Large-scale Robust Deep AUC Maximization: A New Surrogate Loss and Empirical Studies on Medical Image Classification

Zhuoning Yuan1, Yan Yan , Milan Sonka1 , Tianbao Yang

Lipisha Chaudhary

Challenges

- Data Issues:
  - Imbalanced Data
  - Not enough data (Small Datasets)
  - Hidden biases
- Data Imbalance:
  - Is a common issue in real world examples
  - Can cause serious problems like degradation of model performance.
  - With sensitive data (especially medical) ethical issues can come into picture

Area Under the ROC Curve (AUC)

- Widely used metric for measuring classification performance
- AUC is more suitable for handling imbalanced data distribution since maximizing AUC aims to rank the prediction score of any positive data higher than any negative data
- In medical classification tasks the AUC score is the default metric for evaluating

Area Under the ROC Curve (AUC)

- An Example of Sensitivity of AUC

Example 1		Example 2		Example 3	
Prediction	Ground Truth	Prediction	Ground Truth	Prediction	Ground Truth
0.9	1	0.9	1	0.9	1
0.8	1	**0.41**(↓)	1	**0.41**(↓)	1
0.7	1	0.7	1	**0.40**(↓)	1
0.6	0	0.6	0	**0.49**(↓)	0
0.6	0	**0.49**(↓)	0	**0.48**(↓)	0
0.47	0	0.47	0	0.47	0
0.47	0	0.47	0	0.47	0
0.45	0	0.45	0	0.45	0
0.43	0	0.43	0	0.43	0
0.42	0	0.42	0	0.42	0
⋮	⋮	⋮	⋮	⋮	⋮
0.1	0	0.1	0	0.1	0
Acc=0.92		Acc=0.92 (—)		Acc=0.92 (—)	
AUC=1.00		AUC=**0.89** (↓)		AUC=**0.78** (↓)	

- Example 1: shows that all positive instances rank higher than negative instances and two negative instances are misclassified to positive class.
- Example 2: 1 positive and 1 negative instances are misclassified.
- Example 3: 2 positive instances are also misclassified as negative class.
- AUC drops dramatically as the ranks of positive instances drop but meanwhile Accuracy remains unchanged

Problem to be solved

- Make deep learning paradigms more practical and efficient for real-word applications (i.e. medical image classification)
- Previous works achieved great results on large-scale medical dataset for breast cancer screening
  - Used conventional Convolutional Neural Network to classify begin/malignant type of cancer
  - AUC was used to evaluate the performance
- "Can we design a generic method that can further improve the performance of DL on these medical datasets without relying on domain knowledge"?
  - In other words, can it provide good performance on large-scale medical dataset by maximizing AUC

Solution Proposed

- Instead of minimizing the cross-entropy loss, maximize the AUC score
- New margin-based min-max surrogate loss function for AUC score
    - More robust than the conventional AUC Square function
- Extensive empirical studies of proposed method on four difficult medical image classification tasks:
    - Classification of chest x-ray images for identifying many threatening diseases
    - Classification of images of skin lesions for identifying melanoma
    - Classification of mammogram for breast cancer screening,
    - Classification of microscopic images for identifying tumor tissue.

