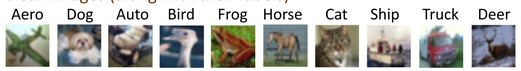
# Towards Achieving Adversarial Robustness Beyond Perceptual Limits

Sravanti Addepalli*, Samyak Jain*, Gaurang Sriramanan, Shivangi Khare, R.Venkatesh Babu
Video Analytics Lab, Indian Institute of Science, Bangalore, India

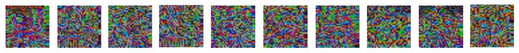
## Preliminaries: Threat Model and Terminology

- Threat model considered: Moderate-ε $\ell_\infty$ norm bound of 16/255
- Human prediction is referred to as the *Oracle Label*
- Types of perturbations within the defined threat model:
  - ***Oracle-Invariant*** images: Do not change Oracle prediction
    - Adversarial examples generated from Normally trained models
    - Adversarial examples at low perturbation bounds (ε = 8/255)

Clean Images (along with their labels)




Adversarial Perturbations generated from a **Normally Trained** (NT) Model



**Oracle-Invariant** Adversarial examples (along with NT model predictions)

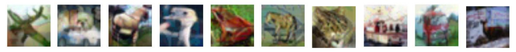

  - ***Oracle-Sensitive*** images: Flip the Oracle prediction
    - Adversarial examples generated from Adversarially Trained models at large perturbation bounds (ε = 32/255)

Clean Images (along with their labels)

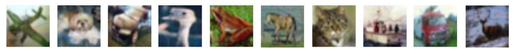

Adversarial Perturbations generated from an **Adversarially Trained** (AT) Model

**Oracle-Sensitive** Adversarial examples (along with AT model predictions)


**Partially Oracle-Sensitive** Adversarial examples




**Oracle-Invariant** Adversarial examples




## Goals and Evaluation Metrics

- Robustness against all attacks within ε = 8/255
  - Auto-Attack (AA) and Guided Adversarial Margin Attack (GAMA) PGD-100
- Robustness to Oracle-Invariant samples within ε = 16/255
  - Gradient-Free Attacks

Clean Images (along with their labels)


Adversarial examples generated using **Square Attack** (along with AT model predictions)


- Robustness-Accuracy trade-off

## Scaling existing AT methods to larger ε bounds


| Method | Attack ε (Training) | Clean | GAMA (8/255) | AA (8/255) | GAMA (16/255) | Square (16/255) |
|---|---|---|---|---|---|---|
| PGD-AT | 8/255 | 81.12 | 49.03 | 48.58 | 15.77 | 26.47 |
| PGD-AT | 16/255 | 64.93 | 46.66 | 46.21 | **26.73** | **32.25** |

Significant drop (16%) in Clean Accuracy when compared to Adversarial Training at ε = 8/255

## Results

| Method | Clean | GAMA 8/255 | AA 8/255 | SQ+RS 16/255 | GAMA 16/255 | AA 16/255 |
|---|---|---|---|---|---|---|
| **CIFAR-10 (ResNet-18), 110 epochs** | | | | | | |
| FAT | **84.36** | 48.41 | 48.14 | 23.22 | 15.18 | 14.22 |
| PGD-AT | 79.38 | 49.28 | 48.68 | 25.43 | 18.18 | 17.00 |
| AWP | 80.32 | 49.06 | 48.89 | 25.99 | 19.17 | 18.77 |
| ATES | 80.95 | 49.57 | 49.12 | 26.43 | 18.36 | 16.30 |
| TRADES | 80.53 | 49.63 | 49.42 | 26.20 | 19.27 | 18.23 |
| ExAT + PGD | 80.68 | 50.06 | 49.52 | 25.13 | 17.81 | 19.53 |
| ExAT + AWP | 80.18 | 49.87 | 49.69 | 27.04 | 20.04 | 16.67 |
| AWP | 80.47 | 50.06 | 49.87 | 27.20 | 19.66 | 19.23 |
| Ours | 80.24 | **51.40** | **50.88** | **29.56** | **22.73** | **22.05** |
| **CIFAR-10 (ResNet-34), 110 epochs** | | | | | | |
| AWP | 83.89 | 52.64 | 52.44 | 27.69 | 20.23 | 19.69 |
| OA-AT (Ours) | **84.07** | **53.54** | **53.22** | **30.76** | **22.67** | **22.00** |
| **CIFAR-10 (WRN-34-10), 110 epochs** | | | | | | |
| AWP | 85.19 | 55.87 | 55.69 | 31.27 | 24.04 | 23.46 |
| AWP+ | 85.10 | 56.07 | 55.87 | 31.36 | 23.79 | 23.27 |
| OA-AT (Ours) | **85.67** | **56.45** | **55.93** | **33.89** | **25.21** | **24.05** |

