Reactive and Proactive Measures for Adversarial Defense

Ashwath Shetty
assh00001

Pankhuri Vanjani
pava00002

Shreyash Arya
shar00001

Abstract

In this project we aim to study the effects of reactive and proactive attacks as an adversarial defense particularly on a base reactive defense (PCL as termed in [14]) that maximally separates features in intermediate layers in a deep learning model. Also, we study the effects of image transformations on feature space and adversarial example transferability. The code is publicly available on GitHub [1].

1. Introduction

With increasing popularity and applicability of machine/deep learning techniques, a study for its limitations and defence against these limitations is of paramount importance. Throughout the course, we have been seen that how easily machine learning systems can be broke and with wide usage of such systems in critical infrastructures, applications involving adversarial defense intrigued us the most to follow as a project. Also, [7] showed that we can't get away from adversarial examples, which was concerning.

From literature review, we saw a division into reactive (that transforms the input) and proactive (that alters the underlying training procedure) defenses but the combination of both was not present. This inspired us to explore this idea which aligns with hashing in modern cryptography. The intuition is a transformation hidden from an adversary that may reduce the transferability of the adversarial example, which would be followed by an additional layer of security provided by the proactive defense.

In this project, we work with black box threat models. We aim to train a proactive + reactive defense with feature transforms and analyze the robustness to adversarial attacks of these measures. Firstly, we analyse and evaluate metrics by augmentating the training framework with a transformed input and then, we visually analyze how training with a transformed input affects the distribution/location of features in projected 2D space.

2. Related work

Rising trend of research in achieving robustness in machine learning models has motivated researchers to get strongest attacks as well as design best defense methods to break those attacks. There are following two major defence categories:

Proactive Defenses

In the proactive defenses, recent work has explored modification of network architecture or new training techniques in defenses for Deep Learning. Previous works [16] [3] [10] have used self-supervised learning for vision tasks to improve the robustness. [6] and [11] explored learning and adversarial training, which jointly trains the model with clean and adversarial images proven to be effective. Some extended work along this path by [12] has shown good results by using ensemble adversarial training to mitigate attacks. Besides these, current research work has exploited the latent feature space and it's decision boundary to achieve higher accuracy against the adversarial attacks. In this direction [4] has increased linear regions of classifier and increased the distance from decision boundary by regularizing scheme. Other works restricting the feature space in convex polytopes by [17], [8], [15], [14] have also proved to be robust.

Reactive Defenses

Several works have proposed the idea of modifying the input via various transformations as a defense to improve robustness. In the work by [18] transformations in fourier space has been used as reactive defenses achieving good results. Chen et. al. [2] extended the work to a variety of transforms such as transformations in color geometric space.

Motivated by the above ideas, we extend the work of [15] and [14] which are based on proactive defense strategy in combination with reactive defenses for images. We aim to explore the combination of proactive and reactive defenses for robustness against adversarial attacks by analysing the separation of different classes in the dataset in latent feature space.

[1] https://github.com/pankhurivanjani/Proactive-and-Reactive-Measures-for-Adversarial-Defense

3. Methodology

Our overall pipeline can be summarized by the following: we integrated multiple transforms into the training pipeline of an existing reactive framework and observed how each transform provides additional robustness to black-box attacks. In the following subsections, we describe the reactive approach we built on top of and the image transformations we used. The threat model for our setup is a scenario where the adversary has only black-box access to our model.

3.1. Proactive Measures: Loss function

We decided to choose a modern proactive defense as a base paper (PCL) [14], particularly because we were interested in the idea of ensuring the feature space of a neural network was well partitioned as defense, we were also interested in seeing if training with input transformations or transformations in the feature space heavily changes those partitions and how it affects on transferability of adversarial examples and this work provided a perfect setup to this.

