IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.60

Automated Radiology Report Generation Using Vision-Language Models

Sandhiya.N
Department of Computer Science and Engineering,
Sengunthar Engineering College (Autonomous), Tiruchengode, India
sandhiyanagaraj2608@gmail.com
Ashok Kumar.K
Department of Computer Science and Engineering,
Sengunthar Engineering College (Autonomous), Tiruchengode, India

Publication History
Manuscript Reference: IRJCS/RS/Vol.13/Issue03/CSMR26.MRCS10139
Research Article | Open Access | Double-Blind Peer Reviewed Article ID: IRJCS/RS/Vol.13/Issue03/CSMR26.MRCS10139
Received: 30, January 2026, Revised: 13, February 2026, Accepted: 28 February 2026 Published Online: 25 March 2026
https://www.irjcs.com/volumes/Vol13/iss-03/60.CSMR26.MRCS10139.pdf
Article Citation: Sandhiya,Ashok(2026),Automated Radiology Report Generation Using Vision-Language Models. IRJCS:International Research Journal of Computer Science, Volume 13,Issue 03 of 2026 pages 444-449
Doi:> https://doi.org/10.26562/irjcs.2026.v1303.60
BibTeX Key Sandhiya@2026Automated Orcid: https://orcid.org/0009-0004-9398-7488
IRJCS papers should be cited as IRJCS (International Research Journal of Computer Science, AM Publications, India 2026, ISSN 2393-9842, https://doi.org/10.26562/ijrcs.2025.v1303.60 The journal's official abbreviation is IRJCS.

Abstract: Developing and evaluating an automatic radiology report generation system involves designing a model that can process clinical images and produce complete, clinically oriented radiology reports using advanced vision language techniques. Clinical imaging modalities such as chestradiographs, CT scans, and MRI studies are fed into a frame work that integrates a deep visual encoder with a transformer-based language decoder, enabling the system to translate visual features into structured narratives resembling standard radiology reporting practices. These narratives include detailed descriptions of findings along with concise impressions. The system is fine-tuned using publicly available, de- identified datasets containing paired medical images and expert-written reports, allowing it to learn visual patterns, medical terminology, and reporting conventions across abroad range of scenarios from normal examinations to complex pathologies. The quality of the generated reports is assessed through qualitative review by radiologists and clinically trained evaluators, focusing on narrative clarity, coherence, professional tone, completeness of important findings, mention of relevant comparisons or follow-up suggestions, and clinical relevance, including accuracy of interpretations and appropriateness of certainty levels. Through this evaluation, the analysis captures strengths such as fluent language generation, standardized structure, rapid draft creation, and strong performance on common or straightforward cases, as well as limitations including missed subtle findings, occasional hallucinated details, repetitive phrasing, and reduced reliability in rare or highly complex conditions. The outcome provides a clear understanding of where AI-generated reports can safely function as assistive drafts that enhance efficiency and communication, and where strong human oversight remains essential, guiding best-practice recommendations for integrating generative models into real-world radiology workflows.
Keywords: Vision transformers, language models, radiology report, decoder

1. INTRODUCTION

Radiology is an essential diagnostic discipline that supports the majority of clinical decision-making processes across modern healthcare systems. As patient volumes increase and imaging modalities continue to advance in resolution and complexity, radiologists face significant pressure to interpret large numbers of studies with high accuracy and limited turnaround time. Traditional radiology reporting remains a fully manual, cognitively demanding process wherein the radiologist carefully examines imaging data, identifies abnormalities, describes anatomical structures, synthesizes differential diagnoses, and communicates findings through a structured narrative report. Although effective, this process is time-consuming and subject to inter-reader variability, stylistic inconsistency, and potential human fatigue. These challenges have motivated research into automated approaches to assist in generating standardised, clinically coherent radiology reports.

INTRODUCTION TO AUTOMATED RADIOLOGY REPORT GENERATION (ARRG)

Automated Radiology Report Generation (ARRG) represents an emerging frontier in medical AI, where deep learning systems are designed not only to interpret medical images but also to produce full narrative reports similar to those written by radiologists. Traditional computer-aided diagnosis (CAD) tools mainly provide simple outputs such as detection of a specific disease or classification of an image as normal or abnormal. However, modern clinical practice requires detailed, structured documentation that communicates subtle findings, compares changes across time, and supports clinical decision-making. ARRG aims to bridge this gap by generating high-quality, human-like reports directly from medical images. This is achieved using advanced Vision–Language Models (VLMs) that combine visual feature extraction with natural-language generation capabilities.

IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.60

As hospitals generate millions of radiological images every year, ARRG systems have the potential to significantly reduce radiologists' workload, minimize reporting delays, and improve access to diagnostic services in resource-limited settings. The technology can serve as a powerful assistive tool, offering draft reports that radiologists can review and refine, thereby improving efficiency without compromising diagnostic safety. This section highlights the growing relevance of ARRG in the evolution of AI-assisted medical diagnosis.

VISION–LANGUAGE MODELS FOR RADIOLOGY: CAPABILITIES AND CHALLENGES

Vision–Language Models (VLMs) represent a transformative advancement in AI, capable of linking visual understanding with natural-language generation. In radiology, VLMs use powerful image encoders such as CNNs or Vision Transformers to extract detailed features from X-rays, CT scans, or MRI images. These features are then passed into a language decoder, often based on Transformer architectures, which generates structured textual descriptions. This allows the model to perform tasks that mirror radiologist behavior, including describing anatomical structures, identifying abnormalities, and summarizing impressions. Despite their potential, VLMs in radiology also face significant challenges. Medical images often contain subtle patterns that require expert-level interpretation, making them harder to learn compared to natural images. Additionally, public datasets may contain incomplete or noisy reports, limiting the accuracy of supervised learning. Rare diseases and complex multi-pathology cases are especially challenging because they are underrepresented in training data. Another concern is hallucination where the model generates findings not present in the image. Furthermore, ensuring clinical reliability requires extensive validation and rigorous human over sight. Thus, while VLMs offer sophisticated capabilities, their responsible use in radiology demands careful attention to safety, bias, interpretability, and robustness.

RESEARCH MOTIVATION AND OBJECTIVES OF THE STUDY

The primary motivation behind this study is the growing demand for efficient, accurate, and scalable radiology reporting solutions. With increasing imaging volumes and a shortage of radiologist in many regions, healthcare systems face significant bottle necks in timely report generation. Automated Radiology Report Generation (ARRG) offers a promising pathway to address these challenges by providing AI-generated draft reports that radiologists can refine. This reduces workload, speeds up reporting, and supports clinicaldecision-making.Theobjectiveofthisresearchistodevelopanend-to-endVLM- based ARRG system that processes medical images and produces full, structured radiology reports. Additionally, the study aims to evaluate the system using a qualitative framework that goes beyond traditional lexical metrics. Key research goals include assessing the narrative fluency of generated reports, measuring completeness of findings, and determining clinical relevance through expert review. The study also seeks to identify specific strengths such as performance on normal cases and weaknesses, such as difficulty with complex pathologies. These insights will inform guidelines for integrating ARRG systems into real-world workflows, ensuring they function as reliable assistive tools rather than standalone diagnostic systems. Ultimately, the research contributes to the safe and effective adoption of AI-driven reporting technologies in modern radiology.

2. LITERATU REREVIEW

Radiology Report Generation (RRG) is a rapidly evolving field at the intersection of computer vision and natural language processing. Recent advancements in Multimodal Large Language Models (MLLMs) have significantly improved the ability to generate clinically meaningful reports from medical images. A recent survey highlights that MLLMs combine visual understanding (LVMs) and language reasoning (LLMs) to improve clinical workflows, but still face challenges like hallucination, bias, and data scarcity

