AITA
ARTIFICIAL INTELLIGENCE
THEORY AND APPLICATIONS

■RESEARCH ARTICLE

Report Generation from X-RAY Images: An Evaluation with Transformer Architectures

Bengü Fetiler[a†], **Ömer Atılım Koca** [b], **Volkan Kılıç** [c]

[a] Department of Artificial Intelligence and Machine Learning, Manisa Celal Bayar University, 45120, Yunusemre, Manisa, Türkiye
[b] Department of Computer Engineering, İzmir Bakırçay University, 35665 Menemen, İzmir, Türkiye
[c] Department of Electrical-Electronics Engineering, İzmir Katip Çelebi University, 35620 Çiğli, İzmir, Türkiye
[†] bengu.fetiler@cbu.edu.tr, corresponding author

RECEIVED AUGUST 04, 2025
ACCEPTED SEPTEMBER 24, 2025

CITATION Fetiler, B., Koca, Ö. A. & Kılıç, V. (2025). Report generation from X-RAY images: An evaluation with transformer architectures. *Artificial Intelligence Theory and Applications, 5(*2), 1-10.

Abstract

The automatic generation of medical reports from chest X-RAY images has attracted increasing attention due to its capability to enhance diagnostic accuracy and reduce workload in clinical decision support. The latest advancements in medical report generation, particularly with encoder-decoder models, emphasize their ability to integrate visual information with textual reports. However, these models face several challenges, including the generation of generic statements, the failure to capture detailed pathological findings, and the production of inconsistent reports. In this study, the effectiveness of Vision Transformer and Convolutional Vision Transformer encoders combined with GPT2-based (Generative Pre-trained Transformer) decoders are investigated for the task of chest X-RAY report generation. Their ability to capture radiological findings and generate clinically meaningful reports is evaluated through comparative analyses conducted under diverse experimental configurations on IU X-RAY (Indiana University X-RAY) dataset. Experimental results on the IU X-RAY dataset demonstrated that the ViT-GPT2 model achieved superior performance, with a BLEU-1 score of 0.356, a METEOR score of 0.171, and a CIDEr score of 0.374, outperforming CNN–RNN baselines. These results confirm the potential of transformer-based models to generate clinically meaningful and linguistically coherent radiology reports.

Keywords: convolutional vision transformer, image processing, medical report generation, natural language processing, radiography, vision transformer

1. Introduction

Medical report generation is a significant task at the intersection of natural language processing and computer vision. Its main objective is to generate diagnostic reports of medical images, such as radiological images. This approach shows potential for reducing the workload on radiologists, improving diagnostic consistency, and enabling timely decision-making [1-3]. Traditional deep learning-based approaches primarily adopt encoder-decoder frameworks, often utilizing Convolutional Neural Network (CNN) to extract spatial information and Recurrent Neural Network (RNN) [4-11] or Transformers [12-17] for generation of textual descriptions. For instance, a study in the literature employed a VGG19-based (Visual Geometry Group 19 Layer) CNN encoder

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than AITA must be honoured. Abstracting with credit is permitted, and providing the material is not used for any commercial purposes and is shared in its entire and unmodified form. Request permissions from info@aitajournal.com

Artificial Intelligence Theory and Applications, ISSN: 2757-9778. ISBN: 978-605-69730-2-4 © 2025 İzmir Bakırçay University

combined with a hierarchical LSTM (Long-Short Term Memory) decoder and a co-attention mechanism to jointly attend to visual and semantic features [18]. In a recent approach, a multi-attention mechanism was integrated with one-shot pruning to reduce model complexity; however, its clinical coherence remained limited [19]. Another study incorporated hybrid reinforcement learning rewards and multi-linear attention mechanisms, aiming to improve the quality of generated reports [20]. However, these methods are often constrained by insufficient cross-modal alignment, limited capacity to capture detailed pathological findings, and an excessive reliance on recurrent normal patterns.

