Sontje Ihler* and Felix Kuhnke

AUC margin loss for limited, imbalanced and noisy medical image diagnosis – a case study on CheXpert5000

https://doi.org/10.1515/cdbme-2023-1165

Abstract: The AUC margin loss is a valuable loss function for medical image classification as it addresses the problems of imbalanced and noisy labels. It is used by the current winner of the CheXpert competition. The CheXpert dataset is a large dataset (200k+ images), however datasets in the range of 1k-10k medical datasets are much more common. This raises the question if optimizing AUC margin loss also is effective in scenarios with limited data. We compare AUC margin loss optimization to binary cross-entropy on limited, imbalanced and noisy CheXpert5000, a subset of CheXpert dataset. We show that AUC margin loss is beneficial for limited data and considerably improves accuracy in the presence of label noise. It also improves out-of-box calibration.

Keywords: Computer-Aided-Diagnosis, Deep Learning, CheXpert, Noisy Labels, Label Imbalance

1 Introduction

Computer-aided diagnosis (CAD) from X-ray, CT, and MRI images is becoming a valuable support tool for diagnosing and treating clinical pathologies. With the success of deep learning for image classification more and more applications are within reach of becoming standard tools for radiologist. The objective of binary/multi-label image classification is the maximization of the area under the ROC curve (AUC or AUROC). Albeit the common practice to optimize these classification tasks using the binary cross entropy (BCE), it would seem to be the winning strategy to directly maximize the AUC score. Due to the nature of AUC being a score loss, it is not possible to optimize the AUC score using gradient descent (required for neural network optimization). Therefore, Yuan et al. recently proposed the AUC margin (AUCM) loss which is a surrogate loss to indirectly optimize the AUC score [1]. The authors provide extensive ablation studies of their work on common computer vision benchmarks like CIFAR10/100, and STL10 and show that AUCM improves accuracy in the presence of imbalanced and noisy labels, which are very common problems in CAD, and they are currently leading the CheXpert competition [2]. They combined 25 models to win this competition and added additional label smoothing to handle label uncertainty (and therefore also label noise). As they do not include an ablation study on CheXpert in their work, the effects of the various components, including AUCM, are not clear. In addition, the original CheXpert dataset (200k+ images) is huge compared to common medical datasets, where 1k-10k datasets are much more common. This raises the questions of the impact of AUCM optimization without additional noise handling strategies and if AUCM generalizes to a limited data scenario. We therefore investigate AUCM in more details in a limited data scenario and study AUCM on the CheXpert5k [3], a subset of the Stanford CheXpert dataset [2], meant for the study of limited data scenarios. **Our contributions are (1) a comparison study of AUCM to BCE for limited, noisy and imbalanced X-ray image classification and (2) we show that AUCM is beneficial also for limited data especially for noisy labels.**

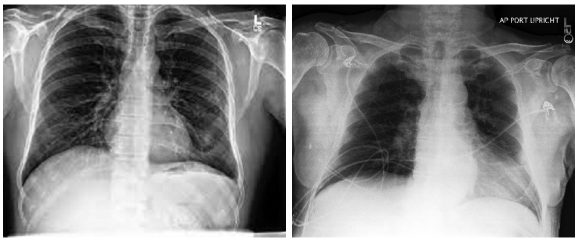


Fig. 1: Sample images from the X-ray image classification CheXpert [2] dataset.

2 CheXpert5k and label noise

CheXpert5000 (CheXpert5k) [3] is a subset of the public CheXpert dataset [2] to explicitly study limited data scenarios. It provides five different training and validation sets sampled from the official CheXpert dataset with 5000 training samples each. CheXpert5k only has samples in frontal and AP view. The label statistics of the 5k training sets are similar to the official CheXpert training set. The official CheXpert validation

*Corresponding author: Sontje Ihler, Leibniz Universität Hannover, Institut für Mechatronische Systeme e-mail: sontje.ihler@imes.uni-hannover.de
Felix Kuhnke, Leibniz Universität Hannover, Institut für Informationsverarbeitung

Open Access. © 2023 The Author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.

