Reactive and Proactive Measures for Adversarial Defense

Ashwath Shetty, Pankhuri Vanjani, Shreyash Arya

UNIVERSITÄT DES SAARLANDES

Problem Statement

Proactive

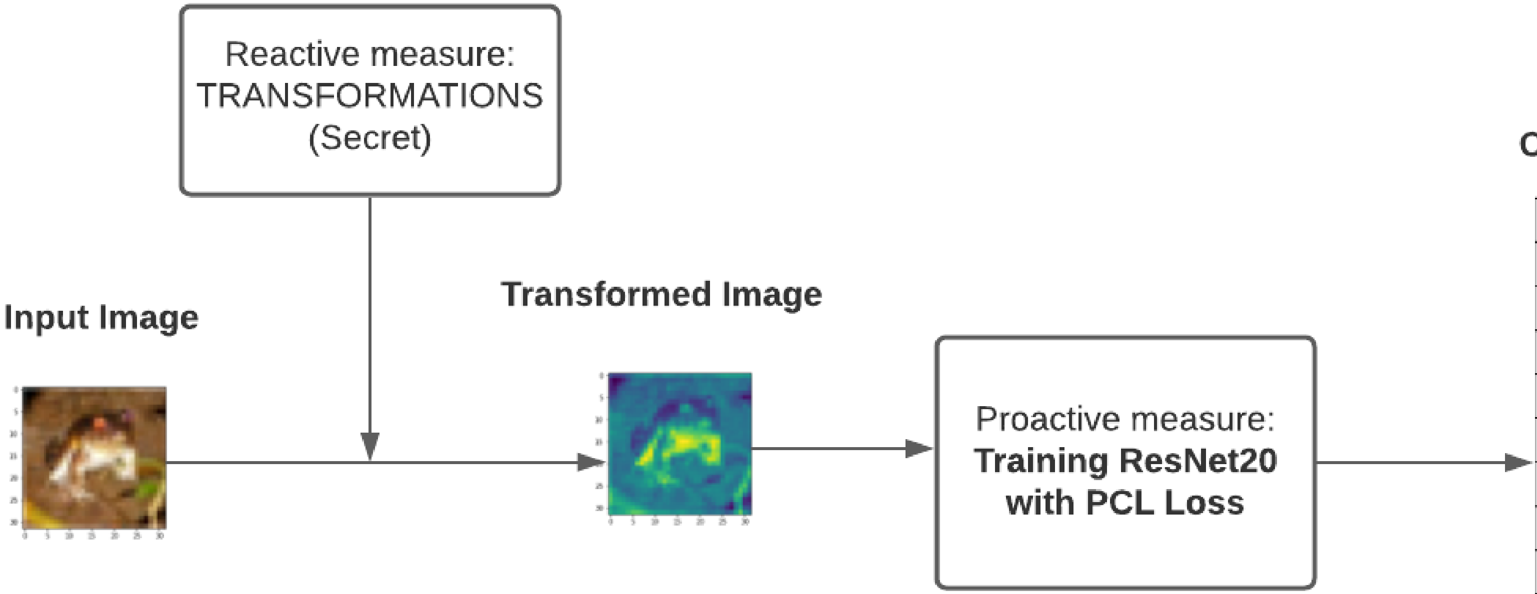
Alter the underlying architecture or learning procedure

E.g. by adding more layers, ensemble/adversarial training or changing the loss/activation functions

Reactive

Modify the inputs during testing time, using image transformations to counter the effect of adversarial perturbation

