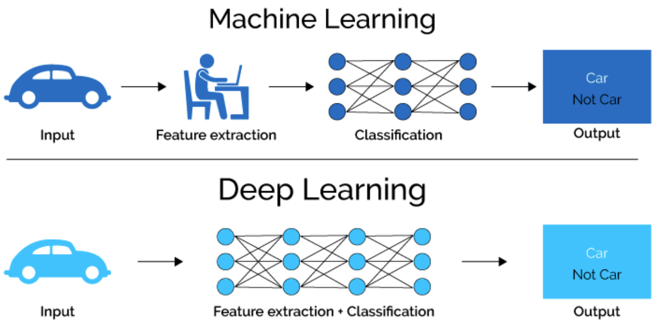
Deep AUC Maximization and Applications in Medical Image Classification

Tianbao Yang

Department of Computer Science
The University of Iowa

Outline

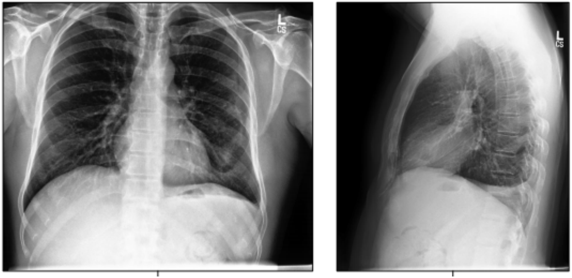
Deep Learning is Everywhere



- Image Recognition beats human
- AlphaGo beats human champion
- ...

AI for Medical Image Classification

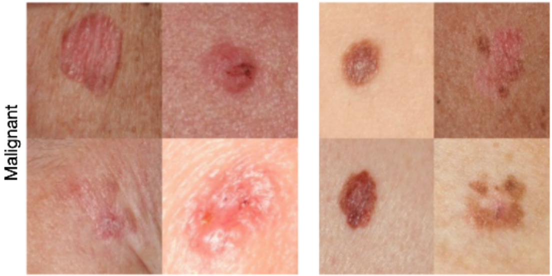
Radiologist-level Interpretation of X-ray images



Irvin, et al. (AAAI, 2019), reported AUC>0.90

AI for Medical Image Classification

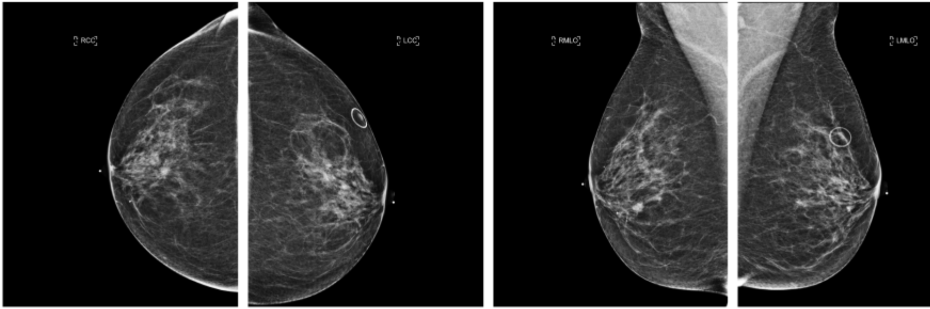
Dermatologist-level classification of skin cancer



Esteva et al. (Nature, 2017), reported AUC$>$0.91

AI for Medical Image Classification

Radiologist-level Screening of Breast Cancer



Wu, et al. (IEEE T. Medical Imaging, 2020), reported AUC=0.895

Keys to Success for Medical AI

- Large-scale Datasets (100,000+ $\sim$ 1,000,000 images)
- Domain-specific techniques (e.g., network structures)

Can we design a generic method to further improve the performance without relying on domain knowledge?

