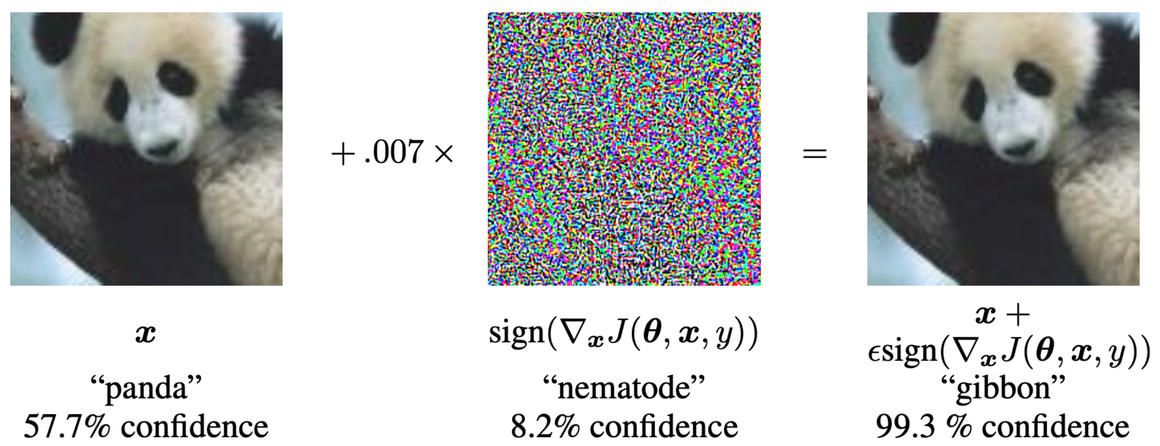
Functional Adversarial Attacks

Cassidy Laidlaw and Soheil Feizi

Department of Computer Science

Adversarial Examples



Consider a classifier $g : \mathcal{X}^n \to \mathcal{Y}$ from a feature space $\mathcal{X}^n$ to a set of labels $\mathcal{Y}$. Given an input $\mathrm{x} \in \mathcal{X}^n$, an adversarial example is a slight perturbation $\widetilde{\mathrm{x}}$ of x such that $g(\widetilde{\mathrm{x}}) \neq g(\mathrm{x})$; that is, $\widetilde{\mathrm{x}}$ is given a different label than x by the classifier.

Adversarial Threat Models

How does one define a "slight perturbation"? A *threat model* defines a set of imperceptible transformations for a natural input. We argue that existing threat models do not encompass the full range of perturbations that are imperceptible.

Additive (ℓ_p) Threat Model

$$(x_1, \ldots, x_n) \to (x_1 + \delta_1, \ldots, x_n + \delta_n)$$

- The usual threat model used for adversarial examples
- Each feature is perturbed by adding a small amount δ_i
- The norm of all the amounts is bounded, e.g. for the ℓ_2 norm $\|(\delta_1, \ldots, \delta_n)\|_2 < \epsilon$

