Enhancing Multi-Label Long-Tailed Classification on Chest X-Rays through ML-GCN Augmentation

HyeRyeong Seo*[1†] MinHyuk Lee*[1‡] WooJin Cheong*[1†] HyeKyung Yoon*[1†]
SoHyung Kim*[1†] MyungJoo Kang[1‡]
[1]Seoul National University, South Korea
{202222972, 356min, mousejr1128, yhk04150, ks000225, mkang}@snu.ac.kr

Abstract

The classification of multi-label thoracic images presents a considerable challenge due to the severe intrinsic imbalances inherent in the dataset. During the testing phase, the model encounters both predominant (head) and less frequent (tail) classes, demanding not only proficiency in image feature extraction but also a comprehensive understanding of label relationships. Traditional medical image classifiers have historically relied on exploiting a small number of dominant head classes. Nevertheless, this approach often yields suboptimal classification outcomes. To resolve this issue, we propose an enhanced version of the Multi-Label Graph Convolutional Network (ML-GCN). Our approach integrates the incorporation of experts, each focusing on distinct aspects of the input dataset, class-balanced sampling, Log-Sum-Pooling (LSE pooling), an attention layer, and regularization through KL divergence. By synergistically applying these techniques, our model significantly outperforms the baseline vanilla ML-GCN, capitalizing on nuanced architectural adjustments. Through this comprehensive approach, we effectively demonstrate the versatility of our model in addressing the specific task of multi-label long-tailed classification within the realm of chest X-ray datasets. Furthermore, our methodology exhibits promising potential for extension to a diverse array of datasets characterized by long-tailed distributions, establishing a strong foundation for its application within various domains. In order to ensure the reproducibility of this study, we will make the source code publicly available: github.com/lisaseo9704/2023_ICCVW_CVAMD_NCIA500

1. Introduction

Deep Learning (DL) has emerged as a transformative force within diverse medical domains, due to its unparalleled capability for rapid and accurate disease diagnosis. The emergence of COVID-19 has underscored the pivotal importance of expeditious and reliable diagnostic procedures, particularly in the domain of Chest X-ray images. Contemporary DL models possess the ability to pinpoint precise areas of pathological conditions and provide cogent explanations for predictions when analyzing individual X-ray images. As a result, these DL models are progressively being leveraged to assist medical practitioners to arrive at precise diagnostic determinations.

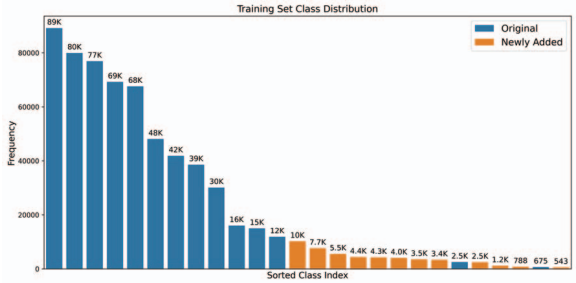
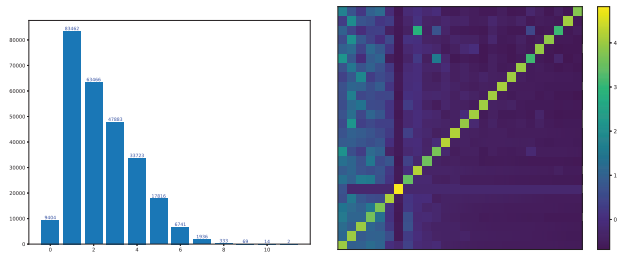


Figure 1: Distribution of the imbalanced chest X-ray dataset

However, the efficacy of previous methods diminishes when confronted with imbalanced datasets harboring a substantial number of classes. Notably, a bias towards simplicity may appear, wherein models tend to favor well-represented head classes, potentially compromising the overall fairness of the classification process. The competition's provided chest X-ray dataset exemplifies this issue acutely, featuring a remarkable imbalance with 26 distinct classes, as depicted in Figure 1. To mitigate this inherent bias, we approach the given task from a multifaceted perspective.

*Equal contribution. †IPAI, Seoul National University, South Korea. ‡Department of Mathematical Sciences, Seoul National University, South Korea. Correspondence to: MyungJoo Kang <mkang@snu.ac.kr>.
Proceedings of IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Paris, France, 2023.



(a) (b)

Figure 2: (a) The number of labels in a single image (b) Head class (except 'no finding') diseases were diagnosed with other class ones. The arrangement follows a descending order, commencing with the head class and extending towards the tail class. Subsequently, the data was normalized on a column-wise basis.

In contrast to multi-class classification, multi-label classification introduces the possibility of co-occurring labels. Analysis of the dataset reveals that out of $264,849$ images, $171,983$ (more than half) manifest two or more labels concurrently, as vividly depicted in Figure 2(a). This high incidence of co-occurrence underscores the complex label relationships. Notably, many of the prominent head class diseases exhibit diagnostic connections with other classes, as shown in Figure 2(b). Intriguingly, 'No finding,' a leading head class, stands out as an exception, never co-occurring with other diseases, and thus previously overlooked.

