Batch-wise Logit-Similarity: Generalizing Logit-Squeezing and Label-Smoothing

Ali Shafahi, Amin Ghiasi, Mahyar Najibi, Furong Huang, John Dickerson, Tom Goldstein

Intro

- We introduce logit-similarity, a generalization of label-smoothing and logit-squeezing which shows how cheap regularization methods can increase adversarial robustness.
- Our version of logit-squeezing applies a batch-wise penalty and allows penalizing the logits aggressively.
- We experimentally show that, with the correct choice of hyper-parameters (standard deviation of Gaussian augmentation, and Logit-Similarity coefficient), regularized models can be as robust as adversarially trained models. Unlike adversarial training, regularization methods are efficient and robust against ℓ_2 attacks in addition to ℓ_∞.

Logit-Squeezing

Logit-Squeezing (L-SQ on example) is penalizing the magnitude of logits while training

$$\underset{\theta}{\text{minimize}}\ l(x,y,\theta) + \beta\|z(x)\|_2$$

We propose Batch-wise Logit-Squeezing (L-SQ on batch):

$$\underset{\theta}{\text{minimize}} \sum l_b(x_b,y_b,\theta) + \frac{\beta}{b_n}\|Z(x_b)\|_F$$

b_n → batch-size

CIFAR-10 Logit-Squeezing Results

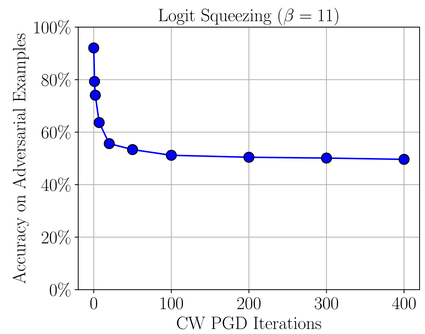
defense	Test	PGD attacks on the xent 20-xent	50-xent	200-xent	PGD attacks on the CW 20-CW	50-CW	200-CW
$\beta = 5$	92.45%	43.26%	43.25%	38.86%	45.50%	37.82%	33.91%
$\beta = 10$	**92.68%**	52.55%	45.18%	40.83%	47.48%	41.39%	37.87%
$\beta = 11$	92.08%	**58.51%**	**55.91%**	**53.87%**	**55.63%**	**53.56%**	**50.44%**
7step-AdvT	87.25%	45.84%	45.39%	45.32%	46.90%	46.66%	46.48%

CIFAR-100 Logit-Squeezing Results

defense	PGD attacks on the xent and CW loss 20-xent	200-xent	20-cw	200-cw
$\beta = 1$	23.89%	18.99%	11.91%	9.00%
$\beta = 5$	30.91%	26.00%	19.80%	15.79%
$\beta = 7$	**31.99%**	**30.05%**	**25.92%**	**23.87%**
2step-AdvT	17.08%	16.49%	17.80%	17.52%
7step-AdvT	22.76%	22.42%	23.12%	22.95%

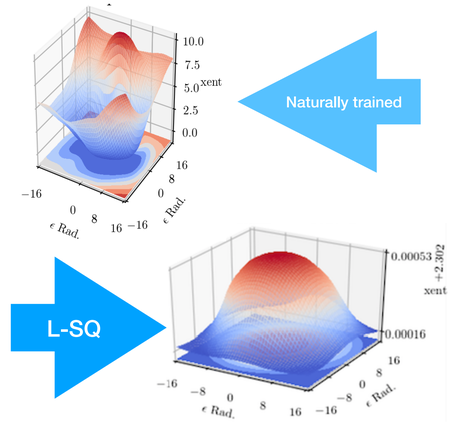
Robustness vs # of PGD Iterations

The following plot shows the accuracy w.r.t. # of PGD iterations for our CIFAR-10 L-SQ batch $\beta = 11$ WRN32-10 model.



Why Logit-Squeezing works

By aggressive logit-squeezing the loss landscape w.r.t. the input is flattened.



Logit-Similarity

If our hypothesis about why Logit-Squeezing works is correct, we should be able to get similar behavior by clustering the logits to be similar and close to any scalar γ.

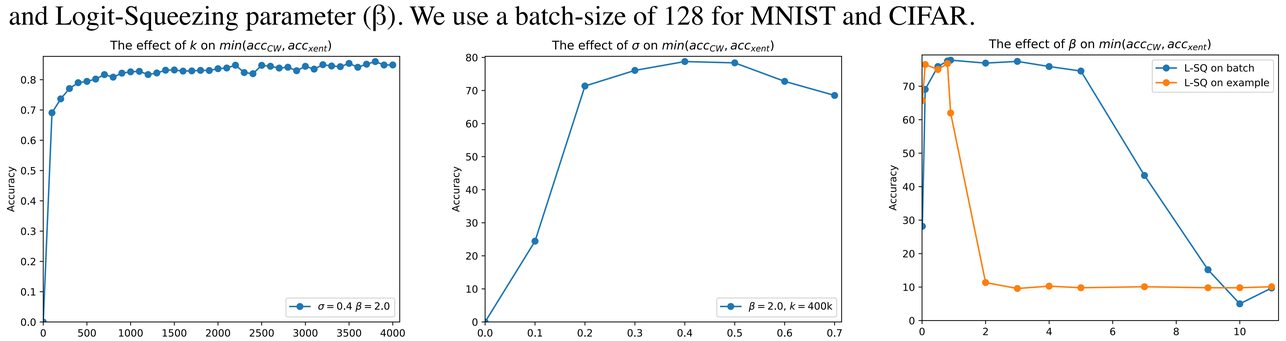
$$\underset{\theta}{\text{minimize}} \sum_b l_b(x_b,y_b,\theta) + \beta'/b_n\|Z(x_b) - \gamma\|_F$$

CIFAR-10 Logit-Similarity (γ = 1)

defense	PGD attacks on the xent and CW loss 20-xent	200-xent	20-cw	200-cw
$\beta' = 5$	44.00%	30.59%	40.77%	28.65%
$\beta' = 10$	47.28%	35.67%	45.52%	34.56%
$\beta' = 11$	**56.36%**	**49.79%**	**56.74%**	**50.33%**
7step-AdvT	45.84%	45.32%	46.90%	46.48%

Ablation Study on MNIST

This study shows the effect of number of training iterations (k), standard deviation of Gaussian augmentation (σ) and Logit-Squeezing parameter (β). We use a batch-size of 128 for MNIST and CIFAR.



Robustness on Other Attacks

- 10 Random restarts 100-PGD results in 45.27% accuracy.
- While 7-PGD trained model on ℓ_∞ adversaries achieves 15.36% robustness against ℓ_2 perturbations (ε=1.5×255), our model (β'=11) achieves 54.99%.
- The same model preserves 88.13% accuracy against grad-free attacks (SPSA, #iters=20, #instances=248).

Label-Smoothing

Label smoothing refers to making the "one-hot" label vectors into "one-warm" vectors to promote clustering of logits:

$$y_{warm} = y_{hot} - \lambda \times (y_{hot} - \frac{1}{N_c})$$

CIFAR-10 Label-Smoothing Results

defense	PGD attacks on the xent and CW loss Test	20-xent	20-cw
$\lambda = 0.9$	92.60%	43.30%	39.76%
$\lambda = 0.95$	**92.88%**	43.00%	41.29%
7step-AdvT	87.25%	**45.84%**	**46.90%**