Multi-label long-tailed disease recognition on Chest X-ray images

Graduation Thesis Final Report

Dang Minh Anh

Nguyen Manh Dung

Nguyen The Trung Kien

Supervisor

Associate Professor Phan Duy Hung

Bachelor of Artificial Intelligence

Hoa Lac campus - FPT University

August 2023

Acknowledgement

We would like to thank our instructor, Dr. Phan Duy Hung for his patience and time, and for instructing and advising us enthusiastically.

We would like to thank all at my University, FPT, for giving us the best environment to study and grow over the years.

We would like to thank our classmates in AI1504, for letting us meet amazing people and learn a lot from them.

We always remember our family's encouragement and support. Thanks to them, we have the will, the energy and the confidence to pursue our goals.

Abstract

Machine learning and deep learning recently have many big achievements in computer vision and there is a trend to apply deep learning in diagnostic medical images, for example, Chest-Xray classification. Since most large-scale image classification benchmarks contain single-label images with a mostly balanced distribution of labels, many standard deep learning methods fail to accommodate the class imbalance and co-occurrence problems posed by the long-tailed multi-label nature of tasks like disease diagnosis such as Chest-Xray classification. Compared to conventional single-label classification problem, multi-label recognition is often more challenging due to issues called the dominant of negative samples (when we treat multi-label classification as series of binary classification) and the long tail distribution of positive samples. In this thesis, we modified the orignial binary cross entropy loss to get a new loss function called class-aware balanced loss which can solve two previous problems in ChestXray14 dataset. We train Swin Transformer model on Chest-Xray14[1] dataset with our new loss and archive the best AUC score compared to other SOTA algorithms.

Keywords: Multi-Label, Chest X-ray, Long-tailed distribution, Class-Aware Loss

Table of Contents

List of figures

List of tables

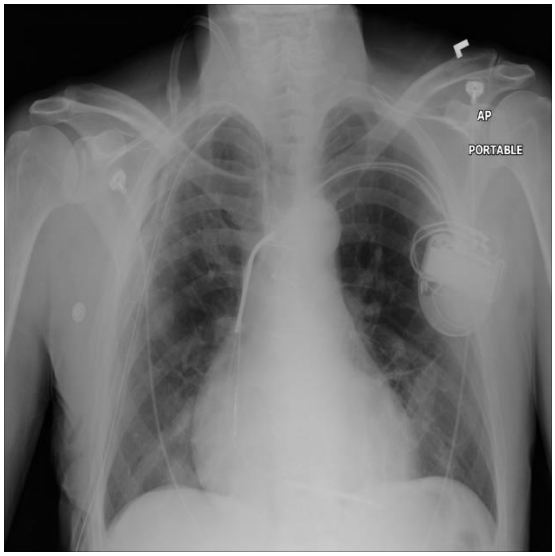
List of abbreviations and acronyms

Abbreviations	Meaning
DL	Deep learning
CNN	Convolutional neural network
ViT	Vision Transformer
ReLU	Rectified Linear Unit
GELU	Gaussian Error Linear Unit
AUC	Area Under the ROC curve
mAP	Mean Average Precision
BCE	Binary cross entropy
FL	Focal loss
ASL	Asymmetric loss

1. Introduction

1.1. Problem & Motivation

Improving the speed and accuracy of clinical diagnostics in the medical field has always been a matter of concern at all the times, now due to the limited number of professionally trained staff and the cost that the diagnosis is made. guess there are many disadvantages. In recent years, with remarkable development and being applied in many fields, deep learning (DL) has emerged as a powerful tool in medical image diagnosis. With recent advances in representation learning, DL can automatically learn and extract meaningful features from large datasets, making them particularly well-suited to analyzing complex medical images and has shown promising results when applied to a wide range of medical image modalities. Chest Xray image classificatioin using deep learning is one of the most concern problem now, however, unlike traditional image classification tasks, where the image is single-label and label distribution is relatively balanced, chest X-ray classification in real-world applications is a multi-label classification task which mean an image can have multiple labels due to the fact that a patient can have several pathologies at a time and there are also correlation between pathologies (Figure 1 show an example of chest Xray disease recognition as multi-label classification problem). To model the correlation of different labels is a challenge in deep learning.



Atelectasis
Cardiomegaly
Effusion
Infiltration
Mass
Nodule
Pneumonia
Pneumothorax
Consolidation
Edema
Emphysema
Fibrosis
Pleural Thick
Hernia

Figure 1: Multi label chest Xray classification (red is positive label, other is negative)

Another challenge in medical image classification is the negative dominant issue (illustrate in Figure 2), positive samples always appear with very small frequency compared to negative samples, thus leading to the imbalance between negative and

positive samples within a class. In multi-label context, negative dominant issues even become worse if every class is imbalanced.

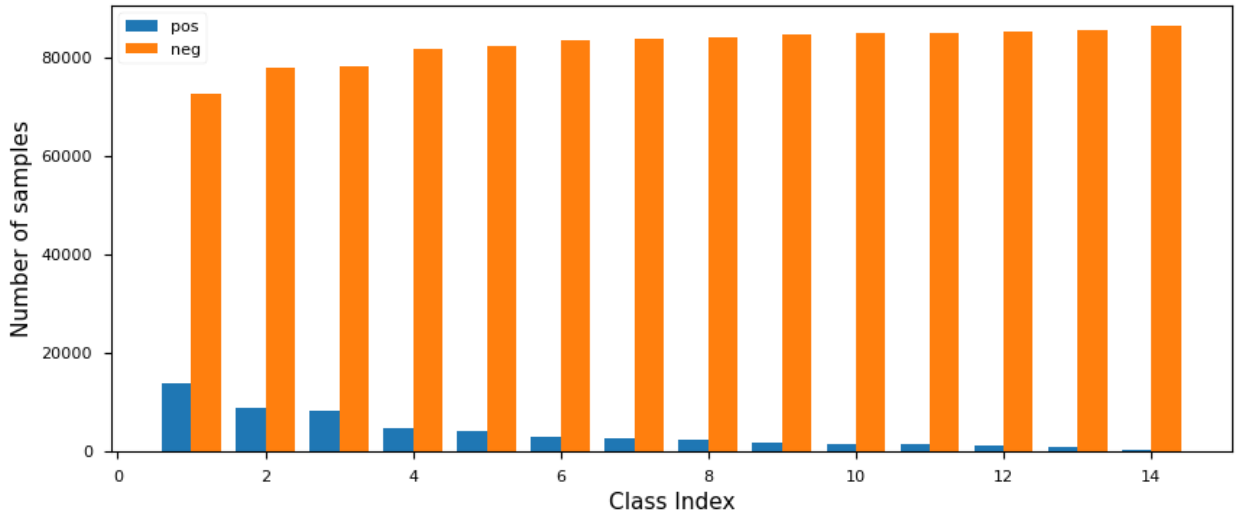


Figure 2: Illustration of negative dominant issue

Besides that, Chest Xray classification also suffers from long tail label distribution (Figure 3), meaning a small portion of classes account for massive training samples. In contrast, other classes have only a few numbers of training data. The imbalanced class distribution eventually makes the model easily biased to head classes with a large portion of training data, leading to poor performance on tail classes as well as the overall performance of the system. This issue is a challenge for applying deep learning in the real world, therefore, a massive number of studies have been conducted in recent years.

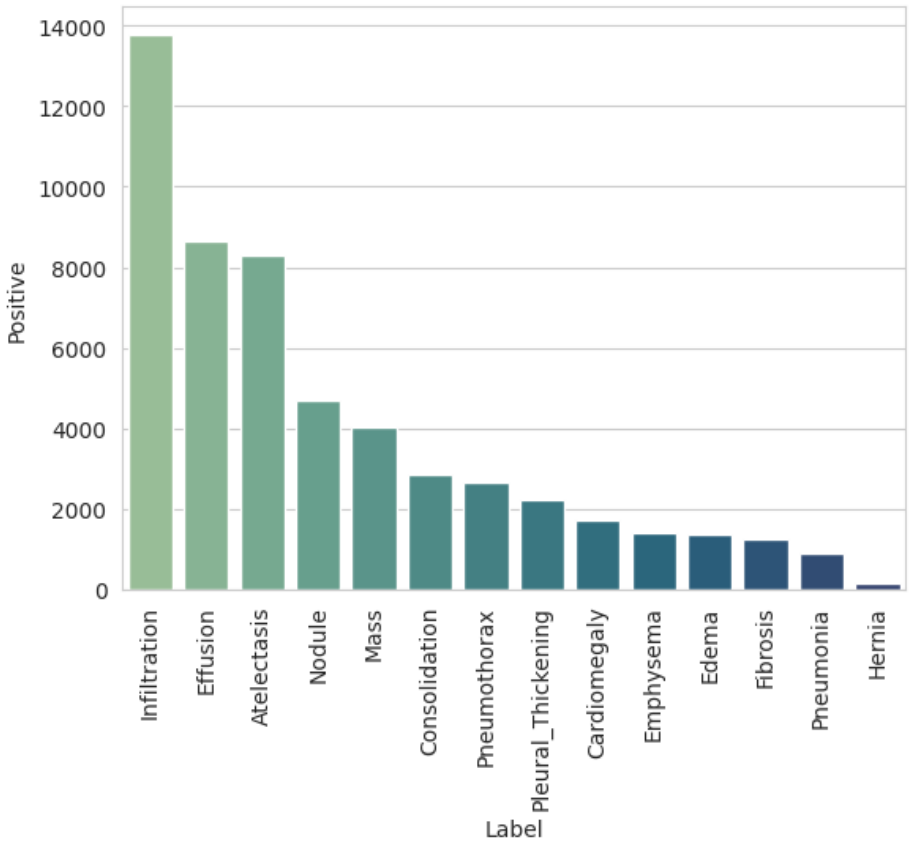


Figure 3: Long-tail distribution demonstration on the Chest Xray14 dataset

1.2. Related Works

1.2.1. Multi-label Chest X-ray

Because of the large amount of chest radiography, many researchers have done a lot of study to apply deep learning to identify thorax disorders on chest radiograph. For instance, Bar et al. [2] colleagues investigated the ability of DCNN models to distinguish between various pathologies and used these models to categorize eight thoracic disorders on a limited dataset of chest x-rays. Majdi et al. [3] suggested a fine-tuned DenseNet-121 to categorize CXR pictures into lung nodules and cardiomegaly disorders. The experiment made use of images from the CheXpert dataset. This model has outstanding results in detecting lung nodules and enlarged heart. Cicero et al. [4] classified frontal chest radiograph pictures into the following categories using the GoogleNet model: normal, consolidation, cardiomegaly, pulmonary edema, pneumothorax, and pleural effusion. The study demonstrated that the DCNN model may perform well even when trained on a small medical dataset. Rasheed et al. [5] used a logistic regression classifier with CXR images to investigate the utility of Machine Learning for the diagnosis of COVID-19. To accelerate learning and choose the characteristics that would produce the

