Multi-Modal Foundation Models for Medicine

Yaozhong Huang, Yunming Chen, Yichong Tian, Ning Yang
Mentor: Dr. Narges Razavian

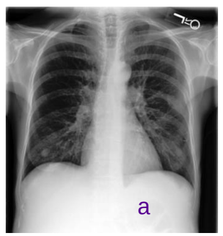
Context

Primary Goal 🎯: To enhance the robustness and informativeness of medical image representations by implementing multi-modality pre-training and co-training techniques that integrate image encoders with large language model capabilities, thereby utilizing the comprehensive data available within Electronic Health Records.

Methodology 🧠:

1. Generate image embedding using Image Encoder.
2. Pass the text through the of LLM and use the last layer of generated embedding as the training target for image embedding.
3. Employ Contrastive Learning to train the image encoder and attention pooling layers.

Data 🔍: The MIMIC Chest X-ray (MIMIC-CXR) Database is a large publicly available dataset of chest radiographs with free-text radiology reports and structured labels. It contains totally 377,110 images, and 227,835 corresponding radiographic studies. In addition, there are 14 dimensions in radiographs with labels.



FINAL REPORT
EXAMINATION: CHEST (PA AND LAT)

INDICATION: ___F with new onset ascites // eval for infection

TECHNIQUE: Chest PA and lateral

COMPARISON: None.

FINDINGS:

There is no focal consolidation, pleural effusion or pneumothorax. Bilateral nodular opacities that most likely represent nipple shadows. The cardiomediastinal silhouette is normal. Clips project over the left lung, potentially within the breast. The imaged upper abdomen is unremarkable. Chronic deformity of the posterior left sixth and seventh ribs are noted.

IMPRESSION:

No acute cardiopulmonary process.

b

subject_id	study_id	Atelectasis	Cardiomegaly	Consolidation	Edema	Enlarged Cardiomediastinum	Fracture	Lung Lesion	Lung Opacity	No Finding	Pleural Effusion	Pleural Other	Pneumonia	Pneumothorax	Support Devices
10000032	50414267									1.0					
10000032	53189527									1.0					

c

Figure 1: Data Overview (a) Chest X-ray (b) Radiology Report (c) Diagnostic labels

Related Works

Google's PaLM-E and OpenAI's CLIP represent transformative advancements in multi-modal learning within machine learning. PaLM-E's unique approach to interpreting images as sequences parallels text processing and showcases the expanding adaptability of LLMs. CLIP complements his by employing contrastive learning to associate images with textual descriptions, thereby achieving an interpretation of visual content that mirrors human cognition.

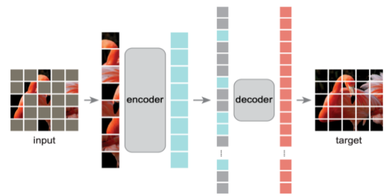
In medical imaging, these LLMs signal a promising direction for enhancing image interpretation. The integration of detailed HER data with image analysis could lead to greater diagnostic accuracy. The Masked Autoencoder (MAE), with its proficiency in reconstructing images from incomplete data, stands out for its potential to discern intricate patterns vital in medical diagnostics. Our research leverages MAE's capabilities to address the complex nature of medical images, aiming to strengthen diagnostic processes with comprehensive data insights.

References:

1. Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, Ross Girshick. 2021. Masked Autoencoders Are Scalable Vision Learners. arXiv:2111.06377v3 [cs.CV] https://arxiv.org/abs/2111.06377
2. Alec Radford et al. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV] https://arxiv.org/abs/2103.00020

Approaches

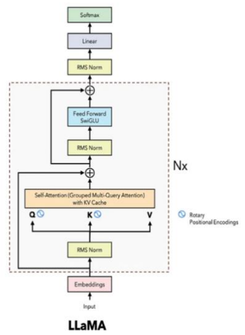
Image Encoder: MAE (Masked AutoEncoder)



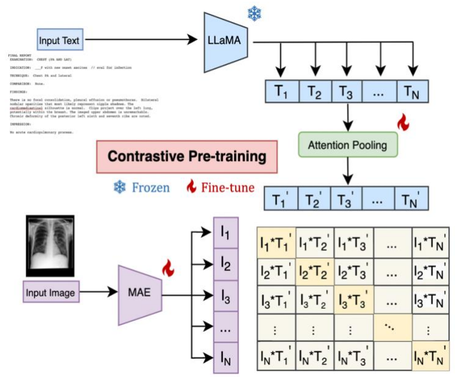
- Mask random patches of an input image.
- The encoder is trained to predict and reconstruct the missing pixels.
- Aid in learning deep image features and understanding the image's structure and context.

