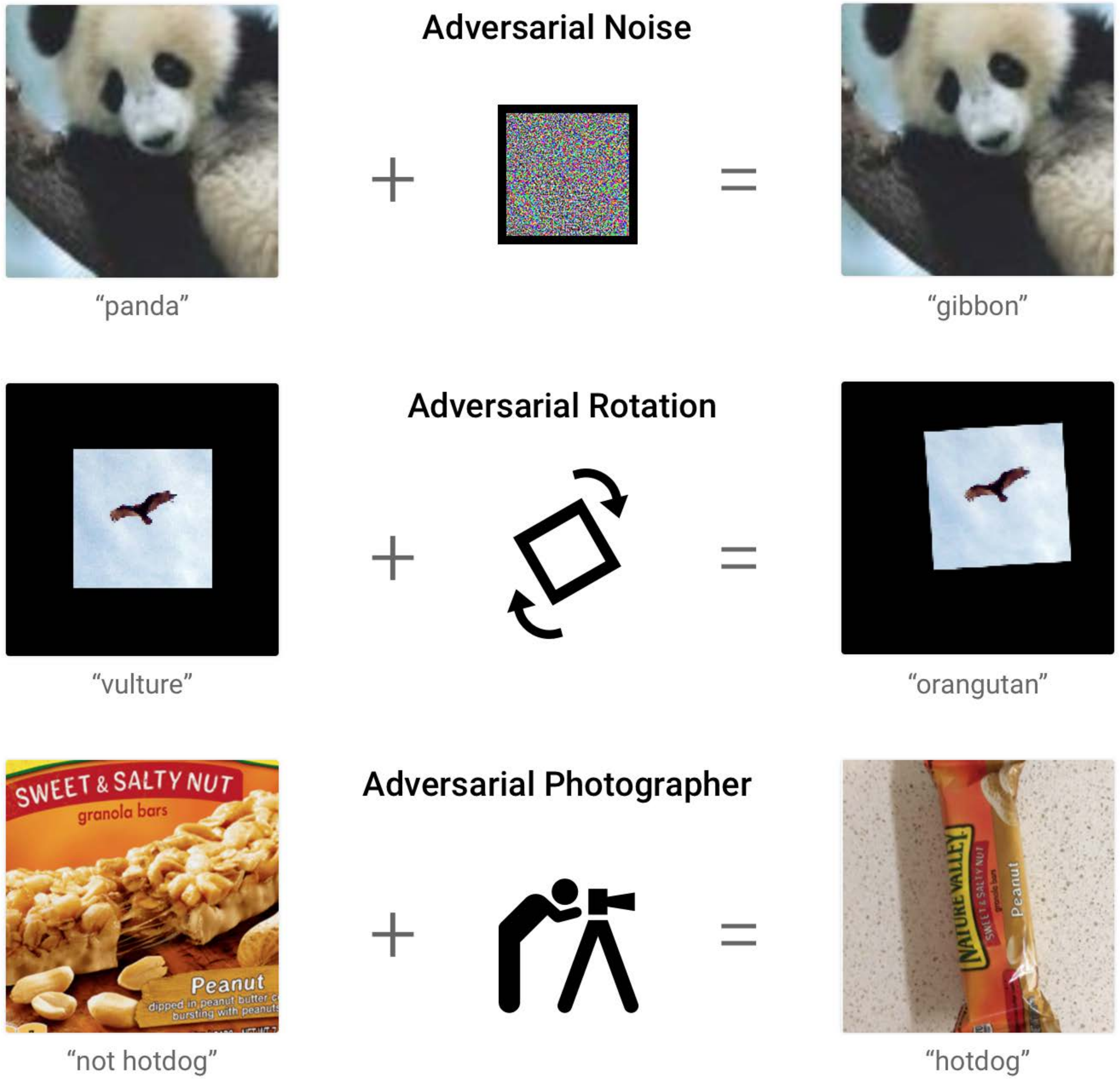
Distribution-restrained Softmax Loss for the Model Robustness

