DUIB: Defending against Adversarial Attacks and Common Corruptions

Myeongseob Ko
Department of Electrical
Computer Engineering
Virginia Tech
Blacksburg, VA 24061
myeongseob@vt.edu

Hoang Anh Just
Department of Electrical
Computer Engineering
Virginia Tech
Blacksburg, VA 24061
just@vt.edu

Abstract

The one big reason for machine learning (ML) not yet taken over every aspect in our life is its lack of guaranteeing reliability and safety to its users. With the increased studies in attacks to deep neural networks (DNNs), their performance are greatly deteriorated and are often unable to defend. Additionally, deep neural networks often perform poorly on examples that occur naturally in real-world, which are called common corruptions. There are many works proposing superior defenses to attacks. However, they only consider single attack scenario, which is impractical and its byproduct often unintentionally opens vulnerability to another attack. Therefore, we are interested in defenses that can defend against multiple attacks, including common corruptions, which is often regarded as universal robustness. The current state-of-the-art work, perceptual adversarial training (PAT), is first to propose effective method to defend against unseen attacks. However, the performance, though considerably higher than other techniques, is still overall worsened. Therefore, we propose a framework to improve the current PAT defense by leveraging information theory technique, which is information bottleneck. In particular, we propose adopting Hilbert-Schmidt independence criterion (HSIC) to PAT in a novel way. To show the efficacy of our method, we evaluate the method through multiple adversarial attacks and common corruption examples.

1 Introduction

Machine learning has been widely used in a variety of tasks with unprecedented improvement. However, there are a lot of studies that well-trained machine learning models are susceptible to adversarial perturbations which are carefully designed by an attacker or to natural perturbations (i.e., noise, blur, weather) appearing in the physical world. Adversarial samples are crafted by the attacker to misclassify the trained models, and many studies have focused on increasing classifiers' robustness against these adversarial attacks [7, 14, 18]. On the other hand, typical natural perturbations are regarded as common corruptions, which were proposed by [11]. This category of attacks have given rise to another, concurrent line of research which is robustness against common corruptions [6, 10, 3]. Despite the works being developed in parallel, the intersection of the common ground is lacking in proper studies, i.e. achieving a robust model against *both attacks at the same time*.

Much research has been done in improving robustness against only one specific adversarial threat model (i.e., L_2, L_∞, etc.) [13]. However, a model trained for a certain type of attack is vulnerable to another types of adversarial attacks, which opens a space for an adversary to attack again [5]. Laidlaw et al. [13] propose to define a perceptual threat model to defend against a wide range of adversarial attacks based on the perceptual metric. The intuition behind is that human vision is robust

Advanced Machine Learning - Prof. Ming Jin (Spring 2022).

to different types of perturbations.

Given the perceptual distance d$(x_1; x_2)$ between input samples x_1 and x_2, the perceptual threat model can be formalized. [13] found that the perceptual bound can include various adversarial threat models (e.g., Lp and spatial attacks). However, as the perceptual bound is not tight, the general robustness over different kinds of adversarial threat models and common corruptions is still low.

In this paper, we propose an approach to increase the robustness against various adversarial threat models and common corruptions at the same time. The current design of perceptual metric [19] is similar to human perception but not well designed for robust human vision. Therefore, if we can extract the semantic perceptual features to calculate the perceptual bound between two images from a perceptual threat model, the trained model can be more robust to carefully crafted or natural attacks. Information bottleneck (IB) [15] can be considered as a way of selecting crucial features by decreasing the mutual information between input and hidden features, and increasing the mutual information between the hidden features and output.

$$IB(\theta) = I(Z, Y; \theta) - \beta I(Z, X; \theta). \tag{1}$$

$$I(Z, Y; \theta) = \int dxdy\, p(z, y|\theta) log \frac{p(z, y|\theta)}{p(z|\theta)p(y|\theta)}. \tag{2}$$

Instinctively, the first term in equation (1) pushes Z to be effective to Y; and the second term compresses the impact of input X into Z, which provide a minimal sufficient statistics of X for predicting Y. The IB is attractive as we can find a good representation of input X, but the main drawback of the IB is that computing mutual information is computationally challenging. There are multiple studies [1, 16, 2] to address this issue by computing approximate mutual information, which are still unstable if we adopt it for training machine learning models. Recent works [17, 8] show that the Hilbert-Schmidt independence criterion (HSIC) can be an tractable solution if we leverage the HSIC as an regularize to the objective function. It is well studied that IB can help to increase the robustness against adversarial attacks, but We show that adding the HSIC to the main threat model only improve the robustness against a certain types of attacks (i.e., L_p) and decrease the robustness over a variety of attacks. We demonstrate that adding HSIC to the perceptual threat model and the main model at the same time can increase the defense performance over general threat models and common corruptions. The followings are our contribution:

1. We propose an improved perceptual threat model based on HSIC theorem, which is more robust against a variety of adversarial attacks and natural perturbations.
2. We empirically show that our approach can produce a machine learning model with high robustness diverse unseen attacks including L1, L2, recoloring, JPEG attacks, and common corruptions on CIFAR-10.

