Max-Mahalanobis Linear Discriminant Analysis Networks

Tianyu Pang[1], Chao Du[1] and Jun Zhu[1]

[1]Department of Computer Science and Technology, Tsinghua University, Beijing, China

ICML | 2018

Motivitions

A typical feed-forward deep neural network (DNN) is a combination of a nonlinear transformation from the input x to the latent feature vector z and a linear classifier acting on z to return a prediction for x. Our work is proposed under the two motivitions:

1. Compared to the nonlinear transformation part, the linear classifier part is under-explored, which is by default defined as a softmax regression (SR).
2. DNNs with a SR classifier are vulnerable to adversarial attacks, where human imperceivable noises can be crafted to fool a high-accuracy network.

Thus, we attempt to design a network with a novel linear classifier part substituted for SR, expecting for better performance.

Inspirations

In the binary-class classification cases, Efron (1975) shows that if the input pair (x, y) distributes as

$$P(y = i) = \pi_i,\ P(x|y = i) = \mathcal{N}(\mu_i, \Sigma), \quad (1)$$

where $i \in \{0, 1\}$, then logistic regression (LR) is less efficient than linear discriminant analysis (LDA). The relative efficiency of LR to LDA can be represented as $\mathrm{Eff}_p(\zeta, \Delta)$, where $\zeta = \log(\frac{\pi_0}{\pi_1})$, and $\Delta = [(\mu_1-\mu_0)^\top\Sigma^{-1}(\mu_1-\mu_0)]^{\frac{1}{2}}$ is the Mahalanobis distance of two Gaussian components. Generally, larger values of $|\zeta|$ or Δ imply lower values of $\mathrm{Eff}_p(\zeta, \Delta)$.

Max-Mahalanobis Distribution

We consider the multi-class cases, L is #class, $[L] = \{1, \cdots, L\}$. Under a linear transformation on the input, the distribution assumption (1) can be standardized and extended to

$$P(y = i) = \pi_i,\ P(x|y = i) = \mathcal{N}(\bar{\mu}_i, I), \quad (2)$$

where $i \in [L]$, $\sum_{i=1}^{L}\pi_i = 1$ and $\sum_{i=1}^{L}\bar{\mu}_i = 0$. Then the decision boundary obtained by LDA between class i and j is decided by the Fisher's linear discriminant function $\lambda_{i,j}(x) = 0$.

In the adversarial setting, the nearest adversarial example x^* w.r.t the normal example x must be located on the decision boundary. We randomly sample a normal example of class i as $x_{(i)}$, i.e., $x_{(i)} \sim \mathcal{N}(\mu_i, I)$, and denote its nearest adversarial counterpart on the decision boundary $\lambda_{i,j}(x) = 0$ as $x^*_{(i,j)}$. There is $\hat{y}(x_{(i)}) = i, \hat{y}(x^*_{(i,j)}) = j$ or $\hat{y}(x_{(i)}) = j, \hat{y}(x^*_{(i,j)}) = i$, where $\hat{y}(\cdot)$ refers to the LDA classifier. We define the distance between $x_{(i)}$ and $x^*_{(i,j)}$ as $d_{(i,j)}$, then there is:

Theorem 1 *The expectation of the distance $d_{(i,j)}$ is a function of the Mahalanobis distance $\Delta_{i,j}$:*

$$\mathbb{E}[d_{(i,j)}] = \sqrt{\frac{2}{\pi}}\exp(-\frac{\alpha_{i,j}^2}{2}) + \alpha_{i,j}[1 - 2\Phi(-\alpha_{i,j})],$$

where $\alpha_{i,j} = \frac{1}{2}\Delta_{i,j} + \zeta_{i,j}/\Delta_{i,j}$, and $\Phi(\cdot)$ is the normal cumulative distribution function. Further there is $\partial\mathbb{E}[d_{(i,j)}]/\partial\Delta_{i,j} > 0$.