Our solution: Deep AUC Maximization

Keys to Success for Medical AI

- Large-scale Datasets (100,000+ $\sim$ 1,000,000 images)
- Domain-specific techniques (e.g., network structures)

Can we design a generic method to further improve the performance without relying on domain knowledge?

Our solution: Deep AUC Maximization

Is Deep Learning Trustworthy?

Not Yet

TECH

Google Photos labeled black people 'gorillas'

Jessica Guynn USA TODAY

Published 1:15 p.m. ET Jul. 1, 2015 | **Updated 2:10 p.m. ET Jul. 1, 2015**

CNN BUSINESS Markets Tech Media Success Video

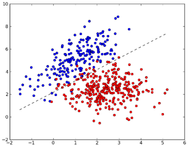
Sorry, kids: Apple's new Face ID isn't meant for you

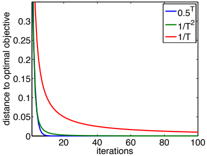
by Sara Ashley O'Brien @saraashleyo

September 27, 2017: 3:38 PM ET

What is Wrong?

Three Key Components of Machine Learning





Data

Model

Optimization

Data Imbalance

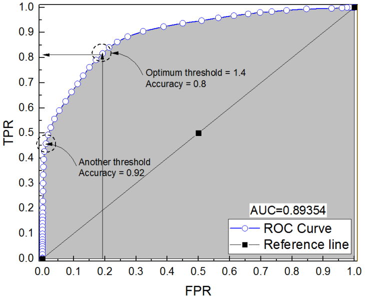
- Rare Disease Identification
- Terrorist Identification
- Credit Card Fraud Detection
- ...

How to measure a Classifier's performance?

- Accuracy
  - not suitable for imbalanced data
- Precision, Recall, F-measure
  - often used in information retrieval
- Area under ROC Curve (AUC)
  - default metric in medical analytics

ROC Curve (Receiver Operating Characteristic Curve)

ROC curve: True Positive Rate vs False Positive Rate



- World War II for measuring the ability of a radar receiver operator
- Medicine, radiology, biometrics
- Diagnostic analysis

Interpretation of AUC

$$\text{True-AUC}(h) = \Pr(h(\mathbf{x}) \geq h(\mathbf{x}')|y = 1, y' = -1)$$

- Equivalent to Wilcoxon Statistics (Hanley and McNeil,1982)
- Suitable for Imbalanced Data
- h: prediction model (e.g., deep neural network)
- $\mathbf{x}, \mathbf{x}'$ random data

AUC Maximization is much more Difficult

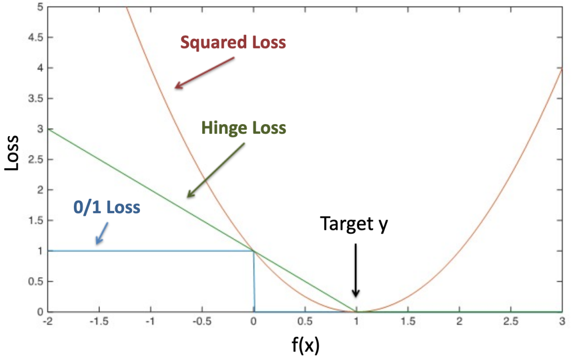
Example 1		Example 2		Example 3	
Prediction	Ground Truth	Prediction	Ground Truth	Prediction	Ground Truth
0.9	1	0.9	1	0.9	1
0.8	1	**0.41**(↓)	1	**0.41**(↓)	1
0.7	1	0.7	1	**0.40**(↓)	1
0.6	0	0.6	0	**0.49**(↓)	0
0.6	0	**0.49**(↓)	0	**0.48**(↓)	0
0.47	0	0.47	0	0.47	0
0.47	0	0.47	0	0.47	0
⋮	⋮	⋮	⋮	⋮	⋮
0.1	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** (↓)	