Our Approach



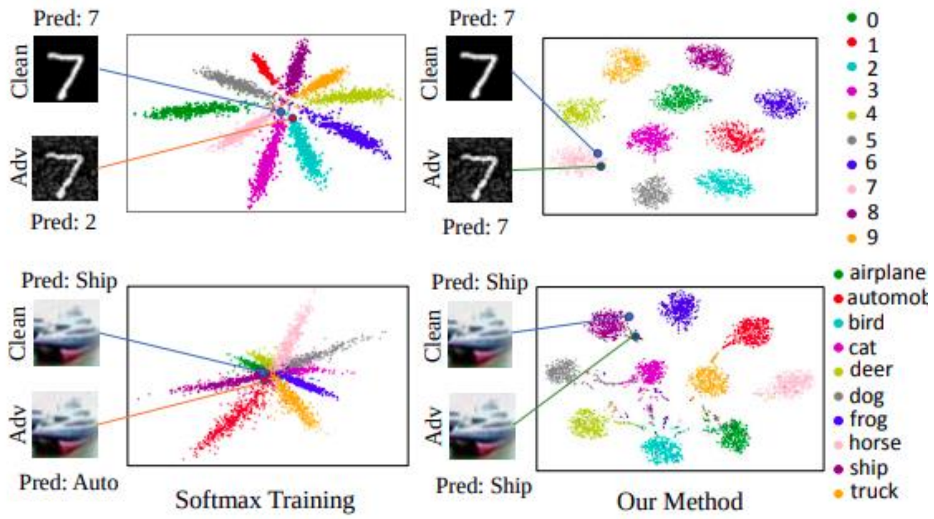
Classification Labels

index	label
0	**airplane** (0)
1	**automobile** (1)
2	**bird** (2)
3	**cat** (3)
4	**deer** (4)
5	**dog** (5)
6	**frog** (6)
7	**horse** (7)
8	**ship** (8)
9	**truck** (9)
...	...

Threat model : adversary with only black box access

Proactive Defense (PCL)

Enforcing Class Separation in Feature Space provides additional adversarial robustness.



Source: https://github.com/aamir-mustafa/pcl-adversarial-defense

PCL Loss function

$$\mathcal{L}_{\mathrm{CE}}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{r} -\log \frac{\exp\left(\boldsymbol{w}_{y_i}^T \boldsymbol{f}_i + \boldsymbol{b}_{y_i}\right)}{\sum_{j=1}^{k} \exp\left(\boldsymbol{w}_j^T \boldsymbol{f}_i + \boldsymbol{b}_j\right)}$$

Standard Cross Entropy Loss

f_i : penultimate layer outputs
w_{yi} : ground truth class label
w_j , b_j are the weights and biases for the j^{th} output neuron.

PCL Loss function

$$\mathcal{L}_{\mathrm{PC}}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i} \left\{ \boxed{\left\| \boldsymbol{f}_i - \boldsymbol{w}^c_{y_i} \right\|_2^2} - \frac{1}{k-1} \sum_{j \neq y_i} \left(\left\| \boldsymbol{f}_i - \boldsymbol{w}^c_j \right\|_2^2 + \left\| \boldsymbol{w}^c_{y_i} - \boldsymbol{w}^c_j \right\|_2^2 \right) \right\}$$

Prototype conformity loss

Minimize Intra-Class Distance

f_i : penultimate layer outputs
w_{yi} : ground truth class label for i^{th} example
w_j^c : Cluster centre for class j
w_j , b_j : weights and biases for the j^{th} output neuron.

PCL Loss function

$$\mathcal{L}_{\mathrm{PC}}(\boldsymbol{x}, \boldsymbol{y}) = \sum_i \left\{ \left\| \boldsymbol{f}_i - \boldsymbol{w}^c_{y_i} \right\|_2^2 - \boxed{\frac{1}{k-1} \sum_{j \neq y_i} \left(\left\| \boldsymbol{f}_i - \boldsymbol{w}^c_j \right\|_2^2 + \left\| \boldsymbol{w}^c_{y_i} - \boldsymbol{w}^c_j \right\|_2^2 \right)} \right\}$$

Prototype conformity loss

Maximize Intra-Class Distance

f_i : penultimate layer outputs
w_{yi} : ground truth class label for i^{th} example
w_j^c : Cluster centre for class j
w_j , b_j : weights and biases for the j^{th} output neuron.