DATASETS AND BENCHMARKS FOR RADIOLOGY REPORT GENERATION

The progress of Automated Radiology Report Generation (ARRG) has been strongly shaped by the availability of large-scale, publicly accessible datasets that pair medical images with corresponding radiology reports. Among these, MIMIC-CXR stands out as the most widely used benchmark. It contains more than 370,000 chest X-ray images and over 220,000 associated free-text reports from Beth Israel Deaconess Medical Centre. The dataset provides a rich source of radiological language, diverse patient conditions, and multiple view types, making it ideal for supervised training of vision–language models. Its derivative, MIMIC-CXR-JPG, offers images in JPEG format with standardized preprocessing, enabling faster experimentation. Another notable dataset is IU X-Ray, which contains frontal and lateral X-ray pairs with structured reports, though its relatively small size limits generalization. Beyond chest imaging, however, publicly available datasets remain limited. While several private or hospital-specific datasets exist for CT and MRI, they are generally restricted due to privacy and institutional regulations. This restricts the development of ARRG systems capable of handling multi-modality or cross-anatomy imaging. Researchers have also developed task-specific benchmarks to evaluate report completeness, clinical accuracy, and fact consistency. Tools such as CheXpert labels, RadGraph annotations, and RadEntity extraction frameworks help quantify clinical correctness by mapping the textual output to structured medical concepts. Despite this progress, several benchmark limitations persist. Reports vary widely in structure, detail, and writing style, leading to inconsistency in ground-truth labels. Many datasets contain incomplete or ambiguous reports, which introduce noise during training. Furthermore, rare and complex pathologies are underrepresented, making it difficult for models to learn comprehensive diagnostic patterns. These challenges emphasize the need for more diverse, multi-center, multimodal datasets, and standardized annotation protocols. Thus, dataset quality and representation remain central issues in advancing robust ARRG systems.

MODEL ARCHITECTURES: FROM CNN–RNN TO VISION–LANGUAGE TRANSFORMERS

Early ARRG systems used CNN–RNN encoder–decoder architectures, where CNNs (e.g., ResNet, DenseNet) extracted image features and RNNs (LSTM/GRU) generated text.Attentionmechanismsimprovedregionfocusbutstruggledwithlong-textcoherence and structured reporting.

IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.60

The introduction of Transformers improved performance by enabling long-range dependency modeling and better contextual understanding. Vision Transformers (ViT), Swin Transformers, and hybrid CNN–ViT encoders enhanced visual representation, while transformer decoders generated structured reports. RecentVision– Language Models (VLMs) such as BLIP-2, LLaVA-Med, and Med-PaLM integrate visual encoders with large language models, producing more coherent and clinically relevant reports. However, challenges like hallucination and weak image-text grounding still persist.

KNOWLEDGE-ENHANCED AND MULTIMODAL APPROACHES

To improve clinical accuracy, recent ARRG systems incorporate medical knowledge and multimodal data. Knowledge-enhanced models use ontologies like UMLS and RadLex to ensure correct terminology and reduce hallucination. Structured approaches such as RadGraph represent entities and relationships, improving factual consistency. Models like KERP use hierarchical templates to guide report generation. Multimodal systems integrate additional inputs such as patient history, prior scans, and EHR data, enabling better reasoning and temporal analysis. Despite improved reliability, these approaches increase complexity and require well-aligned datasets.

EVALUATION METHODS FOR RADIOLOGY REPORT GENERATION

Evaluation of ARRG systems is challenging due to clinical complexity. Traditional metrics such as BLEU,ROUGE, and CIDEr measure textual similarity but fail to capture clinical correctness. Domain-specific methods like CheXpert label comparison and RadGraph evaluation assess clinical accuracy by focusing on medical entities and relationships. Tools such as RadCliQ and factuality scores further improve evaluation. Expert evaluation by radiologists remains essential but is time-consuming. Recentimage- aware evaluation methods attempt to ensure alignment between generated reports and visual evidence. A combination of automated and expert evaluation is considered most reliable.

EVOLUTION FROM CNN–RNN TO HYBRID ARCHITECTURES

Initial ARRG systems used CNN–RNN models such as CNN-LSTM and HRNN for report generation. While effective for basic tasks, these models struggled with long-term dependencies, redundancy, and lack of contextual understanding. Attention mechanisms improved performance but did not eliminate hallucination issues. The inability to capture deep semantic relationships limited their clinical usefulness. These limitations led to the transition toward Transformer-based and hybrid architectures, which offer better scalability, coherence, and reasoning capability.

3. EXISTINGSYSTEM

Radiology report generation is performed entirely manually by trained radiologists. After examining chest X-ray images, the radiologist must identify abnormalities, describe anatomical structures, interpret findings, and write detailed narrative reports. This manual process is time-consuming, labour-intensive, and prone to human errors, especially when dealing with large numbers of studies. The quality and style of reports can vary between radiologists, leading to inconsistencies in terminology and reporting structure. Additionally, human fatigue may result in missed subtle findings, delayed diagnosis, or incomplete documentation. Although some hospitals use structured templates or automated text extraction tools, these approaches still rely heavily on human interpretationandlackadvancedAIsupportforvisualunderstandingorautomatedreport drafting. Therefore, the existing system struggles with scalability, efficiency, and standardization in modern high-volume clinical environments.

