Boundary thickness and robustness in learning models

Chao Chen

LEHIGH UNIVERSITY

Notations

In a classification task, the goal is to train a classifier $f(x): \mathcal{X} \rightarrow [0,1]^C$.

$f(x)_i$ is the posterior (after the softmax) for the class i, $\Pr(y = i|x)$.

$g_{ij}(x) = f(x)_i - f(x)_j$ is the difference of posterior of x in class i and j.

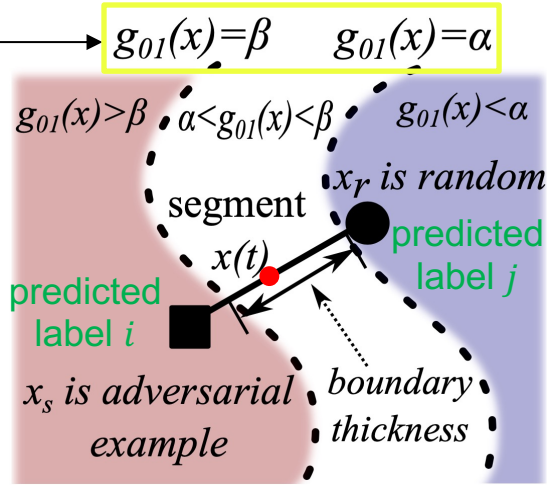
Boundary Thickness

For $\alpha \in (-1,1)$ and $\beta \in (-1,1)$, and a distribution over pairs of points $(x_r, x_s) \sim p$, the predicted labels of x_r and x_s are i and j, respectively. The boundary thickness of $f(\cdot)$ is:

$$\Theta(f,\alpha,\beta,p) = \mathbb{E}_{(x_r,x_s)\sim p}\left[|x_r - x_s| \int_0^1 \mathbf{I}\{\alpha < g_{ij}(x(t)) < \beta\} dt\right],$$

where $g_{ij}(x) = f(x)_i - f(x)_j$, $\mathbf{I}\{\cdot\}$ is the indicator function, and $x(t) = tx_r + (1-t)x_s, t \in [0,1]$.

1. $x(0) = x_s,\ x(1) = x_r$. Thus, $x(t)$ is the line segments between x_r and x_s.
2. $\mathbf{I}\{\alpha < g_{ij}(x(t)) < \beta\}$ checks if $x(t)$ locates within the "boundary" of
3. $\int_0^1 \mathbf{I}\{\alpha < g_{ij}(x(t)) < \beta\} dt$ "counts" the prob of $x(t)$ in the "boundary".
4. $|x_r - x_s|$ measures the distance between x_r and x_s.
5. $\Theta(f,\alpha,\beta,p)$ measures the distance between $g_{ij}(x) = \beta$ and $g_{ij}(x) = \alpha$.



Boundary Thickness

For $\alpha \in (-1,1)$ and $\beta \in (-1,1)$, and a distribution over pairs of points $(x_r, x_s) \sim p$, the predicted labels of x_r and x_s are i and j, respectively. The boundary thickness of $f(\cdot)$ is:

$$\Theta(f, \alpha, \beta, p) = \mathbb{E}_{(x_r, x_s)\sim p}\left[|x_r - x_s| \int_0^1 \mathbf{I}\{\alpha < g_{ij}(x(t)) < \beta\} dt\right],$$

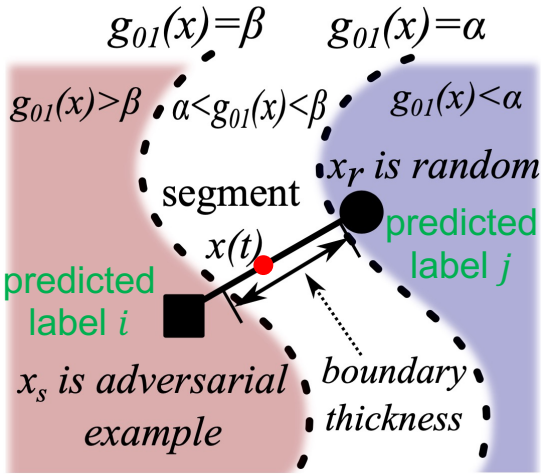
where $g_{ij}(x) = f(x)_i - f(x)_j$, $\mathbf{I}\{\cdot\}$ is the indicator function, and $x(t) = tx_r + (1-t)x_s, t \in [0,1]$.

