Adversarial Concurrent Training: Optimizing Robustness and Accuracy Trade-off of Deep Neural Networks

Elahe Arani*
elahe.arani@navinfo.eu
Fahad Sarfraz*
fahad.sarfraz@navinfo.eu
Bahram Zonooz
bahram.zonooz@gmail.com

Advanced Research Lab
NavInfo Europe
Eindhoven, The Netherlands

Abstract

Adversarial training has been proven to be an effective technique for improving the adversarial robustness of models. However, there seems to be an inherent trade-off between optimizing the model for accuracy and robustness. To this end, we propose *Adversarial Concurrent Training* (ACT), which employs adversarial training in a collaborative learning framework whereby we train a robust model in conjunction with a natural model in a minimax game. ACT encourages the two models to align their feature space by using the task-specific decision boundaries and explore the input space more broadly. Furthermore, the natural model acts as a regularizer, enforcing priors on features that the robust model should learn. Our analyses on the behavior of the models show that ACT leads to a robust model with lower model complexity, higher information compression in the learned representations, and high posterior entropy solutions indicative of convergence to a flatter minima. We demonstrate the effectiveness of the proposed approach across different datasets and network architectures. On ImageNet, ACT achieves 68.20% standard accuracy and 44.29% robustness accuracy under a 100-iteration untargeted attack, improving upon the standard adversarial training method's 65.70% standard accuracy and 42.36% robustness.

1 Introduction

Deep neural networks (DNNs) have emerged as a predominant framework for learning multiple levels of representation, with higher levels representing more abstracts aspects of the data [2]. The better representation has led to the state-of-the-art performance in many challenging tasks in computer vision [24, 42], natural language processing [2, 45] and many other domains [14, 30]. However, despite their pervasiveness, recent studies have exposed the lack of robustness of DNNs to various forms of perturbations [11, 15, 38]. In particular, adversarial examples which are imperceptible perturbations of the input data carefully crafted by adversaries to cause erroneous predictions pose a real security threat to DNNs deployed in critical applications [25].

 * Equal contribution.

The intriguing phenomenon of adversarial examples has garnered a lot of attention in the research community [46] and progress has been made in both creating stronger attacks to test the model's robustness [5, 9, 28, 44] as well as defenses to these attacks [26, 27, 49]. However, Athalye et al. [1] show that most of the proposed defense methods rely on obfuscated gradients which is a special case of gradient masking and lowers the quality of the gradient signal causing the gradient-based attack to fail and give a false sense of robustness. They observe adversarial training [27] as the only effective defense method. The original formulation of adversarial training, however, does not incorporate the clean examples into its feature space and decision boundary. On the other hand, Jacobsen et al. [19] provide an alternative viewpoint and argue that the adversarial vulnerability is a consequence of narrow learning, resulting in classifiers that rely only on a few highly predictive features in their decisions. We have not yet developed a full understanding of the major factors that contribute to adversarial vulnerability in DNNs and consequently, the optimal method for training robust models remains an open question.

A recent variant of adversarial training, TRADES [49], adds a regularization term on top of the standard cross-entropy loss which forces the model to match its embeddings for the clean example and the corresponding adversarial example. However, there might be an inherent tension between the objective of adversarial robustness and that of standard generalization [41]. Therefore, combining these optimization tasks into a single model and forcing the model to completely match the feature distributions of the adversarial and clean examples may lead to sub-optimal solutions. We, therefore, hypothesize that considering the optimization for adversarial robustness and generalization as two distinct yet complementary tasks and encouraging more exhaustive exploration of the input and parameter space can lead to better solutions.

In this paper, we propose adversarial concurrent training (ACT) for training a robust model in conjunction with a natural model in a collaborative manner (Fig.1a). The goal is to utilize the task-specific decision boundaries to align the feature space of the robust and natural model in order to learn a more extensive set of features that are less susceptible to adversarial perturbations. To this end, ACT closely intertwines the training of a robust and natural model by involving them in a minimax game inside a closed learning loop. The adversarial examples are generated by determining regions in the input space where the discrepancy between the two models is maximum. In the subsequent step, each model minimizes a supervised learning loss which optimizes the model on its specific task in addition to a mimicry loss that aligns the two models. Our formulation consists of bi-directional knowledge distillation between the clean and adversarial domain, enabling them to collectively explore the input and parameter space more extensively. Furthermore, the supervision from the natural model acts as a regularizer which effectively adds a prior on the learned representations and leads to semantically meaningful features that are less susceptible to off-manifold perturbations introduced by adversarial attacks.

