IMPROVING MULTI-LABEL CHEST X-RAY CLASSIFICATION USING MODERN CNNS AND TRANSFORMERS

Prakrut Bhaisare (22169)

B.Tech,Mathematics and Computing, 4th Year

ABSTRACT

This report presents a comprehensive comparative study of deep learning architectures and training strategies for multi-label classification of thoracic diseases from chest X-rays. We implement and evaluate three distinct models: a baseline ResNet-50 replicating the ChestX-ray14 paper, an EfficientNet-B3, and a Vision Transformer (ViT), across multiple loss functions and image resolutions. Our results demonstrate that modern architectures significantly outperform the baseline, with EfficientNet-B3 achieving the highest mean AUC of 0.795 using Weighted BCE loss and 512px images. The study reveals that proper loss function selection and adequate image resolution are critical factors for optimal performance, with Weighted BCE consistently outperforming Focal Loss across all configurations.

***Index Terms*—** Chest X-ray, Multi-label Classification, Weakly-Supervised Localization, Deep Learning, ResNet, EfficientNet, Vision Transformer, Focal Loss, Computer-Aided Diagnosis

1. INTRODUCTION

Chest X-rays remain one of the most accessible and cost-effective radiological examinations worldwide. The automation of their interpretation using deep learning has significant potential to enhance radiologists' workflow, improve diagnostic consistency, and increase accessibility in resource-limited settings. A fundamental challenge in this domain is the scarcity of large-scale datasets with pixel-level annotations, as producing detailed annotations requires substantial expert time and resources.

The ChestX-ray14 dataset {**Wang2017**} addressed this challenge by providing over 112,000 frontal-view X-ray images with image-level labels for 14 thoracic diseases, automatically extracted from radiological reports using Natural Language Processing. The accompanying paper proposed a weakly-supervised framework capable of both disease classification and localization using only image-level labels.

This project conducts a comprehensive comparison of three architectural paradigms and multiple training strategies:

1. **Baseline CNN:** ResNet-50 with custom transition layers and LSE pooling (paper replication)
2. **Modern CNN:** EfficientNet-B3 with compound scaling
3. **Transformer:** Vision Transformer (ViT-Base) with self-attention mechanisms

Additionally, we evaluate the impact of different loss functions (Weighted BCE vs Focal Loss) and image resolutions (512px vs 384px) on model performance.

2. DATASET AND METHODOLOGY

2.1. Dataset

We utilize the ChestX-ray14 dataset [1], comprising 112,120 frontal-view chest X-ray images from 30,805 unique patients. The dataset includes 14 thoracic disease categories: *Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Pleural_Thickening,* and *Hernia*. The data is partitioned at patient level into training (70%), validation (10%), and test (20%) sets, with significant class imbalance characteristic of medical datasets.

2.2. Model Architectures

2.2.1. Baseline ResNet-50

We faithfully replicate the DCNN framework from the original paper [1]:

- **Backbone:** ResNet-50 pre-trained on ImageNet
- **Custom Components:** Transition layer (1×1 conv), LSE pooling ($r = 10$)
- **Loss:** Weighted Cross-Entropy for class imbalance handling
- **Input:** 512×512 pixels

2.2.2. EfficientNet-B3

Leveraging modern CNN design principles [2]:

- **Architecture:** Compound scaling optimized backbone
- **Features:** 1536-dimensional feature space
- **Training:** AdamW optimizer, advanced augmentation, class weights
- **Input:** 512×512 pixels (primary), 384×384 pixels (ablation)

2.2.3. Vision Transformer

Exploring attention-based architectures [3]:

- **Architecture:** ViT-Base with patch size 16
- **Features:** 768-dimensional embedding space
- **Training:** AdamW with cosine annealing, gradient clipping
- **Input:** 512×512 pixels, 384×384 pixels.

2.3. Loss Functions

We compare two approaches for handling class imbalance:

2.3.0.1. Weighted Binary Cross-Entropy (Weighted BCE)

$$L_{W-BCE} = -\frac{1}{N}\sum_{i=1}^{N}\left[w_p y_i \log(\hat{y}_i) + w_n(1-y_i)\log(1-\hat{y}_i)\right] \tag{1}$$

where w_p and w_n are class weights inversely proportional to class frequencies.

2.3.0.2. Focal Loss

$$L_{Focal} = -\frac{1}{N}\sum_{i=1}^{N}\Big[\alpha(1-\hat{y}_i)^{\gamma} y_i \log(\hat{y}_i) + (1-\alpha)\hat{y}_i^{\gamma}(1-y_i)\log(1-\hat{y}_i)\Big] \tag{2}$$

with focusing parameter γ and balancing parameter α.

