Interpreting Adversarially Trained Convolutional Neural Networks

Tianyuan Zhang, Zhanxing Zhu

ICML 2019.

Outline

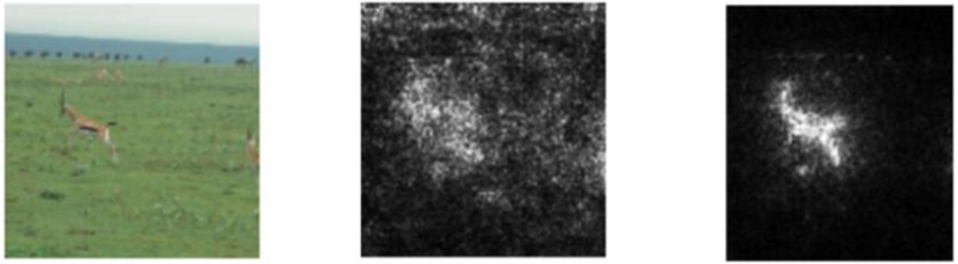
- Model visualization
  - SmoothGrad
- Experiment
  - Visualization results
  - Generalization performance on transformed data
    - Stylizing images
    - Saturation
    - Patch-Shuffing

SmoothGrad

- Averages gradients from Gaussian noisy images to alleviate noises in gradient explanation

$$E = \frac{\partial S_c(x)}{\partial x}.$$

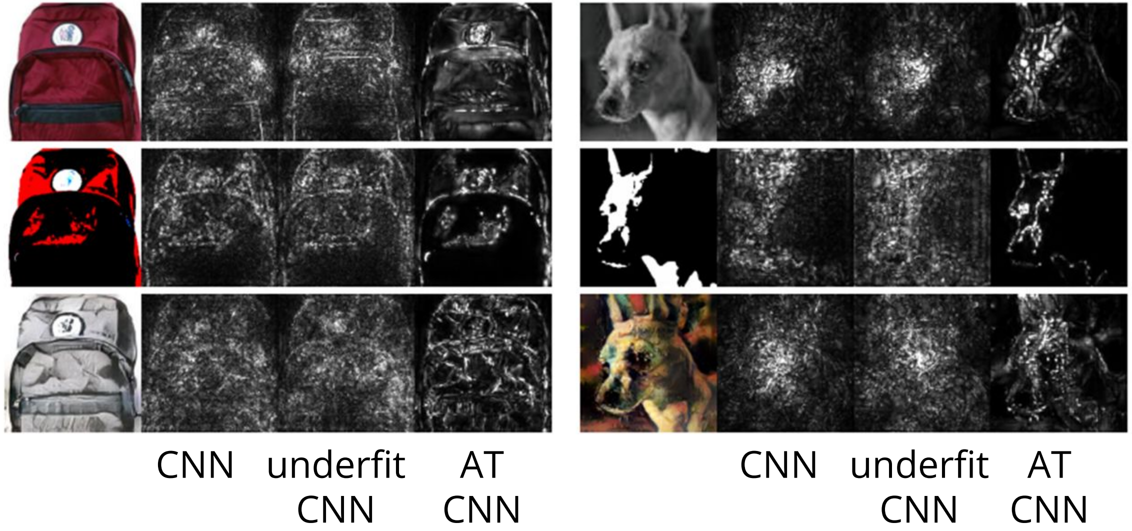
$$E = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial S_c(x_i)}{\partial x_i}, \quad x_i = x + g_i,$$



Gazelle (瞪羚) Typical SmoothGrad

Visualization results

- Generate Sensitive map through SmoothGrad method
- AT-CNN models successfully capture the shape information of the object, providing a more interpretable prediction



