Classification of Chest X-ray Image into Multiple Disease Findings Simultaneously

Sanskar Kejriwal
EEE Dept.
IIT Guwahati
k.sanskar@iitg.ac.in

Pratyush Ranjan
EEE Dept.
IIT Guwahati
r.pratyush@iitg.ac.in

Riddhi Agrawal
EEE Dept.
IIT Guwahati
a.riddhi@iitg.ac.in

Devansh Sharma
EEE Dept.
IIT Guwahati
s.devansh@iitg.ac.in

Daksh Kaushik
BSBE Dept.
IIT Guwahati
k.daksh@iitg.ac.in

***Abstract*—Medical image classification poses unique challenges due to the long-tailed distribution of diseases, the co-occurrence of diagnostic findings, and class imbalance. This paper deep dives into some of the existing solutions and tries to summarize their results and findings, along with listing their limitations. It also proposes a novel solution that has ensembled different convolutional neural networks (CNNs), each augmented with an attention layer, and suggest further improvements that can be done to increase its performance on benchmark dataset for Chest X-rays.**

I. Introduction

The field of medical image classification has gained significant attention due to the increasing recognition of the potential of artificial intelligence in healthcare. Most of the existing solutions has faced challenges majorly in multi label classification, co-occurance of labels and in correct classification of rare diseases due to high class imbalance. Existing work has incorporated one or combination of these techniques like Data re-sampling, Data re-weighting, Data Augmentation and Transfer Learning which has its own merits and demerits and can be further improved.

The novel idea is built by ensembling three convolutional neural networks (CNNs) - ResNet, MobileNet, and VGG, which are most effective in chest X ray disease classification compared to other models present, each augmented with an **location based attention layer**. The attention layer plays a crucial role in identifying the specific region of an image that strongly contributes to the classification of a particular diagnosis. Our ensemble model leverages the strengths of these simple yet powerful CNN architectures while introducing the concept of attention, allowing the network to focus on crucial image regions and patterns.

II. Related Work

A. Long-Tailed Classification of Thorax Diseases on Chest X-Ray[1]

1) Introduction: The authors tackle the **long-tailed** distribution of findings in chest radiography, conducting a study on long-tailed learning methods for thorax disease classification in X-rays. They create a benchmark dataset, combining NIH-CXR-LT and MIMIC-CXR-LT datasets, featuring different class imbalances. Their aim is to improve classification accuracy for both common and rare conditions, addressing the issue of skewed class distribution in chest X-ray diagnoses.

2) Dataset Construction: The authors curated the **NIH-CXR-LT** and **MIMIC-CXR-LT** datasets by introducing five new rare disease findings mined from radiology reports. These datasets were split into training, validation, test, and balanced test sets, with extreme class imbalance.

3) Evaluation and Results: The authors assessed multiple long-tailed learning methods on their benchmark datasets using a pretrained **ResNet50** architecture and Adam optimizer. They evaluated performance on both balanced and imbalanced test sets, finding that class-balanced re-weighting and classifier re-training effectively enhanced accuracy for rare classes.

4) Conclusion: In summary, this study introduces a thorough benchmark for long-tailed learning in thorax disease classification from chest X-rays. It offers valuable insights into the efficacy of various long-tailed learning approaches and underscores the significance of mitigating class imbalance in medical image classification.

B. Multi-Label Classification on Chest X-ray Images with Transformers[6]

1) Introduction: While conventional methods for automated chest X-ray image diagnosis relied on CNN architectures, recent research demonstrates that transformers, commonly used in natural language processing, can surpass CNN-based models in computer vision tasks. This paper introduces a **state-of-the-art multi-label classification** model with the Swin Transformer as the backbone, significantly improving performance on the **ChestX-ray14** dataset.

2) Swin Transformer: The Swin Transformer tackles obstacles in using transformers for computer vision. Unlike language tasks, images have diverse scales, making fixed-size tokens problematic. High-res images also need pixel-level predictions, which are computationally intensive for existing transformers.

