Attributional Robustness Training using Input-Gradient Spatial Alignment

Mayank Singh[1*], Nupur Kumari[1*], Puneet Mangla[1,2], Abhishek Sinha[1#], Balaji Krishnamurthy[1] , Vineeth N Balasubramanian[2]

[1]Media and Data Science Research, Adobe, Noida, INDIA
[2]Indian Institute of Technology, Hyderabad, INDIA

*Equal contribution
#Work done at Adobe

ECCV'20 ONLINE
23-28 AUGUST 2020

Speaker Bio - Puneet Mangla

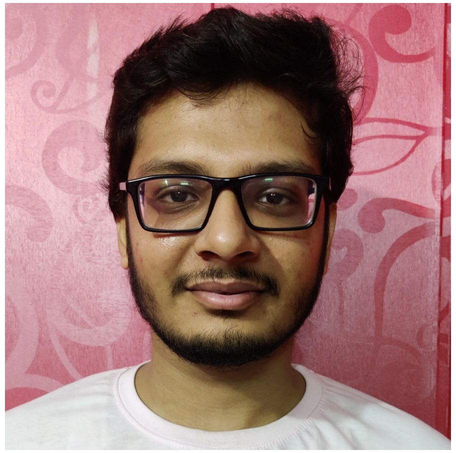
Final year Computer Science Undergraduate,
Member of *Lab1055: Machine Learning and Vision Group,*
IIT Hyderabad, India

Summer Intern 2019 & 2020 at
Adobe Media and Data Science Research Labs, Noida, India

