Weighted Focal Loss: An Effective Loss Function to Overcome Unbalance Problem of Chest X-ray14

Ruoxi Qin, Kai Qiao, Linyuan Wang, Lei Zeng, Jian Chen and Bin Yan*

National Digital Switching System Engineering and Technological Research Center, Zhengzhou, China, 450001
ybspace@hotmail.com

Abstract. X-ray film has been applied widely in the elementary diagnosis and screening of thorax diseases. The rapid development of deep learning has improved automatic diagnosis of thorax disease. But different thorax disease in database inevitably exist the problem of unbalanced quantity because of the different morbidity in real life. So, the restricted factor is not just the architecture of the models but the unbalanced problem of database. What’s more, different from the same problem of natural images, this unbalanced problem is not just the unbalanced quantity but also the unbalanced complexity of diagnosis between samples. In this study, we introduced the Focal Loss and proposed the W-FL (weight Focal Loss) function which makes the model pay more attention to the difficult samples and those kinds of diseases which have small quantity and protect the loss and computed gradients from overwhelming by easy negatives. We evaluate the efficiency of W-FL function on the public Chest X-ray14 database using prevailing per-trained models including AlexNet, Inception-v3 and ResNet. We exceed at least 2% average AUC value on 14 different thorax diseases than the W-CEL function which is commonly proposed to solve the unbalanced problem in this database.

1. Introduction
X-rays film is the most common type of radiology examination in the world and a particularly challenging example of multi-label classification in medical diagnosis. Making up nearly 45% of all radiological studies, the x-ray achieves global ubiquity as a screening tool for pathologies including thorax diseases. Automatically detecting the abnormalities with high accuracy system could greatly enhance the diagnosis’s process. But the real challenge is the difficulty for classification on some diseases which even the expert radiologist may misdiagnose, let alone the multifarious diseases and independence between them in each film. Especially for the thorax diseases whose fundamental diagnosis and screening mostly depend on the x-ray films, the diagnosis drives the demand of accurate and quick diagnosis through a multitude of pathological outcomes.

With the development of deep learning, more and more people want to use this method to build high precision computer-aided diagnosis (CAD) systems, and more and more experiments proved the effectiveness of deep learning to achieve this goal [1-5]. The appearance of hospital-scale Chest X-ray14 database, which 112,120 Chest X-ray images are individually labelled with up to 14 different thoracic diseases makes it possible for deep learning combining with the diagnosis of thorax diseases closely.

Mined from the text radiological reports via NLP techniques [6], the Chest X-ray14 database inevitably suffer from the unbalance-class problem because of the different morbidity in real life.
Table 1 shows the proportion of abnormalities. Among abnormalities, the unbalanced problem is also serious while the “Infiltration” makes up most quantity of the data (14.05%) and the “Hernia” occupies the least part of database (0.16%). On the contrast, half proportion of this database is “No Finding”. There are many methods to deal with unbalance-class problem of database [7]. Some previous studies showed results on cost sensitive learning of deep neural networks [8-10] and more work is about data level method such as oversampling and undersampling [11-12]. There also many study about this problem of medical decision [13]. Although there are lots of methods, we can’t ignore the unbalanced complexity of diagnosis – ideally with the use of Focal loss function to overcome the complicated unbalanced problem between samples.

Some notable work about this database includes a baseline work about Chest X-ray14 using transfer learning model [6]. A partial solution using LSTMs to extract interdependencies among target labels in predicting 14 pathologic patterns from chest x-rays [14]. A new network named CheXNet with 121-layers trained on Chest X-ray14 achieves state of the art on all 14 diseases [15]. Compared with baseline work, all progresses on this database pay attention to the improvement of network architecture but ignore the unbalanced problem. W-CEL function proposed in [6] is the only method which is also used in other network focusing on the unbalanced problem. The definition of W-CEL is shown as below:

$$
\beta = \begin{cases}
\beta_p = \frac{|P| + |N|}{|P|} & y = 1 \\
\beta_N = \frac{|P| + |N|}{|N|} & y = 0
\end{cases}
$$

In this function, the $P$ defined as the total numbers of “1” in a batch of labels while the $N$ is the total numbers of “0”. $y$ means each element of labels. Through the statistic quantity of “0” and “1” in each label, the W-CEL function only makes the model focus on the unbalanced problem between abnormalities and normal samples.