3) Proposed Method: The model uses the **Swin Transformer** for multi-label classification, dividing the input image into patches, each treated as a token. Linear embedding layers transform patch features, and Swin transformer blocks are applied while token count is maintained. Patch merging layers reduce tokens in deeper layers, with 14 MLP heads branching from the shared section for classification.

4) Results and Conclusion: The 3-layer headed SwinCheX model attains state-of-the-art performance on the ChestX-

ray14 dataset, with an **average AUC score of 0.810**, surpassing DenseNet in detection accuracy. This highlights the effectiveness of vision transformers for multi-label classification in chest X-ray images. Future research can explore additional vision transformer architectures to evaluate their performance further.

C. Multi-Label Chest X-Ray Classification via Deep Learning[5]

1) Introduction: This paper aims to create a lightweight solution for detecting 14 chest conditions from X-ray images to enhance clinical support and patient care. The approach combines CNN models with image and non-image features like age and gender. Performance is assessed using various CNN models, including **CustomNet, DenseNet121, ResNet-50, InceptionV3, and Vgg16**, with metrics like accuracy and AUROC.

2) Approach and Methods: This study's approach centers on creating a CNN classifier for chest X-ray image labeling, with evaluation across diverse datasets to ensure robustness and generalization. **Data augmentation** techniques are used to enhance dataset size and quality, mitigating overfitting. Transfer learning from **ImageNet-trained** models aids in capturing relevant X-ray image features during CNN model training.

3) Results and Conclusion: The performance evaluation on unseen test data indicates that DenseNet121 outperforms other models, with an **AUROC of 0.78** and an accuracy of 87%. While CustomNet, ResNet-50, InceptionV3, and Vgg16 yield promising results, they exhibit slightly lower AUROC values and accuracies. However, challenges in accurately predicting positive cases of some diseases are attributed to imbalanced training data.

D. Multi-Label Chest X-Ray Classification via Dual Weighted Metric Loss[3]

1) Introduction: In this study, authors introduce a novel deep learning-based classification framework for multi-label chest X-ray images. To tackle label correlation and imbalance issues, they present a **dual-weighted metric loss function** that considers image-label relationships at both image and disease category levels. Their model is evaluated on the Chest X-ray14 dataset, surpassing existing models with superior performance based on average AUC scores.

2) Methods: The authors introduce a novel classification framework for multi-label chest X-ray images, employing **ConvNeXt** for visual feature extraction and **BioBert** for semantic vectors. These features are mapped into a common metric space, and a dual-weighted metric loss function is introduced to consider image-label relationships. The model's loss function combines two components: weighted **multi-label classification loss (LMC)** and **dual-weighted metric loss (LMetric)**, with LMC computed using binary cross-entropy.

3) Results and Conclusion: The authors experiment with the Chest X-ray14 dataset comprising 112,120 frontal X-ray images of 14 chest diseases. Their model outperforms others, achieving an average **AUC score of 0.826**. The introduced loss function enhances classification accuracy by considering the relationships between visual features and semantic vectors.

III. Proposed Work

A. Introduction

This study[2] explores an ensemble approach using three base models (**VGG19, MobileNet, and InceptionResNetV2**) and incorporates location based attention mechanisms to enhance image classification accuracy. The goal is to leverage VGG19's deep features, MobileNet's efficiency, and InceptionResNetV2's complexity. Attention layers are added to each base model to improve contextual awareness, focusing on specific input regions. The ensemble of these attention-augmented models offers a promising strategy for image classification, with potential for further improvements.

B. Network Architecture

The network architecture employed in this research comprises an ensemble of three base models—VGG19, MobileNet, and InceptionResNetV2—enhanced with **attention layers** to boost their image classification capabilities. By integrating all three base models into the ensemble, we aim to create a powerful predictive tool that benefits from the diversity of the constituent models. We have used **location-based attention**[4]. In location-based attention, the input undergoes processing through a CNN to generate feature maps, capturing information at various positions and scales. Subsequently, these feature maps determine attention weights for each position by employing a 1×1 convolution, resulting in scalar values. These scalar values serve as **attention weights**, influencing the feature maps through a weighted sum operation. This process produces a representative output, emphasizing important information based on the calculated attention weights for each position.The proposed model architecture, illustrated in Figure 1, outlines the structure and design of the model discussed in this study.