Solution Proposed

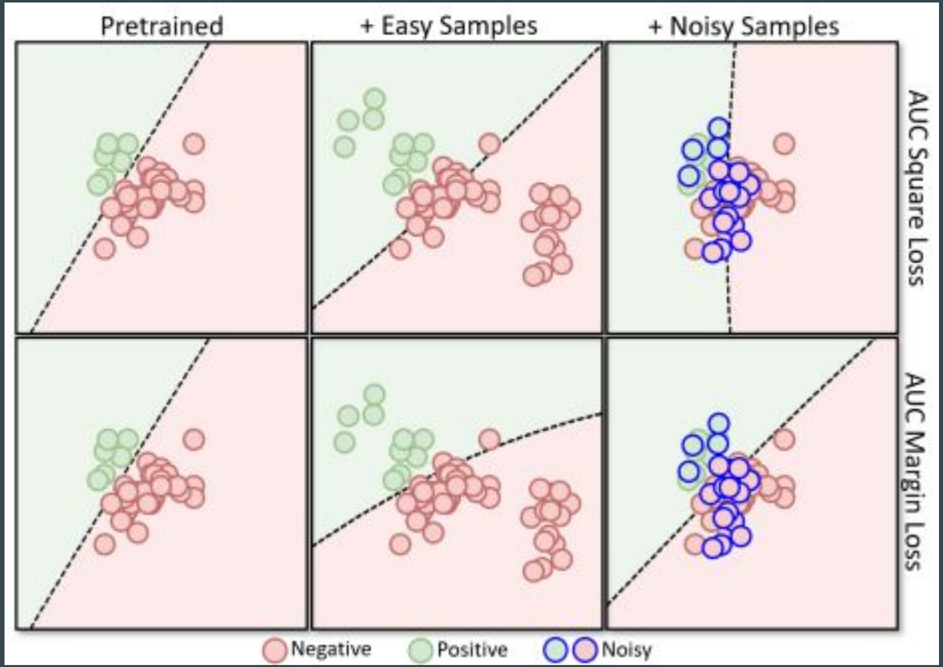
- **D**eep **A**UC **M**aximization (DAM):
  - A deep learning method by AUC maximization
  - DAM is used exclusively for medical image classification
  - AUC is a standard performance measure for medical image classification
    - Directly optimizing AUC could achieve a better performance for learning a deep neural network
- AUC maximization is much more challenging than minimizing misclassification error since AUC is much more sensitive to model change.

AUC Square Loss

- The foremost challenge for AUC maximization is to determine a surrogate loss for the AUC score.
- A naive way is to use a pairwise surrogate loss based on the definition of the AUC score
  - AUC maximization is that it needs to optimize the pairwise loss between two instances from different classes
- A generic pairwise loss on training data suffers from a severe scalability issue, which makes it not practical for DL on large-scale datasets
- But AUC square loss has adverse effect when trained with easy data and is sensitive to the noisy data.

AUC Margin Loss

- Addresses the two issues from AUC Square Loss:
  - Adverse effect:
    - Training affected by easy data
    - Sensitivity towards Noisy Data
- Example:
  - Top row: optimizing the AUC square loss
  - Bottom row: optimizing the new AUC margin loss (proposed loss)
  - First column: initial decision boundary
  - Middle column: easy examples to the training set
  - Last column: noisily labeled data



AUC Maximization

- Notations:
  - $\mathbb{I}(\cdot)$: Indicator function of a predicate
  - $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$: set of training data
  - x_i : input training example (eg. image)
  - $y_i \in \{1, -1\}$: corresponding label (e.g., the indicator of a certain disease)
  - $w \in \mathbb{R}^d$: parameters of the deep neural network to be learned
  - $h_w(x) = h(w, x)$: the prediction of the neural network on an input data x
  - $L(w; x, y) = l(h_w(x), y)$: Loss function; where $l(\hat{y}, y)$ is a surrogate loss function of the misclassification error (e.g., cross-entropy loss)
  - $[s]_+ = \max(s, 0)$

AUC Maximization

- Background on Scalable AUC Maximization:
  - Existing works consider the following definition of AUC:

$$AUC(h) = Pr(h_w(x) \geq h_w(x')|y = 1, y' = -1)$$
$$AUC(h) = \mathbb{E}(\mathbb{I}(h_w(x) - h_w(x') \geq 0)|y = 1, y' = -1)$$

  - Indicator function is replaced by a convex surrogate loss

$$AUC(h) = \mathbb{E}[\underbrace{l(h(x) - h(x'))}_{\text{Surrogate Loss Function}}|y = 1, y' = -1)$$

  - Final AUC formulation used by the existing works:

$$\min_{w \in \mathbb{R}^d} \frac{1}{N_+ N_-} \sum_{x \in S_+} \sum_{x \in S_-} l(h_w(x) - h_w(x'))$$

    - Where:
      - S_+, S_- : is the set of positive and negative examples
      - N_+, N_- : Denotes their size

AUC Maximization

- Issues with optimizing surrogate loss:
    - High cost; for large datasets the complexity worse as $O(n^2)$
    - Optimization focuses only on linear models
    - Not suitable for distributed optimization
- AUC Square Loss:
    - To solve the scalability problem, we optimize the AUC square loss
    - Use the square loss instead of the surrogate loss

