RadTextAid: A CNN-Guided Framework Utilizing Lightweight Vision-Language Models for Assistive Radiology Reporting

Mahmud Wasif Nafee, Tasmia Rahman Aanika, Taufiq Hasan
mHealth Lab, Department of Biomedical Engineering
Bangladesh University of Engineering and Technology (BUET), Dhaka-1205, Bangladesh.
Email: taufiq@bme.buet.ac.bd

Abstract

Deciphering chest X-rays is crucial for diagnosing thoracic diseases such as pneumonia, lung cancer, and cardiomegaly. Radiologists often work under significant workloads and handle large volumes of data, which can lead to exhaustion and burnout. Advanced deep learning models can effectively generate draft radiology reports, potentially alleviating the radiologist's workload. However, many current systems create reports that include clinically irrelevant or redundant information. To address these limitations, we propose RadTextAid, a novel multi-modal framework for generating high-quality, clinically relevant radiology reports. Our approach integrates VLMs for natural language generation, augmented by disease-specific tags derived from a CNN analyzing chest X-ray images to identify key pathological features. A key feature within our framework is the pre-processing of the radiology report training dataset. This removes routine, repetitive, or non-informative phrases commonly found in chest X-ray reports and ensures that the model focuses its learning on clinically meaningful content, which expert radiologists qualitatively validated. Experimental results show that our system yields an absolute improvement of 4.8% in terms of BERTScore and 3.16% in terms of the F1-cheXbert metric compared to a state-of-the-art model. Thus, the results demonstrate that the proposed RadTextAid framework not only improves the detection of abnormalities from chest X-ray images but also enhances the overall quality and coherence of generated reports, thus paving the way toward more efficient and effective radiology reporting.

Introduction

Interpreting chest X-ray images is a critical, time-sensitive task essential for the timely diagnosis and treatment of numerous medical conditions. Radiologists typically invest considerable time meticulously analyzing each chest X-ray image to prevent misdiagnoses that could adversely affect patient outcomes. However, this thorough approach often leads to inefficiencies, particularly in settings characterized by high patient volume or a shortage of qualified radiologists, which is common in low- and middle-income countries (LMICs). [Parag and Hardcastle2022]

Deep learning and natural language processing (NLP) - based algorithms can automatically generate draft radiology reports, potentially improving reporting efficiency while retaining accuracy. However, a central issue in report generation is the accurate detection and precise documentation of abnormalities while differentiating them from routine physiological and pathological observations. Publicly available datasets predominantly contain information on routine observations, with limited references to abnormalities or comprehensive medical histories. This abundance of routine findings in the training data poses significant obstacles for state-of-the-art Vision Language Models (VLMs) employed in automated report generation, as these models rely on the token loss during training. The similarity of captions in these datasets diminishes the model's ability to generate differential annotations concerning clinical abnormalities in their reports. Consequently, the models lack penalties for either overlooking or inaccurately predicting pathological conditions, leading them to generate text primarily related to routine observations.

To address the issue of inadequately generating pathological findings, one viable approach is to train the VLM exclusively on texts pertaining to clinical abnormalities rather than on the entirety of the report. This targeted training would enhance the model's capacity to generate detailed pathological findings, necessitating greater attention and precision. Alternatively, implementing a Convolutional Neural Network (CNN)-based classifier to first identify chest X-ray abnormalities could serve as a mechanism to guide the VLM through targeted prompts, thereby enhancing overall reporting accuracy. However, to our knowledge, lightweight VLMs guided by CNN models have not been adequately studied in the literature for radiology report generation.

In this study, we propose RadTextAid, a novel framework designed to train a lightweight multi-modal model to assist radiologists in automatic report generation by focusing specifically on clinical abnormalities. The proposed pipeline consists of three distinct models: (i) The PaliGemma multi-modal vision language model is trained to generate relevant findings; (ii) The Llama 3.1 8b model, which extracts only the pathological findings to inform the training corpus; and

(iii) a CNN-based chest X-ray image classifier (CheXNet) to produce pathological tags for prompt guidance of the vision language model. Our main contributions can be summarized as follows.

- We propose a novel, efficient pipeline to train vision language models for domain-specific tasks such as X-ray report generation.
- We investigate the state-of-the-art large language models (LLMs) to extract clinically relevant information from medical reports with few-shot prompting.
- We explore the potential of CNN-generated disease labels from chest X-ray images to construct prompts to guide the vision language model.