Generalization performance: Stylizing images

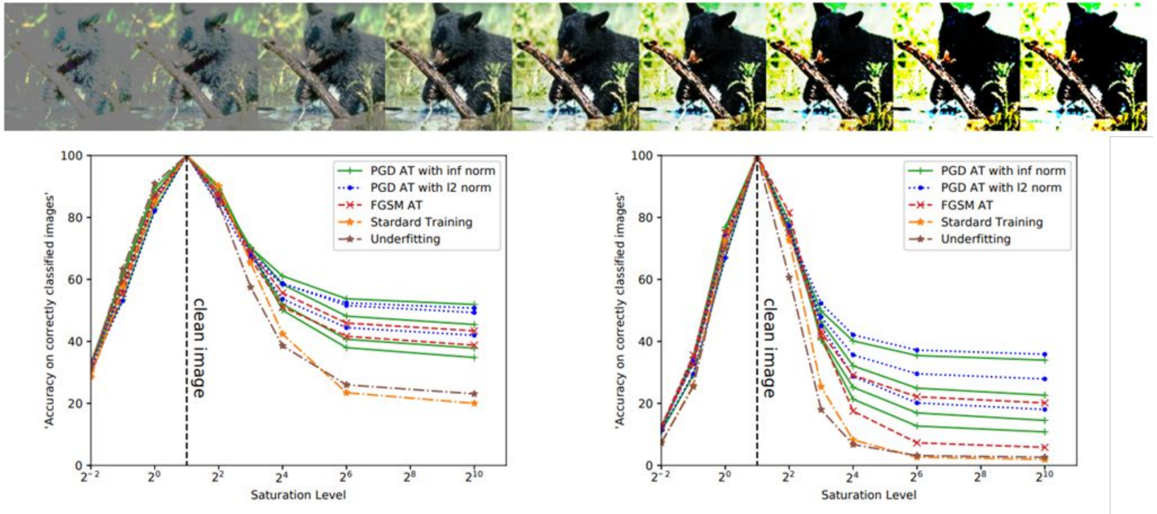
- Test model accuracy on stylized images through style transfer network
- AT-CNNs achieve higher accuracy on stylized ones with textures being dramatically changed.

w/o style　　with style

DATASET	CALTECH-256	STYLIZED CALTECH-256	TINYIMAGENET	STYLIZED TINYIMAGENET
STANDARD	**83.32**	16.83	**72.02**	7.25
UNDERFIT	69.04	9.75	60.35	7.16
PGD-l_∞: 8	66.41	19.75	54.42	18.81
PGD-l_∞: 4	72.22	21.10	61.85	20.51
PGD-l_∞: 2	76.51	21.89	67.06	19.25
PGD-l_∞: 1	79.11	22.07	69.42	18.31
PGD-l_2: 12	65.24	20.14	53.44	19.33
PGD-l_2: 8	69.75	21.62	58.21	20.42
PGD-l_2: 4	74.12	**22.53**	64.24	**21.05**
FGSM: 8	70.88	21.23	66.21	15.07
FGSM: 4	73.91	21.99	63.43	20.22

Generalization performance: Saturation

- Test model accuracy on saturated images
- AT-CNNs performs better on saturated images, reveals AT-CNNs are less sensitivity to texture loss.