Chen Li
Inspur Electronic Information Industry

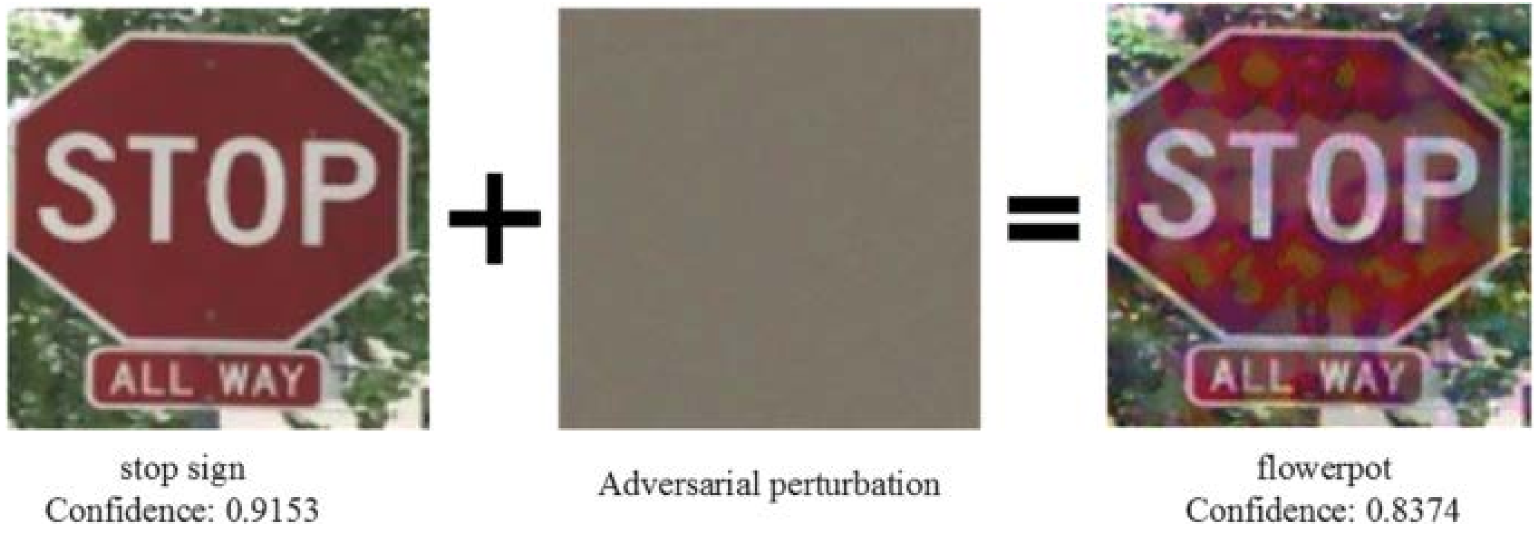
AISafety-SafeRL 2023 (IJCAI-23)

Introduction

- AI models are vulnerable

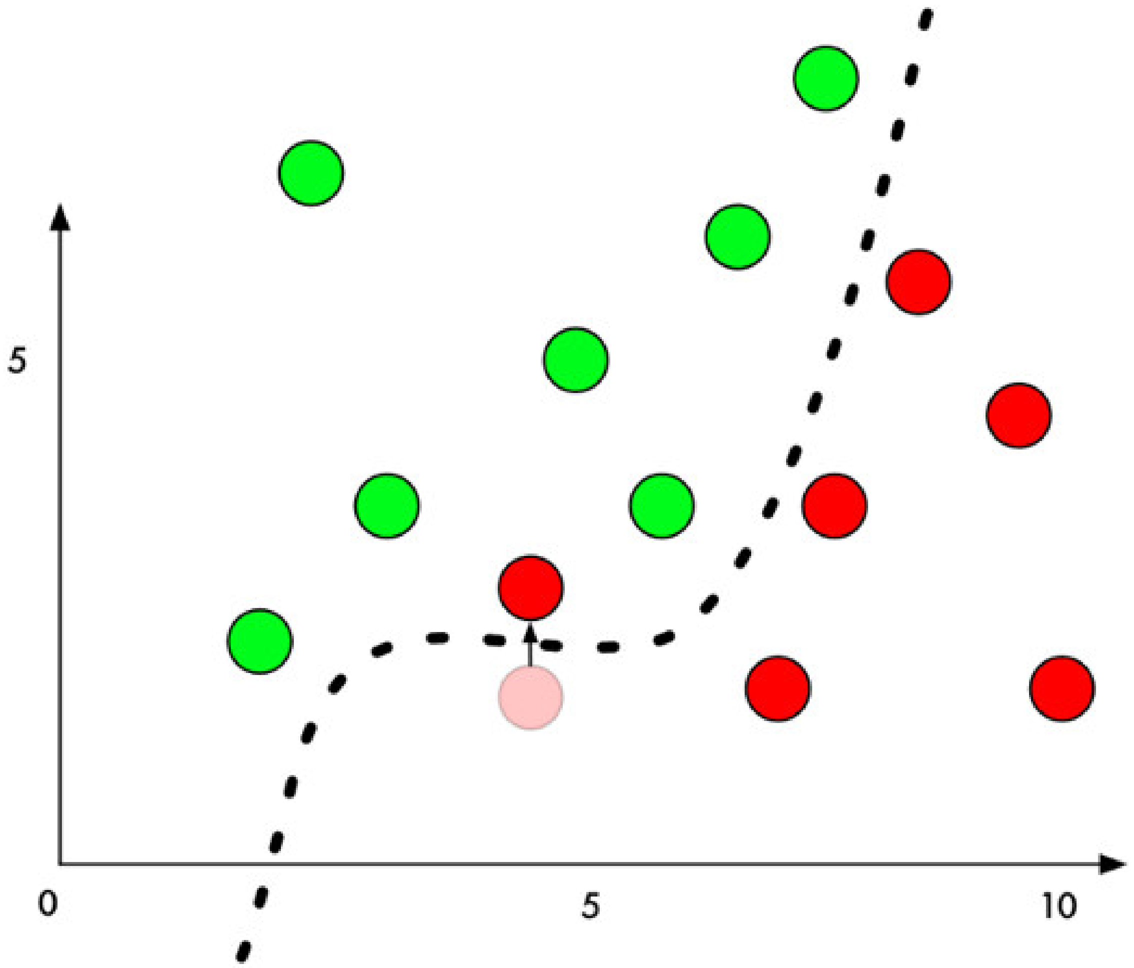


- Robustness plays a important role in AI safety



Fei Wu, et al. EURASIP Journal on Wireless Communications and Networking, 173 (2020)

