Multimodal Fusion of CNN and LLM Embeddings for Chest X-Ray Classification

Anonymous Author(s)
Affiliation
Address
email

Abstract

Accurate detection of abnormalities in Chest X-rays is crucial for timely diagnosis and treatment. While multimodal models give us strong results, they are often computationally expensive to train. In this work, we propose a framework that fuses CNN embeddings from X-ray images with LLM-based semantic embeddings from radiology reports. The fused representation is processed through a multi-layer perceptron network to perform both binary as well as multilabel classification. Experiments show that our approach improves metrics compared to unimodal baselines while requiring fewer compute resources, making it suitable for a resource constrained environment.

Keywords: Large Language Models, Convolutional Neural Networks, Chest X-rays, Medical Imaging

1 Introduction

Advancements in healthcare AI have improved quick diagnosis. LLMs have been used to generate reports based on various medical data. CNNs have been widely applied for feature extraction and disease detection. CNNs are effective for extracting visual features from images, while textual information provides complementary semantic context. Similarly unimodal LLMs capture text semantics, but cannot leverage the information images can provide. Multimodal LLMs can harness both capabilities and give good results but they are more expensive to train. We propose a framework which combines a pretrained CNN and a pretrained LLM. The former helps in visual feature extraction while the latter helps in semantic information extraction from radiology reports. The embeddings are concatenated and then are fed as an input to a Multi Layer Perceptron network, performing classification of the abnormalities. The rest of the paper is organized as follows: Section 2 reviews related literature, Section 3 describes the methodology, dataset, and experiments, Section 4 presents the results and analysis, and Section 6 concludes the work.

2 Literature review

Extensive research has been conducted on the use of X-ray scans for disease classification. In particular, deep CNN models such as ResNet, VGG, and DenseNet variants are popular throughout the literature and have been reported to have high accuracy in disease classification tasks [1, 2, 3]. There have also been positive results in this area in using transformers for disease classification task. An example is the SwinCheX model that uses a Swin Transformer as a base for the multilabel classification task [4]. Similarly, another paper experimented with transformer based fusion model of multi view X-ray image classification [5]. LT-ViT is an example of a Vision Transformer based model for multi-label classification of chest X-rays [6]. Despite this research, X ray scans lack contextual information such as patient history and metadata that becomes important for an accurate diagnosis of a patient. They also suffer from ambiguity due to overlap of visual features for many diseases.

Recently, there has been increasing research on the use of text-based models for disease classification, where datasets include radiology reports, patient metadata, and other medical text. One study was based on use of BERT in classifying radiology reports related to Orthopaedic trauma for injuries [7]. MCN-BERT is another text-based model that combines medical concept normalization with BERT for symptom-based disease classification [8]. While text reports can provide radiologist's interpretation and subtle findings, text-based models suffer from performance degradation when there is missing context. Our work explores a framework by fusing CNN and LLM embeddings for chest X-ray classification.

3 Methodology

3.1 Dataset

For our experiments we used the publicly available Indiana Chest X-ray dataset [9]. The dataset contains multi-view Chest X-ray images and radiology reports of 3,955 patients. The Chest X-ray scans for each patient contain frontal and lateral views. The radiology report is presented in XML format with the information enclosed in tags. The patient history and findings is contained in four sections, namely a comparison section containing prior patient information, the indication section which details symptoms or reasons for the examination, the findings section which lists radiological observations and the impression section outlines the final diagnosis. Additionally, the reports contain a 'MeSH' tag that specifies the conditions such as Cardiomegaly, Calcinosis and so on. For our experiments we extract the labels for our dataset from the MeSH tag of the radiology report.

On analysis of the reports, we found that many of the data points included explicit mention of the disease being present in the patient along with the mention of symptoms and conditions. To prevent data leakage that would cause the model to cheat, we masked all mentions of any of the diseases in the text.