Addressing these aforementioned issues, we added four experts of the same architecture to ML-GCN which enables the input data to be interpreted from distinct viewpoints. Since there is a possibility that all of the experts could be biased toward head classes, we used RIDE loss [30] and KL-divergence as a regularizer for experts to encourage the learning of inconsistent semantic features respectively. Moreover, by implementing a sampling method in our model, we were able to enhance the performance to some extent in the tail classes.

Furthermore, the utilization of ML-GCN enabled us to uncover label dependencies, a pivotal aspect given the prevalent co-occurrence patterns in multi-label classification tasks.

Our contributions can be summarized as follows:

- We empirically demonstrate the effectiveness of sampler, pooling, attention layer and KL-divergence functioning as a regularizer in the context of multi-label long-tail classification using chest X-ray datasets. This suggests the potential applicability of our methodology to a wide range of datasets characterized by long-tailed distributions.
- The capability of ML-GCN to discern label dependencies and its harmonization with the newly introduced experts have resulted in a substantial enhancement in the overall performance of the model.

2. Related work

Long-Tailed Classification Within the realm of extensive medical image datasets, imbalances and long-tailed distributions are widespread, characterized by the abundance of samples in head classes while tail classes contend with scarcity and associated challenges. A variety of strategies have emerged to navigate the complexities of long-tailed classification.

One common technique involves dataset resampling [4, 10, 23, 37]. This approach entails oversampling tail classes while undersampling head classes. However, this methodology often triggers overfitting in tail classes, rendering insufficient training for the head classes. Recent advancements have been aimed at mitigating concerns related to overfitting. For instance, [21] sought balance by introducing synthetic minority samples derived from majority samples. However, we chose not to adopt this approach due to the possible presence of noise in synthetic X-ray images, which might hinder the effectiveness of training.

Another avenue addresses label ratios and their reflection in the loss function [22, 8, 26, 3]. Focal loss [22] fine-tunes cross-entropy by assigning lower weights to easily learnable data and prioritizing challenging or misclassifiable instances. Given that tail classes manifest less frequently in mini-batches, higher learning rates are assigned, while head classes' frequent appearance calls for lower learning rates during each training step. However, the intrinsic attributes of long-tailed distributions can lead to challenges, with higher learning rates disproportionately affecting tail classes and thus impeding proper model training.

"Decoupled training," an additional technique, aims to disentangle representation learning from classifier learning [20, 19, 38, 6, 29, 33, 9]. Notably, [20] demonstrated the efficacy of this approach by highlighting that classifiers trained on long-tailed distributions tend to exhibit smaller norms, particularly for tail classes. Appropriate countermeasures are then applied to redress this imbalance.

Enhancements in long-tailed classification performance have also been realized through model improvements. Ensemble methodologies, such as [36] have recently gained prominence. These often involve a shared feature-extraction backbone accompanied by multiple experts for label classification [2, 7, 30, 35]. SADE [35] has three expert, each trained with different loss functions. The first is inclined towards the original long-tailed distribution, the second towards uniform class distribution, and the third

towards inversely long-tailed class distribution. This design fosters SADE's robustness across diverse test class distributions. [11] harnessed two classifiers, a Classification Head (CLS Head), and a Supervised Contrastive Learning Head (SCL Head), to glean image representations of normal (head class) and abnormal (tail class) instances through contrastive learning. This approach yielded impressive results in thorax datasets.

Multi-Label Classification Multi-Label Classification entails predicting zero or multiple classes for a given input instance (image). Within the multiple-instance learning (MIL) framework, PCAM [32] computes the likelihood of disease lesions within extracted features. While proficient in thoracic disease classification, PCAM's focus on head classes impeded its suitability for scenarios featuring numerous classes, often resulting in overfitting when the class count exceeds 20. Moreover, numerous studies have pursued patch-based methodologies to comprehend thorax images. For instance, [1] partitioned images into non-overlapping patches, selecting those that encapsulate the most informative features. Although computationally efficient, this approach overlooks label dependencies. Additionally, [27] used Swin Transformer blocks for multi-label classificaion on chest X-ray images surpassing DNet [12] which employed location-aware Dense Networks (DNetLoc) to integrate image data and spatial information for tail classes, based on location-aware Dense Networks (DNetLoc).

In contradistinction to binary multi-class classification, binary multi-label classification grapples with mutually non-exclusive labels, necessitating label correlations. In medical domains, such correlations are frequent due to inherent human nature. Consequently, prior research often centers around pathological prior knowledge to formulate models [14, 16].