Choice of p:

In the paper, x_r with label i is uniformly chosen from the training set.

x_s with random target label $j \neq i$ is an ℓ_2 adversarial sample.

[Specific instances of p recovers margin and mixup regularization.]



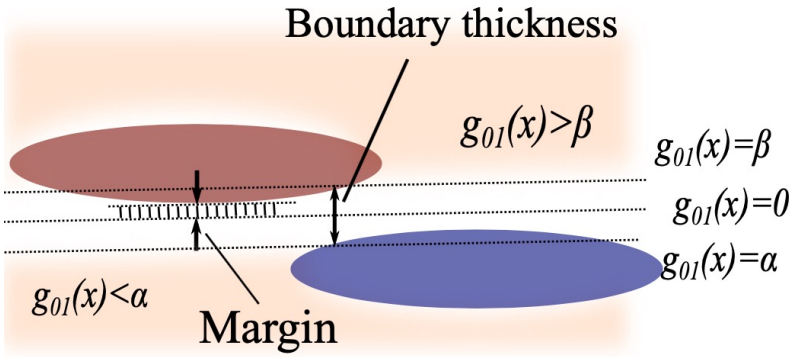
Boundary Thickness

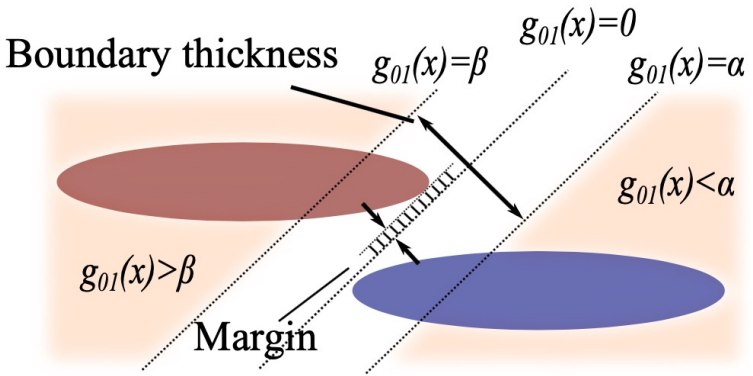
A binary linear classifier $f(x) = [f(x)_0, f(x)_1] = [\sigma(w^T x + b), 1 - \sigma(w^T x + b)]$, where $\sigma(\cdot)$ is the sigmoid function.

Same data is used to train the linear model in both figures.

Same margin (minimal distance to $g_{01}(x) = 0$) for both models.

The model in right has higher boundary thickness.





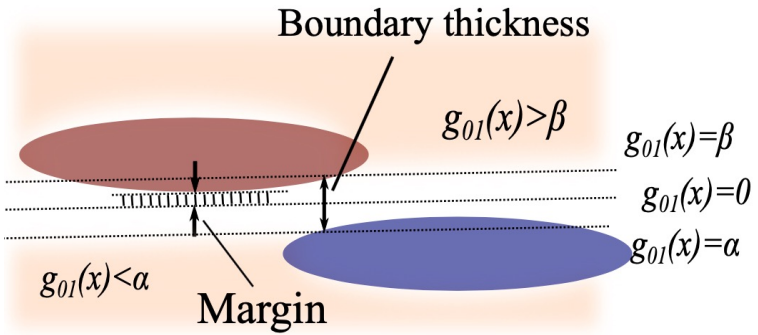
Boundary Thickness – Binary Linear Classifier

$$\Theta(f,\alpha,\beta,p)=\mathbb{E}_{(x_r,x_s)\sim p}\left[|x_r-x_s|\int_0^1 \mathbf{I}\{\alpha<g_{ij}(x(t))<\beta\}dt\right],$$

A binary linear classifier $f(x)=[f(x)_0,f(x)_1]=[\sigma(w^Tx+b),1-\sigma(w^Tx+b)]$, where $\sigma(\cdot)$ is the sigmoid function. The adversarial sample x_s of x_r should satisfy $x_s-x_r=cw, c\in\mathbb{R}$.

