Defending Against Universal Attacks Through Selective Feature Regeneration

Tejas Borkar¹, Felix Heide²,³, Lina Karam¹,⁴

¹Arizona State University  ²Princeton University  ³Algolux  ⁴Lebanese American University
Universal Adversarial Attacks

- Image agnostic and transferable across networks
Defending against Universal Adversarial Attacks

- Selective feature regeneration effectively restores robustness

Proposed defense: Baseline DNN with resilient feature regeneration $\Phi_{reg}$
Ranking CNN Filters Based on Noise Susceptibility

We show:

- Max perturbation level induced in feature map ∝ \( \ell_1 \)-norm of the filter weight \( \|W\|_1 \)

Suppressing perturbations in ranked filters’ output maps

Percentage of suppressed maps in conv-1

- VGG-16: 0.69%
- Googlenet: 0.7%
- Caffenet: 0.56%

Top-1 accuracy

0 0.2 0.4 0.6 0.8
Robustness to Unseen Universal Adversarial Attacks

- Defense trained on only UAP noise samples

Clean image

UAP, NAG, GAP, sPGD

Clean map

Perturbed feature map

Regenerated resilient feature map
Defending Against Universal Attacks Through Selective Feature Regeneration

Robustness to image-agnostic noise:

Adversarial perturbation

Input image

Perturbed image

Adversarial noise

Predictions

PD: croquet ball 77%
Ours: ice cream 50%

FD: croquet ball 10%
Ours: ice cream 83%

HGD: mixing bowl 30%
Ours: ice cream 66%

Robustness to unseen universal attacks:

NAG

GAP

sPGD

Summary:

- Novel $\ell_1$-norm measure identifies and ranks adversarially susceptible feature maps
- Selective regeneration of only the most vulnerable feature maps restores robustness

Code: https://github.com/tsborkar/Selective-feature-regeneration