According to the above analysis, in order to achieve accurate diagnosis on Chest X-ray14, it is necessary for us to overcome the unbalanced problem in this database. In this paper, we excavate the complexity of the Chest X-ray14 data by contrasting the different results using different loss function and improve the FL loss functions [16] to accommodate the multi-labels classification. We apply and improve the FL loss function to achieve a quick and reliable transfer learning on this database. Through the experiments, we find that the mutual influence of unbalance-class and complicated diagnosis make it more difficult for training. Our experiments show a simple extension of loss function to classify multiple diseases outperforms previous transfer learning on Chest X-ray14.

### Table 1. Proportion of different diseases in Chest X-ray14.

| Disease           | Proportion | Disease      | Proportion |
|-------------------|------------|--------------|------------|
| Atelectasis       | 8.16%      | Cardiomegaly | 1.96%      |
| Effusion          | 9.41%      | Emphysema    | 1.78%      |
| Infiltration      | 14.05%     | Nodule       | 4.47%      |
| Pneumonia         | 0.957%     | Fibrosis     | 4.19%      |
| Pneumothorax      | 3.75%      | Mass         | 4.06%      |
| Pleural_Thickening| 2.39%      | Consolidation| 3.30%      |
| Edema             | 1.63%      | Hernia       | 0.161%     |

#### 2. Method

**2.1 Pre-trained model**

Our goal is finding the mutual influence of unbalanced problem in Chest X-ray14 and the effectiveness of Focal Loss function. We tackle this problem by using a transfer learning model trained on ImageNet and comparing the result with other loss function (CEL, W-CEL) to verify our ideas. For the quick transfer learning on Chest X-ray14 dataset, we defined the unified transfer
learning framework as mentioned in [6]. We use the pre-trained models as the instrument of extracting high-dimensional features. To roundly demonstrate the complex interactions between this database and the universality of Focal Loss function, we use the per-trained AlexNet [17] and Inception-v3 [18] and ResNet [19] as our transfer learning models.

We firstly use the per-trained AlexNet model which has 650 thousand neurons and 60 million parameters and reduce the top-5 error rate to 16.4% as one of our baseline networks [17]. As one of the representative Convolutional Neural Networks, AlexNet promotes the development of deep learning in computer vision. To completely verify our efficiency of the Focal loss function is better than W-CEL function, we also make experiment on the ResNet [19], which has the best performance in [6] using transfer learning model.

As a creative architecture recently, Inception-v3 promotes the performance in ways that utilize the added computation as efficiently as possible by suitably factorizing convolutions and aggressive regularization [18]. The architecture of Inception-v3 implements the idea of “factorization into small convolutions”, which can make the network handle more plentiful features.

2.2 Fine-tune network

2.2.1 Transition layer connected to bottleneck layer. We train our fully-connected layers as the decoder of the high-dimensional features. The transition layers help pass down the weights from pre-trained DCNN models and transform the activations into a uniform dimension of output. Through the decoder of high-dimensional features, we can use activations to represent the possibility of each disease. The transition layers have the most of parameters for us to train and the nonlinear expression of that decide the performance of our model.

2.2.2 Loss layer. To make the output more reasonable for multi-label problem, we replace the softmax nonlinearity which is often used in traditional loss layers with a sigmoid nonlinearity, outputting the probability that the image contains each disease in each dimensionality (14 in this database). Without the normalization of softmax, that is the way to furthest address the multi-label issue. In this way we can make each dimensionality of the output represent the probability of each disease among “0” to “1”, which are more reasonable when we calculate the loss with multi-labels. This definition of output transits the multi-label problem into a regression-like loss setting. What’s more, as note in [16], the implementation of the loss layer combines with the sigmoid operation for computing the loss computation, resulting in greater numerical stability. That’s another reason for us to use sigmoid nonlinearity as our activated function in loss layer.

Figure 1. The flow-chart of our transfer learning model

Figure 1 illustrates the DCNN (Deep Convolutional Neural Network) architecture we adapted. As shown in Figure 1, we perform the network surgery on the pre-trained models (using ImageNet [20]),
e.g., AlexNet [17], Inception-v3 [18] and ResNet [19], by leaving out the transition layers and the final prediction layers. Among the aforementioned architecture, we set the amount of the neurons in transition layers as 4096 in Inception-v3 and 2048 in AlexNet.

