On Saliency Maps and Adversarial Robustness

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Puneet Mangla, Vedant Singh, and Vineeth Balasubramanian

Department of Computer Science and Engineering
IIT Hyderabad, India

ECML PKDD 2020

Overview

- ***Robustness and Interpretability*** are important parameters for DNNs

- ***Early works:*** focusing solely on either robustness or interpretability

- **Recent works:** started exploring relation between these notions

    - Robust DNNs exhibit high interpretability

    - DNNs with robust explanations are inherently robust

- ***Our contributions:*** explore new tangible relationship between a saliency maps and adversarial perturbations

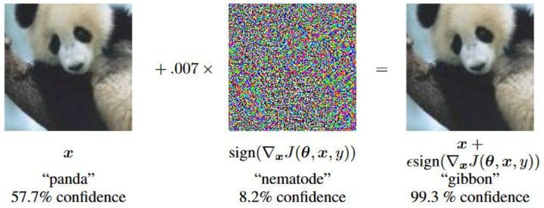
    - propose a new method (SAT) that uses the saliency map while training to improve networks robustness.

    - Experimented on widely used datasets

    - Show that the improvement becomes more pronounced when a better saliency map is used

    - Exploit bounding boxes or segmentation masks as weak saliency to efficiently improve model's robustness

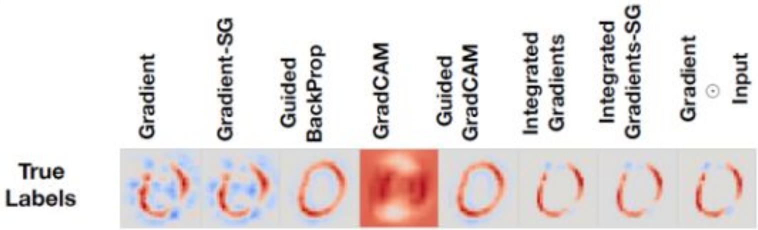
Adversarial Robustness



- **Adversarial Attacks:**
  - Adding imperceptible perturbations to input leading to wrong model predictions.
  - E.g FGSM (Goodfellow et al. 2015), PGD (Madry et al 2018.), stAdv (Xiao et al. 2018)

- **Adversarial Training:**
  - Make models robust by augmenting training data with adversarial perturbations.
  - Popular ones: PGD-AT (Madry et al. 2018) and TRADES (Zhang et al. 2019)

Interpretability



- **Backpropagation based:** importance of each pixel by backpropagating the class score error to the input image

  - E.g Guided Backprop (Springenberg et al. 2015), SmoothGrad (Smilkov et al 2017.), Integrated Gradients (Sundararajan et al. 2017)

- **Activation Based:** use linear combinations of activations of convolutional layers

  - E.g CAM (Zhou et al. 2016), GradCAM (Selvaraju et al 2017.), GradCAM++ (Chattopadhyay et al. 2018)

Coupling Robustness and Interpretability

- **Zhang et al. 2018:** Robust models are more biased towards image shape than its texture and evince more interpretable saliency maps

- **Etmann et al. 2019:** quantified above behavior of robust models by considering the alignment between saliency map and the image as the metric for interpretability

- **Dombrowski et al. 2019, Ghorbani et al. 2019 :** Do robust and interpretable saliency maps imply adversarial robustness ?

Our Work- Can saliency maps be used to induce robustness ?

- **Motivation:** humans tend to learn new tasks in a robust fashion when provided with explanations during their learning phase
  - Eg. a medical student

- **Our hypothesis:** a DNN model that is trained with explanations is less easily fooled by adversarial perturbations.

