How are Saliency Maps Related to Adversarial Perturbations? *

Puneet Mangla (cs17btech11029)
Advisor: Vineeth N Balasubramanian

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Under Review at A* Computer Vision Conference

Abstract

In this work, we provide a different perspective to understand the relationship between visual explanations and adversarial robustness, wherein we seek to ask if there is a plausible way in which we can use a given saliency map of the ground truth class of an image to improve upon a model's robustness.

Recent studies

- Solely focuses on interpretations or robustness.
- Recent, attempted to couple both notions.
- ATCNNs exhibit more interpretable saliency maps than their non-robust counterparts
- Behavior can be quantified mathematically.

Key Contributions

An efficient methodology that uses the saliency maps to mimic adversarial training.

Key contributions

- Improved robustness by using visual explanation methods to obtain saliency maps using 'pseudo-adversarial training'.
- More pronounced when a finer and more interpretable saliency map is used instead.
- Ensemble of saliency maps with adversarial perturbations improves PGD and TRADES AT, taking just half the training time.
- Bounding boxes can be exploited in the same manner.

Notations and Preliminaries

Neural network as $\Phi(.\,;\,\theta) : \mathbb{R}^d \to \mathbb{R}^k$, parametrized by weights θ, $\mathbf{x} \in \mathbb{R}^d$ and outputs a logit, $\Phi^i(\mathbf{x})$.

1. Saliency map as $\mathbf{s} \in [0,1]^d$, where the presence of an object of interest in input $\mathbf{x}$ lies between 0 and 1. Unnormalized saliency map for trained network Φ: $\nabla_{\mathbf{x}}\Phi^{i^*}(\mathbf{x})$, where $i^* = \arg\max_i \Phi^i(\mathbf{x})$.

2. PGD Attack: A more powerful multi-step attack : $\mathbf{x}^0 = \mathbf{x}$,

$$\mathbf{x}^{t+1} = \Pi_{\mathbf{x}+N}\big(\mathbf{x}^t + \alpha\ \mathrm{sign}(\nabla_{\mathbf{x}}\mathcal{L}\left(\Phi(\mathbf{x},\theta), y\right))\big)$$

3. Adversarial Training: make the models robust by matching the training distribution with the adversarial test distribution. Essentially, for AT, the optimal parameter θ^* is given by:

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{(\mathbf{x},y)\sim D}\left[\max_{\delta \in N} \mathcal{L}\left(\Phi(\mathbf{x}+\delta,\theta), y\right)\right]$$

4. Linearized robustness: robustness $\rho(\mathbf{x})$ of a network Φ at image $\mathbf{x}$ as: $\rho(\mathbf{x}) = \min_{j \neq i^*} \frac{\Phi^{i^*}(\mathbf{x}) - \Phi^j(\mathbf{x})}{\|\nabla_{\mathbf{x}}(\Phi^{i^*}(\mathbf{x}) - \Phi^j(\mathbf{x}))\|}$

Our Method: Saliency-Based Adversarial Training (SAT).

Etmann formulate the direction of adversarial perturbation as :

$$\nabla_{\mathbf{x}}(\Phi^{j^*}(\mathbf{x}) - \Phi^{i^*}(\mathbf{x}))$$

Perturbation depends on two quantities: (i) $\nabla_{\mathbf{x}}\Phi^{i^*}(\mathbf{x})$, the saliency map for the true class i^*; and (ii) $\nabla_{\mathbf{x}}\Phi^{j^*}(\mathbf{x})$, the saliency map of $\mathbf{x}$ for class j^*, closest to the in terms of decision boundary. For a binary classifier, one can derive

$$\Phi^{j^*}(\mathbf{x}) = -\Phi^{i^*}(\mathbf{x})$$

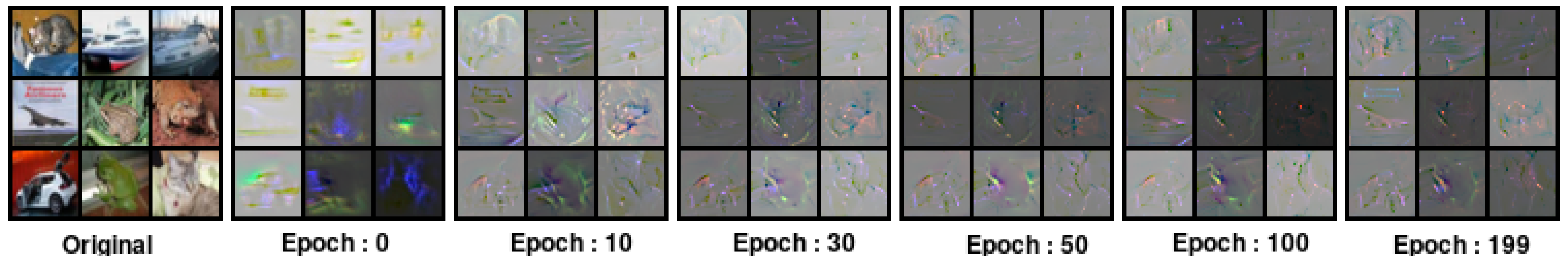
and define the adversarial perturbation direction simply as $-\nabla_{\mathbf{x}}(\Phi^{i^*}(\mathbf{x}))$.