We empirically test the efficacy of our proposed approach and show that ACT provides a better trade-off between robustness and generalization across different datasets (CIFAR-10, CIFAR-100 [23] and ImageNet [24]) and network architectures (ResNet [13] and WideResNet [47]). Our further analyses show that ACT learns a lower complexity model with higher posterior entropy solutions, indicative of convergence to flatter minima. While standard adversarial training reduces the information compression in the learned representations compared to standard training [26], our method shows higher information compression than even standard training. The empirical results coupled with desirable characteristics of models trained with ACT demonstrates the effectiveness of concurrent training for adversarial ro-

bustness. Our results also demonstrate the versatility of ACT to different datasets and network architectures which makes the method applicable across a variety of application

2 Related Work

The discovery of adversarial examples [33] has garnered a lot of interest from the research community. Researchers have proposed various forms of defense methods which include detecting the adversarial examples [8, 10], applying non-linear pre-processing and transformations on the input image, using ensemble method [3, 4, 35, 40], regularization techniques [20, 37, 48] and training on adversarial examples [9, 26, 27, 49]. However, Athalye *et al*. [1] showed that most of the proposed defense methods rely on gradient obfuscation, lowering the quality of the gradient signal, to give a false sense of robustness. They found adversarial training to be an effective method after addressing the issue of gradient obfuscation. Nevertheless, the increase in robustness comes at the cost of generalization. A number of studies even argue that there is an inherent trade-off between robustness and generalization and consider them as contradictory goals [18, 36, 41, 49]. Ilyas *et al*. [18] consider the adversarial vulnerability to be a direct consequence of the model's sensitivity to well generalizing features which are highly predictive yet brittle. Jacobsen *et al*. [19] provide an alternative perspective on adversarial vulnerability and show that DNNs are also excessively invariant to task relevant changes in the input image. They attribute this to narrow learning resulting from the insufficiency of the standard cross-entropy loss to incentivize explaining all class dependent aspects of the input.

On other end, collaborative learning which provides additional supervision signals, has been effective in increasing the robustness to different noise types. Knowledge distillation [17] has been shown to be a general-purpose training paradigm which is more robust to common challenges in the real-world datasets [32]. Han *et al*. [12] use two networks to filter different types of errors introduced by noisy labels. Hendrycks *et al*. [16] show that self-supervision can improve the robustness of the model to adversarial examples, label corruption, and common input corruptions. Based on the aforementioned findings, we hypothesize that adversarial training within a collaborative framework that encourages the model to explore the input and parameter space more extensively can be instrumental in further improving the robustness gains of the standard adversarial training method.

3 Adversarial Concurrent Training

In this section, we first present the overall idea and intuition behind the proposed method and how it aims to address some of the shortcomings of standard adversarial training and then formally define the method and introduce the loss functions used for training the model.

3.1 Proposed Method

Standard adversarial training [27] involves generating an adversarial example x' for each clean example x and then subsequently training the model to assign the same label y to x' using cross-entropy loss. Adversarial training has been proven to confer robustness to the model. However, the standard formulation has a few shortcomings. The model does not receive any pair information indicating that x' is the adversarial counterpart of x and therefore