2.3 Improved Focal Loss for Classification

Recently more and more new kinds of loss function for neural networks training were also developed [21]. About Chest-Xray14 database, they tacitly approve the W-CEL loss function [6] as their method to solve the unbalanced problem about the database. The original intention of W-CEL loss function is solving the unbalanced problem caused by different quantity of pathologies and normality. Inspired by the W-CEL loss function, we creatively use the Focal Loss function [16] which is proposed to promote dense detectors. Furthermore, we modify the Focal loss function to pay attention to the mutual influence between unbalance-class and complex diagnosis. The focal loss function defined as below:

\[ p_t = \begin{cases} p & \text{if } y=1 \\ 1 - p & \text{otherwise} \end{cases} \]

\[ FL(p_t) = -(1 - p_t)\gamma \log(p_t) \]

In the above \( \gamma \in \{0,1\} \) specifies the ground-truth class; \( \gamma \in \{0,5\} \) aims at adjusting the rate by down-weight the loss of easy examples [16]. In our experiment, we set the \( \gamma \) as 2, which is mentioned in [16] that has the best performance. The experiment shows some disadvantage about FL function, to overcome that, we improve the definition about the parameter \( \beta \). Different from the originally definition about \( \alpha \) in two-dimension. We defined \( \beta \) as the corresponding weight similar as the W-CEL function in formula (1). So our modified FL Loss function defined as below:

\[ FL(p_t) = -\beta(1 - p_t)\gamma \log(p_t) \]

Through the modified Focal Loss function, the parameter \( \beta \) and \( \gamma \) solve both sides of unbalanced problem and achieve a good performance with transfer learning models. What’s most important is that we propose a new application of Focal Loss function which we think will be more suitable for the unbalanced problem of Chest X-ray14 database.

3. Experiments and results

3.1 Experiment design

3.1.1 Processing Chest X-ray14 Data. To reduce the operating pressure of computer and make our input image suitable for our transfer learning model, we transform the X-rays images’ size to 299×299 for Inception-v3 and 227×227 for AlexNet and unify them to grayscale images (some image of this data is not grayscale). For the pathology classification, we randomly shuffled the entire dataset into two subgroups: i.e. training (80%) and validation (20%). We make sure that the data of training and validation are strictly separated so that we can have an appraisal about our classification model. About the multi-label setup, we do the same way as is mentioned in [6]. We define a 14-dimensional label vector for each image and the all-zero vector represents the “No Finding” while the factor “1” in labels means the existent of corresponding disease.

3.1.2 Metrics. Considering the relevance between different diseases, pathological diagnosis is not like traditional image classification. Instead of using the common top accuracy judging criteria to define the loss, we set the corresponding Area-Under-Curve (AUC) values as the judging criteria adjusted to multi-class tasks since accuracy metric is associated with notable difficulties in the context of imbalanced data. The corresponding Area-Under-Curve (AUC) values is only calculated on validation to demonstrate the performance of model. We select 4800 images randomly in validation data to
calculate the AUC value. The set of metrics provides a meaningful quantification of this performance to verify the effectiveness of our loss function.

3.2 Classification Result
Comparing the result using CEL with that using W-CEL, Figure 2 shows the interaction between proportion and performance. With W-CEL function, the AlexNet’s performance of the AUC values is connected to the unbalance-class of different diseases. The W-CEL function has already balance the loss between normal examples and abnormal examples and has integrally promotes the performance of examples which has poor proportion (except “Hernia”). But the rising value of AUC connected to the proportion is not unconditional: the performance of “Hernia” which has the poorest proportion even decline in AlexNet and compared with the result of ResNet the performance of the AUC values is not connected to proportion. Those performances verify two aspect disadvantage of W-CEL function: a) the W-CEL function doesn’t pay attention on the different proportion of abnormalities, which will let the examples with poor proportion has bad performance; b) the performance of W-CEL varies greatly. In network with simple architecture, the performance is connected to the proportion while it is unconnected with deeper network and the average value of AUC even decline in Inception-v3. It will get worse performance with the deeper architecture. The poor performance of “Hernia” compared with that using CEL function also proves the unbalance-class problem between abnormalities and prove the inefficiency of W-CEL function. We must address this problem not only in statistics but in pathology, which shows the complicated unbalanced problem in this database. This problem immensely hinders the application of high precision computer-aided diagnosis (CAD) systems on large-scale disease diagnosis.