In this study, the model performances of a Vision Transformer (ViT) and a Convolutional Vision Transformer (CVT) encoder are evaluated for the task of automatic generation of medical reports from chest X-RAY images. Transformer-based models were selected for this study because they can effectively capture global dependencies in medical images and generate fluent, semantically consistent text. Their ability to integrate global and local patterns makes them particularly advantageous for radiological report generation compared to conventional CNN–RNN approaches [21-23]. Each encoder is combined with a GPT2 decoder to generate the medical reports. The ViT leverages pure self-attention mechanisms to model long-range dependencies across the entire image. It enables effective global feature extraction that has demonstrated strong performance in vision-language tasks compared to conventional convolution-based encoders. However, the CVT combines local visual sensitivity with global contextual modelling, which has demonstrated advantages over traditional CNNs in multiple vision–language tasks. The GPT2 (Generative Pre-trained Transformer), fine-tuned for medical domain generation, ensures fluent and coherent narrative output. Our framework is trained and evaluated on the IU X-RAY dataset, which includes a large corpus of chest X-RAY images paired with expert-annotated radiology reports. Experimental results demonstrate that our framework outperforms CNN-RNN methods in performance metrics (BLEU-n (Bilingual Evaluation Understudy), METEOR (Metric for Evaluation of Translation with Explicit Ordering), ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation), CIDEr (Consensus-based Image Description Evaluation)).

The rest of this paper is organized as follows: Section 2 presents the proposed approach, together with the datasets and performance metrics employed. A comprehensive examination and interpretation of the experimental findings are presented in Section 3, while Section 4 concludes the study.

2. Method

This section introduces an end-to-end encoder-decoder framework that automatically generates radiology reports from radiological images. The framework utilizes a ViT or a CVT encoder and a pre-trained GPT2 decoder to create clinically meaningful and coherent reports.

2.1. Vision Transformer-Based Encoder

Figure 1 illustrates the encoder module utilizing a ViT [24], where a transformer architecture is applied directly to image patches to achieve global receptive field modelling. Input chest X-RAY images are resized and divided into fixed, non-overlapping patches. Each patch is linearly embedded and augmented with positional encodings to maintain spatial context. A sequence of transformer encoder layers

processes tokens generated from embedded image patches that encode spatial information. Each layer comprises Multi-Head Self-Attention and Feed-Forward Network blocks. A key advantage of ViT over traditional convolutional encoders is its ability to directly capture long-range dependencies and global structural relationships via self-attention. This is especially beneficial for medical images, where pathological patterns need to be identified across the entire image. The encoder produces a sequence of high-dimensional tokens that comprise the global semantic and spatial information extracted from the X-RAY, preparing them for the decoder.

2.2. Convolutional Vision Transformer-Based Encoder

In the encoder module, a CVT [25] architecture, which integrates convolutional layers with transformer-based attention mechanisms to capture both local and global features,