Related Works

Chest radiography holds the position of being the most prevalent imaging examination worldwide. In contrast to general image captioning tasks that prioritize coherence [Chen et al.2018], medical image captioning necessitates a greater emphasis on accuracy in identifying anomalies and extracting information while still maintaining coherence. [Singh et al.2022] This means the generated report should be comprehensible and convey precise medical information effectively [Srinivasan et al.2020].

Recent developments in computational machine translation [Sirshar et al.2022] have shown that by using a robust sequence model and an LSTM with an attention head, performance improves significantly.. The Contrastive Attention (CA) approach proposed by Liu *et al.* [Liu et al.2021] distills the contrastive information by comparing the current input image with normal X-ray images (from healthy subjects) rather than concentrating just on it. Srinivasan *et al.* [Srinivasan et al.2021] introduces a deep neural network that utilizes a set of chest x-ray images to predict the medical tags and provide a comprehensible radiology report.

Recently, transformer architectures have been used more commonly for radiology report generation to overcome the limitations of Recurrent Neural Networks (RNNs). The CNX-B2 model [Alqahtani et al.2024] is a CNN combined with a transformer network to generate medical reports. The work of [Alfarghaly et al.2021] involves fine-tuning a pre-trained ChexNet model [Rajpurkar et al.2017] to predict specific identifiers from the image, generating weighted semantic features from the pre-trained embeddings of the predicted tags, and producing complete medical reports by conditioning a pre-trained GPT-2 model on the visual and semantic features. In [Mondal et al.2023], the authors introduce EfficienTransNet, an automatic chest X-ray report generation approach based on CNN-Transformers. EfficienTransNet incorporates clinical history or indications to enhance the report generation process and align with radiologists' workflow, which is mostly overlooked in recent research.

In order to enhance the generation of chest x-ray reports, [Nicolson, Dowling, and Koopman2023] propose the exploration of warm starting the encoder and decoder using up-to-date computer vision and natural language processing checkpoints. TrMRG [Mohsan et al.2023], also known as the Transformer Medical report generator, is a comprehensive model that utilizes the Transformer architecture to generate reports. This model incorporates pre-trained computer vision and language models, making it a powerful tool for report generation. [Wang et al.2023] introduces METransformer that incorporates a transformer framework and includes "expert tokens" into both the transformer encoder and decoder, representing many experts.

The challenge of large vision-language models in medical imaging is addressed by [Kim et al.2024] through a Llava-based framework, though its size poses limitations. Their findings showed that lightweight models perform better, prompting us to suggest models like Paligemma and Florence-2 as viable alternatives. While [Alfarghaly et al.2021] and [Srinivasan et al.2021] explore confidence scores of tags for report generation, the potential of structured prompts from tags remains unexplored. Our proposed model aims to bridge these gaps in the literature.

Methodology

Proposed Framework

An overview of the proposed radiology report generation framework is shown in Fig. 1. We begin with a chest x-ray image as input, which flows through a multi-label classifier based on CheXNet-121 [Rajpurkar et al.2017]. This classifier scans the available image for relevant visual features and produces 105 diagnostic tags that indicate the identified disease conditions or abnormalities. These tags provide semantic information about the X-ray findings and are visualized in the system diagram as individual elements, offering clarity and interpretability of the multi-label classification results.

Next, the initial Chest X-ray image and the produced tags (disease labels), which are text outputs, are fed into a multi-modal Vision-Language Model (VLM). In our study, we have considered the Florence-2 [Xiao et al.2023] and PaliGemma [Beyer et al.2024] VLM models. The VLM fuses the visual characteristics learnt from the image with the semantic information obtained by the tags (disease labels), creating a detailed diagnostic report with observations of inferred conditions and abnormalities along with additional clinical insights. The proposed framework can be divided into two components:

1. The multi-label classifier based on CheXNet-121 to produce 105 disease labels.
2. The multi-modal Vision-Language Model (VLM), i.e. Florence 2 or PaliGemma

CNN-based multi-label classifier

The input chest x-ray image is first passed through a CNN model to produce the tags' predictions. Our base model is based on a ChexNet model [Rajpurkar et al.2017], which is essentially a Densenet121 model [Huang, Liu, and Weinberger2016] pre-trained on Chest X-ray14 dataset to identify and localize 14 thoracic