Outline

Necessity of a Surrogate Loss

$$\begin{aligned}\text{True-AUC}(h) &= \Pr(h(\mathbf{x}) \geq h(\mathbf{x}')|y=1, y'=-1) \\ &= \mathrm{E}[\mathbb{I}(h(\mathbf{x}) \geq h(\mathbf{x}'))|y=1, y'=-1]\end{aligned}$$

$$\min_h \text{AUC-Surrogate}(h) = \mathrm{E}[\ell(h(\mathbf{x}) - h(\mathbf{x}'))|y=1, y'=-1]$$

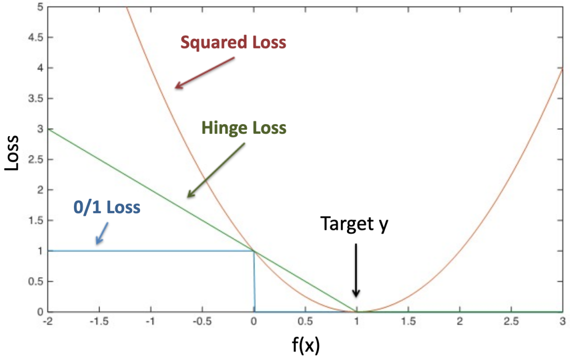


Necessity of a Surrogate Loss

0-1 loss function

$$\begin{aligned}\text{True-AUC}(h) &= \Pr(h(\mathbf{x}) \geq h(\mathbf{x}')|y=1, y'=-1)\\ &= \mathrm{E}[\mathbb{I}(h(\mathbf{x}) \geq h(\mathbf{x}'))|y=1, y'=-1]\end{aligned}$$

$$\min_h \text{AUC-Surrogate}(h) = \mathrm{E}[\ell(h(\mathbf{x}) - h(\mathbf{x}'))|y=1, y'=-1]$$



Challenge of Optimizing a Pariwise Loss

$$\min_{h} \text{AUC-Surrogate}(h) = \mathrm{E}[\ell(h(\mathbf{x}) - h(\mathbf{x}'))|y = 1, y' = -1]$$

Issues:

- High costs: B samples: $O(B^2)$
- Not suitable for online learning: data coming sequentially
- Not suitable for distributed optimization: data in different machines

Related Work: Zhao, Jin, Hoi, Yang (ICML 2011)

- first online AUC maximization
- large buffer, high computation, cannot scale up
- no convergence rate

Square loss Mitigates the Optimization Challenge

Square loss is an exception:

square loss

$$\min_{\mathbf{w}\in\mathbb{R}^d} A(\mathbf{w}) \triangleq \mathrm{E}_{\mathbf{z},\mathbf{z}'}[(h_{\mathbf{w}}(\mathbf{x}) - h_{\mathbf{w}}(\mathbf{x}') - 1)^2 | y = 1, y' = -1] \tag{1}$$

Min-max Reformulation (Ying et al. 2016):

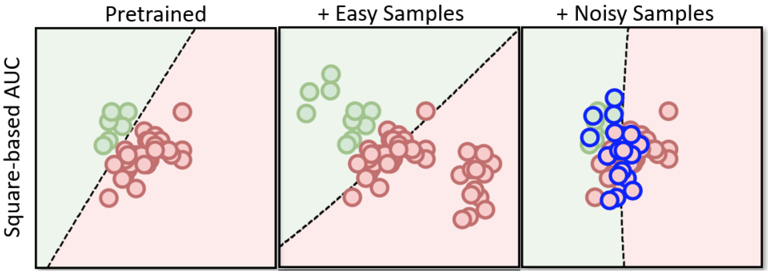
$$\min_{\substack{\mathbf{w}\in\mathbb{R}^d \\ (a,b)\in\mathbb{R}^2}} \max_{\alpha\in\mathbb{R}} f(\mathbf{w}, a, b, \alpha) = \mathrm{E}_{\mathbf{z}}[F(\mathbf{w}, a, b, \alpha, \mathbf{z})], \tag{2}$$

- $\mathbf{z} = (\mathbf{x}, y)$
- Ying et al. (2016): focuses on linear model

Is Square loss Good for AUC Maximization?