3. Method

In this section, we elaborate on our strategies of the given multi-label long-tailed classification task. We draw inspiration from the adjacency matrix, which captures label dependencies, and embrace a multi-modality approach. The pipeline of our model is visualized in Figure 3

3.1. Base line

To address the multi-label classification task, we set ML-GCN [5] as our baseline model. This model utilizes ResNet-101 as its backbone and stacked GCN [34] as a classifier. The inherent capacity of this classifier lies in its adeptness at aggregating substantial label-related information through a graph structure, where nodes symbolize labels and edges signify label co-occurrences. Therefore, the stacked GCN replaces MLP. To offer deeper insights into

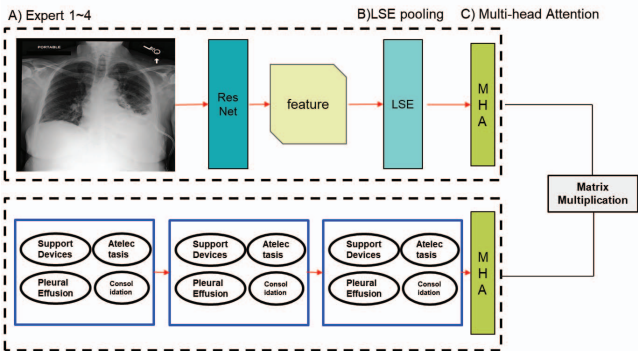


Figure 3: A) The newly added experts extract features from an image, followed by B) LSE pooling applies to these features. Finally, we incorporate (C) a single layer of the Transformer Encoder, implementing a Multi-Head Attention (MHA) mechanism. The purpose of the MHA is twofold: to mitigate information loss and to explicitly represent the inherent features of the data.

the architectural design of our network, the backbone generates a 512-dimensional vector denoted as $\mathbf{u} \in \mathbb{R}^{26}$, whereas the expert module generates the matrix $\mathbf{A} \in \mathbb{R}^{26\times512}$. By applying a sigmoid function to their matrix product, $\mathbf{Au}$ we derive the anticipated probabilities associated with each label. To put it more succinctly, the i^{th} value within $\mathbf{Au}$ represents the logit of the i^{th} label's probability.

This process highlights that the probability associated with a given label solely relies upon the corresponding row in matrix $\mathbf{A}$. The i^{th} row, labeled as $\mathbf{A}_i$, interacts with vector $\mathbf{u}$ through an inner product, resulting in a precise value within $\mathbf{Au}$ for that label's index.

In essence, $\mathbf{A}_i$ serves as a representative feature for its corresponding label. Consequently, we individually examined the 26 rows of matrix $\mathbf{A}$, treating each one as a 512-dimensional representation of a distinct label. Our objective was to enhance this representation to render it more suitable for long-tailed multi-label classification. While drawing inspiration from [20], we encountered the challenge that CXR LT datasets demonstrate high levels of label co-occurrence. Consequently, an alternative approach was necessary to extract correlation information between labels, instead of individually calibrating weights for each label. To address this, we leveraged Multi-Head Attention (MHA) as a means to effectively utilize the correlating information among labels.

Concurrently, we applied a weight-sharing MHA to the $\mathbf{u}$, an output from the backbone. This step aimed to elevate the vector $\mathbf{u}$ into a more potent and informative form. Further elaboration on these processes can be found in section "3.5. Multi-Head Attention".

The original ML-GCN was initially trained on the

Wikipedia dataset using 300-dim GloVe [24]. However, the pre-trained dataset significantly diverged and was incongruent with the chest X-ray dataset that we should use. Therefore, we created new sentences to train GloVe. We compiled positive labels for each data instance and subsequently assembled them into sentences. For instance, if both "No Finding" and "Support Devices" possessed positive labels, we combined them into a sentence in the format of "No Finding" + "Support Devices". This procedure yielded a customized adjacency matrix for the given dataset.

As previously mentioned, we incorporated four experts, each equipped with independent stacked GCNs. Each expert assumed a distinct role to enhance applicability to the multi-label classification task.

3.2. Class-balanced sampling

In light of the properties exhibited by long-tailed distribution datasets, data associated with tail classes often remains underrepresented within mini-batches Consequently, these instances might not be adequately trained, thereby weakening the classifier's robustness. To address this issue, we used balanced data sampling, wherein the number of labels from each class is considered, as opposed to random sampling. This sampling strategy ensures uniformity across classes. However, this method is not likely to be applicable to multi-label long-tailed datasets due to significant relevance of label co-occurrence within such datasets.

Let p_i denote the probability of drawing the i^{th} class label. This value, p_i is mathematically expressed as follows:

$$p_i = \sum_{j=1}^{C} p(j)\, p(i|j) = \sum_{j=1}^{C} p(j)\, \frac{n_{i,j}}{n_j} \tag{1}$$

where $p(j)$ represents the probability of drawing the j^{th} class label, $p(i|j)$ signifies the probability of drawing the i^{th} class label given the j^{th} class, n_j stands for the number of positive labels in the j^{th} class, and $n_{i,j}$ denotes the number of positive labels in the i^{th} class co-occurring with the j^{th} class label.

In datasets with single-label annotations, $n_{i,j}$ is zero when $i \neq j$, simplifying probability calculations based solely on label counts. However, in multi-label datasets, $n_{i,j}$ takes non-zero values even when $i \neq j$. This is especially significant since many labels frequently co-occur with the primary class, leading to an increase in $n_{i,j}$ when i corresponds to the primary class index. Consequently, even the class-balanced sampling distribution skews towards the main class.