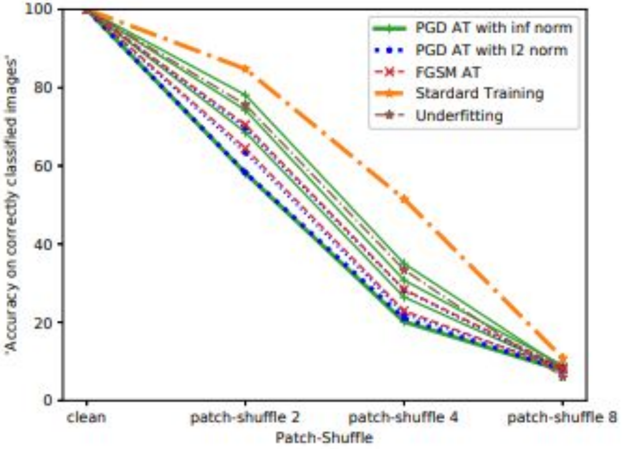


(a) Caltech-256 (b) Tiny ImageNet

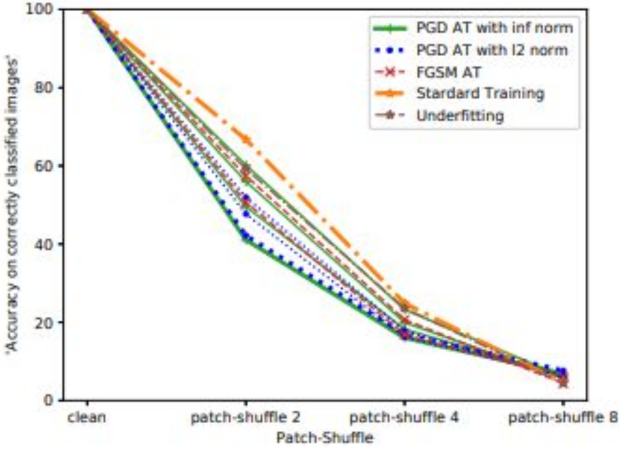
Generalization performance: Patch-shuffing

- Test model accuracy on patch-shuffing images.
- AT-CNNs performs worse on shuffed images,
  reveals AT-CNNs are more baised towards shapes and edges.

(a) Original Image (b) Patch-Shuffle 2 (c) Patch-Shuffle 4 (d) Patch-Shuffle 8



(a) Caltech-256



(b) Tiny ImageNet

Summary

- From both qualitative and quantitative perspectives, The paper implemented a systematic study on interpreting AT-CNNs and normal CNNs.
- AT-CNNs are less sensitive to the texture distortion and focus more on shape information

Bluff: Interavtively Deciphering Adversarial Attacks on Deep Neural Network

Nilaksh Das, Haekyu Park, Zijie J. Wang, Fred Hohman,
Robert Firstman, Emily Rogers, Duen Horng (Polo) Chau
IEEE Visualization Conference 2020.

Background

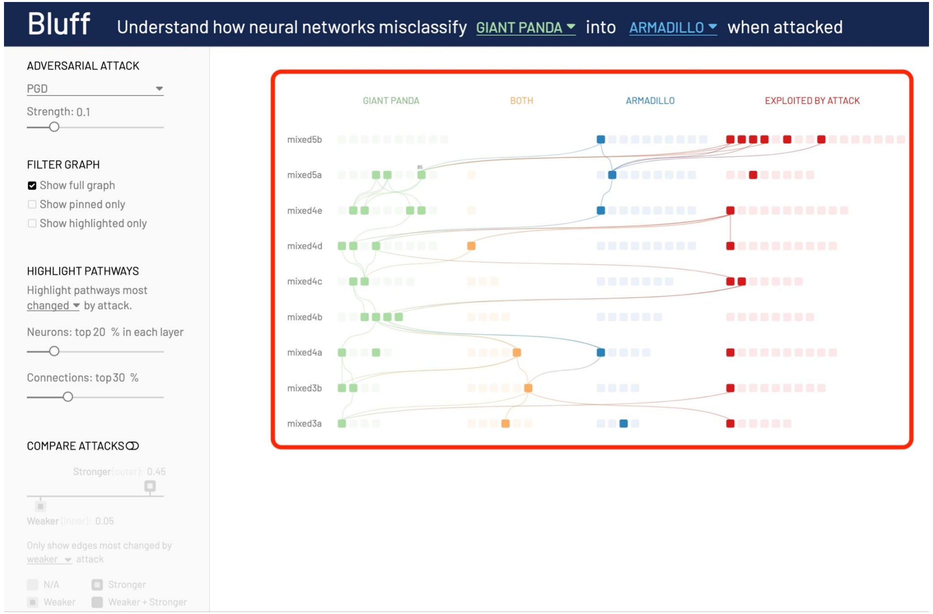
- DNN is vulnerable and complicated
- Want to see that how attack works exploit the model
- To show the connection between attack's strength and neuron's activation patterns