best potential accuracy (ACC), they thought of using a dimensionality reduction strategy. Wang et al. [1] used multiple multi-label DCNN losses and various pooling algorithms to provide a unified weakly supervised multi-label classification framework. Allaouzi and Ahmed [6] conducted experiments using the ChestX-ray14 and CheXpert datasets. Using transfer learning, they extract the features from the pre-trained DenseNet-121. Additionally, a variety of issue transformation approaches are used, including classifier chains, binary relevance, and label powersets. Ait Nasser et al. [7] used ensemble learning to categorize CXR pictures into three categories (normal, lung illness, and heart disease). Data-augmentation techniques were used to enhance the number of samples and prevent overfitting. When combined with data-augmentation approaches, the suggested ensemble learning methodology performed well. Yao et al. [8] employed DenseNet and LSTM to extract features and exploit the statistical label dependencies, respectively, and thus achieved improved diagnosis. Grewal et al. [9] presented a cascaded deep neural network, as well as modeling label dependencies, and examined the selection of loss functions in training as well as the efficiency of cascading. Rajpurkar et al. [10] colleagues built a deep model that combines both dense connections and batch normalization, and it outperformed expert radiologists in identifying pneumonia. Kim et al. [11] classified CXR pictures into three groups (normal, pneumonia, and pneumothorax) using EfficientNet-V2M with transfer learning as an end-to-end technique. Before generating the feeding data to the used model, preprocessing procedures were applied to the pictures from the ChestX-ray14 dataset. Blais and Akhloufi [12] used various models with binary relevance to diagnose chest illnesses in the CheXpert dataset. When combined with Adam optimizer, the Xception DCNN model outperformed other models. Li et al. [13] suggested a unified technique that performs simultaneous illness detection and pathological pattern localization using multi-instance learning and a minimal number of bounding boxes of disease patterns. Guendel et al. [14] introduced the DNetLoc model to increase classification accuracy, which uses high-resolution data and includes pathological pattern spatial information into the classification technique. This method has so far acquired the best average AUC score after being trained on both the PLCO and ChestX-Ray14 datasets. Guan et al. [15] and Wang et al. [16] introduced attention guided solutions exploiting the mutual relationship between labels and the locations of diseases.

1.2.2. Long tail issue.

In many real-world applications, especially in medical image diagnostics, datasets usually follow long-tail distributions, where the number of samples per class varies with a large imbalance factor. This terrible issue limits the applicability of visual recognition in the

medical field. To address this problem, many studies have been conducted in recent years.

The first attempt to re-balance label distribution is the re-sampling [17,18,19] technique. Re-sampling re-balanced classes by adjusting the number of samples per class. Many sampling strategies have been developed until now. Re-sampling is the most widely used method [20] in processing long-tailed distribution image classification in depth learning, mainly including over-sampling [21], under-sampling [22] and mixed sampling [23]. By adding more samples from the tail class, the oversampling approach primarily balances out the disparity between the head class and the tail class. In response to this, Gupta et al. devised the repeated factor sampling approach [24], which adjusts the training data's balance by raising the tail image's sample frequency. To address the issue of class imbalance, Peng et al. [25] presented the soft box sampling approach, which use class perception sampling to determine the replication factor for each picture based on the distribution of labels and replicates the images in accordance with the predetermined number of times. Mixed sampling [23] is a method of achieving sample balance by mixing oversampling and under-sampling. In 2020, Ding et al. [26] proposed a KA integration method of under-sampling and oversampling, under-sampling the majority of classes using the kernel-based adaptive synthesis method and over-sampling the minority classes at the same time, generating a set of balanced datasets to train the corresponding classifiers separately. All these trained classifiers will then vote on the results. Random-balanced sampling includes over sampling and under sampling. Over sampling repeats the samples from minority classes to balance class distribution before training. Under sampling is in the opposite manner, it reduces the number of majority class. Besides random strategy, other sampling methods such as square-root sampling [27] and progressively-balanced sampling [28] have also been developed. However, in multi label context, re-sampling method seem not really balance classes distribution due to label co-occurrence.

Re-weighting attempts to adjust the training loss values for different classes by multiplying them with different weights. The most intuitive method is to directly use label frequencies of training samples for loss. However, this approach may not work well for long-tail issues. Class-Balanced loss (CB) [29] introduced a novel concept of *effective number* to approximate the expected sample number of different classes, an exponential function of their training label number. Following this, CB loss enforces a class-balanced re-weighting term, inversely proportional to the effective number of classes. Besides class-driven re-weight, Lin et al. [30] propose an instance-driven loss called Focal loss, which improves cross-entropy loss by down-weight losses assigned to well-classes examples. So, it can assign higher weights to the harder tail classes but lower weights to

the easier head classes. Hermans et al. [31] suggested the triplet loss function and employed gradient descent to train samples with subtle variations in 2017. Cao et al. [32] presented label-distribution-aware margin loss (LDAM) in 2019, in which the model learns the original feature representation before re-weighting. A novel tolerance regularization approach to mitigate gradient over suppression was developed in 2020 to alleviate distribution balance loss [33]. In the year 2021, an enhancement to equalization loss, known as equalization loss v2 [34], was developed. This version introduced a novel method of adjusting gradient weights, which involves boosting the importance of positive gradients while diminishing the significance of negative gradients during model training for individual subtasks. The seesaw loss [35] method readjusts the gradients of different classes by employing factors that mitigate and compensate for imbalances. In the case of LADE [36], a loss related to label distribution decoupling is introduced. This serves to separate the learning model from the skewed distribution seen during training, allowing the model to be fine-tuned for varying test class distributions based on available test label frequencies.

Inspired by the intuition that regardless of re-sampling or re-weighting, the final result is making the change in model parameters, Cui et al [37] propose a new idea of re-balancing directly in parameter space. Decoupling [30] introduces decoupled learning scheme that separately trains the representation module and classifier. Experimental results show that imbalance issues might not affect representation learning, so we can achieve a robust recognition model by only fine-tuning the classifier.

1.3 Contribution

In this thesis, we focus on solving the negative impact of long-tail imbalanced data on Chest X-ray image classification tasks. We will try several methods to find the best approach robust to class imbalance on Chest X-ray images. Our contribution is that we modified the original loss function to get a new loss function that adapt with long tail distribution in ChestXray14 dataset and get higher performance compared to other previous works.

2. Methodology

Image classification is a supervised learning problem: define a set of target classes (objects to identify in images) and train a model to recognize them using labeled example photos. Early computer vision models relied on raw pixel data as the input to the model. However, raw pixel data alone doesn't provide a sufficiently stable representation to encompass the myriad variations of an object as captured in an image. The position of the object, background behind the object, ambient lighting, camera angle, and camera focus all can produce fluctuation in raw pixel data; these differences are significant enough that they cannot be corrected for by taking weighted averages of pixel RGB values. To model objects more flexibly, classic computer vision models added new features derived from pixel data, such as color histograms, textures, and shapes. The downside of this approach was that feature engineering became a real burden, as there were so many inputs to tweak.

A breakthrough in building models for image classification came with the discovery that a convolutional neural network (CNN) could be used to progressively extract higher- and higher-level representations of the image content. Instead of preprocessing the data to derive features like textures and shapes, a CNN takes just the image's raw pixel data as input and "learns" how to extract these features, and ultimately infer what object they constitute. Various types of CNN architecture have been proposed to extract better features and improve the task’s performance. CNN was SOTA architecture in vision task until 2020, when a mechanism called ‘attention’ was applied to vision task by the introduction of vision transformer models. These transformers have become one of the best backbones used in image classification now.