Upper Bound for Robustness

We define the robustness of the classifier as

$$\mathrm{RB} = \min_{i,j\in[L]} \mathbb{E}[d_{(i,j)}].$$

According to Theorem 1, there is $\mathrm{RB} \approx \overline{\mathrm{RB}} = \min_{i,j\in[L]} \Delta_{i,j}/2$. Let $\mu = \{\mu_i | i \in [L]\}$, $\|\mu\|_2$ be $\max_i \|\mu_i\|_2$. The following theorem gives a tight upper bound for $\overline{\mathrm{RB}}$ w.r.t μ:

Theorem 2 *Assume that $\sum_{i=1}^{L}\mu_i = 0$ and $\|\mu\|_2^2 = C$. Then we have*

$$\overline{\mathrm{RB}} \leq \sqrt{\frac{LC}{2(L-1)}}.$$

The equality holds if and only if

$$\mu_i^\top\mu_j = \begin{cases} C, & i = j, \\ C/(1-L), & i \neq j, \end{cases} \quad (3)$$

where $i, j \in [L]$ and $\mu_i, \mu_j \in \mu$.

We denote any set of means that satisfy the optimal condition (3) as μ^*. We define the distribution of assumption (2) with $\mu = \mu^*$ as Max-Mahalanobis distribution (MMD).

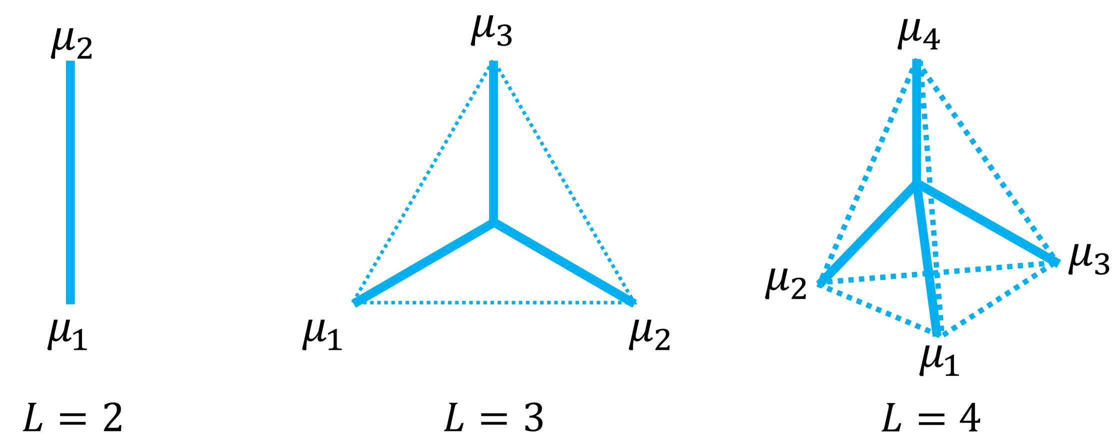

Figure 1: MMD under different values of L. $L = 2$, straight line; $L = 3$, equilateral triangle; $L = 4$, regular tetrahedron.

The MM-LDA Network

According to above analysis, we propose the Max-Mahalanobis linear discriminant analysis (MM-LDA) network. Specifically, considering the joint distribution $Q_\theta(z, y)$ induced by the network with parameters θ. We denote the MMD as $P(z, y)$, $\mathbb{H}(P, Q)$ as the cross-entropy function. Then the training objective for MM-LDA networks could be designed as

$$\begin{aligned}\mathbb{H}(Q_\theta, P) &= \mathbb{E}_{(z,y)\sim Q_\theta}[-\log P(y|z) - \log P(z)] \\ &= \mathbb{E}_{(z,y)\sim Q_\theta}[-\log P(y|z)] + \mathbb{E}_{z\sim Q'_\theta}[-\log P(z)].\end{aligned}$$

Here Q'_θ is the marginal distribution of Q_θ for z. Since we are focusing on classification tasks, we assume for tractability that the marginal distribution $Q'_\theta(z)$ is consistent with it of the MMD, i.e., $P(z)$. Thus, minimizing $\mathbb{H}(Q_\theta, P)$ equals to minimizing $\mathbb{E}_{(z,y)\sim Q_\theta}[-\log P(y|z)]$, which further leads to a similar loss function with SR networks under the MC approximation, and the only difference is that for MM-LDA networks $P(y|z)$ is obtained by LDA classifier rather than SR.

Experiments

Class-biased Datasets

Class-biased datasets (both training and test sets) are constructed by randomly sampling each data point of class i from CIFAR-10 with probability α_i. For a fair comparison, we still use uniform class priors $\pi_k = 1/L$ when using MM-LDA networks.