The core idea of their work is to ensure that the representations learnt by the network are well clustered for each class. They do this by adding a prototype conformity loss (equation 2) to the standard cross entropy loss (equation 1) in their final formulation (equation 3)

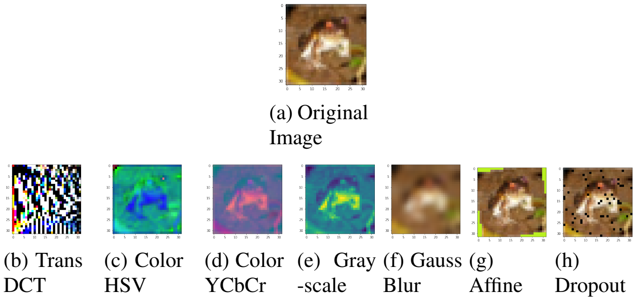
$$\mathcal{L}_{\mathrm{CE}}(\boldsymbol{x},\boldsymbol{y})=\sum_{i=1}^{r}-\log\frac{\exp\left(\boldsymbol{w}_{y_i}^{T}\boldsymbol{f}_i+\boldsymbol{b}_{y_i}\right)}{\sum_{j=1}^{k}\exp\left(\boldsymbol{w}_{j}^{T}\boldsymbol{f}_i+\boldsymbol{b}_{j}\right)} \tag{1}$$

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

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{y})=\mathcal{L}_{\mathrm{CE}}(\boldsymbol{x},\boldsymbol{y})+\mathcal{L}_{\mathrm{PC}}(\boldsymbol{x},\boldsymbol{y}) \tag{3}$$

(In Equation 3, f_i is the output of the feature by the network for the ith example, w_i is the cluster center for ith class (which is learned during training). Term1 adds a penalty on the distance between f_i and its corresponding class cluster center. Term2 adds a negative penalty (increasing distance) between f_i and the other class centers. Thus together, this term tries to ensure learnt features are well clustered.

In the paper [14], it is demonstrated that adding these terms gives an additional boost to adversarial robustness as each class is better separated. Though we could not wholly replicate the extent of the result of the paper, the better-separated feature space, and the slight improvement in accuracy can be seen in figures and table.


(a) Original Image

(b) Trans DCT (c) Color HSV (d) Color YCbCr (e) Gray -scale (f) Gauss Blur (g) Affine (h) Dropout

Figure 1. Effects of applied tranformations on the example frog image from CIFAR10

3.2. Transformations for Reactive Defenses

For the visualization of the following transformations, please refer to the Figure 1.

- *Discrete Cosine Trasform (DCT)*

  To represent the image in frequency domain, DCT decomposes a signal into cosine wave components. Some of recent works [5] [1] have shown capability of DCT based compression to weaken the the adversarial attacks under the universal adversarial perturbations. 2-D DCT Transform can be represented as 4

$$y_k=2\sum_{n=0}^{N-1}x_n\cos\left(\frac{\pi k(2n+1)}{2N}\right) \tag{4}$$

  discrete cosine transform (DCT), which decomposes a signal into cosine wave components, to represent a natural image in frequency space.

- *IMGAUG*

  ImgAug library [9] is used for the for augmenting and transforming the images for adversarial defences varying different properties. We use the following three transformations:

  * Gaussian Blur: It augments the image with blur using a gaussian kernel. For our experiments, we have used the default setting of sigma = 3.0.

  * Affine: Affine transforms the image and relocates the pixels with geometrical and semantic information intact. An image is rotated by -20 to 20 degress and filled up with a random RGB color for all newly created pixels. Chen et. al. [2] also uses affine transformation with colors to explore adversarial attacks in another (image transformed) space.

  * Dropout: In this transformation, certain fraction for pixels in the image is set to zero. In our case, we drop only 5% of the pixels from the total image which found

to be the optimal value from visual and experimental evaluations.

- *Color*

  One of the reason, we thought color space would be good choice for a transform, was that in each different color space different features stand out for a class, which could lead to different features being learnt for the model, but as this tranformation is spatially aligned we had low expectactions from it. These are the color spaces we used.

  * HSV : HSV (for hue, saturation, value; also known as HSB, for hue, saturation, brightness) are alternative representations of the RGB color model, designed in the 1970s by computer graphics researchers to more closely align with the way human vision perceives color-making attributes. In these models, colors of each hue are arranged in a radial slice, around a central axis of neutral colors which ranges from black at the bottom to white at the top.

  * YCbCr: It is a family of color spaces used as a part of the color image pipeline in video and digital photography systems. Y is the luma component and CB and CR are the blue-difference and red-difference chroma components. Y (with prime) is distinguished from Y, which is luminance, meaning that light intensity is nonlinearly encoded based on gamma corrected RGB primaries

  * Grayscale : In this transformation, the value of each pixel is a single sample representing only an amount of light, that is, it carries only intensity information.