3.2 Experiments

The initial experiments were carried out on the task of binary classification of data points into Normal and Abnormal, followed by experiments on multilabel classification of data points into Atelectasis and Calcinosis. For binary classification, first experiment involved averaging the CNN embeddings of frontal and lateral images and then concatenating it with the semantic embedding of the report. The second experiment is same as above but we concatenate all the three embeddings. For the multilabel classification, the first step was to extract image and text embeddings using pretrained models for which torchXrayVision's DenseNet121 and the BioBERT model were used, respectively. These embeddings were then passed through a simple MLP layer for classification.To evaluate the performance of various modalities, we conducted the following experiments:

1. Experiments using only multi-view images (unimodal): Multi-view images were fused together using a small attention fusion module on the embeddings and passed through an MLP layer.
2. Experiments using only Text (unimodal): the text embeddings were directly passed through an MLP layer.
3. Experiments combining Image and text data (multimodal): the image and text embeddings were concatenated and passed through an MLP layer.

The attention module was used to teach the model which features are most informative and combine them. The attention fusion module projects inputs into queries, keys, and values via linear layers, computes scaled dot-product attention with masking, applies softmax to obtain attention weights, and combines the weighted values into a single fused embedding normalized by the number of valid inputs.

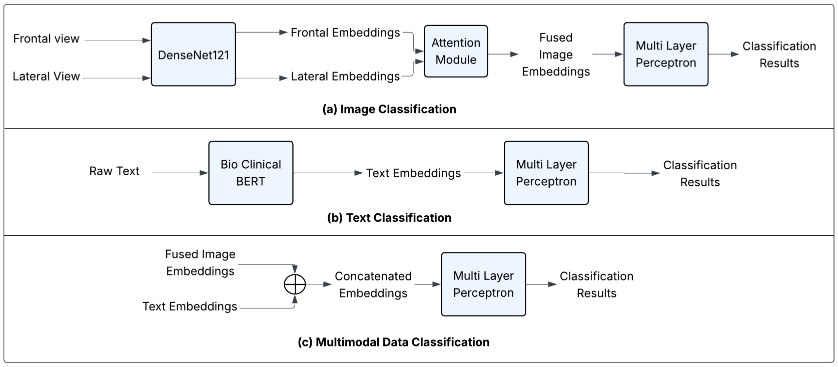

Figure 1: Block diagram for multilabel classification

4 Results and metrics

In the initial experiments involving binary classification, we observed improvement in metrics when we concatenated all the three embeddings compared to when we averaged the embeddings and concatenated it with the text embeddings(Table 1. Table 2 reports the results of the three experimental settings involving multi class classification. The multimodal model consistently achieves superior performance compared to the unimodal baselines. In particular, it records substantially higher AUROC, Precision, Recall, and F1 scores, indicating that leveraging multimodal information provides a clear advantage for the classification task. These improvements are statistically significant across runs, underscoring the robustness of the ensemble approach.

Table 1: Performance comparison of binary classification experiments

Fusion Strategy	Accuracy	F1-score	AUC
Averaging + Concatenation	93.8%	0.95	0.98
Concatenation Only	94.4%	0.96	0.98

Table 2: Performance comparison of multilabel classification experiments

Modality	AUROC	Precision	Recall	F1 Macro
Image	0.81 ± 0.03	0.76 ± 0.04	0.76 ± 0.03	0.76 ± 0.03
Text	0.86 ± 0.01	0.80 ± 0.01	0.80 ± 0.01	0.80 ± 0.01
Multimodal	0.92 ± 0.01	0.87 ± 0.02	0.87 ± 0.01	0.87 ± 0.01

Additionally, a small experiment was carried out to check the complementary nature of the text models and image models. In Figure 2 we use a Venn diagram to illustrate the complementarity between models in terms of exact match accuracy (all labels predicted correctly for an instance). This visualization is intended to provide qualitative intuition.