1. **Bias Probability 1** has $\alpha = (0.1, 0.2, 0.3, \cdots, 1.0)$.
2. **Bias Probability 2** has $\alpha = (0.2, 0.2, \cdots, 0.2, 1.0)$.

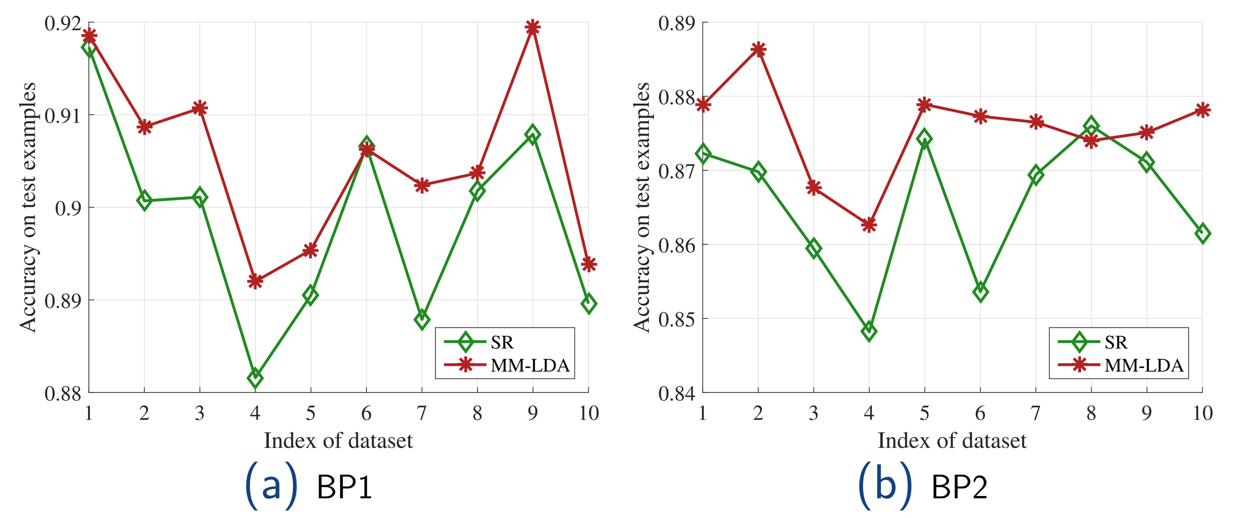

Figure 2: Each index corresponds to a counterpart of class-biased datasets under the bias probability.

Adversarial Setting

1. **SAT** fine-tunes the classifiers on the adversarial examples with the same value of perturbation.
2. **HAT** fine-tunes the classifiers on the adversarial examples with various values of perturbation from $[0.02, 0.20]$.

Table 1: Classification accuracy (%) on adversarial examples of MNIST and CIFAR-10. Res. refers to Resnet-32.

Pert.	Model	MNIST				CIFAR-10			
		FGSM	BIM	ILCM	JSMA	FGSM	BIM	ILCM	JSMA
0.04	Res.(SR)	93.6	87.9	94.8	92.9	20.0	5.5	0.2	65.6
	Res.(SR)+SAT	86.7	68.5	98.4	-	24.4	7.0	0.4	-
	Res.(SR)+HAT	88.7	96.3	**99.8**	-	30.3	5.3	1.3	-
	Res.(MM-LDA)	**99.2**	**99.2**	99.0	**99.1**	**91.3**	**91.2**	**70.0**	**91.2**
0.12	Res.(SR)	28.1	3.4	20.9	56.0	10.2	4.1	0.3	20.5
	Res.(SR)+SAT	40.5	8.7	88.8	-	88.2	6.9	0.1	-
	Res.(SR)+HAT	40.3	40.1	92.6	-	44.1	8.7	0.0	-
	Res.(MM-LDA)	**99.3**	**98.6**	**99.6**	**99.7**	**90.7**	**90.1**	**42.5**	**91.1**
0.20	Res.(SR)	15.5	0.3	1.7	25.6	10.7	4.2	0.6	11.5
	Res.(SR)+SAT	17.3	1.1	69.4	-	**91.7**	9.4	0.0	-
	Res.(SR)+HAT	10.1	10.5	46.1	-	40.7	6.0	0.2	-
	Res.(MM-LDA)	**97.5**	**97.3**	**96.6**	**99.6**	89.5	**89.7**	**31.2**	**91.8**

Model	MNIST	CIFAR-10
Res.(SR)	0.38	**7.13**
Res.(MM-LDA)	**0.35**	8.04

Table 2: Error rates (%) on the normal examples in test sets.

