Interpretability and Adversarial Attacks

Ugo Tanielian

March 28, 2022

Outline

Motivation

Attacks

SOTA Defenses

Beyond adversarial robustness

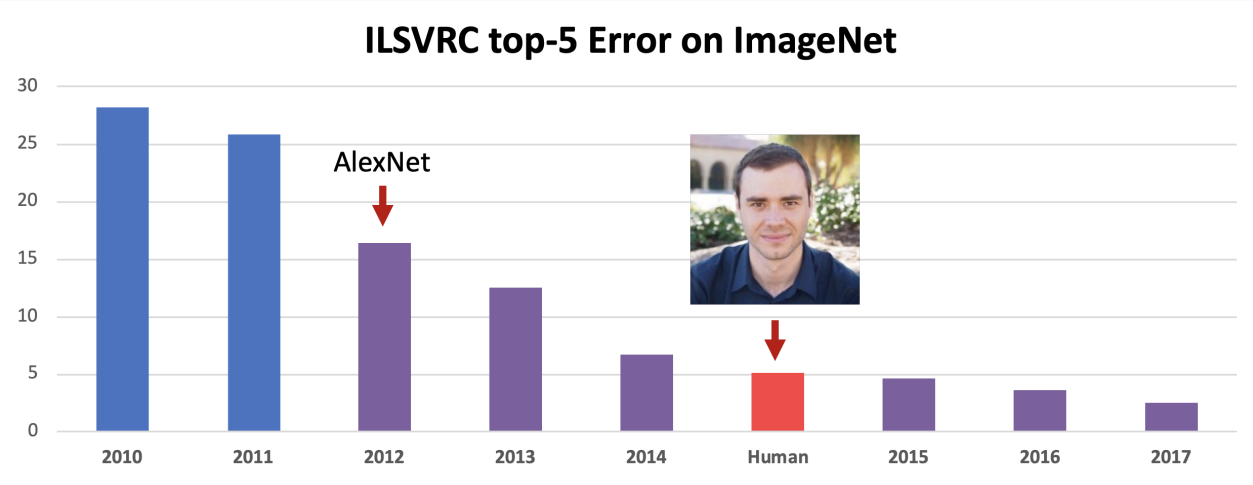
Just for fun

Interesting information

Motivation

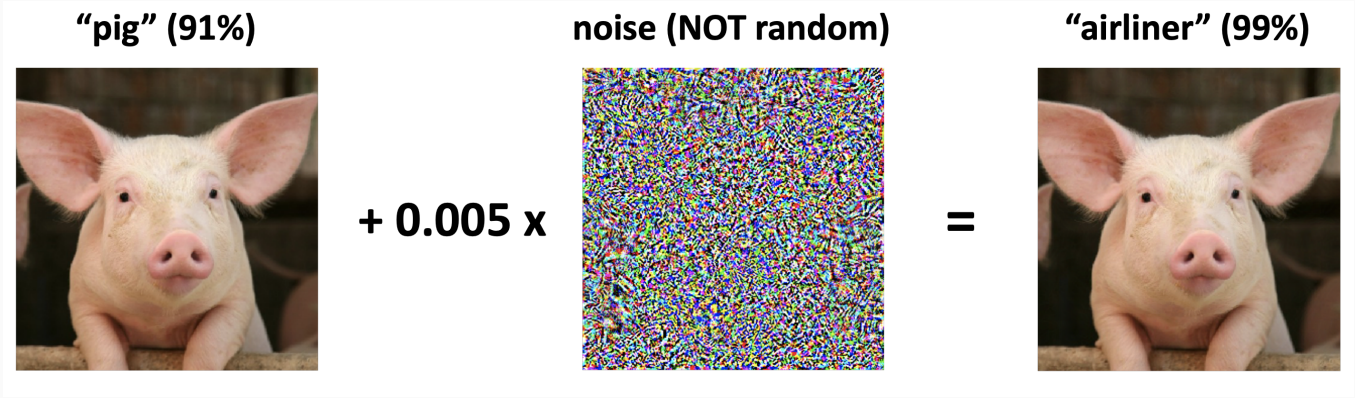
Deep Learning for Images: A sucess story ?