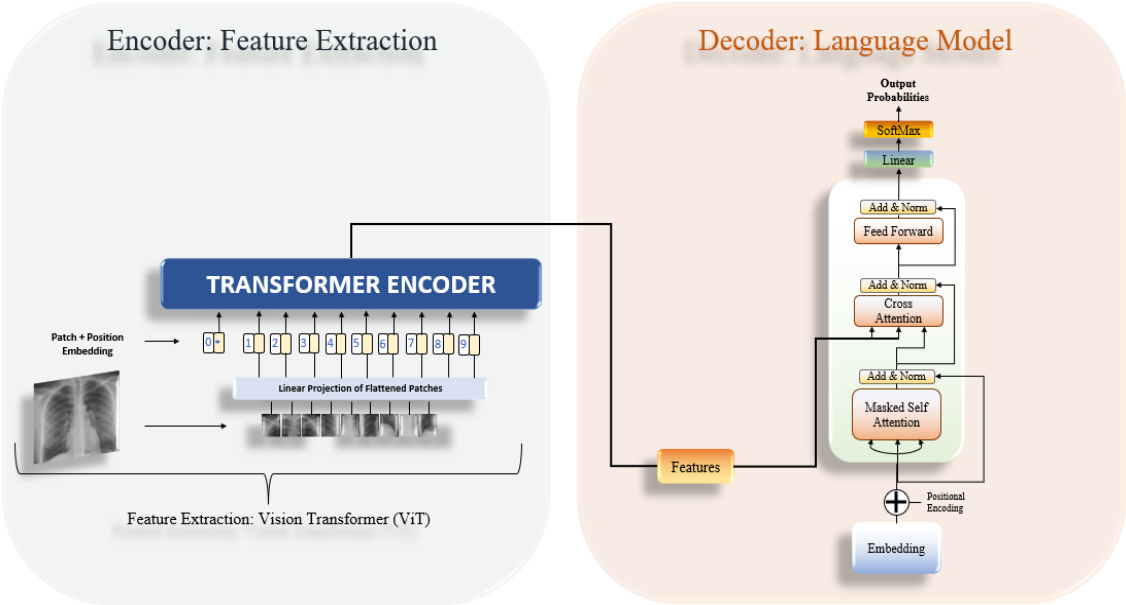


Figure 1. Proposed Framework.

was incorporated. Here, input chest X-RAY images are resized and then processed through multiple convolutional layers to capture basic spatial patterns, while maintaining local information. These convolutional layers serve as a front-end feature extractor, generating a dense feature map that retains spatial resolution and fine-grained visual cues. The resulting feature maps are then partitioned into patch tokens, which are linearly projected and supplemented with positional encodings to preserve the spatial arrangement. These tokens are fed into a transformer encoder stack, where each layer comprises Multi-Head Self-Attention and Feed-Forward Network components. The incorporation of convolutional operations before the transformer blocks introduces valuable inductive biases, particularly translation equivariance and spatial locality. These properties are particularly advantageous for medical imaging, where fine-grained pathological features are often clustered in specific regions of the image. The hybrid architecture of CVT enables the encoder to model short-range texture patterns using convolution, while simultaneously capturing long-range dependencies through self-attention. Finally, the encoder generates a sequence of high-dimensional tokens that encode comprehensive semantic and spatial information, which are then passed to the decoder for report generation.

2.3. GPT-Based Large Language Model

The decoder in our framework leverages pre-trained GPT2 [26] architecture, an autoregressive language model recognized for generating coherent and contextually accurate text. The visual tokens from the ViT encoder are integrated into the GPT2 decoder using cross-attention mechanisms. This enables the decoder to generate text conditioned on both previously generated words and visual features. The decoder sequentially generates the report, producing tokens from left to right. In our experiments, report generation is performed using greedy decoding and beam search to analyze the effects of different decoding strategies on report quality. Greedy decoding is computationally efficient because it selects the token with the highest likelihood at each step. However, this strategy may lead to suboptimal sequences since it does not consider alternative candidate tokens. Conversely, beam search keeps multiple possible hypotheses ("beams") at each step, evaluating various candidate paths to select the most probable complete sequence. This enables the model to explore various phrasings

Table 1. Comparison of different transformer-based models.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-	CIDEr
VIT-GPT2	0.356	0.219	0.148	**0.104**	**0.171**	0.261	0.314
CVT-GPT2	**0.362**	**0.222**	**0.149**	**0.104**	0.163	0.271	0.241
VIT-GPT2 (beam search)	0.301	0.191	0.133	0.097	0.146	**0.274**	**0.374**
CVT-GPT2 (beam search)	0.355	0.212	0.147	0.103	0.164	0.270	0.251

and sentence structures, often resulting in semantically more detailed reports and higher evaluation scores. In this study, GPT-2 was used in its general pre-trained form without additional medical domain adaptation. This decision was made to focus primarily on assessing the impact of transformer-based encoders on report generation performance. Accordingly, the experimental setup was configured to provide reliable and comparable results. All transformer-based models were trained for 100 epochs on an NVIDIA GeForce RTX 3090 GPU (24 GB), ensuring stable convergence and reliable performance.

