OPEN

An aggregate method for thorax diseases classification

Bayu Adhi Nugroho

A common problem found in real-world medical image classification is the inherent imbalance of the positive and negative patterns in the dataset where positive patterns are usually rare. Moreover, in the classification of multiple classes with neural network, a training pattern is treated as a positive pattern in one output node and negative in all the remaining output nodes. In this paper, the weights of a training pattern in the loss function are designed based not only on the number of the training patterns in the class but also on the different nodes where one of them treats this training pattern as positive and the others treat it as negative. We propose a combined approach of weights calculation algorithm for deep network training and the training optimization from the state-of-the-art deep network architecture for thorax diseases classification problem. Experimental results on the Chest X-Ray image dataset demonstrate that this new weighting scheme improves classification performances, also the training optimization from the EfficientNet improves the performance furthermore. We compare the aggregate method with several performances from the previous study of thorax diseases classifications to provide the fair comparisons against the proposed method.

The traditional classifiers such as an SVM, a decision tree or a logistic regression require feature engineering to perform classification. The better features chosen during the feature engineering will produce more accurate classification performance. Despite the necessity of the feature engineering, a neural network has the advantages of performing end-to-end training to output a final classification's prediction. The removal of feature engineering in the neural network will reduce the risk of the use of incorrect features for classification. To develop a neural network for medical diagnosis, patient data are necessary; however, the positive class is minority and the negative class is majority. The neural network is biased to the majority class and has poor performance on the minority class. The common methods to balance the number of positive and negative class for the traditional classifiers is by the use of undersampling and oversampling approaches. After applying those methods, the numbers of each training-pattern are equal. The more sophisticated approach might involve algorithmic technique to perform cost-sensitive training¹. "The effective number of samples is defined as the volume of sample"². Cui et al.² developed a metric to determine the effective samples and reformulated the loss function based on the numbers of effective samples in the positive and negative classes. The work of Wang et al.³ contributes to provide the availability of a novel Chest X-Ray dataset. Both^{3,4} use the same formulation to balance the dataset, however the method from^{3,4} is different than² to deal with the imbalance problem. Cui et al.'s approach² treats the contributions of training patterns to the loss function equally for all the output nodes, this is contrary to the Wang and Gundel et al.'s methods^{3,4} which use the distinct weights from positive and negative classes as the multipliers in the loss-function. Although better classification performance can potentially be achieved by Cui et al.², the approach only addresses effective samples² and the imbalances of positive-negative classes have not been tackled. In this paper, a novel weights function for focal-loss is proposed to address the imbalance of positive-negative classes, which tackles the classification correctness in both positive and negative samples when training the neural networks. The performance of the proposed focal-loss function is evaluated by performing Chest X-Ray classification which is involved with imbalanced data³. We also propose the use of EfficientNet⁵ with progressive image resizing under two-stage training in complement with the proposed loss-function. The motivation to use EfficientNet is to inspect the outcome of the proposed loss function into different architecture scaling. The aggregate of the proposed loss-function and the two-stage EfficientNet training achieved 2.10% improvement, which is measured with the area under the receiver-operating-characteristic curve (AUROC). Also, heatmap visualization shows that the proposed aggregate approach can achieve better coverage of the diseases. According to Baltruschat et al.⁶ the current state-of-the-art performance for the dataset³ classification performance was achieved by Gündel et al.⁴. Further research by Guan et al.⁷ which use three-stage training procedures reported better performance than Gündel et al.⁴. However, the work⁷ did not share the split-sets which is critical for the performance evaluation, also the re-implementation by another party in github⁸ reported lower results. We also notice that the re-implementation⁸ of⁷ did not share identical sets with the work of Gündel

Curtin University, Bentley, WA 6102, Australia. email: BayuAdhi.Nugroho@postgrad.curtin.edu.au

et al.⁴. Baltruschat et al.⁶ noticed that different split-sets would lead to different performances for the dataset³. To have the fair benchmarks, we report several results from various split-sets options for the performance evaluation. We perform three split-sets experiments configuration setup, which aims to provide better evaluation and have the comprehensive analysis; the first is by the use of “official” split from⁹, the second is under fivefolds cross-validation configuration which has also been used in the work of Baltruschat et al.⁶, and the last one is by the use of identical splits from the public Github-page^{8,10}. We achieve state-of-the-art results for the classification problem of the Chest X-Ray dataset³, measured under these three split-sets configurations. This research is mainly contributing to the improvement of medical image classification problem. We tackle the imbalance problem within Chest X-Ray dataset and propose the advancement of the use of state-of-the-art neural net architecture for the final classification performance; The EfficientNet with two-stage training. In summary, our contribution is to propose an approach which can combine weights calculation algorithm for deep-network and the optimization of training strategy from the state-of-the-art architecture. The “Introduction” section of this paper provides a brief introduction and overview of the research. The “Method” section mainly discusses the existing classification approach and the proposed method. The “Experiments and results” section presents the results from experiments. The “Discussion” section give a more in-depth discussion about the outcome, then The “Conclusion” section explains the conclusion from the research.