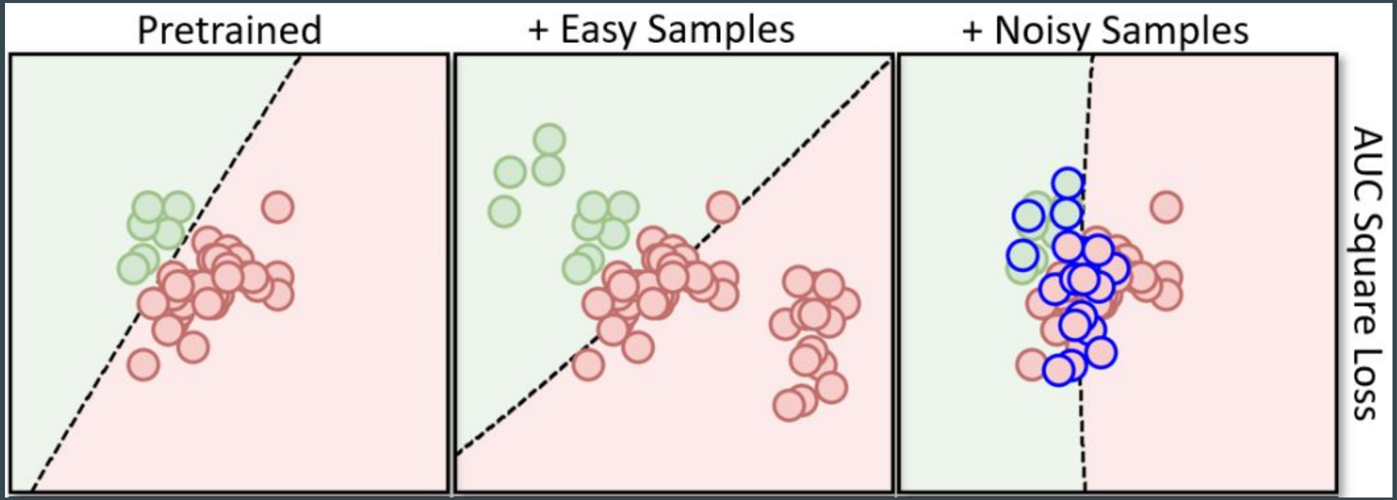
$$l(h_w(x) - h_w(x')) = (1 - h_w(x) + h_w(x'))^2$$

- Objective function:

$$\min_{\substack{w \in \mathbb{R}^d \\ (a,b) \in \mathbb{R}^2}} \max_{\alpha \in \mathbb{R}} f(w, a, b, \alpha) = \mathbb{E}_z[F(w, a, b, \alpha, ; z)]$$

AUC Square Loss

- Is AUC Square Loss the answer for AUC maximization?
  - Sensitive to Noisy Data
  - Adverse Effect on easy data



AUC Margin Loss

- Reformulation of AUC Square loss:
  - AUC Square Loss: [1]

$$A_s(w) = \mathbb{E}[(1 - h_w(w) + h_w(x'))^2 | y = 1, y' = -1]$$

$$= \underbrace{\mathbb{E}[(h_w(x) - a(w))^2 | y = 1]}_{A_1(w)} + \underbrace{\mathbb{E}[(h_w(x') - b(w))^2 | y = 1]}_{A_2(w)} + \underbrace{(1 - a(w) + b(w)^2}_{A_3(w)}$$

$$= A_1(w) + A_2(w) + (1 - a(w) + b(w))^2$$

  - Where:

$$a(w) = \mathbb{E}[h_w(x) | y = 1]$$

$$b(w) = \mathbb{E}[h_w(x') | y' = 1]$$

Derivation for [1] in Appendix B - Robust Deep {AUC} Maximization: {A} New Surrogate Loss and Empirical Studies on Medical Image Classification

AUC Margin Loss

Reference:
$$= A_1(w) + A_2(w) + (1 - a(w) + b(w))^2$$

- $A_1(w)$, $A_2(w)$ aims to minimize the variance of prediction scores on positive and negative data respectively
- $A_3(w)$ aims to push the mean prediction scores of positive and negative examples to be far away
- Issue: the square term can cause the same adverse effect as the AUC square loss
- Solution: Replace $A_3(w)$ with a squared hinge function:

$$max_{\alpha\geq 0}\{2\alpha(m - a(w) + b(w)) - \alpha^2\} = (m - a(w) + b(w))_+^2$$

  - Where:
    - ‘m’ is a hyper-parameter that specifies desired margin between a(w) and b(w)

AUC Margin Loss

$$A(w) = A_1(w) + A_2(w) + \max_{\alpha \geq 0} 2\alpha(m - a(w) + b(w)) - \alpha^2$$