Saliency Based Adversarial Training : Motivation

- An adversarial perturbation,**e**, which is intended as a perturbation to input **x** which results in a change of predicted label, can be modeled as follows:

$$\arg\max_i \Phi^i(x+e) \neq \arg\max_i \Phi^i(x) \tag{5}$$

$$\iff \exists j \neq i^* : \Phi^j(\mathbf{x}+e) > \Phi^{i^*}(\mathbf{x}+e) \tag{6}$$

$$\iff \exists j \neq i^* : e^T \cdot (\nabla_{\mathbf{x}}\Phi^j(\mathbf{x}) - \nabla_{\mathbf{x}}\Phi^{i^*}(\mathbf{x})) > \Phi^{i^*}(\mathbf{x}) - \Phi^j(\mathbf{x}) \tag{7}$$

Adversarial perturbation definition

$$\arg \inf_{e \in \mathbb{R}^d} \{\|e\| : \arg\max_i \Phi^i(x+e) \neq \arg\max_i \Phi^i(x)\}$$

Etmann et al. 2019

$$\Phi^i(\mathbf{x}+e) \approx \Phi^i(\mathbf{x}) + e^T \cdot \nabla_{\mathbf{x}}\Phi^i(\mathbf{x})$$

Saliency Based Adversarial Training : Motivation

$$\exists j \neq i^* : e^T \cdot (\nabla_{\mathbf{x}} \Phi^j(\mathbf{x}) - \nabla_{\mathbf{x}} \Phi^{i^*}(\mathbf{x})) > \Phi^{i^*}(\mathbf{x}) - \Phi^j(\mathbf{x})$$

The infimum over **llell**, which provides a minimal perturbation to change the class label, is achieved by choosing **e** as a multiple of $\nabla_x(\Phi^j(x) - \Phi^{i*}(x))$.

- The direction of adversarial perturbation then becomes

$$\nabla_x(\Phi^j(x) - \Phi^{i*}(x)).$$

- This perturbation direction depends on two quantities:

(i) $\nabla x \Phi^{i*}(x)$ the saliency map for the true class **i***

(ii) $\nabla x \Phi^j(x)$, the saliency map of **x** for class **j** for which the infimum of **e** is attained.

Case of Binary Classifiers

- A binary classifier **h : x→ {−1,1}** given by: **h=sign(Φ(x,θ))**, where **Φ(x,θ))** represents the logit of the positive class.
- Let **Φ'(x)** denotes the logit of negative class, then

$$P(y=+1|\mathbf{x}) = \frac{1}{1+\exp^{-\Phi(\mathbf{x},\theta)}} \qquad P(y=-1|\mathbf{x}) = \frac{1}{1+\exp^{-\Phi'(\mathbf{x},\theta)}}$$

$$P(y=-1|\mathbf{x}) = 1 - P(y=+1|\mathbf{x}) = \frac{1}{1+\exp^{\Phi(\mathbf{x},\theta)}}$$

- Thus, logit score of the negative class is −Φ(x,θ)
- So, the direction of adversarial perturbation $\nabla_x(\Phi^j(x)-\Phi^{i*}(x))$ becomes $-\nabla x\Phi^{i*}(x)$
- **The negative of saliency map, gives us the direction of adversarial perturbation in case of binary classifier**

Case of Multi-class Classifiers

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

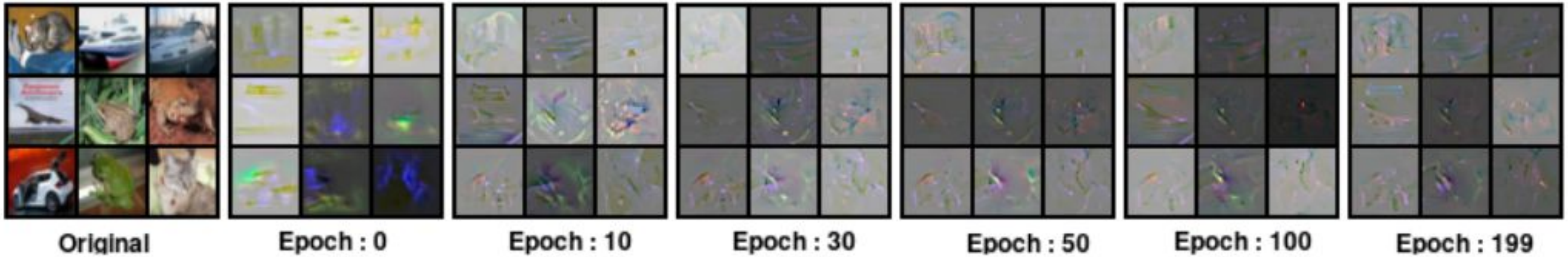
- The multi-class case would require finding the class **j** for which the infimum of **||e||** is attained.
- To avoid this computational overhead, we rely on $\nabla_x \Phi^{i*}(x)$ alone, and simply propose the use of $-\nabla_x \Phi^{i*}(x)$ as the direction of perturbation.

