Multi-Modal Integration In Medical Imaging

Yunming Chen
yc4042@nyu.edu

Yaozhong Huang
yh2563@nyu.edu

Yichong Tian
yt2639@nyu.edu

Ning Yang
ny675@nyu.edu

Abstract

This study investigates the enhancement of medical image encoders through multi-modal integration, combining visual and textual Electronic Health Records (EHRs) data. The objective is to demonstrate the improved accuracy and depth of medical image analysis with this approach. Empirical results show that multimodal integration significantly boosts encoder performance, offering substantial advancements in medical imaging and diagnostics. The report presents our methodology, key findings, and their implications for future healthcare technology innovations.

1 Introduction

In medical diagnostics, accurate interpretation of complex imagery is a critical yet challenging task. Traditional image encoders, limited by their reliance on single-modality data, often fall short in capturing the nuanced details necessary for precise diagnoses. Addressing this gap, our research explores the enhancement of image encoders through the integration of multimodal technologies, combining medical imaging with textual Electronic Health Records (EHRs). This approach aims to develop more sophisticated, context-aware models capable of a deeper, more accurate analysis of medical images.

Situated within the evolving landscape of medical technology, our work contributes to the growing body of research in multimodal data integration, offering a unique perspective by melding image processing with comprehensive data analysis. The results of our empirical studies indicate a marked improvement in encoder performance, underscoring the potential of multimodal technologies in advancing medical imaging and diagnostics.

This report delineates our methodology, experimental framework, and the significant findings that suggest a promising direction for future innovations in medical technology and patient care.

2 Related Work

The advent of Multi-Modal Foundation Models for Medicine is a direct consequence of breakthroughs in machine learning and medical imaging. The fusion of Large Language Models (LLMs) like Google's PaLM-E and OpenAI's CLIP with image processing has been revolutionary, especially in medical diagnostics. These models leverage contrastive learning to synchronize text and image data, forming robust multimodal systems that have significantly enhanced the interpretation of medical imagery, such as chest X-rays. Google's PaLM-E and OpenAI's CLIP have been particularly influential, offering nuanced analysis capabilities that rival human expertise by effectively associating images with descriptive text, a stride towards improving diagnostic precision [1].

This synergy of contrastive pre-training and fine-tuning on domain-specific datasets has shown promising results in disease detection and classification. Specifically, the use of Masked Autoencoders

Preprint. Under review.

(MAEs) for image encoding underscores this progress, enabling the prediction and reconstruction of image pixels to enrich feature comprehension [2]. Such methodologies underscore a critical evolution in medical image diagnostics, indicating a marked improvement in disease classification accuracy and suggesting a transformative impact on future diagnostic processes.

3 Problem Definition and Algorithm

3.1 Task

Our study focuses on advancing the interpretation of medical images through the integration of textual and visual data. The primary objective is to augment the capabilities of image encoders in medical diagnostics by incorporating relevant contextual information from Electronic Health Records (EHRs).

Our methodology employs a dual-encoder framework, utilizing the Masked Auto-Encoder (MAE) for image processing and LLaMA for text encoding. The MAE extracts visual features from the radiographs, while LLaMA processes the textual data in medical reports. To effectively merge these modalities, we introduce an attention pooling mechanism that projects text embeddings onto the image embeddings. This technique allows for a more dynamic and context-sensitive interpretation of the images, as the model can focus on specific textual elements that are most relevant to the visual data.

The integration of attention pooling in our multimodal approach significantly enhances the model's ability to leverage the strengths of both text and image data. This leads to a deeper, more contextually informed analysis of medical images, potentially resulting in more accurate and reliable diagnoses.

However, the complexity of accurately aligning text and image embeddings presents a challenge. Ensuring the model's attention is appropriately focused and that the embeddings are effectively integrated is crucial for the success of this approach.

