Supplementary material: Towards Achieving Adversarial Robustness by Enforcing Feature Consistency Across Bit Planes

Sravanti Addepalli*, Vivek B.S.*, Arya Baburaj, Gaurang Sriramanan, R.Venkatesh Babu
Video Analytics Lab, Department of Computational and Data Sciences
Indian Institute of Science, Bangalore, India

1. Details on Architecture and Training

In this section, we present details related to the architecture of models used and the impact of change in hyperparameters.

1.1. Architecture details for Fashion-MNIST and MNIST

We use a modified LeNet architecture for all our experiments on Fashion-MNIST and MNIST datasets. This architecture has two additional convolutional layers when compared to the standard LeNet architecture [8]. Architecture details are presented in Table-1.

Table 1: Network architectures used for Fashion-MNIST and MNIST datasets. Modified LeNet is used for training the model and Net-A is used as a source for generating black-box attacks.

Modified LeNet (M-LeNet)	Net-A
{conv(32,5,5) + Relu}×2	Conv(64,5,5) + Relu
MaxPool(2,2)	Conv(64,5,5) + Relu
{conv(64,5,5) + Relu}×2	Dropout(0.25)
MaxPool(2,2)	FC(128) + Relu
FC(512) + Relu	Dropout(0.5)
FC + Softmax	FC + Softmax

1.2. Impact of change in Hyperparameters

In this section, we study the effect of variation in the hyperparameter λ (Eq. (2) in main paper). For CIFAR-10 dataset, we set the initial value of λ to be 1, and multiply this by a constant factor every 25 epochs (3 times over 100 epochs). We present the results obtained by changing the rate of increase in λ for CIFAR-10 dataset in Fig-1. As the rate increases, accuracy on clean samples reduces, and accuracy on adversarial samples increases. The clean accuracy saturates to about 70%, and accuracy on adversarial samples saturates to approximately 40%. The best trade-off between both is obtained at a rate of 15, where the clean accuracy is 75.28% and adversarial accuracy is 40.6%. However, for a fair comparison with PGD training and other existing methods, we select the rate at which clean accuracy matches with that of PGD-AT. Hence, the selected hyperparameter is 9.

We use a similar methodology for hyperparameter selection in MNIST and Fashion-MNIST datasets as well. For these datasets, we set a fixed value of λ and do not increase it over epochs. The value of λ is selected such that the accuracy on clean samples matches with that of a PGD trained model.

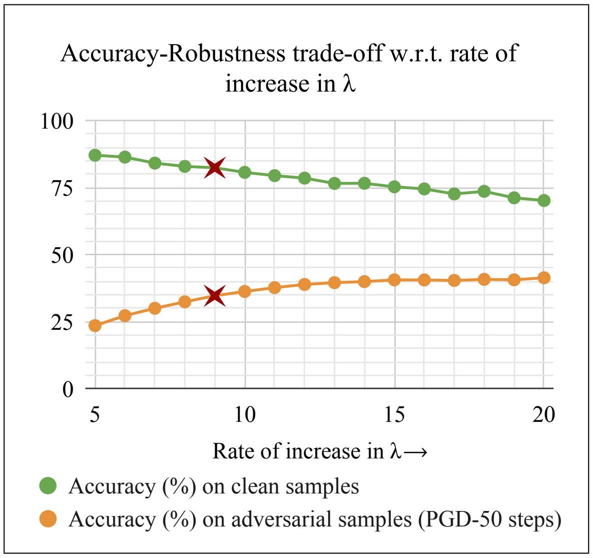

Figure 1: Plot of recognition accuracy ($\%$) on clean samples and PGD samples versus the rate of increase in hyperparameter(λ) used for BPFC training. The selected setting (9) is highlighted using a cross mark.

*Equal contribution

Table 2: **Fashion-MNIST**: Recognition accuracy (%) of models in a white-box attack setting.

Training method	**Clean**	**FGSM**	**IFGSM 40 steps**	**PGD (n-steps)** 40	100	1000
FGSM-AT	93.0	89.9	25.3	15.5	15.1	15.0
RSS-AT	87.7	81.2	77.5	72.0	71.8	71.8
PGD-AT	87.4	81.4	80.2	79.1	79.0	79.0
NT	92.0	16.6	2.4	0.3	0.3	0.3
Mixup	91.0	37.7	0.1	0.0	0.0	0.0
BPFC (**Ours**)	87.1	73.1	70.2	68.0	67.7	67.7

Table 3: **MNIST**: Recognition accuracy (%) of models in a white-box attack setting.