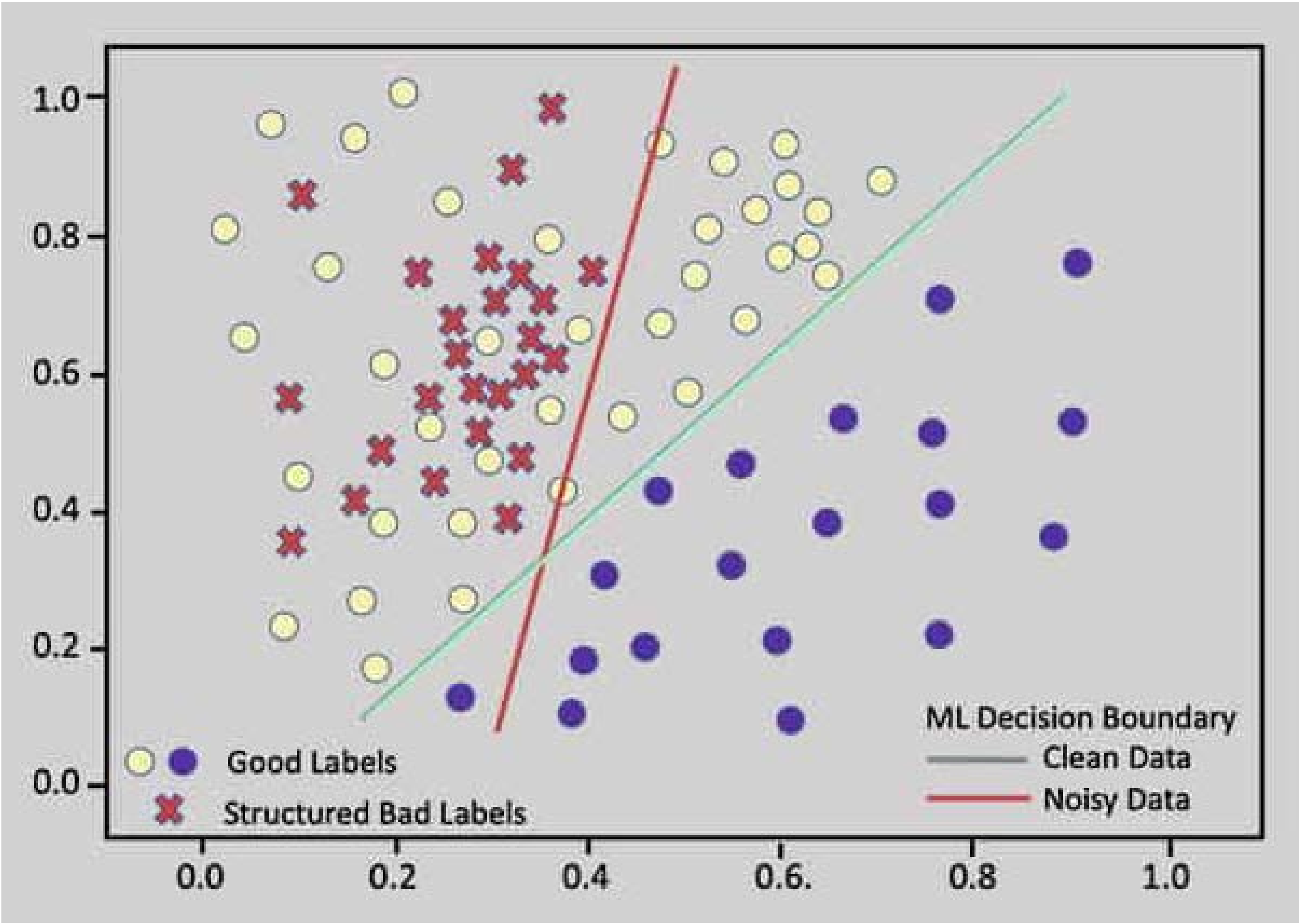
Introduction



$$\mathbf{x}^* = \arg\max_{\mathbf{x}}(a_i(\mathbf{x}) - R_\theta(\mathbf{x}))$$

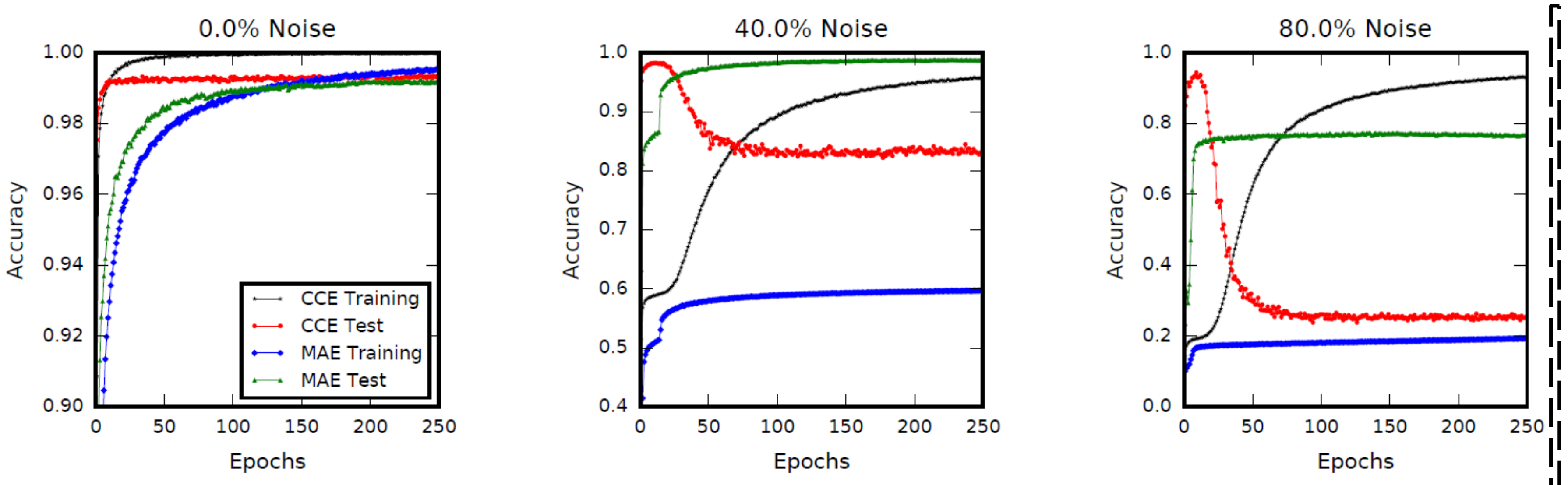
$$\tilde{X}^n_{adv} = X^{n-1}_{adv} - \alpha * \operatorname{sign}(\nabla_{X^n}(J(X^n|y_{\text{target}}))),$$