PCL Loss function

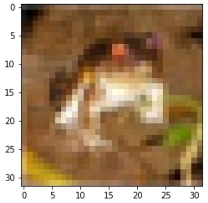
$$\mathcal{L}_{\mathrm{CE}}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{r} -\log \frac{\exp\left(\boldsymbol{w}_{y_i}^T \boldsymbol{f}_i + \boldsymbol{b}_{y_i}\right)}{\sum_{j=1}^{k} \exp\left(\boldsymbol{w}_j^T \boldsymbol{f}_i + \boldsymbol{b}_j\right)}$$

$$\mathcal{L}_{\mathrm{PC}}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i} \left\{ \|\boldsymbol{f}_i - \boldsymbol{w}_{y_i}^c\|_2^2 - \frac{1}{k-1} \sum_{j \neq y_i} \left(\|\boldsymbol{f}_i - \boldsymbol{w}_j^c\|_2^2 + \|\boldsymbol{w}_{y_i}^c - \boldsymbol{w}_j^c\|_2^2 \right) \right\}$$

PCL Training Loss function:

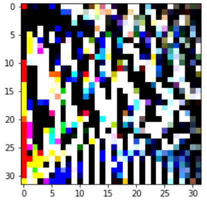
$$\boxed{\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{L}_{\mathrm{CE}}(\boldsymbol{x}, \boldsymbol{y}) + \mathcal{L}_{\mathrm{PC}}(\boldsymbol{x}, \boldsymbol{y})}$$

Reactive Defense (Transformations)



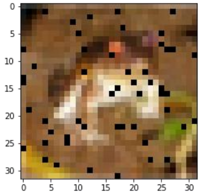
Original Image

Frequency Domain



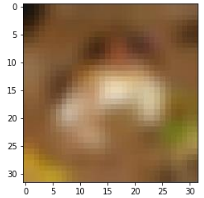
DCT

Probabilistic

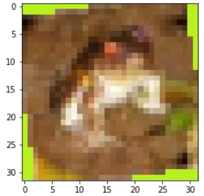


Dropout

Geometric & Image Filter

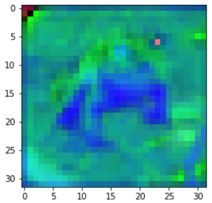


Gaussian Blur

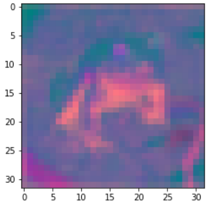


Affine

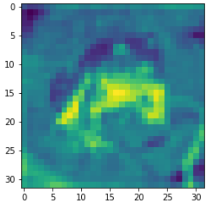
Color Space



HSV



YCbCr



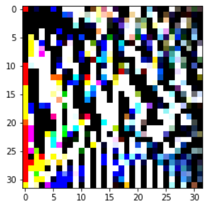
Grayscale

Experimental Settings

- Dataset: CIFAR-10
- Model Architecture: Resnet 20
- Baselines: Softmax (L_{ce}) training, PCL ($L_{ce}+L_{pcl}$) training
- Epsilon values of attacks: 8/255 and 16/255
- Black box attack examples: Madry Labs[1]

1. https://github.com/MadryLab/cifar10_challenge?utm_source=catalyzex.com

Results: Discrete Cosine Transform (DCT)

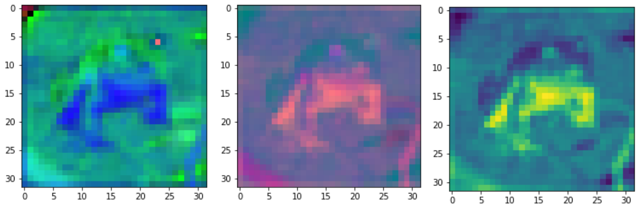


Transformations	Clean Accuracy	Black Box Accuracy eps=8/255	Black Box Accuracy eps=16/255
Baseline			
Softmax (Lce)	**90.13**	9.06	3.110
PCL (Lce+Lpcl)	89.69	12.24	3.970
Fourier domain			
DCT + Softmax (Lce)	81.540	61.88	45.75
DCT+PCL	80.990	**64.600**	**48.74**
Color Space			
GreyScale+ Softmax (Lce)	88.85	9.510	5.29
GreyScale+PCL	88.050	11.59	6.29
HSV+ Softmax (Lce)	90.68	27.820	16.99
HSV+PCL	90.450	28.940	17.24
YCrCb + Softmax (Lce)	90.800	8.740	4.54
YCrCb + PCL	90.410	9.110	4.05