Recent Publications at
WACV 2020, ECML-PKDD 2020, ECCV 2020

Research Interests
Robustness, Limited Supervision, GANs

Others:
Binge-watcher, music and a food lover

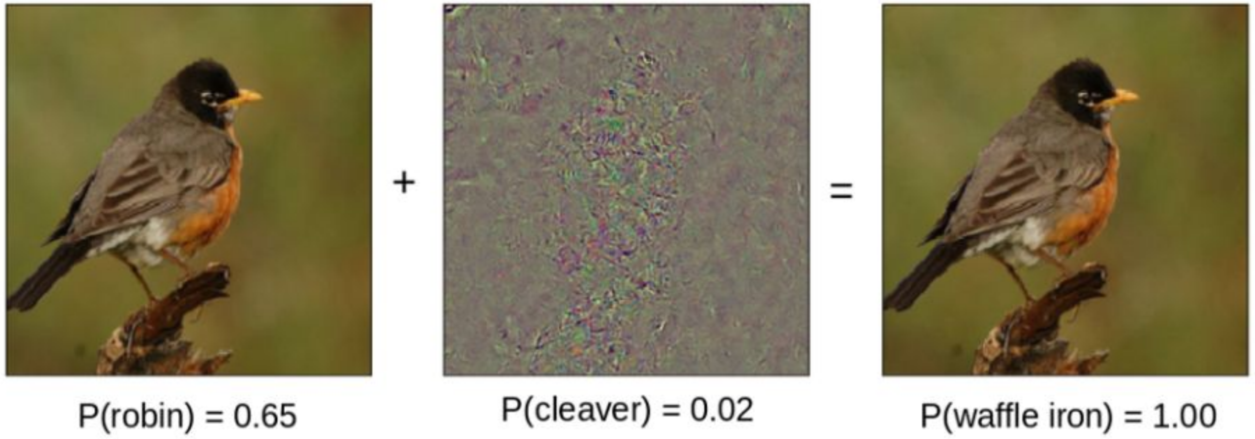


Puneet Mangla
CSE, IIT Hyderabad
@puneet2k

More at:
https://puneet2000.github.io

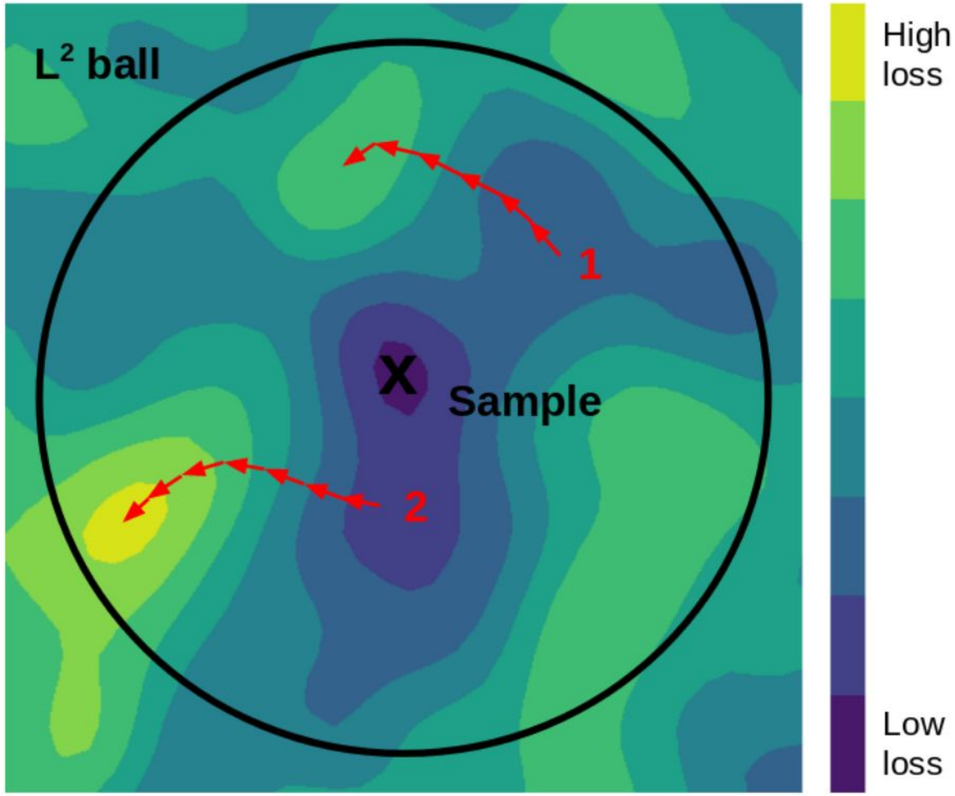
Adversarial Attacks

Imperceptible perturbations fooling model's prediction



Left: natural image. Middle: Adversarial perturbation found by PGD attack against ResNet50 model, size of perturbation is magnified x100 to be more visible. Right: adversarial example.

Projected Gradient Descent (PGD)



Projected gradient descent with restart. 2nd run finds a high loss adversarial example within the L^2 ball. Sample is in a region of low loss.

Adversarial Training

- simply putting the PGD attack inside your training loop.
- “ultimate data augmentation”
- create specific perturbations that best fool our model and classify them correctly

Regular Training

$$\min_{\theta} \mathcal{L}(x, y; \theta)$$

Adversarial Training

$$\min_{\theta} \max_{\delta \in \Delta} \mathcal{L}(x + \delta, y; \theta)$$

Image Attribution Methods

Integrated Gradient Attribution Map

$$IG(x, f(x)_i) = (x - \overline{x}) \odot \int_{t=0}^{1} \nabla_x f(\overline{x} + t(x - \overline{x}))_i dt$$

Explanation techniques that aim to highlight relevant input features responsible for model's prediction e.g.

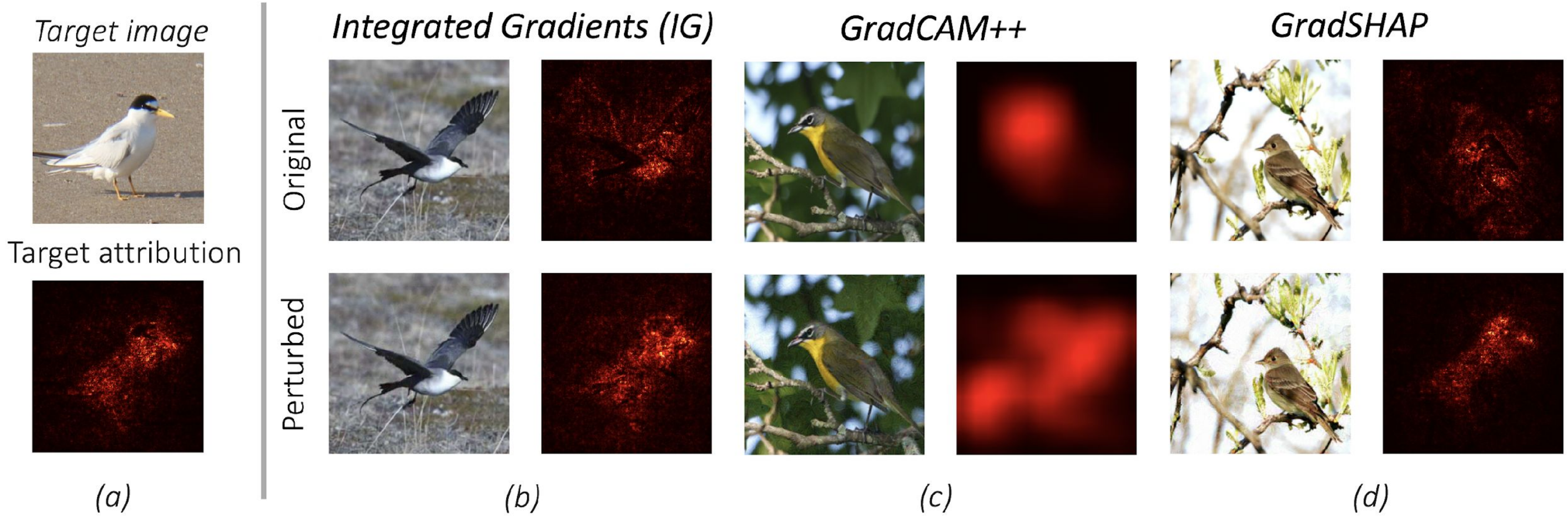
- Integrated Gradients [1]
- Gradient [2]
- GradCAM++ [3]
- GradSHAP [4]