- In the last decade, Deep Learning has achieved great successes in computer vision



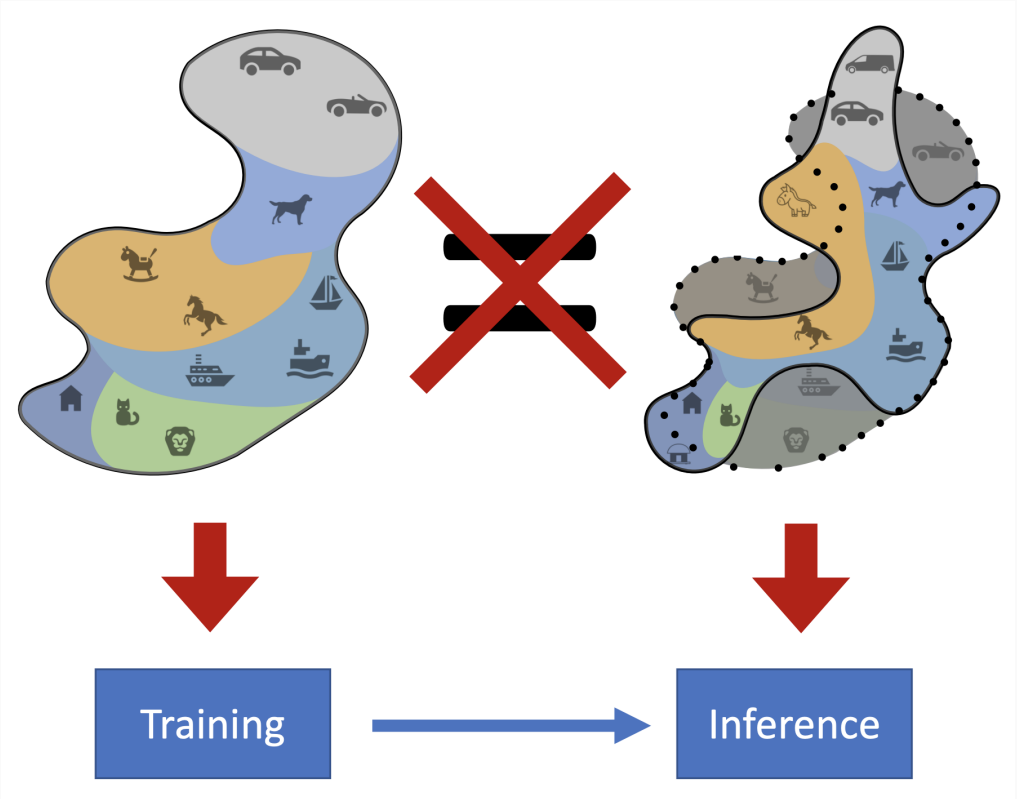
- What does it mean to below the human bias ?
- Are we chasing the right metric ?
- Does it mean we can really trust these models in real environments ? when human safety is at stake ? (e.g. self-driving cars)

Accuracy vs Robustness ?



A generalization / data issue ?

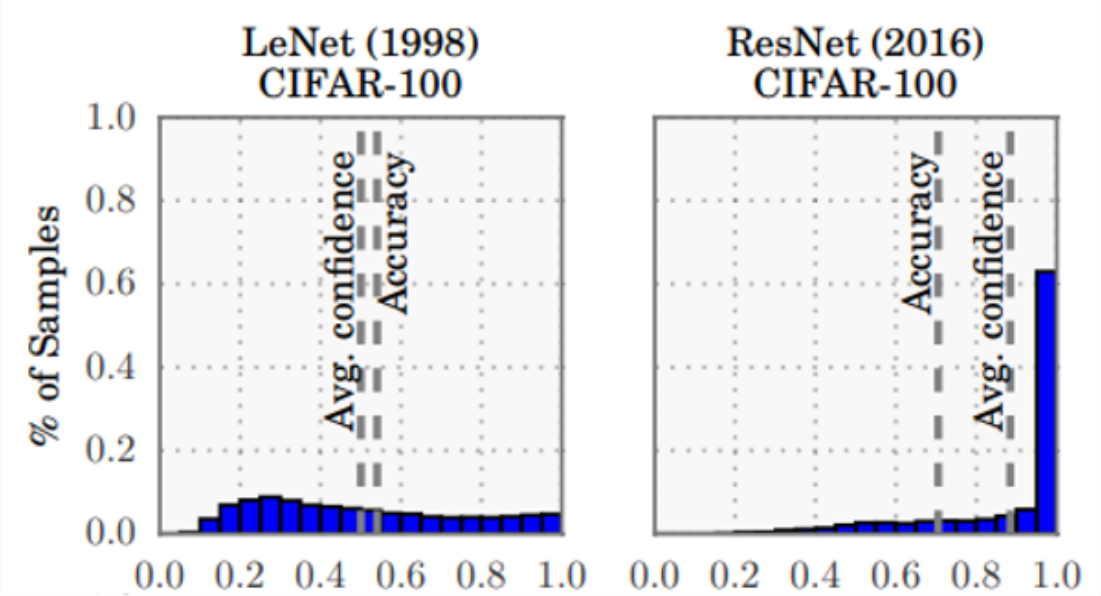
More generally, the assumption that train and test distribution are the same is wrong in general