2.1. Overview pipeline

In this thesis, we using base version of Swin Transformer as feature extractor. Model can be divided into two parts include feature extractor and heads of classification. Input images go through Swin Transformer to extract image embedding, then these embeddings continue go through a fully connected network as in normal image classification framework to get predicted probability. Pipeline as in Figure 4:

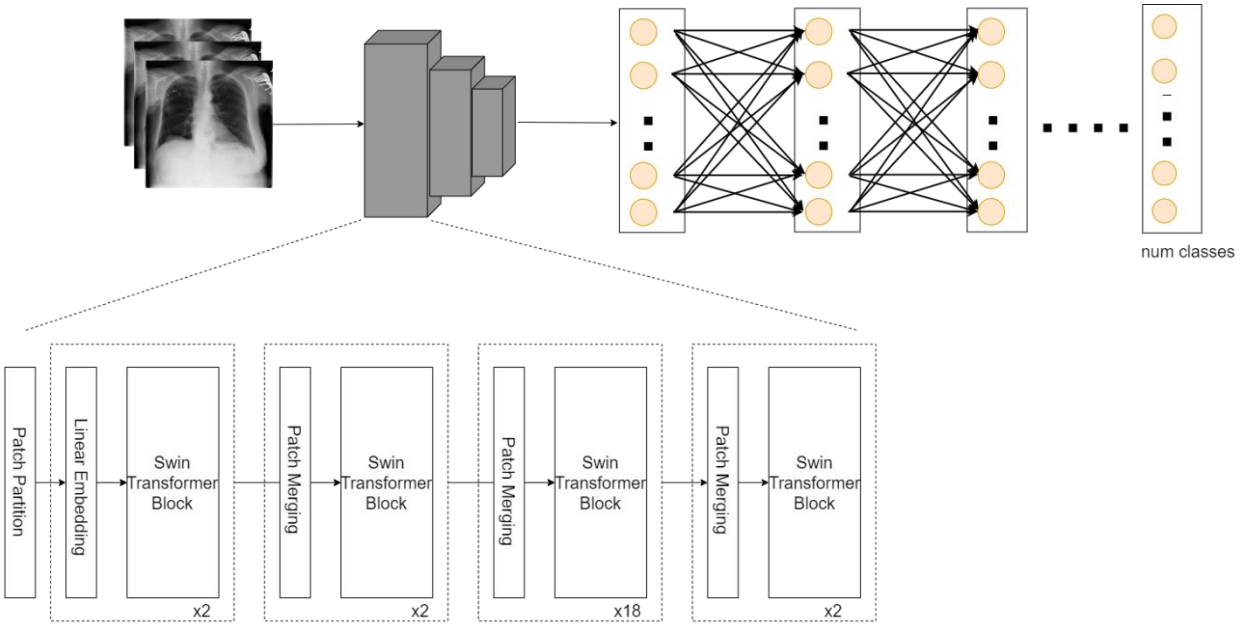


Figure 4: Pipeline of model

For training the network, we use decoupling strategy which means feature extractor and classifier are trained separately. First, for training the feature extractor, we run model util model become converges. After that, we freeze the features extractor and then train the head of classifier. These strategy is proved to improve model performance in [20].

2.2. Swin Transformer

2.2.1 Transformer Architecture

The transformer is a powerful architecture and was first introduced in [38] in the domain of natural language processing by Google Brain. Since its first appearance, researchers have developed a variety of variants that have achieved SOTA. Architecture of transformer that introduced in [38] is in Figure 5.

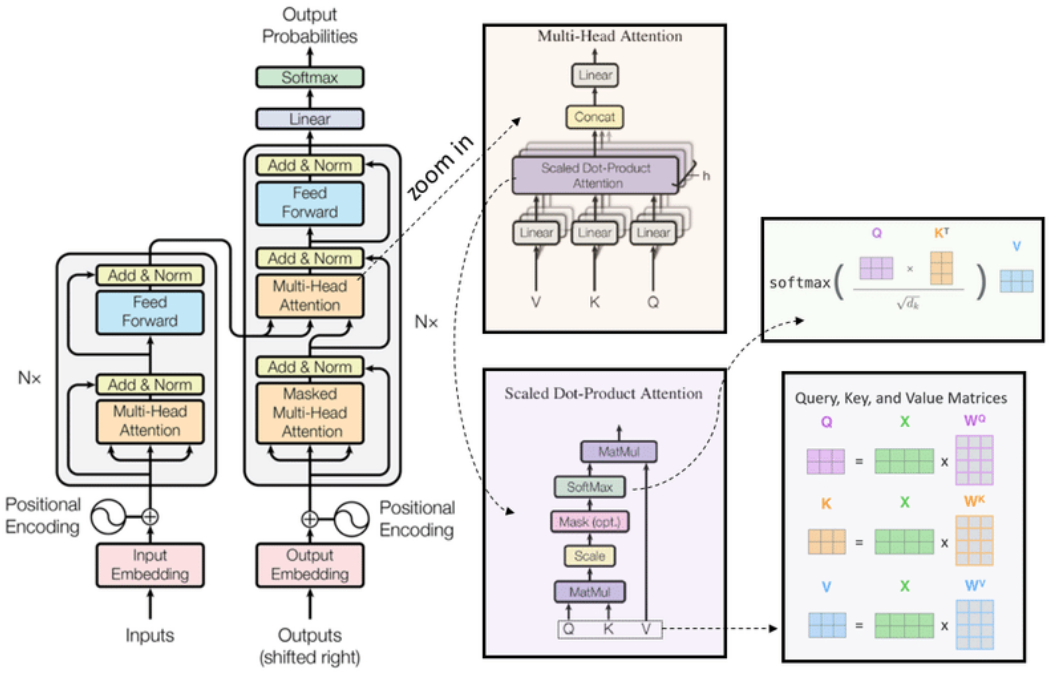


Figure 5: Transformer architecture

Position and order of words are the essential parts of any language. They define the grammar and thus the actual semantics of a sentence. But the Transformer architecture, each word in a sentence simultaneously flows through the Transformer's stack, The model itself doesn't have any sense of position or order for each word. Consequently, there is still the need for a way to incorporate the order of the words into our model.

One possible solution to give the model some sense of order is to add a piece of information to each word about its position in the sentence which is called positional encoding. The positional encoding mechanism used in the original paper is as in Figure 6.

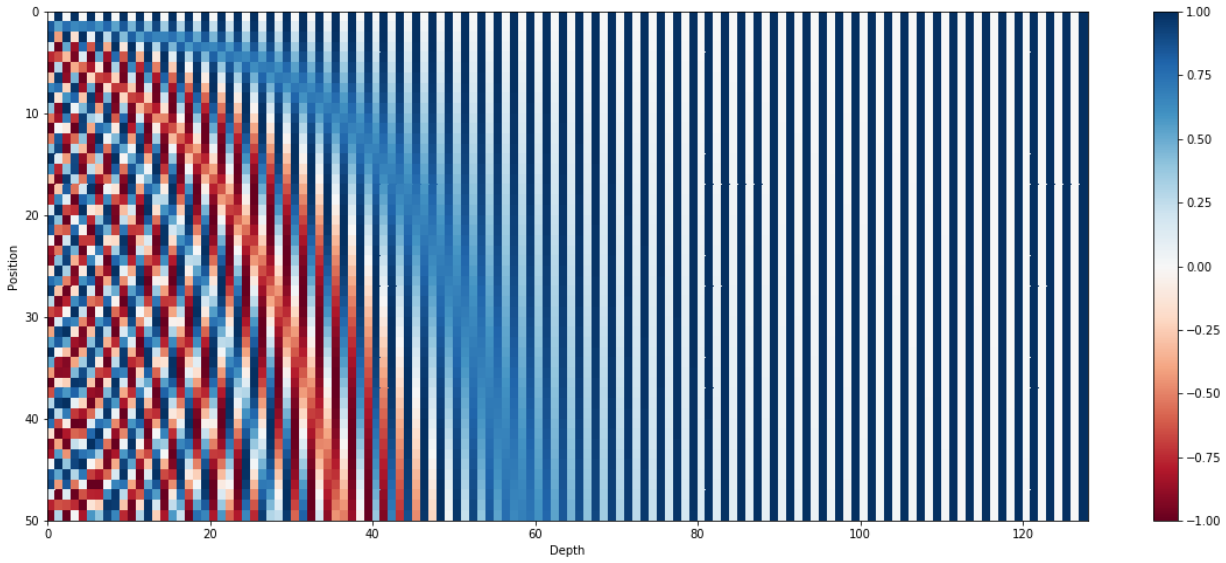


Figure 6: Example of 128-dimentional positional encoding for a sentence with max length of 50.

2.2.2 Swin Transformer

With the breakthrough success of transformer in the field of natural language processing, researchers have continued to find ways to apply this mechanism to many other fields such as computer vision, signal processing, and so on. The ViT [39] model introduced Transformers to computer vision. The standard transformer receives an 1D token embeddings as input. To apply transformer to 3D images, an image is split into fixed size of patches, then transform each patch by linear projection after flattening and then add positional embedding to create input vector. ViT architecture overview is shown in Figure 7. There is a slight modification in the basic transformer block structure of ViT with layer norm is the first layer.