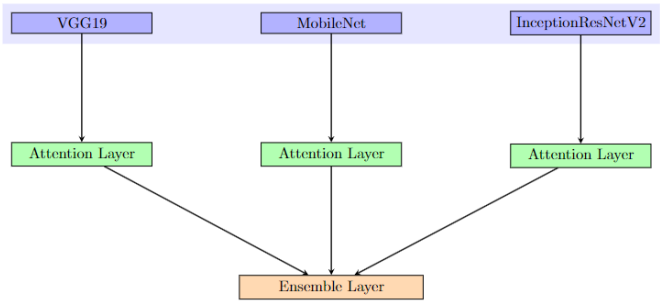

Fig. 1. Network Architecture

C. Cost Functions

1) Binary Cross-Entropy Loss : This is commonly used for binary classification, however it can also be adapted for multilabel classification tasks. To apply Binary Cross-Entropy Loss in this context, we typically employ a neural network with a sigmoid activation function in the output layer, with one neuron for each label or class. Each label's loss is then calculated independently using the **Binary Cross-Entropy** formula. During training, the model's weights are updated for each label's loss, allowing it to learn to predict the likelihood of each label's presence. When making predictions, a threshold (often set at 0.5) is used to determine whether an example belongs to a specific label based on the predicted probability.

2) Mean Absolute Error (MAE) Loss: Unlike traditional MAE for regression tasks, the multi-label version involves computing the absolute error for each label separately and then averaging across instances and labels.

D. Training Metrics

1) Binary Accuracy: Binary Accuracy is a critical metric for assessing the effectiveness of our model in binary classification tasks, where the objective is to correctly classify instances into one of two categories. This metric quantifies the ratio of **correctly predicted binary outcomes to the total number of instances**. In the context of multi-label classification, for each label, you determine whether the model correctly predicts the presence or absence of that label. You can then compute the average binary accuracy across all labels to get an overall measure for your multi-label classification task.

2) Adam Optimizer: For the training of our models, we utilize the **Adam optimizer**. This optimization algorithm combines elements from both stochastic gradient descent (SGD) and root mean square propagation (RMSprop). It is an effective choice for updating model parameters during training and ensuring convergence.

IV. Experimental Details

A. Datasets

This dataset comprises over 100,000 de-identified chest X-ray images from more than 30,000 patients, including those with advanced lung diseases. The data represents NLP analysis of radiology reports and may include areas of lower confidence in diagnoses. As a simplifying assumption, we assume that based on the size of the dataset, the dataset is accurate in diagnoses.

One of the difficulties of this problem involves the lack of a **"diagnosis confidence"** attribute in the data. In addition to a chest X-ray, diagnosis involves patient presentation and history.

The image set involves diagnoses that were scraped from radiology reports and is a multi-label classification problem. The diagram below shows the proportion of images with multi-labels in each of the eight pathology classes and the labels' **co-occurrence** statistics.

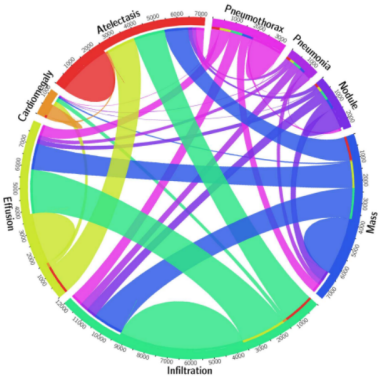

Fig. 2.

