MIMIC-CXR: Chest X-ray Image Classification and Report Generation

DS 5220 Supervised Learning Fall 2024

Professor **Hongyang Ryan Zhang**

December 9th, 2024

Yuandi Tang
Northeastern University
440 Huntington Ave
Boston, MA 02115
tang.yuand@northeastern.edu

Alexander Koehler
Northeastern University
440 Huntington Ave
Boston, MA 02115
koehler.ale@northeastern.edu

Feng-Jen Hsieh
Northeastern University
440 Huntington Ave
Boston, MA 02115
hsieh.fe@northeastern.edu

Abstract

*This study utilizes the **MIMIC-CXR** dataset to develop an advanced automated system for chest X-ray analysis. Our primary objectives include classifying common chest abnormalities and generating radiologist-style reports. We employ state-of-the-art methodologies, such as advanced CNN architectures including **ResNet18**, **ResNet50**, and **VGG16** for image classification, and **LLaMA-3.2-11B-Vision-Instruct** for multimodal alignment and text generation. The two-stage process includes CNN tuning for optimal model performance, followed by comprehensive evaluations. This research enhances healthcare efficiency by streamlining chest X-ray interpretations and improving diagnostic accuracy.*

1. Introduction

Chest X-rays are pivotal for diagnosing various thoracic conditions. Yet, their interpretation is prone to error, especially for subtle or rare abnormalities. The **MIMIC-CXR** dataset offers a substantial resource for developing AI systems to assist radiologists by reducing their workload and improving diagnostic accuracy.

This project employs CNNs and the advanced multimodal model **LLaMA-3.2-11B-Vision-Instruct** to process chest X-ray images and generate diagnostic reports. By combining image classification with NLP-driven insights, our system aims to enhance clinical workflows and offer a scalable solution for medical imaging challenges.

2. LLaMA-3.2-11B-Vision-Instruct

LLaMA-3.2-11B-Vision-Instruct represents a cutting-edge approach to multimodal AI, integrating vision and language to interpret medical images and generate coherent diagnostic reports. Key features include:

- **Multimodal Understanding**: Processes both images and text, enabling tasks such as image-based question answering and radiology report generation.
- **Pretrained Knowledge**: Trained on extensive datasets, the model excels at medical terminology and complex visual-textual correlations.
- **Contextual Adaptability**: Combines patient history and clinical information with visual data for enhanced diagnostic relevance.
- **Scalability**: Demonstrates zero-shot and few-shot learning capabilities, making it adaptable for varied clinical use cases.

3. Literature Review

Recent advancements in AI for medical imaging have opened new avenues for automating diagnostics and report generation. The integration of deep learning techniques for chest X-ray analysis has made significant strides, offering the potential for machine-driven diagnostics that rival human experts. Below, we explore some of the foundational and recent developments in this field.

3.1. Deep Learning in Chest X-Ray Analysis

CheXNet, introduced by Rajpurkar et al., represents a significant milestone in deep learning for chest X-ray analysis. This 121-layer CNN, based on the DenseNet architecture, was trained on the extensive **ChestX-ray14** dataset. **CheXNet** achieves diagnostic accuracy surpassing that of practicing radiologists, particularly in pneumonia detection. Its success demonstrates the potential of CNNs in critical

healthcare applications and sets a benchmark for subsequent research in automated medical imaging [1]. This was the first notable AI system to outperform radiologists in specific diagnostic tasks, laying the groundwork for many follow-up studies.

3.2. Multimodal AI Systems in Biomedicine

Recent developments have expanded beyond image classification to multimodal AI systems that integrate both visual and textual data. **LLaVA-Med**, for instance, is designed to perform a variety of medical tasks by leveraging both image and language data. Tasks such as visual question answering and diagnostic summarization in medical contexts can benefit from this type of AI integration. LLaVA-Med achieves multimodal alignment efficiently by leveraging large-scale pretrained models that bridge the gap between vision and language. Such systems have demonstrated impressive adaptability across a wide range of biomedical applications and show great promise in automating medical workflows [2].

3.3. Vision-Language Pre-Training for Medical Imaging

The integration of vision and language through pretraining has shown promising results in medical imaging tasks. Models trained with both visual and textual objectives have exhibited a much stronger ability to understand visual data when paired with appropriate text. For instance, the introduction of vision-language transformers (e.g., **VisualBERT**) has facilitated better alignment of image features with corresponding descriptions in radiology reports. This integration not only aids in more effective report generation but also enhances the overall quality of diagnostics [3]. These models, which extend the BERT architecture to include image features, offer a powerful mechanism for bridging the semantic gap between visual and textual modalities.

