Generating Deep Networks Explanations with Robust Attribution Alignment (Supplementary Material)

Guohang Zeng GUOHANGZ@STUDENT.UNIMELB.EDU.AU
Yousef Kowsar KOWSAR.YOUSEF@UNIMELB.EDU.AU
Sarah Erfani SARAH.ERFANI@UNIMELB.EDU.AU
James Bailey BAILEYJ@UNIMELB.EDU.AU

School of Computing and Information Systems
The University of Melbourne, Australia

A. Impact of other adversarial defense approaches

In the submitted paper, we used PGD-ℓ_2 trained model to generate robust attribution maps, then leveraged the saliency maps of PGD adversarially trained models to align the generated attribution maps. The benefits of robustly trained models have been studies by several previous research. For example, (Tsipras et al., 2018) showed that the saliency maps of adversarially robust models align well with human perception. (Zhang and Zhu, 2019) demonstrated that the representation learned by robust models are more biased towards shape-based features and alleviates texture-based features. (Margeloiu et al., 2020) showed that adversarially trained CNNs are more interpretable for medical imaging diagnosis.

Since previous works all investigated the influence of PGD adversarial training on the interpretable of CNNs, it raises the question whether the phenomenon can be generalized to other adversarial training approaches. In this section, we compare PGD training with other two adversarial defense approaches: Defensive distillation and Jacobian regularization, and see whether other adversarial defense approaches can also be used to generate faithful explanations.

A1. Defensive Distillation

(Papernot et al., 2016) proposed defensive distillation, which leveraged the notion of Knowledge Distillation (Hinton et al., 2015) to make deep neural networks robust against adversarial attacks. (Papernot et al., 2016) first trained a teacher model by standard training manner, then scaled the output logits of the teacher model:

$$f^i_{scaled}(x) = \frac{e^{z_i(x)/T}}{\sum_{j=1}^{K} e^{Z_j(x)/T}}$$

where $f^i_{scaled}(x)$ represents the scaled logits for sample x, $z_i(x)$ denotes the output of teacher model, and the parameters T is a scale factor to control the mean magnitudes of $f^i_{scaled}(x)$. Instead of minimizing the cross-entropy loss between model's output and the ground true one-hot label, Defensive distillation uses the scaled logits derived from teacher model to supervise the student model. The target outputs of student model is defined as:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[L(\theta, x, f_{scaled}(x)) \right]$$

where L denotes the loss function of the model and $f_{scaled}(x)$ denotes the scaled logits. Noticed that the scaled logits have the right target label but with smaller mean input gradients. The motivation of defensive distillation is to make mean input gradients of the model to be small in order to defend adversarial attacks, and the mean input gradients is controlled by the hyperparameter T. (Papernot et al., 2016) showed that doing so improves resilience of a network to small perturbation in the images. However, it is worth noting that Defensive distillation is not a valid defense approach under many adversarial attacks such as Carlini and Wagner attacks (Carlini and Wagner, 2017).

A2. Jacobian Regularization (Gradient masking)

(Ross and Doshi-Velez, 2018) studied input gradient regularization(also known as Jacobian Regularization or Gradient Masking) as a method for adversarial robustness. Jacobian regularization jointly optimize the model by integrating Jacobian regularizer into training:

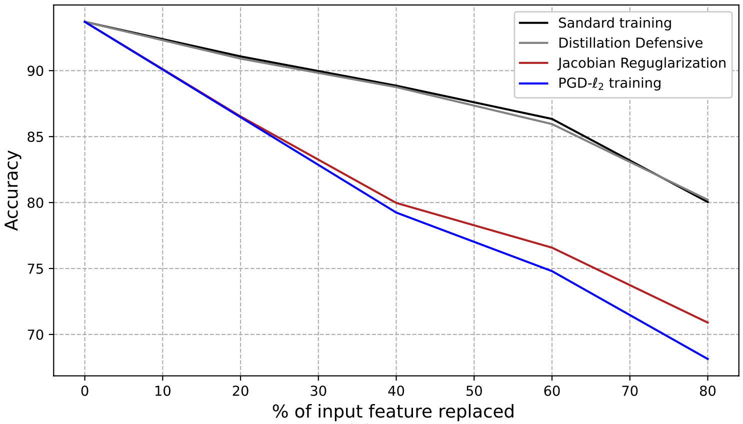
$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[L(\theta, x, y) + \lambda \left\| \nabla_x L\left(f_\theta(x), y\right) \right\|_2 \right]$$

where $\left\| \nabla_x L\left(f_\theta(x), y\right) \right\|_2$ denotes the square of the Forbenius norm of the input-output Jacobian, and λ is a hyper parameter that determines the importance of the Jacobian regularizer. Intuitively, a small adversarial perturbation becomes unlikely to change the output of the model drastically when the magnitude of input-output Jacobian is small. It is shown that Jacobian regularization can be combined with brute-force adversarial training to improve robustness against adversarial attacks.