- Benefits:
  - Robust to easy data
  - Robust to noisy data

DAM with the AUC margin Loss

- AUC margin loss is equivalent to the following min-max optimization:

$$\min_{\substack{w \in \mathbb{R}^d \\ (a,b) \in \mathbb{R}^2}} \max_{\alpha \geq 0} \mathbb{E}_z [F_M(w, a, b, \alpha; z)]$$

- A Proximal Epoch Stochastic Method is used:
  - To update variables w, a, b, α
  - $v = (w, a, b)$ denotes all the primal variables
    - where:
      - a & b are the mean prediction score on positive data and negative data, respectively
      - $\alpha = 1 + b - a$

DAM with AUC Margin Loss

- Algorithm:
  - For every iteration in $t = 1, \ldots, T$
    - We compute gradients for each of the primal variable with parameter as *z.*
    - Update the primal variable: (w, a, b) and α
    - Update model parameters:
      - λ is the standard regularization parameter

Algorithm 1 PESG for optimizing the AUC margin loss

Require: η, γ, λ, T

Initialize $\mathbf{v}_1, \alpha_1 \geq 0$

for $t = 1, \ldots, T$ **do**

Compute $\nabla_{\mathbf{v}} F_{\mathrm{M}}(\mathbf{v}_t, \alpha_t; \mathbf{z}_t)$ and $\nabla_{\alpha} F_{\mathrm{M}}(\mathbf{v}_t, \alpha_t; \mathbf{z}_t)$.

Update primal variables

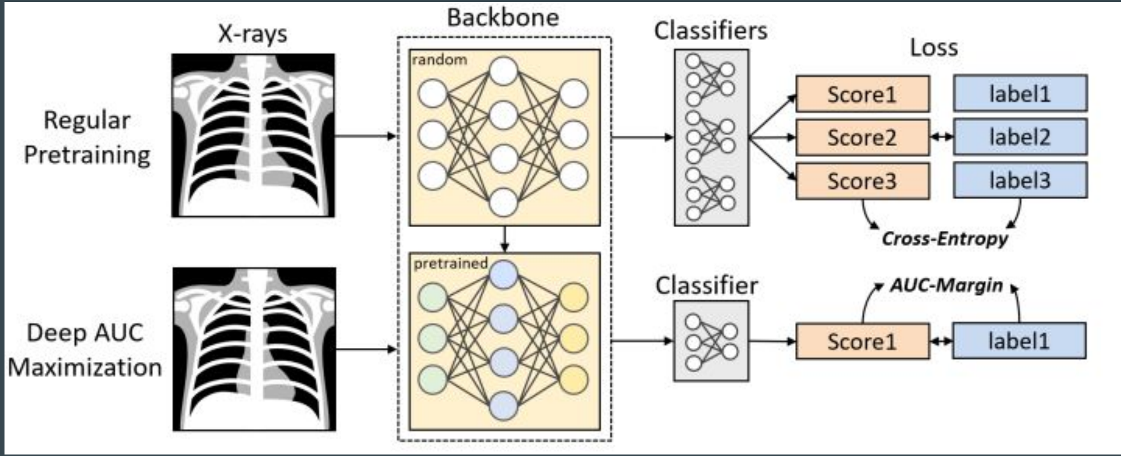
$$\mathbf{v}_{t+1} = \mathbf{v}_t - \eta(\nabla_{\mathbf{v}} F_{\mathrm{M}}(\mathbf{v}_t, \alpha_t; \mathbf{z}_t) + \gamma(\mathbf{v}_t - \mathbf{v}_{\mathrm{ref}})) - \lambda\eta\mathbf{v}_t$$

Update $\alpha_{t+1} = [\alpha_t + \eta\nabla_{\alpha} F_{\mathrm{M}}(\mathbf{v}_t, \alpha_t; \mathbf{z}_t)]_+$.

Decrease η by a factor and update $\mathbf{v}_{\mathrm{ref}}$ periodically

end for

A Two-stage Framework for DAM



- Directly optimizing the AUC margin loss can easily handle the recognition tasks on simple datasets, e.g., CIFAR
- Can be difficult working with complex tasks in medical image classification
- They employ a two-stage framework on difficult medical image classification tasks
    - Includes a pre-training step that minimizes the standard cross-entropy loss
    - An AUC maximization step that maximizes an AUC surrogate loss of the pre-trained CNN for learning all layers with the last classifier layer randomly initialized.