Attribution Attack

Perturbations that can arbitrarily manipulate attribution maps without affecting the model's prediction. Example targeted attribution attack is shown below

$$\arg\max_{\delta \in B_\epsilon} D[A(x+\delta, f(x+\delta)_y), A(x, f(x)_y)]$$

$$\text{subject to: } \arg\max(f(x)) = \arg\max(f(x+\delta)) = y$$



Attributional Robustness Objective

- Minimize $$\max_{\delta \in B_\epsilon} ||A(x + \delta, f(x + \delta)_y) - A(x, f(x)_y)||$$

Where $A(x, f(x)_y)$ is the attribution map of input x with respect to ground truth class y and f is the classification model.

Prior Work

- Robust attribution regularization [5] : This methodology directly regularizes the Attribtional Robustness objective with Integrated Gradients (IG) as the attribution method i.e.

Minimize: $$\arg\max_{\delta \in B_\epsilon} ||IG(x, x+\delta)|| + L_{ce}(x, y)$$

Here L_{ce} is the standard cross entropy loss.

Key Contributions

- We propose *ART*, a new training methodology to learn attributionally robust model by maximizing the spatial correlation between the input and its attribution map.

- We empirically show that the proposed methodology also induces immunity to adversarial perturbations and and common perturbations on standard vision datasets.

- We show that *ART* improves performance on other computer vision tasks such as weakly supervised object localization and segmentation. It achieves state-of-the-art performance in weakly supervised object localization on CUB-200 dataset.

Attributional Robustness Training (*ART*)

Minimize $\max_{\delta \in B_\epsilon} ||g^y(x+\delta) - g^y(x)||$

Where $g^y(x)$ is the gradient attribution w.r.t. input x.

The above term is upper bounded by spatial correlation between attribution map and input and we reduce this upper bound during training for attributional robustness

$$\begin{aligned}
||g^y(x+\delta) - g^y(x)|| &= ||g^y(x+\delta) - (x+\delta) - (g^y(x) - x) + \delta|| \\
&\leq ||g^y(x+\delta) - (x+\delta)|| + ||g^y(x) - x|| + ||\delta|| \\
&\leq ||g^y(x+\delta) - (x+\delta)|| + \max_{\delta \in B_\epsilon} ||g^y(x+\delta) - (x+\delta)|| + ||\delta||
\end{aligned}$$

$$\max_{\delta \in B_\epsilon} ||g^y(x+\delta) - g^y(x))|| \leq 2 \max_{\delta \in B_\epsilon} ||g^y(x+\delta) - (x+\delta)|| + ||\epsilon||$$

ART objective

$$\underset{\theta}{\text{minimize}} \; \underset{(x,y)}{\mathbb{E}} \left[L_{ce}(x+\delta, y) + \lambda \, L_{attr}(x+\delta, y) \right]$$