Method

The existing weights function and network architecture. Wang et al.³ and Gündel et al.⁴ defined the weights, ω_{k+} and ω_{k-} , of the positive and negative samples for the k^{th} pattern.

$$\begin{aligned}\omega_{k+} &= \frac{P_k + N_k}{P_k} \\ \omega_{k-} &= \frac{P_k + N_k}{N_k}\end{aligned}\quad (1)$$

where P_k and N_k are the numbers of positive and negative samples for the k^{th} pattern. However, Cui et al.² used both ω_{k+} and ω_{k-} equally to develop the loss function. In the manuscript, Wang et al.³ and Gündel et al.⁴ did not use an identical dataset. Wang et al.³ used Chest X-Ray 8 which consists of eight classes with 108,948 images. Whilst⁴ used Chest X-Ray 14 which consists of fourteen classes with 121,120 images. However, we can find the results from the method of Wang et al.³ under the Chest X-Ray 14 “official split” configuration in the manuscript of Gündel et al.⁴. Both works implemented Eq. (1) in the loss-function according to the literature^{3,4}. Therefore, we can conclude the Eq. (1) applies to the samples of the training set. Lin et al.¹¹ proposed the focal-loss function:

$$L_{foc}(p) = -\alpha(1 - p)^\gamma \log(p). \quad (2)$$

p is the prediction. In Eq. (2), the parameter α attempts to balance the positive-negative samples while γ adjusted to release the easy samples and to dominate the hard samples, where the easy and hard samples are those classified correctly and incorrectly respectively. Generally, $\gamma \geq 0$; when $\gamma = 0$ focal-loss is the same as an ordinary cross-entropy loss¹¹. The experimental results showed that the easy samples are down-weighted when $\gamma \approx 1$; The samples are further down-weighted when $\gamma > 1$. Determination of α is discussed to demonstrate the impact to the focal loss function (Eq. 2). The parameters chosen as below²:

$$\begin{aligned}\beta &= \frac{(N - 1)}{N} \\ \alpha_k(\beta) &= \frac{1 - \beta}{1 - \beta^{n_k}} \\ N(\beta) &= \sum \alpha_k(\beta)\end{aligned}\quad (3)$$

where n_k is the number of the k^{th} pattern, and N is the number of samples. Conceptually, β is used to adjust the significance of the number of samples. $N(\beta)$ is the sum of all α_k -s which is corresponded to the β value for each k -pattern. $N(\beta)$ is used for normalization with the number of patterns. However, the work from Cui et al.² ignores the negative patterns into the weight calculations, this dismissed the very important variables because the negative patterns from negative classes are commonly to find in the medical images classification problem.

The DenseNet-121 network. DenseNet-121 is popular to perform classification³ with some other methods^{3,4,6,12} which use ResNet¹³. DenseNet¹⁴ and ResNet¹³ utilize different skip-connection approaches to pass features from previous layers to later layers. ResNet¹³ performs a summation of features for the skip-connections while DenseNet¹⁴ performs concatenation from features. After the input layer, DenseNet utilises 7×7 convolution in a stride 2 mode and it uses 3×3 max pooling also in stride 2 mode. Then it concatenate features in the first Dense block. There are four Dense blocks in DenseNet, each Dense block at least consists of six consecutive of a 1×1 convolution layer followed by a 3×3 convolution layer. The numbers of these consecutive 1×1 and 3×3 layers in Dense blocks are depend on the types of DenseNet which are either 121,169,201 or 264 layered DenseNet, but all of these DenseNet configurations have four Dense blocks and the differences are only in the number of consecutive convolution layers within a Dense block. The concatenated features from a Dense block in DenseNet are then downsampled through a transition layer. The transition layer consists of a 1×1 convolutional layer and a 2×2 average pool layer in stride 2 mode. A Dense block in DenseNet is followed by a transition layer consecutively. ChexNet by Rajpurkar et al.¹⁵ initiates the popularity of DenseNet-121 as the backbone

Layers	Output size	DenseNet-121		ChexNet	
	112 × 112	7 × 7 CONV stride 2		7 × 7 CONV stride 2	
	56 × 56	Max Pool stride 2		Max Pool stride 2	
Dense block (1)	56 × 56	1 × 1 CONV	× 6	1 × 1 CONV	× 6
		3 × 3 CONV		3 × 3 CONV	
Transition (1)	56 × 56	1 × 1 CONV		1 × 1 CONV	
	28 × 28	2 × 2 Avg Pool stride 2		2 × 2 Avg Pool stride 2	
Dense block (2)	28 × 28	1 × 1 CONV	× 12	1 × 1 CONV	× 12
		3 × 3 CONV		3 × 3 CONV	
Transition (2)	28 × 28	1 × 1 CONV		1 × 1 CONV	
	14 × 14	2 × 2 Avg Pool stride 2		2 × 2 Avg Pool stride 2	
Dense block (2)	14 × 14	1 × 1 CONV	× 24	1 × 1 CONV	× 24
		3 × 3 CONV		3 × 3 CONV	
Transition (3)	14 × 14	1 × 1 CONV		1 × 1 CONV	
	7 × 7	2 × 2 Avg Pool stride 2		2 × 2 Avg Pool stride 2	
Dense block (4)	7 × 7	1 × 1 CONV	× 16	1 × 1 CONV	× 16
		3 × 3 CONV		3 × 3 CONV	
Classification layer	1 × 1	7 × 7 GLOBAL AVG POOL		7 × 7 GLOBAL AVG POOL	
		1000D SOFTMAX		14D SIGMOID	

Table 1. The layer comparison DenseNet-121 and ChexNet.

network to perform the Chest X-Ray classification. ChexNet¹⁵ consists the sigmoid functions in the last layer. ChexNet changes the output dimension of the final classification layer of DenseNet-121 from 1024 dimension of softmax output into 14 dimension of sigmoid functions. The change from 1024 to 14 nodes reflects the number of classification's labels in the Chest X-Ray dataset³. Table 1 depicts the layer-differences between ChexNet¹⁵ and DenseNet¹⁴.