$$X^n_{adv} = \operatorname{clip}\left(\tilde{X}^n_{adv}, [X - \epsilon, X + \epsilon]\right),$$



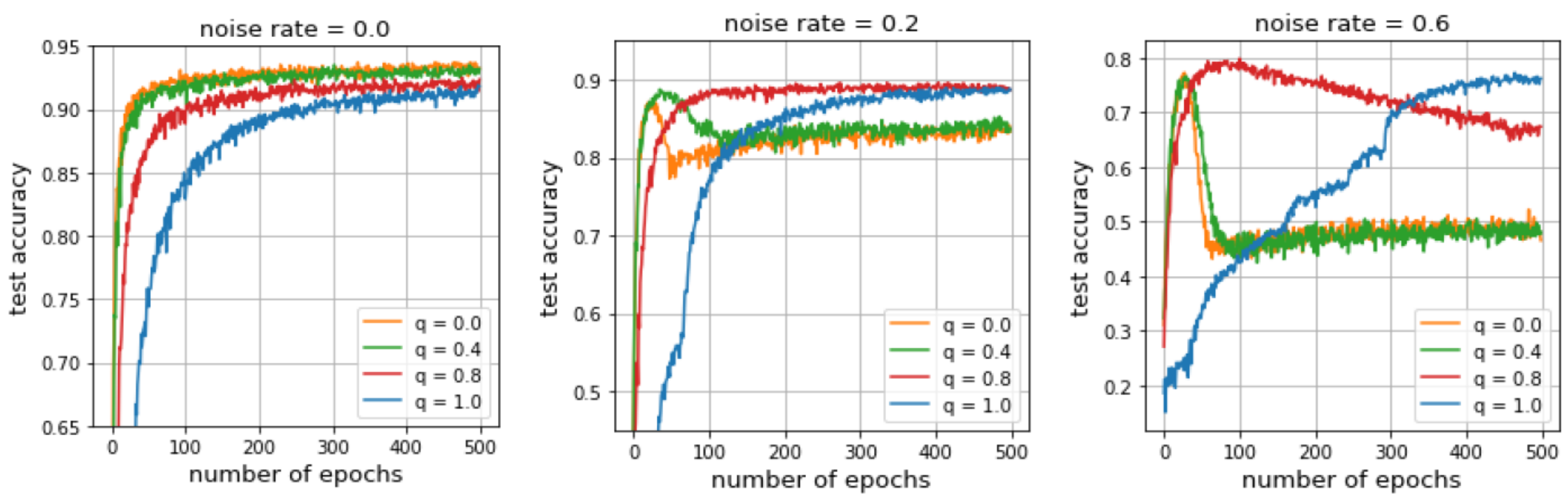
- Supervised Machine Learning requires labeled training data
- Usually, training data can be error-prone, adding ‘label noise’ to training sets
- DNN models should be trained with noisy labels operate effectively

Loss to Robustness



$$\sum_{i=1}^{n} \frac{\partial \mathcal{L}(f(\boldsymbol{x}_i;\boldsymbol{\theta}), y_i)}{\partial \boldsymbol{\theta}} = \begin{cases} \sum_{i=1}^{n} -\frac{1}{f_{y_i}(\boldsymbol{x}_i;\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} f_{y_i}(\boldsymbol{x}_i;\boldsymbol{\theta}) & \text{for CCE} \\ \sum_{i=1}^{n} -\nabla_{\boldsymbol{\theta}} f_{y_i}(\boldsymbol{x}_i;\boldsymbol{\theta}) & \text{for MAE/unhinged loss} \end{cases}$$

Aritra Ghosh, et al. Robust loss functions under label noise for deep neural networks. In AAAI 2017

