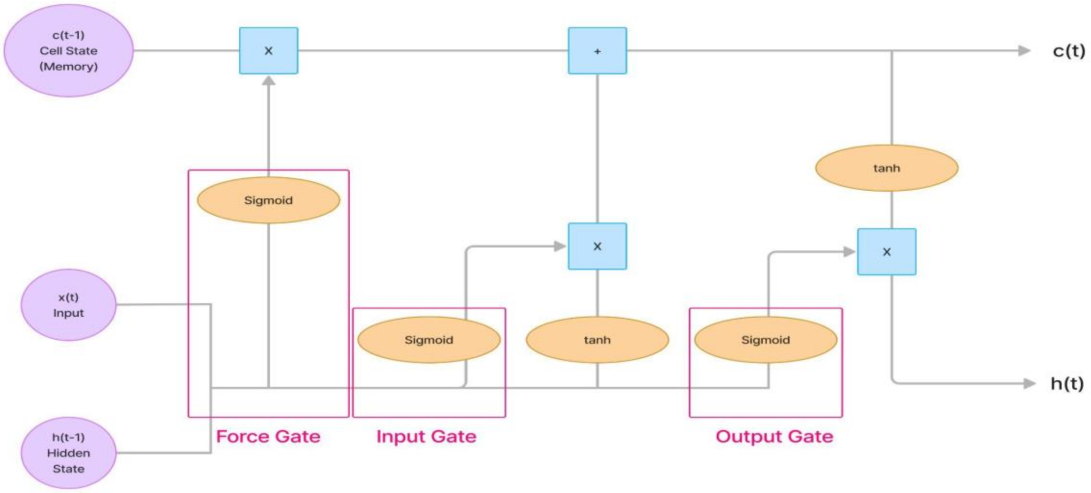

Fig 5: Architecture of LSTM

In contemporary deep learning, long-term dependency management and long-term information retention are made possible by LSTM networks. Because they can recognize complex links among sequential data, they are excellent at comprehending contextual nuances. LSTM networks process data with extended dependencies, advancing domains such as natural language processing, with their forget, input, and output gates. They are therefore essential for complicated activities requiring sequential information processing that is nuanced.

3.2.2 Transformer - Natural language processing (NLP) was transformed by [38] introduction of the transformer model, which included self-attention mechanisms. Transformers can capture dependencies across input sequences in parallel, which greatly improves training and inference efficiency compared to standard sequential models. Transformers are highly effective in capturing long-range dependencies and contextual information that is necessary for tasks such as text summarization and machine translation. They are composed of encoder and decoder components that have several layers of self-attention and feed-forward neural networks as shown in Fig 6. Their outstanding effectiveness and adaptability have established them as a pillar of contemporary NLP, propelling continuous developments in the area.

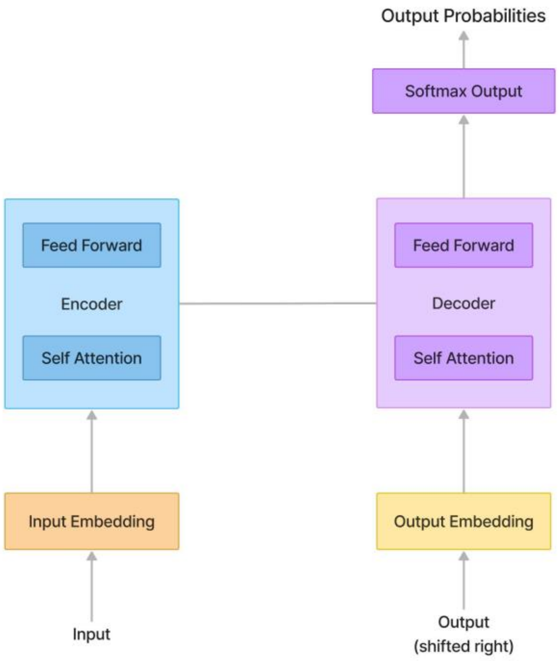

Fig 6: Architecture of Transformer

Transformers provide parallel processing for activities such as language translation by processing input sequences efficiently through the use of self-attention. This technique captures relationships between words at the same time by dynamically weighing word importance. Transformers overcome the limits of sequential processing by using positional encoding to express word order through the use of encoder and decoder components. Feed-forward network integration provides non-linearity to help capture intricate patterns. Transformers use layer normalisation and residual connections to solve problems like vanishing gradients and provide steady performance.