This is a reasonable approximation. Why?

- Consider the multi-class setting as k binary classification problems
- For each individual problem, corresponding logit score of the negative class is $-\nabla_x \Phi^{i*}(x)$
- Assuming that each of the classes **!= i*** is equally likely to be the **j** that minimizes **||e||**
- Approximate average direction of the perturbations across the k binary classification problems.

Saliency based Adversarial Training: Algorithm

- $-\nabla_x\Phi^{i*}(x)$ is given to us in form of saliency map, **s.**
- **We don't have any intermediate perturbations.**
- While adversarial training, during the initial phases, the perturbations computed by the attack methods are random. But with training, they become more class-discriminative.



Saliency based Adversarial Training: Algorithm

To generate intermediate perturbations, we mimic the above observation

- We choose the **ith** component **δ^t[i]** of perturbation as
- Note that saliency map, **s,** can be converted in range {-1,1} by using thresholds.

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

where $\mathbf{z} \in \{-1, 1\}^d$ is sampled randomly, and $0 < \alpha < 1$. D

Leveraging bounding- boxes and segmentation masks

- When additional annotations such as bounding boxes or segmentation masks are available in a dataset,our approach considers these as weak saliency maps for the methodology.

We generate the weak saliency from bounding boxes or segmentation masks as:

$$\tilde{\mathbf{s}}[i] = \begin{cases} 1, & \text{if } \mathrm{i}^{\mathrm{th}} \text{ pixel lies inside bbox or seg masks} \\ -1, & \text{otherwise} \end{cases}$$

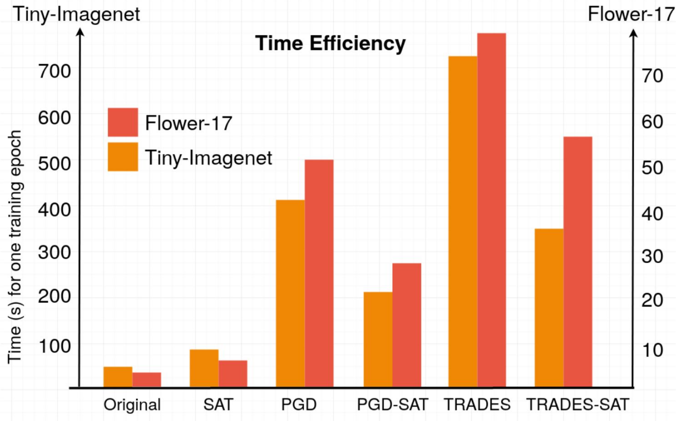
Robustness Results on Tiny-Imagenet and Flower-17

Method	Tiny-Imagenet			FLOWER-17		
	$\epsilon = 1/255$	$\epsilon = 2/255$	$\epsilon = 3/255$	$\epsilon = 1/255$	$\epsilon = 2/255$	$\epsilon = 3/255$
Original	1.04	0.4	0.0	63.2	48.01	34.2
Original + Uniform-Noise	9.45	2.32	0.77	64.56	50.43	36.2
SAT	**9.79**	**2.46**	**0.77**	**66.17**	**52.94**	**38.93**
PGD	18.91	14.34	11.37	72.38	70.4	70.3
PGD + Uniform-Noise	19.57	15.49	11.66	73.52	72.79	72.71
PGD-SAT	**20.56**	**16.38**	**12.91**	**78.67**	**75.73**	**75.00**
TRADES	18.45	16.76	11.09	74.56	73.89	73.67
TRADES + Uniform-Noise	19.96	16.13	12.58	76.47	74.26	74.0
TRADES-SAT	**20.04**	**16.45**	**12.96**	**79.41**	**77.94**	**77.20**