No Really!

- Adverse Effect on Easy Data
- Sensitive to Noisy Data
- Explanation: Consider SGD update



Our Margin-based Surrogate Loss

Decomposition of Square loss:

$$A(\mathbf{w}) = \mathrm{E}[(h_{\mathbf{w}}(\mathbf{x}) - a(\mathbf{w}))^2 | y = 1] + \mathrm{E}[(h_{\mathbf{w}}(\mathbf{x}') - b(\mathbf{w}))^2 | y' = 1] + (1 + b(\mathbf{w}) - a(\mathbf{w}))^2$$

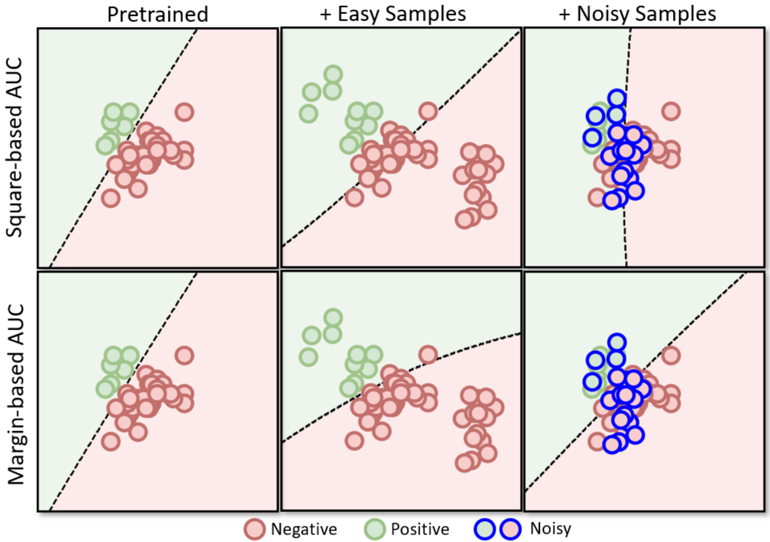
$a(\mathbf{w})$ ($b(\mathbf{w})$): average score of positive data (negative data)

Margin-based Loss: (under prepration)

$$A_1(\mathbf{w}) = \mathrm{E}[(h_{\mathbf{w}}(\mathbf{x}) - a(\mathbf{w}))^2 | y = 1] + \mathrm{E}[(h_{\mathbf{w}}(\mathbf{x}') - b(\mathbf{w}))^2 | y' = 1] + (m + b(\mathbf{w}) - a(\mathbf{w}))_+^2$$

where $[s]_+ = \max(0, s)$, m is a margin parameter.

Margin-based Surrogate Loss is more robust



Outline

Min-max formulation of Margin-based Surrogate Loss

$$A_1(\mathbf{w}) = \mathrm{E}[(h_{\mathbf{w}}(\mathbf{x}) - a(\mathbf{w}))^2 | y = 1] + \mathrm{E}[(h_{\mathbf{w}}(\mathbf{x}') - b(\mathbf{w}))^2 | y' = 1] + (m + b(\mathbf{w}) - a(\mathbf{w}))_+^2$$

Using convex conjugate: $[s]_+^2 = \max_{\alpha \geq 0} 2\alpha s - \alpha^2$

$$\min_{\mathbf{w},a,b} \max_{\alpha \in [0,\infty)} \mathrm{E}_{\mathbf{z}}[F(\mathbf{w}, a, b, \alpha; \mathbf{z})]$$

- $\mathbf{z} = (\mathbf{x}, y)$
- **Non-Convex Strongly-Concave Min-Max Problem**