Tab. 1: Label distribution of one of the five CheXpert5k (imbalanced) training sets [3] and a summary of the results of the original CheXpert paper [2] if label mapping of uncertain labels (u) improves model performance compared to ignoring the uncertain labels.
u-zeros: uncertain labels are mapped to 0; u-ones: uncertain labels are mapped to 1; u-3-classes: optimized as 3-class problem.
++: strong improvement, +: slight improvement, –: strong decrease, -: slight decrease

Label	Atelectasis	Cardiomegaly	Consolidation	Edema	Pleural Effusion
1	813	637	363	1443	2086
0	16	146	368	367	485
u	810	178	666	332	271
not labeled	3361	4039	3603	2858	2158
1	16.3%	12.7%	7.3%	28.9%	41.7%
0	0.3%	2.9%	7.4%	7.3%	9.7%
u	16.2%	3.6%	13.3%	6.6%	5.4%
not labeled	67.2%	80.8%	72.1%	57.2%	43.2%
u-zeros	-	+	-	-	+
u-ones	++	+	–	+	+
u-3-Class	+	++	o	-	+

set is used as test set. It contains 235 images, all manually annotated by three medical experts.

The original CheXpert is a large medical dataset of 200k+ X-ray images which has been automatically annotated using neural language processing on patient files. It provides labels for 14 pathologies. It was introduced as part of the CheXpert Challenge which benchmarks the performance on five pathologies, see Table 1. As CheXpert was annotated using a neural network, the labels are not only 0 and 1 for each pathology being present or not but also u for uncertain and _ for unlabeled to address uncertainty of the language model or abstinence of the predefined pathologies in the patients' files.

Common practice for supervised training on the CheXpert dataset is to map category _ to 0, and map u to 0 or 1 depending on the pathology. This approach is based on the original work [2], where they study how the model's performance improved or decreased for different mappings of the uncertain labels. We provide a summary in Table 1. The mapped uncertain and unlabeled labels will generally have a higher noise level than the confident labels 1 and 0.

3 Experiments

In our experiments we compare models trained on BCE to models trained with AUC margin (AUCM) loss. We explicitly study AUCM loss for label imbalance and label noise. We perform our experiments on CheXpert5k.

We provide area under the receiver operating characteristic curve (AUC or AUROC), area under the precision-recall curve (AUPRC) and the expected calibration error (ECE) on the official CheXpert validation set. The ECE is computed from 10 bins. We report the mean and standard deviation over five runs with different training/validations sets. We also provide the average amount of images the model has seen until convergence (image iterations).

We provide results for the Big Transfer Model BiT50x1 [4] which is a ResNet50 variant with optimized architecture for transer learning and trained on a larger dataset. It significantly outperforms ResNet50 on CheXpert5k (with identical model capacity and memory consumption) and was recommended as a drop-in substitute for ResNet50 [3]. We also provide results for ResNet50 as an established baseline model for comparison.

3.1 Imbalanced and noisy data

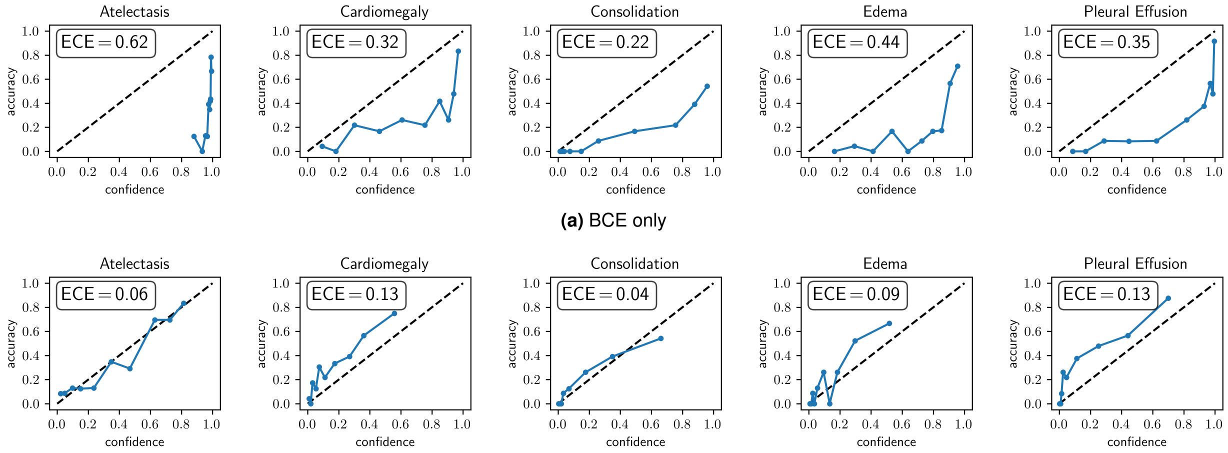
We perform two experiments to investigate AUCM on imbalanced data, and imbalanced data with noise.