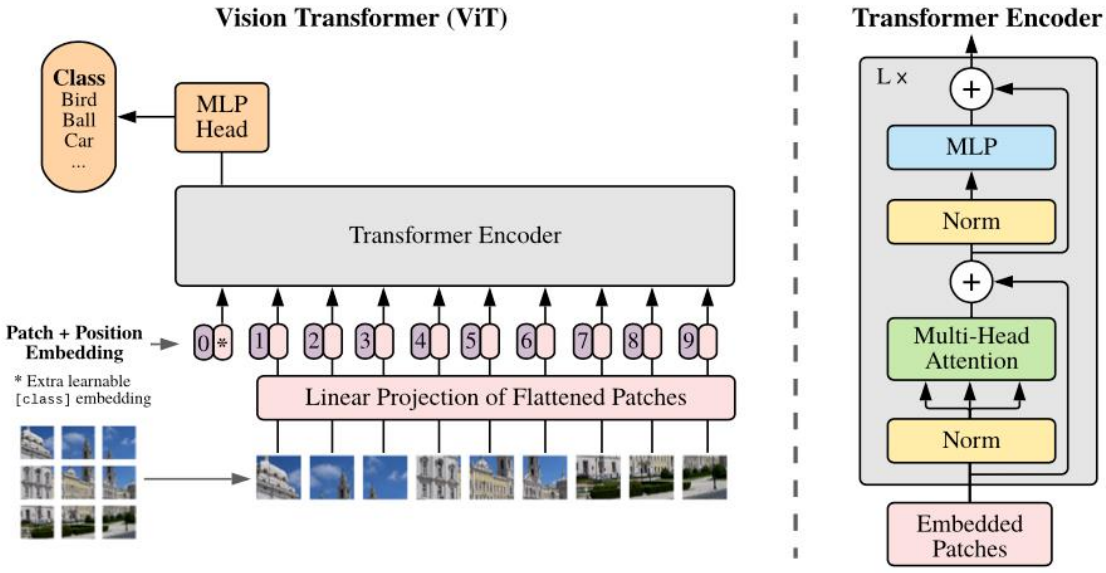


Figure 7: ViT overview [39]

ViT shows its excellent performance in image classification compared to SOTA CNN model with the same amount of data. However, because ViT calculates self-attention for all patches in the entire feature layer, it is not efficient when dealing with high-resolution images.

Liu et al [40] proposed an improved version of vision transformer called Swin Transformer that can solve weakness of ViT when calculated high-resolution images. Swin is an acronym that stands for Shifted window (illustrated in Figure 8b). This shifted window concept is not new to the research community. It has been used in CNNs for many years. It is one of the CNN features that has made it excel in the computer vision realm as it brought about great efficiency. However, it had not been used in Transformers before and Swin Transformer is the first transformer model that applied shifted window mechanism.

Swin Transformer still uses patches as in the ViT model. However, instead of performing global self-attention for all patches of images as in previous work, Swin Transformer uses a local window and performs attention for patches within the window. Shifted windows reduce computational complexity compared to ViT and thus, can be applied to high-resolution images.

Besides, instead of using fixed size of patches as in ViT ($16 \times 16px$), the Swin Transformer first starts with small patches ($4px \times 4px$) in the first layer merges into bigger ones in the deeper layers (as shown in Figure 8a). Patches merging help to gradually integrate information of patches that without local window at early layer. As the model gets deeper patches size is bigger, thus attention is performed on bigger pieces of images. Swin Transformer now serve as the most popular backbone for both image classification and dense recognition tasks.

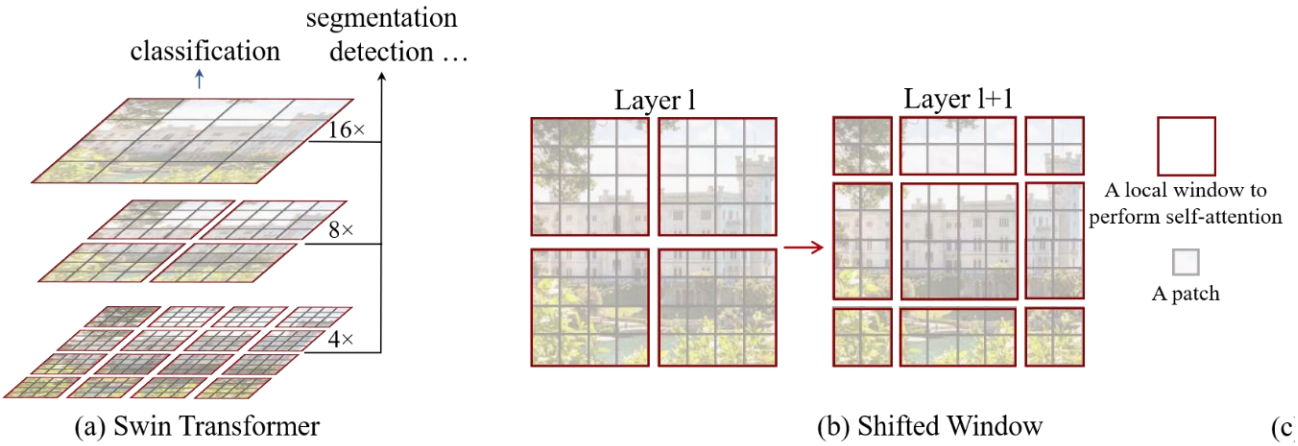


Figure 8: Patches merging(a) and shifted window(b) in Swin Transformer architecture [39]

2.2.3. Activation function

Rectified Linear Unit (ReLU) [41] activation is commonly used in feed-forward neural networks. The definition of ReLU is:

$$ReLU(x) = \max(0, x) \quad (1)$$

The ReLU function has several advantages over other activation functions. It is simple to compute, requiring only a single comparison operation, which makes it faster to evaluate than other activation functions. Additionally, the ReLU function does not suffer from the vanishing gradient problem, which can make it easier to train deep networks. Because of these advantages, the ReLU function has become one of the most widely used activation functions in neural networks.

One weakness of ReLU is that it can produce output values that are either 0 or positive, but never negative. This can make it difficult for the network to model data with negative values. Additionally, the ReLU function can suffer from the so-called "dying ReLU" problem, where some of the neurons in the network can become "dead" and stop producing any output. This can happen when the neurons always receive negative input and are therefore always outputting 0, which can make it difficult for the network to learn.

Gaussian Error Linear Unit (GELU) [22] combines the effect of ReLU, zone out, and dropout. One of ReLU's limitations is that it's non-differentiable at zero - GELU resolves this issue, and routinely yields a higher test accuracy than other activation functions. GELU is now quite popular and is the activation function used in many vision and language models, Swin Transformer architecture use GELU as activation function. Figure 9 shows the comparison of GELU others activation function.

$$GELU(x) = xP(X \leq x) = x\varphi(x) \tag{2}$$
$$\cong 0.5x(1 + \tanh\left[\sqrt{\frac{2}{\pi}}(x + 0.044715x^3)\right]$$

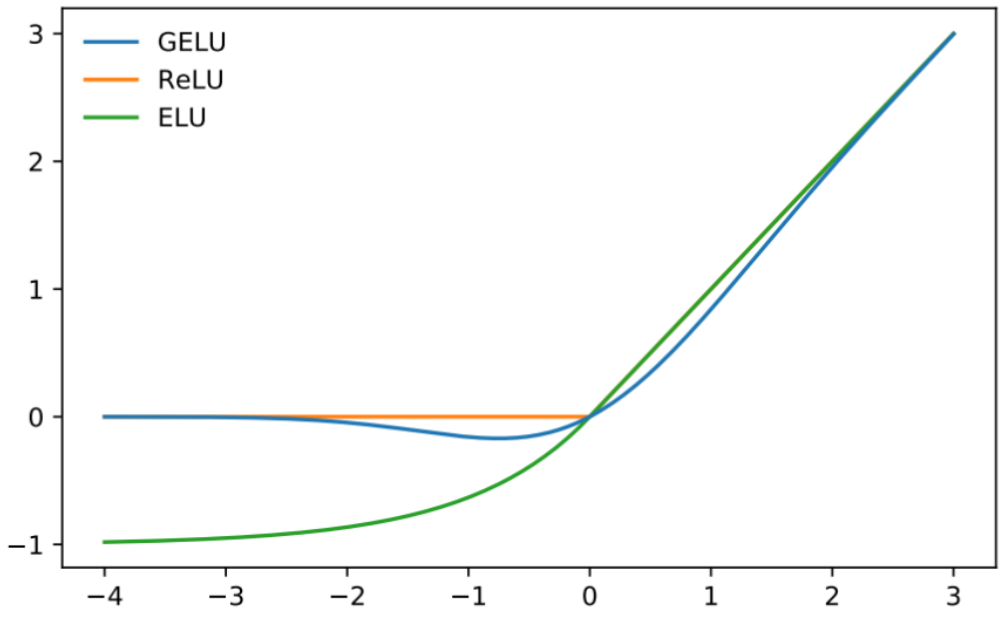


Figure 9: Illustration of GELU compared to ReLU and ELU [42]

2.3. Class-Aware Loss

The most popular way to consider multi-label classification is a series of binary classifications, so binary cross entropy is usually used as loss function.

$$BCE = -yL_+ - (1-y)L_- \tag{3}$$

Where $\begin{cases} L_+ = \log(p) \\ L_- = \log(1-p) \end{cases}$ and L_+, L_- are respectively positive and negative loss parts.

And the total classification loss is sum of binary loss from C labels.

$$L_{total} = \sum_{i=1}^{C} L(p_i, y_i) \tag{4}$$

Where p is the prediction probability of the model, y is the ground truth of a sample. BCE loss optimizers each label independently and does not consider the dependent co-occurrences of labels in each sample. Besides that, BCE loss does not consider the issue of imbalanced data distribution. Furthermore, this loss is symmetric, therefore, negative labels and positive labels will be treated equally and will lead to over-suppression on the negative sides [43], in other words, the model tend to predict all sample as negative, thus decreasing the recall of the models. One possible solution is multiplying loss with its class frequency, but it seems inefficient in case of long-tail distribution.

Another way to solve imbalance is focusing on positive labels, Lin et al [30] introduce focal loss that weighted according to its prediction probability. The definition of focal loss is:

$$\begin{cases} L_+ = (1-p)^{\gamma}\log(p) \\ L_- = p^{\gamma}\log(1-p) \end{cases} \tag{5}$$

Focal loss reweights loss of an example according to its predicted probability, thus focus more on hard sample and down weight loss of easy sample, however, this loss also symmetric and in negative dominant context, despite of the small value loss for negative samples, their large number of negative samples still lead the total contribution of negative sample dominate total loss.