2.4. Training Protocol

All models were trained with comprehensive hyperparameter optimization. The baseline used 100 epochs, while EfficientNet and ViT used early stopping based on validation performance. We employed patient-level data splitting and extensive data augmentation.

3. EXPERIMENTS AND RESULTS

3.1. Evaluation Metrics

Primary evaluation used Area Under the ROC Curve (AUC) for each disease class, with mean AUC as the overall performance metric. Additional metrics included precision, recall, F1-score, accuracy, exact match ratio, and Hamming loss.

3.2. Ablation Studies: Loss Functions and Image Resolution

To understand the impact of key training decisions, we conducted systematic ablation studies evaluating two critical factors: loss function selection and image resolution. We compared Weighted Binary Cross-Entropy against Focal Loss, and assessed performance at 512px versus 384px resolutions. These experiments provide practical guidance for researchers and practitioners in selecting optimal configurations for medical image classification tasks.

Disease	Baseline	EfficientNet	ViT	Best
Atelectasis	0.7078	**0.7409**	0.7301	EfficientNet
Cardiomegaly	0.8273	**0.8875**	0.8665	EfficientNet
Effusion	0.7710	**0.8159**	0.7865	EfficientNet
Infiltration	0.6793	0.6991	**0.6957**	EfficientNet
Mass	0.6991	**0.7928**	0.7486	EfficientNet
Nodule	0.6585	**0.7330**	0.6985	EfficientNet
Pneumonia	0.6462	0.6845	**0.7239**	ViT
Pneumothorax	0.7812	**0.8482**	0.8173	EfficientNet
Consolidation	0.7024	0.7286	**0.7424**	ViT
Edema	0.7911	**0.8329**	0.8451	ViT
Emphysema	0.7474	**0.9218**	0.8081	EfficientNet
Fibrosis	0.7586	0.7819	**0.8144**	ViT
Pleural Thickening	0.7093	**0.7603**	0.7517	EfficientNet
Hernia	0.8373	0.8955	**0.9495**	ViT
Mean AUC	0.7369	**0.7945**	0.7842	EfficientNet

Table 1. Per-class AUC comparison across all models (Weighted BCE, 512px images). Best performance highlighted.

Configuration	EfficientNet	vs Best	ViT	vs Best
Weighted BCE, 512px	**0.7945**	-	0.7842	-
Focal Loss, 512px	0.7459	-6.1%	0.7461	-4.9%
Weighted BCE, 384px	0.7402	-6.8%	0.7285	-7.1%
Focal Loss, 384px	0.7128	-10.3%	0.7063	-9.9%

Table 2. Ablation study: Impact of loss function and image resolution on mean AUC.

Metric	Baseline	Eff-Net	ViT
Mean AUC	0.737	**0.795**	0.784
Accuracy	0.915	0.923	0.911
Exact Match Ratio	0.351	0.362	0.310
Hamming Loss	0.085	**0.077**	0.089
Improvement vs Baseline	-	**+7.8%**	+6.4%

Table 3. Overall performance metrics comparison.

3.3. Key Findings from Ablation Studies

3.3.1. Loss Function Analysis

- **Weighted BCE Superiority:** Weighted BCE consistently outperformed Focal Loss across both architectures and resolutions
- **Magnitude of Difference:** 4.9-6.1% performance gap in favor of Weighted BCE
- **Potential Reasons:** The direct class frequency-based weighting in Weighted BCE may be more effective for the

severe class imbalance in medical datasets compared to Focal Loss's hard example focusing

3.3.2. Image Resolution Impact

- **Higher Resolution Benefits:** 512px images consistently outperformed 384px across all configurations
- **Performance Degradation:** 6.8-10.3% performance drop with lower resolution
- **Clinical Implications:** Higher resolution preserves subtle pathological details crucial for accurate diagnosis

3.3.3. Optimal Configuration

The optimal configuration across all experiments was:

- **Architecture:** EfficientNet-B3
- **Loss Function:** Weighted BCE
- **Image Resolution:** 512px
- **Result:** 0.7945 mean AUC

3.4. Architectural Analysis

3.4.1. EfficientNet-B3 Strengths

EfficientNet demonstrated superior overall performance with particular excellence in:

- **Emphysema detection:** Outstanding AUC of 0.922 (+17.4% vs baseline)
- **Mass detection:** Significant improvement (0.793, +9.4%)
- **Consistent performance:** Best performer in 9 of 14 disease classes

3.4.2. Vision Transformer Insights

The ViT showed competitive performance with unique strengths:

- **Hernia detection:** Exceptional AUC of 0.950 (+11.2% vs baseline)
- **Edema and Fibrosis:** Strong performance in texture-based diseases
- **Training efficiency:** Achieved competitive results with fewer epochs

3.4.3. Baseline Performance

The ResNet-50 baseline provided solid performance but was consistently outperformed by modern architectures, particularly in detecting localized abnormalities and texture-based diseases.