Bluff

- URL:https://poloclub.github.io/bluff/
- INCEPTION V1
- 4 Goals
  - Untangling activation pathways
  - Interpreting multiple pathways
  - Comparing attack characteristics
  - Lower barrier of entry for interpreting and deciphering adversarial attacks

Demo - Overview

- Provides 5 case to test
- Green: Important neurons in original class
- Orange: Important neurons in both classes
- Blue: Important neurons in target class
- Red: Important neurons for adversial image

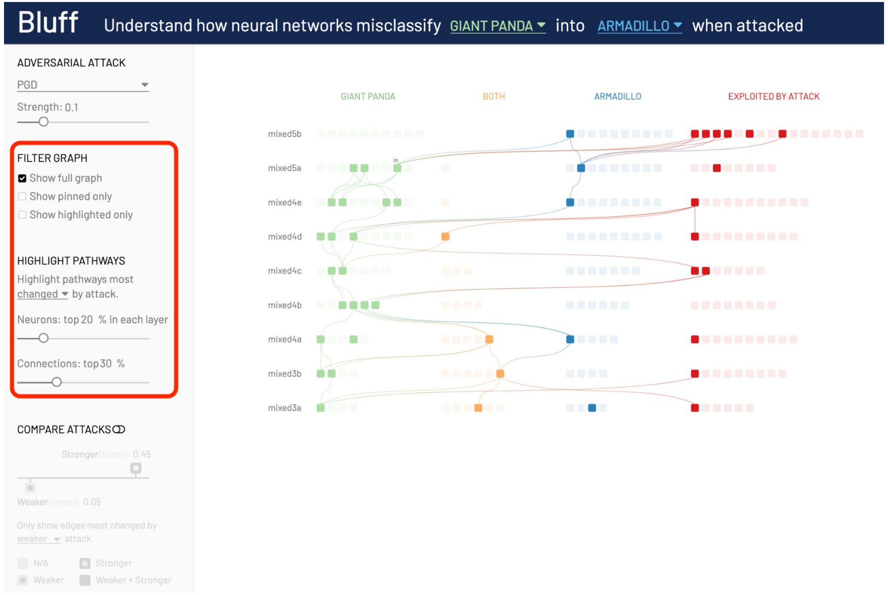


Demo - Attack

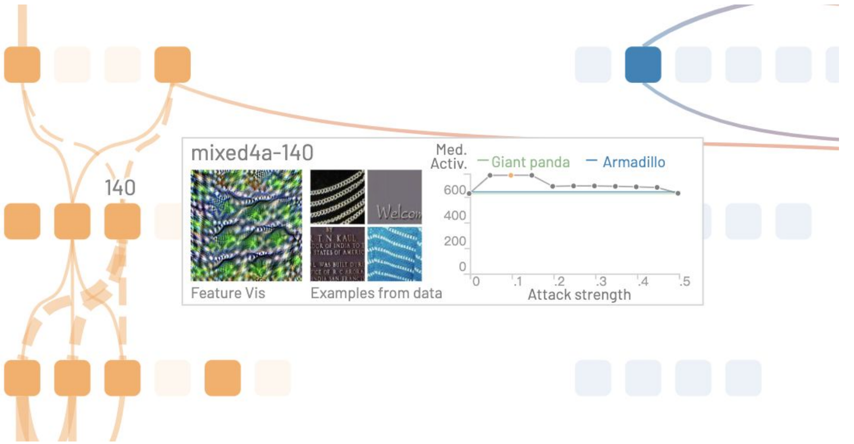
- Method: FGSM/PGD
- Attack strength
- Able to compare two attacks in same method but with different strength

Demo - Graph Viewing

- Choose to show full/pinned/highlighted graph
- Modify the percentage of highlighted neurons/connections
- Shows the neurons that are most activated/changed/excited/inhibited by attack



Demo - In one neuron



- Connection: shows that where the data is from and where to go before and after the nueron
- Pictures: the left one is feature of the neuron and the right one is example dataset
- Chart: the realationship of strength of the attack and the activation pattern to original and target images

Summary

- Bluff shows how the neurons act when an adversarial attack occurs, but it has fixed to the example it provides, so it would be more practical if there is more space for user to select their own original and target images.