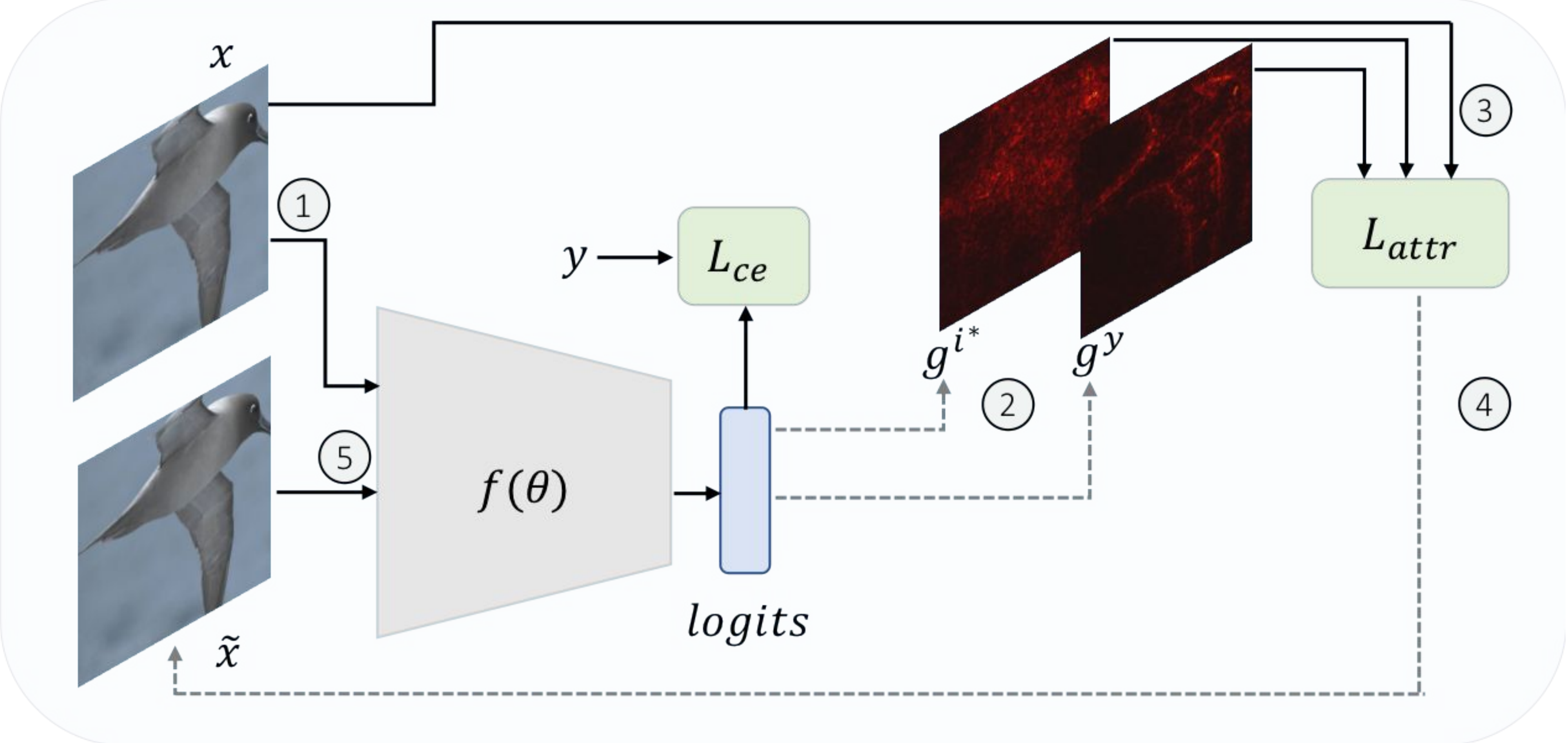
$$\text{where } \delta = \underset{||\delta||_\infty < \epsilon}{\arg\max} \; L_{attr}(x+\delta, y)$$

$$L_{attr}(x,y) = \log\Big(1 + \exp\big(-(d(g^{i^*}(x), x) - d(g^y(x), x))\big)\Big)$$

$$\text{where } d(g^i(x), x) = 1 - \frac{g^i(x).x}{||g^i(x)||_2.||x||_2} \;\; ; \;\; i^* = \underset{i \neq y}{\arg\max} f(x)_i$$

L_{ce} is the standard cross entropy loss and L_{attr} is a triplet loss between input x, its positive anchor $g^y(x)$ and negative anchor $g^{i*}(x)$ to increase the spatial correlation between x and $g^y(x)$.

ART objective



L_{ce} is the standard cross entropy loss and L_{attr} is a triplet loss between input x, its positive anchor $g^{y}(x)$ and negative anchor $g^{i*}(x)$ to increase the spatial correlation between x and $g^{y}(x)$.