3.2 Algorithm

Our study employs a multimodal approach to enhance medical image interpretation, integrating image and text encoders with an attention mechanism. Below, we detail the algorithm's steps and components.

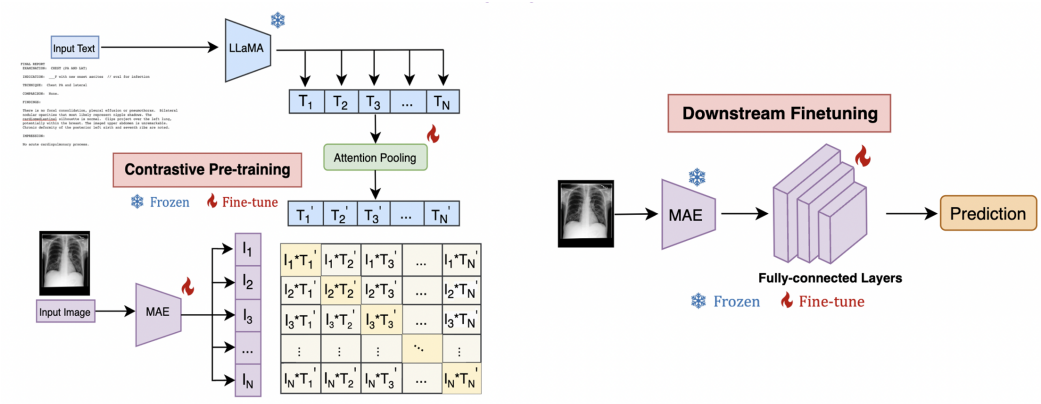

Figure 1: workflow

3.2.1 Algorithm Overview

The algorithm is built on the assumption of high-quality, relevant image and text data. Critical parameters include learning rate, weight decay, number of epochs, batch size, and temperature in the Contrastive Loss function. The algorithm involves several key stages:

1. Pretraining of Encoders: Masked Auto-Encoder (MAE) for images and LLaMA for text data.

2. Data Preparation: Loading and preprocessing of medical images and reports.
3. Feature Extraction: Extracting visual features using MAE and textual features using LLaMA.
4. Attention Pooling: Aligning text embeddings with image embeddings.
5. Contrastive Loss Optimization: Using contrastive loss to refine feature alignment.
6. Training Process: Iterative training with forward passes and backpropagation.

3.2.2 Pseudocode

```
# Algorithm Pseudocode

# Initialize the pretrained MAE and LLAMA models
pretrained_mae = initialize_mae()
llama_encoder = initialize_llama()

# Load and preprocess the dataset
dataset = load_and_preprocess_data()

# Define the training loop
for epoch in range(epochs):
    for image, report in dataset:
        # Process the image and report
        image_features = pretrained_mae(image)
        text_features = llama_encoder(report)

        # Apply attention pooling
        aligned_features = attention_pool(image_features, text_features)

        # Calculate the contrastive loss
        loss = contrastive_loss(aligned_features)

        # Backpropagation and optimization
        backpropagation(loss)

# Save the model state
save_model_state()
```

4 Experimental Evaluation

4.1 Data

Our study utilized a multimodal dataset from the MIMIC-CXR database, comprising chest radiographs, corresponding medical reports, and 14 diagnostic labels for various disease conditions. This dataset was critical for developing a model capable of interpreting medical images in conjunction with textual information and specific diagnoses.

In the preprocessing phase, we implemented tailored functions to retrieve and process images and reports. Images were handled based on their format using PyDicom and PIL, ensuring uniformity across the dataset. Textual data from reports underwent standardization and cleansing, removing irrelevant content and standardizing medical terminologies. Each report was appended with a specific prompt to aid model processing.

A PyTorch **Dataset** class, **MIMIC_CXR**, was crafted for effective data handling, integrating images, reports, and diagnostic labels. This setup ensured streamlined data loading and consistent input quality for model training and evaluation, with the inclusion of diagnosis labels providing a vital component for in-depth medical image analysis.