2 Background

2.1 Perceptual distance

The perceptual distance measures how similar are two images in a way that coincides with human judgement. Zhang et al. [19] find that internal activations of trained models for classification tasks well correspond to human perceptual judgments. Mathmetically, the perceptual distance between image $x1$ and $x2$ can be formalized as :

$$d(x1, x2) \triangleq \|\phi(x1) - \phi(x2)\| \tag{3}$$

Here, $\phi : \chi -> Z$ maps an input $x \in \chi$ to the normalized, flattened internal activations $\phi(x) \in Z$, where $Z \subseteq \mathbb{R}^m$ indicates the set of all possible internal activations. They find that the perceptual distance can be a good surrogate for human vision.

2.2 Perceptual Adversarial Attacks

Now we can define the perceptual adversarial attacks based on the perceptual distance. For a given input x with a true label y, an adversary want to generate a perceptual adversarial image $\tilde{x}$ with a budget ϵ to make a model $f : \chi \to Y$ misclassify:

$$f(\tilde{x}) \neq y,\ d(x, \tilde{x}) = \|\phi(x) - \phi(\tilde{x})\| \leq \epsilon \tag{4}$$

However, the perceptual distance constraint is more complex than L_p constraints. Thus, we can derive the attack formulation based on the Lagrangian relaxation:

$$\max_{\tilde{x}} L(f(\tilde{x}, y)) - \lambda max(0, \left\|\phi(\tilde{X}) - \phi x)\right\| - \epsilon) \tag{5}$$

The perceptual constraint cost is designed to be 0 as long as the generated perceptual adversarial sample is within the perceptual distance ϵ.

2.3 Adversarial Training

Let $f_\theta(\cdot)$ be a classifier parameterized by θ over a distribution of inputs and labels $(x, y) \in D$, and L denote the cross entropy loss. Then the adversarial robust training can be defined as the expected loss calculated by perceptual adversarial samples, i.e.,

$$\min_{\theta} \mathbb{E}_{(x,y)\ D}[\max_{d(\tilde{x},x)\leq\epsilon} L(f(\tilde{x}, y)] \tag{6}$$

The training formulation minimizes the worst-case loss within a perceptual neighborhood which is bounded by the perceptual budget ϵ of a given training point x.

2.4 Hibert-Schmidt Independence Criterion (HSIC)

The Hilbert-Schmidt Independence Criterion (HSIC) is a statistical kernel dependence measure with Hilbert-Schmidt norm proposed by Gretton et al. [17]. HSIC is the Hilbert-Schmidt norm of the cross-covariance operator between the distributions in Reproducing Kernel Hilbert Space (RKHS). Intuitively, two random variables x and y are independent if and only if any bounded continuous function of the two random variables are uncorrelated(i.e., zero covariance). HSIC between two random variables x, y can be defined as:

$$\begin{aligned} HSIC(x, y) = &\mathbb{E}_{xx'yy'}[k_x(x, x')k_y(y, y')] \\ &+ \mathbb{E}_{xx'}[k_x(x, x')]\mathbb{E}_{yy'}[k_y(y, y)] \\ &- 2\mathbb{E}_{xy}[\mathbb{E}_{x'}[k_x(X, X')]\mathbb{E}_{y'}[k_y(y, y')]] \end{aligned}$$

Theorem 1 (Gretton et al. (2005a)[9], Theorem 4)
Denote by F and G RKHSs both with universal kernels, k, l respectively on compact domains X and Y. Assume without loss of generality that $\|s\|_\infty \leq 1$ for all $s \in F$ and likewise $\|t\|_\infty \leq 1$ for all $t \in G$.
Then the following holds: $\|C_x y\|^2_{HS} = 0 \Leftrightarrow X \perp\!\!\!\perp Y$.

Let $(x_i, y_i)_{i=1}^n$ be i.i.d. samples from the joint distribution on χ and Y. The empirical estimate of HSIC is given by:

$$\widehat{HSIC}((x_i, y_i)_{i=1}^n; F, G) = \frac{1}{(n-1)^2}\mathrm{Tr}\,(KHLH), \tag{7}$$

where $K_{i,j} = k(x_i, x_j), L_{i,j} = l(y_i, y_j)$ are kernel matrices for the kernel k and l respectively, and $H_i, j = \delta_{i,j} - \frac{1}{n}$ is a centering matrix. The main result of Gretton et al. [9] is that the empirical estimate $\widehat{HSIC}$ converges to $HSIC$ at a rate of $O(\frac{1}{n^{1/2}})$.