4. EXPERIMENTAL SETTING

The Experimental Setting describes the methodologies and conditions for training and evaluating the image captioning models. This section provides details on the datasets utilized, the evaluation metrics applied to assess model performance, and the overall experimental setup.

4.1 DATASET

The ***Flickr30K*** dataset, referenced from the research paper [1], stands out as an exceptional tool for researchers who want to delve into the study of automatic image description and grounded language understanding. This dataset consists of 30,000 images obtained from Flickr and a vast collection of 158,000 captions that are human-annotated. The dataset provides a diverse and extensive corpus that can be used for training and evaluation purposes. Unlike other datasets, Flickr30K gives researchers the freedom to design their experimental setups as it doesn't impose any fixed split of images for training, testing, and validation. Moreover, the dataset's annotations include detectors for common objects, a color classifier, and a preference towards larger objects, which provide additional contextual information for analysis.

4.2 EVALUATION METRICS

BLEU (Bilingual Evaluation Understudy) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics are used to quantitatively evaluate the accuracy and relevance of the generated captions.

4.2.1 Bleu- An essential tool for evaluating natural language processing activities, particularly machine translation, is the BLEU metric (Bilingual Evaluation Understudy). Using n-gram matching, it assesses the similarities between machine-generated translations and human reference translations by comparing their accuracy and succinctness. The development of algorithms and the improvement of translation models both depend on BLEU. Even with its widespread application, BLEU is not perfect at capturing subtleties such as semantics and context. Nonetheless, it is useful for assessing and enhancing machine-generated text in a variety of NLP and ML fields due to its ease of use and efficacy.

4.2.2 Rouge- Recall-Oriented Understudy for Gisting Evaluation, or ROUGE, is a metric for assessing the comprehensiveness of generated text in automatic text summarization and machine translation systems. ROUGE analyses n-grams and word overlap between system output and reference summaries, in contrast to BLEU, which places more emphasis on precision. This method helps to develop summarization algorithms by allowing for a more nuanced assessment of the algorithms. ROUGE is limited in its ability to assess coherence and fluency, nevertheless. However, the fact that it is so widely used indicates how important it is to the advancement of machine translation and automatic summarization in NLP and ML.

4.3 EXPERIMENTAL SETUP

This section details the specific configurations and conditions under which the image captioning experiments were conducted.

4.3.1 Training process and model optimization - The utilization of the Flickr30K dataset played a crucial role in the experimental setup for the AI image decoding task. It provided a substantial pool of 30,000 samples, each containing five instances accompanied by human commentary, for both training and testing phases. During the training process, the generator function played a vital role in managing the extensive dataset over 50 epochs. This function dynamically regulated the flow of data to the model, thereby optimizing computational efficiency and facilitating iterative refinement. Moreover, the selection of generator tasks enabled the fine-tuning of training dynamics, including adjustments to batch size and implementation of data enhancement strategies, enhancing model generalization and robustness.

4.3.2 Model architecture and hyperparameter tuning - A convolutional neural network (CNN) combined with a recurrent neural network (RNN) was used for the image captioning model. CNN worked as the encoder, extracting visual features from input images through convolutional layers and pooling layers. On the other hand, RNN acted as the decoder, generating captions based on the encoded visual features. Attention mechanisms were added to enhance the model's ability to generate contextually relevant captions. This was done by aligning visual features with corresponding words. The performance of the model was optimized through hyperparameter tuning. This included adjusting the learning rate, batch size, dropout rate, and regularization techniques through grid search or random search methods. By carefully tuning the hyperparameters and designing the architecture, a sophisticated image captioning system was developed. The system is capable of producing accurate and contextually coherent descriptions for diverse images.

5. RESULTS

This section represents the results of the six models on Flickr30K dataset. The training methodology and model optimization strategies played a crucial role in achieving these results, facilitating iterative refinement and enhancing the model's generalization capabilities.

Table 2 below provides a summary of BLEU and ROUGE scores of various models. These outcomes underscore the effectiveness of the approach and lay the groundwork for future research endeavors in automated image captioning. This work holds promise for applications across various domains, from assistive technology to content creation, and has the potential to drive further innovation in human-computer interaction.