On the evaluation set of 236 (sample, class) pairs, the text-only model correctly classified 189 instances, while the image-only model correctly classified 186. Of these, 153 pairs were predicted correctly by both models. Notably, 36 predictions were unique to the text-only model and 33 were unique to the image-only model, indicating that each modality captures complementary information. Only 14 pairs were misclassified by both models. These results suggest that integrating text and image representations can leverage the complementary strengths of the two modalities, thereby reducing errors and improving overall performance.

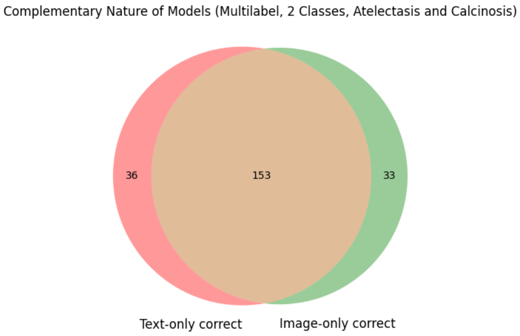

Figure 2: Venn diagram depicting the prediction overlap between text and image classification

5 Limitations

Our study is limited by the relatively small size of the Indiana University dataset, which may affect generalization. Our experiments focused only on a small subset of conditions and employed a simple concatenation-based fusion, without comparisons to more advanced multimodal architectures. Class imbalance remains a challenge, and further work is needed to validate robustness across diverse datasets and real-world settings.

6 Conclusion

We presented a multimodal framework combining CNN-based image features with LLM-based text embeddings for chest X-ray classification. Experiments show that fusion improves over unimodal baselines but struggles with rare conditions due to class imbalance. Our results highlight the promise of multimodal fusion for practical medical applications, while underscoring the need for better strategies to address class imbalance. Furthermore, real-world applications should carefully consider potential biases in training data, fairness across demographic groups, and the risks of misclassification in clinical practice.

References

[1] Wanni Xu, You-Lei Fu, and Dongmei Zhu. Resnet and its application to medical image processing: Research progress and challenges. *Computer Methods and Programs in Biomedicine*, 240:107660, 2023.

[2] Ankita Shelke, Madhura Inamdar, Vruddhi Shah, Amanshu Tiwari, Aafiya Hussain, Talha Chafekar, and Ninad Mehendale. Chest x-ray classification using deep learning for automated covid-19 screening. *SN Computer Science*, 2(4), May 2021.

[3] Keno K. Bressem, Lisa C. Adams, Christoph Erxleben, Bernd Hamm, Stefan M. Niehues, and Janis L. Vahldiek. Comparing different deep learning architectures for classification of chest radiographs. *Scientific Reports*, 10(1), August 2020.

[4] Sina Taslimi, Soroush Taslimi, Nima Fathi, Mohammadreza Salehi, and Mohammad Hossein Rohban. Swinchex: Multi-label classification on chest x-ray images with transformers, 2022.

[5] Dongkyun Kim. Chexfusion: Effective fusion of multi-view features using transformers for long-tailed chest x-ray classification. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. IEEE, October 2023.

[6] Umar Marikkar, Sara Atito, Muhammad Awais, and Adam Mahdi. Lt-vit: A vision transformer for multi-label chest x-ray classification. In *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, October 2023.

[7] A.W. Olthof, P. Shouche, E.M. Fennema, F.F.A. IJpma, R.H.C. Koolstra, V.M.A. Stirler, P.M.A. van Ooijen, and L.J. Cornelissen. Machine learning based natural language processing of radiology reports in orthopaedic trauma. *Computer Methods and Programs in Biomedicine*, 208:106304, September 2021.

[8] Esraa Hassan, Tarek Abd El-Hafeez, and Mahmoud Y. Shams. Optimizing classification of diseases through language model analysis of symptoms. *Scientific Reports*, 14(1), January 2024.

[9] Indiana University. Chest x-rays dataset. Available at: `https://openi.nlm.nih.gov/faq?download=true`, 2021.