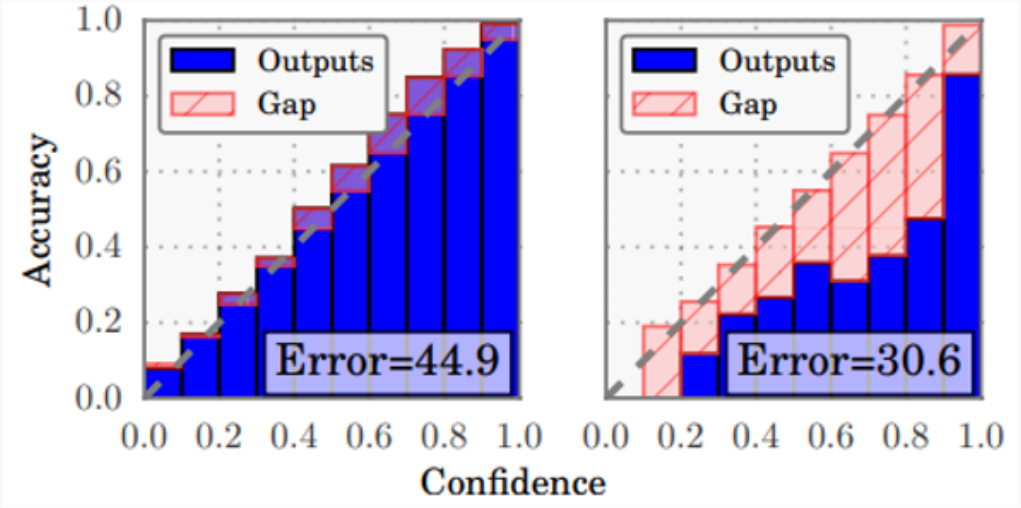


Calibration & Robustness

- Can we trust neural networks ?
- Modern neural networks, unlike those from a decade ago, are poorly calibrated Guo et al. 2017.
- Inputs that are unrobust are more likely to have poorly calibrated predictions Qin et al. 2021.
- Temperature scaling is the simplest, fastest way to remedy the miscalibration phenomenon in neural networks.

Calibration & Robustness: Guo et al. 2017





Even one pixel attacks can work

- The results show that 67.97% of the natural images in Kaggle CIFAR-10 test dataset and 16.04% of the ImageNet (ILSVRC 2012) test images can be perturbed to at least one target class.