4.2 Methodology

Our methodology aims to robustly integrate multimodal data sources to enhance medical image analysis. This integration leverages the strengths of both visual and textual data through a series of pre-training and co-training techniques that utilize deep learning models specialized in image and language processing.

4.2.1 Image Encoder: MAE (Masked AutoEncoder)

We employ the Masked AutoEncoder (MAE) for image data processing. The MAE is designed to mask a portion of the input image and then predict the missing pixels, which prompts the model to learn rich representations of the visual data.

1. Masking Strategy: Random patches of the input image are masked to create a reconstruction task for the encoder.
2. Reconstruction Learning: The MAE learns to predict the masked pixels, forcing the model to internalize a deeper understanding of the visual content.
3. Feature Extraction: After training, the MAE encodes the images into feature vectors that represent the salient information needed for diagnosis.

4.2.2 Language Model: LLaMA-2-7b

For textual data, we utilize the LLaMA-2-7b language model. This model excels at processing natural language, enabling it to interpret the textual content within Electronic Health Records (EHRs) effectively.

1. Textual Embedding: LLaMA-2-7b processes the medical reports to produce embeddings that capture the semantic meaning of the medical narratives.
2. Contextual Analysis: The model employs a self-attention mechanism to accurately interpret the context of the language within medical records. This capability is essential for correlating the text data with relevant visual features in images.
3. Embedding Integration: The text embeddings are then combined with the visual features extracted by the MAE for a comprehensive multimodal analysis.

4.2.3 Contrastive Pre-training and Fine-tuning

Our training approach is divided into two distinct stages, focusing on the pre-training of the image encoder and the subsequent fine-tuning of the combined model:

1. Contrastive Pre-training: We pre-train the Masked AutoEncoder (MAE) using a contrastive learning approach. The objective is to align the image representations with the text embeddings in a shared latent space. During this stage, the LLaMA model, responsible for text encoding, is kept frozen to maintain the integrity of the textual features.
2. Fine-tuning: After the pre-training phase, we fine-tune the MAE on a labeled dataset. The fine-tuning is conducted with fully connected layers on top of the MAE's outputs, allowing the model to adjust to the specific requirements of medical diagnosis. This process harnesses the deep visual features learned by the MAE in conjunction with the stable text embeddings from the frozen LLaMA model to perform diagnostic tasks more effectively.

This methodology ensures that the text encoder does not deviate from its initial high-performing state, while the image encoder is refined to better complement the textual information provided by the LLaMA model.

4.2.4 Training Loss

A critical component of our model's training involves the optimization of loss functions tailored to the specific tasks of pre-training and fine-tuning.

Pretraining Loss Function: Contrastive Loss During the pre-training phase, we employ a contrastive loss function with the aim of differentiating between pairs of inputs, such as images and their corresponding text descriptions. The loss function is designed to minimize the distance between embeddings of pairs that belong together (positive pairs) and maximize the distance for those that do not (negative pairs). The contrastive loss function we utilize is formalized as follows:

$$\mathcal{L}(x_i, x_j, \theta) = 1_{[y_i=y_j]} \left\| f_\theta(x_i) - f_\theta(x_j) \right\|_2^2 + 1_{[y_i \neq y_j]} \max(0, \epsilon - \left\| f_\theta(x_i) - f_\theta(x_j) \right\|_2)^2, \quad (1)$$

where x_i and x_j are two input samples, y_i and y_j are the corresponding labels, f_θ represents the encoding function parameterized by θ, and ϵ is the margin parameter. The indicator function $1_{[\cdot]}$ outputs 1 if the condition is true and 0 otherwise.