The above table shows results on Tiny-Imagenet and Flower-17 datasets where the given bounding boxes and segmentation masks are used as saliency maps. The value of ϵ denotes the maximum l_∞ perturbation allowed in 5-step PGD attack. (More the ϵ, stronger the attack)

Time efficiency of training procedure

- PGD-SAT and TRADES-SAT require much less training time when compared to their vanilla counterparts, while achieving superior performance at the same time.
- In case of vanilla SAT, we observe an increase in robustness without much increase in training time.



Using better saliency maps for training (Cifar-100)

- The saliency maps used in this study were SmoothGrad, Guided Grad-CAM++ and Integrated Gradients.
- **Better explanations improves performance of our trained models.**

Notations **X-Y-Z** (Resnet10-Std.-GBP)
X: Model architecture (Resnet10)
Y: Training procedure (Std.)
Z: Saliency method (GBP)

Method	PGD $\frac{1}{255}$	PGD $\frac{2}{255}$	PGD $\frac{3}{255}$	PGD $\frac{4}{255}$
Original	25.83	7.76	3.35	1.94
Original + Uniform-Noise	33.15	13.50	6.01	3.22
SAT				
Resnet-10 — Std. — GBP	20.53	7.52	3.5	2.12
Resnet-10 — Std. — S.Grad	**39.22**	**19.89**	**9.44**	**4.49**
Resnet-10 — Std. — G.G.CAM++	21.46	8.00	3.53	2.23
Resnet-10 — Std. — I.Grad	36.2	5.43	7.28	3.37
Resnet-10 — Adv. — GBP	34.29	14.73	6.84	4.22
Resnet-10 — Adv. — S.Grad	**40.01**	**21.2**	**10.96**	**4.85**
Resnet-10 — Adv. — G.G.CAM++	34.07	13.18	5.85	3.09
Resnet-10 — Adv. — I.Grad	37.56	16.45	7.55	4.31

Acknowledgements

We are grateful to:

सत्यमेव जयते

Government of India
Ministry of Human Resource
Development

सत्यमेव जयते

Department of Sciences
& Technology
Government of India

Honeywell
THE POWER OF CONNECTED

ArXiv

GitHub

References

- Goodfellow, I., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: ICLR’15
- Madry, A., Makelov, A.,Schmidt,L ., Tsipras, D., Vladu, A.:Towards deep learning models resistant to adversarial attacks. In: ICLR’18
- Xiao, C., Zhu, J.Y., Li, B., He, W., Liu, M., Song, D.: Spatially transformed adversarial examples. In: ICLR’18
- Zhang, H., Yu, Y., Jiao, J., Xing, E.P., Ghaoui, L.E., Jordan, M.I.: Theoretically principled trade-off between robustness and accuracy. In: ICML’19
- Springenberg, J., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. In: ICLR (workshop track) (2015)
- Smilkov, D., Thorat, N., Kim, B., Viégas, F.B., Wattenberg, M.: SmoothGrad : removing noise by adding noise. CoRR (2017)
- Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: ICML’17
- Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR’16
- Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad-cam: Why did you say that? visual explanations from deep networks via gradient- based localization. In: ICCV’17
- Chattopadhyay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad- cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: WACV’18
- Zhang, T., Zhu, Z.: Interpreting adversarially trained convolutional neural networks. In: ICML’18
- Etmann, C., Lunz, S., Maass, P., Schönlieb, C.B.: On the connection between adversarial robustness and saliency map interpretability. In: ICML’19
- Dombrowski, A.K., Alber, M., Anders, C.J., Ackermann, M., Müller, K.R., Kessel, P.: Explanations can be manipulated and geometry is to blame. In: NeurIPS’19
- Ghorbani, A., Abid, A., Zou, J.: Interpretation of neural networks is fragile. In: AAAI’19