3. Experimental Evaluations

3.1. Dataset and Evaluation Metrics

In this study, IU X-RAY [27] dataset is utilized for medical report generation tasks. The dataset comprises 7,470 chest X-RAY images (both lateral and frontal views) linked to 3,955 corresponding radiology reports. These were collected from the radiology department at Indiana University Hospital. Each report typically includes structured sections such as Findings and Impression, providing concise descriptions of diagnostic observations and clinical interpretations. Several pre-processing steps were conducted to ensure consistency and clinical relevance in our modelling. Samples from patients younger than 16 years were excluded to ensure an exclusive focus on adult cases and

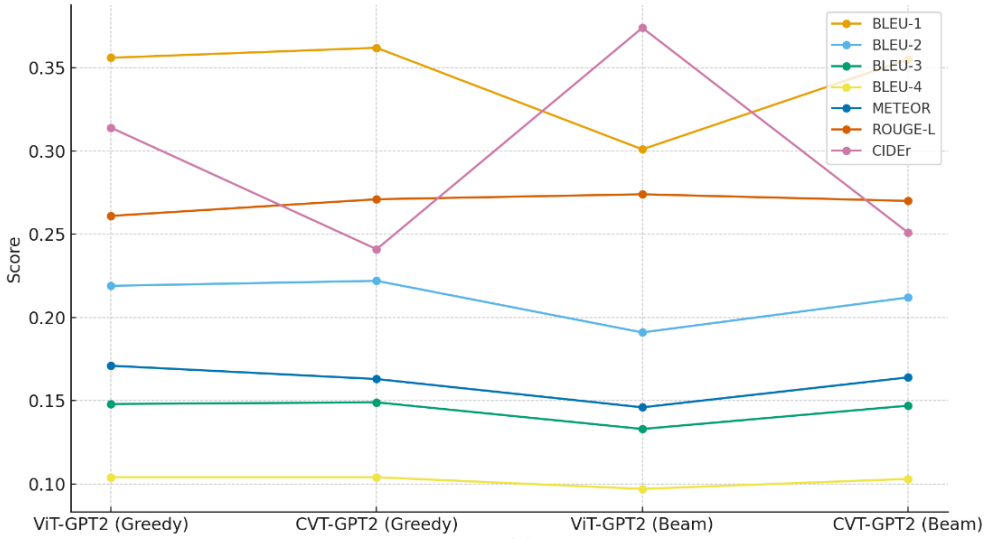


Figure 2. Line chart illustrating model performance comparison (ViT-GPT2 vs. CVT-GPT2; greedy vs. beam search)

to facilitate a more accurate characterization of adult thoracic anatomy and pathology. This approach reduces the variability introduced by pediatric cases, which often differ in anatomical proportions and disease presentation. Furthermore, the original reports were examined and cleaned to remove references to prior examinations or temporal comparisons, such as sentences including phrases like "... compared to the previous study". This step prevents the model from relying on unavailable contextual information and ensures that reports are generated only from the current X-RAY image. In addition, reports were standardized standardized by focusing exclusively the Findings and Impression sections, which contain the most clinically relevant information. Low-quality or corrupted image files were also excluded from the dataset. In addition, the dataset was also reviewed by a radiologist, and incomplete or erroneous records were removed to ensure clinically reliable model inputs prior to training. The models were trained and evaluated on a filtered and standardized version of the IU X-RAY dataset. This setup was designed to highlight image conditioning, temporal independence, and clinically coherent report generation.

The study evaluates the quality and clinical relevance of the automatically generated radiology reports using widely adopted metrics in image-to-text and medical report generation tasks. These include BLEU-n (with n-gram levels from 1 to 4) [28], ROUGE-L [29], METEOR [30], and CIDEr [31]. Each metric captures different linguistic or semantic characteristics of the generated content when compared to the reference reports. BLEU-n quantifies the rate of n-gram overlap between generated and reference reports. While BLEU-1 reflects unigram-level similarity, BLEU-4 incorporates higher-order n-grams, offering a more comprehensive measure of fluency and syntactic accuracy. In contrast to BLEU, METEOR improves the evaluation by incorporating recall in addition to precision, as well as synonym matching. This makes it more suitable for capturing semantically similar phrases that may not match exactly at the lexical level. Similarly, ROUGE-L measures the longest common subsequence between the generated and ground-truth reports. This enables evaluation of their sentence-level structural similarity and narrative coherence. CIDEr is a metric used to evaluate the similarity between generated and reference captions for image captioning. It achieves this by measuring TF-IDF (Term Frequency-Inverse Document Frequency) weighted n-

gram overlap, which effectively captures both the relevance of the generated content and its alignment with human descriptions. Together, these metrics offer a comprehensive evaluation framework for assessing both the linguistic quality and the diagnostic accuracy of the generated radiology reports.