The proposed weights function and network architecture. The normalization of α_k formulated in Eq. (4) is used to weight the k^{th} pattern:

$$\widetilde{\alpha}_k(\beta) = \frac{C}{N(\beta)} \cdot \alpha_k(\beta) \quad (4)$$

where C is the number of class. Although Cui et al.² proposed the grid search to determine β based on their formulation, the separable weights of a positive and negative pattern have not been addressed. In this paper, we integrate the separability of positive and negative patterns into the loss-function in order to improve the classification capability of Cui et al's approach. The hypotheses address the importance of both positive and negative pattern weights to improve end-to-end training.

$$\omega_{k+} = \widetilde{\alpha}_k(\beta) \quad (5)$$

where ω_{k+} are the weights for positive samples of the k^{th} pattern. Equation 5 is an elaboration point between² and our proposed method. We deliberately assign α_k to each sample in k^{th} pattern based on the specified ω_{k+} weights. The work² emphasized the importance of effective samples to define the weights and we have two types of weights ω_{k+} and ω_{k-} come into the proposal. In our proposed approach, $\widetilde{\alpha}_k(\beta)$ from² attempts to determine the weights of only the positively labeled samples, which is given in Eq. (5). Also, we determine the weight of the negative patterns:

$$\omega_{k-} = 1 - \omega_{k+} \quad (6)$$

Experimental results evaluate the performance of the proposed weights in Eqs. (5) and (6) to balance the imbalanced samples.

The weighted cross entropy loss. The formulation for cross entropy loss¹⁶ with the proposed weight is:

$$L_{bce}(p) = \sum_{k=1}^C \omega_k (-y_{true}^k \log(p)) \quad (7)$$

$$\omega_k = \begin{cases} \omega_{k-} & \text{if } y_{true}^k = 0 \\ \omega_{k+} & \text{if } y_{true}^k = 1 \end{cases}$$

where y_{true}^k are the ground-truth labels for each samples in pattern k . To perform the experiments in Section 4, we set the $\omega_{k-} = \omega_{k+}$ for a particular case, the case where we want to see the outcome from Cui et al's² formulation

Stage	Operator	Resolution	Channels	Layers
i	$\mathcal{F}_i$	$\mathcal{H}_i \times \mathcal{W}_i$	$\mathcal{C}_i$	$\mathcal{L}_i$
1	Conv 3×3	224×224	32	1
2	MBConv1, $k3 \times 3$	112×112	16	1
3	MBConv6, $k3 \times 3$	112×112	24	2
4	MBConv6, $k5 \times 5$	56×56	40	2
5	MBConv6, $k3 \times 3$	28×28	80	3
6	MBConv6, $k5 \times 5$	28×28	112	3
7	MBConv6, $k5 \times 5$	14×14	192	4
8	MBConv6, $k3 \times 3$	7×7	320	1
9	Conv 1×1 and pooling and FC	7×7	1280	1

Table 2. The EfficientNet-B0 layer⁵.

into the dataset³ classification problem. The Cross Entropy loss use softmax output by default, whereas the Binary Cross Entropy loss use sigmoid output.

The weighted focal loss. The formulation for focal loss with the proposed weight is:

$$L_{foc}(p) = \sum_{k=1}^C \omega_k (-\alpha (1-p)^{\gamma} y_{true}^k \log(p)) \quad (8)$$

$$\omega_k = \begin{cases} \omega_{k-} & \text{if } y_{true}^k = 0 \\ \omega_{k+} & \text{if } y_{true}^k = 1 \end{cases}$$

The proposed focal-loss attempts to weight both the easy-hard samples and the positive-negative patterns, which are not addressed by Cui et al.'s approach². The proposed focal loss also suits the multiclass classification problem. There is no existing focal-loss method which addresses both effective number of samples and positive-negative patterns weighting.

The progressive image resizing. Progressive image resizing is the procedure to train a single deep network architecture with incremental input sizes in multiple stages of training. The first stage trains the network with the default image size for the network and then followed by the next stage which utilises the bigger size images and the use of the best performance of the pre-trained model from the previous stage. There is no formal definition of the exact number of steps, but the classification performance will improve to some extents and then saturates, and gain diminishes; this is very specific to the classification problems. We report that the third stage of training with progressive image resizing did not improve the performance of the existing Chest X-Ray classification problem. Another functionality from the progressive-image-resizing is to provide another form of augmentation. It (re)trains the model with the augmentations of different sized-inputs. Several works^{17–20} of literature mention that augmentation is a proven method to reduce overfitting. We need to have our final model is risk-free from overfitting, and the two-stage training is our approach to perform that. In summary, we perform the two-stage training to achieve two aims: (1) to improve the classification accuracy, (2) to prevent overfitting.

EfficientNet. The recent work by Tan et al.⁵ introduced EfficientNet. It proposed a formulation to perform grid-search among three prominent aspects of the deep network's architecture: depth, width and input resolution. The depth defines the number of layers, the width defines the number of nodes for each layer, and the input resolution defines the size of the input images. The compound scaling from those three components are then composed into different architectures from EfficientNet-B0 into EfficientNet-B7. The networks use the mobile inverted bottleneck layers similar to^{21,22}, the layers then concatenated to squeeze-excitation layer²³. The ReLu6 function is capped at the magnitude of 6; it was used in MobileNetV2²². However, EfficientNet replaces the use of ReLu6 with Swish. Equation 9 shows the difference among the ordinary ReLu function, the ReLu6²⁴ and the Swish activation function:

$$\begin{aligned} ReLu(x) &= \max(0, x) \\ ReLu6(x) &= \min(\max(0, x), 6) \\ Swish(x) &= x ReLu(x) \end{aligned} \quad (9)$$

