

# Adversarial Defences

## Evaluating Adversarial Robustness

---

Maciej Żelazczyk

January 27, 2021

PhD Student in Computer Science

Division of Artificial Intelligence and Computational Methods

Faculty of Mathematics and Information Science

[m.zelazczyk@mini.pw.edu.pl](mailto:m.zelazczyk@mini.pw.edu.pl)

**Warsaw University  
of Technology**

# Adversarial examples

First described by [Szegedy et al., 2014]:

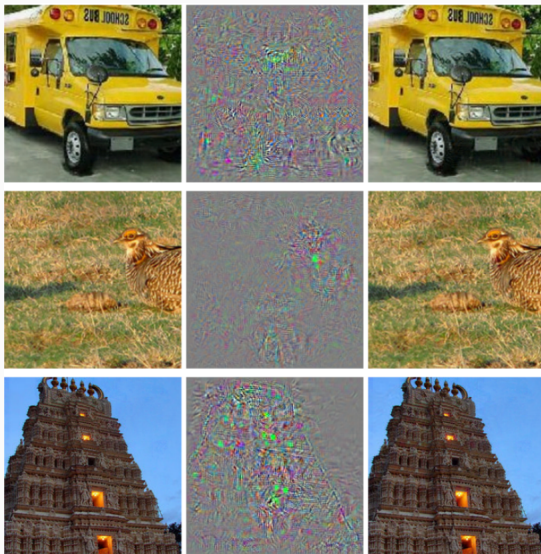
- Denote by  $f : \mathbb{R}^m \rightarrow \{1, \dots, k\}$  a classifier mapping images to labels.
- Assume  $f$  has an associated continuous loss function  $L_f : \mathbb{R}^m \times \{1, \dots, k\} \rightarrow \mathbb{R}^+$ .
- For image  $x \in \mathbb{R}^m$  and target label  $l \in \{1, \dots, k\}$  we formulate an optimization problem:
- Minimize  $\|r\|_2$  subject to:
  1.  $f(x + r) = l$
  2.  $x + r \in [0, 1]^m$ .

# Adversarial examples

Remarks:

- Minimizer  $r$  might not be unique.
- For an arbitrary minimizer denote  $x + r$  by  $D(x, l)$ .
- Informally:  $x + r$  is the image closest to  $x$  classified as  $l$  by  $f$ .
- Problem only non-trivial if  $f(x) \neq l$ .
- $D(x, l)$  approximated by L-BFGS.

# Adversarial examples



Source: [Szegedy et al., 2014]

# Fast gradient sign method

Denote  $\tilde{x} = x + r$ . [Goodfellow et al., 2015] show that for a linear model:

- Activation on adversarial example:  $w^T \tilde{x} = w^T x + w^T r$ , where  $w$  is a weight vector.
- The adversarial perturbation causes the activation to grow by  $w^T r$ .
- This increase can be maximized subject to the constraint  $\|r\|_\infty < \epsilon$  by assigning  $r = \text{sign}(w)$ .
- Assume  $w$  has  $n$  dimensions and the average magnitude of an element of the weight vector is  $m$ , then the activation will grow by  $\epsilon mn$ .
- $\|r\|_\infty$  does not grow linearly in  $n$  but the increase in activation  $\epsilon mn$  is linear in  $n$ .

## Fast gradient sign method

This inspires a method for perturbing a non-linear model:

- Let  $\theta$  be the parameters of the model and  $L(\theta, x, l)$  be the cost function.
- We linearize the cost function around the value of  $\theta$ , obtaining an optimal max-norm constrained perturbation:

$$r = \epsilon \text{sign}(\nabla_x L(\theta, x, l))$$

# Adversarial defenses

An adversarial defense is a method of circumventing adversarial examples. Potential motivations for investigating defenses [Carlini et al., 2019]:

- Defend against an adversary who will attack a system.
- Test the worst-case robustness of a machine learning algorithm.
- Measure progress of machine learning algorithms towards human-level abilities.

A threat model specifies the conditions under which a defense is designed to be secure and the precise security guarantees provided; it is an integral component of the defense itself. Without a threat model, defense proposals are often either not falsifiable or trivially falsifiable [Carlini et al., 2019].

Aspects of a threat model:

1. Adversary goals.
2. Adversarial capabilities.
3. Adversary knowledge.



## Adversary goals

The high-level goal of an adversary can be defined as causing the model to produce erroneous output. The concrete specification will depend on a particular problem. Examples:

- A goal of an adversary might be to induce misclassification, e.g. an example classified as anything but the correct class.
- A different goal would be to misclassify examples from a source class into a target class (*source/target* or *targeted* attack).
- In specific settings, only particular source/target pairs might be interesting, e.g. in security classifying a threat as benign might be the most crucial type of attack.

## Adversarial capabilities

In order to build meaningful defenses, reasonable constraints to need to be imposed on the attacker.

- An unconstrained attacker could, in principle, alter the state of the neural network.
- In more realistic settings, an unconstrained attacker could significantly alter the semantics of the image.
- Defenses usually restrict the attacker to changes of pre-determined magnitude based on a similarity metric  $D$ .
- $l_p$ -norm is a popular choice for  $D$ .
- Restricting perturbations to be small may not always represent a realistic model - e.g. for malware detection, an adversarial program may not care about the magnitude of perturbations.

