Feature Scattering Adversarial Training

Runtian Zhai

MSRA

v-ruzhai@microsoft.com

November 1, 2019

Defense Against Adversarial Attacks Using Feature Scattering-based Adversarial Training

Haichao Zhang & Jianyu Wang, Baidu Research USA, NeurIPS 2019

Main idea:

- They propose to replace PGD used in adversarial training with feature scattering attack: an attack which maximizes the OT distance between the features of clean images and perturbed images.
- Their method achieves over 64% robust test accuracy against PGD on Cifar-10 (reproduced).

Maximizing OT Distance

- During each training iteration, they maximize the OT distance between $\{f(x) : x \in \mathcal{B}\}$ and $\{f(x') : x' \in \mathcal{B}'\}$, where $\mathcal{B}$ is the batch of clean inputs, $\mathcal{B}'$ is the batch of perturbed inputs, and f is a neural network whose last layer is softmax.
- The transport cost between $f(x)$ and $f(x')$ is defined as

$$c(x, x') = 1 - \frac{< f(x), f(x') >}{\|f(x)\|_2 \|f(x')\|_2} \qquad \text{(cosine distance)}$$

- The OT distance (OT loss) is computed with Sinkhorn, and $\mathcal{B}'$ is generated with one step of projected sign gradient ascent.

Adversarial Training with Feature Scattering

For each iteration:

1. Draw a mini-batch $\mathcal{B}$.
2. Random start: Initialize $\mathcal{B}'$ by sampling from a uniform distribution over the perturbation range.
3. Compute OT Loss with Sinkhorn: $L_{OT}(\mathcal{B}, \mathcal{B}')$
4. Update: $\mathcal{B}' \leftarrow Proj(\mathcal{B}' + \alpha \cdot \text{sign}(\nabla_{\mathcal{B}'} L_{OT}(\mathcal{B}, \mathcal{B}')))$
5. Train f over $\mathcal{B}'$ and the original labels with label smoothing.

Label Smoothing Matters

- Standard cross entropy loss is defined as the entropy between $f(x)$ and one-hot vector $(0, \cdots, 0, 1, 0, \cdots, 0)$. Label smoothed cross entropy is the entropy between $f(x)$ and $(\frac{\delta}{K-1}, \cdots, \frac{\delta}{K-1}, 1-\delta, \frac{\delta}{K-1}, \cdots, \frac{\delta}{K-1})$.
- Ablation study on δ:

δ	Robust Test Acc at Epoch 200 (%)
0.0	43.03
0.1	63.67
0.3	67.14
0.5	72.03

No Overfitting

- Unlike standard adversarial training, adversarial training with feature scattering does not suffer from overfitting. Its robust test accuracy does not decrease after learning rate decay.

