Chest X-Ray Feature Pyramid Sum Model with Diseased Area Data Augmentation Method

Changhyun Kim*1, Giyeol Kim†2, Sooyoung Yang‡2, Hyunsu Kimψ2, Sangyool Lee*3, Hansu Cho*3

* SK Telecom
† Gacheon University
‡ ChungAng University
ψ SungKyunKwan University

Abstract

Deep learning has shown considerable promise in medical image analysis, but significant challenges remain. These stem from the inherent complexities of medical images, such as **varying sizes of lesions within the same image and the potential coexistence of multiple diseases**. To address these issues, we propose a novel model combining TResNet [6] with Feature Pyramid Network (FPN) [4]. This model adeptly handles multi-label classification, demonstrating robust performance across a range of lesion sizes. Furthermore, most medical images follow **a long-tail distribution, presenting class imbalance problems**, where the occurrence of one lesion often correlates with the presence of others. Considering these correlations, we introduced a strategy for dealing with the class imbalance issue by augmenting minority classes using bounding box information of the disease. Our proposed approach offers a novel solution for handling the unique challenges in deep learning-based medical image analysis, paving the way for more precise interpretations of complex medical images. The performance of mAP in 26 disease classes has been improved from 32.76% to 33.37% in a single model, and 35.11% in ensemble model.

Introduction

-The CXR-LT Challenge aims to solve the class imbalanced multi-label classification problem of 26 lung diseases. The size and shape of the lesions vary depending on the type of disease.

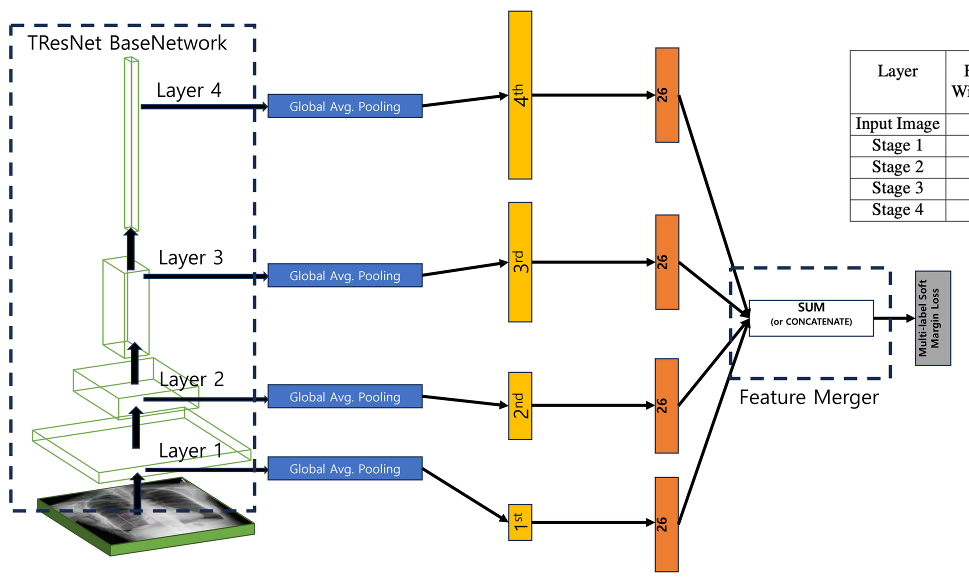
-This study proposes a model called the **Feature Pyramid Sum Model (FPSM)** based on the TResNet BaseNetwork, which takes into account the various sizes of lesions.

-Furthermore, we attempted to address the Long-tail problem by proposing the **Diseased Area Data Augmentation Method (DADAM)** as a data augmentation approach.

-Moreover, we successfully enhanced the performance of the baseline TResNet model by applying various Hyper Parameter Optimization (HPO) techniques and Neural Architecture Search (NAS).

-We randomly split the CXR-MIMIC dataset into an 80-20 ratio for training and validation datasets.

