COSE474-2024F: Final Project
"Ensemble Model for Pneumonia Detection"

Yan Ha

1. Introduction

Pneumonia detection is a vital task in medical imaging that directly impacts patient outcomes. Early and accurate detection ensures timely treatment, which is critical for preventing severe complications or fatalities. However, automated detection using chest X-rays presents several challenges, including data quality variations and subtle visual indicators of pneumonia. Ensemble learning, which combines multiple models, offers a robust solution for improving prediction accuracy. By leveraging diverse architectures, an ensemble can capture complementary features, making it particularly effective for binary classification tasks such as identifying pneumonia (Mooney, n.d.; Wang et al., 2017).

The primary problem in this study is distinguishing between normal and penumonia chest X-rays. The task becomes more challenging in the presence of imbalanced datasets, where pneumonia cases are less frequent than normal cases. Additionally, high sensitivity is required to minimize false negatives, as missing a pneumonia diagnosis can lead to severe outcomes. At the same time, high specificity is needed to reduce false positives, which can lead to unnecessary treatments or tests. Addressing these competing objectives requires a balanced and reliable approach (Wang et al., 2017; Johnson et al., 2019).

This project proposes an ensemble model that integrates ResNet50, DenseNet121, and Vision Transformer (ViT-B/16) to tackle the problem of pneumonia detection. These models are pretrained on ImageNet, enabling the ensemble to leverage learned features effectively. Cross attention is employed to fuse features from CNNs and transformers, while a weighted prediction mechanism ensures the ensemble produces reliable results. The model is evaluated on the Kaggle Chest X-ray Dataset, demonstrating its effectiveness in real-world medical imaging tasks.

2. Methods

2.1. Significance and Novelty

The ensemble model leverages the complementary strengths of CNN-based architectures (ResNet50 and DenseNet121) and the global attention capabilities of Vision Transformer (ViT-B/16). CNNs specialize in capturing local spatial features, while ViT captures long-range dependencies across the entire image. Integrating these features using cross-attention ensures robust feature fusion, which is critical for improving classification performance. This study also incorporates weighted ensemble predictions, which allow the model to prioritize predictions from stronger individual components (Dosovitskiy et al., 2021; He et al., 2016; Huang et al., 2017).

2.2. Main Challenges and Solutions

Balancing the feature representations from diverse architectures was a significant challenge. Cross-attention was implemented to align and fuse features effectively. the limited size of the dataset was another challenge, addressed using transfer learning. ImageNet-pretrained weights provided a strong initialization, allowing the model to generalize better with limited data. Computational constraints, particularly with ViT, were managed by resizing images to 224x224 and optimizing training on GPU resources.

2.3. Main Figure

The architecture of the proposed ensemble model is designed to process chest X-ray images for pneumonia detection. It begins with an input layer where X-ray images are fed into three feature extraction models: ResNet50, DenseNet121, and Vision Transformer (ViT-B/16). Each model independently extracts features using its unique architecture, capturing both local and global patterns. These features are then passed to a cross-attention mechanism, which aligns and fuses them into a unified representation. The fused features are subsequently processed by a weighted ensemble layer, where predictions from each model are combined proportionally based on their predefined weights. The final output is a binary classification indicating either "Normal" or "Pneumonia."

The entire workflow is reproducible with the provided code. The implementation details, including data preprocessing, training configurations, and model architecture, are documented in the accompanying notebook.

3. Experiments

3.1. Dataset

The Kaggle Chest X-ray Dataset was utilized, which consists of 5,863 images divided into two classes: normal and pneumonia cases. The dataset includes both bacterial and viral pneumonia cases, but for this study, the labels were treated as a single pneumonia class. This simplification aligns with the binary classification task. Images were resized to 224x224 pixels to standardize input dimensions for the ensemble models. Normalization was performed using ImageNet-specific mean and standard deviation values to match the pretrained weights used by ResNet50, DenseNet121, and ViT-B/16. Data augmentation techniques, such as random rotations, horizontal flips, and brightness adjustments, were applied to enhance generalization by introducing variability in the training set.

3.2. Computing Resources

All experiments were conducted on Google Colab with GPU acceleration provided by an NVIDIA Tesla T4 GPU. The software stack included PyTorch 1.10+, CUDA 11.x, and an Ubuntu-based environment. This setup ensured sufficient computational power to handle the transformer components of the model and large-scale data augmentation processes during training.

3.3. Experimental Design and Setup

The ensemble model was trained for 10 epochs, with a batch size of 32. The Adam optimizer was used with an initial learning rate of 1e-4 and a weight decay of 1e-5. These hyperparameters were chosen based on a grid search for optimal convergence rates. The dataset was split into training (70%), validation (15%), and test (15%) subsets. Metrics such as accuracy and loss were tracked across epochs to evaluate both convergence and generalization.

The ensemble model employed cross-attention to integrate features from ResNet50, DenseNet121, and ViT-B/16. Each model contributed features of varying complexity and scale, which were aligned and fused using the cross-attention mechanism before producing a weighted final prediction.

