의료 보고서 생성을 위한 비전-언어 모델

Nguyen Duc Toan[1], Muhammad Omer[2], 추현승 [3,4]
[1]성균관대학교 AI 시스템공학과 박사과정
[2]성균관대학교 소프트웨어학과 박사과정
[3]성균관대학교 전자전기컴퓨터공학과 교수
[4]성균관대학교 AI 시스템공학과 교수

austin47@g.skku.edu, omer389@g.skku.edu, choo@skku.edu

Vision-Language Model for Medical Report Generation

Toan Duc Nguyen[1], Muhammad Omer[2], Choo Hyunseung[1,3]
[1]Dept. of AI Systems Engineering, Sungkyunkwan University
[2]Dept. of Computer Science and Engineering, Sungkyunkwan University
[3]Dept. of Electrical and Computer Engineering, Sungkyunkwan University

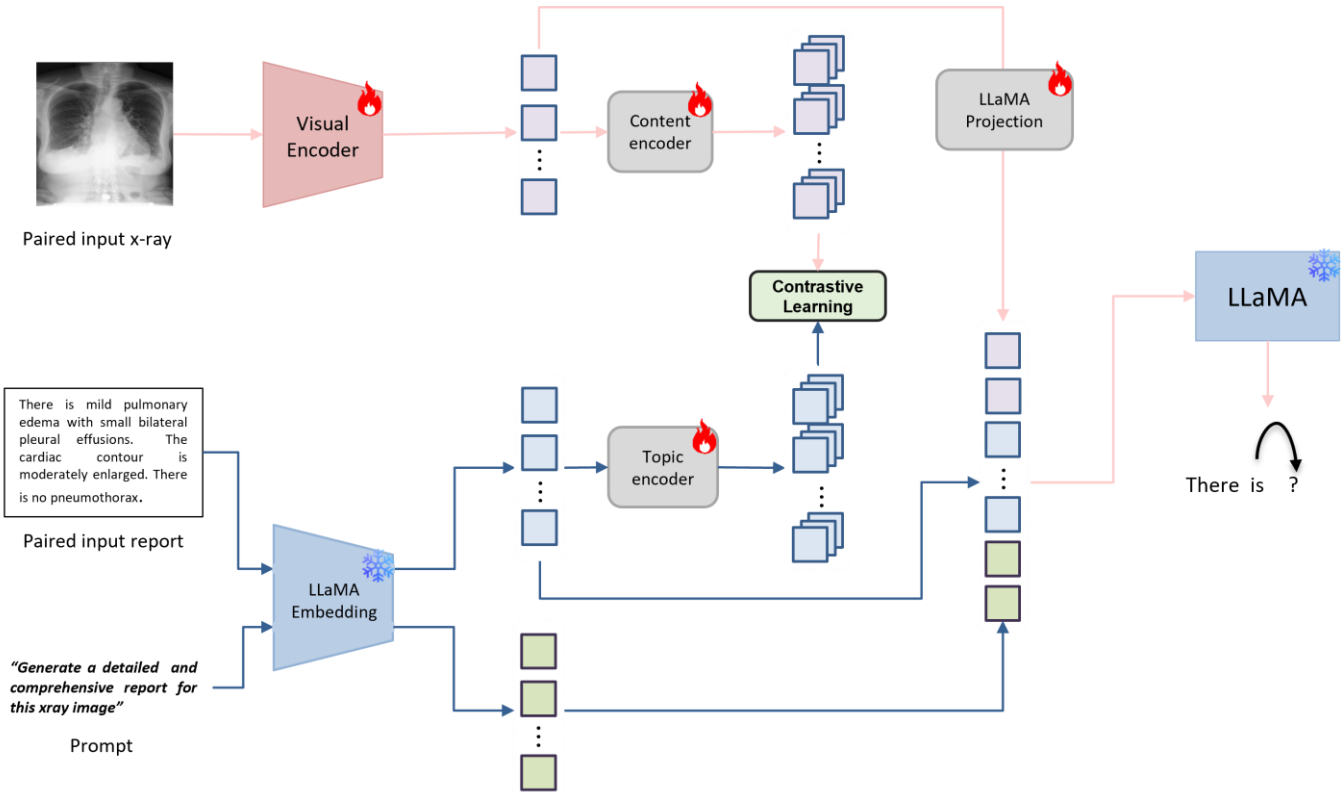
요 약

Medical report generation aims to automatically produce diagnostic reports from medical images, offering the potential to alleviate the workload of radiologists and reduce human error. Recent advances in large vision-language models have shown remarkable capability in bridging visual and textual modalities, making them strong candidates for this task. In this work, we propose a novel medical report generation framework that leverages a visual encoder and a content-topic dual encoding mechanism to capture both fine-grained visual features and high-level semantic structures. The encoded representations are aligned with language embeddings through contrastive learning, followed by a LLaMA-based decoder for fluent and clinically relevant report generation. We evaluate our method on the IU X-Ray dataset, achieving a BLEU-1 score of 0.489, outperforming prior baselines and demonstrating the effectiveness of large vision-language models for medical imaging applications.