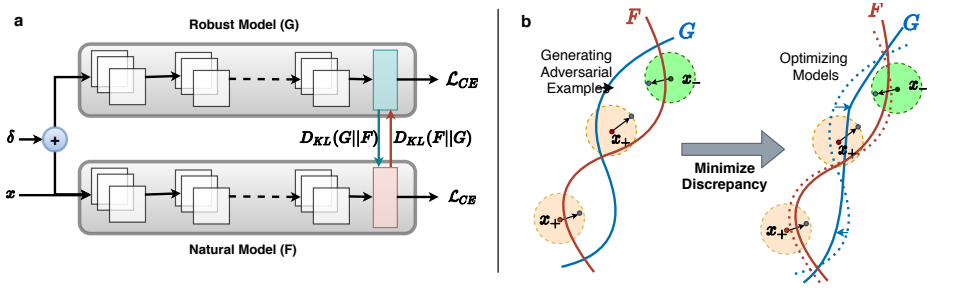

Figure 1: a) Schematic of our proposed method. b) An overview of ACT on a binary classification problem. x_+ and x_- indicate data samples for positive and negative classes respectively whereas circle indicates the allowed ε-bound. First, adversarial examples are generated by identifying discrepancy regions between G and F. The arrow in the circles shows the direction of the adversarial perturbation and the circles show the perturbation bound. In a subsequent step, the discrepancy between the models is minimized. This effectively aligns the decision boundaries and pushes them further from the examples. Best viewed in color.

fails to utilize the semantic similarity between the adversarial and clean examples for learning an optimal embedding. The objective function does not involve explicitly minimizing the generalization on clean examples which can lead to overfitting to the adversarial domain. Furthermore, the model is not incentivized to incorporate all class-dependent features of the input into its decision boundary which leads to narrow learning. One approach is to combine the generalization and adversarial robustness loss into one objective function [21, 49]. However, the goal of adversarial robustness is different from standard generalization [41]. Therefore, combining these two optimization tasks together into a single model and completely matching the feature distributions of the adversarial and clean examples could cause tension between the two tasks and leads to sub-optimal solutions.

We hypothesize that treating the optimization for adversarial robustness and generalization as distinct yet complementary tasks in a way that encourages more exhaustive exploration of the input and parameter space can lead to better solutions. To this end, we propose Adversarial Concurrent Training (ACT) which entails training an adversarially robust model in conjunction with a natural model in a collaborative manner (Fig.1a). The goal is to utilize the task specific decision boundaries to align the feature space of the robust and natural model in order to learn a more extensive set of features which are less susceptible to adversarial perturbations. ACT closely intertwines the training of a robust and natural model by involving them in a minimax game inside a closed learning loop. The adversarial examples are generated by identifying regions in the input space where the discrepancy between the robust and natural model is maximum. In the subsequent step, the discrepancy between the two models is minimized in addition to optimizing them on their respective tasks.

Our approach has a number of advantages. The adversarial perturbations, generated by identifying regions in the input space where the two models disagree, can be effectively used to align the two models. This alignment coupled with pushing the two decision boundaries away from the data samples leads to smoother decision boundaries (Fig.1b). Updating the models based on the disagreement regions combined with optimization on distinct tasks ensures that the two models do not converge to a consensus, and the method does not reduce to self-training. Furthermore, the supervision from the natural model acts as a noise-free

Algorithm 1: Adversarial Concurrent Training Algorithm

Input: Dataset D, Balancing factor α, Learning rate η, Batch size m
Initialize: G and F parameterized by θ and ϕ
while *Not Converged* **do**
1: Sample mini-batch: $(x_1, y_1), ..., (x_m, y_m) \sim D$
2: Compute adversarial examples:
$\delta^* = \arg\max_{\delta \in S} \mathcal{L}_G(\theta, \phi, \delta)$
3: Compute $\mathcal{L}_G(\theta, \phi, \delta^*)$ (Equation 1)
Compute $\mathcal{L}_F(\theta, \phi, \delta^*)$ (Equation 2)
4: Compute stochastic gradients and update the parameters:
$\theta^* \leftarrow \theta - \eta \frac{\partial \mathcal{L}_G}{\partial \theta}$
$\phi^* \leftarrow \phi - \eta \frac{\partial \mathcal{L}_F}{\partial \phi}$
return $\theta*$ *and* $\phi*$

reference for regularizing the robust model. This effectively adds a prior on the learned representations which encourages the model to learn semantically relevant features in the input space. This combined with the requirement on the robust model's prediction to be stable within the epsilon bound encourages the model to select semantically relevant features with stable behavior over a larger region.