The layers of EfficientNet-B0 are depicted in Table 2. The further scaling of EfficientNets B0 into B7 are then defined by the grid-search formula as reported in⁵. After the input layer the EfficientNet use 3×3 spatial convolutional layer in stride 2 mode, then it uses MBConv1 the linear bottleneck and inverted residual layer²². After the MBconv1 layer the network has six consecutive MBConv6 layers with various 3×3 and 5×5 kernel as listed in Table 2. Each MBConv6 has three consecutive layers consist of a 1×1 convolutional layer, a 3×3 or 5×5 depth-wise convolutional layer and another 1×1 convolutional layer. Each MBConv1 has two consecutive layers

Figure 1. Training distribution from official split.

consist of a 3×3 depth-wise convolutional layer and another 1×1 convolutional layer. The final layer consists of 1×1 convolutional, the global average pooling and 1280 nodes of a fully connected layer. Following the previous modification of DenseNet-121 into the specific implementation of the Chest X-Ray³ classification problem, we also modify the final output layer from 1280 nodes into 14 nodes.

The performance evaluation. Suppose we want to have a better perspective about the algorithm performance; we need to apply different metrics to evaluate the results. We apply the AU-PRC (Area Under Precision-Recall Curve) metric for further evaluation; the metric has a different characteristic than AU-ROC (Area Under Receiver-Operating-Characteristics). In the term of baseline, AU-ROC has a fixed baseline of .50 for random classifier and 1 for the perfect classifier, respectively^{25,26}. In contrast, AU-PRC baseline is dynamic since it heavily depends on the ratio between positive and negative samples²⁷. AU-PRC is more sensitive to data distribution. AU-PRC will have the baseline .50 for a random classifier under the equal number of positive and negative samples; when the number of negative samples ten times than positive samples this baseline will decrease to a smaller number .09²⁷. The formulation to calculate the baseline of AU-PRC shown in Eq. (10) is from the literature²⁷.

$$\text{baselineAUPRC} = \frac{\text{positives}}{\text{positives} + \text{negatives}} \quad (10)$$

Suppose we have two classes with an identical value of AU-PRC .50, the interpretation from this particular result will vary for both classes. The .50 AU-PRC is a good result for the class with low positive samples, but it may not be satisfactory for the class with a remarkable number of positive samples.

Experiments and results

The chest X-ray dataset³ is used to evaluate the performance of the proposed method. It contains 112,120 Chest X-Ray images from 30,805 unique patients, and it has multilabel of 14 classes diseases. The image resolution is 1024×1024 with the 8-bit channel. We downsampled the resolution as 224×224 and converted the channel into RGB which can be adopted to our backbone network. Chest X-Ray 14 only consists of frontal-view images. It does not have any lateral-view cases. The number of positive samples for each class is much less than the negative samples as depicted in Fig. 1. In our proposed method, the five hyperparameters β are given in Eq. (11).

$$\beta_1 = 1 - 2.0 \cdot 10^{-6}; \quad \beta_2 = 1 - 2.0 \cdot 10^{-5}; \quad \beta_3 = 1 - 2.0 \cdot 10^{-4}; \quad \beta_4 = 1 - 7.0 \cdot 10^{-4}; \quad \beta_5 = 1 - 2.0 \cdot 10^{-3} \quad (11)$$

where β_2 is determined by Eq. (3). The grid-search determines the other β -s. In the exception of the β_4 , the grid search was performed by changing the β value with standard deviation of 10 from β_2 . The current value of β_4 was chosen because that magnitude is the median between β_3 and β_5 . Also the results obtained by the proposed

Samples	Number
Healthy	60,361
Hernia	227
Pneumonia	1431
Fibrosis	1686
Edema	2303
Emphysema	2516
Cardiomegaly	2776
Pleural thick.	3385
Consolidation	4667
Pneumothorax	5302
Mass	5782
Nodule	6331
Atelectasis	11,559
Effusion	13,317
Infiltration	19,894

Table 3. The imbalance number of samples.

	Wang et al. ³	Yao et al. ¹²	Baseline reproduce ChexNet ¹⁵	Weighted binary cross entropy loss	Baltrusch et al. ⁶	Gündel et al. ⁴	Weighted focal loss $\beta = 0.9998$ DenseNet-121 two-stage	Weighted focal loss $\beta = 0.9998$ EfficientNet-B3 two-stage
Atelectasis	0.700	0.733	0.7541	0.7625	0.763	0.767	0.7820	0.7919
Cardiomegaly	0.810	0.856	0.8787	0.8812	0.875	0.883	0.8845	0.8917
Effusion	0.759	0.806	0.8236	0.8266	0.822	0.806	0.8380	0.8414
Infiltration	0.661	0.673	0.6928	0.6939	0.694	0.709	0.7022	0.7051
Mass	0.693	0.777	0.8053	0.8023	0.820	0.821	0.8329	0.8356
Nodule	0.669	0.724	0.7318	0.7383	0.747	0.758	0.7863	0.8036
Pneumonia	0.658	0.684	0.6980	0.7019	0.714	0.731	0.7338	0.7366
Pneumothorax	0.799	0.805	0.8378	0.8344	0.819	0.846	0.8706	0.8909
Consolidation	0.703	0.711	0.7349	0.7390	0.749	0.745	0.7537	0.7601
Edema	0.805	0.806	0.8345	0.8305	0.846	0.835	0.8534	0.8609
Emphysema	0.833	0.842	0.8666	0.8701	0.895	0.895	0.9413	0.9424
Fibrosis	0.786	0.743	0.7957	0.8040	0.816	0.818	0.8229	0.8408
Pleural thickening	0.684	0.724	0.7456	0.7502	0.763	0.761	0.7970	0.8080
Hernia	0.872	0.775	0.8684	0.8589	0.937	0.896	0.9260	0.9286
Average	0.745	0.761	0.7906	0.7924	0.806	0.807	0.8232	0.8313

Table 4. Comparison results with previous study under the official split.

method is compared to those obtained by the other six methods, Wang et al.³, Yao et al.¹², baseline ChexNet¹⁵, weighted binary cross entropy loss, Balturschat et al.⁶ and Gündel et al.⁴. The comparison is depicted in Table 4.

