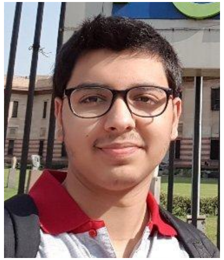
REGroup: Rank-aggregating Ensemble of Generative Classifiers for Robust Predictions

WACV
WAIKOLOA, HI JAN 4-8

Lokender Tiwari

Anish Madan

Saket Anand

Subhashis Banerjee

Project Page: https://lokender.github.io/REGroup.html

WACV Paper ID : 885

INDRAPRASTHA INSTITUTE of INFORMATION TECHNOLOGY DELHI

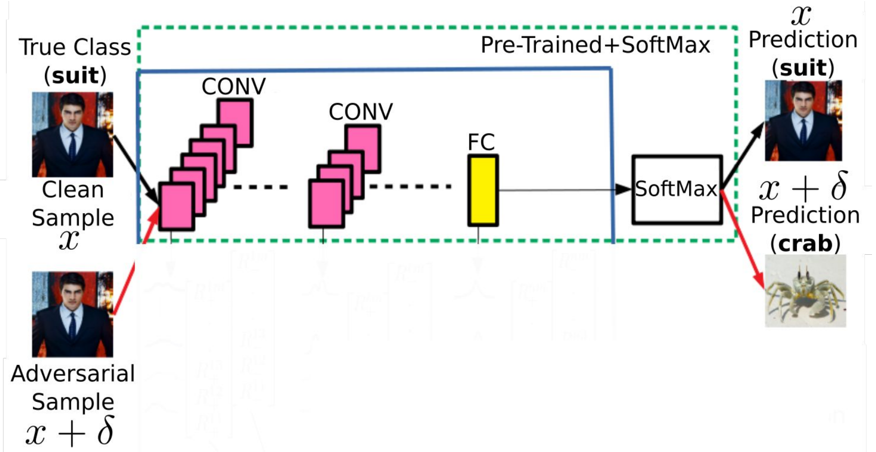
tcs Research

INDIAN INSTITUTE OF TECHNOLOGY DELHI

ashoka UNIVERSITY

Motivation

- Deep Neural Network based image classifiers can be fooled by adversarial samples



- Successful defenses:
    - Adversarial Training [1] : Train classifier using both clean and adversarial samples
    - Input randomization [2] before passing to a classifier
- Require fine-tuning or retraining (computational expensive and time consuming)
    - Adversarial training for full scale ImageNet classification
        - 52 hours on 128 NVIDIA V100 GPUs for ResNet-152 based classifier model [1]

[1] Cihang Xie et al. “Feature Denoising for Improving Adversarial Robustness”. CVPR, 2019
[2] Edward Raff et al. “Barrage of Random Transforms for Adversarially Robust Defense”. CVPR, 2019

Motivation

- Most defense methods [1]
  - are attack specific, architecture specific
  - practically not scalable (e.g., full ImageNet level)

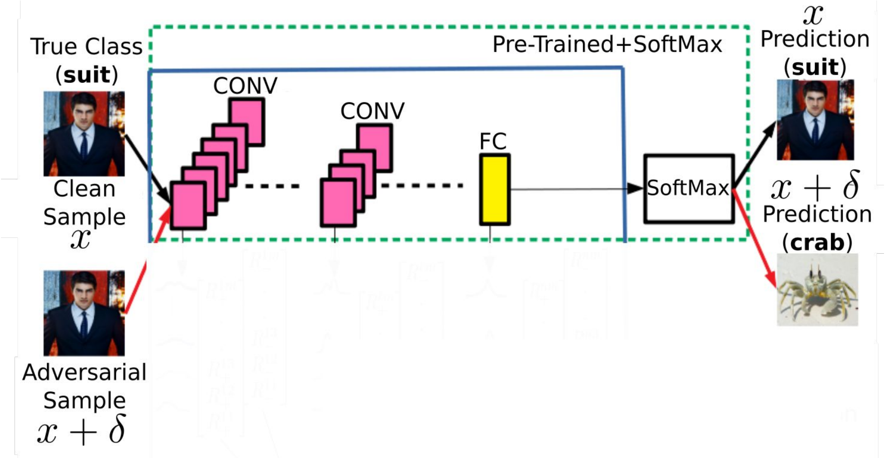
- Need for a defense mechanism
  - agnostic to classifier architectures and the adversarial attack generation method
  - can detect and make correct prediction for adversarial examples
  - easy to scale to large scale classification task