3.4. Quantitative Results

The ensemble model demonstrated consistent improvement across training and validation accuracy. The training and validation loss curves, shown in Figure 1, indicate steady convergence over 10 epochs. Validation accuracy increased from approximately 75% at initialization to 88% at the final epoch, as illustrated in Figure 2. These results demonstrate the ensemble's ability to generalize effectively, despite the dataset's relatively small size.

Figure 1. Training and Validation Loss Curves

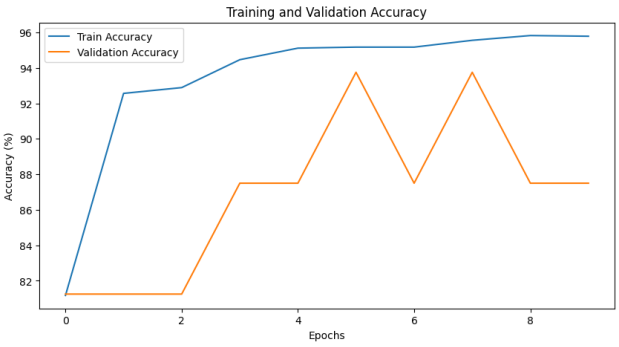

Figure 2. Training and Validation Accuracy Curves

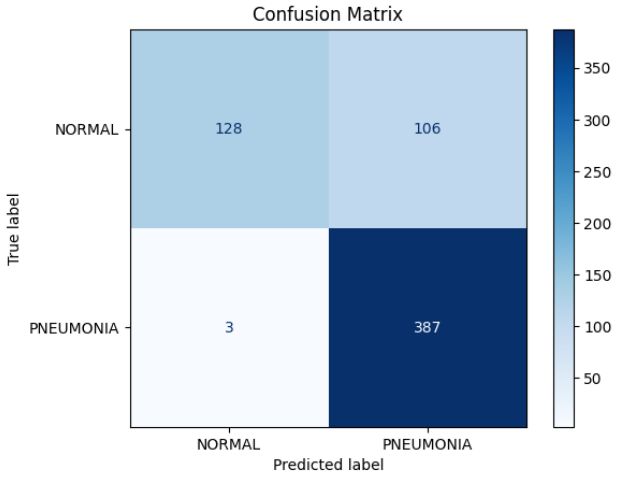

Figure 3. Confusion Matrix

When compared to individual models, the ensemble outperformed ResNet50, DenseNet121, and ViT-B/16 used in

isolation, with an average accuracy improvement of approximately 5%. This underscores the value of feature fusion and weighted ensemble predictions in achieving superior classification performance.

3.5. Discussion

The ensemble model effectively balanced the contributions of CNNs and transformers, leading to robust classification performance. Its ability to integrate localized and global feature representations allowed it to generalize well to unseen data, as evidenced by the steady improvements in validation accuracy. However, the confusion matrix revealed that false negatives were more prevalent than false positives, indicating that subtle pneumonia cases may have been challenging for the model to detect.

Data augmentation mitigated overfitting and enhanced generalization, particularly given the limited size of the dataset. Nonetheless, the lack of diversity in training samples likely constrained the model's capacity to handle rare or ambiguous cases effectively. Computationally, the transformer-based ViT required more resources compared to the CNNs, but this was manageable within the Colab environment.

4. Future Directions

To further improve the model's robustness, future work will focus on training with larger and more diverse datasets, such as NIH Chest X-rays and MIMIC-CXR. These datasets offer richer annotations and multi-class classification opportunities, enabling the ensemble model to tackle more complex medical imaging challenges (Wang et al., 2017; Johnson et al., 2019).

Incorporating clinical metadata, such as patient history or laboratory results, could significantly enhance the model's decision-making process. A multimodal approach, which combines imaging data with associated text-based clinical reports, would provide additional context for predictions. Leveraging advanced models like BioViL, designed for vision-language tasks in the medical domain, could be a promising direction (Boecking et al., 2022).

Given the abundance of unlabeled medical imaging data, semi-supervised learning methods, such as pseudo-labeling or self-supervised learning, could be explored. These techniques allow models to utilize unlabeled data for training, thereby improving generalization without requiring extensive manual annotation (Lee, 2013).

While this study did not include visualization techniques such as Grad-CAM, future work could integrate such tools to provide insights into the regions of the image that influenced the model's predictions. Enhancing interpretability fosters trust among clinicians and enables the identification of biases or shortcomings in the model.

5. References

1. Boecking, B., Usuyama, N., Bannur, S., Castro, D. C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., Poon, H., & Oktay, O. (2022). Making the most of text semantics to improve biomedical Vision–Language processing. In Lecture notes in computer science (pp. 1–21). https://doi.org/10.1007/978-3-031-20059-5-1

2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations. https://arxiv.org/abs/2010.11929

3. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770-778).

4. Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4700-4708).

5. Johnson, A. E. W., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C. Y., ... & Horng, S. (2019). MIMIC-CXR: A large publicly available database of labeled chest radiographs. Nature Scientific Data, 6(1), 1-8.

6. Lee, D.-H. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. ICML Workshop on Challenges in Representation Learning.

7. Mooney, P. (n.d.). Chest X-ray images (pneumonia). Kaggle. https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia

8. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2097-2106).