As a result of co-occurrence, random sampling tends to disproportionately include the head class in mini-batches, exceeding the true label distribution. This exacerbates the tail distribution's imbalance, exceeding the challenges encountered in cases with single-label annotations, thus intensifying the challenges of learning from tail classes.

We delved into two class-balanced sampling strategies. Firstly, the "cycle" method involves sequential label selection, ensuring a uniform distribution of $p(j) = \frac{1}{C}$ across all labels. Random sampling, on the other hand, exhibits a noticeable skew towards the head class. However, cycle sampling equalizes the distribution to $\frac{1}{C}$, ectifying the sampled data distribution. Nevertheless, an imbalance persists due to the uncontrolled nature of $p(i|j)$ arising from co-occurrence.

The second method, 'least-sampled', prioritizes the smallest among previously drawn samples for the subsequent selection. For instance, the initial sample is drawn randomly, while subsequent samples are drawn from the unselected pool of preceding ones. Given that this method takes into account prior samples, a sufficiently large batch size is crucial. Notably, this approach skews the distribution of $p(j)$ towards the tail class. Consequently, while single-label annotations yield a tail-class skewed distribution, CXR-LT experiences an exacerbated imbalance, leading to a head-class skewed distribution even with this approach.

We adopted cycle sampling for our approach. When referring to class-balanced sampling later, we specifically refer to cycle sampling. Our rationale for this decision will be elaborated upon in Section 5.

Moreover, it should be noted that class-balanced sampling tends to under-sample head classes and over-sample tail classes to achieve uniformity across all classes. This approach, however, carries the potential risk of overfitting on tail classes.

Table 1 presents the observed improvements in mean Average Precision (mAP) scores resulting from the utilization of class-balanced sampling as compared to basic (random) sampling. The mAP scores are categorized by the class type: head, middle, and tail. Enhanced mAP scores are evident across all categories, including the tail class, as indicated in Table 1. The anticipated performance enhancements are particularly notable for middle and tail classes. Contrary to initial expectations, we also observed performance improvements in the head class. As previously mentioned, the head class is characterized by a substantial volume of data, which might dominate other classes due to its size and pronounced label co-occurrence. This situation suggests that the model may not effectively learn the features of other classes during random sampling, potentially misattributing these features to the head class, thereby resulting in reduced performance. However, it is noteworthy that the mAP score increase for tail classes was relatively marginal. Furthermore, in our pursuit of optimizing performance, we found it necessary to implement an early stopping mechanism during the training phase to mitigate the risk of overfitting, particularly within tail classes. This measure became imperative as we observed counter-

productive divergence between the progress of head classes and other classes. The intricate balance between enhancing tail classes and potential trade-offs with head and other classes posed a challenge, ultimately making the efficacy of class-balanced sampling for overall performance enhancement less pronounced.

In light of these considerations and to further enhance the model's performance, we explored alternative strategies, which led us to experiment with the application of a regularization technique.

	Head	Middle	Tail
Basic	0.4640537777	0.114901875	0.0939671
Class-Balanced	0.506851888	0.170249875	0.1075862

Table 1: Dividing classes into three sets: *Head*: Top 9 classes, *Tail*: Bottom 9 classes, *Middle*: Classes not in *Head* or *Tail*. And comparing the two sampling methods: *Basic*: mAP score for Random Sampling, *Cycle*: mAP score for Class-balanced Sampling for each classes

3.3. RIDE loss

Ensemble learning utilizing multi-expert models has demonstrated remarkable performance gains on long-tailed datasets [36]. However, a potential concern arises where certain experts may observe the same segments of training data. This similarity in their learnable weights could render the ensemble approach ineffective and devoid of merit. Furthermore, the adoption of over-sampling strategies for long-tailed distributed data introduces an elevated risk of overfitting.

Despite the modest impact on performance and the inherent risk of overfitting, we decided to use class-balanced sampling specifically for tail classes. To preempt the aforementioned risks, we harnessed the RIDE loss [30] which leverages KL-divergence for regularization across each expert.

Technically, KL-divergence is not a metric; rather, it gauges the dissimilarity between two given distributions. Typically, the two distributions closely align through minimizing KL-divergence. However, RIDE operates in the contrary manner. It optimizes experts by striving to disjoint the support of the two distributions as extensively as possible. Additionally, RIDE optimizes classification loss to achieve congruent classification outcomes across diverse, independent experts.

The initial formulation of the RIDE loss entails the summation of KL-divergence for all pairs of experts, as expressed below:

$$\mathcal{L}_{div}(x) = \frac{-1}{n-1} \sum_{i \neq j} KL(\mathbf{p}^{(i)} | \mathbf{p}^{(j)}) \tag{2}$$

where n signifies the number of experts, and $\mathbf{p}^{(i)}$ represents the distribution derived from the i^{th} expert.