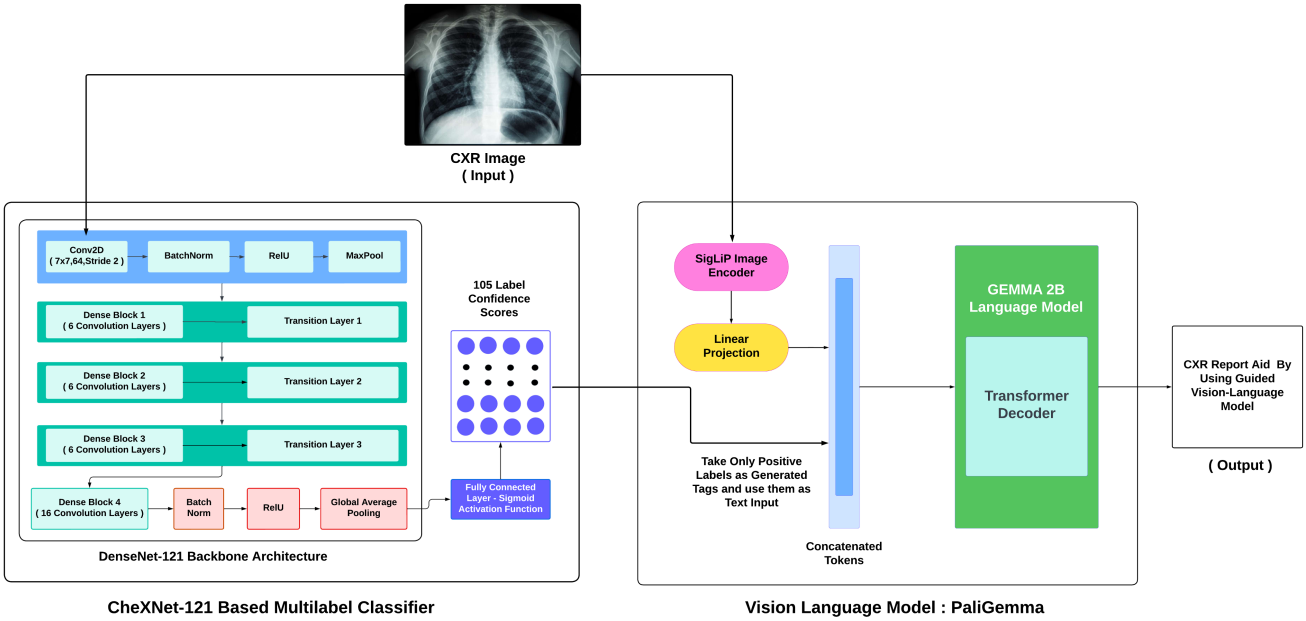


Figure 1: **Our proposed model architecture for RadTextAid**

diseases. However, these 14 tags were insufficient for conditioning the report generation model. Thus, we chose a fine-tuned model to classify the manual tags from the IU-Xray dataset [Demner-Fushman et al.2015] by removing the final layer and adding a new final layer containing 105 nodes for the most occurring manual tags from the dataset as implemented in [Alfarghaly et al.2021].

The positive tags represent the physiological conditions identified in the chest X-ray and serve as valuable inputs for constructing structured prompts. These prompts have the potential to enhance the performance of Vision-Language models in generating accurate and contextually relevant reports.

The model uses the binary cross-entropy (BCE) loss to handle multiple diagnostic labels,

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}\left[y_{ij}\log(\hat{y}_{ij}) + (1-y_{ij})\log(1-\hat{y}_{ij})\right]$$

where N is the number of samples (batch size), M is the number of labels (e.g., $M = 105$ here), y_{ij} is the ground truth binary label for the j-th class of the i-th sample, $\hat{y}_{ij}$ is the predicted probability for the j-th class of the i-th sample, and $\log$ is the natural logarithm.

The multi-modal Vision-Language Model (VLM)

In this step, we use the positive disease tags obtained from the multi-label CNN-based classifier to construct textual prompts. The input image and prompt are passed through a multi-modal VLM. We consider the pre-trained Florence-2 and PaliGemma models due to their lightweight architecture compared to Llava 1.6 (34B parameters) [Liu et al.2024], GPT-4 (200B parameters) [OpenAI2024], GPT-4o Mini (8B parameters) [OpenAI2024], and Qwen2-VL (72B parameters) [Wang et al.2024]. Our method is partly inspired by [Alfarghaly et al.2021], where a conditioned transformer model is used to generate radiology reports.