Connection to Adversarial Robustness

Adversarial examples are calculated by optimizing a loss function L which is large when $f(x) \neq y$:

$$x_{adv} = \underset{x':||x'-x||_p<\epsilon}{\arg\max} \; L(\theta, x', y) \tag{7}$$

where L can be the cross-entropy loss, for example. For an axiomatic attribution function A which satisfies the completeness axiom i.e. $\sum_{j=1}^{n} A(x)_j = f(x)_y$, it can be shown that $|f(x)_y - f(x')_y| < ||A(x) - A(x')||_1$, as below:

$$\begin{aligned} |f(x)_y - f(x')_y| &= |\textstyle\sum_{j=1}^{n} A(x)_j - \sum_{j=1}^{n} A(x')_j| \\ &\leq \textstyle\sum_{j=1}^{n} |A(x)_j - A(x')_j| \\ &= ||A(x) - A(x')||_1 \end{aligned} \tag{8}$$

Comparison with prior state-of-the-art methods

Dataset	Approach	Attributional Robustness		Accuracy	
		IN	K	Natural	PGD-40 Attack
CIFAR-10	Natural	40.25	49.17	95.26	0.
	PGD-10 [6]	69.00	72.27	87.32	44.07
	ART	**92.90**	**91.76**	89.84	37.58
SVHN	Natural	60.43	56.50	95.66	0.
	PGD-7 [6]	39.67	55.56	92.84	50.12
	ART	**61.37**	**72.60**	95.47	43.56
GTSRB	Natural	68.74	76.48	99.43	19.9
	IG Norm [5]	74.81	75.55	97.02	75.24
	IG-SUM Norm [5]	74.04	76.84	95.68	77.12
	PGD-7 [6]	86.13	88.42	98.36	87.49
	ART	**91.96**	**89.34**	98.47	84.66
Flower	Natural	38.22	56.43	93.91	0.
	IG Norm [5]	64.68	75.91	85.29	24.26
	IG-SUM Norm [5]	66.33	79.74	82.35	47.06
	PGD-7 [6]	**80.84**	84.14	92.64	69.85
	ART	79.84	**84.87**	93.21	33.08

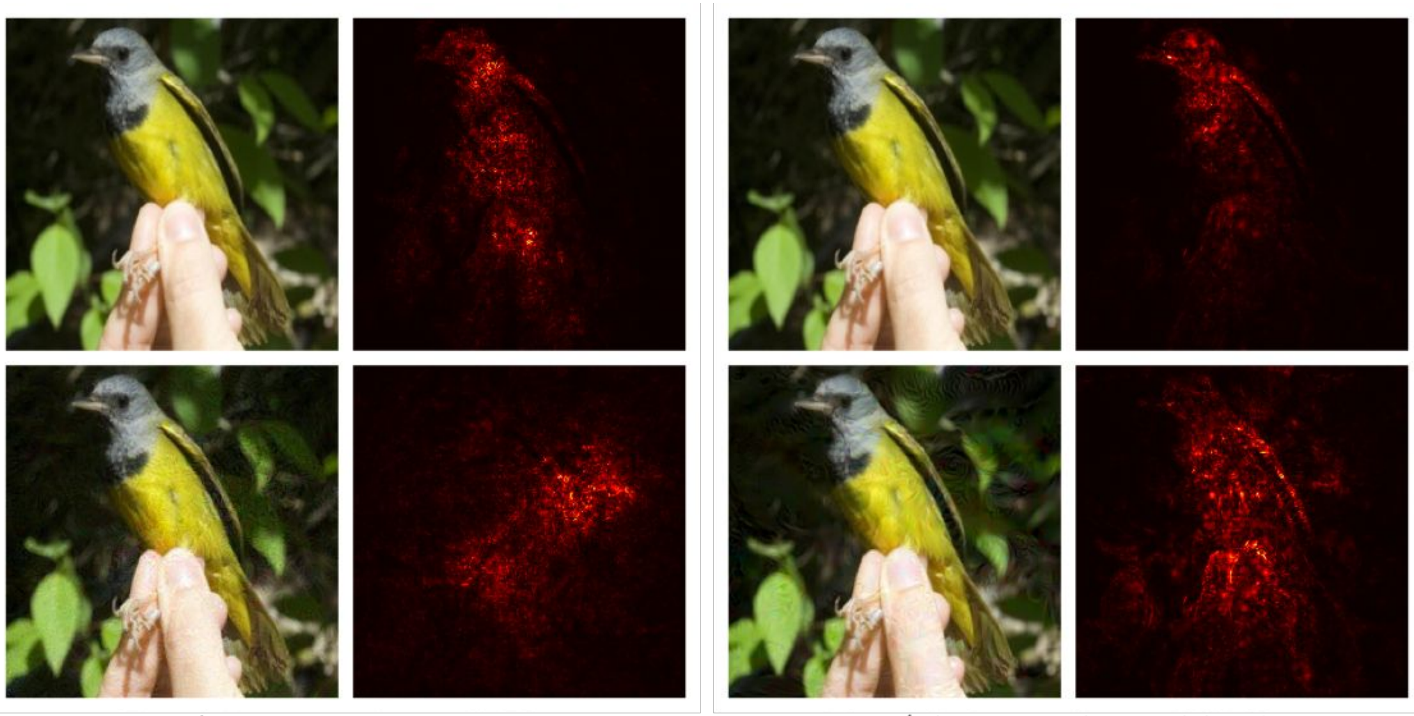
Kendall's coefficient (K): measure of similarity of ordering when ranked by values