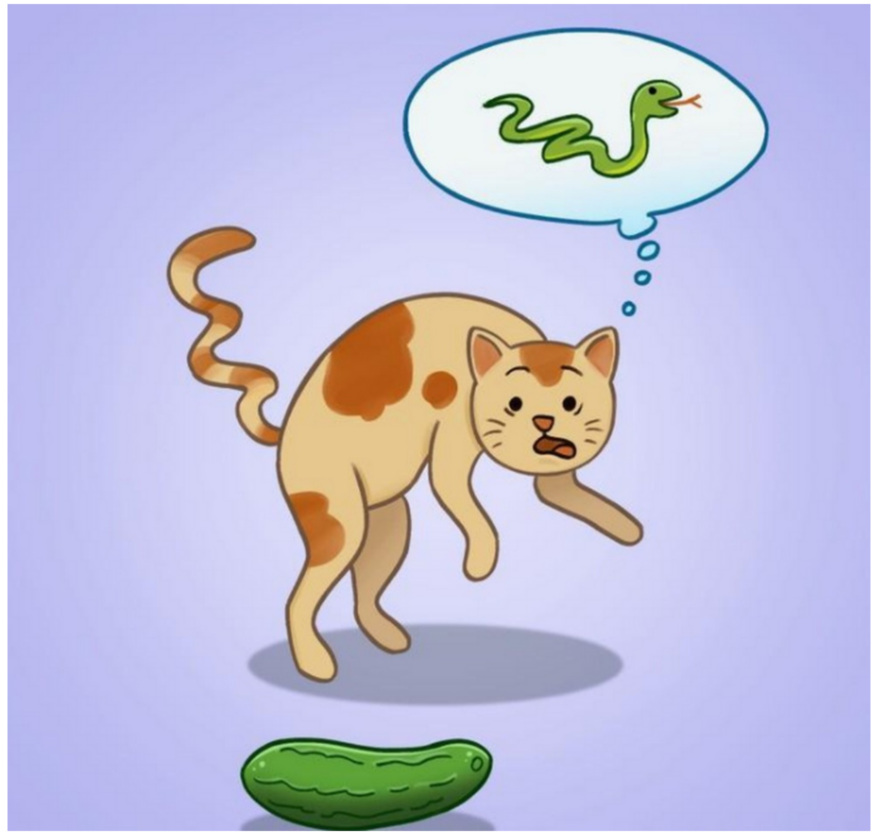


$$\mathcal{L}_q(f(\boldsymbol{x}), \boldsymbol{e}_j) = \frac{(1 - f_j(\boldsymbol{x})^q)}{q}$$

Zhilu Zhang, et al. Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels. arXiv 2018

- CE are implicitly weighed more than samples with predictions that agree more with provided labels in the gradient update
- MAE treats every sample equally, which makes it more robust to noisy labels
- MAE can concurrently cause increased difficulty in training, and lead to performance drop
- GCE is a trade-off loss function between performance and robustness