4. PROPOSED SYSTEM

The proposed system uses MedGemma, an advanced AI model that automatically creates radiology reports from chest X-ray images. It combines three main parts: a visual encoder that analyzes the X-ray, across-modal module that connects image features with text, and a medical language model that writes the report. The system identifies important image details, such as abnormalities and anatomical structures, and ensures the generated text matches the visual evidence, avoiding incorrect or imaginary findings. It also uses medical labels and consistency checks to improve accuracy and reduce errors. The output includes properly structured “Findings” and “Impression” sections similar to a real radiologist’s report. This AI system works end-to-end and helps radiologists by giving fast, reliable draft reports, improving efficiency, consistency, and safety in clinical workflows.

Context-Aware Visual Feature Encoding

The system employs a transformer-based visual encoder to extract context-rich representations from chestradiographs. Instead of relying solely on local texture patterns, the encoder captures global anatomical dependencies and inter-regional relationships. This enables effective identification of subtle abnormalities such as interstitial markings and mild opacities. The use of pretrained representations further enhances robustness across varying imaging conditions and patient variability.

Cross-Modal Clinical Alignment Mechanism

A key contribution of the proposed system lies in its cross-modal alignment module, which bridges visual features and medical language space. Through attention-driven fusion, the model dynamically associates image regions with clinically relevant terminology.Thismechanismensuresthatgenerateddescriptionsaregroundedinactual visual evidence, thereby reducing semantic drift and hallucination. Additionally, it facilitates implicit localization of pathological regions without requiring explicit bounding box annotations.

Structured Medical Language Generation

The aligned multimodal features are decoded using a transformer-based medical language model optimized for clinical text generation. The decoder produces multi-sentence reports with hierarchical structure, capturing both detailed observations and high-level diagnostic summaries. A label-guided consistency mechanism is integrated to enforce agreement between predicted clinical conditions and generated text. This results in improved factual correctness, coherent narrative flow, and adherence to professional reporting standards.

IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.60

Advanced Reporting and Data Analytics

Thealignedmultimodalfeaturesaredecodedusingatransformer-basedmedicallanguage model optimized for clinical text generation. The decoder produces multi-sentence reports with hierarchical structure, capturing both detailed observations and high-level diagnostic summaries. A label-guided consistency mechanism is integrated to enforce agreement between predicted clinical conditions and generated text. This results in improved factual correctness, coherent narrative flow, and adherence to professional reporting standards.

System Significance and Novelty

The proposed framework distinguishes itself by combining domain-specific multimodal learning with structured clinical reasoning. It enhances report reliability through grounded generation and consistency-aware decoding. The system demonstrates potential for real-time deployment as a radiology decision-support tool, improving reporting efficiency while maintaining diagnostic integrity. Furthermore, its modular design allows extension to other imaging modalities and integration with clinical metadata for future enhancements.

Implementation Details and Workflow

The proposed system is implemented using a deep learning framework with support for multimodal processing, leveraging Med-Gemma as the core model. The workflow begins with preprocessing of chest X-ray images, including resizing, normalization, and conversion into tensor representations. These images are then passed through the visual encoder to obtain feature embeddings. The generated visual features are integrated with a predefined clinical prompt and processed through the cross-modal alignment module. The transformer-based decoder subsequently generates the radiology report in a sequential manner. Decoding strategies such as controlled sampling or beam search are applied to improve textual coherence and stability. To enhance clinical reliability, post- processing steps are incorporated, including medical label verification and consistency checks between predicted findings and generated text. The system outputs a structured report consisting of standardized sections such as “Findings” and “Impression.” The entire pipeline is designed for efficient inference and can be deployed using GPU acceleration for real-time applications. The modular workflow allows easy integration with hospital systems and supports scalability for large-scale clinical deployment.