Due to its computationally intensive nature, the original RIDE loss was approximated with Equation (3), offering greater practicality. Considering n as the number of experts, the total number of terms in Equation (2) is $(n-1)!$.

$$\mathcal{L}_{div}(x) = \frac{-1}{n-1} \sum_{i} KL(\bar{\mathbf{p}} | \mathbf{p}^{(i)}) \tag{3}$$

where $\bar{p}$ represents the average of distributions across each expert.

This substitution reduces the summation complexity from (2) from $(n-1)!$ to n, thus enhancing computational efficiency.

We used the binary cross-entropy (BCE) loss as our classification loss function. Ultimately, the comprehensive loss formulation is as follows:

$$\mathcal{L}_{total}(x, y) = \mathcal{L}_{clf}(x, y) + \lambda \mathcal{L}_{div}(x) \tag{4}$$

The value of λ holds significance, as an excessively large value can undermine classification performance. Opting for a moderately small, positive value is prudent. (In our study, we set λ to 0.2.)

By adhering to this approach, we aim to mitigate the potential pitfalls associated with expert overlap and overfitting, fostering a more robust and effective ensemble learning strategy.

3.4. Pooling

Pooling serves as a pivotal and effective technique for extracting pertinent information from X-ray images, which initially incorporated max pooling. However, we replaced max pooling with Log-Sum-Pooling (LSE) [25] to precisely identify disease-specific regions. The formulation for LSE pooling is presented as follows:

$$x_p = \frac{1}{\gamma} \cdot log\Big[\frac{1}{P} \cdot \sum_{(i,j) \in P} exp(\gamma \cdot x_{i,j})\Big] \tag{5}$$

where x_p denotes the resultant feature subsequent to LSE pooling, (i, j) signifies the spatial coordinates of the extracted feature set P from the pre-trained Resnet [13]. The hyper-parameter γ is introduced, where its value is predetermined as 0.5.

In contrast to the earlier work of [31], we introduced LSE pooling directly following the feature extractor, eschewing the inclusion of an intermediary transition layer. This pooling methodology was uniformly applied across all experts within our model.

3.5. Multi-Head Attention

Considering the attention-free nature of the vanilla ML-GCN architecture, we postulated that incorporating Multi-Head Attention (MHA) could enhance the original model's

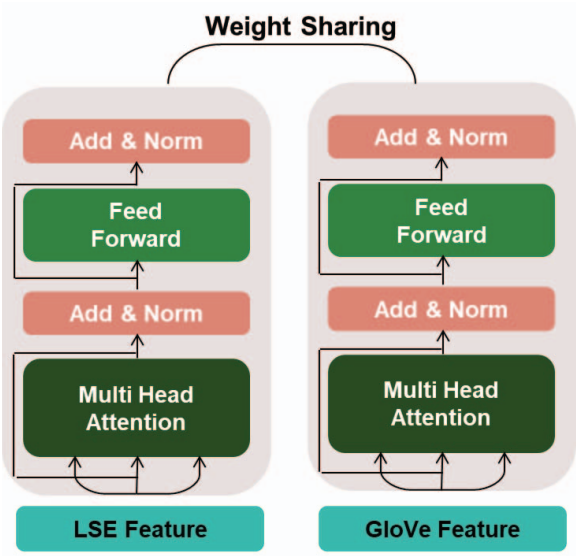


Figure 4: A layer of Transformer Encoder applied to each modality

ability to grasp intricate relationships between the input thorax image and the GloVe vector. To address this hypothesis, we introduced a singular layer of Transformer Encoder [28], with its weight shared between the pooled image vector and the GloVe vector. By doing so, MHA becomes capable of capturing correlations between the image representation and label dependencies, thereby facilitating a richer understanding of the data.

$$Attention(Q_Q, K_Q, V_Q) = softmax(\frac{Q_Q \cdot K_Q^T}{\sqrt{d_K}} V_Q) \quad (6)$$

For clarity of this notation, Q indicates both the pooled vector derived via LSE pooling, and the GloVE vector encapsulating label dependencies. Meanwhile, d_K denotes the dimension of the K_Q vector.

$$MHA(Q_Q, K_Q, V_Q) = Concat(head_1, \ldots, head_h)W$$
$$head_i = Attention(Q_Q, K_Q, V_Q)$$

This process segregates the input vector into multiple heads, permitting attention score calculation within each head. By doing so, the model gains the capability to capture intricate relationships within the input vector, consequently enhancing its capacity to represent the input vector more comprehensively.Referencing Table 2, our implemented approach incorporating MHA showcased discernible improvements in the model performance. Furthermore, it is confirmed that MHA of each expert actually contributed to enhancing their performance. Additionally, we validated that the introduction of MHA for each expert indeed contributed to performance enhancement. It's noteworthy, however, that alternative methodologies involving multiple encoders exhibited no significant impact on the model's performance, as evidenced by the observed results.

This strategic integration of MHA demonstrates its potential to significantly bolster the capabilities of the ML-GCN framework. By imbuing the model with the capacity to discern complex relationships within the data, MHA offers an avenue for substantial performance improvements, as showcased by the empirical evidence.

4. Experiment

4.1. Dataset

To address the growing need for automated and accurate disease diagnosis from radiological images, extensive datasets like MIMIC-CXR [18] and ChestX-ray8 [31] have been made available..