Let $\tilde{g}(\cdot)\coloneqq 2\sigma(\cdot)-1$. When $[\alpha,\beta]\subset[g_{01}(x_r),g_{01}(x_s)]$, the thickness of the binary linear classifier is:

$$\Theta(f,\alpha,\beta)=\left(\tilde{g}^{-1}(\beta)-\tilde{g}^{-1}(\alpha)\right)/|w|$$



Boundary Thickness – Binary Linear Classifier

$$\Theta(f,\alpha,\beta,p)=\mathbb{E}_{(x_r,x_s)\sim p}\left[|x_r-x_s|\int_0^1 \mathbf{I}\{\alpha<g_{ij}(x(t))<\beta\}dt\right],$$

A binary linear classifier $f(x)=[f(x)_0,f(x)_1]=[\sigma(w^Tx+b),1-\sigma(w^Tx+b)]$, where $\sigma(\cdot)$ is the sigmoid function. The adversarial sample x_s of x_r should satisfy $x_s-x_r=cw, c\in\mathbb{R}$.

Proof:

Define a new substitute variable u and corresponding du, $g_{01}(u)$:

$$u=tw^Tx_r+(1-t)w^Tx_s+b$$

$$du=w^Tx_r-w^Tx_s=w^T(-cw)dt=-c|w|^2dt$$

$$g_{01}(x(t))=f(tx_r+(1-t)x_s)_0-f(tx_r+(1-t)x_s)_1=2\sigma(tw^Tx_r+(1-t)w^Tx_s+b)-1=2\sigma(u)-1=\tilde{g}(u)$$

Thus,

$$\Theta(f,\alpha,\beta,p)=c|w|\mathbb{E}\left[\int_0^1 \mathbf{I}\{\alpha<f(tx_r+(1-t)x_s)_0-f(tx_r+(1-t)x_s)_1<\beta\}dt\right]$$

$$\Theta(f,\alpha,\beta,p)=c|w|*\left(\frac{1}{-c|w|^2}\right)\mathbb{E}\left[\int_{w^Tx_s+b}^{w^Tx_r+b}\mathbf{I}\{\alpha<\tilde{g}(u)<\beta\}du\right]$$

$$\Theta(f,\alpha,\beta,p)=\frac{1}{|w|}\mathbb{E}\left[\int_{w^Tx_r+b}^{w^Tx_s+b}\mathbf{I}\{\alpha<\tilde{g}(u)<\beta\}du\right]$$

Boundary Thickness – Binary Linear Classifier

$$\Theta(f,\alpha,\beta,p)=\mathbb{E}_{(x_r,x_s)\sim p}\left[|x_r-x_s|\int_0^1 \mathbf{I}\{\alpha<g_{ij}(x(t))<\beta\}dt\right],$$

A binary linear classifier $f(x)=[f(x)_0,f(x)_1]=[\sigma(w^Tx+b),1-\sigma(w^Tx+b)]$, where $\sigma(\cdot)$ is the sigmoid function. The adversarial sample x_s of x_r should satisfy $x_s-x_r=cw, c\in\mathbb{R}$.

Proof:

$$\Theta(f,\alpha,\beta,p)=\frac{1}{|w|}\mathbb{E}\left[\int_{w^Tx_r+b}^{w^Tx_s+b}\mathbf{I}\{\alpha<\tilde{g}(u)<\beta\}du\right]$$

Since $\tilde{g}(\cdot)=2\sigma(\cdot)-1$ is monotonically increasing:

$$\Theta(f,\alpha,\beta,p)=\frac{1}{|w|}\mathbb{E}\left[\int_{w^Tx_r+b}^{w^Tx_s+b}\mathbf{I}\{\tilde{g}^{-1}(\alpha)<u<\tilde{g}^{-1}(\beta)\}du\right]$$

$$\Theta(f,\alpha,\beta,p)=\frac{1}{|w|}\mathbb{E}[\min(\tilde{g}^{-1}(\beta),w^Tx_s+b)-\max(\tilde{g}^{-1}(\alpha),w^Tx_r+b)]$$

When $[\alpha,\beta]\subset[g_{01}(x_r),g_{01}(x_s)]$, $\tilde{g}^{-1}(\beta)<\tilde{g}^{-1}\big(g_{01}(x_s)\big)=\tilde{g}^{-1}\big(\tilde{g}(w^Tx_s+b)\big)=w^Tx_s+b$ and $\tilde{g}^{-1}(\alpha)>w^Tx_r+b$:

$$\Theta(f,\alpha,\beta,p)=\frac{\tilde{g}^{-1}(\beta)-\tilde{g}^{-1}(\alpha)}{|w|}$$

Boundary Thickness generalizes margin

Example: for a binary SVM, we choose α and β as the values of $\tilde{g}(\cdot)$ at two support vectors, such that $\tilde{g}^{-1}(\alpha) = -1$ and $\tilde{g}^{-1}(\beta) = 1$. Then the boundary thickness is

$$\Theta(f, \alpha, \beta) = \frac{\left(\tilde{g}^{-1}(\beta) - \tilde{g}^{-1}(\alpha)\right)}{|w|} = \frac{2}{|w|},$$

which is the same as the margin of SVM.