- (DCT) + PCL Training outperforms the baseline and other transformations on black box attack examples
  - Transform significantly alters image -> weakens black box attack
- Drop in clean accuracy
  - Image compression -> Loss of Information

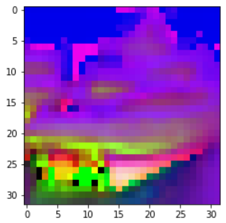
Trade-off due to image alteration!!

Results: Color space transform

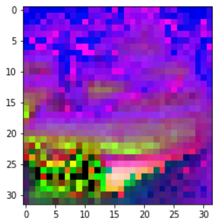


Transformations	Clean Accuracy	Black Box Accuracy eps=8/255	Black Box Accuracy eps=16/255
Baseline			
Softmax (Lce)	**90.13**	9.06	3.110
PCL (Lce+Lpcl)	89.69	12.24	3.970
Fourier domain			
DCT + Softmax (Lce)	81.540	61.88	45.75
DCT+PCL	80.990	**64.600**	**48.74**
Color Space			
GreyScale+ Softmax (Lce)	88.85	9.510	5.29
GreyScale+PCL	88.050	11.59	6.29
HSV+ Softmax (Lce)	90.68	27.820	16.99
HSV+PCL	90.450	28.940	17.24
YCrCb + Softmax (Lce)	90.800	8.740	4.54
YCrCb + PCL	90.410	9.110	4.05

- HSV color space gives a significant boost
  - Adversarial examples -> more perceptible

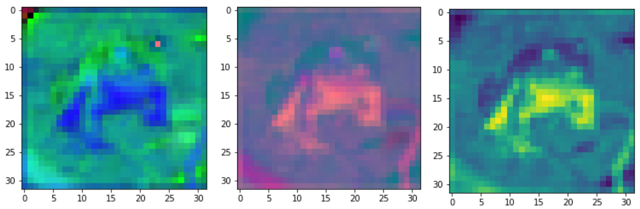


Without attack



With adv attack

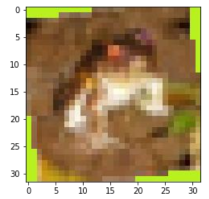
Results: Color space transform



Transformations	Clean Accuracy	Black Box Accuracy eps=8/255	Black Box Accuracy eps=16/255
Baseline			
Softmax (Lce)	**90.13**	9.06	3.110
PCL (Lce+Lpcl)	89.69	12.24	3.970
Fourier domain			
DCT + Softmax (Lce)	81.540	61.88	45.75
DCT+PCL	80.990	**64.600**	**48.74**
Color Space			
GreyScale+ Softmax (Lce)	88.85	9.510	5.29
GreyScale+PCL	88.050	11.59	6.29
HSV+ Softmax (Lce)	90.68	27.820	16.99
HSV+PCL	90.450	28.940	17.24
YCrCb + Softmax (Lce)	90.800	8.740	4.54
YCrCb + PCL	90.410	9.110	4.05

- HSV color space gives a significant boost
  - Adversarial examples -> more perceptible
- No effect of YCrCb and Grayscale
  - Spatial alignment of image features