## Adversary knowledge

The threat model states what kind of knowledge the attacker is assumed to have:

- White-box: complete knowledge of the model and its parameters.
- Black-box: no knowledge of the model or its parameters.
- Black-box access encompasses varying degrees of the ability to query the model and the kind of data the attacker is able to obtain from it (e.g. class predictions vs. class probabilities).
- Grey-box: partial knowledge of the model and its parameters.
- It is not reasonable to assume the defense algorithm can be held secret - Kerckhoffs' principle.
- Even in white-box settings, not all information has to be available to the adversary (e.g. encryption key vs. encryption algorithm).

Perhaps the most important point in evaluating adversarial defenses:

- Once a specific threat model is defined, the evaluation of an adversarial defense should focus on *adaptive adversaries* - adversaries which are adapted to the specific defense and attempt to invalidate it.
- This answers the question: *What attack would break this defense?*
- Showing effectiveness of a defense on a standard attack with default hyperparameters or on any sort of attack not adapted to the specific defense is of very limited utility.
- Any defense should be analyzed from the point of view of an adversary.

## Examples of defenses

Adversarial training [Szegedy et al., 2014]:

- Feed the adversarial examples back to the network in training.
- Improves robustness but is not successful against strong attacks.
- [Goodfellow et al., 2015] apply this in the FGSM setting, training with the following loss:

$$\tilde{L}(\theta, x, l) = \alpha L(\theta, x, l) + (1 - \alpha)L(\theta, x + \epsilon \text{sign}(\nabla_x L(\theta, x, l)), l)$$

- Still, only partially robust against one specific threat model (e.g.  $l_\infty$ ).

Gradient masking [Tramèr et al., 2018]:

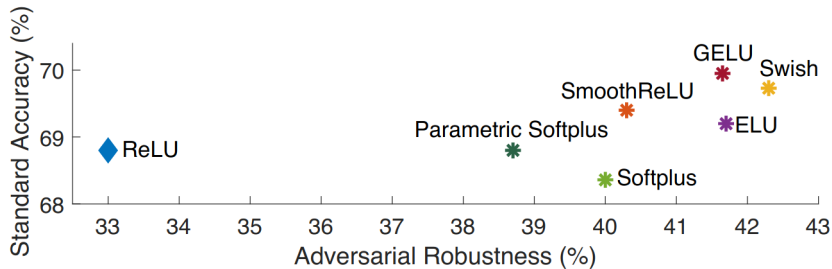
- The defense manipulates the models' gradients and thus prevent gradient-based attacks from succeeding.
- Gradients can still be recovered via black-box input-label queries [Papernot et al., 2017]...
- ... or via a different loss function [Athalye et al., 2018].

## Examples of defenses

Smooth adversarial training [Xie et al., 2020]:

- Based on adversarial training.
- Replace ReLU with smooth activation functions, e.g.:  
 $\text{swish}(x) = x * \text{sigmoid}(x)$ ;  $\text{GELU}(x) = x * \Phi(x)$ , where  $\Phi(x)$  is the CDF of the standard normal distribution, etc.
- Enhances ResNet-50's robustness from 33.0% to 42.3%, while also improving accuracy by 0.9% on ImageNet.
- Helps EfficientNet-L1 achieve 82.2% accuracy and 58.6% robustness on ImageNet, outperforming the previous state-of-the-art defense by 9.5% for accuracy and 11.6% for robustness.

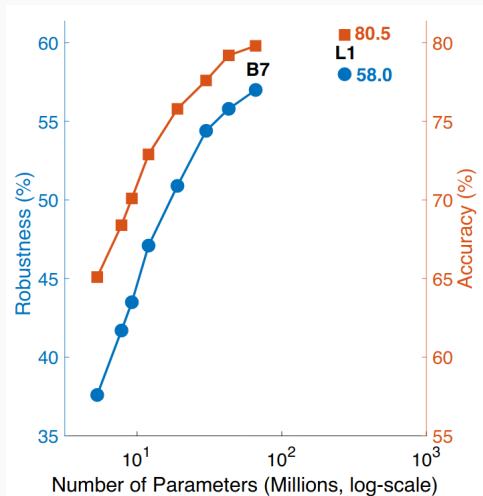
## Examples of defenses



Source: [Xie et al., 2020]



# Examples of defenses



Source: [Xie et al., 2020]

## Examples of defenses

	Accuracy (%)	Robustness (%)
Prior art [30]	72.7	47.0
EfficientNet+SAT	<b>82.2 (+9.5)</b>	<b>58.6 (+11.6)</b>

Source: [Xie et al., 2020]

## Examples of defenses

However:

- Adversarial examples for training generated using PGD-1 with perturbation  $\epsilon = 4$  and attack step size  $\beta = 4$ .
- Adversarial robustness evaluated using PGD-200 with  $\epsilon = 4$  and attack step size  $\beta = 1$ .
- Only one type of attack used both in training and evaluation.
- No obvious signs of adaptation.



Athalye, A., Carlini, N., and Wagner, D. (2018).

**Obfuscated gradients give a false sense of security:  
Circumventing defenses to adversarial examples.**

In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 274–283. PMLR.



Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., Goodfellow, I., Madry, A., and Kurakin, A. (2019).


**On evaluating adversarial robustness.**




Goodfellow, I., Shlens, J., and Szegedy, C. (2015).

**Explaining and harnessing adversarial examples.**


In *International Conference on Learning Representations*.

 Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., and Swami, A. (2017).

**Practical black-box attacks against machine learning.**


 Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2014).

**Intriguing properties of neural networks.**

 Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., and McDaniel, P. (2018).

**Ensemble adversarial training: Attacks and defenses.**

*In International Conference on Learning Representations.*

 Xie, C., Tan, M., Gong, B., Yuille, A., and Le, Q. V. (2020).

**Smooth adversarial training.**