Top-k intersection (IN): measures the percentage of common indices in top-k values of attribution map of x and $\tilde{x}$.

Qualitative Analysis of Attribution Robustness

Natural Model | ART Model

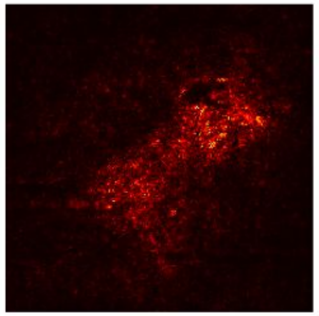
Original / Perturbed



Top-k intersection : 2.74
Kendall correlation : 30.72

Top-k intersection : 20.76
Kendall correlation : 57.36

Target attribution



Target image

Robustness to common perturbations

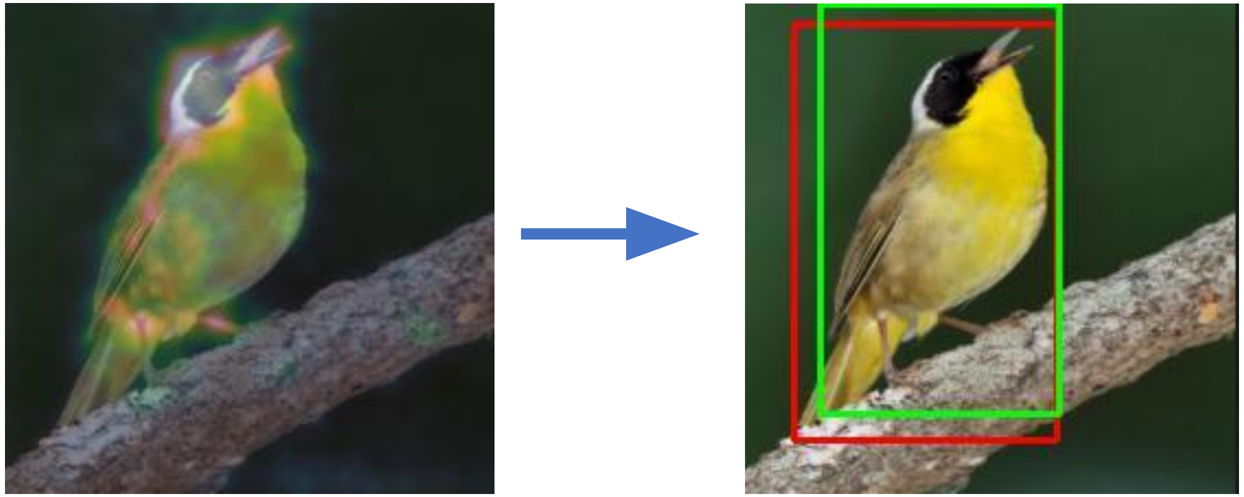
CIFAR-10-C: consists of perturbed images of 15 common-place visual perturbations at five levels of severity

Table 3: Top-1 accuracy of different models on perturbed variants of test-set (GN:Gaussian noise; SN: Shot noise; IN: Impulse noise; DB: Defocus blur; Gl-B: Glass blur; MB: Motion blur; ZB: Zoom blur; S: Snow; F: Fog; B: Brightness; C: Contrast; E: Elastic transform; P: Pixelation noise; J: JPEG compression; Sp-N: Speckle Noise)

Models	**GN**	**SN**	**IN**	**DB**	**Gl-B**	**MB**	**ZB**	**S**	**F**	**B**	**C**	**E**	**P**	**J**	**Sp-N**
Natural	49.16	61.42	59.22	83.55	53.84	79.16	79.18	84.53	**91.6**	**94.37**	**87.63**	84.44	74.12	79.76	65.04
PGD-10	83.32	84.33	73.73	83.09	81.27	79.60	82.07	82.68	68.81	85.97	57.86	81.68	85.56	85.56	83.64
ART	**85.44**	**86.41**	**77.07**	**86.07**	**81.70**	**83.14**	**85.54**	**84.99**	71.04	89.42	56.69	**84.72**	**87.64**	**87.89**	**86.02**

Downstream task of WSOL

- Weakly Supervised Object Localization (WSOL): detecting objects when only class label information of images is available.