3.2 Formulation

We formulate our proposed method, ACT, as a concurrent training of an adversarially robust model G parametrized by θ and a natural model F parametrized by ϕ (see Fig.1a). Each model is trained with two losses: a task specific loss and a mimicry loss. The standard cross-entropy loss ($\mathcal{L}_{CE}$) is used as the task specific loss and the Kullback-Leibler Divergence (D_{KL}) is used as the mimicry loss to align the output distributions of the models. The robust model minimizes the convex combination of the cross-entropy loss on adversarial examples and the D_{KL} between the output distributions of the robust model on adversarial examples and the natural model on clean examples.

$$\mathcal{L}_G(\theta, \phi, \delta) = (1-\alpha)\mathcal{L}_{CE}(G(x+\delta;\theta), y) + \alpha D_{KL}(F(x;\phi)||G(x+\delta;\theta)) \tag{1}$$

where the perturbation δ is sampled from a set of allowed perturbations S bounded by ε. The tuning parameter $\alpha \in [0,1]$ plays key role on balancing the importance of task specific and alignment errors. The natural model uses a similar loss function which minimizes the cross-entropy loss on clean examples.

$$\mathcal{L}_F(\theta, \phi, \delta) = (1-\alpha)\mathcal{L}_{CE}(F(x;\phi), y) + \alpha D_{KL}(G(x+\delta;\theta)||F(x;\phi)) \tag{2}$$

The training procedures involves first finding the adversarial examples by maximizing the robust model loss $\mathcal{L}_G$ with respect to δ within the set of allowed perturbation S, and then subsequently minimizing the loss functions for each model $\mathcal{L}_G$ and $\mathcal{L}_F$ (see Algorithm 1). This results in an approximate minimax optimization:

$$\begin{cases} \min_\theta E_{(x,y)\in D} \max_{\delta \in S} \mathcal{L}_G(\theta, \phi, \delta) \\ \min_\phi E_{(x,y)\in D} \mathcal{L}_F(\theta, \phi, \delta) \end{cases} \tag{3}$$

		0.1	0.3	0.5	0.7	0.9	1.0
Standard model (F)	A_{nat}	95.29±0.10	95.31±0.10	94.88±0.13	94.42±0.10	90.12±0.63	10.00
	A_{rob}	3.57±1.06	3.81±1.29	2.90±0.26	3.36±0.45	10.85±2.04	0.00
Robust model (G)	A_{nat}	85.94±0.11	86.13±0.14	86.33±0.22	86.24±0.18	84.87±0.91	10.00
	A_{rob}	48.93±0.28	49.40±0.59	50.62±0.86	51.37±0.41	55.12±1.07	0.00

Table 1: Effect of α hyperparameter on ACT (ResNet-18 trained on CIFAR-10).

Note that the natural model F is only used during training, and at inference, only the robust model G is used. Therefore, ACT does not incur any additional inference cost compared to standard adversarial training.

4 Empirical Validation

In this section, we empirically evaluate the the effectiveness of our proposed method and study the characteristics of the models.

4.1 Experimental Setup

We evaluate the performance ACT on different datasets (CIFAR-10, CIFAR-100 [23] and ImageNet [24]) and network architectures (ResNet [13] and WideResNet [47]). For all our experiments, we normalize the images between 0 and 1 and apply random cropping with reflective padding of 4 pixels and random horizontal flip data augmentations. For training, we use stochastic gradient descent (SGD) with 0.9 momentum, 200 epochs, batch size 128, and an initial learning rate of 0.1, decayed by a factor of 0.2 at epochs 60, 120 and 150. Unless explicitly mentioned, the results for ACT refers to the performance of the robust model G. For Madry and TRADES, we follow the training scheme used in [49]. For generating adversarial examples during training, we use the projected gradient decent (PGD) as a universal first order adversary [27] with $\varepsilon = 0.031$, step size $\eta = 0.007$, and the perturbation steps $K = 10$. For evaluation, we set $\eta = 0.003$ and test for different perturbation steps. For a fair comparison, we use $1/\lambda = 5$ for TRADES which achieves the highest robustness for ResNet-18 in [49]. In our experiments, TRADES achieves both better robustness and generalization than reported in the original work [49].

We train each method with 3 different random seeds and report the average and one standard deviation performance. A_{nat} refers to standard accuracy on clean examples whereas A_{rob} refers to accuracy on adversarial examples (reported in percentage). Unless otherwise stated, A_{rob} shows the worst performance on a PGD-20 attack with 5 random initialization.