For this competition, we were furnished with the expanded MIMIC-CXR-JPG [17].It offers a substantial collection of thorax images in a JPG format, originating from MIMIC-CXR. This dataset includes both structured labels and textual reports. Notably, MIMIC-CXR-JPG presents a distinct advantage in terms of convenience, as compared to MIMIC-CXR, which supplies the dataset in a DICOM format. While MIMIC-CXR-JPG is comprised of 14 classes, the dataset provided for this competition integrated an additional 12 classes. Consequently, our task entailed the classification of an extremely imbalanced dataset encompassing a total of 26 classes, known as the CXR-LT [15].

4.2. Evaluation metrics

In the context of this competition, the assessment of model performance employed three key metrics: mean Average Precision (mAP), mean Area under the Receiver Operating Characteristic curve (mAUC), and mean F1 score (mF1). Out of these metrics, mAP served as the primary evaluation criterion, selected due to its suitability for appraising performance within the long-tailed dataset context.

4.3. Implementation details

Each stacked GCN block comprised two layers of Graph Convolutional Networks (GCNs), with dimensions of 1024 and 2048 for each respective layer. The word embedding methodology followed the comprehensive approach delineated in 3.1. This procedure entailed the transformation of each label node within the graph into a 300-dimensional vector. The backbone of our ML-GCN model was ResNet-101, pre-trained on ImageNet. The final Convolutional Neural Network (CNN) block of ResNet-101, in conjunction with the stacked GCN, constituted a single expert. As elucidated, we incorporated a total of 4 experts, each attuned to extract distinctive facets of information from the in-

put images. Additionally, we integrated Log-Sum-Pooling (LSE) and Multi-Head Attention (MHA) into our model. The hyperparameter γ for LSE pooling was specifically set to 0.5.

Prior to processing, the input images underwent bilinear interpolation resizing to dimensions of $1024x1024$. To foster balanced class representation, we employed class-balanced sampling. The Adam optimizer was chosen, characterized by a learning rate of $1e-4$ and a beta values set at $[0.9, 0.99]$. Finally, for the RIDE loss, λ was assigned a value of 0.2.

In total, the architecture was meticulously designed to compel the model to comprehend varied perspectives that efficiently extract pertinent information from input images. Additionally, this design aimed to prevent any undesirable information loss throughout the processing pipeline.

4.4. Diversity of experts

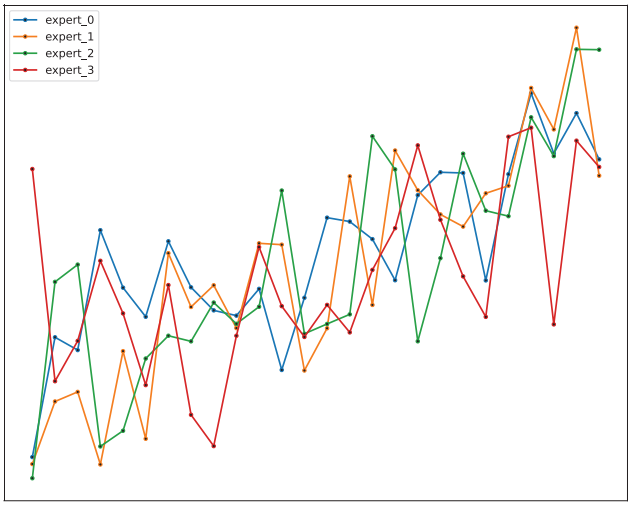


Figure 5: The norm values of each label for the experts are arranged in descending order, ranging from the head class to the tail class.

KL-divergence was leveraged as a training mechanism for our experts, enabling them to discern distinct features within the tail class data. In order to empirically validate our hypothesis that experts would specialize in different aspects, we conducted a thorough analysis of the weight values assigned to each label within the stacked Graph Convolutional Networks (GCNs). The weight profiles for all experts were visualized in Figure 5, revealing a semblance among the graphs. However, our assertion that the experts underwent diverse training is substantiated by the variance observed in their respective norms. In essence, this indicates that these experts are attuned to identifying disparate segments within the latent vectors. Remarkably, despite this diversity, the classification outcomes remained consistent.

4.5. Results

Model	Development Phase
ML-GCN	0.178
ML-GCN(2 experts)	0.229
ML-GCN(4 experts) + s	0.271
ML-GCN(4 experts) + s + RIDE	0.273
ML-GCN(4 experts) + s + RIDE + LSE + MHA	0.276

Table 2: The mean Average Precision(mAP) scores of our approaches were computed during the developmental phase. The variable s denotes the utilization of class-balanced sampling.

Model	Test Phase
ML-GCN(4 experts) + s + RIDE + LSE + MHA	0.279

Table 3: During the testing phase, the mAP score was evaluated to gauge the efficacy of our model.