Stochastic Primal Dual (SPD) Method

Consider

$$\min_{\mathbf{w}\in W}\max_{\alpha\in\Omega} f(\mathbf{w},\alpha) = \mathrm{E}_{\mathbf{z}}[f(\mathbf{w},\alpha,\mathbf{z})]$$

SPD

$$\mathbf{w}_{t+1} = \prod_{W}[\mathbf{w}_t - \eta_t \nabla_w f(\mathbf{w}_t,\alpha_t,\mathbf{z}_t)], \quad \alpha_{t+1} = \prod_{\Omega}[\alpha_t + \eta_t \nabla_\alpha f(\mathbf{w}_t,\alpha_t,\mathbf{z}_t)]$$

$$\text{return} : \widehat{\mathbf{w}} = \sum_{t=1}^{T}\mathbf{w}_t/T, \quad \widehat{\alpha} = \sum_{t=1}^{T}\alpha_t/T$$

Previous Studies Focus on Convex-Concave Problems

Weakly Convex and Strongly Concave Problems

$$\min_{\mathbf{w}\in W} \max_{\alpha\in\Omega} f(\mathbf{w},\alpha) = \mathrm{E}_{\mathbf{z}}[f(\mathbf{w},\alpha,\mathbf{z})]$$

- Weakly convex in $\mathbf{w}$: $f(\mathbf{w},\alpha) + \frac{\gamma}{2}\|\mathbf{w}\|^2$ is convex
- Smooth function is weakly convex
- Solving this problem by following proximal-point framework
- Successively Solve

$$(\mathbf{w}_{k+1},\alpha_{k+1}) \approx \arg\min_{\mathbf{w}} \max_{\alpha} f(\mathbf{w},\alpha) + \gamma\|\mathbf{w}-\mathbf{w}_k\|^2$$

Algorithm 1 Stagewise-SPD

```
1: for s = 1, 2, ..., S do
2:     Let f_s(w, α) = f(w, α) + γ||w − w^(s)||^2
3:     (w^(s+1), α^(s+1)) = SPD(f_s, w^(s), α^(s), η_s, T_s)
4:     η_{s+1} ∝ 1/(s + 1), T_{s+1} ∝ (s + 1)
5: end for
6: Return (w̄^(S+1), ᾱ^(S+1))
```

1: **for** $s = 1, 2, ..., S$ **do**
2: Let $f_s(\mathbf{w}, \alpha) = f(\mathbf{w}, \alpha) + \gamma\|\mathbf{w} - \mathbf{w}^{(s)}\|^2$
3: $(\mathbf{w}^{(s+1)}, \alpha^{(s+1)}) = SPD(f_s, \mathbf{w}^{(s)}, \alpha^{(s)}, \eta_s, T_s)$
4: $\eta_{s+1} \propto 1/(s+1)$, $T_{s+1} \propto (s+1)$
5: **end for**
6: Return $(\bar{\mathbf{w}}^{(S+1)}, \bar{\alpha}^{(S+1)})$

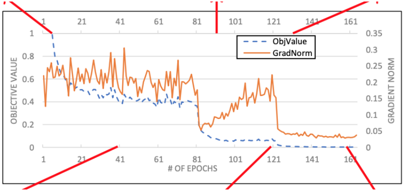
- $O\left(\frac{1}{\epsilon^4}\right)$ for finding ϵ-stationary point (Yan et al. NeurIPS 2020)
- The complexity matches a lower bound (Arjevani et al. 2019)
- Analysis involves duality gap of regularized function

Improved Rate

$$\min_{\mathbf{w}} P(\mathbf{w}) := \max_{\alpha \in \Omega} f(\mathbf{w}, \alpha)$$

- Smooth Condition: $P(\mathbf{w})$ is smooth
- Polyak- Lojasiewicz (PL) condition (Allen-Zhu et al. 2019):

$$\mu(P(\mathbf{w}) - \min_{\mathbf{w}} P(\mathbf{w})) \leq \|\nabla P(\mathbf{w})\|^2$$



picture courtesy Allen-Zhu et al (2019)