3.4. Transfer Learning in Healthcare Applications

Transfer learning has proven valuable in addressing emergent medical challenges. Models like DenseNet and ResNet, which were pre-trained on general image datasets such as ImageNet, have been fine-tuned for specific medical tasks such as COVID-19 diagnosis, demonstrating state-of-the-art performance. The ability to leverage pre-trained models significantly reduces the need for large labeled datasets, which are often scarce in healthcare, while improving model robustness in real-world applications. Transfer learning has been particularly beneficial in rapidly deploying diagnostic tools during healthcare crises, such as the COVID-19 pandemic [4]. This highlights the efficiency and adaptability of transfer learning techniques in the context of medical imaging.

3.5. Gaps and Future Directions

Despite significant progress, several challenges remain:

- **Unified Pipelines:** There is a growing need for AI models that integrate the entire diagnostic workflow, from image analysis to report generation, in a seamless manner. This holistic approach would mimic radiologists' workflows more effectively.
- **Generalization:** Improving model performance across diverse datasets and medical imaging modalities remains a challenge. Currently, many models are highly specialized to specific datasets, limiting their broader applicability.
- **Dataset Bias:** Heavy reliance on specific datasets introduces bias, making it difficult for models to generalize across varying demographics and imaging techniques.
- **Rare Disease Detection:** Models often perform poorly when diagnosing rare conditions due to the inherent imbalance in datasets.

In response to these gaps, our work proposes a unified framework that integrates advanced CNN architectures with state-of-the-art language models. This integrated approach aims to provide a comprehensive solution for automated chest X-ray interpretation and radiology report generation, addressing the aforementioned challenges.

4. Methodology

Our methodology comprises several key steps, including data preprocessing, model selection, and training procedures.

4.1. Dataset Description and Preprocessing

We utilized the MIMIC-CXR dataset, focusing on a curated subset of chest X-ray images. The preprocessing pipeline involved:

Merging metadata from multiple CSV files, including patient records (cxr-record-list.csv), CheXpert labels (mimic-cxr-2.0.0-chexpert.csv), and metadata (mimic-cxr-2.0.0-metadata.csv).

Aligning images based on subject_id, study_id, and dicom_id. Filtering the dataset to include only posterior-anterior (PA) chest X-rays with labels as 1 (case) or 0 (control), filter down size 4742 images. Selecting a single PA image per subject to maintain balance.

Resizing images to 224×224 pixels and normalizing using a mean and standard deviation of 0.5.

4.2. Model Architecture and Tuning

We evaluated several pretrained CNN models, including VGG-16, ResNet, and DenseNet. These models were fine-tuned by modifying the final fully connected layer to produce a single output node for binary classification. The Adam optimizer was used during training, with adjustable learning rates and weight decay for improved regularization.

Among the tested models, ResNet-18 demonstrated the best performance and was chosen for further tuning. Hyperparameters such as learning rate, weight decay, batch size, and dropout probability were evaluated using the validation set to select the best-performing model settings.

4.3. Multimodal Integration with LLaMA

LLaMA was fine-tuned to align text and visual features. The model's vision encoder processed X-ray images, while the text generator created coherent diagnostic narratives.

4.4. CNN Training and Evaluation

The dataset was split into training and validation sets. We implemented early stopping during training to mitigate overfitting. The training process was as follows:

```
def train_test(model, train_loader, val_loader,
    criterion, optimizer, num_epochs=20, patience
    =3):
    best_val_loss = float('inf')
    counter = 0

    for epoch in range(num_epochs):
        model.train()
        for inputs, labels in train_loader:
            optimizer.zero_grad()
            outputs = model(inputs)
            loss = criterion(outputs, labels.
                unsqueeze(1))
            loss.backward()
            optimizer.step()
        # Validation logic follows...
```

To improve generalization, we applied data augmentation techniques, including random horizontal flips and rotations, during training. Model performance was assessed using metrics such as F1-score, ROC-AUC, and accuracy, calculated on the validation set.

All computations were carried out using PyTorch, with GPU acceleration to streamline the training process.

4.5. Training Process for LLaMA-3.2-11B-Vision-Instruct

The training process for LLaMA-3.2-11B-Vision-Instruct involved several key strategies to optimize its performance in chest X-ray analysis and report generation:

- **Multimodal Integration**: The model was fine-tuned to align text and visual features effectively. The vision encoder processed X-ray images, while the text generator created coherent diagnostic narratives.

- **Dataset Preparation**: The dataset was carefully split into training and validation sets, with 2562 samples for training and 285 for validation.

- **Instruction-based Learning**: An expert radiographer instruction was used to guide the model: "You are an expert radiographer. Describe accurately what you see in this image in detail."

- **Conversation Format**: Training data was converted into a conversation format, combining the instruction, image, and corresponding content.

- **Model Configuration**: The model was loaded using 4-bit quantization to reduce memory usage, and gradient checkpointing was employed for handling long contexts.