A multi-class classifier can be reasoned as a binary-classifier in one-vs-all setting and in $-\nabla_{\mathbf{x}}\Phi^{i^*}(\mathbf{x})$ positive values corresponds to presence of all classes except i^* and negative to presence of true class i^*.

In adversarial training, during the initial phase of training the perturbations computed by the attack methods are random. But with training, as weights become optimal, they become more class-discriminative (See Figure). In order to mimic the above behavior of the perturbation over training, we choose the i^{th} component $\delta^t[i]$ of perturbation δ^t at time t as:

$$\delta^t[i] = \begin{cases} \mathbf{z}[i], & \text{with probability } \alpha^t \\ -\mathbf{s}[i], & \text{with probability } 1-\alpha^t \end{cases}$$

where $\mathbf{z}$ is a random vector of 1s and -1s and $0 < \alpha < 1$. Thus at time t, input $\mathbf{x}$ is perturbed as : $x = x + \epsilon_0 \cdot \delta^t$, where ϵ_0 is l_∞ max perturbation.



Conclusion and Future Work

Our work opens rather a new direction to enhance robustness of DNNs by exploiting saliencies. Future work in this direction may aim to improve SAT by inferring about class closest to true one in terms of decision boundary and approximating the directions better. The work also highlights the need for fine and interpretable saliency map annotations to be provided as part of computer vision datasets.

Experiments and Results

We perform the following experiments. Detailed experiments can be seen in writeup. GBP: Guided-Backpropagation; G.CAM++: Grad-CAM++.

Method	$\epsilon = \frac{1}{255}$	$\epsilon = \frac{2}{255}$	$\epsilon = \frac{3}{255}$	$\epsilon = \frac{4}{255}$
Original	25.83	7.76	3.35	1.94
Original + Uniform-Noise	33.15	13.50	6.01	3.22
SAT (Weak saliency)				
Resnet-10 ∣ Std. ∣ G.CAM++	31.98	11.93	5.48	2.89
Resnet-10 ∣ Adv. ∣ G.CAM++	32.88	13.3	6.31	4.0
Resnet-10 ∣ Adv. ∣ G.CAM++	32.96	12.07	5.2	3.04
SAT (Fine saliency)				
Resnet-10 ∣ Std. ∣ GBP	20.53	7.52	3.5	2.12
Resnet-10 ∣ Adv. ∣ GBP	**34.29**	**14.73**	**6.84**	**4.22**
PGD	45.75	40.13	35.41	31.01
PGD + Uniform-Noise	**50.0**	**42.67**	36.10	30.64
PGD-SAT				
Resnet-10 ∣ Std. ∣ GBP	46.87	41.11	35.77	30.75
Resnet-10 ∣ Adv. ∣ GBP	48.33	42.2	**36.33**	**31.66**
Resnet-10 ∣ Adv. ∣ G.CAM++	47.28	41.72	35.99	31.05
TRADES	47.21	42.2	37.03	32.96
TRADES + Uniform-Noise	**51.90**	42.85	37.30	31.71
TRADES-SAT				
Resnet-10 ∣ Std. ∣ GBP	48.63	42.91	37.79	33.19
Resnet-10 ∣ Adv. ∣ GBP	48.74	42.99	**37.83**	**33.23**
Resnet-10 ∣ Adv. ∣ G.CAM++	48.99	**43.05**	37.4	32.84

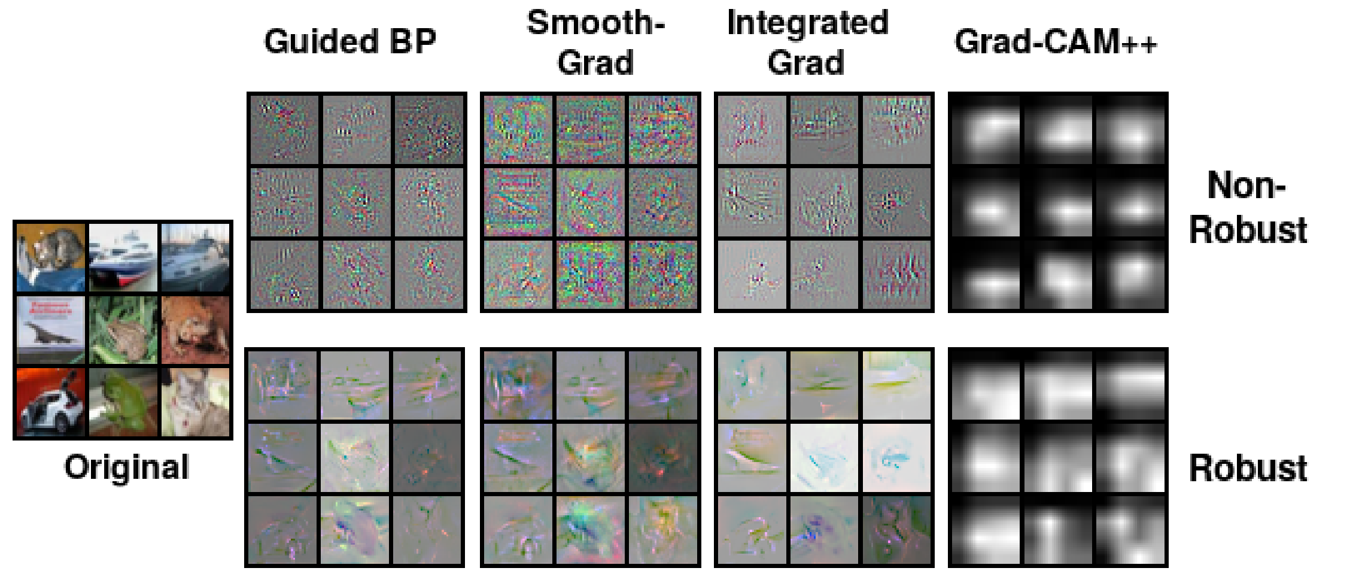
Results of improved robustness on CIFAR-100 dataset.