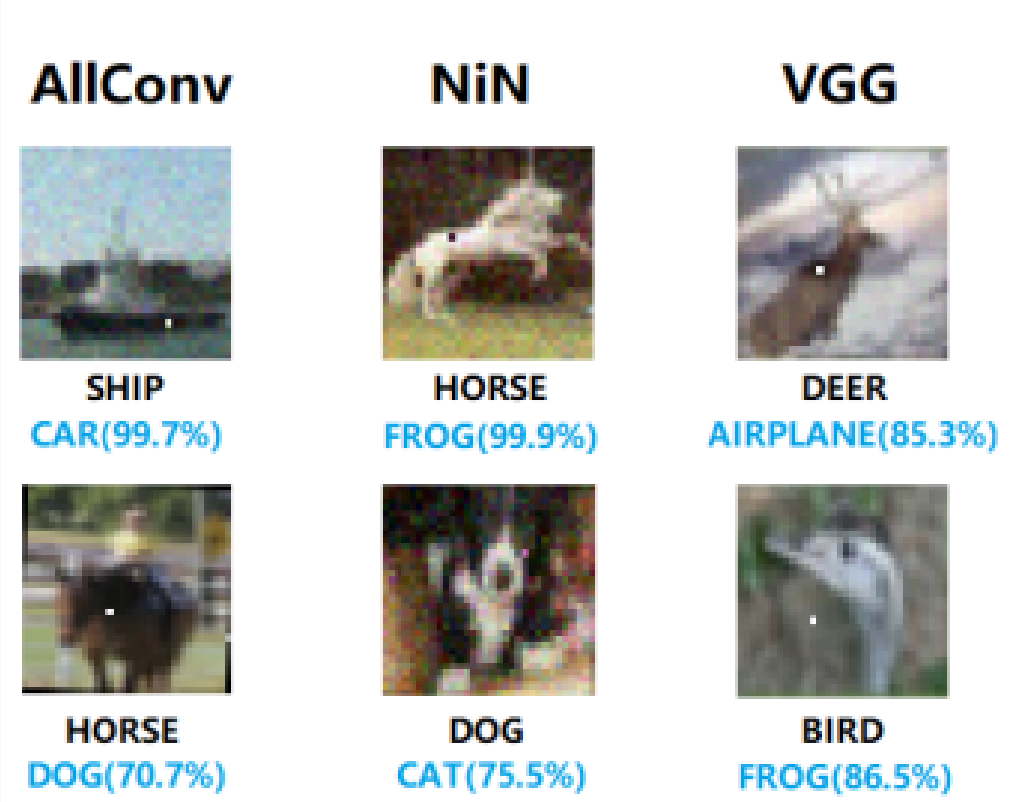


Figure 1: Su et al. 2019

Attacks

Attack onthology

- Poisoning Attack: Contamination during the training phase
  - Data Injection
  - Data Modification
  - Logic Corruption
- Evasion Attack: Malicious samples during testing phase.
  - White Box
  - Black Box
- Exploratory Attack: Gaining knowledge about the algorithm
  - Model inversion
  - Model extraction
  - Inference Attack (data P training set ?)

Poisoning attacks

It is an attack type that takes advantage of your ML model **during training (as opposed to evasion attacks).**

- The goal is to corrupt the training set so that generalization is impacted.
- Poisoning attacks come in two flavors — those targeting your availability or integrity (“backdoor” attacks).
- Backdoor attacks are much more sophisticated. They leave your classifier functioning exactly like it should — with just one exception: a backdoor. A backdoor is a type of input that the model’s designer is not aware of, but that the attacker can leverage Chen et al. 2017.