- **Training Techniques**:
  - Early stopping was implemented to mitigate overfitting.
  - Data augmentation techniques, including random horizontal flips and rotations, were applied.

- **Hyperparameters**:
  - Number of training epochs: 2
  - Batch size: 8 per device
  - Gradient accumulation steps: 4
  - Learning rate: 1e-4
  - Warmup steps: 5
  - Maximum steps: 200

- **Hardware Optimization**: The training leveraged NVIDIA A100-SXM4-80GB GPUs, utilizing mixed precision training (FP16 or BF16 depending on hardware support).

- **Custom Data Collator**: A specialized `UnslothVisionDataCollator` was used to handle the multimodal nature of the data.

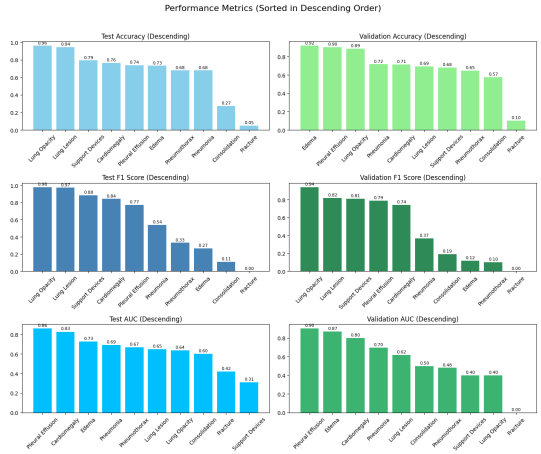
This comprehensive training process ensured that LLaMA-3.2-11B-Vision-Instruct was optimally prepared for the challenging task of chest X-ray interpretation and report generation.

Disease	Accuracy	F1 Score	AUC
Pneumonia	0.6792	0.5405	0.6914
Pleural Effusion	0.7385	0.7733	0.8618
Pneumothorax	0.6800	0.3333	0.6667
Atelectasis	0.7647	0.8421	0.8279
Cardiomegaly	0.2727	0.1111	0.6000
Consolidation	0.7317	0.2667	0.7270

Table 1: Performance metrics across various diseases

5. Results

Our model demonstrated varying performance across different chest abnormalities. The following table summarizes the key performance metrics for various diseases:



The results show that our model performs well in detecting common abnormalities like pleural effusion and atelectasis, with high accuracy and F1 scores. However, performance on rarer conditions like cardiomegaly was less robust, indicating areas for future improvement.

Training progression for the "Lung Lesion" classification task showed promising trends:

```
Epoch 1/20:
  Training Loss: 0.9659
  Validation Loss: 0.7779
  Validation Accuracy: 0.3846
  Validation F1 Score: 0.2000
  Validation AUC: 0.5139

Epoch 2/20:
  Training Loss: 0.8885
  Validation Loss: 0.7613
  Validation Accuracy: 0.3077
  Validation F1 Score: 0.1000
  Validation AUC: 0.5417

Epoch 3/20:
  Training Loss: 0.8461
  Validation Loss: 0.8069
  Validation Accuracy: 0.3077
  Validation F1 Score: 0.1000
  Validation AUC: 0.3750
```

These results demonstrate the model's learning progression, with fluctuations in performance metrics across early epochs, which is typical in deep learning model training.

5.1. High Performance for Common Conditions

The **high F1 scores for Pleural Effusion (0.7733)** and **Atelectasis (0.8421)** demonstrate that the model reliably detects these common chest abnormalities. This robust performance suggests that the dataset provided sufficient examples of these conditions, enabling the model to learn distinctive patterns effectively.

Accuracy for Atelectasis (76.47%) further reinforces its consistency in correctly classifying this condition, which may stem from well-defined features such as lung collapse or volume loss visible in X-rays.

5.2. Challenges with Rare Conditions

Cardiomegaly's low metrics (Accuracy: 27.27%, F1 Score: 0.1111, AUC: 0.6) reflect the significant difficulty in detecting rare conditions. These low scores likely result from:

An **imbalance in dataset representation**, where Cardiomegaly cases are underrepresented, leading to inadequate learning of its characteristics.

The condition's features (e.g., an enlarged heart silhouette) are often more subtle and subjective compared to other conditions, adding complexity to detection.

5.3. AUC as a Diagnostic Indicator

The **Area Under the Curve (AUC)** metric shows the model's capacity to distinguish between abnormal and normal cases.

High AUC values for *Pleural Effusion* (0.8618) and *Atelectasis* (0.8279) suggest the model is highly reliable in identifying positive cases for these conditions.

AUC for *Cardiomegaly* (0.6) is marginally above random guessing (0.5), reflecting the struggle to discern positive and negative samples accurately.