4.2 Effect of α hyperparameter

Table 1 shows the effect of the balancing factor α on the robustness and generalization of the natural and robust models. The extra supervision signal from each of the model affects both the robustness and generalization performance of the models. The adversarial robustness of the robust model generally increases as we increase α value. Interestingly, considerable robustness is transferred to the natural model as well without explicitly being trained on adversarial examples and this transfer increases for higher α values. For our subsequent experiments, we use $\alpha = 0.9$.

	Dataset	Defense	A_{nat}	A_{rob}			Minimum Perturbation
				PGD-20	PGD-100	PGD-1000	
ResNet-18	CIFAR-10	Madry	**85.11±0.19**	50.53±0.01	47.67±0.18	47.51±0.16	0.03782±0.00024
		TRADES	83.49±0.38	53.79±0.29	52.15±0.26	52.12±0.26	0.04279±0.00066
		ACT	84.33±0.27	**55.83±0.18**	**53.73±0.19**	**53.62±0.19**	**0.04454±0.00069**
	CIFAR-100	Madry	58.36±0.10	24.48±0.16	23.10±0.20	23.02±0.23	0.01961±0.00010
		TRADES	56.91±0.46	28.88±0.16	27.98±0.17	27.96±0.19	0.02353±0.00014
		ACT	**61.56±0.46**	**31.14±0.16**	**29.74±0.15**	**29.71±0.14**	**0.02462±0.00017**
WRN-28-10	CIFAR-10	Madry	87.26±0.20	49.76±0.06	46.91±0.10	46.77±0.06	0.04412±0.00083
		TRADES	86.36±0.26	53.52±0.17	**50.73±0.18**	**50.63±0.17**	0.04714±0.00018
		ACT	**87.58±0.16**	**54.94±0.14**	50.66±0.11	50.44±0.13	**0.05601±0.00031**
	CIFAR-100	Madry	**60.77±0.16**	24.92±0.23	23.56±0.26	23.46±0.24	0.02084±0.00011
		TRADES	58.10±0.17	28.49±0.08	**27.50±0.23**	**27.44±0.23**	0.02395±0.00011
		ACT	60.72±0.18	**28.74±0.14**	27.32±0.00	27.26±0.01	**0.02595±0.00016**

Table 2: Comparison of ACT with prior defense models under various white-box attacks.

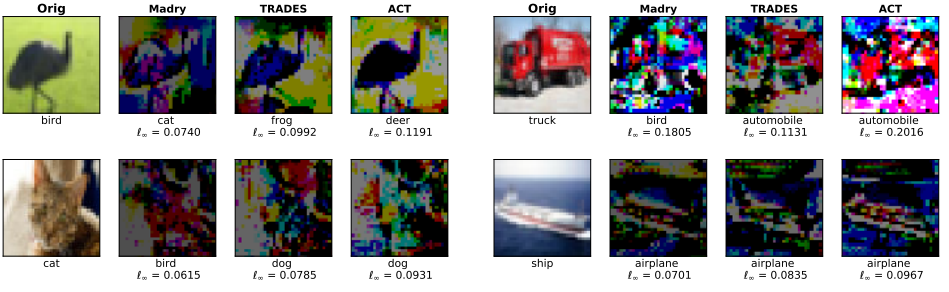

Figure 2: Minimum perturbations required to fool the robust models trained with different defense methods on ResNet-18 and CIFAR-10. The label of each image shows the predicted class along with the ℓ_∞ distance of the adversarial example from the clean example. Note that the perturbations are multiplied by 5 to highlight the visual differences.

4.3 Comparison with prior work

As our method adapts standard adversarial training in a collaborative learning framework, original formulation by Madry [27] is included as baseline. Furthermore, TRADES [49] is included to show the effectiveness of optimization for robustness and generalization as two distinct yet complementary tasks instead of combining them into a single model. Table 2 shows the effectiveness of ACT across different datasets and network architectures under various white-box attacks. Specifically, for ResNet-18, ACT significantly improves the robustness. In instances where Madry has better generalization, the difference in the robustness is considerably larger.

We also evaluate the average minimum perturbation required to successfully fool the defense methods. We apply the $FGSM^k$ attack in foolbox [31] which returns the smallest adversarial perturbation under the ℓ_∞ distance. Table 2 shows that ACT consistently requires higher perturbation in images on average across the different datasets and network architectures. Fig.2 provides examples of the required perturbations to fool each defense. ACT requires a higher degree of perturbation in the semantically relevant regions of the image.