Backbone network training. Since we use the DenseNet 121¹⁴ as the primary backbone network, the availability of pre-trained ImageNet can be used for the classification. Here we used the pre-trained weights from Imagenet to develop the network. The dataset³ for base metrics, including the same training, validation, test splitting set from⁹. We refer the split-set⁹ as “official split”. Summers⁹ has two groundtruth files as label, they are train_val_list.txt which consists of 86,524 samples and test_list.txt which consists of 25,596 samples. Baltrusch et al.⁶ emphasized that different splitting of dataset³ has significantly impact to the classification performance. Since the splitting of training and test data is exactly the same, the benchmark is fair. Figure 1 and Table 3 show that the class distribution is imbalance since the positive and negative samples are very different. We use a single Titan V with 12 Gb GPU memory to develop the network, where 24 hours with 25 epochs of training are required. We only train the Densenet-121 in single-stage training cycle and do not perform progressive image resizing.

Because we aim to improve the overall classification performance, we also modify the architecture of backbone network from DenseNet-121 into EfficientNet⁵. The approach is mainly to expand the performances from the proposed cost-sensitive loss function into better architecture. We are limited only to the use of the EfficientNet-B0 and the EfficientNet-B3 networks for the experiments. This is mainly because we have achieved the peak of

the computing-resources limits, and to perform the experiments over all the EfficientNet architectures are not feasible at this stage. Consecutive EfficientNets training requires extensive computations; due the scaling of the image sizes, the depth and also the width of the network. In the other hand, the approach of progressive image resizing only take into account the aspect of image sizes into computational resources; it ignores the depth and the width of the network. To train the EfficientNets, we use the Tesla v100 with 32 Gb of GPU memory. For each network, we performed the two-stage training procedure with progressive-image-resizing as previously discussed. On the first stage we train the network with pre-trained model from ImageNet, then on the second stage we train the network with the best performing model from first stage. The important finetune is the size of the image input; the first stage we use the default input size from the network then we doubled the input size on the second stage. This has been implemented with size of 224×224 on the first stage of EfficientNet-B0 and 448×448 on the second stage EfficientNet-B0, also 300×300 on the first stage of EfficientNet-B3 and 600×600 on the second stage EfficientNet-B3. We reduce the batch size into half, from 32 on the first stage to 16 on the second stage. The reduced batch size is mainly to ensure the batched-images for each step on each epoch will fit into the GPU's memory boundary. The two-stage training with progressive-image-resizing has successfully improved $\pm 1\%$ to the classification outputs between first stage and second stage for each model.

The baseline network. We reproduce ChexNet¹⁵ based on²⁸. The experiments performed by our proposed method and the other methods^{3,4,6} are based on the training and test split in⁹ are reported in Table 4. However, Rajpurkar et al.¹⁵ never share the split-set for the public. The use of official split⁹ results in lower performance than reported in Rajpurkar et al.¹⁵. We use the ADAM optimizer as in¹⁵ to develop the neural network of which the optimization is converged at epoch 11. Other researches also used ADAM^{4,6} and stochastic gradient descent³.

Weighted binary cross entropy with effective number of samples. This experiment is an adoption of Cui et al.'s² method into the Chest X-Ray dataset³ classification problem. In Cui et al.'s approach², the balanced weights between positive and negative is not used; the weights are computed based on the effectiveness number of samples. Cui et al.'s² used the Eq. (3) to compute the weights. We perform this experiment to provide evidence of performances which come from² versus the one come from our approach. In this experiment, we use binary-cross-entropy as loss function and combine with the weighting into the loss-function. We set the $\omega_{k-} = \omega_{k+}$ for the implementation of Eq. (7) for this particular case, since² ignores the balanced positives-negatives. The best performance classification for the model is also achieved on epoch 11, similar to Baseline. The comparison results with our other experiments are shown in Table 4. This method perform only slightly better than the baseline with the 79.24% area under ROC curve.

Weighted focal loss with positive and negative pattern. In this experiment, we use the loss function¹¹ which is integrated with the focal loss and the proposed weighting. We choose the value of α value based on¹¹ which is between $[.25, .75]$; we found that $\alpha = 0.5$ and $\gamma = 1$ is the best of focal-loss hyperparameters for our proposed method. We use the RANGER (Rectified Adam and LookAhead) optimizer, which requires a smaller number of training epochs to converge. The optimizer converges at epoch 5. We deliberately assign the two-stage training to prevent overfitting and also to improve the performance. This method achieves 82.32% area under ROC curve with two-stage DenseNet-121 and 83.13% with two-stage EfficientNet-B3. The training time took 71 min for one epoch on the first stage with 32 batch size, 180 min for one epoch on the second stage with 16 batch size. The test-time took 15 min with 8 batch size for the first-stage and took 27 min with 8 batch size for the second-stage.

