Weight-Covariance Alignment for Adversarially Robust Neural Networks

Panagiotis Eustratiadis[1] Henry Gouk[1] Da Li[1,2]
Timothy Hospedales[1,2]

[1]School of Informatics, University of Edinburgh

[2]Samsung AI Center, Cambridge

Contributions

- We contribute a Stochastic Neural Network (SNN) for adversarial robustness
  - Trained on clean samples only
  - Using anisotropic noise
- We derive a theoretic bound for the margin between clean and adversarial performance
  - We propose a simple technical implementation for tightening that bound

Weight-Covariance Alignment (WCA)

Our SNN architecture:

$$h(\vec{x}) = \vec{w}^T(f(\vec{x}) + \vec{z}) + b, \quad \vec{z} \sim \mathcal{N}(0, \Sigma) \tag{1}$$

$f(\vec{x}) \longrightarrow$ any feature extractor (e.g., ResNet, VGG, etc.)

The theoretic bound:

$$G^h_{p,\epsilon}(\vec{x}, y) \leq \frac{\Delta^{\tilde{h}}_p(\vec{x}, \epsilon)}{\sqrt{2\pi \vec{w}^T \Sigma \vec{w}}} \tag{2}$$

$G^h_{p,\epsilon}(\vec{x}, y) \longrightarrow$ the difference in probability of misclassification when the network is, and isn't under attack

$\Delta^{\tilde{h}}_p(\vec{x}, \epsilon) \longrightarrow$ the magnitude by which an adversarial perturbation, δ, causes the output of $\tilde{h}$ to change

WCA as a Loss Term

$$G^h_{p,\epsilon}(\vec{x}, y) \leq \frac{\Delta_p^{\tilde{h}}(\vec{x}, \epsilon)}{\sqrt{2\pi \vec{w}^T \Sigma \vec{w}}}$$

To maximize $\vec{w}^T \Sigma \vec{w}$ we devise a simple loss term:

$$\mathcal{L} = \mathcal{L}_{\mathrm{C}} - \mathcal{L}_{\mathrm{WCA}}$$

Where $\mathcal{L}_{\mathrm{C}}$ is the classification loss (e.g., cross entropy) and:

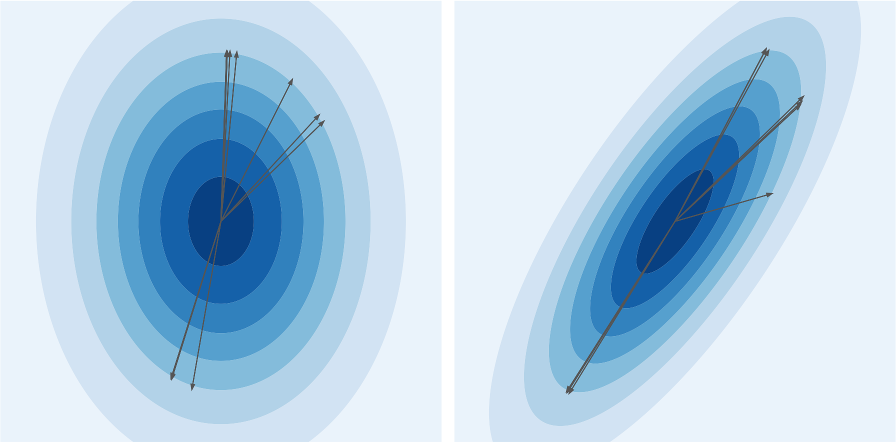
$$\mathcal{L}_{\mathrm{WCA}} = \sum_{i=1}^{C} \ln(\vec{w}_i^T \Sigma \vec{w}_i)$$

THE UNIVERSITY of EDINBURGH

Results on CIFAR

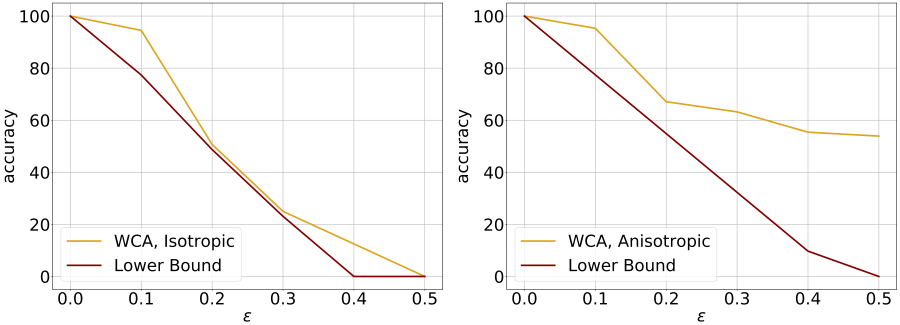
	CIFAR-10			CIFAR-100		
Method	Clean	FGSM	PGD	Clean	FGSM	PGD
Adv-BNN	82.2	60.0	53.6	~ 58.0	~ 30.0	~ 27.0
PNI	87.2	58.1	49.4	~ 61.0	~ 27.0	~ 22.0
L2P	85.3	62.4	56.1	~ 50.0	~ 30.0	~ 26.0
SE-SNN	92.3	74.3	-	-	-	-
IAAT	-	-	-	63.9	-	18.5
WCA	**93.2**	**77.6**	**71.4**	**70.1**	**51.5**	**42.7**

Empirical Observations About WCA



Left: WCA Isotropic, Right: WCA Anisotropic

Empirical Evaluation of the Bound



Thank you

THE UNIVERSITY of EDINBURGH