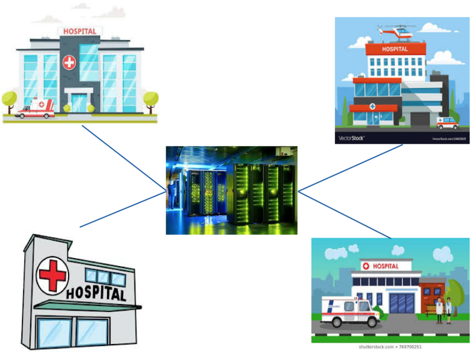
Algorithm 3 Stagewise-SPD

1: **for** $s = 1, 2, ..., S$ **do**
2: Let $f_s(\mathbf{w}, \alpha) = f(\mathbf{w}, \alpha) + \gamma\|\mathbf{w} - \mathbf{w}^{(s)}\|^2$
3: $(\mathbf{w}^{(s+1)}, \alpha^{(s+1)}) = SPD(f_s, \mathbf{w}^{(s)}, \alpha^{(s)}, \eta_s, T_s)$
4: $\eta_{s+1} \propto \exp(-(s+1)),\ T_{s+1} \propto \exp((s+1))$
5: **end for**
6: Return $(\bar{\mathbf{w}}^{(S+1)}, \bar{\alpha}^{(S+1)})$

- $O\left(\frac{1}{\mu^2\epsilon}\right)$ for finding ϵ-optimal point (Liu et al. ICLR 2019, Guo et al. 2020b)
- The complexity in terms of ϵ is optimal, matching the lower bound (Hazan and Kale 2011)
- Replace SPD with other algorithms (e.g., Stoc-Extragradient, AdaGrad, Variance Reduced)

Federated Deep AUC Maximization

- Data at one site is biased
- Data might not be shared
- Federated Learning



Federated Deep AUC Maximization

Guo et al. 2020a (ICML): First work on federated min-max learning

$$\min_{\substack{\mathbf{w}\in\mathbb{R}^d \\ (a,b)\in\mathbb{R}^2}} \max_{\alpha\in\mathbb{R}} f(\mathbf{w}, a, b, \alpha) = \frac{1}{K}\sum_{k=1}^{K} f_k(\mathbf{w}, a, b, \alpha),$$

- Federated Learning: communication complexity is critical
- $f_k(\mathbf{w}, a, b, \alpha) = \mathrm{E}_{\mathbf{z}^k}[F_k(\mathbf{w}, a, b, \alpha; \mathbf{z}^k)]$, $\mathbf{z}^k = (\mathbf{x}^k, y^k) \sim \mathbb{P}_k$
- K: total number of machines

Federated Deep AUC Maximization

Local Primal-Dual Method: Communication Perioadically

Each machine does initialization: $\mathbf{v}_0^k = \mathbf{v}_0, \alpha_0^k = \alpha_0$

for $t = 0, 1, ..., T-1$ **do**

Each machine k updates its local solution in parallel:

$$\mathbf{v}_{t+1}^k = \arg\min_{\mathbf{v}} \left[\nabla_{\mathbf{v}} F_k(\mathbf{v}_t^k, \alpha_t^k; \mathbf{z}_t^k)^T \mathbf{v} + \frac{1}{2\eta}\|\mathbf{v} - \mathbf{v}_t^k\|^2 + \frac{1}{2\gamma}\|\mathbf{v} - \mathbf{v}_0\|^2 \right],$$

$$\alpha_{t+1}^k = \alpha_t^k + \eta \nabla_\alpha F_k(\mathbf{v}_t^k, \alpha_t^k; \mathbf{z}_t^k),$$

if $t+1 \bmod I = 0$ **then**

$$\mathbf{v}_{t+1}^k = \frac{1}{K}\sum_{k=1}^K \mathbf{v}_{t+1}^k, \ \alpha_{t+1}^k = \frac{1}{K}\sum_{k=1}^K \alpha_{t+1}^k, \qquad \diamond \text{ communicate}$$

end if

end for

Return $\tilde{\mathbf{v}} = \frac{1}{K}\sum_{k=1}^K \frac{1}{T}\sum_{t=1}^T \mathbf{v}_t^k$.