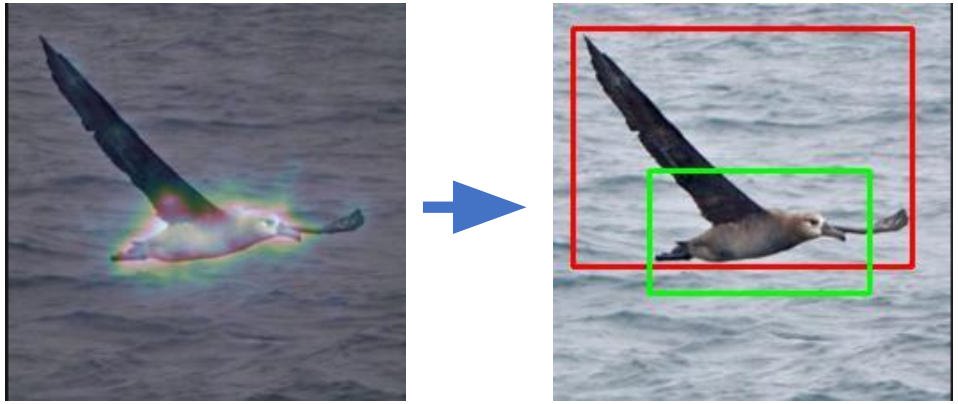


Exploits attribution map to localize the object and detect bounding box

Prior Art on WSOL

- Attention-based dropout layer for weakly supervised object localization (ADL) [7]

ADL aims at solving the problem of attribution map focusing only on the most discriminative region of the image and thus missing the complete object.



Weakly Supervised Image Localization (WSOL)

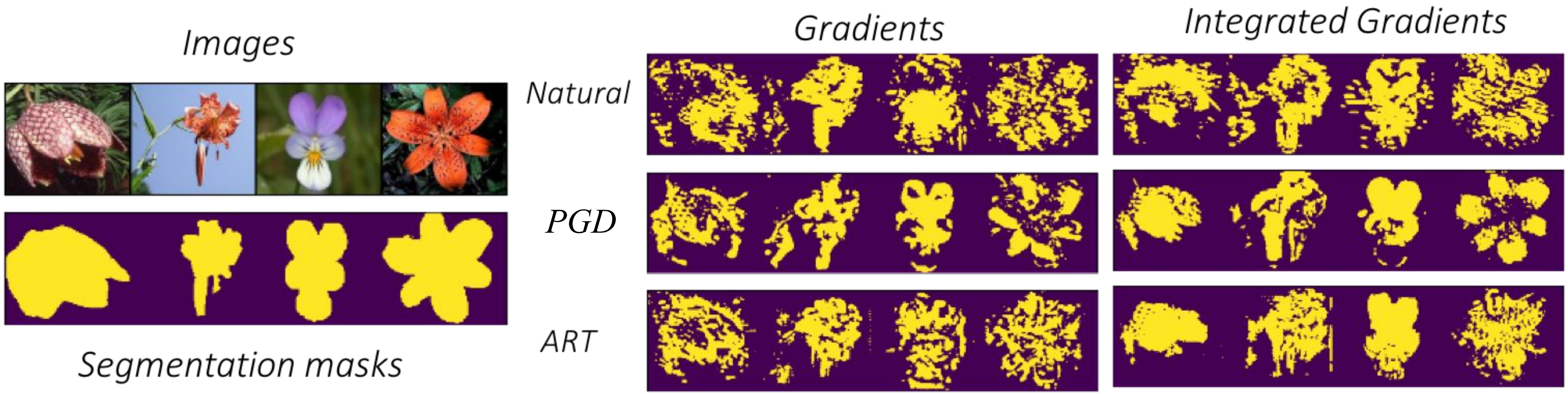
Model	Method	**Saliency Method**				Top-1 Acc
		Grad		CAM		
		GT-Known Loc	Top-1 Loc	GT-Known Loc	Top-1 Loc	
ResNet50-SE	ADL [7]	-	-	-	62.29*	80.34*
ResNet50	ADL#	52.93	43.78	56.85	47.53	80.0
	Natural	50.2	42.0	60.37	50.0	81.12
	PGD-7 [6]	66.73	47.48	55.24	39.45	70.3
	ART	**82.65**	**65.22**	58.87	46.02	77.51
VGG-GAP	ADL#	63.18	43.59	69.36	50.88	70.31
	Natural	72.54	53.81	48.75	35.03	72.94
	ART	**76.50**	**57.74**	52.88	40.75	74.51