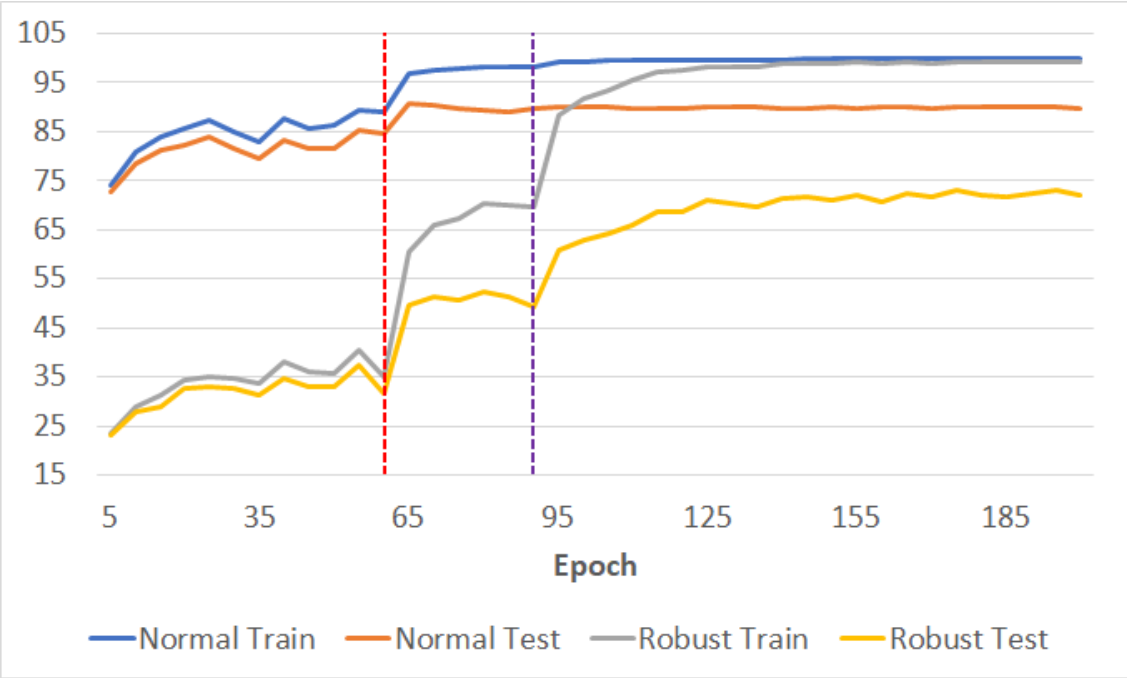
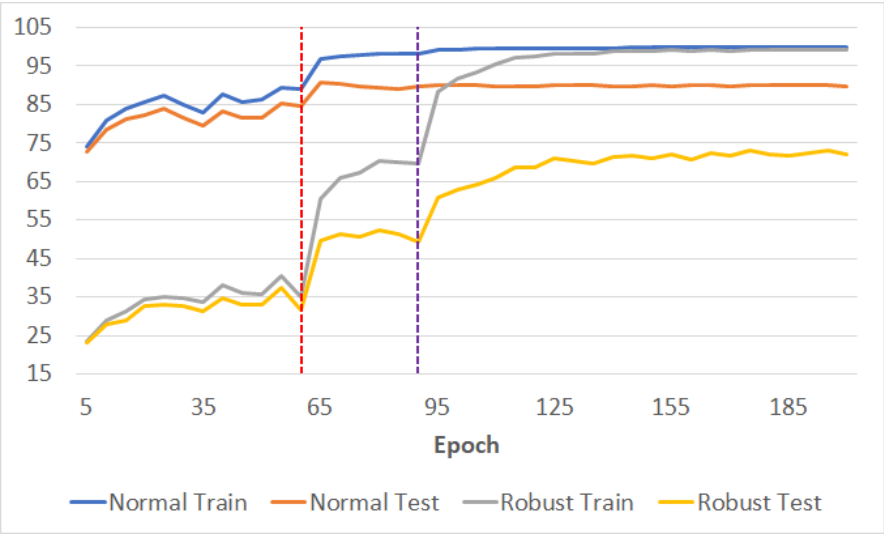
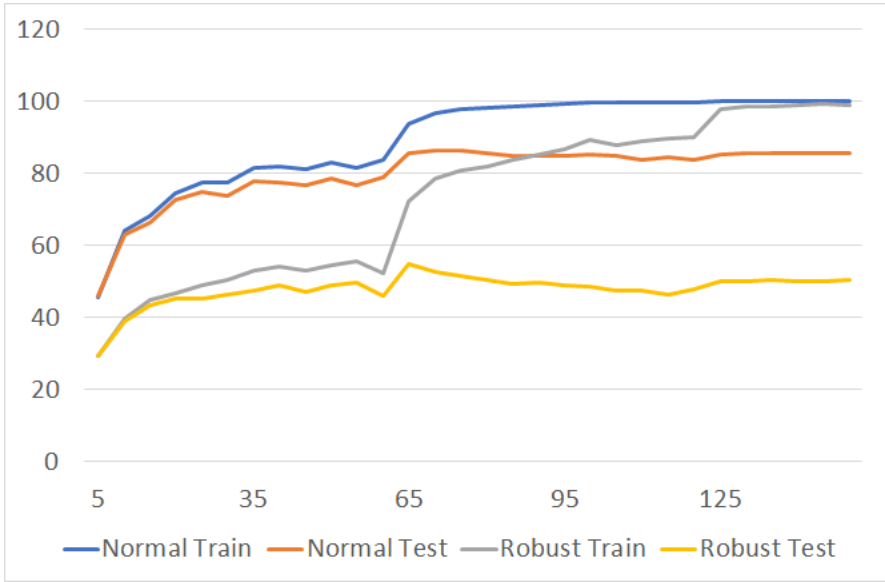


Figure: $\delta = 0.5$. LR decays at Epochs 60 and 90.