Florence-2

Florence-2 is a recent VLM [Xiao et al.2023] (0.7B parameters) developed for various vision and vision-language tasks with a unified sequence-to-sequence architecture. It uses a Dual Attention Vision Transformer (DaViT) [Ding et al.2022] as its vision encoder to process images into token embeddings $\mathbf{V} \in \mathbb{R}^{N_v \times D_v}$, where N_v represents the number of visual tokens and D_v their dimensionality. An extended version of the tokenizer combines these embeddings with task-related text prompts $\mathbf{T}_{\text{prompt}} \in \mathbb{R}^{N_t \times D}$.

The model uses a multi-modal transformer-based encoder-decoder, with input $\mathbf{X} = [V', \mathbf{T}_{\text{prompt}}]$, where $\mathbf{V}'$ is the dimensionally aligned projection of $\mathbf{V}$. It uses a standard cross-entropy language modeling objective for training:

$$\mathcal{L} = -\sum_{i=1}^{|y|}\log P_\theta(y_i \mid y_{<i}, x),$$

where, y represents the target sequence, and x combines visual and textual inputs.

Florence 2 is trained on FLD-5B, a large-scale dataset with 126M images and more than 5 billion annotations (text, region-text pairs, text-phrase-region triplets), this empowers

the model to learn multiple levels of spatial and semantic granularity (PLN).

PaliGemma

PaliGemma is a unified VLM [Beyer et al.2024] for general-use multi-modal functionalities based on SigLIP ViT-So400m image encoder [Zhai et al.2023] and the Gemma-2B decoder-only language model [Team et al.2024] with fewer than 3 billion parameters. The SigLIP encoder uses a contrastive pretraining method with sigmoid loss to produce image embeddings, achieving SOTA clip-level visual representation quality. The textual inputs to the Gemma-2B model are processed using a SentencePiece tokenizer, and then by autoregressive decoding, image tokens, and text prompts can be combined into a single sequence. A linear projection aligns the SigLIP output with Gemma-2 B's input space, facilitating seamless integration. The model is then further trained on higher-resolution images and domain-focused data in the final stages, which leads to better accuracy. For PaliGemma, an input sequence takes the following mathematical form:

$$\begin{aligned}\text{tokens} = [&\text{image tokens} \ldots, \text{BOS}, \text{prefix tokens} \ldots, \\ &\text{SEP}, \text{suffix tokens} \ldots, \text{EOS}, \text{PAD} \ldots]\end{aligned}$$

Here, BOS= Beginning of Sentence token, EOS = End of Sentence token, PAD= Padding token , SEP = Separator token

Experiments

Datasets

MIMIC-CXR Database v2.0.0 This publically available, large-scale MIMIC-CXR Database v2.0.0 is intended to facilitate medical imaging research, especially in the area of chest radiography. The dataset contains 377,110 JPG format images and structured labels derived from the 227,827 free-text radiology reports associated with these images [Johnson et al.2024].

Indiana University Chest X-rays and Reports (OpenI) Accessible via OpenI, the Indiana University Chest X-rays dataset is an extensive collection of chest radiographs with their corresponding diagnostic reports intended to facilitate medical imaging research and education. The dataset contains $7,470$ pairs of images and reports covering a broad spectrum of both common and uncommon thoracic disorders [Demner-Fushman et al.2015].

Chest X-ray images from both datasets were utilized as inputs, with the corresponding findings from each image serving as outputs to fine-tune the proposed model. The manual tags for finetuning the CNN classifier were provided by [Alfarghaly et al.2021].

Radiology Report Training Dataset To ensure the accuracy and relevance of the generated diagnostic reports, we pre-process the radiology reports by eliminating the usage of routine or repetitive words and phrases generally contained in chest x-ray reports. These words and phrases are common in normal (healthy) cases and can cause skewed results where the VLM generates many general terms to reduce the training loss. To address this, the dataset is adjusted by utilizing a filtering process with the use of the pre-trained **Llama 3.1-8B model** [Sam and Vavekanand2024] as shown in Fig.2.