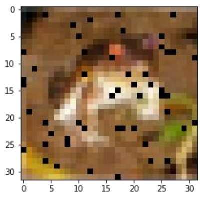
Results: Affine transform



Transformations	Clean Accuracy	Black Box Accuracy eps=8/255	Black Box Accuracy eps=16/255
Baseline			
Softmax (Lce)	**90.13**	9.06	3.110
PCL (Lce+Lpcl)	89.69	12.24	3.970
Geometric			
Affine+ Softmax (Lce)	88.63	48.170	27.660
Affine+PCL	88.650	**50.570**	**29.22**
Pixel Dropout			
Dropout(5%)+ Softmax (Lce)	90.68	35.100	16.520
Dropout(5%)+ PCL	89.930	35.400	17.95
Blur			
Gaussian Blur + Softmax (Lce)	84.640	28.390	13.31
Gaussian Blur+ PCL	84.330	27.130	14.56

- Affine transformation (fixed)-> demonstrates a significant boost
  - Spatial dis-alignment of important features in between trained and black-box model

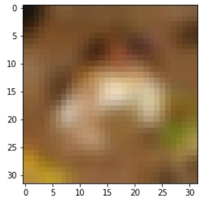
Results: Dropout



Transformations	Clean Accuracy	Black Box Accuracy eps=8/255	Black Box Accuracy eps=16/255
Baseline			
Softmax (Lce)	**90.13**	9.06	3.110
PCL (Lce+Lpcl)	89.69	12.24	3.970
Geometric			
Affine+ Softmax (Lce)	88.63	48.170	27.660
Affine+PCL	88.650	**50.570**	**29.22**
Pixel Dropout			
Dropout(5%)+ Softmax (Lce)	90.68	35.100	16.520
Dropout(5%)+ PCL	89.930	35.400	17.95
Blur			
Gaussian Blur + Softmax (Lce)	84.640	28.390	13.31
Gaussian Blur+ PCL	84.330	27.130	14.56

- Dropout: Some improvement in black box accuracy!!
  - network cannot over-rely on a particular pixel for classification