Model	MNIST	CIFAR-10
Res.(SR)	8.56	0.67
Res.(MM-LDA)	**16.32**	**2.80**

Table 3: Average minimal distortions (C&W attack).

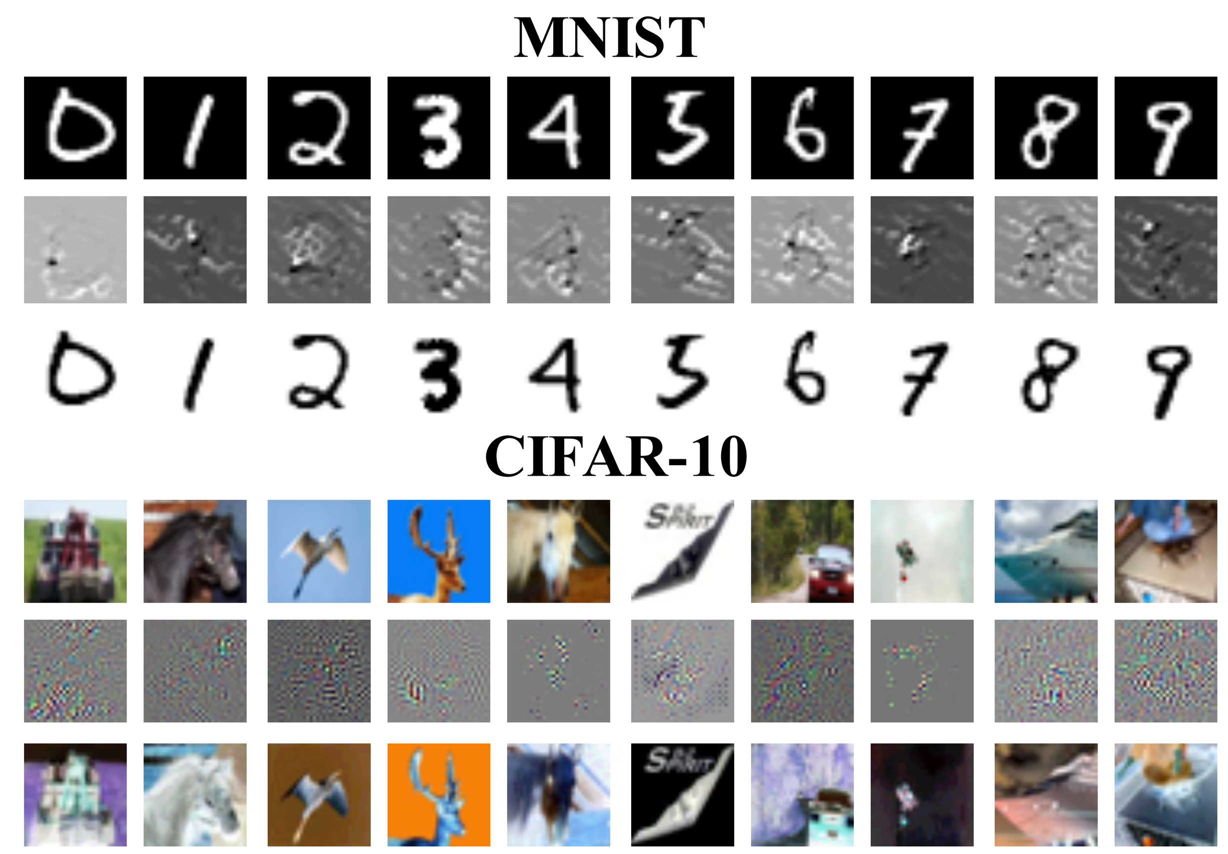

Figure 3: 1st row: normal examples; 2nd row: adversarial noises on SR nets; 3rd row: those on MM-LDA nets.

Thirty-fifth International Conference on Machine Learning (ICML 2018), Stockholm, Sweden