Poisoning attacks

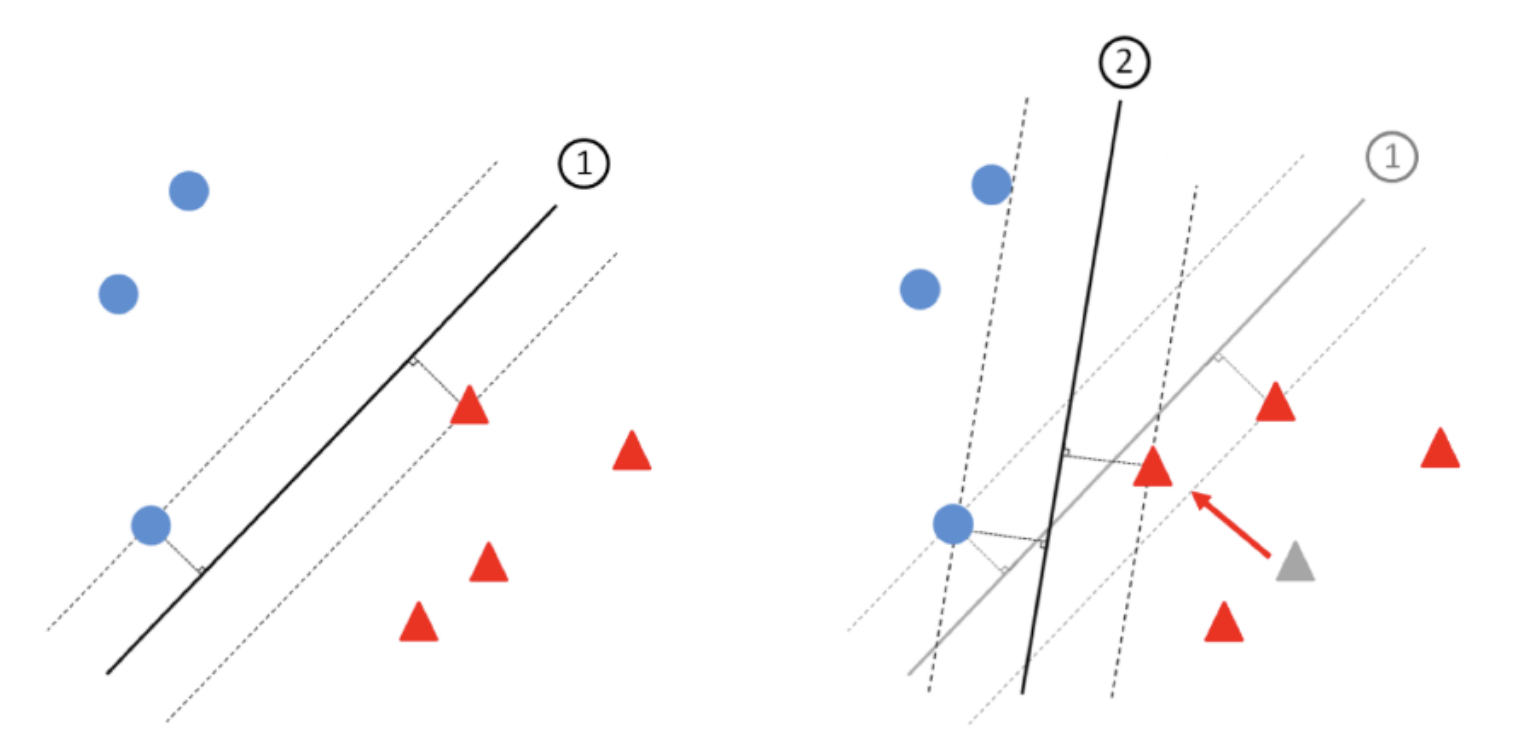
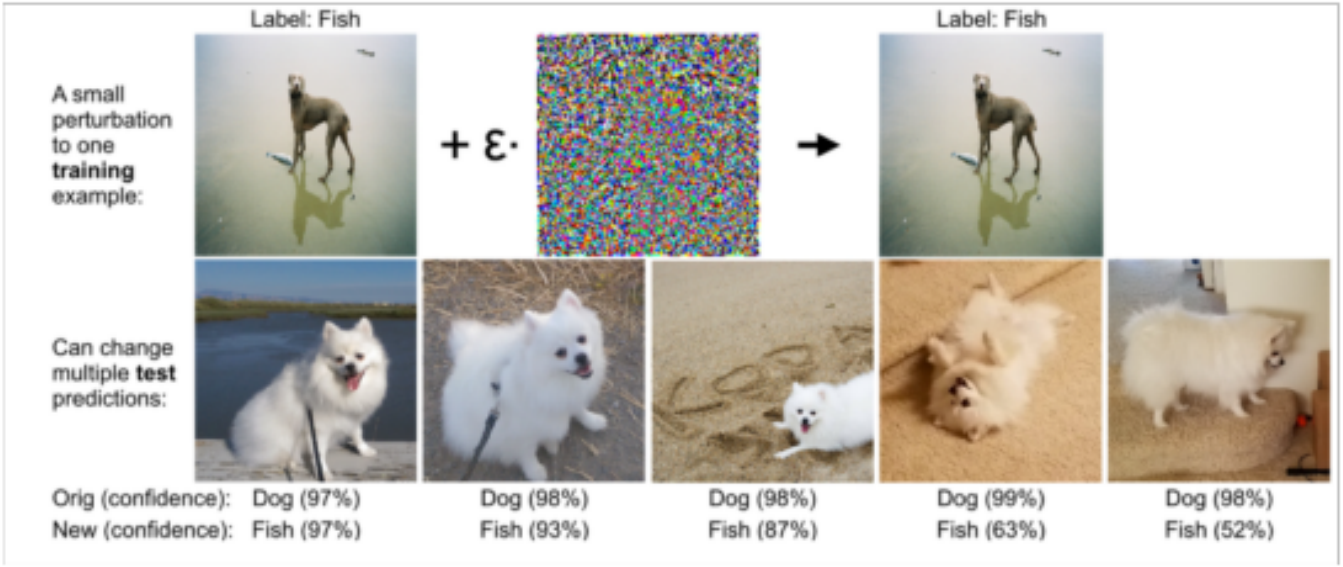