Functional Threat Model

$$(x_1, \ldots, x_n) \to (f(x_1), \ldots, f(x_n))$$

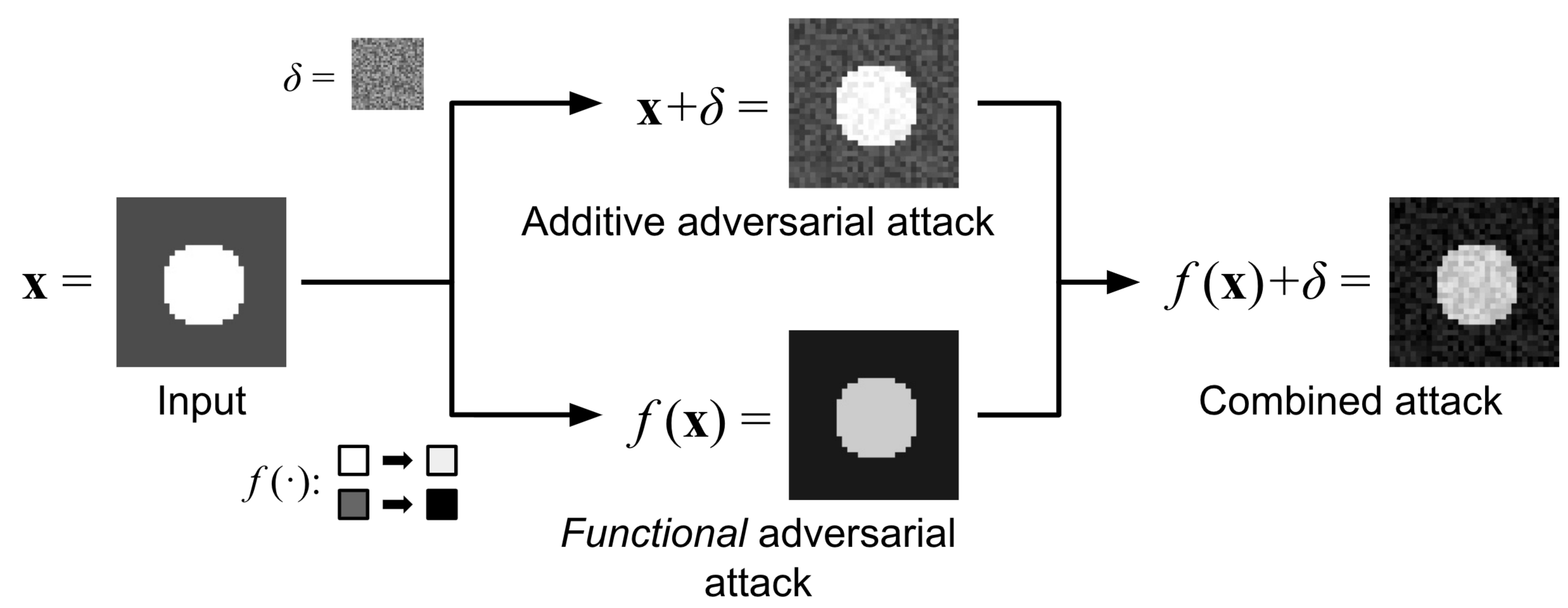
- We propose a new class of threat models for adversarial attacks called *functional threat models*
- Adversarial examples are generated by applying a *single* function f to all features of the input
- The uniformity of the perturbation makes the change less perceptible, allowing for larger absolute modifications

Combined Threat Model

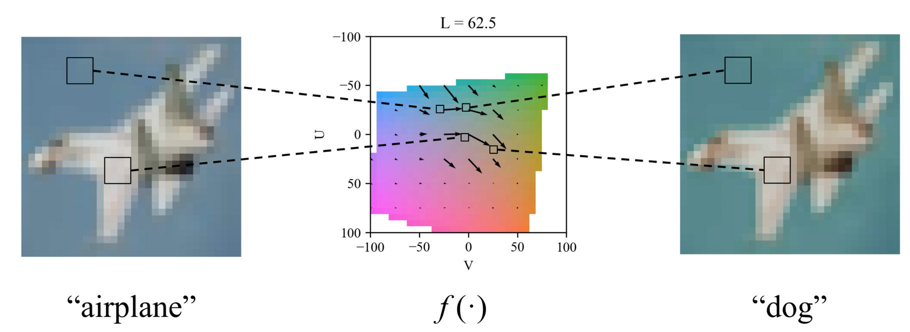
$$(x_1, \ldots, x_n) \to (f(x_1) + \delta_1, \ldots, f(x_n) + \delta_n)$$

- Functional threat models can be combined with additive or other existing threat models
- We prove that the combined threat model encompasses more potential perturbations than the union of the constituents