Training method	**Clean**	**FGSM**	**IFGSM 40 steps**	**PGD (n-steps)** 40	100	1000
FGSM-AT	99.4	89.6	29.4	13.8	4.9	3.7
RSS-AT	99.0	96.4	93.1	93.0	90.9	90.4
PGD-AT	99.3	96.2	94.9	95.4	94.3	94.1
NT	99.2	82.7	0.5	0.0	0.0	0.0
Mixup	99.4	58.1	0.2	0.0	0.0	0.0
BPFC (**Ours**)	99.1	94.4	92.0	91.5	86.6	85.7

Table 4: **PGD attack with multiple random restarts:** Recognition accuracy (%) of different models on PGD adversarial samples with multiple random restarts in a white-box setting.

Training method	**CIFAR-10** PGD 50-steps			**Fashion-MNIST** PGD 100-steps			**MNIST** PGD 100-steps		
# restarts :	1	100	1000	1	100	1000	1	100	1000
PGD-AT	45.3	44.9	44.9	80.6	79.7	79.6	92.9	90.9	90.6
RSS-AT	45.2	44.7	44.7	74.1	73.4	73.2	88.7	86.3	85.6
BPFC (**Ours**)	35.6	35.1	34.9	69.4	68.4	68.3	84.0	80.5	79.9

2. Details on Experimental Results

In this section, we present additional experimental results to augment our observations and results presented in the main paper.

2.1. White-box attacks

2.1.1 Bounded attacks

Detailed results on Fashion-MNIST and MNIST white-box attacks are presented in Tables-2 and 3 respectively. The proposed method achieves significantly better robustness to multi-step adversarial attacks when compared to Normal training (NT), FGSM-AT and Mixup. The robustness to multi-step attacks using the proposed approach is comparable to that of PGD-AT and RSS-AT models, while being faster than both approaches.

Table 5: **DeepFool and C&W attacks (Fashion-MNIST):** Average ℓ_2 norm of the generated adversarial perturbations is reported. Higher ℓ_2 norm implies better robustness. Fooling rate (FR) represents percentage of test set samples that are misclassified.

Training method	**DeepFool** FR (%)	Mean ℓ_2	**C&W** FR (%)	Mean ℓ_2
FGSM-AT	94.34	1.014	100.0	0.715
PGD-AT	90.70	3.429	100.0	2.142
RSS-AT	91.22	2.762	99.9	1.620
NT	94.07	0.467	100.0	0.406
Mixup	92.22	0.226	100.0	0.186
BPFC (**Ours**)	90.94	3.620	100.0	1.789

Table 6: **DeepFool and C&W attacks (MNIST):** Average ℓ_2 norm of the generated adversarial perturbations is reported. Higher ℓ_2 norm implies better robustness. Fooling rate (FR) represents percentage of test set samples that are misclassified.

Training method	**DeepFool** FR (%)	Mean ℓ_2	**C&W** FR (%)	Mean ℓ_2
FGSM-AT	99.36	3.120	100.0	1.862
PGD-AT	95.97	5.316	100.0	3.053
RSS-AT	94.41	4.894	98.2	2.725
NT	99.15	1.601	100.0	1.427
Mixup	91.82	0.518	100.0	0.498
BPFC (**Ours**)	97.47	6.289	100.0	3.041

We run the PGD attack with multiple random restarts on a random sample of 1000 test set images, equally distributed across all classes. This experiment is done to ensure that the achieved robustness is not due to gradient masking. The results with random restarts are presented in Table-4. Here, the overall accuracy is computed as an average over the worst-case per-sample accuracy, as suggested by Carlini *et al.* [2]. A 50-step PGD attack is performed on CIFAR-10 dataset, and a 100-step attack is performed on Fashion-MNIST and MNIST datasets. The degradation from 100 random restarts to 1000 random restarts is insignificant across all datasets, indicating the absence of gradient masking. Degradation from a single run to 100 random restarts is also insignificant for CIFAR-10 and Fashion-MNIST. However, the degradation is larger for MNIST, similar to the trend observed with PGD-AT and RSS-AT models. It is to be noted that the results corresponding to this experiment may not coincide with those reported in Table-1 in the main paper, and Tables-2 and 3 in the Supplementary, as we consider only a sample of the test set for this experiment.