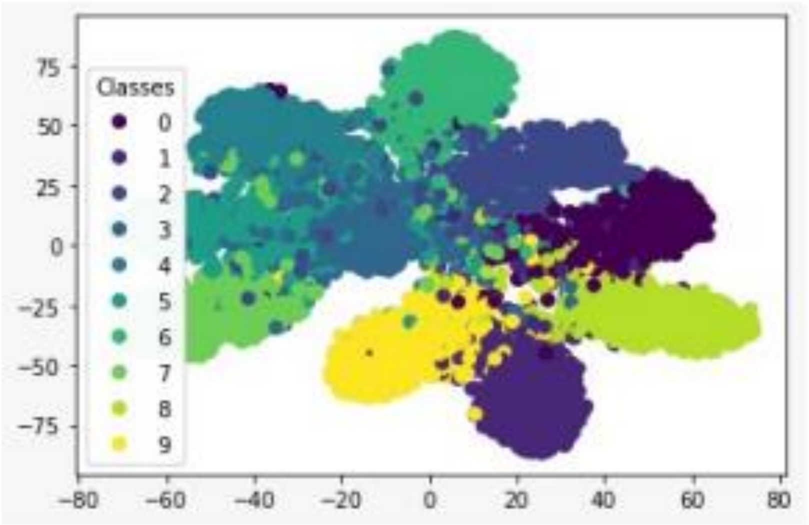
Results: Gaussian Blur



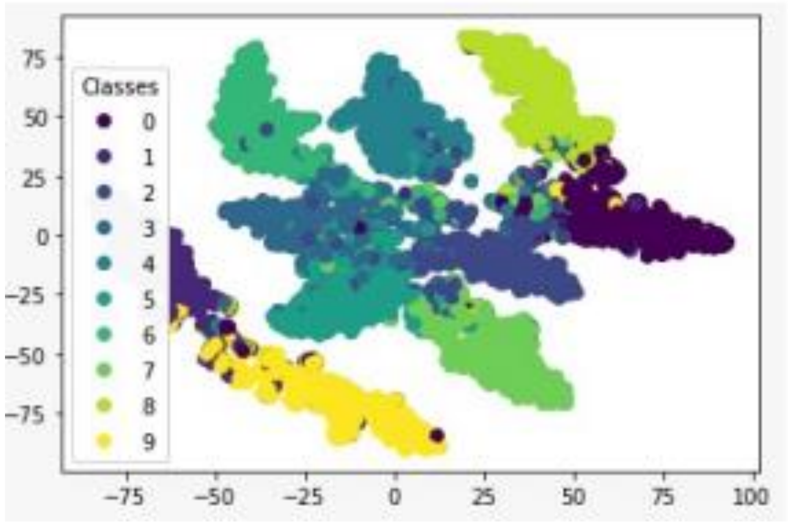
Transformations	Clean Accuracy	Black Box Accuracy eps=8/255	Black Box Accuracy eps=16/255
Baseline			
Softmax (Lce)	**90.13**	9.06	3.110
PCL (Lce+Lpcl)	89.69	12.24	3.970
Geometric			
Affine+ Softmax (Lce)	88.63	48.170	27.660
Affine+PCL	88.650	**50.570**	**29.22**
Pixel Dropout			
Dropout(5%)+ Softmax (Lce)	90.68	35.100	16.520
Dropout(5%)+ PCL	89.930	35.400	17.95
Blur			
Gaussian Blur + Softmax (Lce)	84.640	28.390	13.31
Gaussian Blur+ PCL	84.330	27.130	14.56

- Gaussian Blurring: 3x Improvement in black box accuracy
    - Blurring affects the weights learnt by the network
    - Distributes importance around nearby pixels

Class Feature Map from Penultimate Layer



Softmax Training

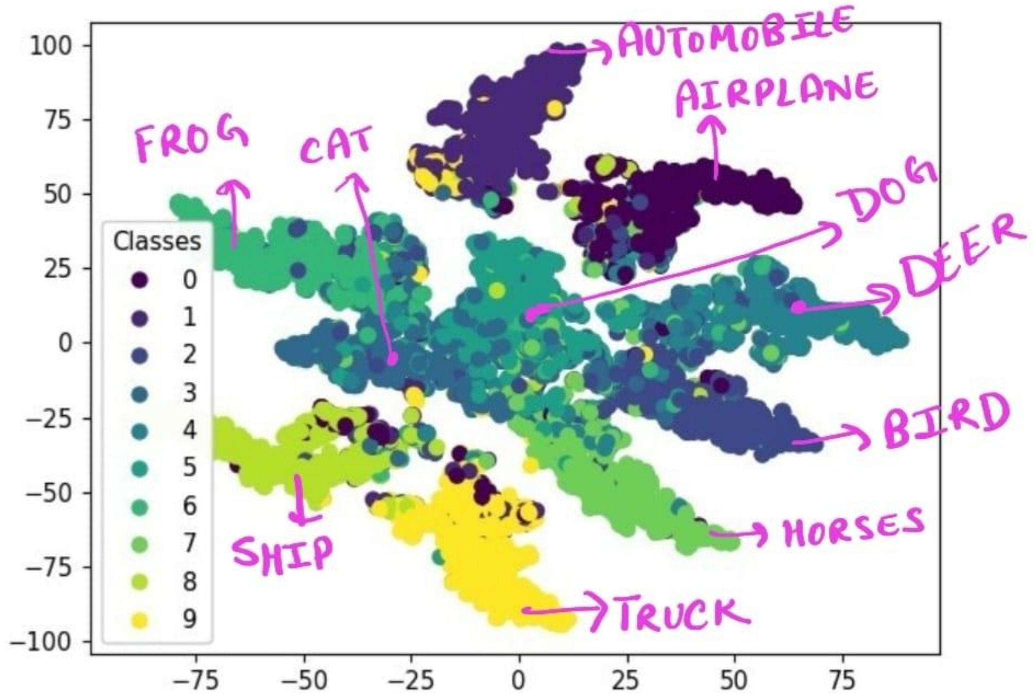


PCL Training

- Better separation with PCL training as compared to Softmax training (in all cases)

- With PCL: 1-3 % increase in bbox accuracy (in all cases)

Class Feature Map from Penultimate Layer



PCL Training

Feature plots of cat, dog and deer class overlapping the most

Potential cause of misclassification after adversarial attacks!