We further verify the effectiveness of our method under black-box attacks. Table 3 shows the transferability of adversarial examples generated using a PGD-20 attack on the surrogate model to the target models trained with different defense methods. ACT shows higher robustness to black-box attacks transferred from Madry and TRADES.

Def \ Sur	CIFAR-10				CIFAR-100			
	Natural	Madry	TRADES	ACT	Natural	Madry	TRADES	ACT
Madry	**15.93**	49.50	35.88	**35.34**	43.13	75.71	61.50	60.29
TRADES	18.23	35.84	46.49	37.11	44.25	60.80	71.20	**59.97**
ACT	16.93	**33.65**	**35.14**	44.04	**40.80**	**57.17**	**57.32**	68.61

Table 3: Comparison of ACT with prior defenses (Def) under black-box PGD-20 attack. Surrogate models (Sur) are source models that provide gradients for adversarial attacks. Values indicate the success rate of adversarial attack hence a lower number shows higher robustness.

Method	A_{nat}	A_{rob}		
		PGD-10	PGD-50	PGD-100
Madry	65.70	42.13	42.29	42.36
Free (m=4) [33]	64.45	43.52	43.39	43.40
FGSM [43]	60.90	43.46	-	-
ACT	**68.20**	**44.06**	**44.24**	**44.29**

Table 4: Comparison of ACT with prior defenses under untargeted PGD attack with $\varepsilon = 2/255$ on the ImageNet dataset.

4.4 Results on ImageNet

We further demonstrate the effectiveness of our approach on the challenging ImageNet classification task [24]. As our approach is designed for untargeted adversarial training, following prior work on untargeted attacks on ImageNet dataset [33, 43] we train a Resnet-50 model to be robust to untargeted PGD attack with the following parameters: $\varepsilon = 2/255$, $\eta = 1.0$, and $K = 10$. For our experiments, we use four Tesla V100 GPUs with $\alpha = 0.5$, batch size of 128 on each GPU, and train for 100 epochs with an initial learning rate of 0.1 decayed by a factor of 0.1 at 30, 60 and 90 epochs. For a fair comparison, we also train Madry [27] under the same experimental setup. Table 4 shows that ACT improves both the generalization and robustness over the standard adversarial training method and its faster variants. As PGD attack is sensitive to the initial randomization, there can be small fluctuations in the final robustness. Therefore, the results for the different PGD attack essentially show that robustness is maintained even as we increase the number of steps.

4.5 Gradient obfuscation

Athalye et al. [1] showed that most of the proposed defense methods give a false sense of security by reducing the quality of the gradient signal and that these defenses can be circumvented by using gradient approximation techniques. Therefore, it is important to perform a number of sanity checks to ensure that a proposed adversarial defense does not rely on gradient obfuscation. These checks include ensuring that white-box attacks are at least as strong as black-box attacks and that an unconstrained iterative gradient-based attack with an unlimited number of iterations should be completely successful [26]. Our evaluations show that black-box attacks are substantially weaker than the corresponding white-box attacks (Tables 2 and 3). Table 5 shows that the robustness of the model monotonically decreases as we increase the allowed perturbation level for a PGD-100 attack. This shows that the gradients of our method do not impair the ability of the gradient-based attacks through gradient obfuscation.

Defense	1	5	10	15	20	25	50	100
Madry	**81.90**	63.53	36.69	16.61	7.01	3.31	0.22	0
TRADES	80.23	65.27	42.51	**23.35**	11.17	5.40	0.29	0
ACT	81.47	**67.15**	**42.98**	22.45	**11.18**	**5.74**	**0.42**	0

Table 5: Accuracy of ResNet-18 with different ϵ values and a fix number of steps (PGD-100) conducted on CIFAR-10.

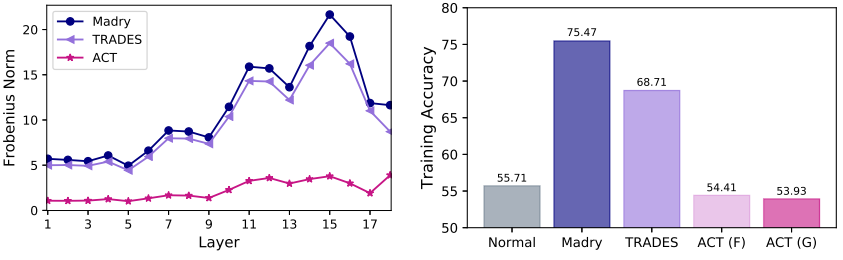

Figure 3: (Left) Comparison of the Frobenius norms of the weights matrix in learnable layers of ResNet-18. (Right) Comparison of the degree to which the different models with frozen learned representation can fit random binary labels.

