DECISION-BASED ADVERSARIAL ATTACKS:
RELIABLE ATTACKS AGAINST BLACK-BOX MACHINE LEARNING MODELS

Article

Brendel, W., Rauber, J., & Bethge, M. (2017). Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. arXiv preprint arXiv:1712.04248.

Authors

Wieland Brendel, Jonas Rauber & Matthias Bethge

Presented by

LIETARD Bastien, Rousselot Cyriaque & Iken Omar

Master 2 Data Science

Université de Lille

January 30, 2022

Adversarial Attacks

Find perturbations that lead a model to fail

Interests :
Foresee and prevent mismatching
Avoid misclassification to a specific label
Adapt models to be more robust

Ideal objectives :
(Human) imperceptible perturbation
Realizable in real-world applications
Robust to defense

Usual types of attack

Gradient-based :
Use of model details : *gradient of the Loss*
Ex. : Carlini & Wagner (2016)
Defense : non-differentiable elements

Score-based :
Use prediction scores (class probabilities...)
$\rightarrow$Numeric estimation of gradient
Ex. : Chen et al. (2017)
Defense : stochastic elements

Transfer-based :
Use of *information like training data*
Train a substitute to synthesize adversarial samples
Ex. : Papernot et al. (2017)
Defense : augmented dataset with perturbed samples

These situations are sometimes **unrealistic** or **too weak** !

Decision-based Attacks

In real world scenarios, class probabilities or logits are hardly available.

Model's information and training data are rarely accessible.

Need to build stronger attacks that would need more complex defenses.

Minimal changes to make the perturbation imperceptible.

II. Boundary Attack 1/4

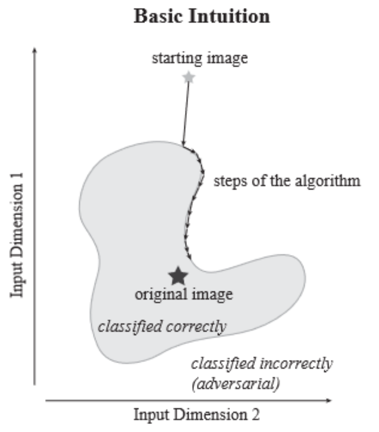


Figure 1: Illustration of a boundary attack

Algorithm 1: Simple version of the boundary attack

Data: Original image o, adversarial criterion $c(.)$, decision model $d(.)$
Result: adversarial example $\tilde{o}$ such that the distance $d(o, \tilde{o}) = ||o - \tilde{o}||_{\mathbf{2}}^{\mathbf{2}}$ is minimized
Initialization: $k = 0$, $\tilde{o}^{\mathbf{0}} \sim \mathcal{U}(0, 1)$
while $k <$ *maximum number of steps* **do**
  draw random perturbation from proposal distribution $\eta_k \sim \mathcal{P}(\tilde{o}^{k-\mathbf{1}})$;
  if $\tilde{o}^{k-\mathbf{1}} + \eta_k$ *is adversarial* **then**
    $\tilde{o}^k \leftarrow \tilde{o}^{k-\mathbf{1}} + \eta_k$
  else
    $\tilde{o}^k \leftarrow \tilde{o}^{k-\mathbf{1}}$

The parameter $c(.)$ can be choose as misclassification (predicts an incorrect label), or targeted misclassification (predict a specific incorrect label) . It is versatile.

- ✓ Untargeted : Random Initialization
- ✓ Targeted : Initialized from a point classified as the target

Efficiency depends on P the proposal distribution. How to choose it ? It is an optimization problem with constraints :

- ✓ The perturbed output is in the domain

$$\tilde{o}_i^{k-1} + \eta_i^k \in [0, 255] \tag{1}$$

- ✓ The size of the perturbation is controlled by a parameter δ

$$||\eta^k||_2 = \delta . d(o, \tilde{o}) \tag{2}$$

- ✓ The distance between the goal and the perturbed output is improved by a parameter ϵ

$$d(o, \tilde{o}^{k-1}) - d(o, \tilde{o}^{k-1} + \eta^k) = \epsilon . d(o, \tilde{o}^{k-1}) \tag{3}$$

Hyperparameters of this attacks are ϵ and δ. They are adjusted with the geometry of the boundary in order to get to 50% misclassified perturbations.