Table 2 underscores that the mAP score of the vanilla ML-GCN was the lowest among the considered configurations. To assess the efficacy of experts, we introduced two additional experts to the vanilla ML-GCN. This augmentation led to a substantial improvement in the model's performance metrics. However, the model still demonstrated instability when classifying tail classes. As a remedy, we applied class-balanced sampling and integrated two more experts. This comprehensive approach, targeting both head and tail classes, led to a noteworthy surge in the mAP score, as evident from Table Table2. Subsequently, our exploration delved into the application of regularization techniques, specifically employing the RIDE loss. This refinement yielded a modest yet discernible enhancement in performance. Lastly, by implementing Log-Sum-Pooling (LSE) and incorporating an additional Multi-Head Attention (MHA) layer, we observed a further incremental boost in performance.

In comparison to the vanilla ML-GCN, our model exhibited remarkable performance gains. Notably, in the test phase (Table 3), we achieved a final mAP of 0.279, akin to the mAP score achieved during the developmental phase. This substantiates our claim that the model demonstrates not only robustness but also a degree of generalizability, as it exhibited consistent performance across distinct evaluation phases.

5. Limitations

While our approaches exhibited a progressive increase in mAP scores, there remain areas of further refinement. The inclusion of multiple experts, class-balanced sampling,

RIDE loss, LSE pooling, and MHA demonstrated their appropriateness for the task at hand. However, certain limitations are discernible, as outlined below:

- The generation of new sentences (as described in 3.1) aimed at identifying label co-occurrences proved to be suboptimal. The form or content of these sentences might not have been conducive to the model's requirements, possibly leading to a scarcity of relevant information. To address this, an alternative strategy could involve leveraging the text files present within the MIMIC-CXR-JPG dataset, thereby potentially mitigating this information deficit.

- An analysis of Table 2 suggests that the efficacy of the RIDE loss was comparatively lower than that of other methods. While it is challenging to categorically assert that employing a regularizer was ill-suited, the results shown in Figure 5. warrant its consideration. Thus, an avenue for enhancement lies in exploring alternative regularizers or augmenting the RIDE loss with additional loss terms.

- In evaluating the performance of the RIDE loss for regularization, an alternative sampling strategy ('Least-Sampled', as indicated in Table 4) was investigated in place of class-balanced sampling ('Cycle', as indicated in Table 4). 'Cycle' involves repetitive and uniform label sampling, while 'Least-Sampled' selects the labels that have been sampled as few as possible. Although 'Least-Sampled' leads to a more balanced sample distribution compared to class-balanced sampling due to its uniformity, it also has a tendency to over-sample tail classes, potentially exacerbating overfitting. The significant disparity between the most frequent label and the least frequent label by a factor of around 160 underscores the long-tailed nature of the dataset. Consequently, even under 'Cycle' sampling, the support device to Pneumoperitoneum ratio differs by approximately 10 times. Applying 'Least-Sampled' narrows this gap to around 4 times. Intuitively, under the presumption that our model had been adequately regularized to resist overfitting, the performance should have witnessed a substantial improvement under 'Least-Sampled'. This is due to the model learning the tail class labels more frequently. Nonetheless, as highlighted in Table 4, the mAP score under 'Cycle' outperforms that under 'Least-Sampled'. This outcome suggests that while regularization did contribute to modest performance enhancement, it did not entirely succeed in alleviating overfitting, in reference to the results in Table 2

Taking into account of future endeavors, we intend to address the aforementioned aspects.

Least-Sampled	Cycle
0.236	0.271

Table 4: Comparison between the 'Least-Sampled' and 'Cycle-Sampled'

	Devices	Pneumoperitoneum	ratio
Cycle Sampling	200	21	9.52
Least Sampling	168	43	3.90

Table 5: The average number of labels assigned through sampling by 512 instances

6. Conclusion

Conclusively, the classification of multi-labels within thoracic images represents a formidable and intricate challenge, particularly under the circumstances of a highly imbalanced dataset. The task necessitates a comprehensive approach that encompasses both the recognition of label interdependencies and the extraction of image features. In the realm of label dependencies, we harnessed the foundational ML-GCN model. To address the latter facet, we undertook advancements and refinements to the ML-GCN architecture, thereby elevating the overall performance metrics. Foremost among these enhancements was the strategic incorporation of sampling techniques and the augmentation of expert perspectives, demonstrating their pivotal role in enhancing model efficacy. Additionally, we ascertained the utility of regularization techniques and pooling mechanisms, which proved pivotal for further augmenting model performance. Furthermore, our extensive experiments affirm the potential of class-balanced sampling, RIDE loss, and KL-divergence as versatile tools, capable of bolstering the performance of diverse baseline models when confronted with the challenges of distinct datasets in similar tasks.

Acknowledgement

This work was supported by the NRF grant [2012R1A2C3010887] and the MSIT/IITP ([1711117093], [2021-0-00077], [No. 2021-0-01343, Artificial Intelligence Graduate School Program(SNU)]).

References

[1] Benjamin Bergner, Christoph Lippert, and Aravindh Mahendran. Iterative patch selection for high-resolution image recognition. *arXiv preprint arXiv:2210.13007*, 2022.