2.1.2 Unbounded attacks

The results with unbounded attacks (DeepFool [10] and Carlini-Wagner (C&W) [3]) for Fashion-MNIST and MNIST datasets are presented in Tables-5 and 6 respectively. We select the following hyperparameters for C&W attack on Fashion-MNIST and MNIST datasets: search steps = 9, max iterations = 500, learning rate = 0.01. For DeepFool attack, we set the number of steps to 100 for both Fashion-MNIST and MNIST datasets.

The average ℓ_2-norm of the generated perturbations to achieve approximately 100% fooling rate using C&W attack is higher with the proposed approach when compared to most other approaches, with the exception of PGD-AT, whose average ℓ_2-norm is marginally higher. DeepFool attack does not achieve 100% fooling rate for Fashion-MNIST and MNIST datasets, as was the case with CIFAR-10 (ref: Section-5.3.3, main paper). However, since the fooling rates of the proposed approach are comparable to, or greater than that of PGD-AT and RSS-AT, we can make a fair comparison between the required ℓ_2-norm for achieving the given fooling rate across these approaches. We observe that the proposed approach is more robust to DeepFool attack, when compared to both of these approaches.

2.2. Black-box attacks

Multi-step attacks such as I-FGSM are known to show weak transferability across models in a black-box setting [7]. Dong *et al.* [4] introduced a momentum term in the optimization process of I-FGSM, so as to increase the transferability of the generated adversarial samples. This attack is referred to as the Momentum Iterative FGSM (MI-FGSM) attack.

The results corresponding to black-box multi-step PGD and MI-FGSM attacks are presented in Tables-7 and 8 respectively. We consider two source models for black-box attacks on each of the models trained: one with the same architecture as the target model, and second with a different architecture. For Fashion-MNIST and MNIST, the architecture of the second model (Net-A) is presented in Table-1. For CIFAR-10, we consider a second model with VGG-19 [11] architecture. The proposed approach achieves a significant improvement in robustness to adversarial samples with respect to Normal Training (NT) and Mixup, and comparable results with respect to the adversarial training methods, across all the datasets.

Table 7: **PGD Black-box attacks:** Recognition accuracy (%) of different models on PGD black-box adversaries. Columns represent source model used for generating the attack. 7-step attack is used for CIFAR-10 and 40-step attack is used for Fashion-MNIST and MNIST

Training method	**CIFAR-10**		**Fashion-MNIST**		**MNIST**	
	VGG19	ResNet18	Net-A	M-LeNet	Net-A	M-LeNet
FGSM-AT	85.85	85.61	94.27	91.52	79.8	74.11
RSS-AT	80.92	80.82	84.71	83.91	95.19	96.27
PGD-AT	81.37	81.22	85.16	85.71	96.52	96.69
NT	16.86	0	27.10	0.33	4.64	0.03
Mixup	30.16	29.53	49.07	60.71	31.4	58.25
BPFC (**Ours**)	80.42	80.15	81.45	83.00	95.31	95.91

Table 8: **MI-FGSM [4] Black-box attacks:** Recognition accuracy (%) of different models on MI-FGSM black-box adversaries. Columns represent source model used for generating the attack. 7-step attack is used for CIFAR-10 and 40-step attack is used for Fashion-MNIST and MNIST

Training method	**CIFAR-10**		**Fashion-MNIST**		**MNIST**	
	VGG19	ResNet18	Net-A	M-LeNet	Net-A	M-LeNet
FGSM-AT	76.44	74.22	94.61	92.11	79.95	73.92
RSS-AT	80.21	80.10	84.61	84.02	96.11	95.28
PGD-AT	80.47	80.59	84.98	85.58	95.56	95.34
NT	12.98	0.04	28.28	4.69	12.48	1.93
Mixup	35.74	25.22	50.60	63.32	43.72	62.98
BPFC (**Ours**)	79.04	79.04	81.33	82.70	94.03	94.46

2.3. Adaptive attacks

In this section, we explain the adaptive attacks used in this paper in greater detail. We utilize information related to the proposed regularizer to construct potentially stronger attacks when compared to a standard PGD attack. We maximize the following loss function to generate an adaptive attack corresponding to each data sample x_i:

$$L_i = \lambda_{ce}\, ce(f(x_i), y_i) + \lambda_g \|g(x_i) - g(q(x_i))\|_2^2 - \lambda_{LSB}\|x_i - q(x_i)\|_2^2 \quad (1)$$