The intuitive theoretical background and the evidence from experiment. Since part of our approach inherits the strength of focal loss¹¹, and also the class-balanced approach². We can have further theoretical analysis from the proposed approach intuitively based on¹¹ and². The main distinction of focal-loss with binary cross entropy loss is the existence of α and γ parameter. Cui et al. mention on the paper: "the class-balanced term can be viewed as an explicit way to set α in focal loss based on the effective number of samples"². However, Lin et al. also mention "a common method for addressing class imbalance is to introduce a weighting factor $\alpha \in [1, 0]$ for class 1 and $1 - \alpha$ for class - 1"¹¹. We can inference those two statements into our elaboration in Eq. (6). The experiments provide a further evidence for the theory. The improvement from the change of the proposed formula is $\pm 1\%$ under the test-set according to the Table 5. Both experiments were performed with $\alpha = 0.5$ and $\gamma = 1.0$ for the Focal Loss's parameters. Both Tables 5 and 8 use an identical split⁸. The training-set consists of 78,468 images, the validation-sets consists of 11,219 images, the test-sets consists of 22433 images. The training time took 38 min for one epoch with 32 batch size and the test time took 12 min with 8 batch size.

The imbalance metric evaluation.

Table 6 and Fig. 2 show the advancement from the proposed method in comparison with previous work¹⁵ also with the baseline retrieved from the dataset. We calculate the baseline of AU-PRC metric directly from the dataset's distribution of positives and negatives samples with the use of Eq. (10). The bold fonts show the top-scores achieved between a same split-set configuration. The hernia has the lowest number of positive samples in the distribution. Despite being the most minority, the proposed algorithm results from hernia a couple of hundred higher AU-PRC than the baseline; shown in Table 6 and Fig. 2.

Pathology	Focal loss DenseNet-121		Weighted focal loss $\beta = 0.9998$ DenseNet-121	
	Validation	Test	Validation	Test
Cardiomegaly	0.9155	0.9096	0.9092	0.9090
Emphysema	0.9140	0.9178	0.9056	0.9327
Edema	0.9141	0.8851	0.9147	0.8917
Hernia	0.8614	0.9135	0.9067	0.9404
Pneumothorax	0.8896	0.8663	0.8973	0.8749
Effusion	0.8822	0.8762	0.8792	0.8827
Mass	0.8622	0.8430	0.8655	0.8514
Fibrosis	0.8277	0.8219	0.8313	0.8308
Atelectasis	0.8191	0.8079	0.8228	0.8259
Consolidation	0.8247	0.8007	0.8224	0.8043
Pleural thicken.	0.8219	0.7874	0.8214	0.7910
Nodule	0.7823	0.7751	0.7888	0.7756
Pneumonia	0.7722	0.7504	0.7586	0.7698
Infiltration	0.7061	0.7073	0.7113	0.7166
Average	0.8424	0.8330	0.8453	0.8427

Table 5. The improvement of the proposed weight calculation.

Pathology	Reproduce ChexNet ¹⁵		Weighted focal loss $\beta = 0.9998$ EfficientNet-B3 two-stage		Baseline AU-PRC from distribution
	Official	Identical ⁸	Official	Identical ⁸	
Cardiomegaly	0.3288	0.2880	0.3444	0.3127	0.0247
Emphysema	0.3125	0.2948	0.4706	0.4515	0.0224
Edema	0.1497	0.1455	0.2048	0.1835	0.0205
Hernia	0.0785	0.0311	0.4147	0.5372	0.0020
Pneumothorax	0.3683	0.2929	0.4965	0.4242	0.0472
Effusion	0.5012	0.5149	0.5428	0.5439	0.1187
Mass	0.2887	0.2847	0.3355	0.3357	0.0515
Fibrosis	0.0773	0.0886	0.1231	0.1405	0.0150
Atelectasis	0.3208	0.3262	0.3664	0.3859	0.1030
Consolidation	0.1488	0.1479	0.1736	0.1692	0.0416
Pleural thicken.	0.1109	0.1159	0.1831	0.1754	0.0301
Nodule	0.1899	0.2025	0.2595	0.2919	0.0564
Pneumonia	0.0448	0.0381	0.0609	0.0484	0.0127
Infiltration	0.3891	0.3342	0.4067	0.3592	0.1774
Average	0.2364	0.2218	0.3130	0.3114	0.0517

Table 6. The AU-PRC improvement.

Discussion

In order to provide more insights of the effect from different splits into the classification performance, several split-sets have been taken into performance evaluation. The standard procedure is to follow the “official” splits⁹ and we report the results in Table 4. To the best of our knowledge, only⁶ reported the performance evaluation of a random fivefolds cross validation from the Chest X-Ray dataset³ and we report the results from the proposed method in Table 7. There are other split-sets which are considered “non-standard” settings, these splits are from github pages. Ren⁸ is the third party re-implementation of⁷, and also¹⁰ is the third party re-implementation of¹⁵. However, after further investigation⁸ and¹⁰ are under the identical training, validation and testing sets. We report the results with the custom-sets^{8,10} in Table 8. The results in Table 9 are the improvements made in compare with the most recent research⁷. The one-by-one comparison for each disease with latest research⁷ as listed in Table 8. We achieve better performances in compare with the work of Guan et al.⁷ and we propose technically more simple approach to achieve the results. Since the diversity of split-sets is a well-known problem for the dataset’s³ evaluation, the use cross validation is a fair method to follow. Baltruschat et al.⁶ is the only work that reported performing cross-validation to the dataset³, we achieve better performance in fivefolds cross validation experiment than the work of Baltruschat et al.⁶. The class-activation-mapping (CAM) method^{29,30} visualizes the

Figure 2. The area under precision-recall curve.