1. Introduction

Medical report generation (MRG) is an important task in medical artificial intelligence that aims to automatically generate detailed textual descriptions from medical images such as chest X-rays, CT scans, or MRIs. Unlike image classification, which predicts only a label, MRG requires generating long, coherent narratives that describe both normal findings and abnormalities, while adhering to clinical terminology. This makes the task more complex due to the need for accurate visual understanding, medical knowledge, and natural language fluency. Early studies often adopted encoder- decoder architectures, where convolutional neural networks (CNNs) or vision transformers (ViTs) extracted image features and recurrent neural networks (RNNs) or transformers generated the text [1,2]. Although these approaches demonstrated progress, they often struggled with generating diverse, clinically precise reports, frequently producing generic or repetitive sentences [3].

In this work, we propose a novel medical report generation framework that integrates a dual-level encoding strategy with a large language model (LLM) backbone. Specifically, our model employs a visual encoder to extract image features, followed by content and topic encoders that capture both fine-grained and high-level semantic cues. These representations are aligned with text embeddings from LLaMA through contrastive learning, ensuring robust cross-modal



(Figure 1) Overview of the proposed framework. A visual encoder extracts features from the chest X-ray, which are further processed by content and topic encoders to capture fine-grained clinical details and high-level semantic structure. Contrastive learning aligns the encoded representations with LLaMA embeddings, and a projection layer adapts the features for the LLaMA decoder

understanding. Finally, a LLaMA-based decoder generates coherent and clinically relevant reports, benefiting from both visual grounding and topic-guided supervision.

2. Methodology

An overview of the proposed architecture is shown in Figure 1. The model consists of four main components: (1) a visual encoder, (2) a dual semantic encoding module comprising content and topic encoders, (3) a contrastive alignment mechanism with LLaMA embeddings, and (4) a LLaMA embedding and decoder for report generation.

Given a chest X-ray image, we first extract high-dimensional visual features using a pretrained vision transformer (ViT) backbone, denoted as the visual encoder. The encoder produces a sequence of image tokens that capture both local and global information relevant to radiological interpretation. These representations serve as the foundation for subsequent semantic encoding. The image representations are then projected into the LLaMA embedding space using a LLaMA projection layer, which adapts the encoded features to the input requirements of the LLaMA decoder.

To disentangle different levels of medical semantics, we introduce a dual encoding strategy: a content encoder and topic decoder. The content encoder refines the visual tokens into semantically meaningful representations that correspond to fine-grained clinical observations (e.g., "enlarged heart," "opacity in the left lung"). This pathway emphasizes local image- text alignment. In parallel, the topic encoder integrates LLaMA embeddings derived from section-level or topic-level tokens (e.g., "Findings," "Impression"). This pathway captures global report structure and ensures that the generated text aligns with the common organization of radiology reports.

Both encoders transform the input into latent spaces that represent different but complementary aspects of the report, enabling the model to balance local accuracy with high-level coherence.

To bridge the gap between visual and textual modalities, we employ a contrastive learning objective between the outputs of the encoders and the LLaMA embeddings. Specifically, image-derived representations from the content and topic encoders are aligned with corresponding textual embeddings from the LLaMA encoder. This process ensures that semantically similar visual- text pairs are mapped closer in the shared latent space, while unrelated pairs are pushed apart. During training, the image is fed into the visual encoder, while the paired report text is encoded using the frozen LLaMA encoder to obtain textual embeddings. The model learns to align these paired representations through a contrastive loss that minimizes the distance between embeddings of matched image- report pairs and maximizes it for mismatched pairs within the batch.

The model is trained with a combination of objectives: contrastive and language modelling. The contrastive loss ensures effective alignment between image and text embeddings across modalities. The language modeling loss (cross-entropy) supervises the decoder to generate accurate and fluent text based on ground truth reports. By jointly optimizing these objectives, the model learns to both align visual and textual semantics and generate relevant reports.

3. Implementation Details

3.1 Dataset

We evaluate our method on the IU X-Ray dataset [4], a widely used benchmark for medical report generation. The dataset consists of 7,470 chest X-ray images paired with 3,955 radiology reports. Each report typically contains two sections: Findings, describing observable details from the image, and Impression, summarizing the diagnostic conclusion. Following prior works, we split the dataset into training, validation, and test sets with a ratio of 70%, 10%, and 20%, respectively. Standard preprocessing steps are applied, including resizing images to 224×224, normalizing pixel intensities, and tokenizing reports using the LLaMA tokenizer.