![Figure 2. Interaction between proportion and performance. Top: Alexnet with W-CEL Bottom: ResNet with W-CEL](image-url)
The complexity and unbalance problem mentioned above, we improve the performance when we train the model from scratch. Compared with the result with W-CEL, we can find some correlation between the pathologies’ proportion and the change of different pathologies’ AUC value. With the less proportion of the data, Focal Loss universally works better, which solve a degree of problem about unbalance proportion.

We use the FL Loss function more effective on this problem, we improve the classification result using different pre-trained model with different loss function.

|                  | Pre-trained AlexNet | Pre-trained Inception-v3 | Pre-trained ResNet-50 |
|------------------|---------------------|--------------------------|-----------------------|
| CEL W-CEL FL W-FL CEL W-CEL FL W-FL CEL W-CEL FL W-FL |
| Atelectasis      | 0.665 0.715 0.721 0.709 0.683 0.666 0.679 0.697 0.620 0.626 0.702 0.678 |
| Infiltration     | 0.611 0.58 0.623 0.597 0.558 0.583 0.598 0.567 0.567 0.577 0.615 0.605 |
| Pneumothorax     | 0.663 0.724 0.722 0.713 0.703 0.782 0.766 0.695 0.615 0.654 0.762 0.721 |
| Edema            | 0.538 0.852 0.835 0.826 0.706 0.616 0.785 0.792 0.826 0.824 0.830 0.829 |
| Emphysema        | 0.610 0.683 0.714 0.745 0.643 0.610 0.763 0.754 0.581 0.541 0.781 0.708 |
| Fibrosis         | 0.576 0.723 0.760 0.739 0.675 0.586 0.745 0.736 0.659 0.689 0.688 0.734 |
| Consolidation    | 0.453 0.722 0.706 0.729 0.598 0.667 0.776 0.753 0.687 0.736 0.725 0.762 |
| Effusion         | 0.760 0.786 0.783 0.790 0.803 0.691 0.810 0.778 0.709 0.755 0.820 0.800 |
| Pneumonia        | 0.345 0.550 0.666 0.580 0.337 0.566 0.639 0.597 0.616 0.653 0.710 0.679 |
| Pleural Thickening| 0.587 0.64 0.696 0.723 0.548 0.524 0.670 0.682 0.650 0.606 0.679 0.712 |
| Cardiomegaly     | 0.623 0.79 0.848 0.865 0.584 0.668 0.758 0.813 0.618 0.636 0.847 0.757 |
| Nodule           | 0.529 0.547 0.591 0.596 0.588 0.543 0.631 0.585 0.504 0.522 0.603 0.544 |
| Mass             | 0.611 0.652 0.688 0.649 0.612 0.527 0.684 0.679 0.54 0.570 0.711 0.640 |
| Hernia           | 0.714 0.569 0.516 0.551 0.711 0.561 0.489 0.526 0.532 0.554 0.472 0.504 |
| Average          | 0.591 0.680 0.705 0.701 0.625 0.610 0.705 0.689 0.623 0.638 0.712 0.692 |

The results of all those four loss functions with three different models are demonstrated in Table 2. Compared with the result with W-CEL, we can find some correlation between the pathologies’ proportion and the change of different pathologies’ AUC value. With the less proportion of the data, Focal Loss universally works better, which solve a degree of problem about unbalance-class. But we can’t ignore the phenomenon that the AUC value of “Hernia” which has the least proportion and is easy to diagnose (having good performance with CEL function) decreases by 19.5 instead of increasing. That may because the γ make the model give more weight on the complexity example without a parameter to balance the complexity and unbalance-class problem making even the easy examples are overwhelming because of the poor proportion. That also shows the unbalanced problem in statistics and pathology. The decline of the AUC values is not only because the unbalance-class of different diseases, but also because the complexity of diagnosis. The combine of those problems makes the classification of diseases more difficult than that of nature images.

To use the FL Loss function more effective on this problem, we improve the definition about the parameter β. We use the loss function in formula (4). With the modified FL loss function, our performance on “Hernia” outperforms that without parameter α (increase by 3.5% on AlexNet and increase by 8.7% on Inception-v3). The introduction of parameter β makes the performance of model more balanced without enhances the performance entirely. But we can’t ignore the advantage it takes. We believe that through the painstaking work on selecting parameter, the modified FL loss function can further improve the performance when we train the model from scratch.