Pathology	Baltrusch et al. ⁶	Weighted focal loss $\beta = 0.9998$ EfficientNet-B3 two-stage
Cardiomegaly	89.8 ± 0.8	90.6 ± 2.4
Emphysema	89.1 ± 1.2	94.6 ± 1.2
Edema	88.9 ± 0.3	90.3 ± 0.9
Hernia	89.6 ± 4.4	92 ± 1.3
Pneumothorax	85.9 ± 1.1	91.2 ± 1.2
Effusion	87.3 ± 0.3	88.5 ± 0.5
Mass	83.2 ± 0.3	86.9 ± 1.1
Fibrosis	78.9 ± 0.5	82.2 ± 2.5
Atelectasis	79.1 ± 0.4	83.3 ± 0.7
Consolidation	80.0 ± 0.7	80.9 ± 0.5
Pleural thicken.	77.1 ± 1.3	82.7 ± 1.3
Nodule	75.8 ± 1.4	81.7 ± 1.4
Pneumonia	76.7 ± 1.5	77 ± 1.9
Infiltration	70.0 ± 0.7	72.8 ± 4.5
Average	82.2 ± 1.1	85.3 ± 0.6

Table 7. Results from fivefolds cross-validation.

discriminative-features from the deep-network's last layer in the form of heatmap localisation, the more heatmap visualization match the groundtruth bounding-box from dataset means the network has better understanding of the images. We visualize our classification performances with heatmap from CAM method in Table 10. We

Pathology	Third party ⁸ of Guan et al. ⁷	Weighted focal loss $\beta = 0.9998$ DenseNet-121		Weighted focal loss $\beta = 0.9998$ EfficientNet-B3	
		Stage 1	Stage 2	Stage 1	Stage 2
Cardiomegaly	0.9097	0.9090	0.9142	0.9137	0.9144
Emphysema	0.8905	0.9327	0.9425	0.9471	0.9558
Edema	0.9185	0.8917	0.8930	0.9021	0.9071
Hernia	0.9064	0.9404	0.9427	0.9357	0.9409
Pneumothorax	0.8794	0.8749	0.8854	0.9003	0.9092
Effusion	0.8843	0.8827	0.8885	0.8899	0.8923
Mass	0.8707	0.8514	0.8568	0.8596	0.8669
Fibrosis	0.8208	0.8308	0.8458	0.8526	0.8657
Atelectasis	0.8225	0.8259	0.8307	0.8350	0.8397
Consolidation	0.8210	0.8043	0.8115	0.8124	0.8208
Pleural thicken.	0.8127	0.7910	0.8011	0.8041	0.8136
Nodule	0.7691	0.7756	0.8144	0.8043	0.8293
Pneumonia	0.7614	0.7698	0.7726	0.7721	0.7703
Infiltration	0.7006	0.7166	0.7178	0.7297	0.7363
Average	0.8405	0.8426	0.8512	0.8542	0.8616

Table 8. Identical split comparison⁸.

(A)								
Name	Hernia	Pneumonia	Fibrosis	Edema	Emphysema	Cardiomegaly	Pleural Thick.	Pneumothorax
Rate	+3.45%	+0.89%	+4.49%	−1.14%	+6.53%	+0.47%	+0.09%	+2.98%
(B)								
Consolidation	Mass	Nodule	Atelectasis	Effusion	Infiltration	Average		
−0.02%	−0.38%	+6.02%	+1.72%	+0.80%	+3.57%	+2.10%		

Table 9. The improvement rate.

obtain the bounding-boxes as the annotation groundtruth for only 8 (eight) classes which are available from the file BBox_List_2017.csv⁹. The annotations consists of 984 images, and the number of samples for each class is not distributed evenly. Table 10 shows that the networks which are equipped with the proposed method read the area of the disease better than the baseline.

We found the third party re-implementation⁸ reported lower performances than reported in the paper⁷.

Guan et al.⁷ did not provide the official code and split-sets. The critical classification problems for the dataset³, different splits will lead to different performances⁶.

Conclusion

We proposed an aggregate of novel weighting function to formula the focal-loss function in complement with the two-stage training of EfficientNet, a state-of-the-art neural network architecture. We aim to improve the classification capability. Existing approaches of weighting function did not address the sample characteristics of both the positive-negative and easy-hard. The proposed weighting function attempts to improve the classification capability by address both the sample characteristics, which are ignored by the existing methods. The proposed approach provides a better decision boundary to the multiclass classification problem since the proposed approach addresses the imbalances of both positive-negative and hard-easy samples, also the use of recent network architecture scale-up the performances better. The proposed approach is able to improve the classification rates by 2.10% than the latest research's outputs which is measured in the area under the receiver operating characteristic curve (AUROC). The proposed method also achieved state-of-the-art results under three distinct experiments setup, currently the results are the best improvements for the Chest X-Ray dataset being used. Since the proposed approach only addresses multiclass classification problem and multilabel classifications are not tackled, future research will target on multilabel problems. The proposed approach will be further validated.

Pathology	Baseline Reproduce ChexNet ¹⁵	Weighted Focal Loss DenseNet-121	Weighted Focal Loss EfficientNet-B3
Atelectasis
Cardiomegaly
Effusion
Infiltration
Mass
Nodule
Pneumonia
Pneumothorax

Table 10. The heatmap from different methods.

Data availability

The dataset for this research is available publicly in the Chest X-ray NIHCC repository: <https://nihcc.app.box.com/v/ChestXray-NIHCC>.

Code availability

The code for to reproduce the research is available publicly at the repository: <https://github.com/bayu-ladom-ipok/weOpen>.