Overview



ReColorAdv: Functional Attack on Image Colors



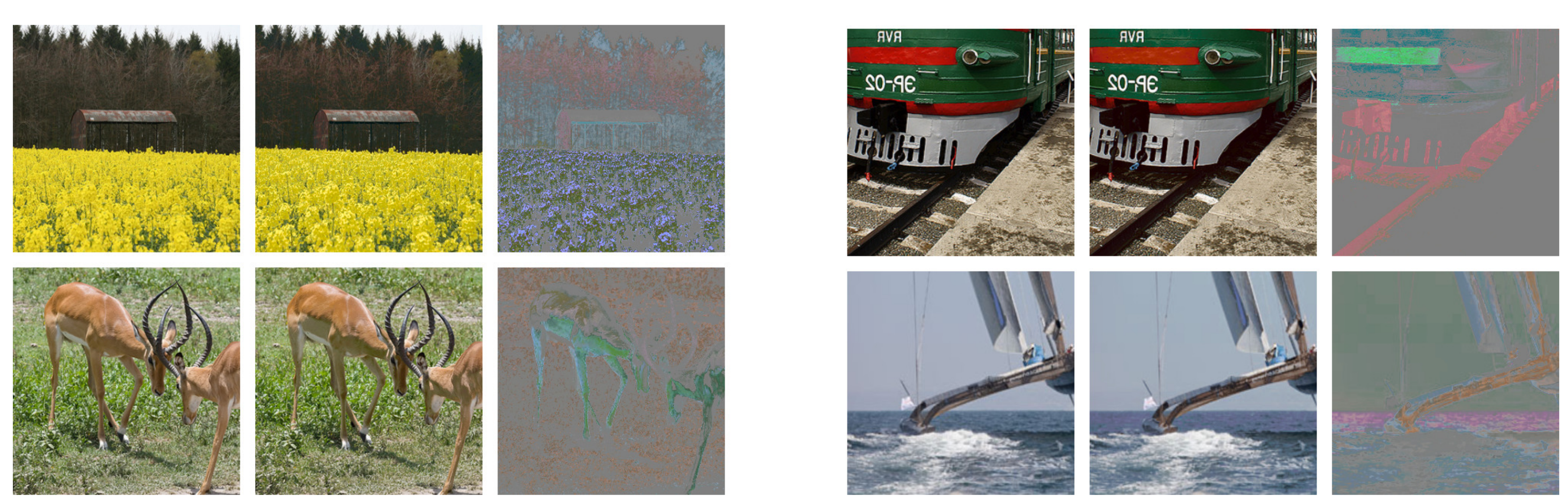
ReColorAdv is a novel adversarial attack against image classifiers that leverages a functional threat model. ReColorAdv generates adversarial examples by uniformly perturbing each pixel x_i in the input image x with a function $f : \mathcal{C} \to \mathcal{C}$:

$$x_i = (c_{i,1}, c_{i,2}, c_{i,3}) \in \mathcal{C} \subseteq [0, 1]^3 \to \widetilde{x}_i = (\widetilde{c}_{i,1}, \widetilde{c}_{i,2}, \widetilde{c}_{i,3}) = f(c_{i,1}, c_{i,2}, c_{i,3})$$

Regularization and Scope

- Perturbation function $f(\cdot)$ is bounded to prevent it from modifying any color by too large of an amount
- PGD with smoothing term encourages similar colors to be perturbed in similar ways
- Works with different color spaces including RGB and CIELUV (perceptually accurate)
- Can be combined with other attacks such as Carlini and Wagner's [1] and spatially-transformed adversarial examples [2]

Examples on ImageNet (left-to-right: original, adversarial example, perturbation)