Method



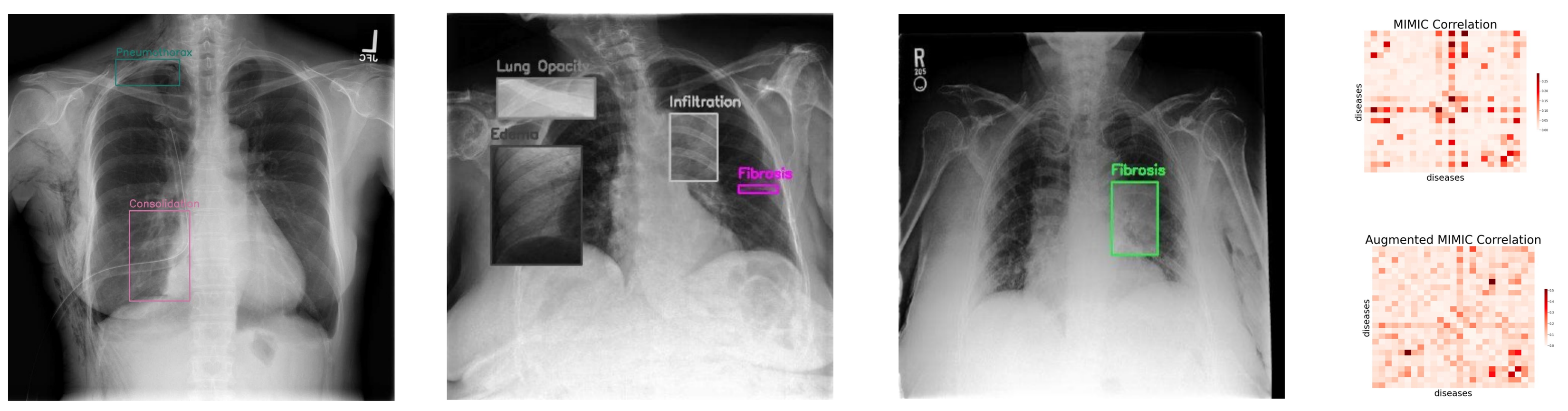
Layer	Feature map Width × Height	Channels	Feature Dimension
Input Image	448 × 448	3	N/A
Stage 1	112 × 112	76	76 (+2)
Stage 2	56 × 56	152	152 (+2)
Stage 3	28 × 28	1216	1216 (+2)
Stage 4	14 × 14	2432	2432 (+2)

Figure 1. Chest X-Ray Feature Pyramid Sum Model (CXR-FPSM)

-We extract layer-specific feature maps, considering various organ sizes, and convert them into a single 1D-channel vector.

-(Optional) Furthermore, by incorporating binary embedding, we can provide an additional indicator of whether the data is synthetic or not.

- We compute the multi-label classification loss from features extracted on each different layers and aggregate the classifier outputs according to the resolution of features, achieving an ensemble effect.



(a) Original patch image (b) augmented with normal image (c) augmented with disease image (d) Corr. Coefficients

Figure 2. Illustrative images using Diseased Area Data Augmentation Method (DADAM)

-(a) Original disease patch images from the VinDR-CXR dataset.

-(b) (CutMix approach) Cut & paste disease patches onto normal images using bounding box information from the VinDR-CXR, NIH Chest X-ray, and ChexDet datasets.

-(c) (Mixup approach) Images created by attaching a randomly selected disease patch to a disease image along with another disease patch that has the highest correlation, with a transparency of 0.6.

-(d) Augmented MIMIC dataset has similar correlation coefficient matrices as MIMIC datasets showing preservation of cooccurrence in different diseases. (Diagonal components are all set to zero and did the absolute process on all matrix components)

Experiment Result

Used Feature Pyramid	F1	mAP	Recall	Pre-cision	AUC
4^{th}	27.13	32.77	75.79	20.46	82.65
$4^{th} + 2^{nd}$	27.24	32.79	76.20	20.53	82.62
$4^{th} + 3^{rd} + 2^{nd}$	27.51	**33.10**	**76.98**	20.87	**82.83**
$4^{th} + 3^{rd} + 2^{nd} + 1^{st}$	**27.68**	32.76	76.06	**20.90**	82.60

(a) use_binary_enc=0, use_concatenate=0

Used Feature Pyramid	F1	mAP	Recall	Pre-cision	AUC
4^{th}	27.18	32.92	**76.47**	20.45	82.81
$4^{th} + 2^{nd}$	27.18	32.59	76.28	20.52	82.57
$4^{th} + 3^{rd} + 2^{nd}$	**27.36**	**33.04**	76.14	20.68	**82.81**
$4^{th} + 3^{rd} + 2^{nd} + 1^{st}$	27.35	32.94	75.73	**20.69**	82.67

(b) use_binary_enc=1, use_concatenate=0

Used Feature Pyramid	F1	mAP	Recall	Pre-cision	AUC
4^{th}	27.23	32.66	76.39	20.54	82.69
$4^{th} + 2^{nd}$	26.97	30.92	74.32	20.33	81.26
$4^{th} + 3^{rd} + 2^{nd}$	27.12	31.96	74.91	20.42	81.75
$4^{th} + 3^{rd} + 2^{nd} + 1^{st}$	26.88	31.61	75.57	20.22	81.61