3 Proposed Method

In this section, we present our method which leverages the perceptual threat model to generate the adversarial images within a bound ϵ and the HSIC as a regularizer to extract semantic perceptual features. The key intuition behind is that even though adversarial attacks and common corruptions are perturbing or corrupting the images, they still carry the main structure of the subject in those images based on human perception. Furthermore, the perceptual bound proposed by Laidlaw et al. [13] can cover other types of threat models but not tight enough. Therefore, we are proposing to adjust the perceptual bound based on the HSIC theorem to improve the robustness. The key question is how to control the bound while preserving the effectiveness of perceptual distance.

3.1 Improved Perceptual Adversarial Training via DUIB

We define the improved perceptual adversarial attacks based on HSIC theorem. For a given input x with a true label y, an adversary can generate an adversarial image $\hat{x}$ with a given budget ϵ to make a model $f : \chi \rightarrow Y$ misclassify:

$$\max_{\tilde{x}} L(f(\tilde{x}, y) - \lambda max(0, \|\phi(\tilde{x}) - \phi x\| - \epsilon) - (\lambda_x \sum_{j=1}^{M} HSIC(\tilde{x}, \tilde{Z}_j) - \lambda_y \sum_{j=1}^{M} HSIC(y, \tilde{Z}))$$

Here, λ_x and λ_y denote the hyper-parameters for HSIC. The HISC allows an adversary to adjust the bound based on estimated information budget (i.e., $(\lambda_x \sum_{j=1}^{M} HSIC(\tilde{x}, \tilde{Z}_j) - \lambda_y \sum_{j=1}^{M} HSIC(y, \tilde{Z}_j))$).

Then, the final training formulation is defined as:

$$\begin{aligned}\min_{\theta} \mathbb{E}_{(x,y)\sim D}[\max_{d(\tilde{x},x)\leq\epsilon} & L(f(\tilde{x}, y)) \\ & - (\lambda_x \sum_{j=1}^{M} HSIC(\tilde{x}, \tilde{Z}_j) + \lambda_y \sum_{j=1}^{M} HSIC(\tilde{Y}, \tilde{Z}_j)) \\ & + (\lambda_{\hat{x}} \sum_{i=1}^{T} HSIC(\tilde{x}, \hat{Z}_i) + \lambda_{\hat{y}} \sum_{i=1}^{T} HSIC(Y, \hat{Z}_i))]\end{aligned}$$

Here, Z and $\hat{Z}$ represent the hidden features extracted from a perceptual threat model g and a main model f respectively. As the perceptual model and the main model can be different, we use different notations for each model: the total number of layers for calculating HSIC in the perceptual threat model and the main model are M and T respectively. Note that we also leverage the HSIC as a regularizer to the base model f, as HSIC encourages to remove redundant or noisy information contained in $\tilde{x} \in \chi$, while retaining the discriminative nature of the classifier.

4 Results and Analysis

After implementing our proposed method, we evaluate it on a wide range of unseen adversarial attacks which are provided in [13], such as L_∞, L_2, recoloring [12], perceptual projected gradient descent (PPGD), which is an application of [14], and Lagrangian perceptual attack (LPA), which is similar to [4]. From Table 1, we observe that our proposed methods PAT-DUIB with bounds 0.7 and 1.0 outperform the current state-of-the-art work achieving the mean accuracy throughout all attacks of 34.12% and 36.44%, respectively. The PAT-DUIB with bound 0.5 reaches accuracy similar to current SOTA.

Additionally, we evaluate our proposed defense against the common corruption examples. The methods are evaluated using the relative mean corruption error (mCE) metric proposed in [11]. We consider the common corruption examples by first splitting the common corruption attacks into four

Training	Bound	Clean	Linf	L2	Recolor	PPGD	LPA	Mean
PAT [SOTA]	0.7/1.0	71.6	28.7	33.3	67.5	26.6	9.8	33.18
PAT with IB	0.5	87.8	33.8	39.5	54.8	16.5	5.7	30.06
PAT with IB	0.7	85.3	33.5	38.2	63.9	13.0	3.5	30.42
PAT with IB	1.0	78.1	30.5	34.4	67.9	15.4	6.0	30.84
PAT-DUIB	0.5	83.2	35.9	41.6	67.1	12.0	3.4	32
PAT-DUIB	0.7	73.8	34.8	39.5	67.2	24.3	16.4	**36.44**
PAT- DUIB	1.0	73.7	30.5	35.8	68.5	17.7	18.1	**34.12**

Table 1: Accuracies against adversarial attacks for Perceptual Adversarial Training (PAT) models with our proposed method on CIFAR-10.