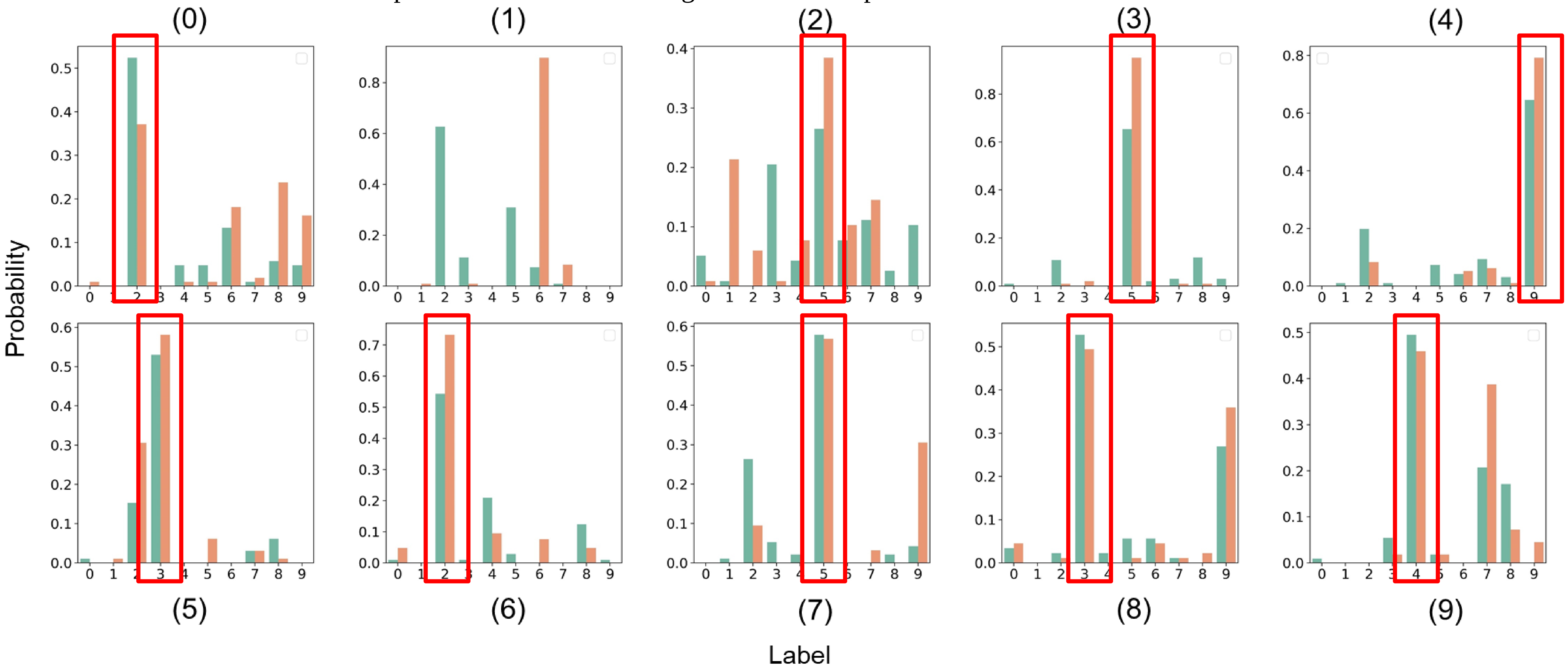
AI Roubustness v. s. Human Robustness



Relationship Between Adversarial Examples and 2nd Softmax

Green bar: the probabilities of softmax outputs after attack

Red bar: the probabilities of second largest softmax outputs before attack



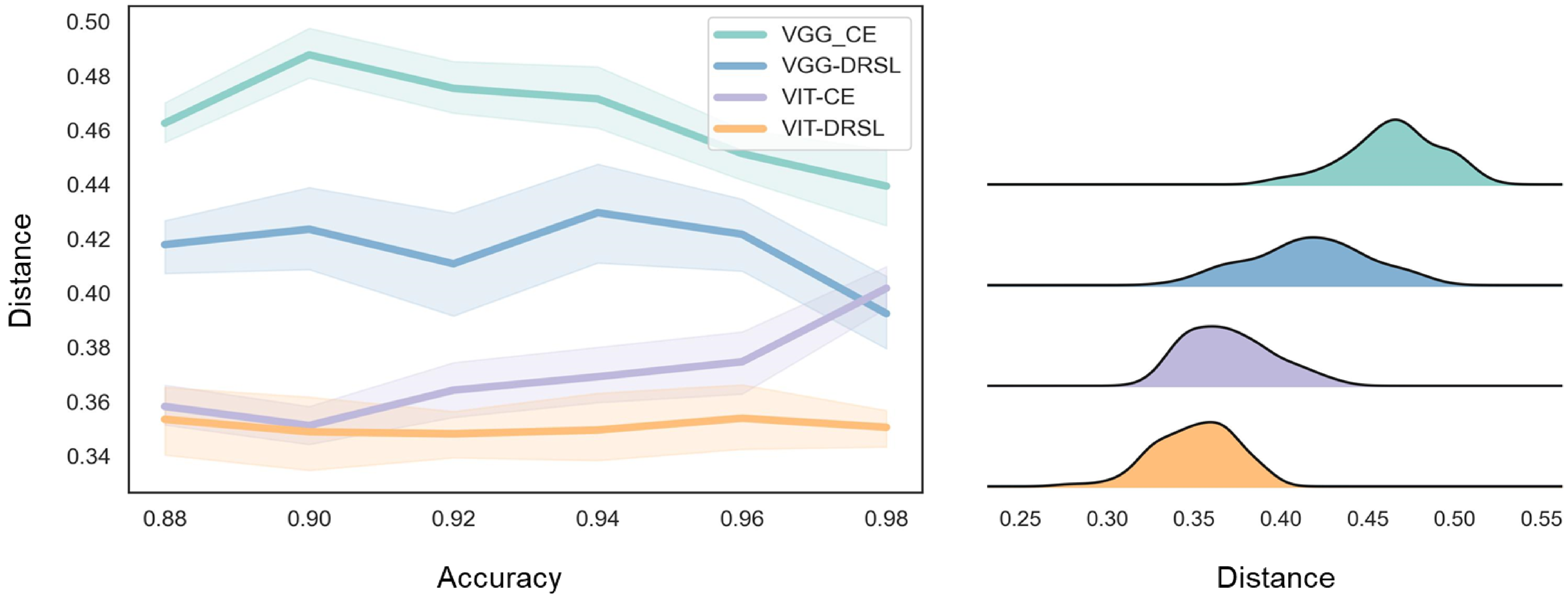
- After attacks, the second largest softmax probabilities trends to be the largest one
- One hypothesis: as the second largest softmax probabilities decrease, it becomes more difficult for an adversarial attack to manipulate them into becoming the largest one

Distribution-Restrained Softmax Loss Function (DRSL)