Proper Network Interpretability Helps Adversarial Robustness in Classification

Akhilan Boopathy, Sijia Liu, Gaoyuan Zhang, Cynthia Liu, Pin-Yu Chen, Shiyu Chang, Luca Daniel

ICML 2020

Outline

- Model visualization
  - CAM-type method(CAM)
  - pixel sensitivity map(IG)
- Interpretability-Aware Robust Training
- Experiments
  - forseeen attacks
  - unforseen attacks
  - attacks against interpretability

Model Interpretation method - CAM

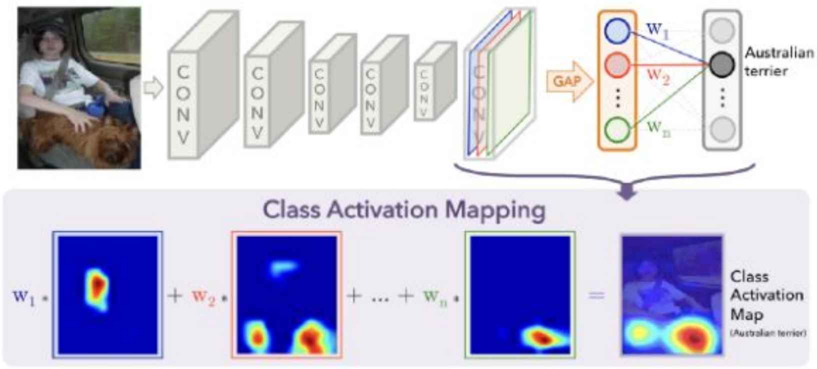
-CAM(class activation map):averages each feature map

$$[I_{\mathrm{CAM}}(\mathbf{x}, c)]_i = (1/u) \sum_{k \in [K]} w_k^c A_{k,i}, \ i \in [u], \qquad (1)$$

-GradCAM:use gradient of the classification score w.r.t feature as weight.

$$w_k^c = \frac{1}{u} \sum_{i=1}^{u} \frac{\partial f_c(\mathbf{x})}{\partial A_{k,i}}.$$

$\mathbf{A}_k \in \mathbb{R}^u$ for channel $k \in [K]$. w_k^c



Model Interpretation method - IG

- IG(integrted gradient):

  similar as SmoothGrad, change gaussian noised images with interpolations between input x and a baseline image a.(usually pick zero vector as a)

$$E = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial S_c(x_i)}{\partial x_i},\ x_i = x + g_i,$$

SmoothGrad

$$[I_{\mathrm{IG}}(\mathbf{x}, c)]_i = (x_i - a_i)\sum_{i=1}^{m}\frac{\partial f_c(\mathbf{a} + \frac{i}{m}(\mathbf{x} - \mathbf{a}))}{\partial x_i}\frac{1}{m},\ i \in [d],$$