4. Experimental Setting

The experiments have been done on CIFAR-10 dataset, which contains 60000 colored images belonging to ten classes of size 32*32. The machine learning architecture used is Resnet20 with loss function and optimizer as PCL (cross-entropy + proximity) loss and Stochastic Gradient Descent(SGD) repectively. During model training in all experiments, models are trained for 100 epochs with learning rate 0.01 for cross-entropy and 0.5 for proximity loss. We performed our experiments on GPU cluster provided by the Saarland Univerity at memory usage of 4GB and on basic Google Colab hardware.

For black box attack examples, we have generated results for two values of epsilon 8/255 and 16/255 (more stronger), from MadryLab CIFAR10 Challenge [13] using their secret model attacks which our model is not aware of. For baseline evaluation, the resnet20 model is trained with cross entropy loss and PCL (cross entropy + proximity) loss on RGB images. As evaluation metric, we use the classification accuracy of the model on clean and adversarial examples on the CIFAR-10 testset.

5. Results

Summary of results can be found in Table 1. From the accuracy values obtained on black box attacks over different epsilon values, it can be observed that reactive + proactive measures outperforms the baseline proactive measure against the black box attacks. It shows the usefulness of these transformations in combating adversarial attacks. It has also been observed that PCL training gives 1-3% additional improvement in accuracy in each case over it's counterpart based on vanilla softmax training.

For our baseline reactive approach, we trained [14] with the parameters provided in their repository, we unfortunately we were not able to replicate the results from the paper completely, but the amount of improvement we got, served the purpose of our analysis.

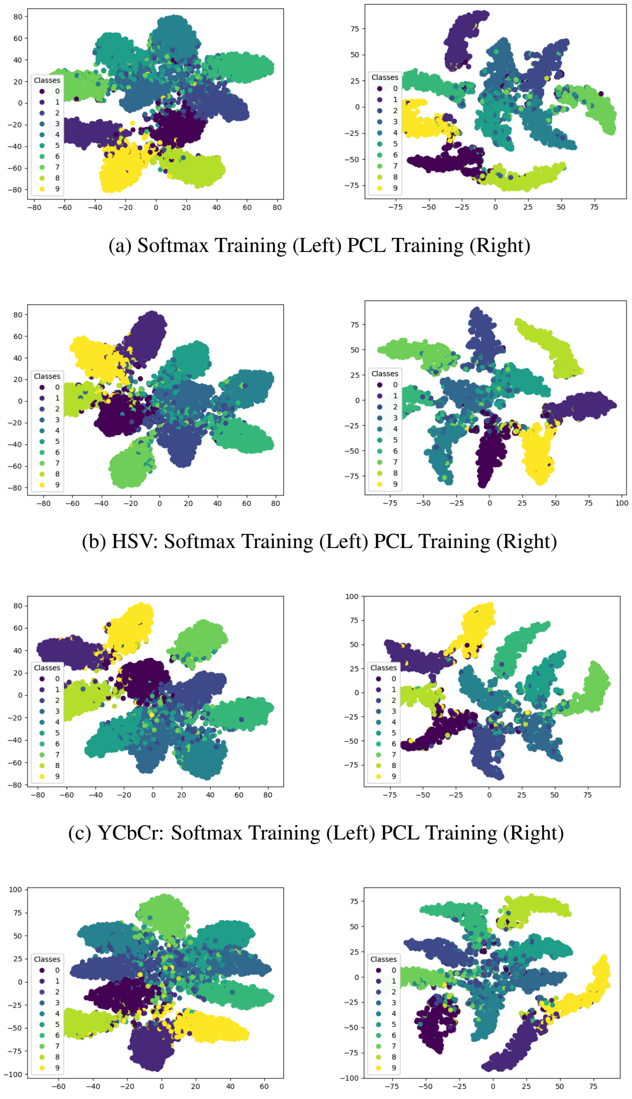
Among all the transformations, DCT gave the best results on black box attacks achieving accuracy as high as 64.6% for epsilon value of 8/255 and 48.74% on even stronger attack of epsilon value 16/255. But it comes at the cost of accuracy in clean images which dropped around 81%. This is due to the fact that DCT performs image compression which leads to loss of information in images.