Summary and Improvements

- Training with transformation in general gave an additional boost in black box accuracy.
- Tranforms like DCT which significantly alters the image gave us the best result.
- Training with PCL loss improved feature clusters and gave 1-3% improvement
- Feature Positions seemed not to have a major contribution (supplementary slides)

Summary and Improvements

- Training with transformation in general gave an additional boost in black box accuracy.
- Tranforms like DCT which significantly alters the image gave us the best result.
- Training with PCL loss improved feature clusters and gave 1-3% improvement
- Feature Positions seemed not to have a major contribution (supplementary slides)

- Effect with adversarial training
- Analysis on individual class accuracy
- Adaptive Attack on System

Thank you!

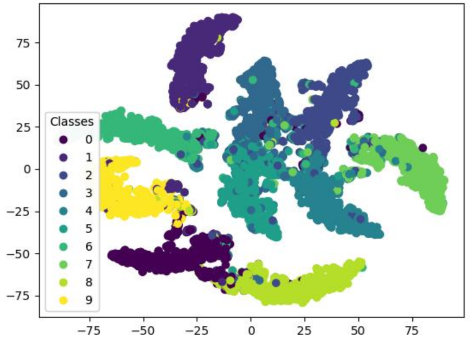
References

- *Aamir Mustafa, Salman Khan, Munawar Hayat, Roland Goecke, Jianbing Shen, and Ling Shao. Adversarial defense by restricting the hidden space of deep neural networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3385–3394, 2019. 1, 2, 3*
- https://github.com/MadryLab/cifar10_challenge

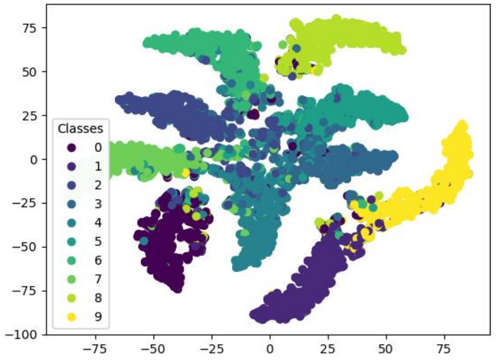
Supplementary Material

How positions of features in a plot would affect adversarial robustness?

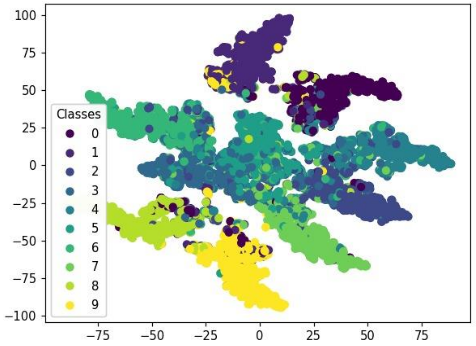
- Initially we thought networks which learnt different feature mappings would be more robust.
- But this turns out not to be the case
- For example, the grayscale transform, significantly moves the position of class 9,7, but still shows no improvement in robustness