$$L(f(x;\theta),y) = -1_y^T \log(softmax(f(x;\theta)))$$

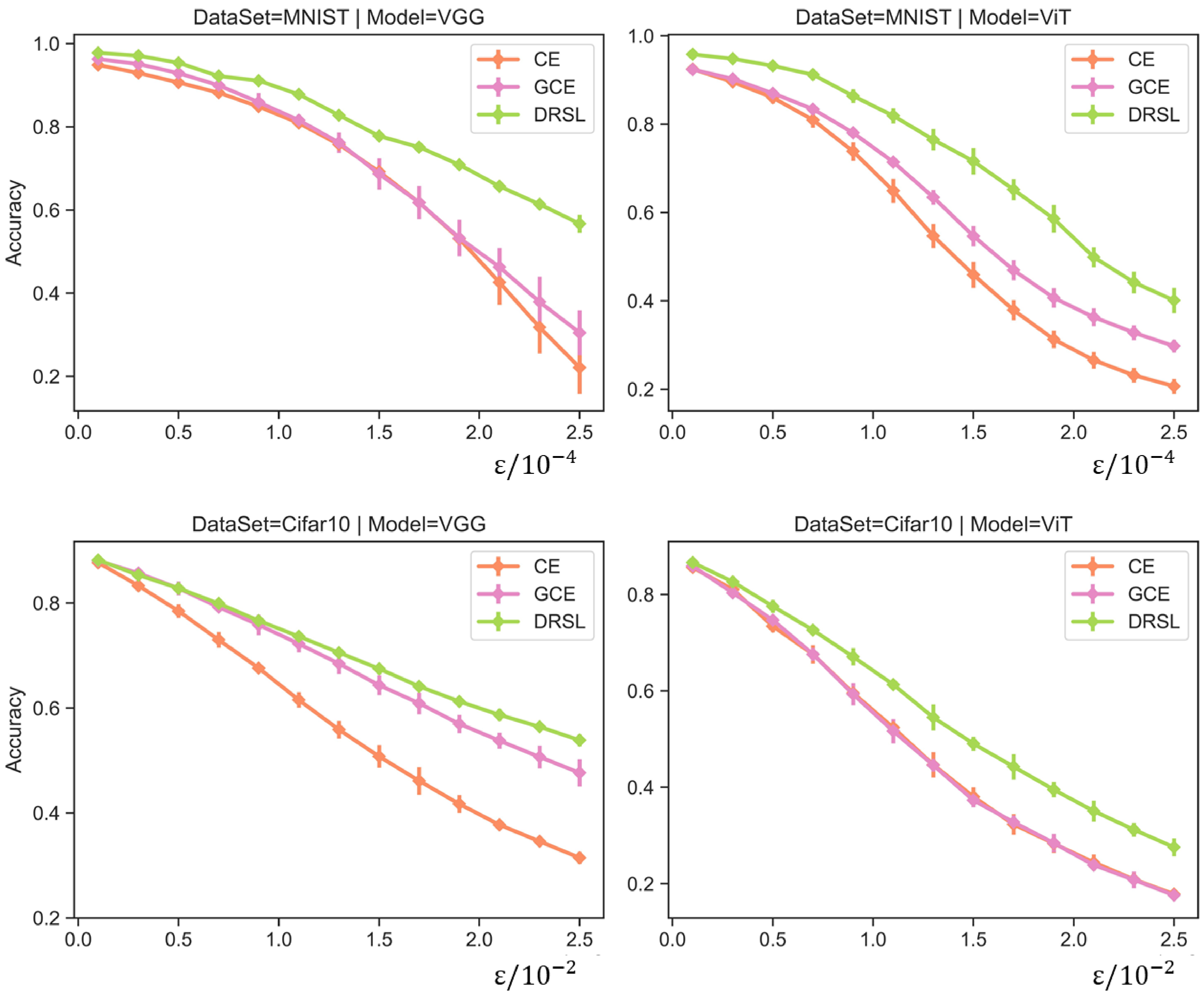
$$\Rightarrow$$

$$L(f(x;\theta),y) = -1_y^T \log(softmax(f(x;\theta))) + \tau \cdot d(softmax(f(x;\theta)), avg)$$



- Model accuracy has nothing to do with the softmax distribution
- DRSL restrain the softmax distribution