[2] Jiarui Cai, Yizhou Wang, and Jenq-Neng Hwang. Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 112–121, 2021.

[3] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

[4] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority oversampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[5] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5177–5186, 2019.

[6] Peng Chu, Xiao Bian, Shaopeng Liu, and Haibin Ling. Feature space augmentation for long-tailed data. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 694–710. Springer, 2020.

[7] Jiequan Cui, Shu Liu, Zhuotao Tian, Zhisheng Zhong, and Jiaya Jia. Reslt: Residual learning for long-tailed recognition. *IEEE transactions on pattern analysis and machine intelligence*, 45(3):3695–3706, 2022.

[8] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019.

[9] Alakh Desai, Tz-Ying Wu, Subarna Tripathi, and Nuno Vasconcelos. Learning of visual relations: The devil is in the tails. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15404–15413, 2021.

[10] Andrew Estabrooks, Taeho Jo, and Nathalie Japkowicz. A multiple resampling method for learning from imbalanced data sets. *Computational intelligence*, 20(1):18–36, 2004.

[11] Qingji Guan, Zhuangzhuang Li, Jiayu Zhang, Yaping Huang, and Yao Zhao. Joint representation and classifier learning for long-tailed image classification. *Image and Vision Computing*, page 104759, 2023.

[12] Sebastian Guendel, Sasa Grbic, Bogdan Georgescu, Siqi Liu, Andreas Maier, and Dorin Comaniciu. Learning to recognize abnormalities in chest x-rays with location-aware dense networks. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 23rd Iberoamerican Congress, CIARP 2018, Madrid, Spain, November 19-22, 2018, Proceedings 23*, pages 757–765. Springer, 2019.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Dat Q. Tran Dat T. Ngo Ha Q. Nguyen Hieu H. Pham, Tung T. Le. Interpreting chest x-rays via cnns that exploit hierarchical disease dependencies and uncertainty labels. 2019.

[15] Gregory Holste, Song Wang, Ziyu Jiang, Thomas C Shen, George Shih, Ronald M Summers, Yifan Peng, and Zhangyang Wang. Long-tailed classification of thorax diseases on chest x-ray: A new benchmark study. In *MICCAI Workshop on Data Augmentation, Labelling, and Imperfections*, pages 22–32. Springer, 2022.

[16] M. Ko Y. Yu S. Ciurea-Ilcus C. Chute H. Marklund B. Haghgoo R. L. Ball K. Shpanskaya J. Seekins D. A. Mong S. S. Halabi J. K. Sandberg R. Jones D. B. Larson C. P. Langlotz B. N. Patel M. P. Lungren A. Y. Ng J. Irvin, P. Rajpurkar. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *AAAI*, 2019.

[17] Alistair Johnson, Matt Lungren, Yifan Peng, Zhiyong Lu, Roger Mark, Seth Berkowitz, and Steven Horng. Mimic-cxr-jpg-chest radiographs with structured labels. *PhysioNet*, 2019.

[18] Pollard Tom J. Berkowitz Seth J.-Greenbaum Nathaniel R. Lungren Matthew P. Deng Chih-ying Mark Roger G. Horng Steven Johnson, Alistair E. W. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. pages 2052–4463, 2019.

[19] Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations*, 2020.

[20] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019.

[21] Jaehyung Kim, Jongheon Jeong, and Jinwoo Shin. M2m: Imbalanced classification via major-to-minor translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13896–13905, 2020.

[22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[23] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550, 2008.

[24] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[25] Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1713–1721, 2015.

[26] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems*, 33:4175–4186, 2020.

[27] Sina Taslimi, Soroush Taslimi, Nima Fathi, Mohammadreza Salehi, and Mohammad Hossein Rohban. Swinchex: Multi-label classification on chest x-ray images with transformers. *arXiv preprint arXiv:2206.04246*, 2022.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[29] Tao Wang, Yu Li, Bingyi Kang, Junnan Li, Junhao Liew, Sheng Tang, Steven Hoi, and Jiashi Feng. The devil is in classification: A simple framework for long-tail instance segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 728–744. Springer, 2020.

[30] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X Yu. Long-tailed recognition by routing diverse distribution-aware experts. *arXiv preprint arXiv:2010.01809*, 2020.

[31] 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*, pages 2097–2106, 2017.

[32] Wenwu Ye, Jin Yao, Hui Xue, and Yi Li. Weakly supervised lesion localization with probabilistic-cam pooling. *arXiv preprint arXiv:2005.14480*, 2020.

[33] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2361–2370, 2021.

[34] Si Zhang, Hanghang Tong, Jiejun Xu, and Ross Maciejewski. Graph convolutional networks: a comprehensive review. *Computational Social Networks*, 6(1):1–23, 2019.

[35] Yifan Zhang, Bryan Hooi, Lanqing Hong, and Jiashi Feng. Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition. *Advances in Neural Information Processing Systems*, 35:34077–34090, 2022.

[36] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[37] Zizhao Zhang and Tomas Pfister. Learning fast sample re-weighting without reward data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 725–734, 2021.

[38] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16489–16498, 2021.