Complexity Result

Federated Stochastic AUC Maximization: under PL condition

Alg.	Setting	Iteration Compl.	Comm. Compl.
Liu et al. (2019, ICLR)	Single	$O(1/(\mu^2\epsilon))$	-
Naive Parallel	Distributed	$O(1/(K\mu^2\epsilon))$	$O(1/(K\mu^2\epsilon))$
Ours	Distributed	$O(1/(K\mu^2\epsilon))$	$O(1/(\mu^{3/2}\epsilon^{1/2}))$

Guo et al. 2020a (ICML)

Outline

Experiments: the 1st Place at CheXpert Competition

the 1st Place

Stanford ML Group (Andrew Ng)
150+ submissions worldwide

Leaderboard

Will your model perform as well as radiologists in detecting different pathologies in chest X-rays?

Rank	Date	Model	AUC	Num Rads Below Curve
1	Aug 31, 2020	DeepAUC-v1 *ensemble*	**0.930**	2.8
2	Sep 01, 2019	Hierarchical-Learning-V1 (ensemble) *Vingroup Big Data Institute* https://arxiv.org/abs/1911.06475	0.930	2.6
3	Oct 15, 2019	Conditional-Training-LSR *ensemble*	0.929	2.6

Experiments: the 1st Place at CheXpert Competition

Data Set

- 224,316 chest X-rays images of 65,240 patients
- 14 common chest radiographic observations
- Only 5 selected diseases for evaluation
  - Atelectasis, Cardiomegaly, Consolidation, Edema, Pleural Effusion

Experiments: Deep AUC vs other methods

Table: Averaged Testing AUC Scores on CheXpert. NBRC means the # of radiologists out of 3 are beaten by AI algorithms.

Model	AUC	NRBC	Rank
Stanford Baseline (Irvin et al. 2019)	0.9065	1.8	85
YWW (Ye et al. 2020)	0.9289	2.8	5
Hierarchical Learning (Pham et al. 2020)	0.9299	2.6	2
DeepAUC (Ours)	**0.9305**	**2.8**	**1**

Experiments: Melanoma Classification Competition

Top 1% rank (ranked 33 out of 3314 teams) at **Kaggle Competition**

- May 27, 2020 - August 10, 2020
- 33,126 training images, with only 584 malignant melanoma samples
- Our AUC is **0.9438 vs 0.9490** of top 1
- 10 models vs 18 models of top 1 for ensemble

Melanoma Classification, AUC losses vs Other losses

Table: Comparison of Testing AUC on Melanoma dataset for Optimizing EffecientNetB5. TTA (30) means that the results are averaged over 30 times of evaluation on different test-time augmented data.

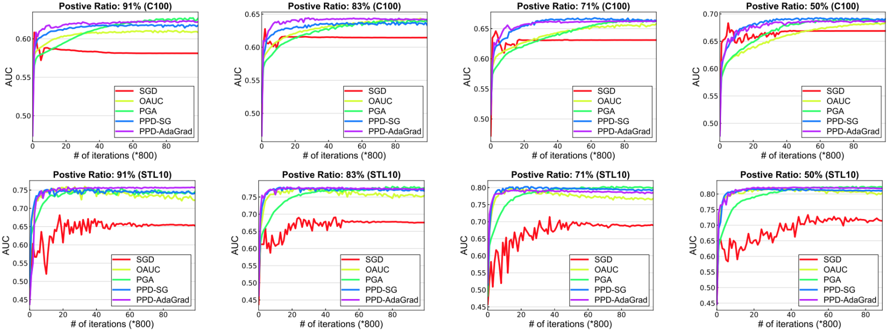
	wo/ TTA		**w/ TTA(30)**	
Loss	**Public**	**Private**	**Public**	**Private**
CE	0.9391	0.9285	0.9447	0.9345
Focal	0.9412	0.9266	0.9424	0.9303
AUC-S	0.9482	0.9332	0.9502	0.9364
AUC-M	**0.9497**	**0.9357**	**0.9503**	**0.9393**