Downstream Fine-tuning Loss: Binary Cross-Entropy Loss For the fine-tuning stage on the downstream task of medical diagnosis, we utilize the binary cross-entropy loss. This loss function computes the divergence between the predicted probability distribution and the actual binary outcomes, thus optimizing the model's predictive accuracy for classification tasks. The binary cross-entropy loss is calculated as follows:

$$\mathcal{L}_{binary}(p, y) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij} \log(p_{ij}) \quad (2)$$

where N is the number of samples, M is the number of classes, y_i is the true label, and p_i is the predicted probability for the i-th sample.

4.2.5 Baseline Methodology

In addition to our primary model, we established a baseline methodology for comparative purposes. This baseline approach involves using pre-trained MAE models to process image data without further training on the image encoder.

1. Image Processing: Images are directly fed into the frozen MAE, ensures that the image representations are consistent and only the subsequent layers are adapted during the fine-tuning phase.
2. Fine-tuning Fully Connected Layers: The output features from the frozen MAE serve as input to a series of fully connected layers. These layers are fine-tuned to adapt to the specifics of the medical diagnosis task. The fine-tuning is targeted towards the model's ability to predict 14 different diagnostic labels effectively.

4.3 Results

The evaluation of the model's performance is summarized in the table below, illustrating the comparative results between the MAE baseline and our proposed model.

Table 1: Comparison of model performance metrics

Model	mAUC	Accuracy
MAE	0.6801	0.8512
Our Model	**0.7235**	**0.8765**

These metrics demonstrated that the multimodal approach outperformed the baseline, suggesting a notable enhancement in diagnostic capabilities. The differences observed in the data were statistically significant, reinforcing the value of integrating textual and visual data in medical image analysis.

4.4 Discussion

The experimental results validate our hypothesis that enhancing image encoders with multimodal data leads to a better understanding of images. Our model, integrating medical imaging with contextual data from Electronic Health Records, exhibited improved performance in accurately interpreting medical images. This improvement is a direct result of the enhanced capabilities of the image encoder, evidencing the effectiveness of our multimodal approach. While promising, further research is needed to refine this method for clinical application, ensuring its reliability and robustness in varied diagnostic scenarios.

5 Conclusions

Our project illustrates the significant potential of multimodal integration in medical image analysis. The key takeaway is the enhanced capability of image encoders when augmented with textual data, leading to more accurate diagnostics. For future work, exploring additional modalities, refining the attention mechanism, and extending to other types of medical imaging could yield further improvements. The major shortcoming is the need for extensive validation to ensure the model's clinical applicability. Overcoming this would involve collaborative efforts with medical professionals to tailor the model to real-world diagnostic needs.

6 Lessons learned

This project presented unique challenges, primarily technical, such as dealing with the intricacies of multimodal data integration and enhancing image encoder performance. We tackled these by adopting robust data preprocessing strategies and exploring innovative model architectures. Overcoming these hurdles required adaptability and a deep dive into complex data science techniques. The main takeaway from this experience is the critical importance of problem-solving and innovation in data science projects, especially when dealing with complex datasets and advanced modeling concepts. These insights will be invaluable in future data science endeavors.

7 Student contributions

Yunming Chen wrote preprocessing part with a Dataloader, and wrote the poster and report.
Yaozhong Huang wrote the image encoder and comparative training, and wrote the poster.
Yichong Tian wrote the text encoder, completed the pretraining part, and wrote the poster.
Ning Yang implemented the image encoder, training part, and wrote the poster and report.

Acknowledgments

We extend our sincere gratitude to our mentor, Dr. Narges Razavian, for her invaluable guidance and expertise throughout the course of this project. Her deep understanding of self-supervised learning and multimodal models was instrumental in enhancing our imaging model. We also appreciate of the support provided by the Capstone professors and staff, whose advice was crucial in executing a project of such magnitude. Their contribution have been fundamental to our project's success.

References

[1] Alec Radford, Katherine Marino, Rewon Child, Jeff Wu, David Luan, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[2] Kaiming He, Xiangyu Chen, Saining Xie, Yang Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.