After DCT, the best performance has been observed in case of affine transformation, there is three times improvement (28.390) from the baseline model (9.06) for epsilon 8/255 and nine times improvement (3.110 to 27.66) for the robustness in case of softmax model for epsilon 16/255. Intuitively, affine remaps the image pixels to different locations while preserving the geometrical and semantic information. Also, random color-fill to the new pixels provides additional robustness as shown color transformations.

In the color space, transforms for the YCrCB and grey scale colour space transform, we did not observe a strong change in the BBox accuracy of the model after training with the transform. This may be due to the fact that these transformations preserve spatial alignment and hence features. Surprisingly, the HSV color space gave a considerable boost in robustness (from 12.24 to 28.940), a possible reason could be that the adversarial image was more clearly visible in the HSV colorspace.

Pixel dropout and gaussian blur transformations also show significant improvements (four to five times) over the baseline where dropout performs superior than the gaussian blur. Although, the gaussian blur performs similar to the HSV color transformation for eps = 8/255 but less accurate on more stronger attack (eps = 16/255) as shown in Figure 2 and(a)(Right) PCL and Figure 4 (a)(Right). Gaussian blur enforces smoothness which removes the edge features important for image classification. In case of dropout, random pixels are dropped which the model can't keep learning specific features from the specific pixels leading to improved robustness.

In all feature plots from penultimate layers we can observe PCL training based maps gave better separation of classes than softmax training which is attributed to it's robustness property. In Figure 7 which has been obtained from affine tranformation and PCL training it has been observed the feature plot of cat, dog and deer are overlapping in most of plots. This overlap can be a potential cause of misclassification after adversarial attacks.



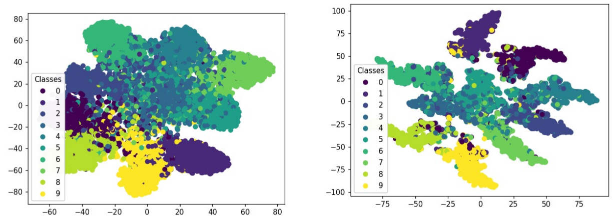
(a) Softmax Training (Left) PCL Training (Right)

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

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

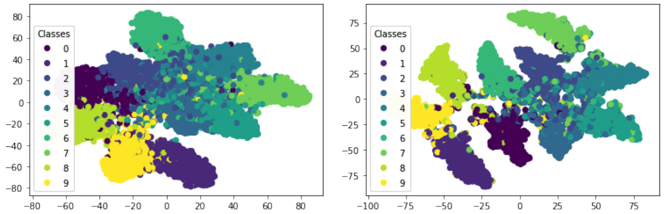
(d) Grayscale: Softmax Training (Left) PCL Training (Right)

Figure 2. Color Space transform penultimate layer feature maps



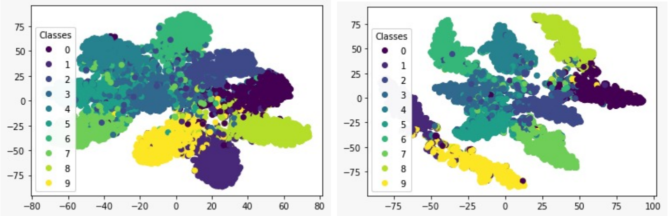
(a) DCT: Softmax Training (Left) PCL Training (Right)

Figure 3. DCT transform penultimate layer feature maps



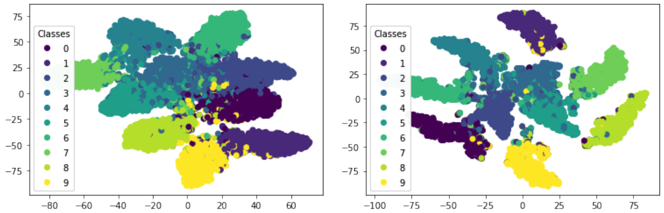
(a) Gaussian Blur: Softmax Training (Left) PCL Training (Right)

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)

Figure 6. Pixel Dropout: penultimate layer feature maps

6. Conclusion and Future work

We observed that training with transformed input gives additional robustness against adversarial attacks. Tranformations that spatially change the input like the DCT transform or the affine tranform gave us the strongest improvements. We observed a general trend of feature clusters considerably changing after tranformations which indicates improvements in adversarial robustness.