Experiments: Margin loss vs Other losses

Table: Testing AUC on benchmark datasets with DenseNet121.

data (imratio)	CE	Focal	AUC-S	AUC-M
Cat&Dog (1%)	0.718±0.018	0.713±0.009	0.803±0.018	**0.809±0.016**
CIFAR10 (1%)	0.698±0.017	0.700±0.007	0.745±0.010	**0.760±0.006**
STL10 (1%)	0.641±0.032	0.660±0.027	0.669±0.070	**0.703±0.030**
CIFAR100 (1%)	0.588±0.011	0.591±0.017	0.607±0.010	**0.614±0.016**
Cat&Dog (10%)	0.893±0.004	0.879±0.005	0.901±0.002	**0.902±0.001**
CIFAR10 (10%)	**0.898±0.005**	0.879±0.005	0.889±0.002	0.887±0.005
STL10 (10%)	0.820±0.015	0.819±0.010	0.825±0.013	**0.846±0.015**
CIFAR100 (10%)	0.710±0.007	0.705±0.007	0.720±0.003	**0.723±0.006**

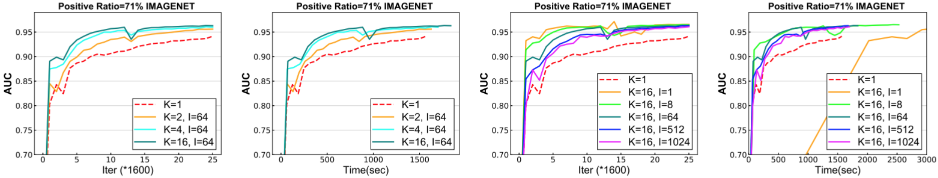
Experiments: Convergence Speed



Liu et al. 2020 ICLR

- blue and purple are our algorithms exploring PL condition
- green is our algorithm without exploring PL condition
- red is standard SGD

Experiments: Federated Deep AUC



(a) Fix I, vary K (b) Fix K, vary I

- more machines converge faster
- can communicate periodically without sacrificing performance

Guo et al. 2020a, ICML

Conclusions

Our Achivements

- Margin-based loss is better than Square Loss
- Stochastic Non-Convex Min-Max Optimization with Fast Rates
- Communication Efficient Federated Learning Algorithms
- The 1st Place at CheXpert Competition

Open Problems:

- Consistency of Margin-based Loss
- Improve Convergence for Non-Convex Min-Max Optimization
- Reduce Communication Complexity of Federated Deep AUC
- Other Medical Datasets

Acknowledgements

- Zhuoning Yuan, Yan Yan, Zhishuai Guo, Mingrui Liu, Hassan Rafique, Qihang Lin
- NSF Career Award

Thank You!

Questions?

References

- Rafique et al (2019). "Non-Convex Min-Max Optimization: Provable Algorithms and Applications in Machine Learning."
- Yan et al. (2020, NeurIPS). "Optimal Epoch Stochastic Gradient Descent Ascent Methods for Min-Max Optimization."
- Liu et al. (2019, ICLR). "Stochastic AUC Maximization with Deep Neural Networks."
- Guo et al. (2020, ICML). "Communication-Efficient Distributed Stochastic AUC Maximization with Deep Neural Networks."
- Guo et al. (2020). "Fast Objective and Duality Gap Convergence for Non-convex Strongly-concave Min-max Problems."

References I