A3. Experiment results

In this section, we investigate whether other adversarial defense approaches such as Defensive distillation and Jacobian regularization can also be also used to generate faithful explanations, or the effect of improving interpretability is just a special case of PDG training. We conduct experiments on CIFAR10 dataset and use ResNet18 as the model architecture. As for the hyperparameters of the two adversarial defense approach, the temperature T set to 100 for Distillation defensive and the factor λ for Jacobian Regularization is set 0.5. The rest of hyperparameters are same as the submitted paper.

For the standard trained model and adversarial trained models, we use vanilla saliency maps as the attribution method to generate explanations, then compare the impact of different adversarial defense approaches on the faithfulness of saliency maps. As shown in Figure 1, Defensive distillation doesn't improve the faithfulness of saliency maps, the ROAR performance of defensive distillation is quite close to the standard training model. Jacobian regularization has similar effect to improve the faithfulness of interpretability, while the effect is slighter lesser than PGD training.



(a) CIFAR10

Figure 1: ROAR evaluation on CIFAR10, which shows the impact of different adversarial defense approaches on the faithfulness of saliency maps.

Figure 2 shows the saliency maps of standard trained model compared with Defensive distillation, Jacobian regularization and PGD training. As shown in the figure, saliency maps from the Jacobian regularization trained model and PGD trained model all align with human perception (Tsipras et al., 2018), while the defensive distillation does not have such effect. Interestingly, the visualization result align with the ROAR evaluation shown in Figure 1: only those saliency maps are human understandable can generate faithful explanations.

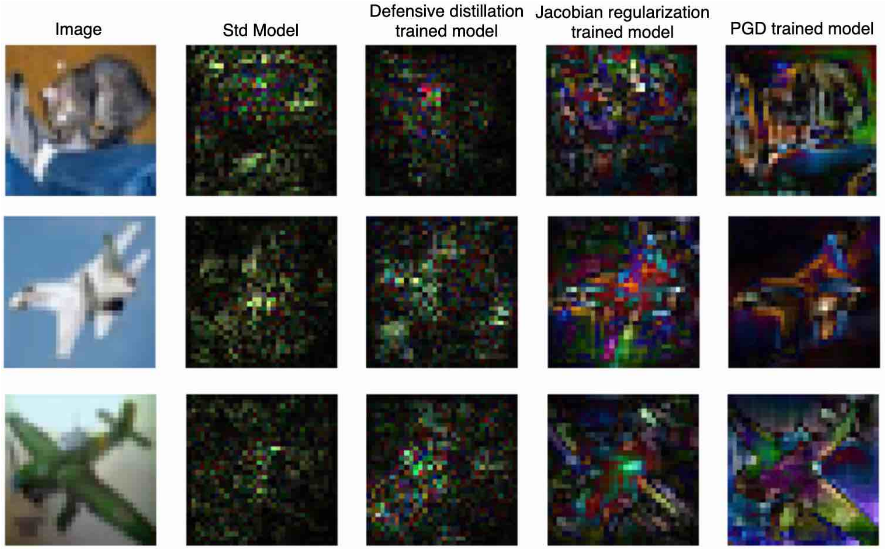


Image
Std Model
Defensive distillation trained model
Jacobian regularization trained model
PGD trained model

Figure 2: Saliency maps of standard trained model and adversarially trained models.

B. Impact of PGD hyperparameters

As discussed in Section 4.4 of the submitted paper, we used PGD-ℓ_2 adversarial trained models to generate faithful explanations. PGD training iteratively computes an adversarial example as:

$$x'_{t+1} = \Pi_{clip}\left(x'_t + \alpha\,\mathrm{sgn}\left(\nabla_x L(\theta, x, y)\right)\right)$$

where $\Pi_{\text{clip },\varepsilon}$ is a clips function computed as:

$$\Pi_{\text{clip, },\varepsilon}(x) = \begin{cases} x + \frac{x'-x}{\|x'-x\|_p}\varepsilon & \text{if } \|x'-x\|_p > \varepsilon \\ x' & \text{otherwise} \end{cases}$$

where the degree of robustness of adversarial examples can be controlled by two hyperparameters: radius ϵ and iteration size. The perturbed adversarial example remains a valid input within a ϵ-ball surrounding the benign sample x, and the distance can be defined as different ℓ_p norm distances. In this section 3.2, we used PGD-ℓ_2 with to generated faithful explanations to align the generated atttribution due to it has the best performance. In this section, we show the impact of PGD hyper-parameters with different ℓ_p norm on explanation's faithfulness.