The quantized image corresponding to x_i is denoted by $q(x_i)$. We consider $f(.)$ as the function mapping of the trained network, from an image x_i, to its corresponding softmax output $f(x_i)$. The corresponding pre-softmax output of the network is denoted by $g(x_i)$. The ground truth label corresponding to x_i is denoted by y_i. The first term in the above equation is the cross-entropy loss, the second term is the BPFC regularizer proposed in this paper, and the third term is an ℓ_2 penalty term on the magnitude of k LSBs. We consider the value of k to be the same as that used for training the models (ref: Section-5.1 in the main paper). The coefficients of these loss terms are denoted by λ_{ce}, λ_g and λ_{LSB} respectively.

Maximizing the cross-entropy term leads to finding samples that are misclassified by the network. Maximizing the BPFC loss results in finding samples which do not comply with the BPFC regularizer imposed during training. Minimizing the third term would help find samples with low magnitude LSBs, which are possibly the points where the defense is less effective. The objective of an adversary is

Table 9: **CIFAR-10**: Recognition accuracy (%) of the model trained using the proposed approach on adversarial samples generated using adaptive attacks.

Adaptive attack	**Loss coefficients**			**n-step Adaptive attack**		
	λ_{ce}	λ_g	λ_{LSB}	7	20	50
PGD	1	0	0	41.72	35.74	34.68
Variation in λ_g (λ_{ce} = 1 and λ_{LSB} = 0)	1	0.1	0	41.67	35.65	34.61
	1	1	0	41.49	35.42	34.52
	1	10	0	42.15	36.14	35.30
Variation in λ_g (λ_{ce} = 0 and λ_{LSB} = 0)	0	0.1	0	41.65	35.62	34.58
	0	0.5	0	41.54	35.45	34.41
	0	1	0	64.35	59.95	59.16
	0	10	0	42.15	36.15	35.30
Variation in λ_{LSB} (λ_{ce} = 1 and λ_g = 0)	1	0	1	42.00	35.96	34.89
	1	0	10	48.49	41.40	39.60
Variation in λ_{LSB} (λ_{ce} = 1 and λ_g = 1)	1	1	1	41.67	35.47	34.52
	1	1	10	46.07	37.54	35.79

Table 10: **Fashion-MNIST**: Recognition accuracy (%) of the model trained using the proposed approach on adversarial samples generated using adaptive attacks.

Adaptive attack	**Loss coefficients**			**n-step Adaptive attack**		
	λ_{ce}	λ_g	λ_{LSB}	40	100	500
PGD	1	0	0	68.03	67.75	67.71
Variation in λ_g (λ_{ce} = 1 and λ_{LSB} = 0)	1	1	0	69.41	69.22	69.19
	1	10	0	76.72	76.44	76.46
	1	25	0	78.95	78.8	78.79
Variation in λ_g (λ_{ce} = 0 and λ_{LSB} = 0)	0	1	0	80.45	80.23	80.2
	0	10	0	80.46	80.24	80.22
	0	25	0	80.46	80.22	80.18
	0	50	0	80.46	80.22	80.18
Variation in λ_{LSB} (λ_{ce} = 1 and λ_g = 0)	1	0	1	68.2	67.98	67.95
	1	0	10	71.32	70.98	70.98
Variation in λ_{LSB} (λ_{ce} = 1 and λ_g = 25)	1	25	1	78.96	78.8	78.77
	1	25	10	78.91	78.74	78.76

to cause misclassification, which can be achieved by maximizing only the first term in Eq. (1). However, the proposed defense mechanism could lead to masking of the true solution, thereby resulting in a weak attack. Thus, the role of the remaining terms, which take into account the defense mechanism, is to aid the optimization process in finding such points, if any. The remainder of the algorithm used is similar to that proposed by Madry *et al.* [9]. The results with adaptive attacks for CIFAR-10, Fashion-MNIST and MNIST datasets are presented in Tables-9, 10 and 11 respectively. We consider the following coefficients in Eq. (1) to find a strong adaptive attack:

Table 11: **MNIST**: Recognition accuracy (%) of the model trained using the proposed approach on adversarial samples generated using adaptive attacks.