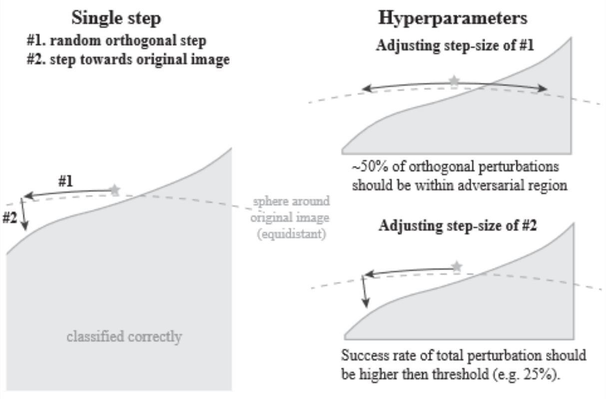


Figure 2: Adjusting step size with the geometry of the boundary

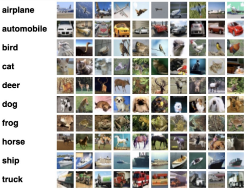
III. Comparison with other Attacks

Datasets

MNIST

[deepai]

CIFAR-10



[paperswithcode]

ImageNet-1000

[image-net]

Settings

Untargeted

Adversarial perturbation flips the label of the original sample to another.

`e.g. dog to cat or car, etc`

Targeted

Adversarial flips the label to a specific target class.

`5 to 0, 6 to 1, etc`

Settings

In the untargeted setting we compare the Boundary Attack against three gradient-based attack algorithms:

Fast-Gradient Sign Method (FGSM)

- ✓ Computes $g = \nabla_o \mathcal{L}(o, c)$ that maximizes the loss $\mathcal{L}$ for the true-label c
- ✓ Seek the smallest ϵ for which $o + \epsilon . g$ is still adversarial.

DeepFool

- ✓ Computes for each class the minimum distance that it takes to reach the class boundary
- ✓ Makes corresponding step in the direction of the class with the smallest distance.

Carlini & Wagner

- ✓ A refined iterative gradient attack
- ✓ Uses the Adam optimizer, multiple starting points to respect a max-based adversarial constraint.

Metric

$$\mathcal{S}_A(M) = median_i \left(\frac{1}{N} \|\eta_{A,M}(o_i)\|_2^2 \right) \tag{4}$$

$\eta_{A,M}(o_i) \in \mathbb{R}^N$: the adversarial perturbation that the attack A finds on model M for the i-th sample o_i.

First setting: Untargeted Attack

Adversarial ≡ any image for which the predicted label is different from the original one.

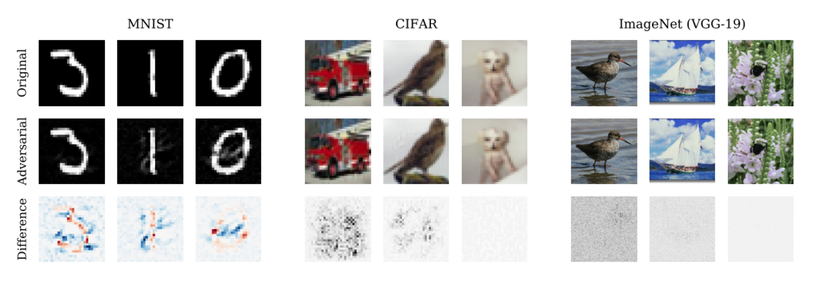


Figure 3: Adversarial examples generated by the Boundary Attack.

Scores

	Attack Type	MNIST	CIFAR	ImageNet VGG-19	ImageNet ResNet-50	ImageNet Inception-v3
FGSM	gradient-based	4.2e-02	2.5e-05	1.0e-06	1.0e-06	9.7e-07
DeepFool	gradient-based	4.3e-03	5.8e-06	1.9e-07	7.5e-08	5.2e-08
Carlini & Wagner	gradient-based	2.2e-03	7.5e-06	5.7e-07	2.2e-07	7.6e-08
Boundary (ours)	decision-based	3.6e-03	5.6e-06	2.9e-07	1.0e-07	6.5e-08

First setting: Untargeted Attack

Here the goal is to synthesize an image that is as close as possible (in L2-metric) to the original image while being misclassified.