	Perturbation Type				
Training	Noise	Blur	Weather	Digital	All
PAT 0.7	0.63742145	0.77235457	0.98663722	0.94674895	0.8357905475
PAT 1.0	0.798926	1.04542925	1.267442333	1.2670728	1.094717596
PAT-DUIB 0.7	0.76257721	0.97374722	1.15889349	1.19441775	1.022408918
PAT-DUIB 0.5	0.541802	0.6857065	0.8683713333	0.8455242	0.7353510083

Table 2: Robustness of classifiers trained with PAT with our proposed method against common corruptions in the CIFAR-10-C dataset. Results are reported as relative mCE (lower is better) [11]

groups in Table 2: noise, blur, weather, and digital. Then, we calculate mCEs for each of these groups. Remarkably, PAT-DUIB with bound 0.5 achieves the lowest mCE error, which in turn results in highest robustiness for common corruption. Other PAT-DUIB moddel, however, does not attain the best score in common corruption evaluation.

From the given results, we notice that with a bigger bound for generating adversarial examples in adversarial training, the model can cover more attacks, which effects in higher performance and can even achieve stronger performance on adversarial attack defenses than the current-state-of-the-art work, (PAT). However, on the other hand, increasing the bound will also result in higher mCE, which consequently, indicates lower robustness of the model on common corruption examples.

Based on the trade-off of the robustness between both of these types of attacks, we recommend using bound 0.7, or depending on the examples we would like to defend from.

Additionally, for ablation study, we evaluated in Table 1 models which would only have their base model containing the information bottleneck regularizer. However, their performace on adversarial attacks is even lower than that of the current SOTA model. This results emphasizes the importance of implementing the information bottleneck in the perceptual model, and it is not sufficient to only add to the base model.

5 Observation and Future Direction

We have proposed a novel method for improving current state-of-the-art adversarial training by implementing dual stage information bottleneck in the perceptual adversarial training. The defense performance against a wide range of adversarial attacks is improved compared to current best adversarial architectures. However, our proposed method still requires improvement in common corruption robustness. Therefore, as a future work we will explore empirical and theoretical analysis on decision boundaries, which will expand our understanding of model predictions on common corruption examples with respect to a given adversarial attack training bound. Based on our experiments, adopting information bottleneck to a clean model improves the performance on common corruptions. Thus, we believe there is space for improvement in this problem for our proposed method.

5.1 Acknowledgement

We would like to thank our advisor Prof. Ruoxi Jia and Prof. Jin Ming for tremendous help throughout the project. We appreciate the help and are looking forward to make this work go further!

References

[1] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

[2] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

[3] Dan A Calian, Florian Stimberg, Olivia Wiles, Sylvestre-Alvise Rebuffi, Andras Gyorgy, Timothy Mann, and Sven Gowal. Defending against image corruptions through adversarial augmentations. *arXiv preprint arXiv:2104.01086*, 2021.

[4] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017.

[5] Congyue Deng and Yi Tian. Towards understanding the trade-off between accuracy and adversarial robustness.

[6] N Benjamin Erichson, Soon Hoe Lim, Francisco Utrera, Winnie Xu, Ziang Cao, and Michael W Mahoney. Noisymix: Boosting robustness by combining data augmentations, stability training, and noise injections. *arXiv preprint arXiv:2202.01263*, 2022.

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

[8] Daniel Greenfeld and Uri Shalit. Robust learning with the hilbert-schmidt independence criterion. In *International Conference on Machine Learning*, pages 3759–3768. PMLR, 2020.

[9] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.

[10] Yong Guo, David Stutz, and Bernt Schiele. Improving corruption and adversarial robustness by enhancing weak subnets. *arXiv preprint arXiv:2201.12765*, 2022.

[11] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[12] Hossein Hosseini and Radha Poovendran. Semantic adversarial examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1614–1619, 2018.

[13] Cassidy Laidlaw, Sahil Singla, and Soheil Feizi. Perceptual adversarial robustness: Defense against unseen threat models. *arXiv preprint arXiv:2006.12655*, 2020.

[14] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[15] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[16] Boxin Wang, Shuohang Wang, Yu Cheng, Zhe Gan, Ruoxi Jia, Bo Li, and Jingjing Liu. Infobert: Improving robustness of language models from an information theoretic perspective. *arXiv preprint arXiv:2010.02329*, 2020.

[17] Zifeng Wang, Tong Jian, Aria Masoomi, Stratis Ioannidis, and Jennifer Dy. Revisiting hilbert-schmidt information bottleneck for adversarial robustness. *Advances in Neural Information Processing Systems*, 34, 2021.

[18] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.

[19] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.