Adaptive attack	**Loss coefficients**			**n-step Adaptive attack**		
	λ_{ce}	λ_g	λ_{LSB}	40	100	500
PGD	1	0	0	91.49	86.6	85.63
Variation in λ_g (λ_{ce} = 1 and λ_{LSB} = 0)	1	1	0	92.99	89.13	88.2
	1	10	0	94.58	91.75	91.04
	1	30	0	94.74	91.99	91.26
Variation in λ_g (λ_{ce} = 0 and λ_{LSB} = 0)	0	1	0	94.8	91.96	91.34
	0	10	0	94.8	91.97	91.34
	0	30	0	94.79	91.98	91.38
	0	50	0	94.79	91.97	91.35
Variation in λ_{LSB} (λ_{ce} = 1 and λ_g = 0)	1	0	1	91.56	86.96	86.01
	1	0	10	93.51	90.11	89.41
Variation in λ_{LSB} (λ_{ce} = 1 and λ_g = 30)	1	30	1	94.72	91.98	91.33
	1	30	10	94.7	91.97	91.36

- $\lambda_{ce} = 1,\ \lambda_g = 0,\ \lambda_{LSB} = 0$
  This corresponds to a standard PGD attack [9], which serves as a baseline in this table. The goal of the remaining experiments is to find an attack stronger than this.
- $\lambda_{ce} = 1,\ \lambda_g =$ variable, $\lambda_{LSB} = 0$
  This case corresponds to using the training loss directly to find adversarial samples. We find that lower values of λ_g lead to stronger attacks, while still not being significantly stronger than baseline. This indicates that addition of the BPFC regularizer does not help in the generation of a stronger attack.
- $\lambda_{ce} = 0,\ \lambda_g =$ variable, $\lambda_{LSB} = 0$
  For CIFAR-10 dataset, this case is able to generate attacks which are as strong as PGD, without using the cross-entropy term. This indicates that the BPFC loss term is relevant in the context of generating adversarial samples. However, addition of this to the cross-entropy term does not generate a stronger attack, as the defense is not masking gradients that prevents generation of stronger adversaries. However, for Fashion-MNIST and MNIST datasets, this attack is weaker than PGD.
- $\lambda_{ce} = 1,\ \lambda_g =$ variable, $\lambda_{LSB} =$ variable
  Next, we consider the case of introducing the third term that imposes a penalty on high magnitude LSBs. Addition of this term with or without the BPFC term does not help generate a stronger attack, indicating that this training regime does not create isolated points

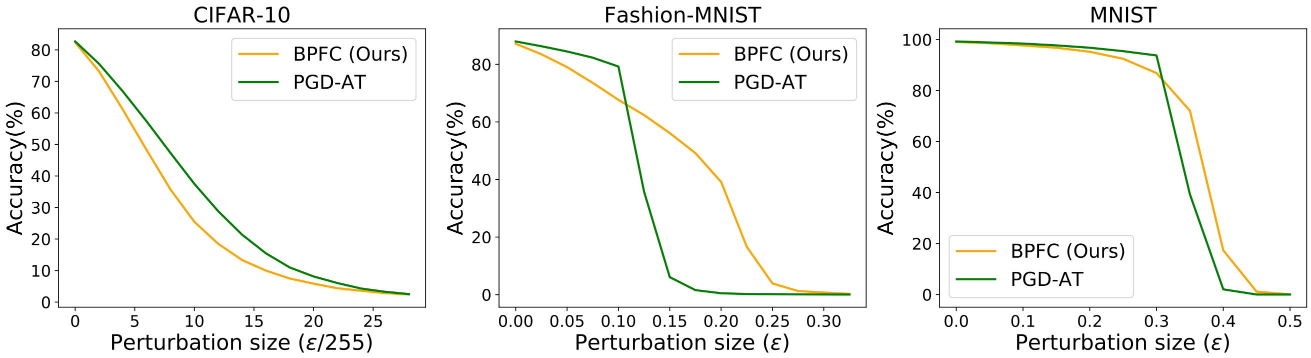

Figure 2: Plot of recognition accuracy (%) on PGD samples generated on test set versus perturbation size of PGD 7-step attack. The model's accuracy is zero for large perturbation sizes indicating the absence of gradient masking.