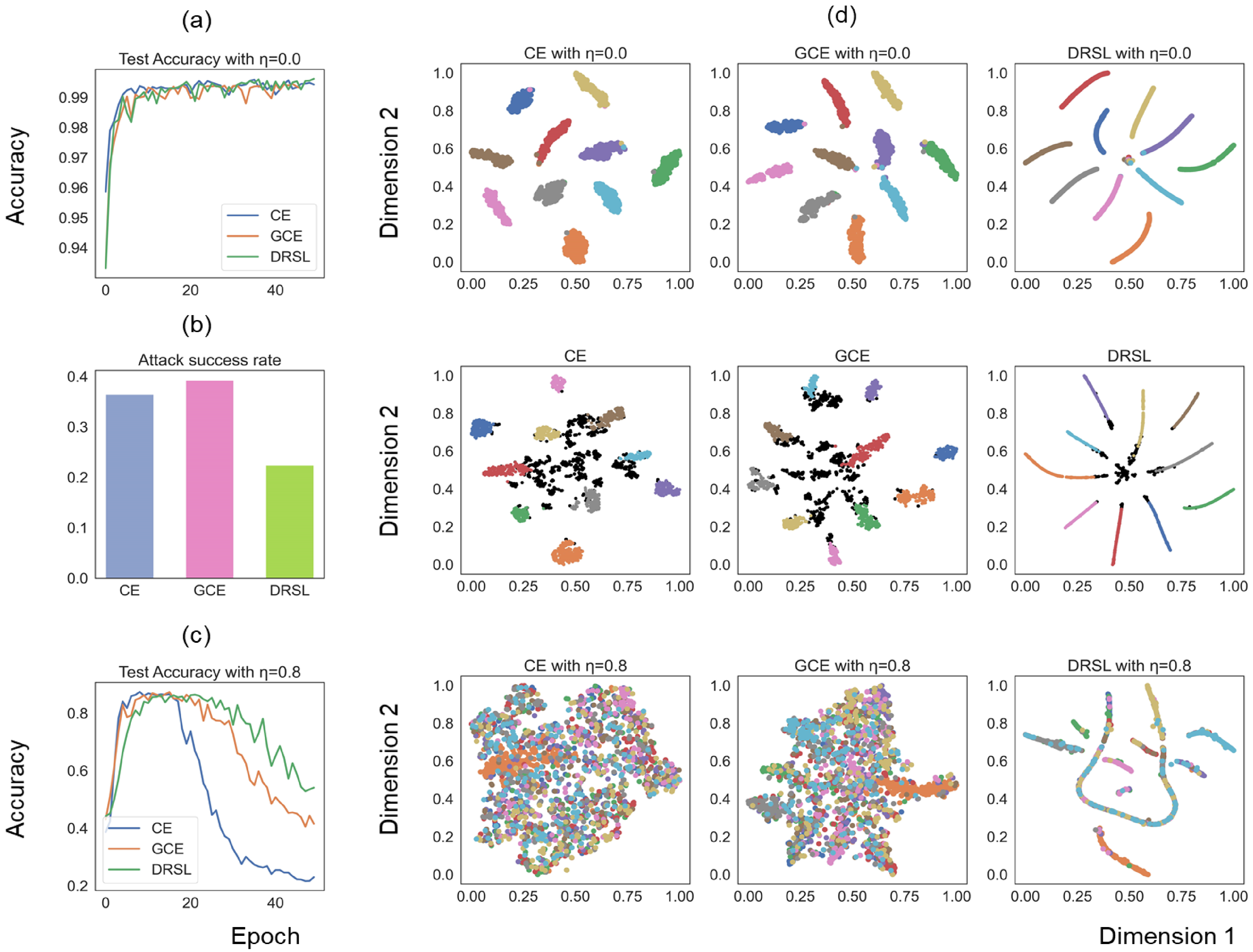
Robustness of CE/DRSL based model



Models/DataSets	CE	GCE	DRSL
ViT/MNIST	0.953±0.0013	0.953±0.0024	0.952±0.0016
VGG/MNIST	0.964±0.0011	0.963±0.0013	0.962±0.0012
VGG/Cifar10	0.894±0.0012	0.896±0.0017	0.893±0.0023
ViT/Cifar10	0.883±0.0014	0.883±0.0016	0.881±0.0017

- DRSL is more robust than other methods in same precision level
- DRSL can be extended to more models

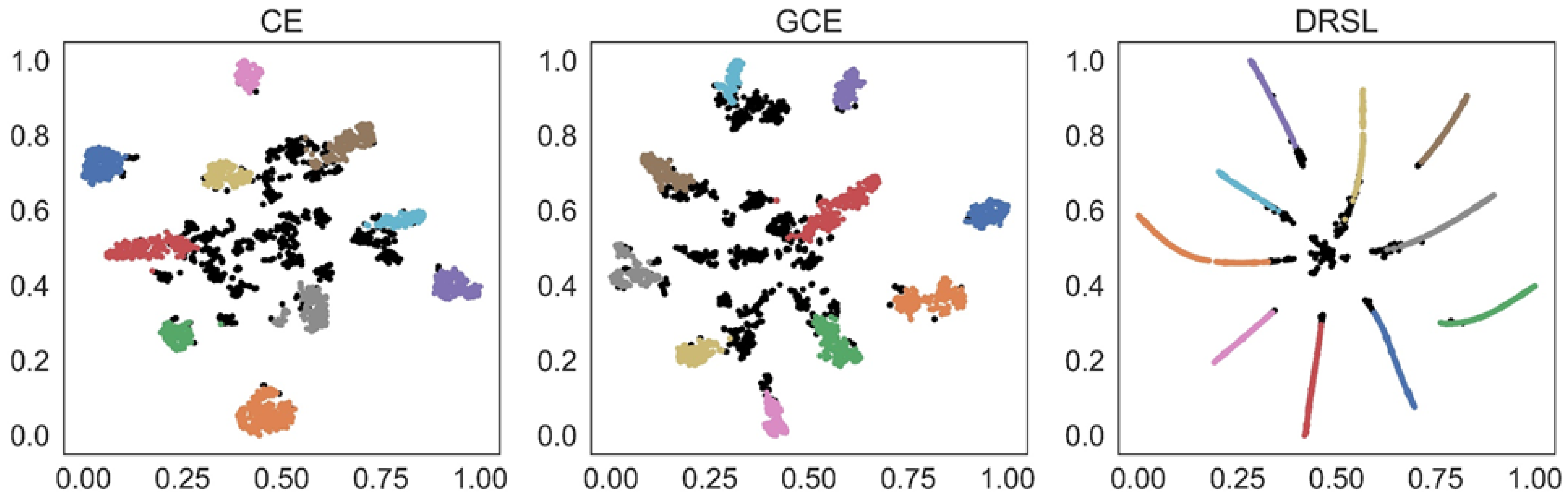
Visualization of Distribution-Restrained Softmax



- All the loss functions achieve a similar precision level
- DRSL is difficult to attack
- DRSL is robust in label noise
- After attacks, adversarial examples of other two models are diffusion in the reduced space, while DRSL's are concentrated at the tip of clusters

Conclusions

$$L(f(x;\theta),y) = -1_y^T \log(softmax(f(x;\theta))$$
$$+\tau \cdot d(softmax(f(x;\theta)), avg)$$



- We identified a significant factor that affects the robustness of models: the distribution characteristics of softmax values for non-real label samples
- the results after an attack are highly correlated with the distribution characteristics
- After the distribution diversity of softmax is suppressed in loss function, a significant improvement of model robustness were found
- DRSL can be applied not only to classification models but also to other softmax-inclusive models, such as generative models, which inspires us to further investigate and explore of the method

Thank You

Thanks to all authors of the present work