Asymmetric loss (ASL) [44] is an improvement of focal loss that introduces two separate gamma coefficients: γ_-, γ_+

$$\begin{cases} L_+ = (1-p)^{\gamma_+}\log(p) \\ L_- = p^{\gamma_-}\log(1-p) \end{cases} \tag{6}$$

Different gamma coefficient can treat the positive class and negative class differently due to the nature of imbalanced data, thus they set $\gamma_- > \gamma_+$ to account for more contribution of positive labels. However, this loss does not consider the imbalance between positive labels of different classes.

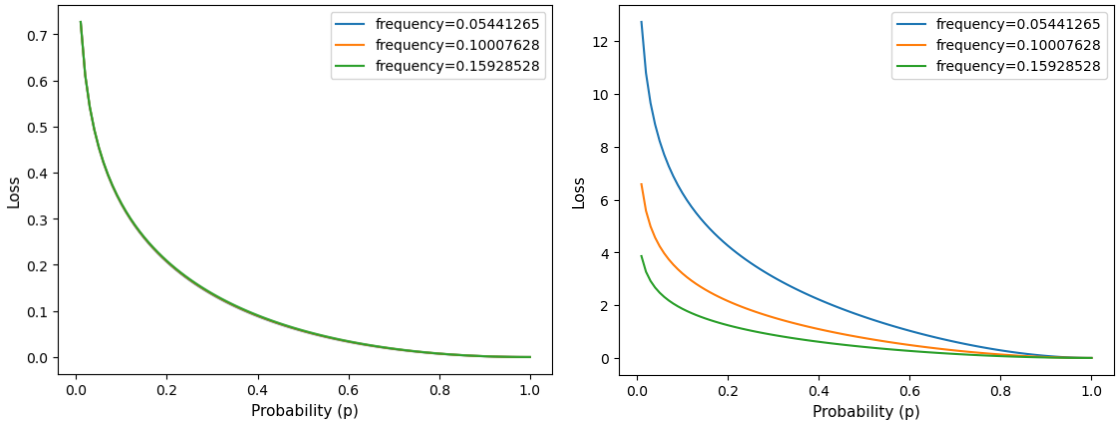


Figure 10: Negative (left) and positive (right) loss of different class.

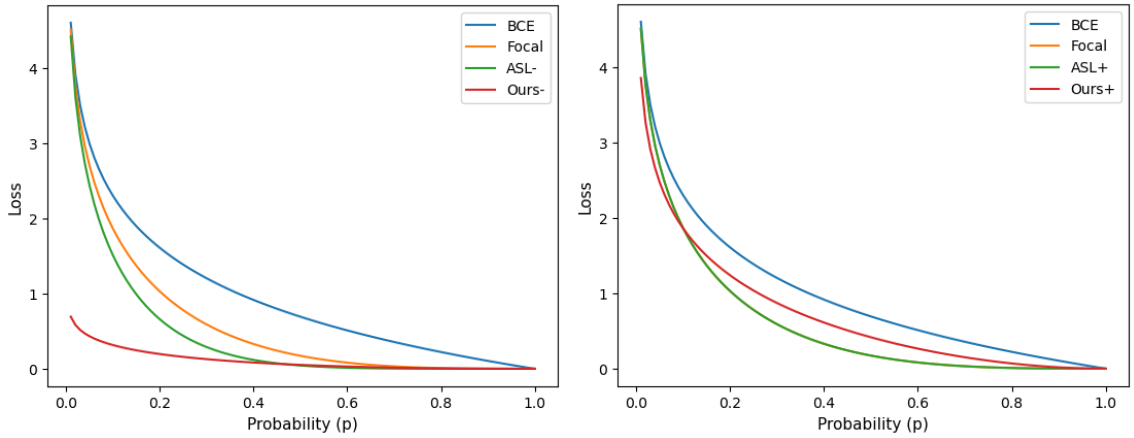


Figure 11: Negative (left) and positive (right) loss of different *loss functions.*

In the long-tail distribution, serious imbalance of positive sample of head classes compared to tail class lead to the dominate of positive samples of head classes, thus reduce the performance of model on tail classes. Inspired by focal loss and re-weight method that balanced loss contribution by multiply each sample with the probability of its class, we propose a new loss function called Class-Aware loss that weighted sample

according to their skewness between positive and negative and their predicted probability. Loss of a class is calculated as follow:

$$\begin{cases} L_+ = w^+ log(p) \\ L_- = w^- \log(1-p) \end{cases} \quad (7)$$

$$w^+ = (1-\alpha_i{}^{1-p})$$
$$w^- = 1-(1-\alpha_i)^p$$
$$\alpha_i = \frac{P_i}{P_i + N_i}$$

Where P_i, N_i is the number of positive samples and negative samples of class i.

With this definition, positive and negative samples of each class are weighted with different focus level according to its class imbalanced. Finally, total loss is the weighted sum of each class loss, w_i is the weighted factor and base on the ratio of positive samples between different classes. Therefore, positive, and negative labels as well as positive labels of different classes are treated differently.

$$L_{total} = \sum_{i=1}^{C} w_i L(p_i, y_i) \quad (8)$$

Figure 10 shows that the class with more frequent has lower loss value and loss value of positive sample inversely proportional to its class frequency while negative samples have almost same loss value for all class. Beside that, negative and positive loss of the same class are also different but negative loss of different classes are almost the same. This is totally suitable with multi-label long-tailed situation, while positive loss of different class should be different due to the long tailed distribution and negative loss of different class should be the same as the number of negative samples of each class is approximately equal as we observed in Chest Xray14 [1] dataset. As a result, class-aware loss equals the contribution of each label in total loss, thus enhancing the prediction probability for positive sample. Figure 11 show the comparison of our loss function with previous loss function used in multi label classification problem, we can see positive loss of our method as the combination of BCE and FL loss, weight loss by the prediction probability while keeping loss of high confidence samples. Compared to other loss functions, our loss function treats negative samples very differently from positive samples to reduce the impact of negative dominant issues, loss of negative samples is very small.

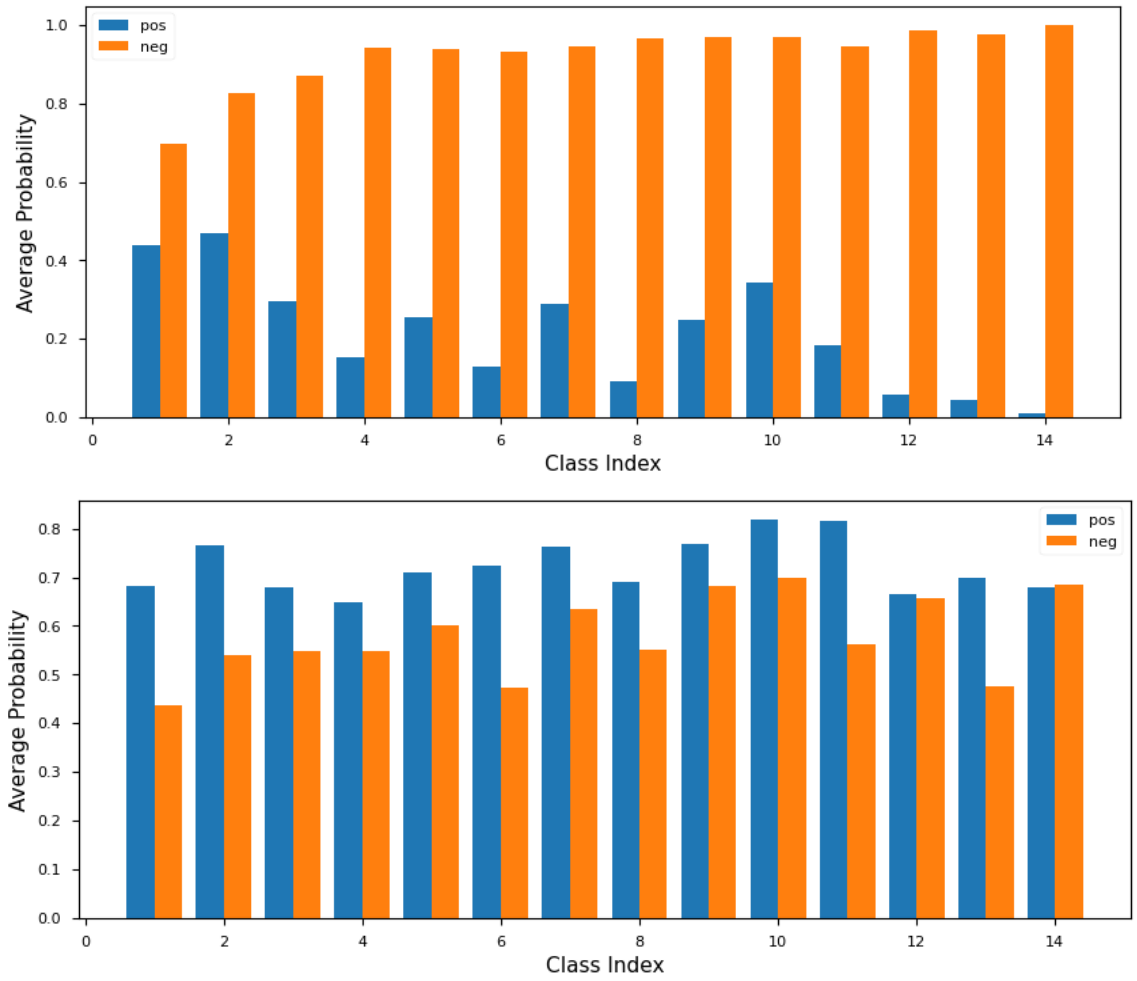


Figure 12: Average prediction probability when using BCE loss(top) and our loss(bottom)

As we can see in Figure 12, the BCE loss lead model focuses more on negative samples, thus prediction is usually low confidence probability and negative average probability is much bigger than positive average probability. While our loss pays attention to both positive and negative samples, negative average probability and positive average probability are approximately equal.