3.1 Training Setup

We initialize the visual encoder with a ViT pretrained on ImageNet, while the LLaMA model is initialized from LLaMA-2-7B with frozen weights to preserve its pretrained language knowledge. The content encoder, topic encoder, and LLaMA projection layer are trained from scratch. The pretrained models are used from Huggingface library. The model is optimized using the AdamW optimizer with a learning rate of 1e-4 and weight decay of 1e-6. A batch size of 8 is used, and training is conducted for 5 epochs. To stabilize contrastive learning, we use a temperature parameter of 0.07. The total loss is a sum of the language modeling loss (cross-entropy) and the contrastive loss, with weights of 0.5 and 0.5, respectively. All experiments are conducted on two NVIDIA A100 GPUs with 80GB memory.

3.3 Evaluation Metrics

We adopt standard natural language generation metrics widely used in medical report generation to evaluate the quality of generated reports. BLEU-n (n=1- 4) [5] measures the precision of n-grams between generated and reference reports, capturing surface-level similarity. METEOR [6] considers both precision and recall, while also accounting for synonym and stemming matches, making it more sensitive to semantic similarity. ROUGE-L [7] evaluates the longest common subsequence between generated and reference texts, focusing on fluency and content overlap. Finally, CIDEr [8] is designed for consensus-based evaluation by comparing generated reports to multiple references.

4. Results

Table 1 presents the performance of our proposed model compared with recent state-of-the-art methods, including R2Gen [9] and R2GenGPT [10], on the IU X-Ray dataset. Our model achieves the highest BLEU-1 score of 0.489, slightly surpassing R2GenGPT (0.488) and R2Gen (0.47). For higher-order BLEU scores, our method demonstrates competitive performance, obtaining a BLEU-3 score

of 0.229, which is the best among all methods. Although BLEU-2 and BLEU-4 are marginally lower than R2GenGPT, our model consistently outperforms previous baselines in METEOR (0.22), ROUGE-L (0.379), and CIDEr (0.453), indicating stronger semantic alignment and clinical relevance.

<Table 1> performance comparison of the proposed method with other baselines on the IU X-Ray dataset

Method	R2Gen	R2GenGPT	Ours
BLEU-1	0.47	0.488	**0.489**
BLEU-2	0.304	**0.316**	0.314
BLEU-3	0.219	0.228	**0.229**
BLEU-4	0.165	**0.173**	0.165
METEOR	0.187	0.211	**0.23**
ROUGE	0.371	0.377	**0.379**
CIDR	-	0.438	**0.453**

5. Conclusion

In this work, we presented a novel vision-language framework for medical report generation that combines a dual semantic encoding strategy with contrastive alignment and a large language model backbone. By jointly modeling fine-grained clinical observations through a content encoder and high-level structural semantics through a topic encoder, our approach effectively grounds visual features in medical text. The integration of contrastive learning further enhances the alignment between image representations and LLaMA embeddings, leading to more coherent and clinically relevant report generation. Experimental results on the IU X-Ray dataset demonstrate that our model achieves competitive performance across multiple metrics.

Acknowledgements

This work was partly supported by the Korea government (MSIT), IITP, Korea, under the ICT Creative Consilience program (IITP-2025-RS-2020-II201821, 40%), (RS-2024-00459512, Development of Brain Disease (Stroke) Prediction Model based on Fundus Image Analysis, 20%), (RS-2021-II212068, Artificial Intelligence Innovation Hub, 20%), (RS-2019-II190421, Artificial Intelligence Graduate School Program(Sungkyunkwan University), 20%)

References

[1] Jing, Baoyu, Pengtao Xie, and Eric Xing. "On the automatic generation of medical imaging reports." arXiv preprint arXiv:1711.08195 2017

[2] Li, Yuan, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. "Hybrid retrieval-generation reinforced agent for medical image report generation." Advances in neural information processing systems 31, 2018

[3] Xue, Yuan, Tao Xu, L. Rodney Long, Zhiyun Xue, Sameer Antani, George R. Thoma, and Xiaolei Huang. "Multimodal recurrent model with attention for automated radiology report generation." In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 457-466. Cham: Springer International Publishing, 2018

[4] Demner-Fushman, Dina, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. "Preparing a collection of radiology examinations for distribution and retrieval." Journal of the American Medical Informatics Association 23, no. 2 2015

[5] Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. "Bleu: a method for automatic evaluation of machine translation." In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp. 311-318. 2002.

[6] Banerjee, Satanjeev, and Alon 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, pp. 65-72. 2005

[7] Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." In Text summarization branches out, pp. 74-81. 2004

[8] Vedantam, Ramakrishna, C. Lawrence Zitnick, and Devi Parikh. "Cider: Consensus-based image description evaluation." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4566-4575. 2015

[9] Chen, Zhihong, Yan Song, Tsung-Hui Chang, and Xiang Wan. "Generating radiology reports via memory-driven transformer." arXiv preprint arXiv:2010.16056 2020

[10] Wang, Zhanyu, Lingqiao Liu, Lei Wang, and Luping Zhou. "R2gengpt: Radiology report generation with frozen llms." Meta-Radiology 1, no. 3 2023