The margin of general model f on dataset $\mathcal{D} = \{x_k\}_{k=0}^{n-1}$ is defined as

$$\text{Margin}(\mathcal{D}, f) = \min_k \min_{j \neq y_k} |x_k - \text{Proj}(x_k, j)|$$

where $\text{Proj}(x, j) = \arg \min_{x' \in S(i_x, j)} |x' - x|$ is the projection onto the decision boundary $S(i_x, j)$.

$S(i, j) = \{x \in \mathcal{X} : f(x)_i = f(x)_j\}$ is the decision boundary between classes i and j.

Boundary Thickness generalizes margin

$$\text{Margin}(\mathcal{D}, f) = \min_{k} \min_{j \neq y_k} |x_k - \text{Proj}(x_k, j)|$$

$$\Theta(f, \alpha, \beta, p) = \mathbb{E}_{(x_r, x_s) \sim p} \left[|x_r - x_s| \int_0^1 \mathbf{I}\{\alpha < g_{ij}(x(t)) < \beta\} dt \right]$$

Q: When will Boundary thickness reduces to margin?

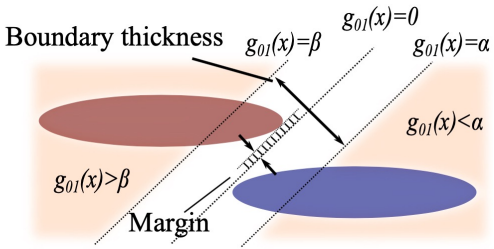
A: $\alpha = 0, \beta = 1$, and p is chosen that x_s is the projection $\text{Proj}(x_r, j)$ for the worse case class j.

Formally, choose x_r randomly from $\mathcal{D} = \{x_k\}_{k=0}^{n-1}$, with predicted label $i = \arg\max_l f(x)_l$. For another class $j \neq i$, choose $x_s = \text{Proj}(x_r, j)$

$$\min_{x_r} \min_{j \neq i} |x_r - x_s| \int_0^1 \mathbf{I}\{\alpha < g_{ij}(x(t)) < \beta\} dt = \text{Margin}(\mathcal{D}, f)$$

Left-hand side is a “worst-case” of boundary thickness.

Usually impractical to compute the margin but more straightforward to measure the thickness.



Boundary Thickness and Robustness

For non-adversarial training:

Train without weight decay, train with standard weight decay, mixup training.

(Robustness from low to high.)

Learning rate is initiated with 0.1, and at epoch 100 and 150, reduce to 0.01 and 0.001.

Use $\alpha = 0, \beta = 0.75$ and ℓ_2 PGD-20 to evaluate thickness.

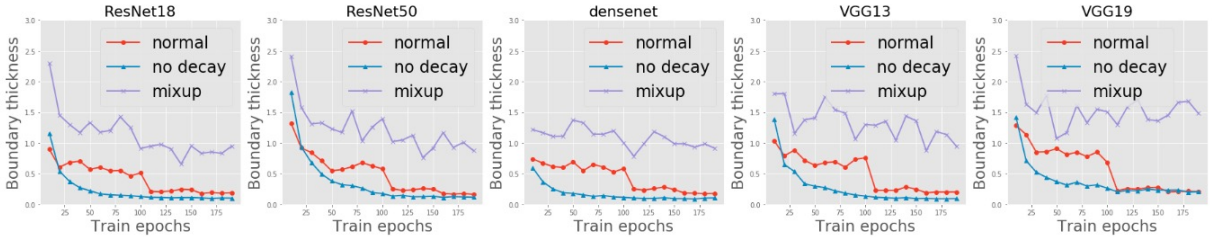
For adversarial training:

Five different settings: large learning rate, ℓ_2 regularization, ℓ_1 regularization, early stop, cutout.