B. Training Details

In the training details for our model on Kaggle, a **learning rate of 0.0005** is employed, influencing the step size during optimization iterations. The synthetic batch size, set to 256, determines the total number of samples which will be used in **Gradient Accumulation**. With a **training batch size of 16**, each iteration processes a specific number of training samples, impacting the speed and memory efficiency of the training process. The default optimizer is 'adam,' a widely used optimization algorithm that adapts learning rates dynamically. **Accumulation steps**, calculated as the synthetic batch size divided by the batch size, enable the accumulation of gradients before performing a backward pass, useful in scenarios with limited GPU memory. The validation batch size is set to twice the training batch size, ensuring a robust evaluation on unseen data. The training is configured for four epoch, with **early stopping patience** set to 2 epochs. Moreover, incorporating a **learning rate plateau** patience of 2 epochs allows for adaptive adjustments to the learning rate in the event of performance plateaus during training. The utilization of 8 workers for parallel data loading and preprocessing highlights an emphasis on optimizing efficiency. Notably, it's essential to acknowledge that each epoch demands approximately 3 hours on the Kaggle platform, underscoring the substantial computational requirements inherent in the training process.

C. Baseline Methods

In the realm of machine learning and deep learning research, the pursuit of novel and advanced models is a common goal. However, to gauge the efficacy of these models, it is essential to establish a baseline for comparison. Baseline methods, in this context, serve as rudimentary yet vital reference points against which the performance of more sophisticated techniques can be evaluated. This section elaborates on the baseline methods employed in our research.

1) Base Models and Configuration: We initiate our comparative analysis with a set of three base models, namely VGG19, MobileNet, and InceptionResNetV2.

i) **VGG19**: It is a variant of the VGG (Visual Geometry Group) model, and as the name suggests, it has 19 layers. The

architecture is characterized by its simplicity and uniformity, as it stacks multiple convolutional layers with small receptive fields (3x3), followed by max-pooling layers. The model ends with a few fully connected layers leading to the output layer.

ii) **MobileNet**: It is designed for mobile and embedded vision, utilizes a streamlined architecture with depthwise separable convolutions for lightweight neural networks. Tailored for low latency on resource-constrained devices, the original MobileNet comprises 28 layers, striking a balance between efficiency and real-time performance.

iii) **InceptionResNetV2**: It showcases architectural excellence in convolutional neural networks. Boasting a remarkable depth of 164 layers, this sophisticated model seamlessly integrates the inventive design of the Inception family with the robust residual connections of ResNet. The result is a neural architecture adept at capturing intricate features, rendering it a formidable choice for complex computer vision tasks that demand both depth and sophistication.

In summary, baseline methods are fundamental in our research, providing a foundation for model comparison and serving as benchmarks to evaluate the advancements achieved through the ensemble and attention mechanisms.

D. Evaluation Metrics

In any deep learning research, the choice of evaluation metrics is of paramount importance. These metrics serve as the yardstick for measuring the performance of models and assessing how well they fulfill their intended tasks. In this section, we discuss the evaluation metrics employed in our research, the rationale behind their selection, and how they provide a comprehensive view of model performance.

1) ROC Curve and AUC-ROC: **ROC curves** and the **Area Under the ROC Curve** (AUC-ROC) are vital for evaluating binary classification models, particularly in imbalanced or multi-label scenarios. These metrics illuminate the balance between **true positive** and **false positive** rates across different classification thresholds. ROC curves offer visual clarity on a model's proficiency in distinguishing between positive and negative classes, and AUC-ROC condenses this discriminatory performance into a single numerical value.

In the context of our multi-label classification problem, we conduct label-wise ROC analysis. This involves treating each label as a binary classification task, generating ROC curves and calculating AUC-ROC values for each label independently. This approach helps us assess how well the model predicts each label, which is crucial for understanding its performance in multi-label scenarios.

2) Ensemble Metrics: In the process of ensembling attention models, we use a variety of ensemble techniques, including max voting, simple averaging, and weighted averaging. These techniques enable us to combine the predictions of individual models in different ways and evaluate their effectiveness. By exploring these ensemble techniques, we gain insights into how aggregating predictions from multiple sources influences model performance.