The following table shows the ROAR evaluation of PGD trained model's attribution maps under different PGD hyperparameters. The experiment was conducted on CIFAR10 dataset. We evaluated 20-step and 40-step PGD training. The ϵ were set as 1.0, 1.5, 2.0, 2.5 for PGD-ℓ_2 training and 4/255, 8/255, 16/255 for PGD-ℓ training, respectively. The rest of hyperparameters are the same as the submitted paper.

Table 1: Remove and Retrain(ROAR) evaluation for attribution maps from PGD trained models under different settings

Method	Hyperparameters		Remove and Retrain Rate η			
	ϵ	t	20%	40%	60%	80%
PGD-ℓ_2	1	20	87.48%	81.98%	76.24%	66.43%
	1	40	88.13%	81.68%	75.81%	66.14%
	1.5	20	88.4%	**81.45%**	74.50%	**63.74%**
	1.5	40	**87.43%**	81.94%	74.90%	65.12%
	2	20	88.31%	82.8%	**74.03%**	65.03%
	2	40	87.93%	83.78%	75.74%	64.53%
	2.5	20	89.04%	82.22%	75.1%	63.94%
	2.5	40	88.62%	82.81%	75.01%	65.77%
PGD-ℓ_∞	4/255	20	**88.13%**	84.68%	78.39%	67.54%
	4/255	40	88.61%	**83.7%**	78.64%	68.02%
	8/255	20	88.76%	83.86%	77.26%	66.25%
	8/255	40	89.43%	84.05%	**77.03%**	**65.24%**
	16/255	20	90.75%	86.15%	81.15%	71.66%
	16/255	40	90.21%	85.78%	80.24%	70.93%

The degree of robustness is mainly determined by the radius ϵ. As shown in Table 1, the ROAR performances are close to each other, and the experiment did not show a strong correlation between the degree of robustness and the faithfulness of derived explanation. It is worth noting that PGD-ℓ_2 trained model performed slightly better than PGD-ℓ_∞ trained model.