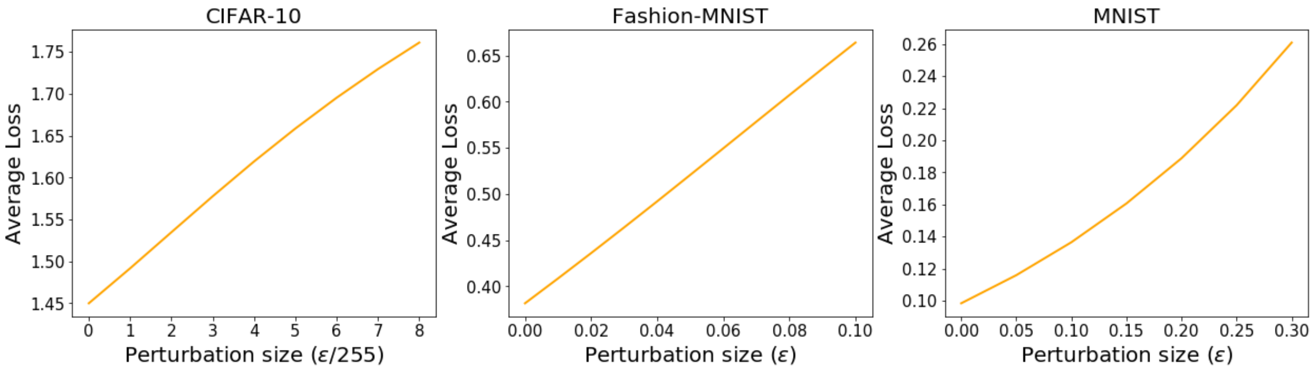

Figure 3: Plot of average loss on FGSM samples generated on test set versus perturbation size of FGSM attack.

in the ℓ_∞-ball around each sample, which correspond to points with low magnitude LSBs. This can be attributed to the addition of pre-quantization noise.

Overall, the adaptive attacks constructed based on the knowledge of the defense mechanism do not lead to stronger attacks. This leads to the conclusion that the proposed defense does not merely make the process of finding adversaries harder, but results in learning models that are truly robust.

2.4. Basic Sanity Checks to verify Robustness

In this section, we present details related to Section-5.7 in the main paper. The plots of accuracy verses perturbation size in Fig.2 demonstrate that unbounded attacks are able to reach 100% success rate. It can be observed that increasing the distortion bound increases the success rate of the attack. Fig-3 shows a plot of the average loss on FGSM samples generated on the test set, versus perturbation size of the FGSM attack. It can be observed that the loss increases monotonically with an increase in perturbation size. These two plots confirm that there is no gradient masking effect [1] in the models trained using the proposed approach.

Table 12: **ImageNet (white-box attacks):** Recognition accuracy (%) of BPFC trained model and PGD-AT model on ImageNet dataset under white-box attack setting. Accuracy is reported on the following untargeted attacks: FGSM attack, I-FGSM 20-step attack (IFGSM), PGD 20-step attack and PGD 100-step attack. Accuracy on PGD 20-step targeted attack (with random targets) is also reported (TPGD).

Training method	**Clean**	**FGSM**	**IFGSM** (20)	**PGD** (20)	**PGD** (100)	**TPGD** (20)
PGD-AT	47.91	24.42	21.52	19.39	19.06	43.43
BPFC (Ours)	40.82	19.97	15.93	13.41	12.82	32.91

2.5. Scalability of the Proposed Method to ImageNet

We report results on ImageNet dataset using the proposed method and PGD-AT in Table-12. The architecture used for both methods is ResNet-50 [6]. We use the PGD-AT pre-trained model from [5] for comparison. We train the proposed method for 125 epochs and decay learning rate by a factor of 10 at epochs 35, 70 and 95. Similar to CIFAR-10, we start with a λ of 1 and step it up by a factor of 9 at epochs 35 and 70. We use a higher step-up factor of 20 at epoch 95 to improve robustness. Since training ImageNet models is computationally intensive, we report results using similar hyperparameters as that of CIFAR-10. However, tuning hyperparameters specifically for ImageNet can lead to improved results. Accuracy on black-box FGSM attack is 47.39% for PGD-AT and 40.41% for the BPFC trained model. We note that the trend in robustness when compared to PGD-AT is similar to that of CIFAR-10, thereby demonstrating the scalability of the proposed approach to large scale datasets.

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. 5

[2] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, and Aleksander Madry. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019. 2

[3] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017. 3

[4] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[5] Logan Engstrom, Andrew Ilyas, Shibani Santurkar, and Dimitris Tsipras. Robustness (python library), 2019. 5

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[7] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016. 3

[8] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 1

[9] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Tsipras Dimitris, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018. 4

[10] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[11] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3