In negative dominant context, model usually prediction with low confident probability, [44] proposed an asymmetric mechanism called probability shifting or negative probability margin that performs hard thresholding of easy negative samples when probability is under a constant value. Shifted probability is define as:

$$p_m = \max(p - m, 0)$$

Where m is negative probability threshold. Combining shifted probability with our loss, Eq (7) become:

$$\begin{cases} L_+ = w^+ log(p) \\ L_- = w^- \log(1 - p_m) \end{cases} \quad (9)$$

3. Experiment results and conclusions

3.1 Data Preparation

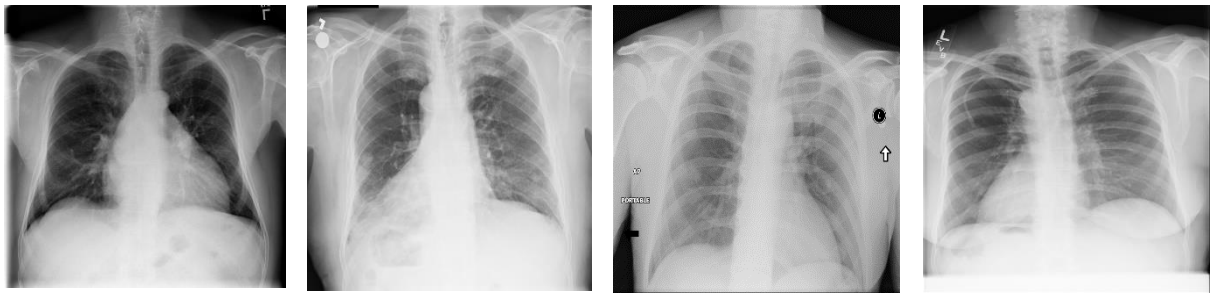


Figure 13: Sample image of Chest Xray14 dataset

For multi-label chest Xray classification, the widely used benchmark Chest-Xray14 [1] is used in the following experiments. Chest-Xray14 has 112,120 frontal X-ray images with disease labels from 30,805 unique patients. Image is grayscale and size of 1024x1024 (example images show in Figure 13). The labels are collected by analyzing radiology reports and are expected to have over 90% accuracy. For a more accurate and objective comparison, we apply the official patient-wise split gathered by Wang et al. [1]. Data distribution is displayed in Figure 14, we can see the long-tailed phenomenon in the positive samples' distribution and the negative dominant phenomenon in negative distribution.

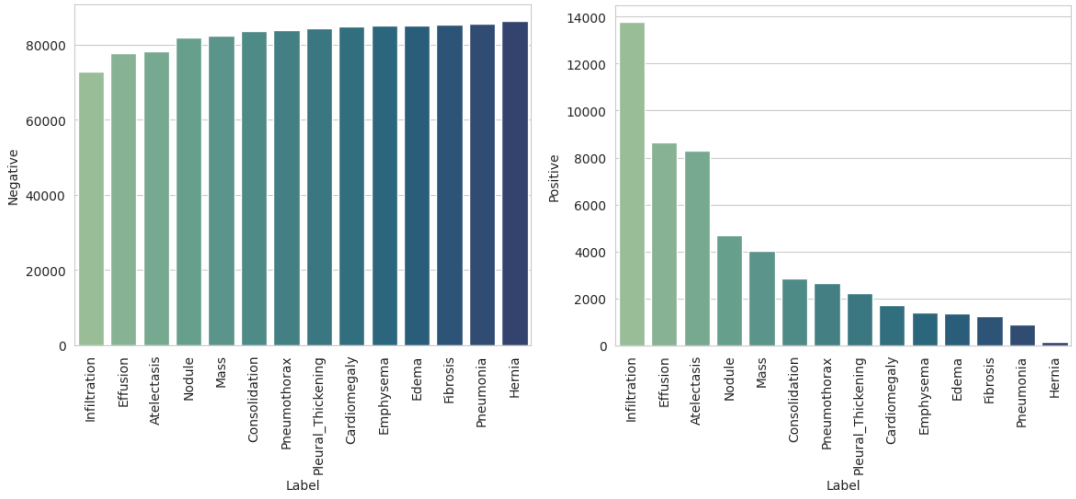


Figure 14: Distribution of negative samples(left) and positive sample(right) of each class.

In multi-label classification, each sample can have one or more positive labels. Figure 15 demonstrates the distribution of the number of positive labels per image. From this distribution, we can see that approximately 70% of samples have no positive labels, the remaining 30% have positive labels include one label and two labels, very rare samples have three labels, almost no image have more than four labels.

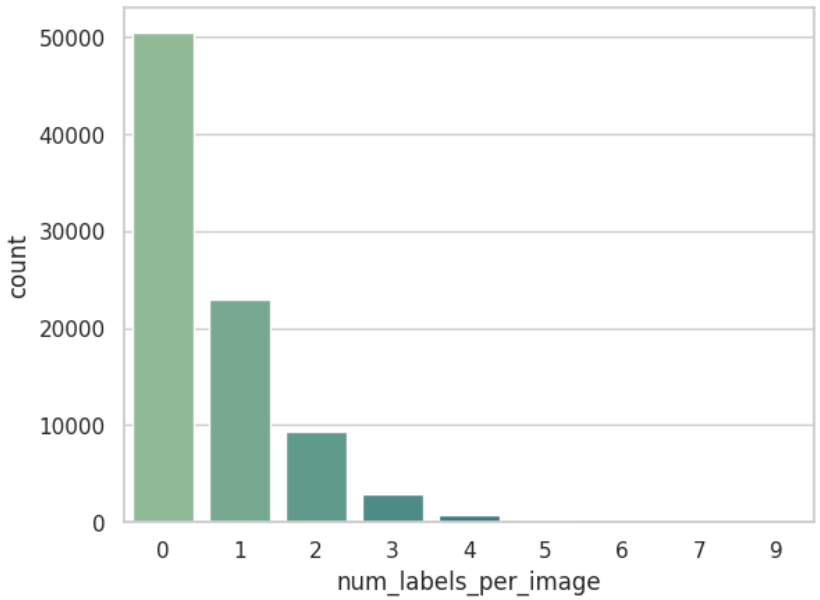


Figure 15: Distribution of number of labels per image

In nature, there are correlations between different diseases, that means one disease can lead to the appearance of other diseases. Figure 16 shows the co-occurrence of 14 classes in the dataset. As we can see, pathology belonging to head classes usually have high co-occurrence to these in tail classes.

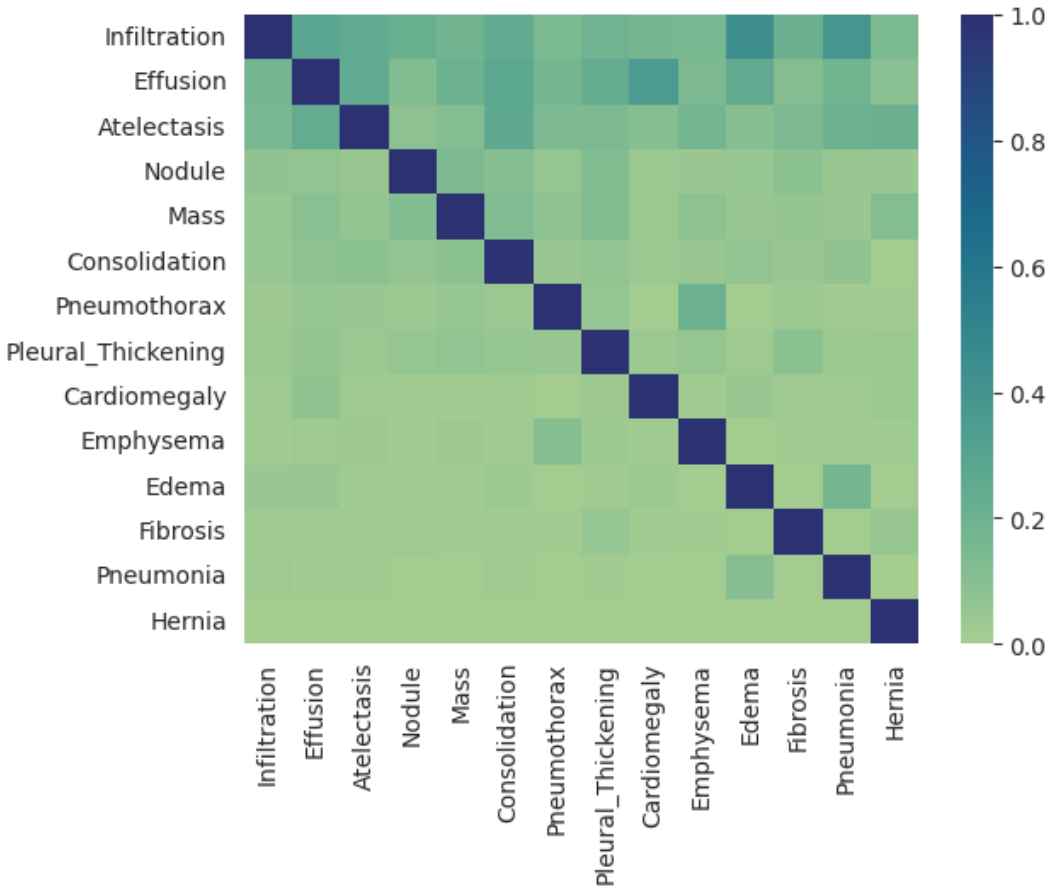


Figure 16: Co-occurrence of labels

3.2 Experiments

3.2.1 Evaluation metric

We use AUC score as the measurement metric which is usually use in Chest Xray classification. To understand AUC, we first need to know about the receiver operating characteristic curve (ROC curve). The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) at different classification thresholds, as shown in Figure 16. TPR can be written as:

$$TPR = \frac{TP}{TP + FN}$$

Where TP, FN are the number of true positives and false negatives.
FPR is:

$$FPR = \frac{FP}{FP + FN}$$

Where FP, TN are the number of false positives and true negatives.