IG

Intepretability-Aware Robust Training

 adversarial examples that intend to fool a classifier could find it difficult to evade interpretation discrepancy

 interpreter is quite sensitive to input perturbations

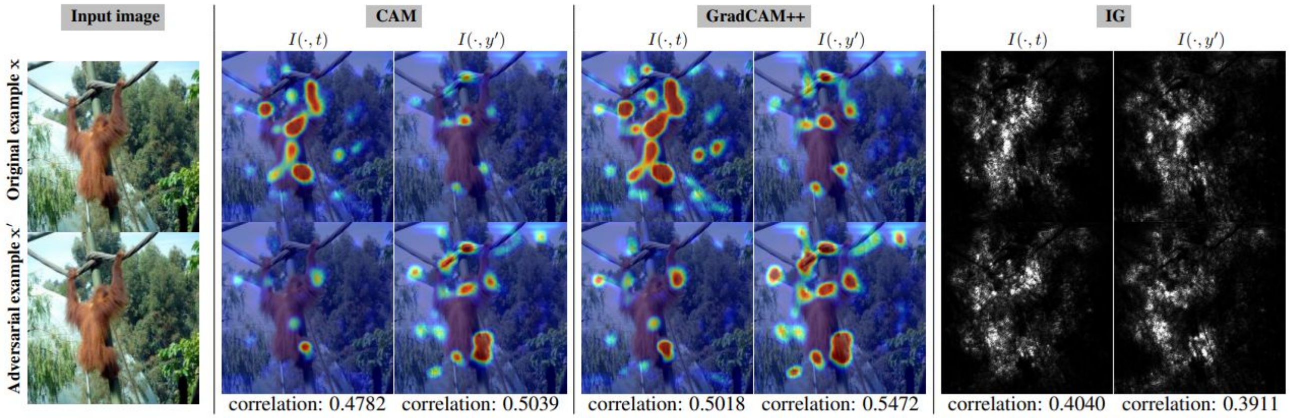


Figure 1. Interpretation (I) of benign ($\mathbf{x}$) and adversarial ($\mathbf{x}'$) image from Restricted ImageNet (Tsipras et al., 2019) with respect to the true label y='monkey' and the target label y'='fish'.

Intepretability-Aware Robust Training

□ adversarial examples that intend to fool a classifier could find it difficult to evade interpretation discrepancy

□ This paper proved that this proposed interpretation discrepancy has a perturbation independent lower bound for any successful adversarial attacks.

Proposition 1. *Given a classifier $f(\mathbf{x}) \in \mathbb{R}^C$ and its interpreter $I(\mathbf{x}, c)$ for $c \in [C]$, suppose that the interpreter satisfies the completeness axiom, namely, $\sum_i [I(\mathbf{x}, c)]_i = f_c(\mathbf{x})$ for a possible scaling factor a.*

interpretation discrepancy classification margin

$$\mathcal{D}_{2,\ell_1}(\mathbf{x}, \mathbf{x}') \geq (1/2)\left(f_y(\mathbf{x}) - f_{y'}(\mathbf{x})\right).$$

$$\mathcal{D}_{2,\ell_1}(\mathbf{x}, \mathbf{x}') = (1/2)\left(\left\|I(\mathbf{x}, y) - I(\mathbf{x}', y)\right\|_1 + \left\|I(\mathbf{x}, y') - I(\mathbf{x}', y')\right\|_1\right).$$

CAM satisfies the completeness axiom

$$[I_{\mathrm{CAM}}(\mathbf{x}, c)]_i = (1/u) \sum_{k \in [K]} w_k^c A_{k,i}, \; i \in [u],$$

Intepretability-Aware Robust Training

- adversial examples are hard to fool interpretation discrepancy, so constraining it helps to prevent misclassification.
- take intepretation discrepancy into training loss

$$\tilde{D}(\mathbf{x}, \mathbf{x}') = (1/2)\left\|I(\mathbf{x}, y) - I(\mathbf{x}', y)\right\|_1 + (1/2)\sum_{i \neq t} \frac{e^{f(\mathbf{x}')_i}}{\sum_{i'} e^{f(\mathbf{x}')_{i'}}} \|I(\mathbf{x}, i) - I(\mathbf{x}', i)\|_1,$$