#Our implementation using officially released code

Qualitative Comparison on WSOL

Comparison of heatmap and estimated bounding box by VGG model trained via our method and ADL[7] on CUB dataset. The red bounding box is ground truth and green bounding box corresponds to the estimated box

Weakly Supervised Image Segmentation



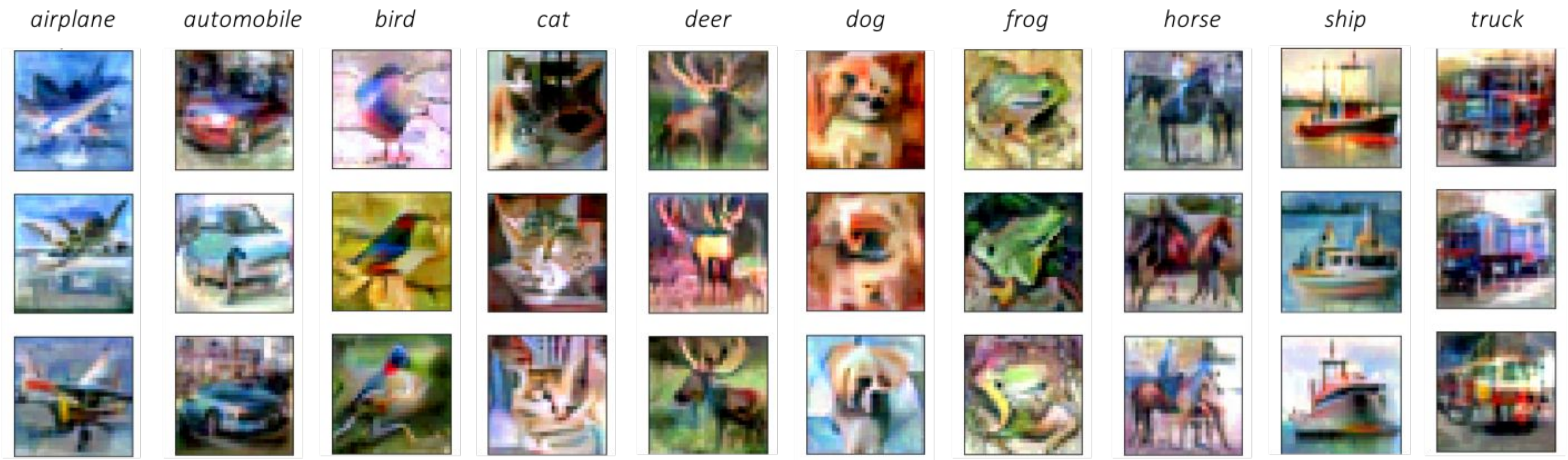
Qualitative example of Gradient Attribution Map on CIFAR-10

Natural

PGD-10 [6]

ART [7]

Image generation through gradient backpropagation



Random samples (of resolution 32 × 32) generated using a CIFAR-10 robustly trained ART classifier as described in [8]

References

1. Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks." ICML (2017).
2. Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." *arXiv preprint arXiv:1312.6034* (2013).
3. Chattopadhay, Aditya, et al. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks." *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018.
4. Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Advances in neural information processing systems*. 2017.
5. Chen, Jiefeng, et al. "Robust attribution regularization." *Advances in Neural Information Processing Systems*. 2019.
6. Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." *arXiv preprint arXiv:1706.06083* (2017).
7. Choe, Junsuk, and Hyunjung Shim. "Attention-based dropout layer for weakly supervised object localization." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.
8. Santurkar, S., Ilyas, A., Tsipras, D., Engstrom, L., Tran, B., Madry, A.: Image synthesis with a single (robust) classifier. In: NeurIPS (2019)

Thank You!

Project Page at : https://nupurkmr9.github.io/Attributional-Robustness/

Code available at: https://github.com/nupurkmr9/Attributional-Robustness