Evaluation Metrics

To evaluate the performance of our fine-tuned VLM, we have used two different performance metrics, the BERTScore and F1-cheXbert. We opt for BERTScore as it is an automated assessment measure for text production. Similar to conventional metrics, BERTScore calculates a similarity score for every token in the candidate sentence in relation to each token in the reference phrase. Instead of relying on exact matches, the system calculates token similarity using contextual embeddings [Zhang et al.2020].

In contrast, the F1-cheXbert [Smit et al.2020] score utilizes the CheXbert[Smit et al.2020] transformer to output selected labels on both original and generated reports and then calculates the F1 score between these two sets of labels.

Implementation Details

The multi-label CNN-based classification model is fine-tuned using TensorFlow, following the methodology outlined in [Alfarghaly et al.2021]. With 32 images per batch and the Adam optimizer, the model is trained end-to-end using mini-batch gradient descent. Binary cross-entropy loss (eq.) was the loss that needed to be examined in this model. We left all the model parameters up for fine-tuning.

On the other hand, the VLMs are implemented using PyTorch. Both models are trained and tested on Intel(R) Xeon(R) CPUs and L4 GPUs provided by Google Colab notebooks. The loss function provided by the VLM packages are used for training the Vision-Language Models. The Adam optimizer, with an initial learning rate of 5×10^{-5} and a batch size of 16 for multi-label classifiers and 4 for the VLMs, is employed for training or fine-tuning. The VLMs were fine-tuned for 10 epochs.

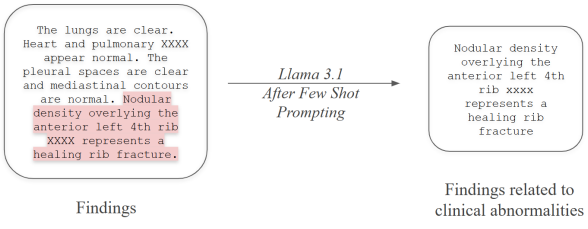


Figure 2: Radiology report pre-processing for training.

Results

Qualitative analysis: Extracting abnormal findings

A Llama 3.1 8b instruct model was used to extract sentences or phrases in the findings that indicated clinical abnormalities. Zero-shot prompting the Llama model with the instruction *"Extract the sentences or phrases that seem to indicate clinical abnormalities"* did not generate fruitful results. Thus, we consulted with experienced radiologists to create a few examples of extracted clinical abnormal findings from the rest of the report. With 12 examples, few-shot prompting the Llama model with the same instruction showed much better results, as shown in Fig. 2. Following this adjustment, two trained radiologists observed 30 reports each and affirmed that the Llama model efficiently eliminated the redundant information and effectively extracted the abnormal findings.

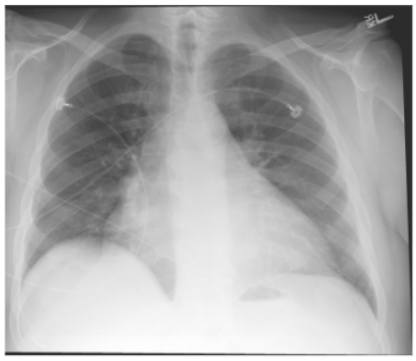
Vision Language model comparison and necessity of prompt guidance

While comparing the two VLMs, we also tested their performances with a generic prompt ("Write a Chest X-ray report") and a tag-specific prompt (such as "Write a Chest X-ray report mentioning cardiomegaly", "Write a Chest X-ray report mentioning pleural effusion, consolidation", etc.) Table 1 shows that Paligemma outperforms Florence-2 in all the BERTScore and the F1-cheXbert metrics. In the cases of both VLMs, we observe that tag-specific prompts enhance the performance of the model. For Florence-2, tag-specific prompts increase the BERTScore (at least 4%) the F1-cheXbert score from 0.5516 to 0.700; that is, it generates almost 14.84% more accurate reports. For Paligemma, both types of prompts result in similar BERTScore metrics, but the F1-cheXbert score shows that the model can generate 2.5% more accurate reports with specific prompt guidance. With the help of this comparative analysis, we can determine that Paligemma with tag-specific prompt guidance should be utilized in our pipeline.