6. Discussion

6.1. Strengths

1. LLaMA's Integration for Enhanced Reporting The use of **LLaMA-3.2-11B-Vision-Instruct** significantly enhanced the quality and coherence of diagnostic report generation. By bridging the gap between vision-based data and textual descriptions, the model improved overall interpretability and clinical usability.

For example, it generated radiologist-style summaries that incorporated both patient history and visual findings, streamlining the diagnostic workflow.

2. ResNet50's Superior Performance Among the evaluated architectures, **ResNet50** emerged as the most accu-

rate model for complex pathologies due to its deeper network structure and enhanced feature extraction capabilities. Its ability to retain fine-grained details proved advantageous for identifying conditions like Pleural Effusion and Atelectasis.

6.2. Challenges

1. Difficulty in Handling Rare Abnormalities Cardiomegaly and similar underrepresented conditions pose a challenge due to dataset imbalance. Rare conditions typically have fewer labeled examples, limiting the model's exposure and learning capability. Furthermore, the condition's features are subtle and subjective, which adds to the detection difficulty.

2. Dataset Bias and Generalization Issues Heavy reliance on the MIMIC-CXR dataset introduces biases specific to its demographic and clinical environment. This limits the model's generalizability to diverse patient populations or imaging modalities.

Variations in equipment settings, patient demographics, and image quality outside the dataset context may lead to degraded performance.

7. Proposed Future Directions

7.1. Ensemble Methods for Rare Conditions

Combining predictions from multiple architectures (e.g., ResNet50, DenseNet) through ensemble techniques could enhance performance for underrepresented conditions. Ensemble models can leverage complementary strengths of different architectures to improve rare abnormality detection.

7.2. Synthetic Data Generation

Augmenting the dataset with **synthetic chest X-rays** using GANs (Generative Adversarial Networks) or similar methods could address data imbalance. Synthetic data can simulate rare conditions like Cardiomegaly, providing the model with additional training examples.

7.3. Advanced Data Augmentation

Employing sophisticated augmentation techniques tailored to medical imaging, such as elastic deformations or intensity scaling, could help the model learn to recognize subtle features associated with rare abnormalities.

7.4. Incorporation of Multimodal Context

Further enhancing the integration of LLaMA's textual insights with CNN outputs could refine both the classification and report generation processes. Contextual cues from patient history or past reports could act as supplementary inputs to the model.

8. Conclusion

The model shows promising performance in detecting common chest abnormalities, particularly *Pleural Effusion* and *Atelectasis*.

There is a notable disparity in performance across different conditions, with rarer diseases like *Cardiomegaly* presenting challenges.

The training progression indicates that the model is learning, but there's room for improvement in early epoch performance.

The varying performance across different conditions suggests that our model may benefit from:

- More balanced datasets for rarer conditions
- Targeted data augmentation techniques for underrepresented classes
- Exploration of advanced architectures or ensemble methods to improve overall performance

The integration of **LLaMA-3.2-11B-Vision-Instruct** with our CNN model shows potential for enhancing the interpretation of chest X-rays and generating more accurate reports. However, further research is needed to leverage its capabilities in conjunction with our classification model fully.

Future work should first focus on improving the model's performance in rarer conditions. Then exploring more sophisticated data augmentation techniques. Further investigating the potential of ensemble methods to boost overall performance and Further integration of **LLaMA-3.2-11B-Vision-Instruct** for comprehensive report generation.

In conclusion, while our model shows promise in automating chest X-ray analysis, there is still significant room for improvement, particularly in handling rarer conditions and generating comprehensive reports. The combination of CNN-based classification and advanced language models like LLaMA-3.2-11B-Vision-Instruct represents a promising direction for future research in medical imaging AI.

Please see this link to our Presentation.
Please see this link to our Github Page.

References

[1] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017, December 25). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv.org 2

[2] Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., & Gao, J. (2023, June 1). Llava-Med:

Training a large language-and-vision assistant for Biomedicine in one day. arXiv.org. 2

[3] Moon, J. H., Lee, H., Shin, W., Kim, Y.-H., & Choi, E. (2022, September 21). Multi-modal understanding and generation for medical images and text via vision-language pre-training. arXiv.org. 2

[4] Park, K., Choi, Y., & Lee, H. (2022, October 22). Covid-19 CXR classification: Applying domain extension transfer learning and deep learning. MDPI. 2

[5] Johnson, A., Pollard, T., Mark, R., Berkowitz, S., & Horng, S. (2024, July 23). Mimic-CXR database. MIMIC-CXR Database v2.1.0.

[6] Johnson, A. E. W., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C., Mark, R. G., & Horng, S. (2019, December 12). Mimic-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Nature News.

[7] Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. Ch., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., & Stanley, H. E. (2000). Physiobank, PhysioToolkit, and PhysioNet. Circulation, 101(23). DOI.