Received: 1 June 2020; Accepted: 6 January 2021

Published online: 05 February 2021

References

- Johnson, J. M. & Khoshgoftaar, T. M. Survey on deep learning with class imbalance. *J. Big Data* **6**, 27. <https://doi.org/10.1186/s40537-019-0192-5> (2019).
- Cui, Y., Jia, M., Lin, T., Song, Y. & Belongie, S. J. Class-balanced loss based on effective number of samples. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16–20, 2019*, Vol. 1, 9268–9277 (2019).
- Wang, X. *et al.* Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3462–3471 (2017).
- Gündel, S. *et al.* 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*, 757–765. <https://doi.org/10.1007/978-3-030-13469-3> (2018).
- Tan, M. & Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In (eds Chaudhuri, K. & Salakhutdinov, R.) *Proceedings of the 36th International Conference on Machine Learning*, vol. 97 of *Proceedings of Machine Learning Research*, 6105–6114 (PMLR, Long Beach, 2019).
- Baltruschat, I. M., Nickisch, H., Grass, M., Knopp, T. & Saalbach, A. Comparison of deep learning approaches for multi-label chest x-ray classification. *Sci. Rep.* **9**, 6381. <https://doi.org/10.1038/s41598-019-42294-8> (2019).
- Guan, Q. *et al.* Thorax disease classification with attention guided convolutional neural network. *Pattern Recogn. Lett.* **131**, 38–45. <https://doi.org/10.1016/j.patrec.2019.11.040> (2020).
- Ren, I. This is a reimplementation of ag-cnn: <https://github.com/ien001/ag-cnn> (2019).
- Summers, R. <https://nihcc.app.box.com/v/chestxray-nihcc> (2017).
- Weng, X., Zhuang, N., Tian, J. & Liu, Y. A pytorch reimplementation of chexnet: <https://github.com/arnoweng/chexnet> (2017).
- Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection 2999–3007. <https://doi.org/10.1109/ICCV.2017.324> (2017).
- Yao, L., Prosky, J., Poblentz, E., Covington, B. & Lyman, K. Weakly supervised medical diagnosis and localization from multiple resolutions. *CoRR* (2018). [arxiv:1803.07703](https://arxiv.org/abs/1803.07703).
- He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016*, Vol. 1, 770–778. <https://doi.org/10.1109/CVPR.2016.90> (2016).
- Huang, G., Liu, Z., van der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *CVPR*, 2261–2269 (IEEE Computer Society, 2017).
- Rajpurkar, P. *et al.* Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *CoRR* (2017). [arxiv:1711.05225](https://arxiv.org/abs/1711.05225).
- Zhang, Z. & Sabuncu, M. R. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3–8 December 2018, Montréal, Canada*, 8792–8802 (2018).
- Cao, K., Wei, C., Gaidon, A., Aréchiga, N. & Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8–14 December 2019* (eds Wallach, H. M. *et al.*) 1565–1576 (Canada, Vancouver, BC, 2019).
- Shorten, C. & Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *J. Big Data* **6**, 60. <https://doi.org/10.1186/s40537-019-0197-0> (2019).
- Wong, S. C., Gatt, A., Stamatescu, V. & McDonnell, M. D. Understanding data augmentation for classification: When to warp? In *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 1–6 (2016).
- Peng, X., Tang, Z., Yang, F., Feris, R. S. & Metaxas, D. N. Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. *CoRR* (2018). [arxiv:1805.09707](https://arxiv.org/abs/1805.09707).
- Tan, M. *et al.* Mnasnet: Platform-aware neural architecture search for mobile. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16–20, 2019*, Vol. 1, 2820–2828. <https://doi.org/10.1109/CVPR.2019.00293> (Computer Vision Foundation / IEEE, 2019).
- Sandler, M., Howard, A. G., Zhu, M., Zhmoginov, A. & Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 4510–4520 (IEEE Computer Society, 2018).
- Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition* (2018).
- Krizhevsky, A. Convolutional deep belief networks on cifar-10 (2010).
- Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36. <https://doi.org/10.1148/radiology.143.1.7063747> (1982).
- Cortes, C. & Mohri, M. AUC optimization vs. error rate minimization. In *Advances in Neural Information Processing Systems 16 [Neural Information Processing Systems, NIPS 2003, December 8–13, 2003, Vancouver and Whistler, British Columbia, Canada]*, (eds Thrun, S. *et al.*) 313–320 (MIT Press, 2003).
- Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE* **10**, 1–21. <https://doi.org/10.1371/journal.pone.0118432> (2015).
- Chou, B. This project is a tool to build chexnet-like models, written in keras: <https://github.com/brucechou1983/chexnet-keras> (2018).
- Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2921–2929. <https://doi.org/10.1109/CVPR.2016.319> (2016).
- Yang, W. *et al.* Towards rich feature discovery with class activation maps augmentation for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16–20, 2019*, Vol. 1, 1389–1398. <https://doi.org/10.1109/CVPR.2019.00148> (2019).

Acknowledgements

This work is supported and funded by Ministry of Religious Affair (MORA) Scholarship, Republic of Indonesia. We also very thankful for NIH Clinical Center for their publicly released dataset.

Author contributions

Author contributions statement B.A.N. performed the works for the manuscript.

Competing interests

B.A.N. is a MORA Scholarship recipient from Sunan Ampel Islamic State University Surabaya—Indonesia, and he is also a research student in Curtin University Bentley - Australia. The scholarship recipient's primary email addresses are: bayu@uinsby.ac.id and bayu.lecture@gmail.com. There are no potential conflict of interests for the manuscript.

Additional information

Correspondence and requests for materials should be addressed to B.A.N.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021