Comparison with captioning baselines

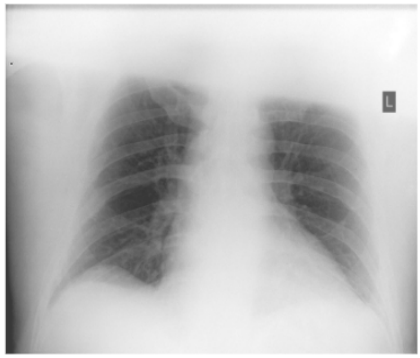
We compared our final methods to three types of models: (1) CNN encoders with RNN decoders. (2) CNN encoders with Transformer decoders. (3) Vision Transformer encoders with LLM decoders. For type-1, we look at attention-based CNN-LSTM architecture mentioned in ref [Sirshar et al.2022] [Liu et al.2021]. For comparison, we follow the architecture of [Sirshar et al.2022]. Type-2 entails CNN as encoders and transformers as decoders. A similar approach was taken in [Alqahtani et al.2024], [Alfarghaly et al.2021], [Mondal et al.2023]. We used the architecture CNX-B2 described in ref [Alqahtani et al.2024] for comparative evaluation. Transformers were used end-to-end for encoding and decoding in type-3 captioning baselines as seen in ref [Wang et al.2023],[Nicolson, Dowling, and Koopman2023]. For our quantitative analysis, we explored the architecture CvT2DistilGPT2 described in ref [Nicolson, Dowling, and Koopman2023].

Table 2 shows the results for both MIMIC-CXR and IU X-Ray. These results show that the proposed method



Reference findings	there is aortic xxxx vascular calcification
Generated findings	there is aortic vascular calcification

(a) Report Aid Example 1



Reference findings	low lung volumes, mild bibasilar atelectasis
Generated findings	low lung volumes mild bibasilar atelectasis and pleural thickening is noted

(b) Report Aid Example 2

Reference findings	nodular densities consistent with chronic granulomatous disease
Generated findings	evidence of prior granulomatous disease

(c) Report Aid Example 3

Figure 3: Example reports generated using the proposed RadTextAid model.

OpenI Dataset					
VLM	**Prompts**	**BERTScore Precision**	**BERTScore Recall**	**BERTScore F1**	**F1-cheXbert**
Florence-2	Generic	0.2467	0.2721	0.2507	0.5516
Florence-2	Tag-specific	0.2753	0.2874	0.2608	0.7000
Paligemma	Generic	0.3181	0.3334	0.3173	0.7666
Paligemma	Tag-specific	0.3273	0.3322	0.3156	0.7916

Table 1: Comparative Analysis of the Two Vision-Language Models.

OpenI Dataset				
Pipeline	**BERTScore Precision**	**BERTScore Recall**	**BERTScore F1**	**F1-cheXbert**
Attention-based CNN-LSTM	0.2174	0.2018	0.2123	0.5451
CNX-B2	0.2194	0.1608	0.2268	0.6484
CvT2-Distil-GPT2	0.3089	0.2996	0.3009	0.7600
Rad-TextAid (Proposed)	0.3273	0.3322	0.3156	0.7916
MIMIC-CXR Dataset				
Pipeline	**BERTScore Precision**	**BERTScore Recall**	**BERTScore F1**	**F1-cheXbert**
Attention-based CNN-LSTM	0.1871	0.1926	0.1852	0.5347
CNX-B2	0.1886	0.1606	0.2265	0.5333
CvT2-Distil-GPT2	0.2939	0.2927	0.2902	0.6813
Rad-TextAid (Proposed)	0.2813	0.3034	0.2975	0.6956

Table 2: Comparative Analysis Against Literature Captioning Baselines.

considerably improves the performance over the baseline methods. When compared with the best existing method, CvT2DistilGPT2, RadTextAid shows an absolute improvement of 4.8% in the NLG metric (BERTScore) and generates 3.16% more accurate reports according to the F1-cheXbert score. Similar trends are observed in the results for MIMIC-CXR. However, the overall performance of all models drops, likely due to the significantly higher average findings length in the MIMIC-CXR dataset compared to the IU X-Ray dataset.

Qualitative results

We present a few qualitative examples of RadTextAid to demonstrate its superiority in Fig. 3. The model, trained using our pipeline, has successfully detected and described all the clinical abnormalities as seen in the original report, as shown in Fig. 3(a) and 3(c). However, in the second example shown in Fig. 3(b), the model mistakenly detects and describes a clinical abnormality not mentioned in the reference report. This error may have occurred due to the image quality. A more robust preprocessing step of the image could help resolve this issue.