Imbalance only: For our first experiment we create imbalanced but non-noisy training sets. We do that by using only the confident labels from CheXpert5k (labels 0 and 1). We mask all other labels, to have no influence on the loss computation. The first training set then consists of 5000 images and 7599 labels, i. e., all confident labels for all five classes (multi-label classification). The label statistics can be seen in Table 1. AUROC, AUPRC and ECE are computed for each class and averaged with equal weights.

Imbalance and label noise: For our second experiment we picked the class with the worst performance when including the uncertain labels in previous studies [2] which is *Consolidation*, see Table 1. Results for Consolidation did not improve when mapping the uncertain labels (u) neither to 0 nor to 1. We can therefore assume that these labels are very noisy when we map them to a single label. In this experiment we map the uncertain labels to 1 and all unlabeled samples to 0

Tab. 2: Results for imbalance only experiments: multi-label AUCM trained only on confident labels i.e. 1 and 0 (approx. no noise). Results are averaged over all five training sets of CheXpert5k and we provide standard deviation. We do see some instability in the training with the original training protocol BCE-3+AUCM where some models degenerate.

Model	Training Stategy	Image Iterations	AUROC	AUPRC	ECE
Bit-50x1	BCE only	34k	0.8225±.0200	0.5982±.0340	0.404±.0234
Bit-50x1	BCE-3+AUCM	15k+68k	0.8208±.0135	0.6061±.0166	**0.1245±.0124**
Bit-50x1	BCE-c+AUCM	34k+48k	**0.8231±.0151**	**0.6055±.0311**	0.2028±.0252
ResNet50	BCE only	184k	0.8007±.0081	0.571±.0224	0.3823±.0242
ResNet50	BCE-c+AUCM	184k+38k	0.7941±.0037	0.5679±.0167	0.2325±.0177



(a) BCE only

(b) BCE-c+AUCM

Fig. 2: Class-wise reliability/calibration plots for BiT-50x1 tested on CheXpert validation set (manual, hence non-noisy labels). AUCM optimization considerably improves calibration for all pathologies.

(following common protocol). We therefore have training sets with 5000 images and 5000 (noisy) labels. Benign to malign ratio is 20.6 to 79.4, i. e., a so called imratio of approx. 0.2.

3.2 BCE vs. AUCM

AUCM training proposes a two-step optimization protocol [1]: first, a model is initially pretrained with BCE loss for a small amount of epochs on the target data, and then fine-tuned with AUCM loss on the target data. For our experiments CheXpert5k training sets are used as target data.

In our study we performed two different training protocols: (BCE-3+AUCM) following the proposed protocol and (BCE-c+AUCM) a variation where we alter the BCE pretraining. In both protocols AUCM fine-tuning is identical and always performed until convergence of validation loss. As suggested, the classifier is reset before AUCM optimization [1].

BCE-3+AUCM: This first protocol follows the recommendations of [1]. The original work uses 400k image iterations (2 epochs) for 200k-CheXpert. With CheXpert5k, the training converges long before that. We therefore chose a pretraining with 3 epochs to mimic the short pretraining.

BCE-c+AUCM: We found that in some cases BCE-3+AUCM leads to model divergence during AUCM optimization. [1] states that BCE pretraining is essential for good results which leads to the assumption that our pretraining might not be long enough. We therefore adapted our BCE pretraining. In our second protocol BCE optimization is performed until convergence of validation loss.

BCE only: We compare our models to models trained only on BCE (until convergence). These models are identical to the models obtained in the first step of BCE-c+AUCM.

3.3 Implementation and Hyperparameters

Our implementation is based on the timm [5] and libauc [6] library. All models and pretrained weights are from timm.

For BCE training, we use ResNet pretrained on ImageNet and Bit-50x1 pretrained on the larger ImageNet21k and BCE loss. Following [3] we use a batchsize of 32, SGD as opti-

Tab. 3: Results for imbalance and noise experiments: Single-class AUCM optimization of BiT50x1 on CheXpert5k for Consolidation only. Uncertain labels are mapped to 1 and not labeled samples are mapped to 0 for training which presumably leads to a high amount of noisy labels in the training set.