In the future, we would like to do further analysis on the

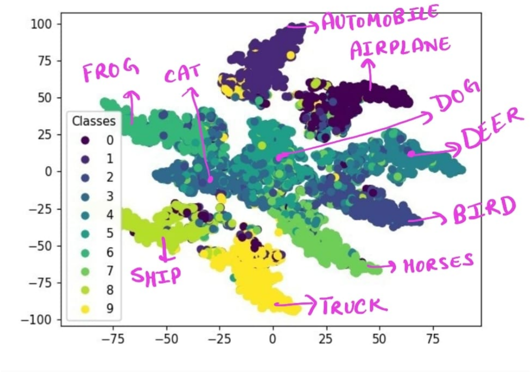


Figure 7. Separation in feature space and Classes overlap for Affine Transform

affect of tranformations with adversarial training, the fine grained effect of individual class accuracies after tranformation, and finally test our current framework on an adaptive attack.

References

[1] Naveed Akhtar, Jian Liu, and Ajmal Mian. Defense against universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3389–3398, 2018. 2

[2] Jiyu Chen, David Wang, and Hao Chen. Explore the transformation space for adversarial images. In *Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy*, pages 109–120, 2020. 1, 2

[3] Tianlong Chen, Sijia Liu, Shiyu Chang, Yu Cheng, Lisa Amini, and Zhangyang Wang. Adversarial robustness: From self-supervised pre-training to fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 699–708, 2020. 1

[4] Francesco Croce, Maksym Andriushchenko, and Matthias Hein. Provable robustness of relu networks via maximization of linear regions. In *the 22nd International Conference on Artificial Intelligence and Statistics*, pages 2057–2066. PMLR, 2019. 1

[5] Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M Roy. A study of the effect of jpg compression on adversarial images. *arXiv preprint arXiv:1608.00853*, 2016. 2

[6] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1

[7] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features, 2019. 1

[8] Matt Jordan, Justin Lewis, and Alexandros G Dimakis. Provable certificates for adversarial examples: Fitting a ball in the union of polytopes. *Advances in Neural Information Processing Systems*, 32, 2019. 1

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
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

Table 1. Results of proactive+reactive defenses on Black Box attacks for CIFAR10 Dataset for epsilon values of 8/25 and 8/255. Lce refers to Cross entropy loss used in vanilla softmax training.

[9] Alexander B. Jung, Kentaro Wada, Jon Crall, Satoshi Tanaka, Jake Graving, Christoph Reinders, Sarthak Yadav, Joy Banerjee, Gábor Vecsei, Adam Kraft, Zheng Rui, Jirka Borovec, Christian Vallentin, Semen Zhydenko, Kil-

ian Pfeiffer, Ben Cook, Ismael Fernández, François-Michel De Rainville, Chi-Hung Weng, Abner Ayala-Acevedo, Raphael Meudec, Matias Laporte, et al. imgaug. https://github.com/aleju/imgaug, 2020. Online; accessed 01-Feb-2020. 2

[10] Minseon Kim, Jihoon Tack, and Sung Ju Hwang. Adversarial self-supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:2983–2994, 2020. 1

[11] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016. 1

[12] Ling Liu, Wenqi Wei, Ka-Ho Chow, Margaret Loper, Emre Gursoy, Stacey Truex, and Yanzhao Wu. Deep neural network ensembles against deception: Ensemble diversity, accuracy and robustness. In *2019 IEEE 16th international conference on mobile ad hoc and sensor systems (MASS)*, pages 274–282. IEEE, 2019. 1

[13] MadryLab. Project title. https://github.com/MadryLab/cifar10_challenge, 2017. 3

[14] 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

[15] Aamir Mustafa, Salman H Khan, Munawar Hayat, Roland Goecke, Jianbing Shen, and Ling Shao. Deeply supervised discriminative learning for adversarial defense. *IEEE transactions on pattern analysis and machine intelligence*, 43(9):3154–3166, 2020. 1

[16] Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. A self-supervised approach for adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 262–271, 2020. 1

[17] Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pages 5286–5295. PMLR, 2018. 1

[18] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. *Advances in Neural Information Processing Systems*, 32, 2019. 1