Conclusion

In this study, we demonstrate the effectiveness of a novel Vision-Language Model (VLM) -based framework for developing automated diagnostic reports aiming to assist chest X-ray image analysis and reporting. Initially, the model uses a CheXNet-121-based multi-label classifier, identifying 105 diagnostic tags from a chest X-ray image associated with different physiological conditions and pathologies. These tags, supplemented with the original image, are input to a multi-modal VLM (e.g., Florence 2, PaliGemma), combining image and semantic information to create a coherent, relevant diagnostic radiology report.

A key innovation of the approach is pre-processing the training corpus, using the Llama 3.1 pre-trained model, eliminating all the redundant and overused vocabulary from the text, meaning that the reports generated will be based on only the abnormalities carrying diagnostic information. This helps reduce duplication, improve clinical value, and provide a strong base for automatic report creation.

The experimental results indicate that our model achieves state-of-the-art performance based on BERTScore and F1-cheXbert metrics. These results demonstrate the model's capability to generate relevant and accurate diagnostic reports. Our comparative analysis reveals that the proposed framework surpasses existing architectures, especially in integrating multi-modal learning techniques, leading to improved outcomes. Additionally, the lightweight models have a unique advantage in terms of scalability in low-resource healthcare settings, where access to a cloud server may not always be available. Overall, the proposed method and the experimental results highlight the effectiveness of our approach in optimizing automatic radiological reporting, which could significantly enhance patient care.

References

[Alfarghaly et al.2021] Alfarghaly, O.; Khaled, R.; Elkorany, A.; Helal, M.; and Fahmy, A. 2021. Automated radiology report generation using conditioned transformers. *Informatics in Medicine Unlocked* 24:100557.

[Alqahtani et al.2024] Alqahtani, F. F.; Mohsan, M. M.; Alshamrani, K.; Zeb, J.; Alhamami, S.; and Alqarni, D. 2024. Cnx-b2: A novel cnn-transformer approach for chest x-ray medical report generation. *IEEE Access* 12:26626–26635.

[Beyer et al.2024] Beyer, L.; Steiner, A.; Pinto, A. S.; Kolesnikov, A.; Wang, X.; Salz, D.; Neumann, M.; Alabdulmohsin, I.; Tschannen, M.; Bugliarello, E.; Unterthiner, T.; Keysers, D.; Koppula, S.; Liu, F.; Grycner, A.; Gritsenko, A.; Houlsby, N.; Kumar, M.; Rong, K.; Eisenschlos, J.; Kabra, R.; Bauer, M.; Bošnjak, M.; Chen, X.; Minderer, M.; Voigtlaender, P.; Bica, I.; Balazevic, I.; Puigcerver, J.; Papalampidi, P.; Henaff, O.; Xiong, X.; Soricut, R.; Harmsen, J.; and Zhai, X. 2024. Paligemma: A versatile 3b vlm for transfer.

[Chen et al.2018] Chen, H.; Zhang, H.; Chen, P.-Y.; Yi, J.; and Hsieh, C.-J. 2018. Attacking visual language grounding with adversarial examples: A case study on neural image captioning.

[Demner-Fushman et al.2015] Demner-Fushman, D.; Kohli, M.; Rosenman, M.; Shooshan, S.; Rodriguez, L.; Antani, S.; Thoma, G.; and Mcdonald, C. 2015. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association : JAMIA* 23.

[Ding et al.2022] Ding, M.; Xiao, B.; Codella, N.; Luo, P.; Wang, J.; and Yuan, L. 2022. Davit: Dual attention vision transformers.

[Huang, Liu, and Weinberger2016] Huang, G.; Liu, Z.; and Weinberger, K. Q. 2016. Densely connected convolutional networks. *CoRR* abs/1608.06993.

[Johnson et al.2024] Johnson, A.; Lungren, M.; Peng, Y.; Lu, Z.; Mark, R.; Berkowitz, S.; and Horng, S. 2024. MIMIC-CXR-JPG - chest radiographs with structured labels.