(c) use_binary_enc=1, use_concatenate=1

Used Feature Pyramid	F1	mAP	Recall	Pre-cision	AUC
4^{th}	27.18	32.76	76.74	20.48	82.58
$4^{th} + 2^{nd}$	26.84	30.99	74.41	20.37	81.12
$4^{th} + 3^{rd} + 2^{nd}$	26.98	32.21	74.93	20.31	81.86
$4^{th} + 3^{rd} + 2^{nd} + 1^{st}$	27.01	31.87	74.83	20.35	81.62

(d) use_binary_enc=0, use_concatenate=1

Table 1. Performance of Feature Pyramid Sum Model (FPSM) Classifier

Experiment Result (continued…)

- A total of 16 experiments were conducted, comprising four variations of binary encoding (synthetic or original) combined with either sum or concatenate operations, and four different feature combinations.

-Overall, the performance of the (a) and (b) cases, where the final features are summed, was superior compared to the sequentially concatenated (c) and (d) cases.

- For the performance metrics of our interest, mAP and AUC, the combination of features with the sum method, specifically when combining the 4th, 3rd, and 2nd features, demonstrated the highest performance (mAP=33.10%, AUC=82.83%).

Data	Image Size	Model	Val Acc	Val mAP
Orig. MIMIC	512	MoCo-v2 (ResNet50)	71	25.7
Orig. MIMIC	384	ViT	78	29.5
Aug. MIMIC Fine-tuning	384	ViT	85	28.1
Aug. MIMIC Scratch	384	ViT	70	26.1
Orig. MIMIC	512	FPSM (Proposed)	N/A	33.1
Aug. MIMIC Fine-tunning	512	FPSM (Proposed)	N/A	33.37

Table 2. Ablation Studies on both Model and Dataset

-The proposed FPSM model outperforms MoCo-v2 (Self-supervised approach) [5] and ViT [8] methods by 7.4% and 3.6% in terms of mAP performance, respectively.

-When applying the proposed DADAM data augmentation, the performance decreases by 1.4% in the case of ViT, whereas it improves by approximately 0.27% in the case of FPSM.

Models	CK score	Acc	AUPRC	AUC
[1] Proposed 1	**0.694**	**0.848**	0.898	0.903
[2] Proposed 2	0.680	0.842	**0.903**	**0.906**
[Base line]TResNet [6]	0.674	0.844	0.894	0.899
CNN Ensemble [1]	0.66	0.797	n/a	n/a
VGG16 [7]	0.532	0.767	n/a	n/a

[1]Feature Pyramid=$4^{th}+3^{rd}+2^{nd}+1^{st}$, 7part label embedding = 0, use_concatenate=0
[2]Feature Pyramid=$4^{th}+3^{rd}+2^{nd}$, 7part label embedding = 1, use_concatenate=0

Table 3. FPSM performance on MURA MSK dataset [9]

-[Model Generalization] The proposed FPSM model also demonstrates superior performance compared to state-of-the-art models in the binary classification of Musculoskeletal abnormalities on the MURA MSK dataset.