training loss worst discripancy

$$\underset{\boldsymbol{\theta}}{\text{minimize}}\ \mathbb{E}_{(\mathbf{x},t)\sim\mathcal{D}_{\text{train}}}\left[f_{\text{train}}(\boldsymbol{\theta}; \mathbf{x}, y) + \gamma \tilde{D}_{\text{worst}}(\mathbf{x}, \mathbf{x}')\right],$$

$$\tilde{D}_{\text{worst}}(\mathbf{x}, \mathbf{x}') := \underset{\|\boldsymbol{\delta}\|_\infty \leq \epsilon}{\text{maximize}}\ \tilde{D}(\mathbf{x}, \mathbf{x} + \boldsymbol{\delta}),$$ misinterpretation(Int)

$$\tilde{D}_{\text{worst}}(\mathbf{x}, \mathbf{x}') := \tilde{D}\left(\mathbf{x}, \mathbf{x} + \underset{\|\boldsymbol{\delta}\|_\infty \leq \epsilon}{\arg\max}\left[f_{\text{train}}(\boldsymbol{\theta}; \mathbf{x} + \boldsymbol{\delta}, y)\right]\right)$$ misclassification(Int2)

Experiment:PGD Attacks

normal training

- tested robustness against PGD attacks on intepretability aware robust trained model

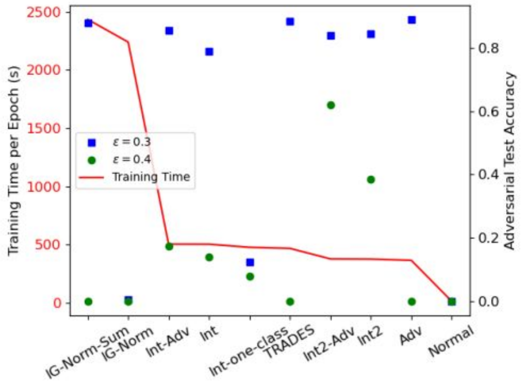


Figure 3. Computation time per epoch and adversarial test accuracy for a `Small` MNIST model trained with different methods.

Method	$\epsilon = 0$	0.05	0.1	0.2	0.3	0.35	0.4
			MNIST, `Small`				
Normal	1.000	0.530	0.045	0.000	0.000	0.000	0.000
Adv	0.980	0.960	0.940	0.925	0.890	0.010	0.000
TRADES	0.970	0.970	0.955	0.930	0.885	0.000	0.000
IG-Norm	0.985	0.950	0.895	0.410	0.005	0.000	0.000
IG-Norm-Sum	0.975	0.955	0.935	0.910	0.880	0.115	0.000
Int-one-class	0.975	0.635	0.330	0.140	0.125	0.115	0.080
Int	0.950	0.930	0.905	0.840	**0.790**	**0.180**	**0.140**
Int-Adv	0.935	0.945	0.905	0.880	**0.855**	**0.355**	**0.175**
Int2	0.950	0.945	0.935	0.890	**0.845**	**0.555**	**0.385**
Int2-Adv	0.955	0.925	0.915	0.880	**0.840**	**0.655**	**0.620**
	$\epsilon = 0$	2/255	4/255	6/255	8/255	9/255	10/255
			CIFAR-10, `WResnet`				
Normal	0.765	0.250	0.070	0.060	0.060	0.060	0.060
Adv	0.720	0.605	0.485	0.330	0.170	0.145	0.085
TRADES	0.765	0.610	0.460	0.295	0.170	0.140	0.100
Int-one-class	0.685	0.505	0.360	0.190	0.065	0.040	0.025
Int	0.735	0.630	0.485	0.365	**0.270**	**0.240**	**0.210**
Int-Adv	0.665	0.585	0.510	0.385	**0.320**	**0.300**	**0.280**
Int2	0.690	0.595	0.465	0.360	**0.290**	**0.245**	**0.220**
Int2-Adv	0.680	0.585	0.485	0.405	**0.335**	**0.310**	**0.285**
			R-ImageNet, `WResnet`				
Normal	0.770	0.070	0.035	0.030	0.040	0.030	0.030
Adv	0.790	0.455	0.230	0.100	0.070	0.060	0.050
Int	0.660	0.570	0.460	0.385	**0.280**	**0.250**	**0.220**
Int2	0.655	0.545	0.480	0.355	**0.265**	**0.205**	**0.170**

Table 1. Evaluation of 200-step PGD accuracy under different perturbation sizes ϵ. ATA with $\epsilon = 0$ reduces to standard test accuracy.