Figures

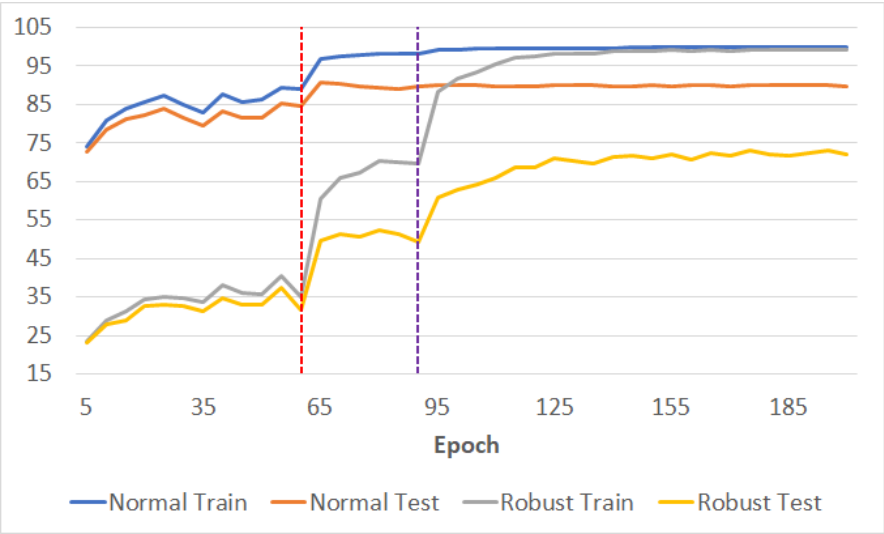


(a) Fea-Scatter

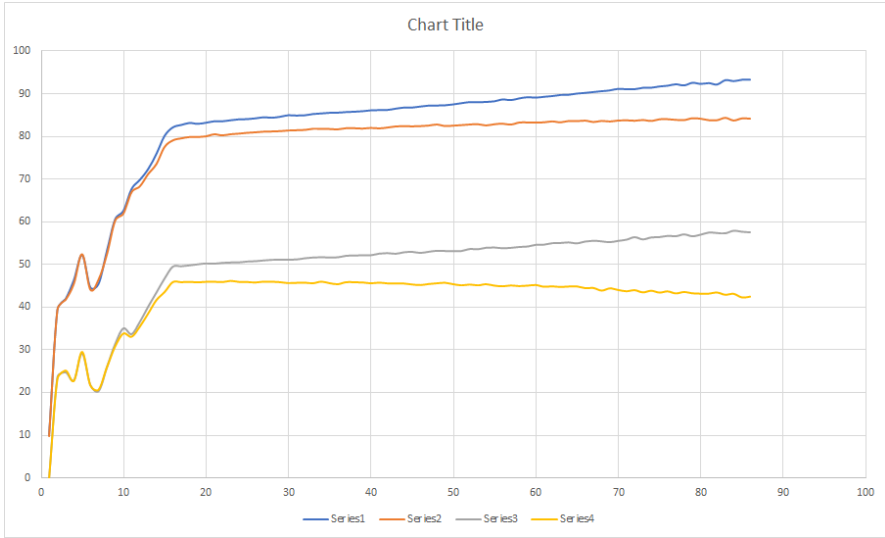


(b) Adv-train

Figures



(c) Fea-Scatter



(d) Fast

Weaknesses

- According to the results in the paper, their model has 68.6% robust test acc against PGD-100 on Cifar-10 but only 60.6% against CW-100. On Cifar-100, their model has 46.3% against PGD-100 but only 30.6% against CW-100. However, normally we believe that CW is weaker than PGD. Is it possible that their method overfits PGD?