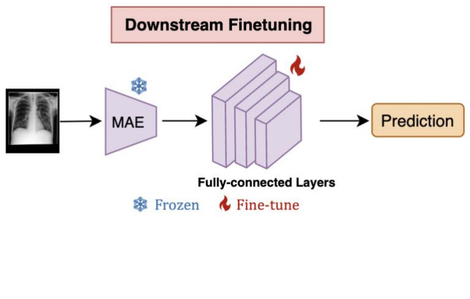
Large Language Model: Llama-2-7b

- Build on a self-attention framework, enabling dynamic importance weighting of words in a sentence for nuanced language processing.
- Integrate RMSNorm and SwiGLU activation for stability and performance improvements.
- Incorporate rotary positional embeddings to better understand and encode textual context.



Workflow

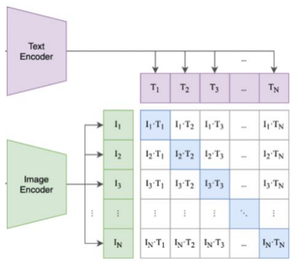




Training Loss

Pretraining loss function: Contrastive Loss

- Aim at differentiating between pairs of inputs, minimizing the distance between embeddings of similar items while maximizing the distance for dissimilar ones, fostering distinguishable feature representations.
- Formula:



$$\mathcal{L}_{\text{cont}}(\mathbf{x}_i, \mathbf{x}_j, \theta) = \mathbb{1}[y_i = y_j] \|f_\theta(\mathbf{x}_i) - f_\theta(\mathbf{x}_j)\|_2^2 + \mathbb{1}[y_i \neq y_j] \max(0, \epsilon - \|f_\theta(\mathbf{x}_i) - f_\theta(\mathbf{x}_j)\|_2)^2$$

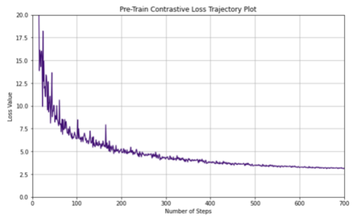
Downstream Classification loss function: Binary Cross-Entropy Loss

- Calculate the divergence between the predicted probabilities and the actual binary outcomes, optimizing the model's predictive accuracy for classification.
- Formula:

$$\text{logloss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij})$$

- N is the number of rows
- M is the number of classes

Results



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

Table 1: mean AUC and Accuracy scores comparison between MAE and Our Model (trained by MIMIC-CXR)

Figure 2: Train curve for self-supervised pre-training, computed by contrastive loss

[Note: Accuracy is determined by considering labels with a probability above 0.5 as indicative of the presence of disease. A case is deemed accurately classified only if all labels are correctly identified. Mean AUC represents the average Area Under the Curve (AUC) across all classes.]

Discussion:

- **Enhanced Performance:** Our model demonstrates superior performance in medical image classification, achieving a higher mean AUC and accuracy than the MAE. This suggests that our model has a better capability to distinguish between various classes of medical images.
- **Effective Learning Curve:** The contrastive learning loss graph shows a steep decline, indicating rapid learning in the initial phase, followed by a plateau which suggests that the model has reached a stable state of learning with less room for improvement.
- **Implications for Medical Diagnostics:** The improved performance of our model in both discriminative feature learning and classification accuracy holds significant potential for advancing medical diagnostics, offering a promising direction for future research and application.

Impact & Future Work

Impact: Our project delves into the use of Large Language Models (LLMs) for self-supervised learning to improve medical image representation. By leveraging the comprehensive information processing power of LLMs, we aim to train more effective medical image encoders, potentially leading to earlier and more accurate disease detection. This approach not only enhances multimodality in medical diagnostics but also promises to broaden the availability of sophisticated diagnostic tools.

Future Work:

1. **Extensive Training with Diverse Data:** Commit to further training our model with an expanded dataset variety and increased training duration to enhance diagnostic accuracy and the model's adaptability to various medical conditions and patient demographics.
2. **Automated Textual Reporting:** Explore the potential for automatically generating descriptive medical reports from images, facilitating preliminary analyses and supporting healthcare professionals in their diagnostic workflows.