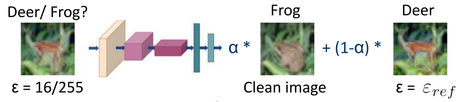
| Method | Clean | GAMA 8/255 | AA 8/255 | SQ+RS 16/255 | GAMA 16/255 | AA 16/255 |
|---|---|---|---|---|---|---|
| **CIFAR-100 (ResNet-18), 110 epochs** | | | | | | |
| AWP | 58.81 | 25.51 | 25.30 | 11.39 | 8.68 | 8.29 |
| AWP+ | 59.88 | 25.81 | 25.52 | 11.85 | 8.72 | 8.28 |
| OA-AT (no LS) | 60.27 | 26.41 | 26.00 | 13.48 | **10.47** | **9.95** |
| OA-AT (Ours) | **61.70** | **27.09** | **26.77** | **13.87** | 10.40 | 9.91 |
| **CIFAR-100 (PreActResNet-18), 200 epochs** | | | | | | |
| AWP | 58.85 | 25.58 | 25.18 | 11.29 | 8.63 | 8.19 |
| AWP+ | **62.11** | 26.21 | 25.74 | 12.23 | 9.21 | 8.55 |
| OA-AT (Ours) | 62.02 | **27.45** | **27.14** | **14.52** | **10.64** | **10.10** |
| **CIFAR-100 (WRN-34-10), 110 epochs** | | | | | | |
| AWP | 62.41 | 29.70 | 29.54 | 14.25 | 11.06 | 10.63 |
| AWP+ | 62.73 | 29.92 | 29.59 | 14.96 | 11.55 | 11.04 |
| OA-AT (no LS) | 65.22 | 30.75 | **30.35** | 16.77 | 12.65 | 11.95 |
| OA-AT (Ours) | **65.73** | **30.90** | **30.35** | **17.15** | **13.21** | **12.01** |
| **SVHN (PreActResNet-18), 110 epochs** | | | | | | |
| **Method** | **Clean** | **GAMA 4/255** | **AA 4/255** | **SQ+RS 12/255** | **GAMA 12/255** | **AA 12/255** |
| AWP | 91.91 | 75.92 | 75.72 | 35.49 | 30.70 | 30.31 |
| OA-AT (Ours) | **94.61** | **78.37** | **77.96** | **39.24** | **34.25** | **33.63** |

## Oracle-Aligned Adversarial Training

- Varying epsilon training schedule with TRADES-AWP loss (CE loss for attack)
- Standard Adversarial training till the perceptual limit of ε = 12/255

$$L = \ell_{CE}(f_\theta(x_i), y_i) + L_{adv},\ L_{adv} = \mathrm{KL}\left(f_\theta(x_i + \widetilde{\delta}_i) || f_\theta(x_i)\right)$$

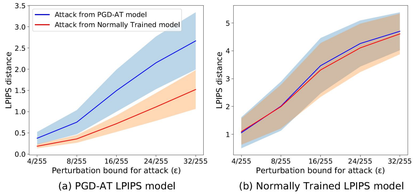
- From ε = 12/255 till 16/255, training on Oracle-Sensitive and Oracle-Invariant samples in alternate iterations
- Sensitivity towards Oracle-Sensitive attacks generated by maximizing CE loss




$$\ell = \ell_{CE}(f_\theta(x_i + \delta_i), y_i),\ \widehat{\delta}_i = \mathrm{PGD}(x_i, y_i, \ell_{CE}, \varepsilon_{ref}),\ \widetilde{\delta}_i = \Pi_\infty(\widehat{\delta}_i, \widetilde{\varepsilon})$$

$$L_{adv} = \mathrm{KL}\left(f_\theta(x_i + \widetilde{\delta}_i) \ \ || \ \ \alpha \cdot f_\theta(x_i) + (1-\alpha) \cdot f_\theta(x_i + \widehat{\delta}_i)\right)$$

- Robustness to Oracle-Invariant attacks




(a) PGD-AT LPIPS model (b) Normally Trained LPIPS model

LPIPS metric computed using an AT model can distinguish well between Oracle-Sensitive (Attack from PGD-AT model) and Oracle-Invariant (Attacks from Normally trained model) examples successfully.

  - Generation of Oracle-Invariant attacks by minimizing LPIPS distance (using the model being trained) between clean and perturbed images

$$\ell = \ell_{CE}(f_\theta(x_i + \delta_i), y_i) - \lambda \cdot \mathrm{LPIPS}(x_i, x_i + \delta_i)$$


## Ablations

| Method | CIFAR-10 Clean | CIFAR-10 GAMA (8/255) | CIFAR-10 GAMA (16/255) | CIFAR-10 Square (16/255) | CIFAR-100 Clean | CIFAR-100 GAMA (8/255) | CIFAR-100 GAMA (16/255) | CIFAR-100 Square (16/255) |
|---|---|---|---|---|---|---|---|---|
| **E1**: OA-AT (Ours) | 80.24 | **51.40** | 22.73 | 31.16 | 60.27 | **26.41** | 10.47 | 14.60 |
| **E2**: LPIPS weight = 0 | 78.47 | 50.60 | 24.05 | 31.37 | 58.47 | 25.94 | 10.91 | 14.66 |
| **E3**: Alpha = 1 | 79.29 | 50.60 | 23.65 | 31.23 | 58.84 | 26.15 | 10.97 | 14.89 |
| **E4**: Alpha = 1, LPIPS weight = 0 | 77.16 | 50.49 | **24.93** | **32.01** | 57.77 | 25.92 | **11.33** | **15.03** |
| **E5**: Without AutoAugment | 80.24 | **51.40** | 22.73 | 31.16 | 58.08 | 25.81 | 10.40 | 14.31 |
| **E6**: With AutoAugment (p=0.5) | 81.59 | 50.40 | 21.59 | 30.84 | 60.27 | **26.41** | 10.47 | 14.60 |
| **E7**: With AutoAugment (p=1) | **81.74** | 48.15 | 18.92 | 28.31 | 60.19 | 25.32 | 9.24 | 13.78 |

**Acknowledgements:** This work was supported by Uchhatar Avishkar Yojana (UAY) project (IISC 10), MHRD, Govt. of India. Sravanti Addepalli is supported by a Google PhD Fellowship in Machine Learning. We are thankful for the support.