Table 2. Comparison of proposed framework with latest studies.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
[17]	0.387	0.245	0.166	0.111	0.164	0.257
[32]	0.361	0.226	0.152	0.106	-	0.187
[33]	0.438	0.298	0.208	**0.151**	–	0.343
[34]	**0.476**	**0.340**	**0.238**	–	–	0.297
VIT-GPT2 (Ours)	0.356	0.219	0.148	0.104	**0.171**	0.314
CVT-GPT2 (Ours)	0.362	0.222	0.149	0.104	0.163	0.241
VIT-GPT2 (beam s., Ours)	0.301	0.191	0.133	0.097	0.146	**0.374**
CVT-GPT2 (beam s., Ours)	0.355	0.212	0.147	0.103	0.164	0.251

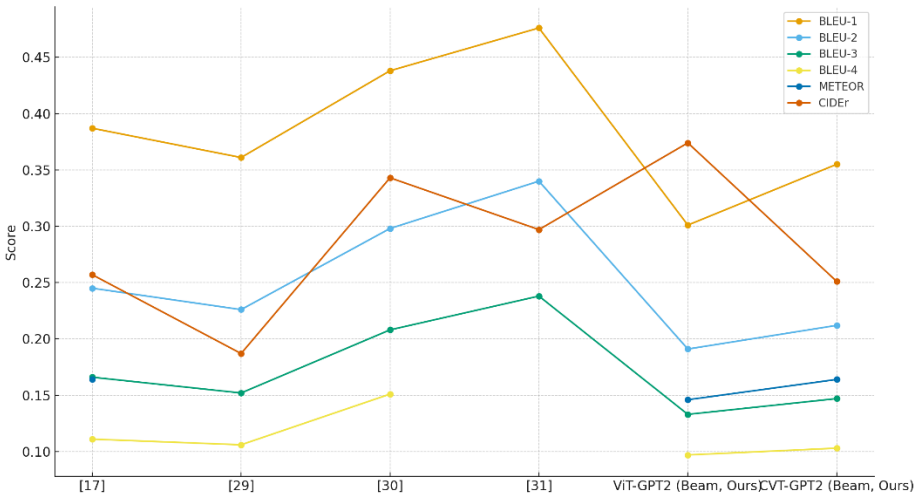


Figure 3. Line chart comparing the proposed models with state-of-the-art baselines on the IU X-RAY dataset (BLEU-n, METEOR, CIDEr).

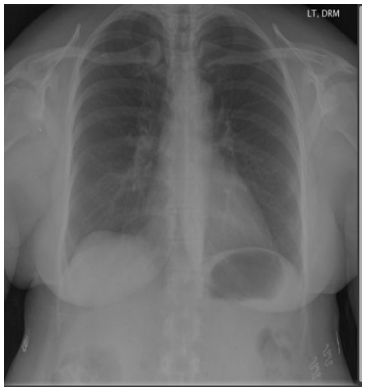
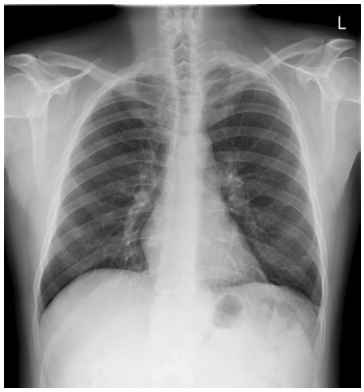
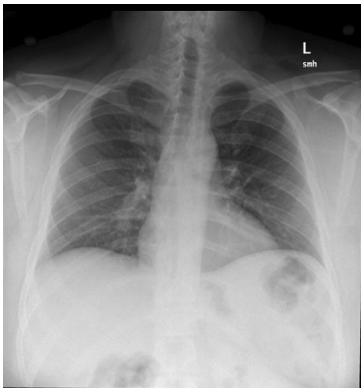
3.2. Result and Discussion