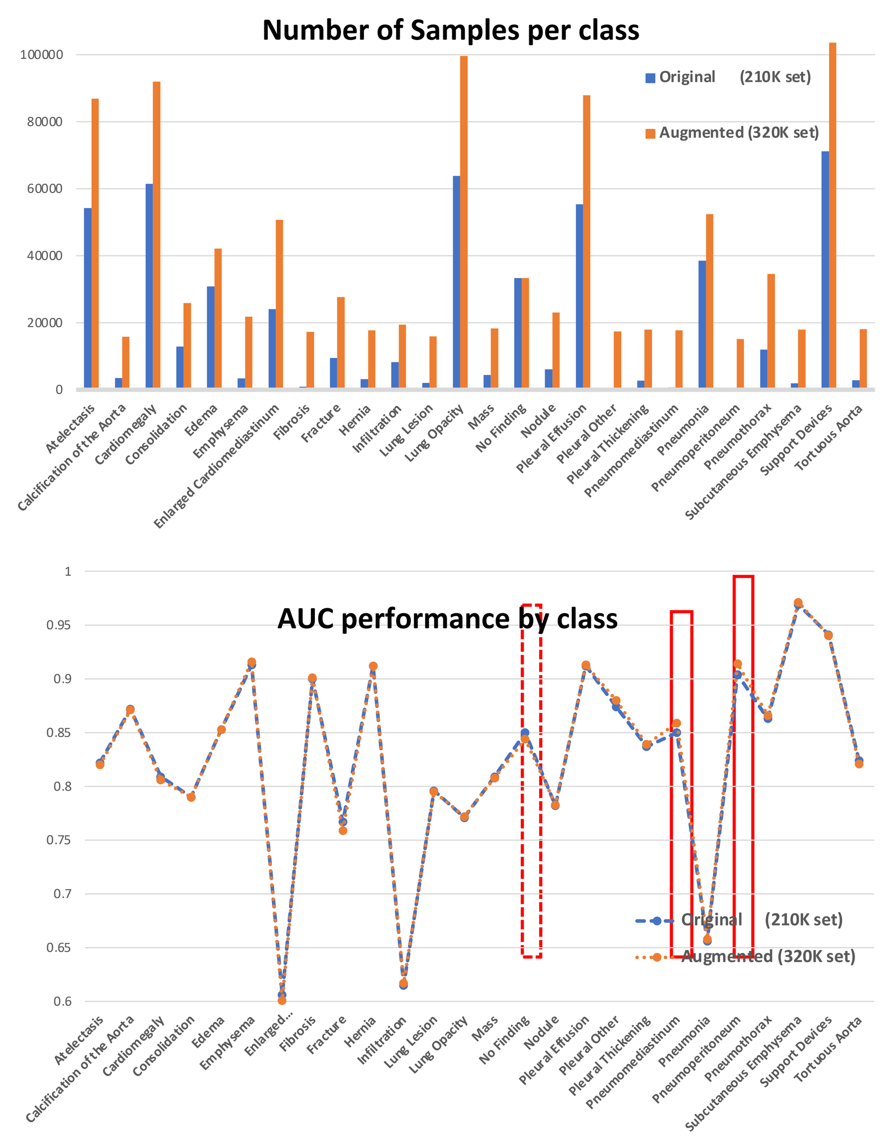


Figure 3. Number of samples and AUC performance on each class

-In the original dataset (210K set), we observed a noticeable 1% improvement in performance for the perfect minority classes such as **'Pneumomediastinum'** and **'Pneumoperitoneum'** when fine-tuning with the augmented dataset (320K set).

-Other minority classes showed less significant changes in performance, suggesting the need for additional experiments with varying sample counts and transparency adjustments in the future.

-As an exception, in the case of **'No Finding,'** there was no change in the number of samples, which led to an overall decrease in performance.

Conclusion & Future work

-We successfully propose a network, "FPSM (Feature Pyramid Sum Model)", for recognizing various-sized lesions in medical images. We compare it with the basenetwork, TResNet, and confirm its superiority in most performance metrics (F1, mAP, Precision, AUC) in medical image recognition.

-To address the long-tail problem, we augmented data for minority classes with more than 15k samples on each class using both CutMix [2] and MixUp [3] techniques with various bounding box labeled datasets. We observed an increase in mAP of approximately 0.27% as a result.

-The model submitted for the challenge achieved a performance of 35.1% at validation and 32.8% at the test phase by ensembling the final output probabilities of three models: 33.10% (FPSM without binary encoding), 33.04% (FPSM with binary encoding), and 33.37% (FPSM with DADAM fine-tuning), using a mean ensemble.

-It is necessary to conduct separate experiments to understand the impact of CutMix (the number of patches being attached) and MixUp (flexibility to control transparency levels) techniques on the performance of minority classes.

References

[1] Dennis Banga and Peter Waiganjo. Abnormality detection in musculoskeletal radiographs with convolutional neural networks (ensembles) and performance optimization. arXiv preprint arXiv:1908.02170, 2019.
[2] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, SanghyukChun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features, 2019.
[3] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization, 2018.
[4] Tsung-Yi Lin, Piotr Doll´ar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection, 2017.
[5] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning, 2020.
[6] Tal Ridnik, Hussam Lawen, Asaf Noy, Emanuel Ben Baruch, Gilad Sharir, and Itamar Friedman. Tresnet: High performance gpu-dedicated architecture, 2020.
[7] AFM Saif, Celia Shahnaz, Wei-Ping Zhu, and M Omair Ah-mad. Abnormality detection in musculoskeletal radiographs using capsule network. IEEE Access, 7:81494–81503, 2019.
[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Syl- vain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.
[9] Aarti Bagul Daisy Ding Tony Duan Hershel Mehta Brandon Yang Kaylie Zhu Dillon Laird Robyn L. Ball Curtis Langlotz Katie Shpanskaya Matthew P. Lungren Andrew Y. Ng Pranav Rajpurkar, Jeremy Irvin. Mura: Large dataset for abnormality detection in musculoskeletal radiographs, 2018.

Acknowledgements

This work was supported by the Korea Medical Device Development Fund grant funded by the Korea government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, the Ministry of Food and Drug Safety) (Project Number:1711194307, RS-2020-KD000093)