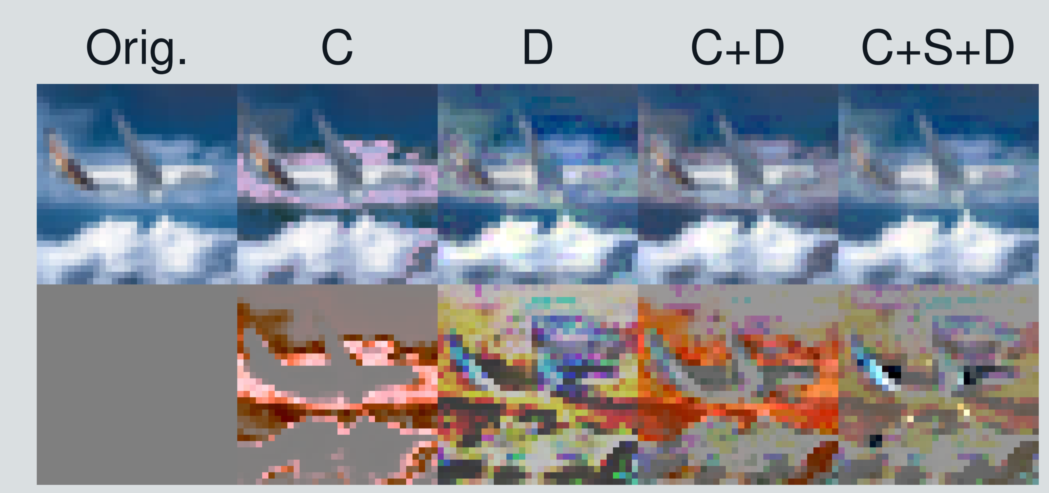
Experiments

CIFAR-10 Accuracy Under Attack

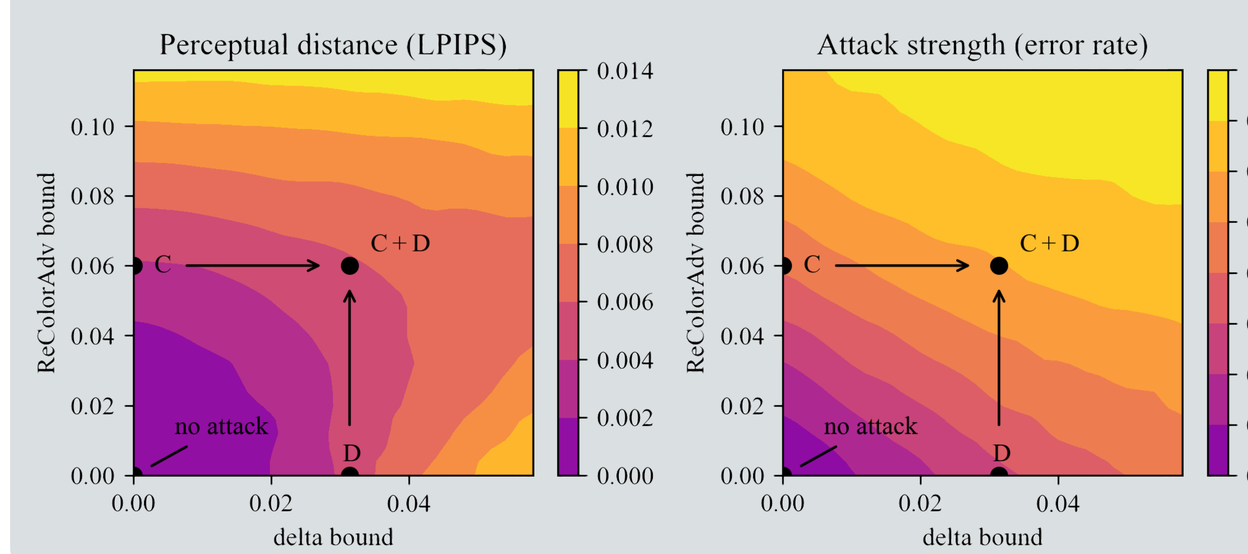
	Defense		
Attack	**None**	**Adv. training**	**TRADES [3]**
C	3.3	45.8	59.2
D	**0.0**	30.1	53.6
S	1.2	26.2	26.6
C+S	0.9	8.7	17.5
C+D	**0.0**	5.2	22.0
S+D	**0.0**	7.6	8.7
C+S+D	**0.0**	**3.6**	**5.7**

C is ReColorAdv attack, **D** is an ℓ_∞ attack, **S** is StAdv attack [2]. Attacks are evaluated separately and combined.

Perceptibility



Combinations of attacks are less perceptible than a single attack. **Above:** unbounded attacks against a TRADES-trained network. **Below:** empirical evaluation using learned perceptual image-patch similarity (LPIPS) [4].



References

[1] 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.

[2] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially Transformed Adversarial Examples. *arXiv preprint arXiv:1801.02612*, 2018.

[3] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically Principled Trade-off Between Robustness and Accuracy. In *ICML 2019*, 2019.

[4] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

Paper

Code