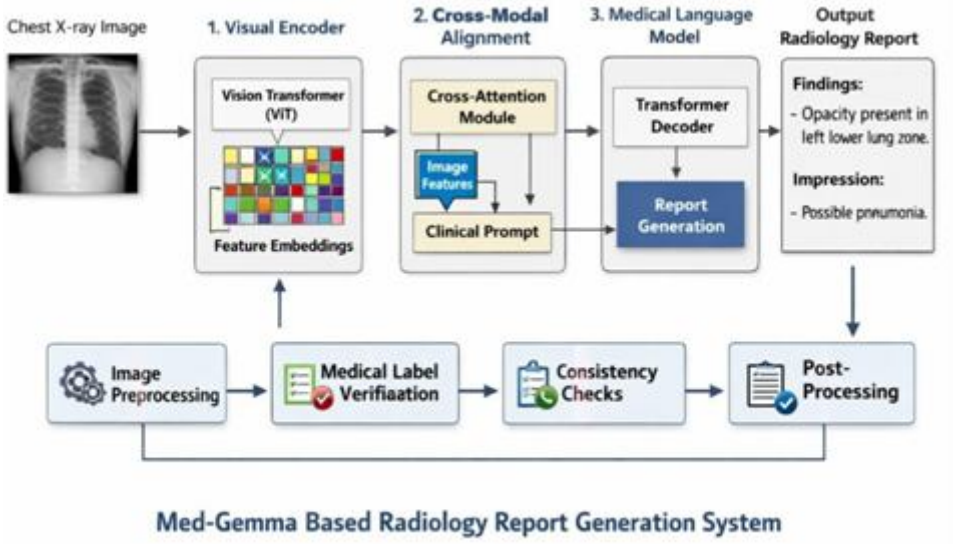

Figure 2. Block diagram

5. RESULTS AND DISCUSSION

Experimental Setup

The proposed system based on Med-Gemma was evaluated using chest X-ray datasets such as MIMIC-CXR. The model was tested on unseen images to assess its ability to generate accurate and clinically relevant radiology reports. Performance was evaluated using both natural language generation metrics and clinical accuracy measures.

Quantitative Evaluation

The performance of the proposed model was measured using standard metrics such as BLEU, ROUGE, and CIDEr. These metrics evaluate the similarity between generated reports and ground-truth reports.

Table 1. Comparison with review literature

Model	BLEU-1	BLEU-4	ROUGE-L	CIDEr
CNN–RNN Model	0.62	0.28	0.45	0.85
Transformer Model	0.68	0.35	0.52	0.98
Proposed (Med-Gemma)	0.74	0.42	0.60	1.15

Qualitative Analysis

The generated reports were analyzed for clinical correctness and structure. The proposed system successfully produced well-organized reports with clear “Findings” and “Impression” sections. Compared to baseline models, it showed better coherence, improved abnormality detection, and reduced redundancy.

ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.60

Clinical Accuracy Evaluation

To assess medical reliability, generated reports were compared using clinical label extraction methods. The proposed system demonstrated higher agreement with ground- truth labels, indicating improved diagnostic relevance.

Discussion

The results indicate that the proposed system outperforms traditional CNN–RNN and transformer-based approaches. The integration of cross-modal alignment and domain- specific knowledge enables more accurate and context-aware report generation. However, minor limitations such as occasional mission of rare findings and dependency on dataset quality were observed.

6. CONCLUSION

This paper presented an automated radiology report generation system using Med- Gemma, a domain-specific Vision–Language Model. The proposed approach integrates visual feature extraction, cross-modal alignment, and transformer-based language generation to produce structured and clinically relevant radiology reports from chest X- ray images. Experimental results demonstrate that the system outperforms traditional CNN–RNN and standard transformer-based models in terms of textual quality and clinical accuracy. The generated reports show improved coherence, reduced redundancy, and better alignment with visual evidence, there by minimizing hallucinated findings. The system effectively generates standardized “Findings” and “Impression” sections, supporting radiologists by providing fast and consistent draft reports. Although challenges such as data dependency and the need for clinical validation remain, the proposed framework serves as a reliable clinical decision-support tool. Overall, this work highlights the potential of domain-adapted multimodal models in enhancing radiology workflows and advancing AI-assisted healthcare systems.