Method	Time (s)	
	CIFAR-10	CIFAR-100
Original	15.25	14.95
SAT		
Resnet-10 ∣ Std. ∣ GBP	18.25	18.11
Resnet-10 ∣ Adv. ∣ GBP	18.18	18.26
PGD	159.97	163.85
PGD-SAT		
Resnet-10 ∣ Std. ∣ GBP	85.15	92.83
Resnet-10 ∣ Adv. ∣ GBP	85.08	93.99

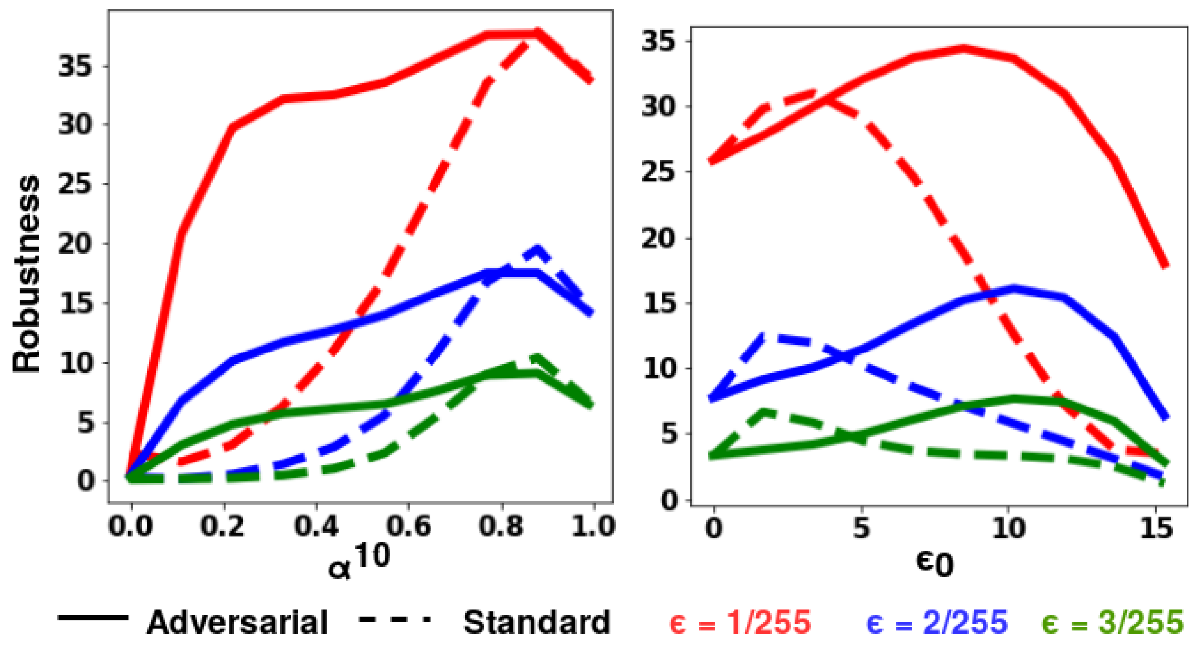
Training time for one epoch in seconds averaged over 10 trials.

Method	$\epsilon = 1\frac{}{255}$	$\epsilon = \frac{2}{255}$	$\epsilon = \frac{3}{255}$
Original	1.04	0.4	0.0
Original + Uniform-Noise	9.45	2.32	0.77
SAT (Weak Saliency)			
Resnet-10 ∣ Std. ∣ G.CAM++	9.35	1.95	0.74
Resnet-10 ∣ Adv. ∣ G.CAM++	9.68	2.30	0.80
SAT (Fine Saliency)			
Resnet-10 ∣ Std. ∣ GBP	3.02	0.0	0.0
Resnet-10 ∣ Adv. ∣ GBP	**10.69**	**2.81**	**1.27**
Bounding Boxes	9.79	2.46	0.77

Results of improved robustness on Tiny-Imagenet dataset.



Saliency maps of robust and non-robust Resnet-10 with various explanation methods on CIFAR-10.



Variation of hyper-parameter α and ϵ_0 on CIFAR-100 dataset.