4.6 Model complexity

The magnitudes of the weights of neural networks can provide an estimate of the model's complexity. Following the analysis presented in [26], we analyze the Frobenius norms of all the weight layers in ResNet-18 for different defense methods trained on CIFAR-10. Fig.3 (left) shows that the Frobenius norm of ACT is considerably lower across all the layers. This provides preliminary evidence that ACT trains lower complexity models than standard adversarial training.

4.7 Information compression

A number of studies on understanding DNNs from an information theory perspective have shown a relationship between the information compression in the learned features and generalization [34, 39]. They relate stronger compression in DNN's hidden states to a stronger bound on generalization. To study the effect of our proposed method on the compression of information in the learned representation, we follow the analysis performed by Lamb et al. [26] whereby we freeze the learned representation of the model and study how successful these frozen representations are in predicting fixed random labels. In particular, we add a 2-layer multi-layer perceptron (MLP) network with 400 and 200 neurons on top of the frozen representations of ResNet-18 models trained on CIFAR-10 with different defense methods and fit them on random binary labels. If the model compresses the information well in the learned representations, it will be more difficult to fit the random binary labels. Thereby, lower accuracy shows better information compression. Lamb et al. [26] showed that standard adversarial training causes the learned representation to be less compressed. To the contrary, Fig.3 (right) suggests that both the natural and robust model trained with ACT has more information compression. Interestingly, the models trained with ACT shows higher information compression compared to standard training (normal). This indicates the efficacy

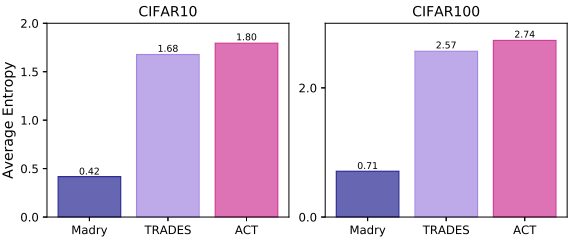

Figure 4: Average entropy over all training samples obtained by a ResNet-18 on CIFAR-10 and CIFAR-100 datasets. ACT converges to higher posterior entropy solutions.

of our proposed approach in capturing more information in the hidden states of the models.

4.8 Entropy regularization

There can be multiple solutions that can fit the training data distribution, but some of these generalize better because of being in wide valleys rather than narrow crevices [6, 23] whereby the predictions do not change drastically with small perturbations. A number of studies have shown that the tendency towards finding these robust minima can be increased by biasing the DNNs towards solutions with higher posterior entropy [6, 29].

We argue that the extra mimicry loss which encourages the model to match the posterior probabilities in ACT has a regularization effect on the logits. The effect is that the model distributes its mass over the secondary classes more uniformly. This can be quantified with the average posterior entropy over the training samples. Fig.4 shows that training with ACT leads to higher posterior entropy solutions. Therefore, the collaborative learning in ACT has a connection to entropy regularisation-based approaches [6, 29] to finding wider minima through mutual probability matching on secondary classes.

5 Conclusion

We proposed *Adversarial Concurrent Training (ACT)* as an efficient approach to training a robust model in conjunction with a natural model. The additional supervision from the natural model allows the robust model to learn richer internal representation which is robust to adversarial perturbations. Our empirical results showed that ACT provides a better trade-off between robustness and generalization across different datasets and network architectures. Furthermore, our analysis suggests that ACT leads to a robust model with lower model complexity, higher information compression in the learned representation, and high posterior entropy solutions indicative of convergence to a flatter minima. The versatility of our proposed approach coupled with the desirable characteristics makes it applicable across a variety of tasks and applications.

References

[1] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint*

arXiv:1802.00420, 2018.

[2] Yoshua Bengio. Deep learning of representations: Looking forward. In *International Conference on Statistical Language and Speech Processing*, pages 1–37. Springer, 2013.