Table 1 presents the performance of four Transformer-based encoder-decoder configurations (ViT-GPT2, CVT-GPT2, and their beam search variants) on the IU X-RAY dataset. Besides, Figure 2 demonstrates a line chart comparison of ViT-GPT2 and CVT-GPT2 models across BLEU-n, METEOR, ROUGE-L, and CIDEr, offering a clearer visualization of performance metrics.The study reports model performance using BLEU-n (with n-gram levels from 1 to 4), METEOR, ROUGE-L, and CIDEr metrics. Among the greedy decoding approaches, CVT-GPT2 achieved the best BLEU-1 (0.362), BLEU-2 (0.222), and BLEU-3 (0.149) scores, suggesting that the convolutional inductive biases in the CVT encoder contributed to improved local and mid-range n-gram consistency in the generated text. However, ViT-GPT2 outperformed CVT-GPT2 in METEOR (0.171) and CIDEr (0.314), indicating higher linguistic quality and more robust semantic

consistency with the reference reports. When beam search was applied, ViT-GPT2 (beam) achieved the highest CIDEr score overall (0.374), demonstrating its ability to generate more diverse and semantically rich reports. This improvement supports the hypothesis that the global receptive field of the ViT encoder, when paired with a broader decoding search space, enhances the capacity of the model to capture pathological patterns and translate them into clinically meaningful language.

Table 2 compares our models against several state-of-the-art baselines from literature on the IU X-RAY dataset. In addition, Figure 3 illustrates a line chart comparing the proposed models with state-of-the-art baselines on the IU X-RAY dataset, providing an intuitive overview of relative strengths and weaknesses. Notably, ViT-GPT2 (beam) achieved the highest CIDEr score (0.374), surpassing all baseline models, and also demonstrated superior performance in METEOR (0.146), indicating its strength in generating lexically diverse and semantically rich radiology reports. While the model shown in the fourth row of Table 2 achieved the best BLEU-n scores among the baselines (BLEU-1: 0.476, BLEU-3: 0.238). However, these models declined in CIDEr and METEOR, metrics that better reflect clinical relevance and information density. In contrast, ViT-GPT2 and CVT-GPT2 models produced more informative and nuanced outputs, as reflected in their higher CIDEr values (0.314 and 0.241 respectively), demonstrating the effectiveness of combining Transformer-based vision encoders with GPT2-style language decoders. These results validate the advantage of leveraging ViT

Table 3. Comparison of generated and reference radiology reports with ViT-GPT2 (beam search) from the IU X-RAY.

		
Reference Report: The heart is within normal size limits. No significant abnormalities are noted in the mediastinum. The lungs appear underinflated but remain clear. No evidence of pleural effusion is present.	**Reference Report:** The size of the heart and the pulmonary vessels are considered normal range. No focal airspace abnormalities are noted in the lungs. There is no signs of pleural effusion and pneumothorax. Stable low left paraspinal / retrocrural lymphadenopathy is observed without interval change	**Reference Report:** Cardiac silhouette and pulmonary blood flow appear normal. There is no evidence of localized infiltrates, pneumothorax, or pleural fluid accumulation.
Generated Report: The heart is normal. There is no consolidation or pneumothorax. There is no	**Generated Report:** The heart is within the normal size range. There is no mediastinal widening. The lungs are completely clear. No large pleural effusion and pneumothorax is seen.	**Generated Report:** The heart maintains a normal size. No notable abnormalities are seen in the mediastinum. Lung fields appear clear.

and CVT backbones together with robust decoders that closely match the reference reports. They also confirm the effectiveness of carefully designed models and beam

search decoding strategies, particularly in capturing complex medical semantics and improving alignment with radiologist-authored references. A comparative analysis of reference radiology reports and automatically generated reports for three sample of chest X-RAY cases is illustrated in
Table 3. In the first sample, both the reference and generated reports correctly describe normal cardiac and mediastinal contours, along with clear and expanded lungs. The generated sentence is ending with “There is no”. This indicates a failure in coherence and completeness of the generation. The second sample illustrates capacity of the model to identify and report common thoracic structures accurately. The generated report correctly notes the absence of pleural effusion, pneumothorax, and mediastinal widening, and it corresponds well to the reference in describing the heart and lungs as normal. The report does not capture the reference’s note on the presence of low left paraspinal/retrocrural adenopathy. The third example demonstrates a generated report identifying normal heart size, an unremarkable mediastinum, and clear lungs.