Table 2: Results table of various AI captioning models

S NO.	MODEL	BLEU SCORE	ROUGE SCORE
1	CNN + LSTM	0.353	0.457
2	RESNET + LSTM	0.564	0.427
3	VGGNET + LSTM	0.282	0.345
4	CNN + Transformer	0.514	0.447
5	RESNET + Transformer	0.396	**0.527**
6	VGGNET + Transformer	**0.576**	0.454

The graphical analysis depicted in Figures 7 and 8 offers an insightful exploration into the performance dynamics of various models within LSTM and Transformer decoding frameworks. Figure 7 reveals that the ResNet model achieves a remarkable Bleu score of 0.564, highlighting its superior capability in generating captions closely mirroring human references. In contrast, VGGNet demonstrates the lowest performance with a Bleu score of 0.282. Moreover, when considering Rouge scores, CNN emerges as the top performer, securing a Rouge score of 0.457, while VGGNet lags with the lowest score of 0.346, particularly evident within the LSTM decoding context. In Figure 8, which illustrates Transformer decoding results, VGGNet emerges as the standout performer with a notable Bleu score of 0.576, contrasting with ResNet's comparatively lower performance at 0.396. Regarding Rouge scores, ResNet excels with a score of 0.527, outperforming other models, while CNN exhibits the least favorable performance with a Rouge score of 0.448 in the realm of Transformer decoding.

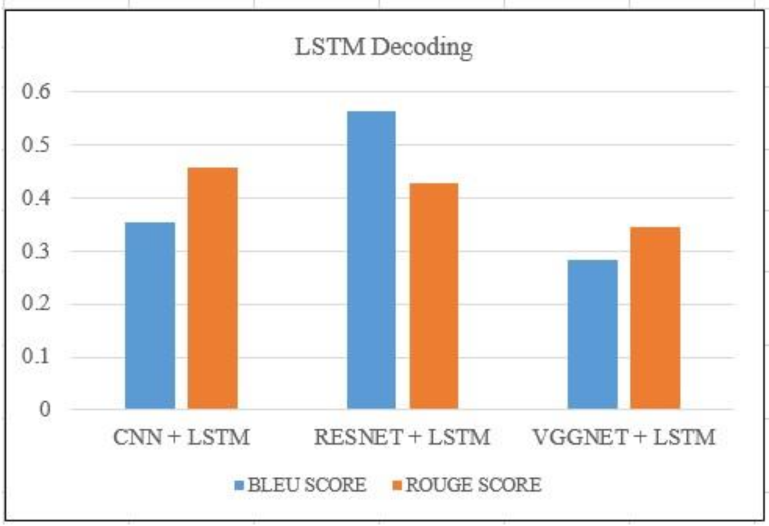

Fig 7: LSTM Decoding Graphical Representation

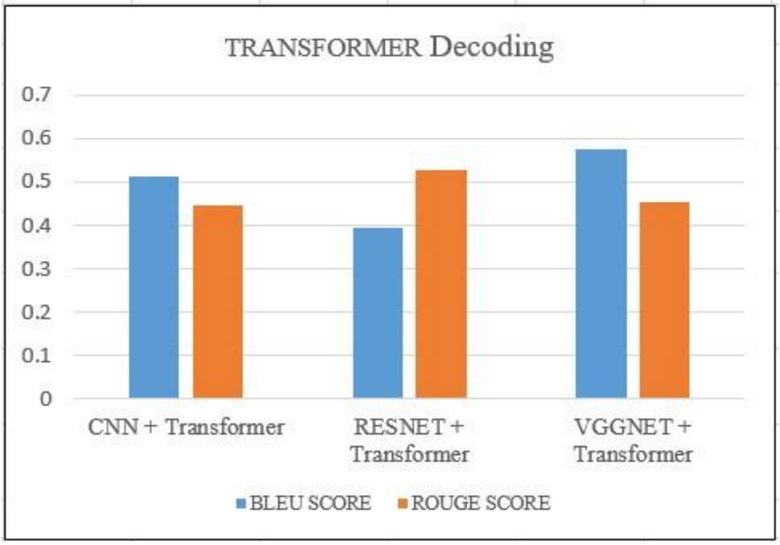

Fig 8: Transformer Decoding Graphical Representation

Given below is the comparison of BLEU and ROUGE scores of different models presented in Table 3, along with the scores in Table 2. The study reveals that the Inception-V3+LSTM model by [31] achieves the highest BLEU score among the examined models, while in [28], avtmNet-based model attains the highest ROUGE score. Whereas, the results in this study shows that the RESNET+LSTM model yields a commendable BLEU score of 0.564 and VGGNET+Transformer with BLEU score of 0.576 , and the RESNET+Transformer model demonstrates a ROUGE score of 0.527. Although the ROUGE score is comparable with those presented in other papers, the BLEU scores of this research stand out favorably.