[3] Arjun Nitin Bhagoji, Daniel Cullina, and Prateek Mittal. Dimensionality reduction as a defense against evasion attacks on machine learning classifiers. *arXiv preprint arXiv:1704.02654*, 2017.

[4] Jacob Buckman, Aurko Roy, Colin Raffel, and Ian Goodfellow. Thermometer encoding: One hot way to resist adversarial examples. *ICLR*, 2018.

[5] 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.

[6] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, 2019.

[7] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

[8] Reuben Feinman, Ryan R Curtin, Saurabh Shintre, and Andrew B Gardner. Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410*, 2017.

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

[10] Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick McDaniel. On the (statistical) detection of adversarial examples. *arXiv preprint arXiv:1702.06280*, 2017.

[11] Keren Gu, Brandon Yang, Jiquan Ngiam, 13 Quoc Le, and Jonathan Shlens. Using videos to evaluate image model robustness. *arXiv preprint arXiv:1904.10076*, 2019.

[12] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in neural information processing systems*, pages 8527–8537, 2018.

[13] K He, X Zhang, S Ren, and J Sun. Deep residual learning for image recognition. computer vision and pattern recognition (cvpr). In *2016 IEEE Conference on*, volume 5, page 6, 2015.

[14] JB Heaton, NG Polson, and Jan Hendrik Witte. Deep learning for finance: deep portfolios. *Applied Stochastic Models in Business and Industry*, 33(1):3–12, 2017.

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

[16] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *arXiv preprint arXiv:1906.12340*, 2019.

[17] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[18] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *arXiv preprint arXiv:1905.02175*, 2019.

[19] Jörn-Henrik Jacobsen, Jens Behrmann, Richard Zemel, and Matthias Bethge. Excessive invariance causes adversarial vulnerability. *arXiv preprint arXiv:1811.00401*, 2018.

[20] Daniel Jakubovitz and Raja Giryes. Improving dnn robustness to adversarial attacks using jacobian regularization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 514–529, 2018.

[21] Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018.

[22] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

[23] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). *URL http://www. cs. toronto. edu/kriz/cifar. html*, 8, 2010.

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[25] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

[26] Alex Lamb, Vikas Verma, Juho Kannala, and Yoshua Bengio. Interpolated adversarial training: Achieving robust neural networks without sacrificing too much accuracy. In *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*, pages 95–103. ACM, 2019.

[27] 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.

[28] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. corr abs/1511.04599 (2015). *arXiv preprint arXiv:1511.04599*, 2015.

[29] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017.

[30] Harry A Pierson and Michael S Gashler. Deep learning in robotics: a review of recent research. *Advanced Robotics*, 31(16):821–835, 2017.

[31] Jonas Rauber, Wieland Brendel, and Matthias Bethge. Foolbox: A python toolbox to benchmark the robustness of machine learning models. *arXiv preprint arXiv:1707.04131*, 2017. URL http://arxiv.org/abs/1707.04131.

[32] Fahad Sarfraz, Elahe Arani, and Bahram Zonooz. Knowledge distillation beyond model compression. *arXiv preprint arXiv:2007.01922*, 2020.

[33] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In *Advances in Neural Information Processing Systems*, pages 3353–3364, 2019.

[34] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

[35] Thilo Strauss, Markus Hanselmann, Andrej Junginger, and Holger Ulmer. Ensemble methods as a defense to adversarial perturbations against deep neural networks. *arXiv preprint arXiv:1709.03423*, 2017.

[36] Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao. Is robustness the cost of accuracy?–a comprehensive study on the robustness of 18 deep image classification models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018.

[37] Cecilia Summers and Michael J Dinneen. Improved adversarial robustness via logit regularization methods. *arXiv preprint arXiv:1906.03749*, 2019.

[38] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[39] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE, 2015.

[40] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

[41] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

[42] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018, 2018.

[43] Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*, 2020.

[44] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. *arXiv preprint arXiv:1801.02612*, 2018.

[45] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3):55–75, 2018.

[46] Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*, 2019.

[47] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[48] Valentina Zantedeschi, Maria-Irina Nicolae, and Ambrish Rawat. Efficient defenses against adversarial attacks. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 39–49. ACM, 2017.

[49] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*, 2019.