[Kim et al.2024] Kim, Y.; Wu, J.; Abdulle, Y.; Gao, Y.; and Wu, H. 2024. Enhancing human-computer interaction in chest x-ray analysis using vision and language model with eye gaze patterns.

[Liu et al.2021] Liu, F.; Yin, C.; Wu, X.; Ge, S.; Zhang, P.; and Sun, X. 2021. Contrastive attention for automatic chest X-ray report generation. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 269–280. Online: Association for Computational Linguistics.

[Liu et al.2024] Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

[Mohsan et al.2023] Mohsan, M. M.; Akram, M. U.; Rasool, G.; Alghamdi, N. S.; Baqai, M. A. A.; and Abbas, M. 2023. Vision transformer and language model based radiology report generation. *IEEE Access* 11:1814–1824.

[Mondal et al.2023] Mondal, C.; Pham, D.-S.; Gupta, A.; Ghosh, S.; Tan, T.; and Gedeon, T. 2023. EfficienTransNet: An automated chest x-ray report generation paradigm. In *Proceedings of the 1st International Workshop on Multimodal and Responsible Affective Computing*. New York, NY, USA: ACM.

[Nicolson, Dowling, and Koopman2023] Nicolson, A.; Dowling, J.; and Koopman, B. 2023. Improving chest x-ray report generation by leveraging warm starting. *Artificial Intelligence in Medicine* 144:102633.

[OpenAI2024] OpenAI. 2024. Gpt-4o system card.

[Parag and Hardcastle2022] Parag, P., and Hardcastle, T. C. 2022. Shortage of radiologists in low to middle income countries in the interpretation of CT scans in trauma. *Banglad. J. Med. Sci.* 21(3):489–491.

[Rajpurkar et al.2017] Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D. Y.; Bagul, A.; Langlotz, C. P.; Shpanskaya, K. S.; Lungren, M. P.; and Ng, A. Y. 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *CoRR* abs/1711.05225.

[Sam and Vavekanand2024] Sam, K., and Vavekanand, R. 2024. Llama 3.1: An in-depth analysis of the next generation large language model.

[Singh et al.2022] Singh, A.; Raguru, J. K.; Prasad, G.; Chauhan, S.; Tiwari, P. K.; Zaguia, A.; and Ullah, M. A. 2022. Medical image captioning using optimized deep learning model. *Computational Intelligence and Neuroscience* 2022.

[Sirshar et al.2022] Sirshar, M.; Paracha, M. F. K.; Akram, M. U.; Alghamdi, N. S.; Zaidi, S. Z. Y.; and Fatima, T. 2022. Attention based automated radiology report generation using CNN and LSTM. *PLoS One* 17(1):e0262209.

[Smit et al.2020] Smit, A.; Jain, S.; Rajpurkar, P.; Pareek, A.; Ng, A. Y.; and Lungren, M. P. 2020. Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert.

[Srinivasan et al.2020] Srinivasan, P.; Thapar, D.; Bhavsar, A.; and Nigam, A. 2020. Hierarchical x-ray report generation via pathology tags and multi head attention. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*.

[Srinivasan et al.2021] Srinivasan, P.; Thapar, D.; Bhavsar, A.; and Nigam, A. 2021. Hierarchical x-ray report generation via pathology tags and multi head attention. In *Computer Vision – ACCV 2020*, Lecture notes in computer science. Cham: Springer International Publishing. 600–616.

[Team et al.2024] Team, G.; Mesnard, T.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Pathak, S.; Sifre, L.; Rivière, M.; Kale, M. S.; Love, J.; et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

[Wang et al.2023] Wang, Z.; Liu, L.; Wang, L.; and Zhou, L. 2023. METransformer: Radiology report generation by transformer with multiple learnable expert tokens.

[Wang et al.2024] Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Fan, Y.; Dang, K.; Du, M.; Ren, X.; Men, R.; Liu, D.; Zhou, C.; Zhou, J.; and Lin, J. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution.

[Xiao et al.2023] Xiao, B.; Wu, H.; Xu, W.; Dai, X.; Hu, H.; Lu, Y.; Zeng, M.; Liu, C.; and Yuan, L. 2023. Florence-2: Advancing a unified representation for a variety of vision tasks. *arXiv preprint arXiv:2311.06242*.

[Zhai et al.2023] Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training.

[Zhang et al.2020] Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. Bertscore: Evaluating text generation with bert.