Table 3: Comparison of AI image captioning models

S No.	PAPER	TECHNIQUE	BLEU SCORE	ROUGE SCORE
1.	Heng Song 2020 [28]	avtmNet	0.331	0.567
2.	Xu et al. 2015 [36]	AlexNet + LSTM	0.243	-
3.	Yao et al. 2017[33]	VGGNet + LSTM	0.326	0.540
4.	Gu et al. 2017 [35]	VGGNet + LSTM	0.300	-

5.	Rennie et al. 2017 [32]	ResNet + LSTM	0.319	0.543
6.	Zhang et al. 2017 [31]	Inception-V3 + LSTM	0.344	0.558
7.	Jin et al. 2015 [37]	VGGNet + LSTM	0.282	-
8.	Mao et al. 2015 [27]	AlexNet, VGGNet + RNN	0.190	-
9.	Our Model	RESNET + Transformer	0.396	0.527
10.	Our Model	VGGNET + Transformer	**0.576**	0.454

A dash (–) in the table indicates results are unavailable.

6. CONCLUSION

In this paper, Transformer based image captioning models and LSTM based image captioning models are presented. The image processing phase includes CNN, ResNet and VGGNet and for the caption generation LSTM and Transformer is utilized. The models are experimented on Flickr30K dataset and for evaluation metric Bleu and Rouge score metric is used. The RESNET+Transformer model performs better than other model with a score of.527 and VGGNET+Transformer gives best Bleu score of .576. In future, novel evaluation metrics and advancing techniques for fine-tuning models on domain-specific data could further improve the performance and applicability of automated image captioning systems. These directions hold promise for pushing the boundaries of human-computer interaction and driving innovation across various domains, from assistive technology to content creation, thereby enriching the capabilities of AI systems to understand and interact with the visual world.

REFERENCES

[1] Hossain, MD Zakir, et al. "A comprehensive survey of deep learning for image captioning." ACM Computing Surveys (CsUR) 51.6 (2019): 1-36.

[2] Yan, Jie, et al. "Exploring better image captioning with grid features." Complex & Intelligent Systems (2024): 1-16.

[3] Rosario, Grant, and David Noever. "PICS: Pipeline for Image Captioning and Search." *arXiv preprint arXiv:2402.10090* (2024).

[4] Jaiswal, Sushma, et al. "Optimized Image Captioning: Hybrid Transformers Vision Transformers and Convolutional Neural Networks: Enhanced with Beam Search."

[5] Yun, Youngsik, and Jihie Kim. "CIC: A framework for Culturally-aware Image Captioning." *arXiv preprint arXiv:2402.05374* (2024).

[6] Liu, Tianrui, et al. "Image Captioning in news report scenario." *arXiv preprint arXiv:2403.16209* (2024).

[7] Ghandi, Taraneh, Hamidreza Pourreza, and Hamidreza Mahyar. "Deep learning approaches on image captioning: A review." *ACM Computing Surveys* 56.3 (2023): 1-39.

[8] Atliha, Viktar. "Improving image captioning methods using machine learning approaches." (2023).

[9] Yu, Boxi, et al. "Automated testing of image captioning systems." *Proceedings of the 31st ACM SIGSOFT International Symposium on Software Testing and Analysis*. 2022.

[10] Wang, Ning, et al. "Controllable image captioning via prompting." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. No. 2. 2023.

[11] Zhong, Zhipeng, Fei Zhou, and Guoping Qiu. "Aesthetically relevant image captioning." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. No. 3. 2023.

[12] Wang, Ning, et al. "Efficient image captioning for edge devices." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. No. 2. 2023.

[13] Yang, Xu, et al. "Exploring diverse in-context configurations for image captioning." *Advances in Neural Information Processing Systems* 36 (2024).