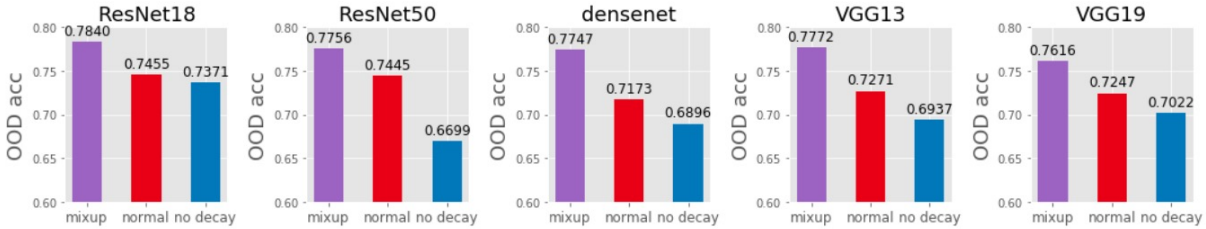
Boundary Thickness and Robustness

For non-adversarial training:

1. Thickness of mixup is larger than normal and no-decay (consistent with robustness).
2. Learning rate decay on epochs 100 and 150 reduces the boundary thickness.
3. OOD (out-of-distribution) robustness corresponds to boundary thickness.



(a) Thickness: mixup > normal training > training without weight decay. After learning rate decays (at both epoch 100 and 150), decision boundaries get thinner.



(b) OOD robustness: mixup > normal training > training without weight decay. Compare with Figure 2a to see that mixup increases thickness, while training without weight decay reduces thickness.

Boundary Thickness and Robustness

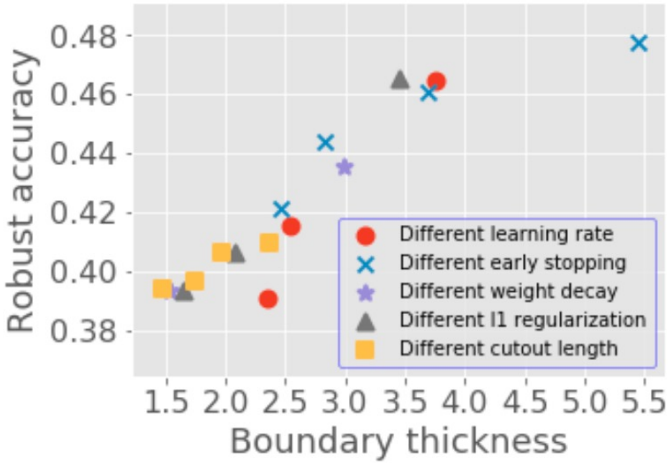
For adversarial training:

1. Robust accuracy is consistent with boundary thickness.

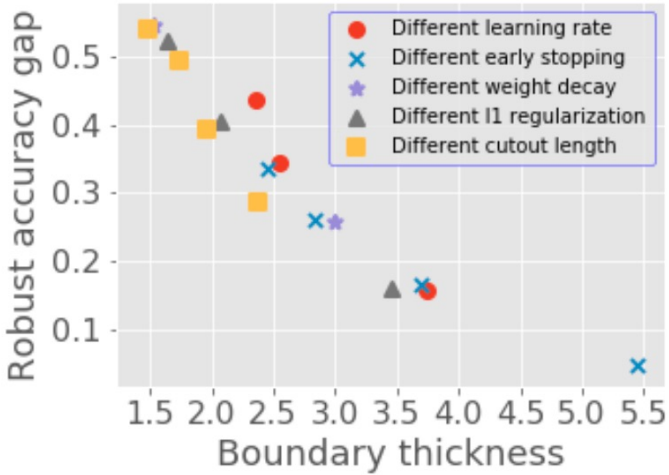
2. Robust accuracy gap is consistent with boundary thickness.

(Increasing boundary thickness reduces overfitting.)

[only models with 90%+ accuracy are plotted.]



(a)



(b)

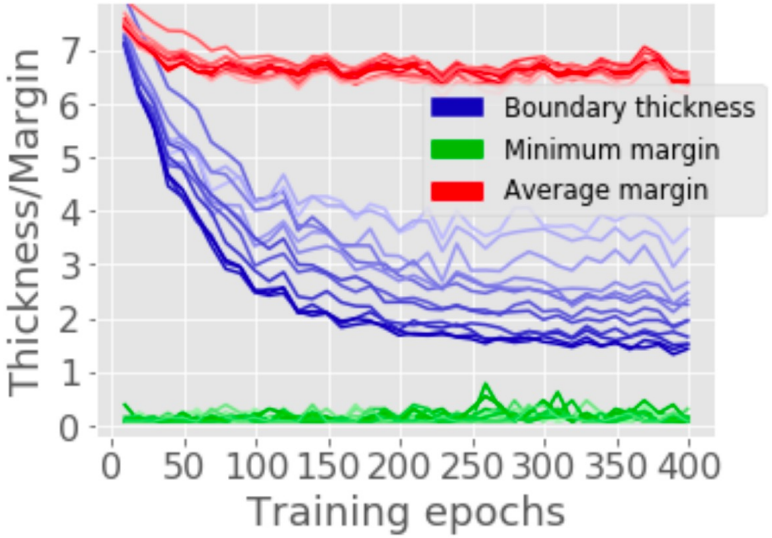
Boundary Thickness and Robustness

Boundary thickness versus margin:

Darker curves represent less robust models.