Empirical Studies

- Extensive empirical studies on the proposed robust DAM method with the AUC margin loss
- Performance on Benchmark datasets:
  - Construct imbalanced dataset from Cat&Dog, CIFAR-10, CIFAR-100 , STL-10
  - Randomly split the training data by class ID into two even portions as the positive and negative classes
  - To make it imbalance remove some samples from the positive class
  - Two popular network used: DesneNet121 and ResNet20
    - ELU activation functions
    - 100 epochs with a stagewise learning rate: initial value of 0.1
    - decaying at 50% and 75% of the total number of training epochs for all experiments
    - $\lambda = 1e^{-4}$
    - Different batch size for datasets

Empirical Studies

- DAM with AUC margin loss (AUC-M) vs. DAM with AUC square loss (AUC-S) DL with two other popular loss functions i.e., cross-entropy loss (CE) and focal loss (Focal) trained by SGD

Dataset	CE	Focal	AUC-S	AUC-M
C2 (D)	0.718±0.018	0.713±0.009	0.803±0.018	**0.809±0.016**
C10 (D)	0.698±0.017	0.700±0.007	0.745±0.010	**0.760±0.006**
S10 (D)	0.641±0.032	0.660±0.027	0.669±0.070	**0.703±0.030**
C100 (D)	0.588±0.011	0.591±0.017	0.607±0.010	**0.614±0.016**
C2 (R)	0.730±0.028	0.724±0.020	0.748±0.007	**0.756±0.017**
C10 (R)	0.690±0.011	0.681±0.011	0.702±0.015	**0.715±0.008**
S10 (R)	0.641±0.021	0.634±0.024	0.645±0.029	**0.659±0.020**
C100 (R)	0.563±0.015	0.565±0.022	0.587±0.017	**0.596±0.016**

D = DenseNet121
R = ResNet20
C2 = Imbalance Cat&Dog
C10 = CIFAR-10
C100 = CIFAR-100
SLT-10 = S10

Empirical Studies

- Medical Image Classification Tasks
    - CheXpert Competition:
        - CheXpert competition is a medical AI competition organized by Stanford ML group
        - Chest X-Ray dataset for detecting chest and lung disease
        - Stats:
            - Train: 224,316 high quality X-ray images from 65,240 patients
            - Validation: 234 images from 200 patients
            - Test: images for 500 patients
        - Only 5 selected diseases for evaluation
            - Atelectasis, Cardiomegaly, Consolidation, Edema, Pleural Effusion

Model	AUC	NRBC	Rank
Stanford Baseline [22]	0.9065	1.8	85
YWW [40]	0.9289	2.8	5
Hierarchical Learning [31]	0.9299	2.6	2
DAM (Ours)	**0.9305**	**2.8**	**1**

Empirical Studies

- Medical Image Classification Tasks
  - Kaggle Melanoma Classification Competition:
    - Stats:
      - 33,126 training images with 584 malignant melanoma images (imbalance ratio=1.76%)
      - 10,892 testing images with an unknown number of melanoma images
      - Testing set is split into public testing set and private testing set at 30%/70% ratio by patient ID

Loss	Public	Private	Public	Private
CE	0.9391	0.9285	0.9447	0.9345
Focal	0.9412	0.9266	0.9424	0.9303
AUC-S	0.9482	0.9332	0.9502	0.9364
AUC-M	**0.9497**	**0.9357**	**0.9503**	**0.9393**
AUC-S (Meta)	0.9495	0.9358	0.9501	0.9409
AUC-M (Meta)	**0.9522**	**0.9380**	**0.9520**	**0.9423**
Our Submission	-	-	**0.9685**	**0.9438**

Empirical Studies

- Medical Image Classification Tasks

Data (imratio)	CE	Focal	AUC-S	AUC-M
DDSM+ (13%)	0.9392	0.9495	0.9469	**0.9544**
PatchCamelyon (1%)	0.8394	0.8556	0.8703	**0.8896**

  - The DDSM+ data is a combination of two datasets namely DDSM and CBIS-DDSM
    - Training: 55,000 mammographic images
    - Test: 13,900 mammographic images
  - PathCamelyon dataset:
    - Training: 294, 912 color images from histopathologic scans of lymph node section
    - Test: 32, 768 color images from histopathologic scans of lymph node section

Thank you