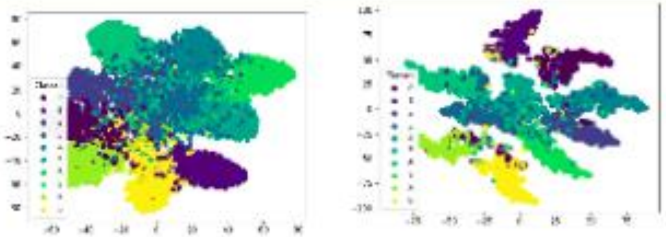
Original



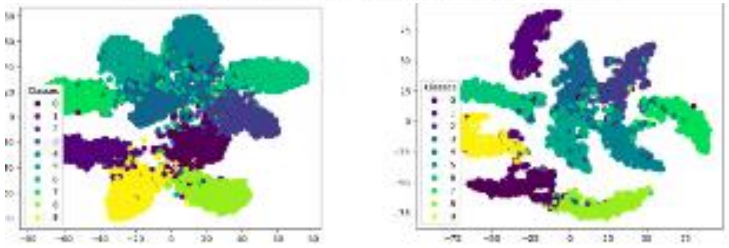
GrayScale



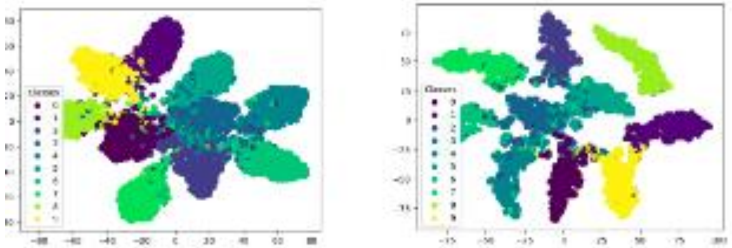
DCT



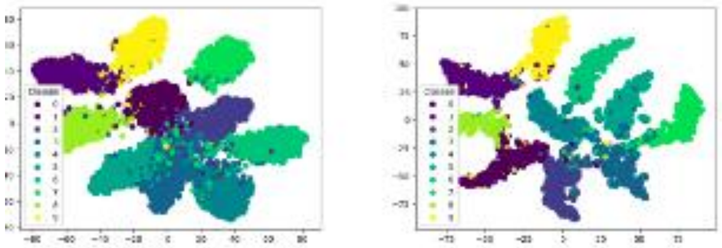
(a) DCT: Softmax Training (Left) PCL Training (Right)



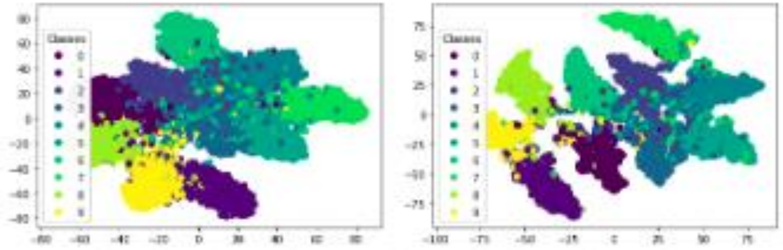
(a) Softmax Training (Left) PCL Training (Right)



(b) HSV: Softmax Training (Left) PCL Training (Right)

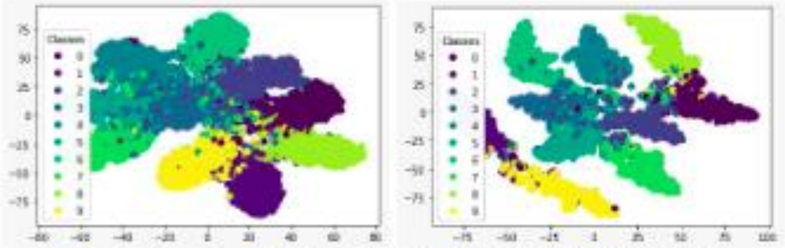


(c) YCbCr: Softmax Training (Left) PCL Training (Right)



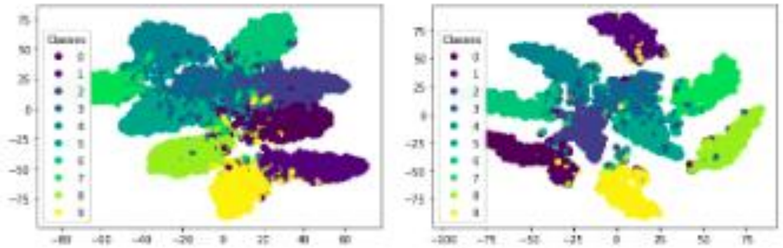
(a) Gaussian Blur: Softmax Training (Left) PCL Training (Right)

Figure 4. Gaussian Blur penultimate layer feature maps



(a) Affine Transform: Softmax Training (Left) PCL Training (Right)

Figure 5. Affine



(a) Dropout : Softmax Training (Left) PCL Training (Right)