4. Conclusion

This study investigated the effectiveness of Vision Transformer (ViT) and Convolutional Vision Transformer (CVT) encoders combined with a GPT2-based decoder for automatic radiological report generation. Extensive experiments on the IU X-RAY dataset demonstrated that the proposed framework is capable of producing reports that are clinically meaningful and linguistically coherent, outperforming conventional CNN–RNN baselines. In particular, the ViT-GPT2 configuration achieved markedly improved performance in terms of semantic alignment and narrative fluency, reflecting the advantage of global receptive field modeling for medical image analysis. The comparative evaluation of decoding strategies further underscored the significance of inference mechanisms. While greedy decoding provided computational simplicity and faster inference, beam search yielded more diverse and semantically rich reports, aligning more closely with expert-authored references. Evaluation across widely adopted metrics (BLEU-n, METEOR, ROUGE-L, and CIDEr) confirmed the robustness of the proposed approach, with CIDEr emerging as the most representative indicator of human-aligned performance due to its ability to capture information density and clinical relevance. The results show the advantages of transformer-based encoders and decoders, which leverage self-attention mechanisms to achieve more powerful semantic alignment and clinically meaningful narrative generation. Overall, the findings demonstrate that Transformer-based vision encoders, when integrated with large generative language models such as GPT2, present considerable potential for advancing the automation of radiological interpretation. This study not only demonstrates the feasibility of transformer-based approaches for radiology report generation but also provides a basis for future work of the study. Future research will focus on expanding to larger and multi-modal datasets, adapting GPT-2 to medical terminology through vocabulary fine-tuning, incorporating reinforcement learning with expert feedback, and conducting extensive clinical validation studies to further enhance diagnostic reliability and clinical integration.

Declarations

Ethical Consideration

This study did not require formal ethical approval because it relied exclusively on publicly accessible, open-source data.

Competing interests

The authors declare that no competing financial interests or personal relationships exist that could have appeared to influence the work reported in this paper.

Funding

This research was supported by the Scientific Research Projects Coordination Unit of Izmir Katip Celebi University (No: 2025-TYL-FEBE-0006).

References

[1] K. Belikova, O. Y. Rogov, A. Rybakov, M. V. Maslov, and D. V. J. S. R. Dylov, "Deep negative volume segmentation," vol. 11, no. 1, p. 16292, 2021.

[2] M. Gurgitano *et al.*, "Interventional Radiology ex-machina: Impact of Artificial Intelligence on practice," vol. 126, no. 7, pp. 998-1006, 2021.

[3] M. Sermesant, H. Delingette, H. Cochet, P. Jaïs, and N. J. N. R. C. Ayache, "Applications of artificial intelligence in cardiovascular imaging," vol. 18, no. 8, pp. 600-609, 2021.

[4] Ö. Çayli, X. Liu, V. Kiliç, and W. Wang, "Knowledge distillation for efficient audio-visual video captioning," in *2023 31st European Signal Processing Conference (EUSIPCO)*, 2023, pp. 745-749: IEEE.

[5] B. Fetiler, Ö. Çaylı, V. J. E. J. o. S. Kılıç, and Technology, "Leveraging Pre-trained 3D-CNNs for Video Captioning," no. 53, pp. 58-63, 2024.

[6] B. Fetiler, Ö. Çaylı, Ö. T. Moral, V. Kılıç, and A. J. A. B. v. T. D. Onan, "Video captioning based on multi-layer gated recurrent unit for smartphones," no. 32, pp. 221-226, 2021.

[7] V. J. S. U. J. o. C. Kılıç and I. Sciences, "Deep gated recurrent unit for smartphone-based image captioning," vol. 4, no. 2, pp. 181-191, 2021.