Class	Mobilenet	Resnet	VGG	Avg Ensemble	Wght Avg Ensemble	Max Vote Ensemble
Atelectasis	0.675	0.635	0.651	0.642	0.691	0.635
Cardiomegaly	0.743	0.633	0.722	0.709	0.758	0.695
Consolidation	0.66	0.639	0.648	0.637	0.678	0.627
A Edema	0.778	0.774	0.776	0.741	0.791	0.735
Effusion	0.728	0.69	0.721	0.678	0.743	0.666
Emphysema	0.731	0.729	0.663	0.717	0.749	0.709
Fibrosis	0.707	0.692	0.682	0.649	0.717	0.633
Infiltration	0.61	0.598	0.604	0.59	0.628	0.574
Mass	0.632	0.598	0.614	0.589	0.644	0.57
Nodule	0.649	0.628	0.635	0.605	0.665	0.584
Pleural Thickening	0.639	0.615	0.6	0.59	0.65	0.567
Pneumonia	0.567	0.56	0.529	0.584	0.592	0.578
Pneumothorax	0.738	0.706	0.706	0.689	0.752	0.672

Fig. 3. AUC-ROC for all labels

3) Hyperparameter Tuning Metrics: Hyperparameters, such as learning rate, batch sizes, and early stopping patience, play a significant role in model training. We monitor training and validation metrics during hyperparameter tuning, looking for signs of overfitting or underfitting and striving to strike the right balance between model complexity and generalization.

V. Results

A. Ensemble

After obtaining the individual outputs from the VGG, ResNet, and MobileNet models with attention layers, we proceed to perform ensemble techniques to harness the collective predictive power of these models. In this study, we explore three ensemble techniques: **Max Vote**, **Simple average**, and **Weighted average**.

We have summarized the AUC-ROC scores of individual models with attention layer and ensemble model with different ensemble techniques in Figure 3. The graph presented in Figure 4 present the result of these Weighted average ensemble technique, which is the most effective one.

For the code, please visit: GitHub Repository

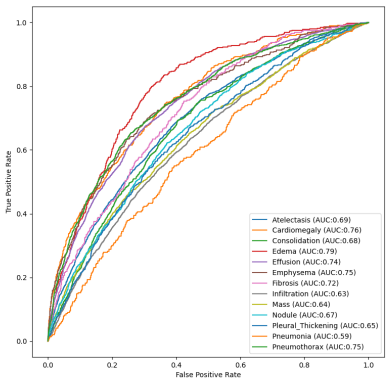

Fig. 4. Weighted Ensemble ROC

VI. Future Work

A. Model Enhancements

- Future research can explore other advanced attention mechanisms, such as multi head attention, to improve further the model's ability to capture intricate details in medical images.
- Investigate the potential of incorporating external knowledge, such as medical ontologies or clinical notes, to enhance the model's understanding of diseases and their associations.
- Explore the use of generative adversarial networks (GANs) for data augmentation, generating synthetic medical images to address the issue of limited training data.

B. Addressing Label Imbalance

- Evaluate our model on different datasets after training them on that. Develop and evaluate techniques specifically designed to handle the long-tailed distribution of diseases, which is common in medical image datasets.
- Investigate the use of semi-supervised and self-supervised learning methods to make better use of limited annotated data for rare diseases.

References

[1] Gregory Holste1 et al. "Long-Tailed Classification of Thorax Diseases on Chest X-Ray: A New Benchmark Study". In: (2022), pp. 1–12. DOI: 10.1007/978-3-031-17027-0_3..

[2] *Introduction to CNNs with Attention Layers*. https://www.ai-contentlab.com/2022/12/introduction-to-cnns-with-attention.html.

[3] Yufei Jin et al. "Deep learning based classification of multi-label chest X-ray images via dual-weighted metric loss". In: (2022), pp. 1–10. DOI: 10.1016/j.compbiomed.2023.106683.

[4] *Location-based Attention*. https://serp.ai/location-based-attention.

[5] Aravind Sasidharan Pillai. "Multi-Label Chest X-Ray Classification via Deep Learning". In: (2022), pp. 1–14. DOI: 10.4236/jilsa.2022.144004.

[6] Sina Taslimi et al. "SwinCheX: Multi-label classification on chest X-ray images with transformers". In: (2022), pp. 1–10. DOI: arXiv:2206.04246v1.