- **REGroup: Rank-aggregating Ensemble of Generative classifiers for robust predictions**

[1] Madry et al, https://www.robust-ml.org/

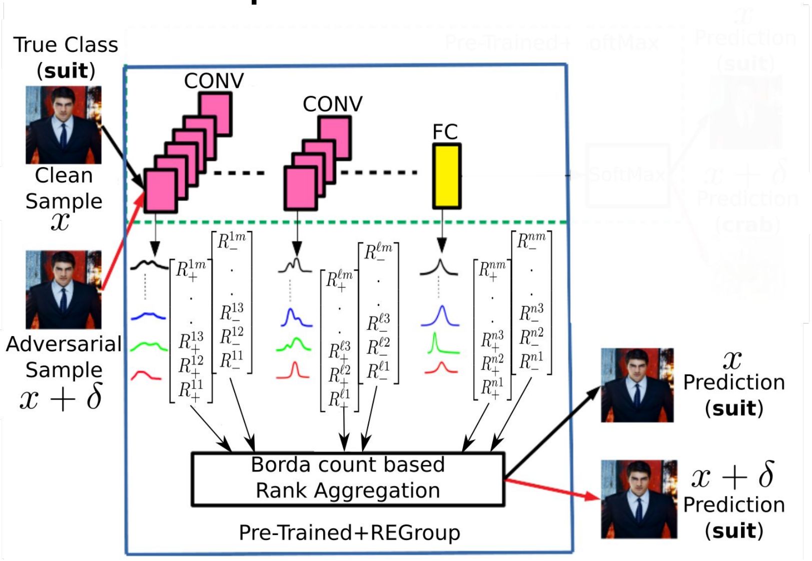
REGroup Overview

Normal Inference Procedure



- SoftMax based final prediction

REGroup inference Procedure



- Layer-wise ranked predictions
- Final prediction based on aggregated rankings

REGroup Overview

1. Generative classifiers
   - Class conditional layer-wise mixture distributions
     - Two distributions per layer, per class
       - Using positive pre-activation neural responses
       - Using negative pre-activation neural responses

One time only

2. Inference step
   - Layer-wise comparison of sample distributions with the class conditional distributions
     - KL-Divergence
   - Make layer-wise ranked predictions
   - Use robust rank aggregation strategy to aggregate ranked predictions
   - Final prediction is the class, with the highest rank

Experiments

- Classifier architectures: VGG19 [1] and ResNet50 [2]
- Dataset: ImageNet [3]
- Adversarial attacks
    - Gradient Based Attacks (White box)
        - Full access to network parameters
    - Gradient Free Attacks (Black box)
        - Classifier is a back-box

[1] Simonyan et al. “Very Deep Convolutional Networks for Large-Scale Image Recognition”, ICLR, 2015
[2] He at al. “Deep Residual Learning for Image Recognition”, CVPR, 2016
[3] Jia Deng et al. “ImageNet: A Large-Scale Hierarchical Image Database”. CVPR, 2009

Experiments

- **Gradient Based Attacks:**
  - Projective Gradient Descent (PGD) [1]
  - DeepFool (DFool) [2]
  - Carlini and Wagner (C&W)[3]
  - Trust Region attack (TR) [4]
  - Color adversarial attack (cAdv) [5]

UN : Untargeted Attack
TA : Targeted Attack
HC : High Confidence (at-least 90%)
ϵ : Adversarial Perturbation Budget
#S : No. of adversarial samples
T1(%) : Top-1 accuracy

				ResNet-50			VGG-19		
		UN /			SMax	REGroup		SMax	REGroup
	Data	TA / HC	ϵ	#S	T1(%)	T1(%)	#S	T1(%)	T1(%)
Clean	V10K	–	–	10000	100	88	10000	100	76
Clean	V2K	–	–	2000	100	86	2000	100	72
Clean	V10C	–	–	417	100	84	392	100	79
PGD	V10K	UN	4 (L_∞)	9997	0	48	9887	0	46
DFool	V10K	UN	2 (L_2)	9789	0	61	9939	0	55
C&W	V10K	UN	4 (L_2)	10000	0	40	10000	0	38
TR	V10K	UN	2 (L_∞)	10000	0	41	9103	0	45
cAdv	V10C	UN	–	417	0	37	392	0	18
PGD	V2K	TA	(L_∞)	2000	0	47	2000	0	31
C&W	V2K	TA	(L_2)	2000	0	46	2000	0	38
PGD	V2K	UN+HC	(L_∞)	2000	0	21	2000	0	19
PGD	V2K	TA+HC	(L_∞)	2000	0	23	2000	0	17