[8] Ö. A. Koca, H. Ö. Kabak, and V. J. T. J. o. S. Kılıç, "Attention-based multilayer GRU decoder for on-site glucose prediction on smartphone," vol. 80, no. 17, pp. 25616-25639, 2024.

[9] Ö. A. Koca, A. Türköz, and V. J. A. B. v. T. D. Kılıç, "Tip 1 Diyabette Çok Katmanlı GRU Tabanlı Glikoz Tahmini," no. 52, pp. 80-86, 2023.

[10] Ö. A. Koca and V. J. A. B. v. T. D. Kılıç, "Multi-Parametric Glucose Prediction Using Multi-Layer LSTM," no. 52, pp. 169-175, 2023.

[11] Ö. T. Moral, V. Kiliç, A. Onan, and W. Wang, "Automated Image Captioning with Multi-layer Gated Recurrent Unit," in *2022 30th European Signal Processing Conference (EUSIPCO)*, 2022, pp. 1160-1164: IEEE.

[12] Ö. Çaylı, V. Kılıç, A. Onan, and W. Wang, "Auxiliary classifier based residual rnn for image captioning," in *2022 30th European Signal Processing Conference (EUSIPCO)*, 2022, pp. 1126-1130: IEEE.

[13] Ö. Çaylı, B. Makav, V. Kılıç, and A. Onan, "Mobile application based automatic caption generation for visually impaired," in *International Conference on Intelligent and Fuzzy Systems*, 2020, pp. 1532-1539: Springer.

[14] X. Liu *et al.*, "Visually-aware audio captioning with adaptive audio-visual attention," 2022.

[15] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 375-383.

[16] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156-3164.

[17] O. Alfarghaly, R. Khaled, A. Elkorany, M. Helal, and A. J. I. i. M. U. Fahmy, "Automated radiology report generation using conditioned transformers," vol. 24, p. 100557, 2021.

[18] N. Kaur, A. J. C. i. B. Mittal, and Medicine, "CADxReport: Chest x-ray report generation using co-attention mechanism and reinforcement learning," vol. 145, p. 105498, 2022.

[19] N. Kaur, A. J. J. o. a. i. Mittal, and h. computing, "CheXPrune: sparse chest X-ray report generation model using multi-attention and one-shot global pruning," vol. 14, no. 6, pp. 7485-7497, 2023.

[20] Z. Xu *et al.*, "Hybrid reinforced medical report generation with m-linear attention and repetition penalty," 2023.
[21] S. J. a. p. a. Singh, "Clinical context-aware radiology report generation from medical images using transformers," 2024.
[22] Q. Pu, Z. Xi, S. Yin, Z. Zhao, and L. J. B. e. o. Zhao, "Advantages of transformer and its application for medical image segmentation: a survey," vol. 23, no. 1, p. 14, 2024.
[23] F. Shamshad *et al.*, "Transformers in medical imaging: A survey," vol. 88, p. 102802, 2023.
[24] A. Vaswani *et al.*, "Attention is all you need," vol. 30, 2017.
[25] H. Wu *et al.*, "Cvt: Introducing convolutions to vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 22-31.
[26] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. J. O. b. Sutskever, "Language models are unsupervised multitask learners," vol. 1, no. 8, p. 9, 2019.
[27] M. Li, R. Liu, F. Wang, X. Chang, and X. J. W. W. W. Liang, "Auxiliary signal-guided knowledge encoder-decoder for medical report generation," vol. 26, no. 1, pp. 253-270, 2023.
[28] 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.
[29] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74-81.
[30] S. Banerjee and A. Lavie, "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*, 2005, pp. 65-72.
[31] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566-4575.
[32] P. Pino, D. Parra, P. Messina, C. Besa, and S. J. a. p. a. Uribe, "Inspecting state of the art performance and NLP metrics in image-based medical report generation," 2020.
[33] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Knowledge-driven encode, retrieve, paraphrase for medical image report generation," in *Proceedings of the AAAI conference on artificial intelligence*, 2019, vol. 33, no. 01, pp. 6666-6673.
[34] X. Huang, F. Yan, W. Xu, and M. J. I. A. Li, "Multi-attention and incorporating background information model for chest x-ray image report generation," vol. 7, pp. 154808-154817, 2019.