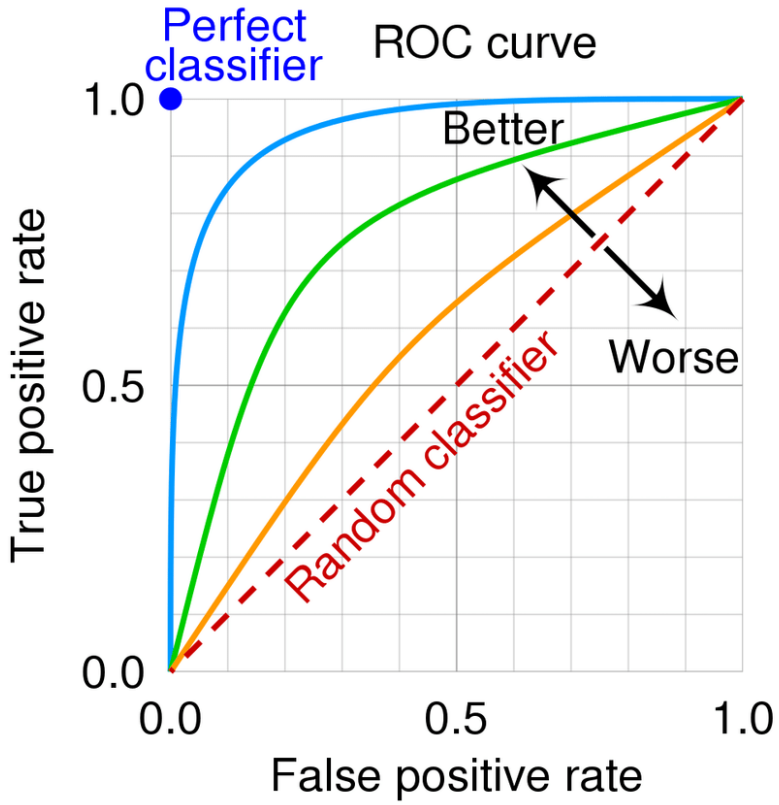


Figure 17: Illustration of ROC curve

AUC is an abbreviation for the area under the ROC curve. AUC measures the likelihood that true positive samples are ranked higher than true negative samples by the magnitude of the area under the ROC curve at all classification thresholds.

Beside AUC, we also use mAP as the second metric to evaluate model performance. The **average precision (AP)** is a way to summarize the precision-recall curve into a single value representing the average of all precisions. The precision-recall curve, commonly plotted on a graph, shows how recall changes for a given precision and vice versa in a computer vision model. A large area under the curve means that a model has both strong recall and precision, whereas a smaller area under the curve means weaker recall or precision. The AP is calculated according to the next equation.

Using a loop that goes through all precisions/recalls, the difference between the current and next recalls is calculated and then multiplied by the current precision. In other words, the AP is weighted average of precision at each threshold, with the different in recall from the preceding threshold serving as the
weight.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$AP = \sum_{n} [Recalls_n - Recalls_{n-1}] * Precision_n$$

Where $Precision_n,\ Recalls_n$ is the respective precision and recall at threshold index n. This value is equivalent to the area under the precision-recall curve (AUPRC).

Mean Average Precision (mAP) for muti-label classification is the mean of APs for all classes in the mAP.

$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k$$

$$AP_k: the\ AP\ of\ class\ k$$

$$n = Number\ of\ classes$$

The mAP incorporates the trade-off between precision and recall and considers both false positive and false negatives. This property makes mAP a suitable metric for most classification and detection applications.

3.2.2 Experiment result

We use Swin Transformer pretrained weights on ImageNet as initial weights and fine-tune on Chest Xray14 dataset. Data augmentation for training includes resizing the original image from 1024x1024 to 256x256 then center crop with size of 224, horizontal flip and random rotation. Dataset is separated as follows: 70% for training, 10% for validation and 20% for testing, test set is official patient-wise split as in [1]. We use a 2-step scheduler and Adam optimizer. The model was implemented using PyTorch and train on A6000 GPUs. Hyperparameters set as in table 1.

Table 1: The parameters of model

Hyper parameters	Value
batch size	32
num epoch	30
initial learning rate	1e-4
optimizer	Adam
weight decay	0.0005
p_margin	0.2

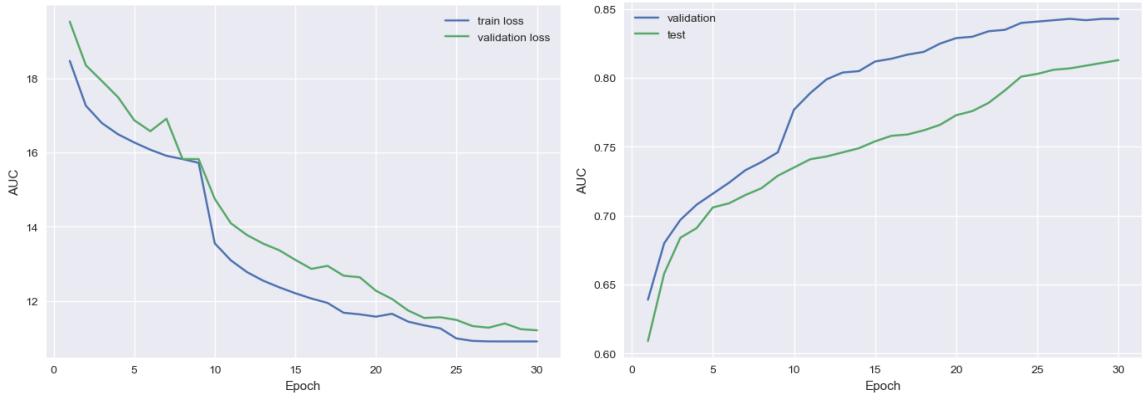


Figure 18: Training and validation loss (left) and AUC score (right)

The changing of loss and AUC per epoch is shown in Figure 17. Table 2 show our model performance across all classes and compare with other previous works. As we can see, our model performs better than other previous models, improving 11.3% AUC score compared to the best model training with the same dataset and 6.4% compared to model training with extra dataset. The ROC curve of our model is shown in Figure 18.

Table 2: AUC score using our method on the official ChestX-ray14 test set.

Method	Wang et al.[1]	Yao et al. [8]	DNet [14]*	Our
Atelectasis	0.7003	0.733	0.767	**0.8395**
Cardiomegaly	0.81	0.856	0.883	**0.9365**
Effusion	0.77	0.806	0.828	**0.8721**
Infiltration	0.6614	0.673	0.709	**0.7642**
Mass	0.6933	0.718	0.821	**0.9131**
Nodule	0.6687	0.777	0.758	**0.8669**
Pneumonia	0.6580	0.684	0.731	**0.7999**
Pneumothorax	0.7993	0.805	0.846	**0.9183**
Consolidation	0.7032	0.711	0.745	**0.8062**
Edema	0.8052	0.806	0.835	**0.8898**
Emphysema	0.8330	0.842	0.895	**0.9520**
Fibrosis	0.7859	0.743	0.818	**0.9008**
Pleural Thick	0.6835	0.724	0.761	**0.8497**
Hernia	0.8717	0.775	0.896	**0.9231**
Mean	0.7451	0.761	0.807	**0.874**

*The method represented in [7] was trained by using more than 180,000 images from the PLCO dataset [45] as extra training data.

Table 3: Comparison of different loss function

Loss function	**AUC score**	**mAP**
BCE	0.804	0.268
Focal	0.821	0.272
Asymmetric	0.835	0.293
Our w/o decoupling	0.870	0.319
Our w/o margin	0.868	0.312
Our with decoupling and margin	**0.874**	**0.338**

Table 3 compares model performance when using others loss function with our loss function. As we can see, our loss function performs better than BCE, Focal, and Asymmetric loss, the best result is archive when combining shifted negative probability and decoupling strategy.