4. DISCUSSION

4.1. Architectural Implications

Our comprehensive analysis reveals several important insights:

4.1.1. EfficientNet's Compound Scaling Advantage

The superior performance of EfficientNet-B3 can be attributed to its compound scaling methodology, which optimally balances network width, depth, and resolution. This appears particularly beneficial for medical images where both local details and global context are important.

4.1.2. ViT's Attention Mechanism Benefits

The Vision Transformer's strong performance in certain categories demonstrates the value of self-attention mechanisms for medical image analysis. The ability to model long-range dependencies appears advantageous for diseases with diffuse patterns like Edema and Fibrosis.

4.1.3. Loss Function Selection Critical

The significant performance difference between Weighted BCE and Focal Loss highlights the importance of proper loss function selection for medical imaging tasks. The direct class frequency-based approach of Weighted BCE appears more suitable for the severe imbalance in medical datasets.

4.2. Clinical Relevance

The performance improvements of modern architectures have direct clinical implications:

- **Emphysema screening:** EfficientNet's 0.922 AUC could support early detection programs
- **Hernia identification:** ViT's 0.950 AUC demonstrates reliability for this condition
- **Mass detection:** Significant improvements aid in cancer screening applications

4.3. Limitations and Challenges

- **Class Imbalance:** All models struggled with rare diseases (Pneumonia, Consolidation)
- **Precision-Recall Trade-off:** High AUC but low F1-scores for some diseases indicate threshold optimization challenges
- **Computational Requirements:** Higher resolution images require more computational resources

5. CONCLUSION AND FUTURE WORK

This comprehensive study demonstrates that modern deep learning architectures significantly outperform traditional CNN baselines for thoracic disease classification in chest X-rays. Our key findings include:

- **EfficientNet-B3 achieves state-of-the-art performance** with 0.795 mean AUC, 7.8% improvement over baseline
- **Weighted BCE outperforms Focal Loss** for medical image classification tasks

- **Higher image resolution (512px) provides significant benefits** over lower resolutions
- **Vision Transformer shows competitive results** with unique strengths in attention-based pattern recognition
- **Architecture selection should consider disease characteristics** - CNNs for localized findings, Transformers for diffuse patterns

5.1. Limitations in Weakly-Supervised Localization

While our models demonstrated strong classification performance, the weakly-supervised bounding box generation using GradCAM revealed significant limitations in precise localization accuracy. As noted in the original ChestX-ray14 paper [1], weakly-supervised methods face inherent challenges in producing precise spatial localizations. The generated bounding boxes often capture the general region of pathology but lack the precision required for clinical localization tasks. This limitation stems from several factors: the class activation maps highlight discriminative regions rather than exact pathological boundaries, the resolution of activation maps is substantially lower than the original images, and the models are optimized for classification rather than precise spatial localization. These findings align with the original paper's observation that "deep convolutional neural network based 'reading chest X-rays' remains a strenuous task for fully-automated high precision CAD systems."

5.2. Future Directions

1. **Loss Function Optimization:** Explore alternative loss functions specifically designed for multi-label medical image classification
2. **Resolution-Adaptive Training:** Investigate progressive resolution training or multi-scale approaches
3. **Improved Localization Methods:** Explore attention mechanisms and spatial transformers for more precise bounding box generation
4. **Weakly-Supervised Refinement:** Implement iterative refinement techniques to improve localization accuracy
5. **Ensemble Methods:** Combine CNN and Transformer architectures to leverage respective strengths
6. **Multi-modal Integration:** Incorporate clinical data and radiological reports

The performance gains demonstrated by modern architectures highlight the ongoing evolution of medical AI and its potential to significantly impact clinical practice through improved diagnostic accuracy and efficiency. However, the limitations in precise localization underscore the need for continued research in weakly-supervised methods or the collection of more comprehensive localization datasets.

Acknowledgments

This study was performed using the ChestX-ray14 dataset. We acknowledge the authors of [1] for making this valuable resource publicly available. Computational resources were provided by CDS Department.

6. REFERENCES

[1] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2097–2106.

[2] Mingxing Tan and Quoc V Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.