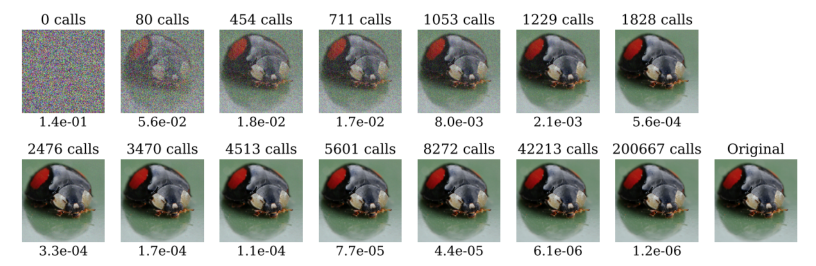


Figure 4: Example of an untargeted attack.

Findings

Boundary Attack is:

1. Competitive with gradient-based attacks in terms of the minimal adversarial perturbations.
2. Very stable against the choice of the initial point

Second setting: Targeted Attack

Here the goal is to synthesize an image that is as close as possible (in L2-metric) to a given image of a tiger cat but is classified as a dalmatian dog.

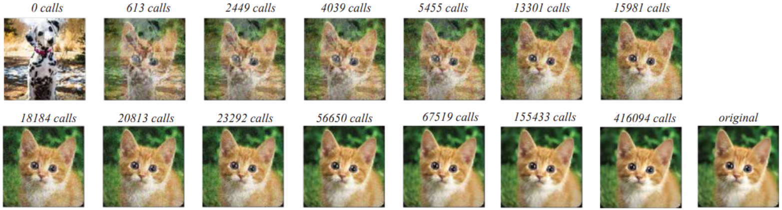


Figure 5: Example of a targeted attack.

To compare the Boundary Attack to Carlini & Wagner:

- ✓ On MNIST and CIFAR a sample with label ℓ gets the target label $\ell + 1$ modulo 10.
- ✓ On ImageNet we draw the target label randomly but consistent across attacks

Scores

	Attack Type	MNIST	CIFAR	VGG-19
Carlini & Wagner	gradient-based	4.8e-03	3.0e-05	5.7e-06
Boundary (ours)	decision-based	6.5e-03	3.3e-05	9.9e-06

Example of attack methods

- ✓ **Gradient masking**: model is modified to yield masked gradients.
- ✓ **Saturated sigmoid network**: an additional regularization term leads the sigmoid activations to saturate.
- ✓ **Defensive distillation**

$$softmax(x, T)_i = \frac{e^{x_i/T}}{\sum_j e^{x_j/T}} \tag{5}$$

1. Train a teacher network as usual but with temperature T.
2. Train a distilled network on the softmax outputs of the teacher.
3. Evaluate the distilled network at temperature $T = 1$ at test time.

Findings

- ✓ Success rate of gradient-based attacks dropped from $\sim 100\%$ down to 0.5%.
- ✓ The size of adversarial perturbations is similar for the distilled/undistilled network.

Scores

		MNIST		CIFAR	
	Attack Type	standard	distilled	standard	distilled
FGSM	gradient-based	4.2e-02	fails	2.5e-05	fails
Boundary (ours)	decision-based	3.6e-03	4.2e-03	5.6e-06	1.3e-05

V. Attacks on Real-World Applications

In the real world, no access to the architecture of the Data. They apply the Boundary Attacks to Black Box models :

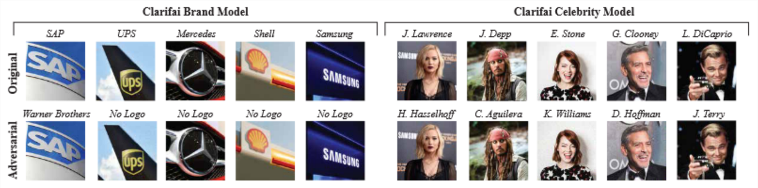


Figure 6: Results on two Black Box Clarifai models to recognize brands and celebrities

The perturbations are more difficult than against ImageNet models. We quantify that by the size of the noise applied ($\sim 1e-2, 1e-3$). For a lot of samples this is humanly indistinguishable.

VI. Conclusion

Boundary attacks follow the decision boundary
between adversarial and non-adv. samples.

- ✓ Allow different attacks (depending on the adv. criterion)
- ✓ Real-world applicable
- ✓ Need very few information/tuning

Going further :

- ✓ learning the proposal distribution.
- ✓ conditioning the proposal distribution to recent history.

Main Article

Brendel, W., Rauber, J., & Bethge, M. (2017). Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. arXiv preprint arXiv:1712.04248.
https://arxiv.org/pdf/1712.04248.pdf