Boundary thickness capture the robustness, while margin cannot.

The minimum margin (usually used) is almost zero in all cases.

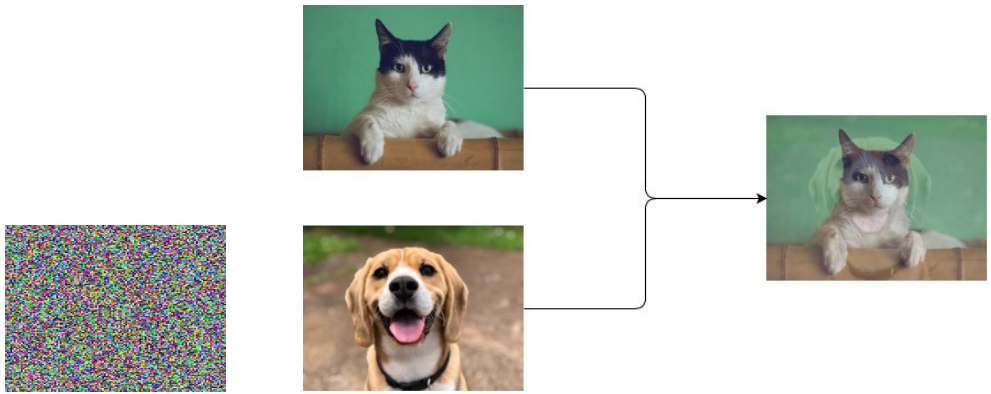


Application of Boundary Thickness

Noisy mixup:

In the ordinary mixup, we obtain mixup samples x by linearly combining two data samples x_1 and x_2.

In the noisy-mixup, one of x_1 and x_2, with some probability p, is replaced by an image that consists of random noise. The label of the noisy image is “NONE.” Specifically, in the CIFAR10 dataset, we let the “NONE” class be the 11-th class.

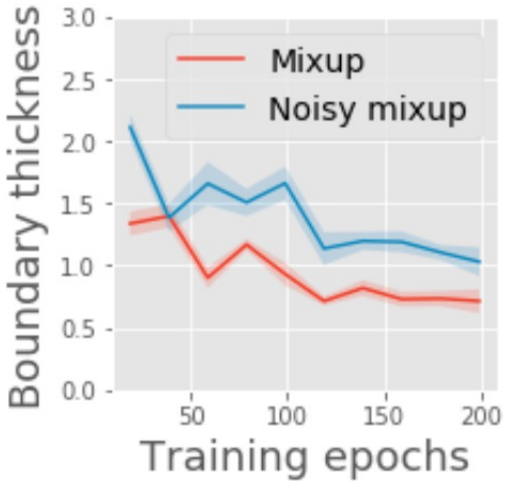


Application of Boundary Thickness

Results:

[For black-box attack, we use ResNet-110 to generate the transfer attack.]

1. Noisy mixup has thicker boundary.
2. Noisy mixup has significantly better test accuracy w.r.t. OOD, black-box and adversarial attacks by sacrificing some clean accuracy.



Dataset	Method	Clean	OOD	Black-box	PGD-20		
					8-pixel	6-pixel	4-pixel
CIFAR10	Mixup	**96.0**±0.1	78.5±0.4	46.3±1.4	2.0±0.1	3.2±0.1	6.3±0.1
	Noisy mixup	94.4±0.2	**83.6**±0.3	**78.0**±1.0	**11.7**±3.3	**16.2**±4.2	**25.7**±5.0
CIFAR100	Mixup	**78.3**±0.8	51.3±0.4	37.3±1.1	0.0±0.0	0.0±0.0	0.1±0.0
	Noisy mixup	72.2±0.3	**52.5**±0.7	**60.1**±0.3	**1.5**±0.2	**2.6**±**0.1**	**6.7**±0.9

Thank you

LEHIGH UNIVERSITY