Tab 1. Classification accuracy on gradient based attacks

[1] Madry, Aleksander, et al. "Towards Deep Learning Models Resistant to Adversarial Attacks." *ICLR*. 2018
[2] Moosavi-Dezfooli et al. "Deepfool: a simple and accurate method to fool deep neural networks." *CVPR*. 2016
[3] Carlini, Nicholas, and David Wagner. "Towards evaluating the robustness of neural networks." *IEEE symposium on security and privacy*, 2017
[4] Yao, Zhewei, et al. "Trust region based adversarial attack on neural networks." *CVPR*, 2019
[5] Bhattad, Anand, et al. "Unrestricted Adversarial Examples via Semantic Manipulation." *ICLR*. 2019

Experiments

- **Gradient Free Attacks** (Black Box Attacks):
    - SPSA Attack [1]
    - Boundary Attack [2]
    - Spatial Attack [3]

				ResNet-50			VGG-19		
					SMax	REGroup		SMax	REGroup
	Data	UN / TA / HC	ϵ	#S	T1(%)	T1(%)	#S	T1(%)	T1(%)
SPSA	V10K	UN	4 (L_∞)	4911	0	71	5789	0	58
Boundary	V10K	UN	2 (L_2)	10000	0	50	10000	0	50
Spatial	V10K	UN	2 (L_2)	2624	0	36	2634	0	30

Tab 2. Classification accuracy on gradient free attacks

UN : Untargeted Attack
ϵ : Adversarial Perturbation Budget
#S : No. of adversarial samples
T1(%) : Top-1 accuracy

[1] Jonathan Uesato et al. "Adversarial Risk and the Dangers of Evaluating Against Weak Attacks". ICML, 2018
[2] Brendel et al. "Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models". ICLR, 2018
[3] Logan Engstrom et al. "Exploring the Landscape of Spatial Robustness". ICML, 2019

Experiments

- Comparison with adversarial training [1] and input randomization method [2]
- BaRT: Barrage of Random Transforms
- EOT : Expectations over Input Transformations
- **Dataset:** Full ImageNet
- PGD Attack
  - Comparison with respect to adversarial attack strength

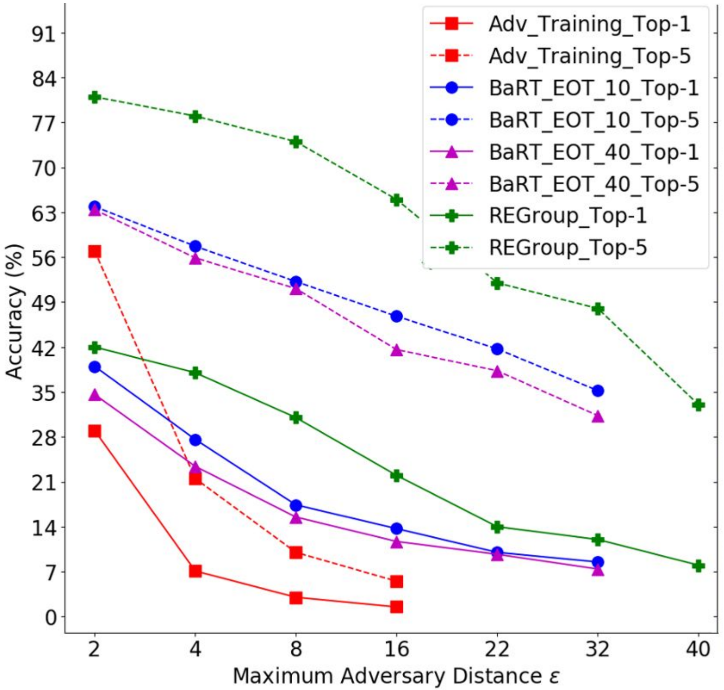


Fig 3. Comparison with adversarial training and fine tuning methods

[1] Kurakin, et al. "Adversarial machine learning at scale." ICLR, 2016
[2] Edward, et al. "Barrage of random transforms for adversarially robust defense." CVPR. 2019.

Thank you