Experiment:unforseen Attacks

- tested robustness against unforseen attacks on intepretability aware robust trained model

Method	Gabor	Snow	JPEG ℓ_∞	JPEG ℓ_2	JPEG ℓ_1
			CIFAR-10, `Small`		
Normal	0.125	0.000	0.000	0.030	0.000
Adv	0.190	0.115	0.460	**0.380**	0.230
TRADES	**0.220**	0.085	0.425	0.300	0.070
IG-Norm	0.155	0.015	0.000	0.000	0.000
IG-Norm-Sum	0.185	0.110	**0.480**	0.375	0.215
Int	0.160	0.105	0.440	0.345	0.260
Int-Adv	0.150	0.120	0.340	0.310	0.235
Int2	0.130	0.115	0.440	0.365	**0.295**
Int2-Adv	0.110	**0.135**	0.360	0.315	0.260

Table 2. ATA on different unforeseen attacks in (Kang et al., 2019). Best results in each column are **highlighted**.

Experiment: Attack Against Interpretability(AAI)

- tested robustness against attack against interpretability on intepretability aware robust trained model

unchanged prediction high discripancy

$$\begin{aligned} \underset{\boldsymbol{\delta}}{\text{minimize}} \quad & \lambda \max\{\max_{j\neq y} f_j(\mathbf{x}+\boldsymbol{\delta}) - f_y(\mathbf{x}+\boldsymbol{\delta}), 0\} - \mathcal{D}_1(\mathbf{x}, \mathbf{x}+\boldsymbol{\delta}) \\ \text{subject to} \quad & \|\boldsymbol{\delta}\|_\infty \leq \epsilon, \end{aligned}$$

Method	$\epsilon = 0.05$	0.1	0.2	0.3	0.35	0.4
MNIST, `Small`						
Normal	0.907	0.797	0.366	-0.085	-0.085	-0.085
Adv	0.978	0.955	0.910	0.857	0.467	0.136
TRADES	0.978	0.955	0.905	0.847	0.450	0.115
IG-Norm	0.958	0.894	0.662	0.278	0.098	0.094
IG-Norm-Sum	0.976	0.951	0.901	0.850	0.659	0.389
Int-one-class	0.874	0.818	0.754	0.692	0.461	0.278
Int	0.982	0.968	0.941	**0.913**	**0.504**	**0.320**
Int-Adv	0.980	0.965	0.936	**0.912**	**0.527**	**0.348**
Int2	0.982	0.967	0.941	**0.918**	**0.612**	**0.351**
Int2-Adv	0.982	0.971	0.950	**0.931**	**0.709**	**0.503**
	$\epsilon = 2/255$	4/255	6/255	8/255	9/255	10/255
R-ImageNet, `WResnet`						
Normal	0.851	0.761	0.705	0.673	0.659	0.619
Adv	0.975	0.947	0.916	0.884	0.870	0.858
Int	0.988	0.974	0.960	**0.946**	**0.939**	**0.932**
Int2	0.989	0.977	0.965	**0.952**	**0.946**	**0.939**

Table 3. Performance of AAI for different values of perturbation size ϵ in terms of Kendall's Tau order rank correlation between the original and adversarial interpretation maps. High interpretation robustness corresponds to large correlation value.

Summary

This paper theoretically and empirically that it is difficult to hide adversarial examples from interpretation.

This paper develops a interpretability-aware robust training method that displays both high classification robustness and high robustness of interpretation.