[14] You, Quanzeng, et al. "Image captioning with semantic attention." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[15] Derkar, Shubham Bhaiyaji, et al. "CaptionGenX: Advancements in Deep Learning for Automated Image Captioning." *2023 3rd Asian Conference on Innovation in Technology (ASIANCON)*. IEEE, 2023.

[16] Bhatnagar, Pooja, Sai Mrunaal, and Sachin Kamnure. "Enhancing Image Captioning with Neural Models." *arXiv preprint arXiv:2312.00435* (2023).

[17] Cornia, Marcella, Lorenzo Baraldi, and Rita Cucchiara. "Explaining transformer-based image captioning models: An empirical analysis." *AI Communications* 35.2 (2022): 111-129.

[18] Suresh, K. Revati, Arun Jarapala, and P. V. Sudeep. "Image captioning encoder–decoder models using cnn-rnn architectures: A comparative study." *Circuits, Systems, and Signal Processing* 41.10 (2022): 5719-5742.

[19] Tiwary, Tejal, and Rajendra Prasad Mahapatra. "An accurate generation of image captions for blind people using extended convolutional atom neural network." *Multimedia Tools and Applications* 82.3 (2023): 3801-3830.

[20] Leotta, Maurizio, Fabrizio Mori, and Marina Ribaudo. "Evaluating the effectiveness of automatic image captioning for web accessibility." *Universal access in the information society* 22.4 (2023): 1293-1313.

[21] Jaiswal, Tarun. "Image captioning through cognitive IOT and machine-learning approaches." *Turkish Journal of Computer and Mathematics Education (TURCOMAT)* 12.9 (2021): 333-351.

[22] Nguyen, Khanh, et al. "Show, interpret and tell: entity-aware contextualised image captioning in wikipedia." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. No. 2. 2023.

[23] Han, Seung-Ho, Min-Su Kwon, and Ho-Jin Choi. "EXplainable AI (XAI) approach to image captioning." *The Journal of Engineering* 2020.13 (2020): 589-594.

[24] Fei, Zhengcong, et al. "Uncertainty-aware image captioning." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. No. 1. 2023.

[25] Staniūtė, Raimonda, and Dmitrij Šešok. "A systematic literature review on image captioning." *Applied Sciences* 9.10 (2019): 2024.

[26] Iwamura K, Kasahara JYL, Moro A, Yamashita A, Asama H (2021) Image captioning using motion-CNN with object detection. Sensors 21(4):1–13

[27] Mao, Chunlan, et al. "Review on research achievements of biogas from anaerobic digestion." *Renewable and sustainable energy reviews* 45 (2015): 540-555.

[28] Song H, Zhu J, Jiang Y (2020) avtmNet: adaptive visual-text merging network for image captioning. Comput Electr Eng 84:1–12

[29] Xiao F, Gong X, Zhang Y, Shen Y, Li J, Gao X (2019) DAA: dual LSTMs with adaptive attention for image captioning. Neurocomputing 364:322–329

[30] Deng Z, Jiang Z, Lan R, Huang W, Luo X (2020) Image captioning using dense net network and adaptive attention. Signal Process Image Commun 85:1–9

[31] Li Zhang, Flood Sung, Feng Liu, Tao Xiang, Shaogang Gong, Yongxin Yang, and Timothy M. Hospedales. 2017.Actor-critic sequence training for image captioning. arXiv preprint arXiv:1706.09601.

[32] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17). 1179–1195.

[33] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2017. Incorporating copying mechanism in image captioning for learning novel objects. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17). IEEE, 5263–5271.

[34] Xu, K, Ba, J, Kiros, R, et al.: 'Show, attend and tell: neural image caption generation with visual attention'. Proc. IEEE Int. Conf. Machine Learning (ICML), Jun, 2015, pp. 2048–2057

[35] Jiuxiang Gu, Gang Wang, Jianfei Cai, and Tsuhan Chen. 2017. An empirical study of language CNN for image captioning. In Proceedings of the International Conference on Computer Vision (ICCV'17). 1231–1240.

[36] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In International Conference on Machine Learning. 2048–2057.

[37] Junqi Jin, Kun Fu, Runpeng Cui, Fei Sha, and Changshui Zhang. 2015. Aligning where to see and what to tell: Image caption with region-based attention and scene factorization. arXiv preprint arXiv:1506.06272.

[38] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).