4. Conclusion
In this paper, we firstly demonstrate the influence of the unbalanced problem in two aspects. According to the result, we find that the influence of unbalanced problem will amplify when more samples with small quantity are introduced to the database and the different result with different loss function shows the interaction of unbalance-class and complexity of diagnosis, which also verify the common loss function’s limitation. To eliminate the negative effects from FL function and solve the problem mentioned above, we creatively combine the FL function with W-CEL function and propose
loss function named W-FL. With that Weighted Focal Loss function our transfer learning models achieve a good performance on Chest X-ray.

Acknowledgments
This work was supported by the National Key R&D Program of China under grant 2017YFB1002502 and National Natural Science Foundation of China (No. 61701089, No.61601518).

References
[1] Lakhani, Paras and Sundaram, Baskaran. Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology, pp. 162326, 2017.
[2] Esteva, Andre, Kuprel, Brett, Novoa, Roberto A, Ko, Justin, Swetter, Susan M, Blau, Helen M, and Thrun, Sebastian. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639):115-118, 2017.
[3] Islam, Mohammad Tariqul, Aowal, Md Abdul, Minhaz, Ahmed Tahseen, and Ashraf, Khalid. Abnormality detection and localization in chest x-rays using deep convolutional neural networks. arXiv preprint arXiv:1705.09850, 2017.
[4] Rajpurkar, Pranav, Hannun, Awni Y, Haghpanahi, Masoumeh, Bourn, Codie, and Ng, Andrew Y. Cardiologist-level arrhythmia detection with convolutional neural networks. arXiv preprint arXiv:1707.01836, 2017.
[5] Grewal, Monika, Srivastava, Muktabh Mayank, Kumar, Pulkit, and Varadarajan, Srikrishna. Radnet: Radiologist level accuracy using deep learning for hemorrhage detection in ct scans. arXiv preprint arXiv:1710.04934, 2017.
[6] Wang, Xiaosong, Peng, Yifan, Lu, Le, Lu, Zhiyong, Bagheri, Mohammadhadi, and Summers, Ronald M. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. arXiv preprint arXiv:1705.02315, 2017.
[7] Buda M, Maki A, Mazurowski M A. A systematic study of the class imbalance problem in convolutional neural networks[J]. 2017.
[8] Salman H Khan, Mohammed Bennamoun, Ferdous Sohel, and Roberto Togneri. Cost sensitive learning of deep feature representations from imbalanced data. arXiv preprint arXiv:1508.03422, 2015.
[9] Vidwath Raj, Sven Magg, and Stefan Wermter. Towards effective classification of imbalanced data with convolutional neural networks. In IAPR Workshop on Artificial Neural Networks in Pattern Recognition, pages 150-162. Springer, 2016.
[10] Yu-An Chung, Hsuan-Tien Lin, and Shao-Wen Yang. Cost-aware pre-training for multiclass cost-sensitive deep learning. arXiv preprint arXiv:1511.09337, 2015.
[11] Gil Levi and Tal Hassner. Age and gender classification using convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 34-42, 2015.
[12] Andrew Janowczyk and Anant Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. Journal of pathology informatics, 7, 2016.
[13] Maciej A Mazurowski, Piotr A Habas, Jacek M Zurada, Joseph Y Lo, Jay A Baker, and Georgia D Tourassi. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. Neural networks, 21(2):427-436, 2008.
[14] Yao, Li, Poblenz, Eric, Dagunts, Dmitry, Covington,Ben, Bernard, Devon, and Lyman, Kevin. Learning to diagnose from scratch by exploiting dependencies among labels. arXiv preprintarXiv:1710.10501,2017.
[15] Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning[J]. 2017.
[16] Lin T Y, Goyal P, Girshick R, et al. Focal Loss for Dense Object Detection[J]. 2017:2999-3007.
[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.

[18] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the Inception Architecture for Computer Vision[C] IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2016:2818-2826.

[19] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C] IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2016:770-778.

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. FeiFei. Imagenet: A large-scale hierarchical image database. In Computer Vision and Pattern Recognition, pages 248–255. IEEE, 2009. 5

[21] Shoujin Wang, Wei Liu, Jia Wu, Longbing Cao, Qinxue Meng, and Paul J Kennedy. Training deep neural networks on imbalanced data sets. In Neural Networks (IJCNN), 2016 International Joint Conference on, pages 4368-4374. IEEE, 2016.