7. FUTURE WORK

Future research can focus on enhancing the proposed system based on Med-Gemma by in corporating additional multimodal inputs such as patient history, laboratory reports, and prior imaging studies to improve clinical reasoning and contextual understanding. Integrating temporal information from sequentials can scan further enable better analysis of disease progression. Another important direction is the inclusion of knowledge-based systems using medical ontologies and structured representation store ducehallucinations and improve factual accuracy. Developing explainable AI techniques, such as attention visualization and region highlighting, can increase transparency and trust among clinicians. Further improvements can be achieved by fine-tuning the model on larger and more diverse medical datasets to handle rare pathologies and improve generalization. Real-time deployment and integration with hospital systems such as PACS and electronic health records (EHR) should also be explored. Finally, incorporating human-in-the-loop validation and reinforcement learning from expert feedback can enhance the reliability and safety of the system, making it more suitable for clinical adoption.

REFERENCES

1. O.Er,N.Yumusak, and F.Temurtas,“Chest diseases diagnosis using artificial neural networks,” Expert Syst. Appl., vol. 37, no. 12, pp. 7648–7655, Dec. 2010.
2. T.Gupte,A.Knack,andJ.D.Cramer,“Mortality from aspiration pneumonia: Incidence, trends, and risk factors, ”Dysphagia, vol.37,no.6,pp.1493–1500,Dec.2022.
3. L.Delrue,R.Gosselin, B.Ilsen, A.Van Landeghem, J.deMey, and P. Duyck, “Difficulties in the interpretation of chest radiography,” in Comparative Interpretation of CT and Standard Radiography of the Chest. Berlin, Germany: Springer, 2011, pp.27– 49.
4. B.Jing,P.Xie, and E.Xing,‘ ‘On the automatic generation of medical imaging reports,” in Proc. 56th Annu. Meeting Assoc. Comput. Linguistics. Melbourne, VIC, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2577–2586. [Online]. Available: https://aclanthology.org/P18-1240
5. I.Allaouzi, M.B.Ahmed, B.Benamrou ,and M.Ouardouz ,“Automatic caption generation for medical images,”in Proc.3rdInt.Conf.SmartCityAppl.NewYork,NY, USA: Association for Computing Machinery, Oct. 2018, pp. 1–6.
6. J.Yuan,H.Liao,R.Luo,andJ.Luo,“Automatic radiology report generation based on multi-view image fusion and medical concept enrichment,” in Medical Image Computing and Computer Assisted Intervention MICCAI 2019 (Lecture Notes in ComputerScience),vol.11769.MidtownManhattan,NYUSA:Springer,2019,pp.721– 729.
7. A.Vaswani, N.Shazeer, N.Parmar, J.Uszkoreit, L.Jones,A.N.Gomez, L.U. Kaiser,andI.Polosukhin, “Attention isallyouneed,”inProc.Adv.NeuralInf.Process. Syst., vol. 30. Red Hook, NY, USA: Curran Associates, 2017, pp. 1–11.
8. J.Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre- training of deepbidirectionaltransformersforlanguageunderstanding,”inProc.Conf.NorthAmer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., vol. 1. Minneapolis, MN, USA: Association for Computational Linguistics, Jun.2019,pp.4171–4186. [Online]. Available: https://aclanthology.org/N19-1423
9. A.Radford,J.Wu,R.Child,D.Luan,D.Amodei,andI.Sutskever,“Language models are unsupervised multitask learners,” OpenAI, San Francisco, CA, USA, Tech. Rep., 2019.
10. C.Raffel, N.Shazeer, A.Roberts, K.Lee, S.Narang, M.Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” J. Mach. Learn. Res., vol. 21, no. 140, pp. 1–67, 2020.
11. V.Sanh,L.Debut,J.Chaumond,andT.Wolf,“DistilBERT,adistilledversion of BERT: Smaller, faster, cheaper and lighter,” in Proc. NeurIPSEMCWorkshop,2019,pp.673–680.
12. W.Wang, F.Wei, L.Dong, H.Bao, N.Yang, and M.Zhou, “MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers,” in Proc. Adv. NeuralInf.Process. Syst., Annu. Conf.NeuralInf. Process. Syst. (NeurIPS), vol. 33, Dec. 2020.

ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.60

13. Y.Xiong,B.Du,andP.Yan,"Reinforced transformer for medical image captioning,' 'in Proc.Int. Workshop Mach. Learn. Med. Imag.MidtownManhattan,NY, USA: Springer, 2019, pp. 673–680.
14. C.Y.Li,X.Liang,Z.Hu,andE.P.Xing, "Knowledge-drivenencode, retrieve, paraphrase for medical image report generation,"in Proc.33rd AAAI Conf.Artif.Intell.31stInnov.Appl.Artif.Intell.Conf.9thAAAISymp.Educ.Adv.Artif.Intell. (AAAI/IAAI/EAAI).PaloAlto,CA,USA:AAAIPress,2019,pp.6666–6673, doi:10.1609/aaai.v33i01.33016666.
15. Z.Chen,Y.Song,T.-H.Chang,andX.Wan,"Generatingradiologyreportsvia memory-driven transformer," in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP). Stroudsburg, PA, USA: Association for Computational Linguistics, Nov. 2020, pp. 1439–1449. [Online]. Avail- able: https://aclanthology.org/2020.emnlp- main.112
16. P.Srinivasan,D.Thapar,A.Bhavsar,andA.Nigam,"HierarchicalX-rayreport generation via pathology tags and multi head attention," in Proc. Asian Conf. Comput. Vis. (ACCV), Nov. 2020, pp. 1–17.
17. F.Liu,C.You,X.Wu,S.Ge,S.Wang,andX.Sun,"Auto-encodingknowledge graph for unsupervised medical report generation,"in Proc. Adv. Neural Inf. Process. Syst. ,A. Beygelzimer ,Y. Dauphin, P.Liang,a ndJ.W.Vaughan,Eds.,2021,pp.16266– 16279. [Online]. Available: https://openreview.net/forum?id=nIL7Q-p7-Sh
18. H. Nguyen, D. Nie, T. Badamdorj, Y. Liu, Y. Zhu, J. Truong, and L. Cheng, "Automated generation of accurate & fluent medical X-ray reports," in Proc. Conf. Empirical Methods Natural Lang. Process.PuntaCana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3552–3569.
19. F. Liu, X. Wu, S. Ge, W. Fan, and Y. Zou, "Exploring and distilling posterior and prior knowledge for radiology report generation,"inProc.IEEE/CVFConf.Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 13748–13757.
20. F.Nooralahzadeh, N.P.Gonzalez, T.Frauenfelder, K.Fujimoto, and M. Krauthammer, "Progressive transformer-basedgenerationofradiologyreports,"inProc. Findings Assoc. Comput. Linguistics (EMNLP). Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 2824–2832. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.241
21. O.Alfarghaly,R.Khaled,A.Elkorany,M.Helal,andA.Fahmy,"Auto-mated radiology report generation using conditioned transformers, "Informat. Med.Unlocked, vol. 24, 2021, Art. no. 100557.
22. D.You,F.Liu,S.Ge,X.Xie,J.Zhang,andX.Wu,"Align Transformer: Hierarchical alignment of visual regions and disease tags for medical report generation," in Proc. 24th Int. Conf., Med. Image Comput. Comput. Assist. Intervent. (MICCAI), Strasbourg, France, Sep./Oct. 2021.
23. Y.Wang,Z.Lin,Z.Xu,H.Dong,J.Tian,J.Luo,Z.Shi,Y.Zhang,J.Fan,and Z.He,"Trustitornot:Confidence-guided automatic radiology report generation,"2021, arXiv:2106.10887.
24. N.Parmar,A.Vaswani,J.Uszkoreit,L.Kaiser,N.Shazeer,A.Ku,andD.Tran, "Image transformer,' 'inProc.35th Int.Conf. Mach.Learn. ,vol.80,J.DyandA.Krause, Eds.,Jul.2018,pp4055–4064.[Online]. Available: https://proceedings.mlr.press/v80/parmar18a.html
25. A.Kolesnikov, A.Dosovitskiy, D.Weissenborn, G.Heigold, J.Uszkoreit, L. Beyer, M. Minderer, M. Dehghani, N. Houlsby, S. Gelly, T. Unterthiner, and X. Zhai, "Animageisworth1616words:Transformersforimagerecognitionatscale,"inProc. 9th Int. Conf. Learn. Represent. (ICLR), 2021.
26. H.Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training data-efficient image transformers & distillation through attention," in Proc. Int. Conf. Mach. Learn., vol. 139, Jul. 2021, pp. 10347–10357.
27. M.Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transform- ers," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 9650–9660.