Model	Training Strategy	Image Iterations	AUROC	AUPRC	ECE
Bit-50x1	BCE only	34k	0.8405±.0172	0.4252±.0483	0.0895±.0200
Bit-50x1	BCE-3+AUCM	15k+8k	0.8258±.0304	0.3872±.0720	0.0604±.0152
Bit-50x1	BCE-c+AUCM	34k+42k	0.**8622±.0143**	**0.4730±.0207**	**0.0531±.0153**

mizer with momentum 0.9, weight decay 2e-5, and learning rate 0.003. During training, a plateau scheduler is used for the learning rate with decay rate 0.1 and patience 10.

For AUCM training, we use the BCE models as described above (pretrained on CheXpert5k). We use the AUCM loss and the corresponding PSEG optimizer from libauc [6]. For the optimizer, we keep the default settings from [6] with weight decay 1e-5 and learning rate 0.003. We again use a batchsize of 32 and a plateau scheduler with decay rate 0.1 and patience 10. We use the same hyperparameters for all runs.

4 Results

We provide results for our imbalance only experiments in Table 2 and Figure 2, and the combined imbalance and noise experiments in Table 3.

Imbalance only: There is a slight improvement (AUROC, AUPRC) in accuracy after AUCM optimization for BiT-50x1 (not for ResNet50). However this change is so small, it is not conclusive. We do see however a drastically improved out-of-box calibration (ECE) while maintaining the same accuracy. The improved calibration can be seen for all pathologies. This is especially interesting for the strongly imbalanced class *Atelectasis* which has only 16 0-labels in all 5000 images. We see that BCE-only Bit-50x1 model (Figure 2a) is less calibrated than in [3] where the models had very good calibration values. We assume this is due to the absence of noise in our training data for this experiment. The noise in the full CheXpert5k might lead to better calibrated models. We found that only short pretraining with BCE (BCE-3+AUCM) leads to deterioration of the model in some cases.

Imbalance and label noise: In our noisy experiment we see a strong improvement in model accuracy with our adapted pretraining strategy (BCE-c+AUCM). AUCM seems to show its strength especially in robustness to label noise. Shorter BCE pretraining (BCE-3+AUCM) unfortunately led to poorer model performance. Calibration is high for all training protocols (also BCE only) right from the start. Impressively, both AUCM optimizations improved out-of-box calibration even further. So overall, BCE-c+AUCM again improved the models' calibration without any loss in accuracy.

5 Conclusion

We find that the AUCM loss is beneficial for our limited data scenario especially in the presence of label noise. BCE-c+AUCM optimization outperforms BCE-only by over 2% in the presence of label noise. We furthermore see improvement in out-of-box calibration for all AUCM optimizations with no loss in accuracy (except for too short BCE-training for Consolidation). We find that an initial longer BCE-training is more robust than a too short training and leads to higher accuracy. We see good results with a fully converged BCE model as a starting point for AUCM optimization. This actually simplifies the original training protocol by removing a hyperparameter.

AUCM optimization from BCE-c models generally resulted in better models than BCE-only models. The concept of AUC optimization is very promising for medical image diagnosis and we would like to see wider application and more in-depth research on this.

6 References

[1] Zhuoning Yuan, Yan Yan, Milan Sonka, and Tianbao Yang. Large-scale robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification. In *Proceedings of ICCV*, pages 3040–3049, 2021.

[2] Jeremy Irvin, Pranav Rajpurkar, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of AAAI*, volume 33, pages 590–597, 2019.

[3] Sontje Ihler, Felix Kuhnke, and Svenja Spindeldreier. A comprehensive study of modern architectures and regularization approaches on chexpert5000. In *Proceedings of MICCAI*, pages 654–663. Springer, 2022.

[4] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Proceedings of ECCV*, pages 491–507. Springer, 2020.

[5] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

[6] Zhuoning Yuan, Zi-Hao Qiu, Gang Li, Dixian Zhu, Zhishuai Guo, Quanqi Hu, Bokun Wang, Qi Qi, Yongjian Zhong, and Tianbao Yang. Libauc: A deep learning library for x-risk optimization., 2022.