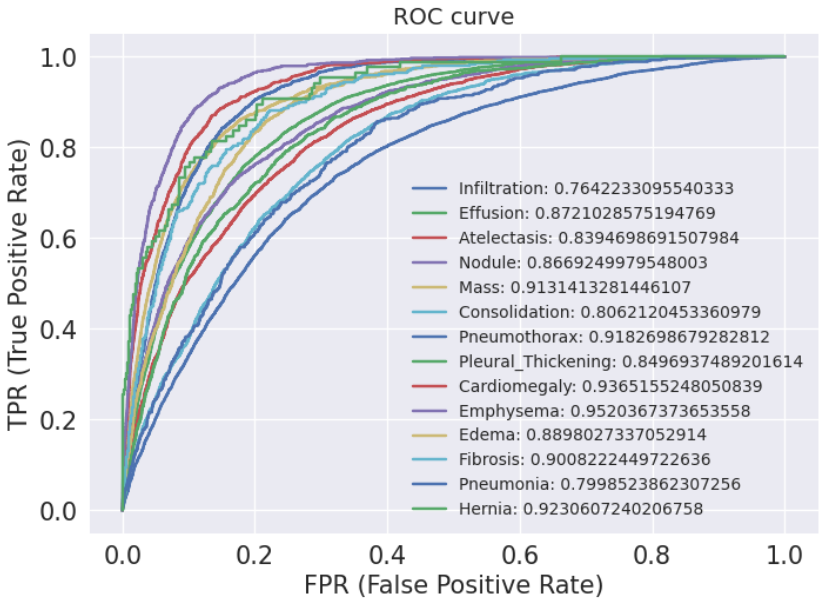


Figure 19: ROC curve of each class

4. Conclusion

In this study, we modify traditional loss function to get a new loss function to address the problem of class imbalance in image classification. The modified loss function was applied on the Swin Transformer model with pretrained weights on the ImageNet dataset, using the base version of Swin Transformer, Swin-B. Our experimental results showed that our loss function achieved the best performance compared to previous loss functions. Our research successfully achieved the initial goal and addressed the research question of solving the problem of class imbalance in image classification. We demonstrated that our loss function can significantly improve the performance of image classification tasks.

However, there are still some limitations to our research. Due to the equipment limitation, we only used the Swin-B version, future research can explore bigger version. Additionally, we only evaluated our loss function on Chest-Xray14 datasets and did not examine its performance on other long-tailed datasets. We hope that our research can inspire further studies in this field and contribute to the development of more effective solutions for class imbalance in image classification.

Reference

[1] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M. and Summers, R.M., 2017. 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* (pp. 2097-2106).

[2] Bar, Y., Diamant, I., Wolf, L., & Greenspan, H. "Deep learning with non-medical training used for chest pathology identification. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 1-9)". Springer, Cham, 2015.

[3] Majdi, M.; Salman, K.; Morris, M.; Merchant, N.; Rodriguez, J. Deep learning classification of chest X-ray images. In Proceedings of the Southwest Symposium on Image Analysis and Interpretation (SSIAI), Albuquerque, NM, USA, 29–31 March 2020; pp. 116–119.

[4] Cicero, M.; Bilbily, A.; Colak, E.; Dowdell, T.; Gray, B.; Perampaladas, K.; Barfett, J. Training and validating a deep convolutional neural network for computer-aided detection and classification of abnormalities on frontal chest radiographs. *Investig. Radiol.* **2017**, *52*, 281–287.

[5] Rasheed, J.; Hameed, A.A.; Djeddi, C.; Jamil, A.; Al-Turjman, F. A machine learning-based framework for diagnosis of COVID-19 from chest X-ray images. *Interdiscip. Sci. Comput. Life Sci.* **2021**, *13*, 103–117.

[6] Allaouzi I, Ahmed MB (2019) A novel approach for multi-label chest x-ray classification of common thorax diseases. IEEE Access 7:64279–64288

[7] Ait Nasser, A.; Akhloufi, M.A. Chest Diseases Classification Using CXR and Deep Ensemble Learning. In Proceedings of the 19th International Conference on Content-Based Multimedia Indexing, Graz, Austria, 14–16 September 2022; pp. 116–120.

[8] Yao, L., Poblenz, E., Dagunts, D., Covington, B., Bernard, D. and Lyman, K., Learning to diagnose from scratch by exploiting dependencies among labels. arXiv 2017. *arXiv preprint arXiv:1710.10501*.

[9] Kumar, P., Grewal, M. and Srivastava, M.M., 2018. Boosted cascaded convnets for multilabel classification of thoracic diseases in chest radiographs. In *Image Analysis and Recognition: 15th International Conference, ICIAR 2018, Póvoa de Varzim, Portugal, June 27–29, 2018, Proceedings 15* (pp. 546-552). Springer International Publishing.

[10] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K. and Lungren, M.P., 2017. Chexnet: Radiologist-level

pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*.

[11] Kim, S.; Rim, B.; Choi, S.; Lee, A.; Min, S.; Hong, M. Deep Learning in Multi-Class Lung Diseases' Classification on Chest X-ray Images. *Diagnostics* **2022**, *12*, 915.

[12] Blais, M.A.; Akhloufi, M. Deep Learning and Binary Relevance Classification of Multiple Diseases using Chest X-ray images. In Proceedings of the 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC), Guadalajara, Mexico, 1–5 November 2021; pp. 2794–2797.

[13] Li, Z., Wang, C., Han, M., Xue, Y., Wei, W., Li, L.J. and Fei-Fei, L., 2018. Thoracic disease identification and localization with limited supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8290-8299).

[14] Guendel, S., Grbic, S., Georgescu, B., Liu, S., Maier, A. and Comaniciu, D., 2019. 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* (pp. 757-765). Springer International Publishing.

[15] Guan, Q., Huang, Y., Zhong, Z., Zheng, Z., Zheng, L. and Yang, Y., 2018. Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification. *arXiv preprint arXiv:1801.09927*.

[16] Wang, H. and Xia, Y., 2018. Chestnet: A deep neural network for classification of thoracic diseases on chest radiography. *arXiv preprint arXiv:1807.03058*.

[17] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, pp.321-357.

[18] Estabrooks, A., Jo, T. and Japkowicz, N., 2004. A multiple resampling method for learning from imbalanced data sets. *Computational intelligence*, *20*(1), pp.18-36.

[19] Liu, X.Y., Wu, J. and Zhou, Z.H., 2008. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *39*(2), pp.539-550.

[20] Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; Kalantidis, Y. Decoupling representation and classifier for long-tailed recognition. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 30 April 2020.

[21] Park, M.; Song, H.J.; Kang, D.O. Imbalanced Classification via Feature Dictionary-Based Minority Oversampling. *IEEE Access* **2022**, *10*, 34236–34245.

[22] Lee, Y.S.; Bang, C.C. Framework for the Classification of Imbalanced Structured Data Using Under-Sampling and Convolutional Neural Network. *Inf. Syst. Front.* **2021**, *24*, 1795–1809.

[23] Ding, H.; Wei, B.; Gu, Z.; Zheng, H.; Zheng, B. KA-Ensemble: Towards imbalanced image classification ensembling under-sampling and over-sampling. *Multimed. Tools Appl.* **2020**, *79*, 14871–14888.

[24] Gupta, A.; Dollar, P.; Girshick, R. Lvis: A dataset for large vocabulary instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5356–5364.

[25] Peng, J.; Bu, X.; Sun, M.; Zhang, Z.; Tan, T.; Yan, J. Large-scale object detection in the wild from imbalanced multi-labels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9709–9718.

[26] Swana, E.F.; Doorsamy, W.; Bokoro, P. Tomek link and SMOTE approaches for machine fault classification with an imbalanced dataset. *Sensors* **2022**, *22*, 3246.

[27] Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A. and Van Der Maaten, L., 2018. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 181-196).

[28] Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J. and Kalantidis, Y., 2019. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*.

[29] Cui, Y., Jia, M., Lin, T.Y., Song, Y. and Belongie, S., 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9268-9277).

[30] Lin, T.Y., Goyal, P., Girshick, R., He, K. and Dollár, P., 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980-2988).

[31] Hermans, A.; Beyer, L.; Leibe, B. In Defense of the Triplet Loss for Person Re-Identification. *arXiv* **2017**, arXiv:1703.07737.

[32] Cao, K.; Wei, C.; Gaidon, A.; Arechiga, N.; Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; p. 32.

[33] Wu, T.; Huang, Q.; Liu, Z.; Wang, Y.; Lin, D. Distribution-balanced loss for multi-label classification in long-tailed datasets. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 162–178.

[34] Tan, J.; Lu, X.; Zhang, G.; Yin, C.; Li, Q. Equalization loss v2: A new gradient balance approach for long-tailed object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1685–1694.

[35] Wang, J.; Zhang, W.; Zang, Y.; Cao, Y.; Pang, J.; Gong, T.; Chen, K.; Liu, Z.; Loy, C.C.; Lin, D. Seesaw loss for long-tailed instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 9695–9704.

[36] Hong, Y.; Han, S.; Choi, K.; Seo, S.; Kim, B.; Chang, B. Disentangling label distribution for long-tailed visual recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6626–6636.

[37] Cui, J., Zhong, Z., Liu, S., Yu, B. and Jia, J., 2021. Parametric contrastive learning. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 715-724).

[38] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. *Advances in neural information processing systems*, *30*.

[39] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. and Uszkoreit, J., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[40] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012-10022).

[41] Hahnloser, R.H., Sarpeshkar, R., Mahowald, M.A., Douglas, R.J. and Seung, H.S., 2000. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *nature*, *405*(6789), pp.947-951.

[42] Hendrycks, D. and Gimpel, K., 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

[43] Wu, T., Huang, Q., Liu, Z., Wang, Y. and Lin, D., 2020. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16* (pp. 162-178). Springer International Publishing.

[44] Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M. and Zelnik Manor, L. 2021. Asymmetric loss for multi-label classification. Proceedings of the IEEE/CVF International Conference on Computer Vision (2021), 82–91.

[45] Gohagan, J.K., Prorok, P.C., Hayes, R.B., Kramer, B.S. and PLCO Project Team, 2000. The prostate, lung, colorectal and ovarian (PLCO) cancer screening trial of the National Cancer Institute: history, organization, and status. *Controlled clinical trials*, *21*(6), pp.251S-272S.