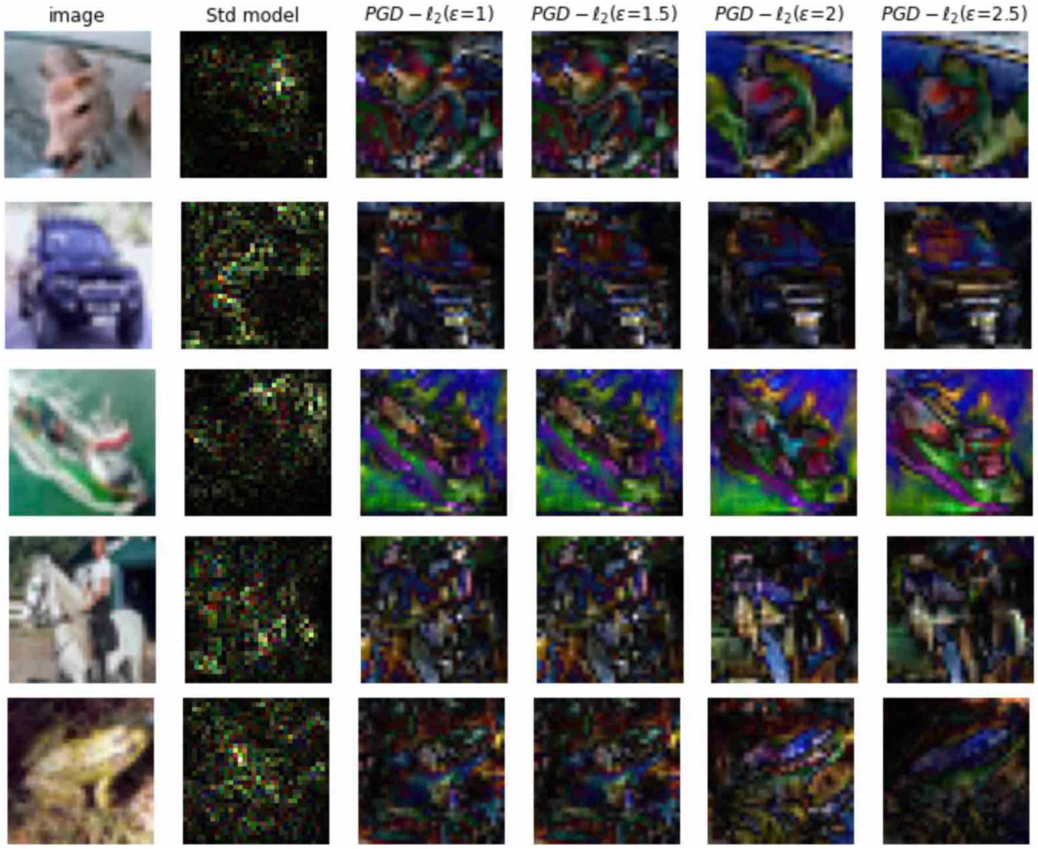


image
Std model
$PGD-\ell_2(\varepsilon=1)$
$PGD-\ell_2(\varepsilon=1.5)$
$PGD-\ell_2(\varepsilon=2)$
$PGD-\ell_2(\varepsilon=2.5)$

Figure 3: Saliency maps of standard trained model and PGD trained models with different ϵ. Saliency maps from higher ϵ models are more similar to the original image.

On the other hand, we found an interesting correlation between the degree of robustness ϵ and the shape of the saliency maps. As shown in Figure 3, saliency maps from higher ϵ models are more similar to the original image. This phenomenon aligns with the findings of (Etmann et al., 2019) that showed an interesting relationship between a model's adversarial robustness, its saliency map and its original input image:

$$\rho(x) \leq \frac{|\langle x, \nabla\Psi(x)\rangle|}{\|\nabla\Psi(x)\|} + \xi \tag{1}$$

where $\rho(x)$ represents the degree of robustness of the model, $\nabla\Psi(x)$ denotes the saliency map explanation, $\frac{|\langle x, \nabla\Psi(x)\rangle|}{\|\nabla\Psi(x)\|}$ is a metric to depict the similarity between the original image x and its saliency map $\nabla\Psi(x)$ (we call it the *alignment* between image and its saliency map),

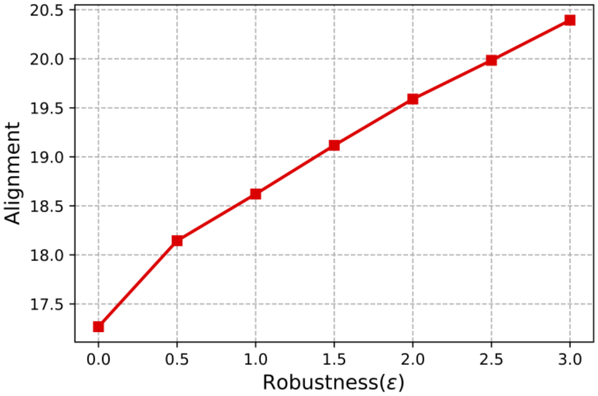


Figure 4: Relation between the alignment $\frac{|\langle x, \nabla\Psi(x)\rangle|}{\|\nabla\Psi(x)\|}$ and the robustness ϵ in PGD training

and ξ is an error term. The above inequality shows that when the robustness (the distance to the decision boundary) grows, saliency map will be more similar to input image.

In (Etmann et al., 2019), the degree of robustness $\rho(x)$ is defined by:

$$\rho(x) = \inf_{e \in X} \{\|e\| : f(x+e) \neq f(x)\} \tag{2}$$

where $f(.)$ denotes the output of a machine learning. Equation 3.6 depicts the minimum distance between data points to the decision boundary of the model. In our work, we choose the radius of PGD training ϵ as another measurement to depict the robustness of adversarially trained models. As shown in Figure 4, as the radius of PGD adversary increases, so does the alignment(the similarity between the input image and its saliency map). In other word, when a model is PGD adversarially trained with higher ϵ, its saliency map will be more similar to the input image, which is according with the visualization result shown in Figure 3.

References

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017.

Christian Etmann, Sebastian Lunz, Peter Maass, and Carola-Bibiane Schönlieb. On the connection between adversarial robustness and saliency map interpretability. *arXiv preprint arXiv:1905.04172*, 2019.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Andrei Margeloiu, Nikola Simidjievski, Mateja Jamnik, and Adrian Weller. Improving interpretability in medical imaging diagnosis using adversarial training. *arXiv preprint arXiv:2012.01166*, 2020.

Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*, pages 582–597. IEEE, 2016.

Andrew Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

Tianyuan Zhang and Zhanxing Zhu. Interpreting adversarially trained convolutional neural networks. In *International Conference on Machine Learning*, pages 7502–7511. PMLR, 2019.