Figure 2: Decision boundary is significantly impacted in this example if just one training sample is changed, even when that sample's class label does not change (right):Miller et al. 2020

Poisoning attacks (2)



[Koh Liang 2017]: Can manipulate **many** predictions with a **single** "poisoned" input

Poisoning defenses

- The most common type of defenses is **outlier detection**, also knows as *"data sanitization" and "anomaly detection"*.
- Sometimes the poison injected is indeed from a different data distribution and can be easily isolated.

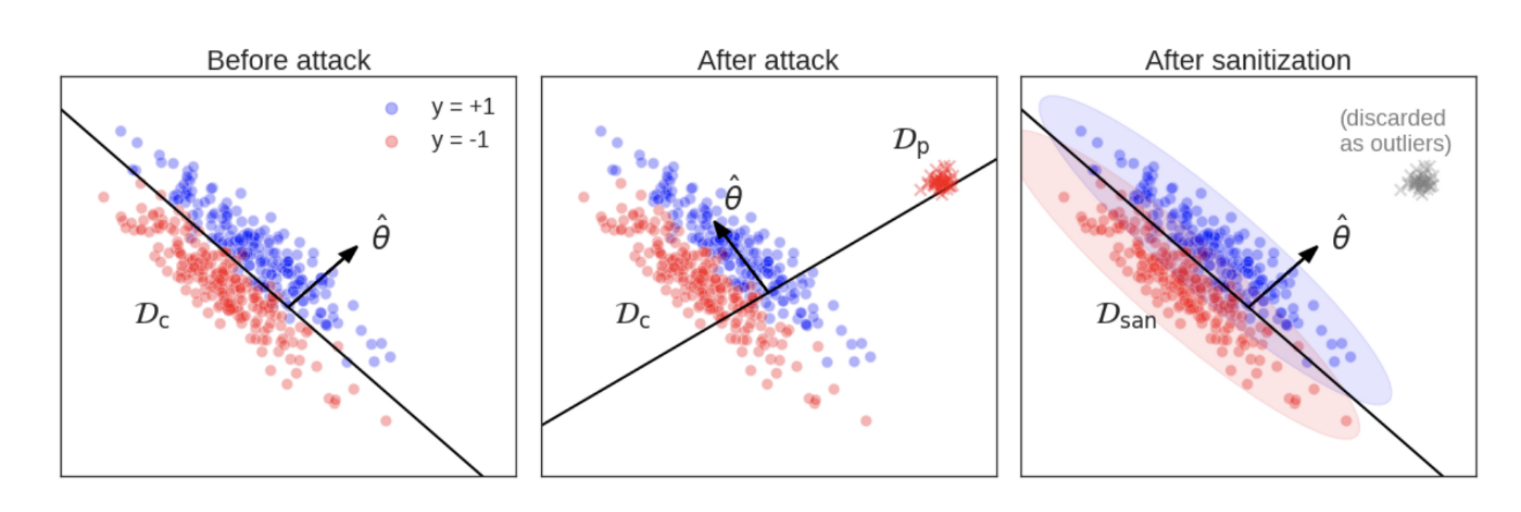


Figure 3: y discarding outliers from $D = D_c \bigcup D_p$: Koh et al. 2021

Evasion attacks

An **evasion attack** happens when the network is fed an "adversarial example" — a carefully perturbed input that looks and feels exactly the same as its untampered copy to a human — but that completely throws off the classifier.

All models can be attacked !

- Video: Adversarial boxes
- Audio: Audio adversiarial examples

Why a model can be attacked ?

1. *Szegedy*: the presence of low-probability “pockets” in the manifold (ie too much non-linearity) and poor regularization of networks.
2. *Goodfellow*: too much linearity in modern machine learning and especially deep learning systems
3. *The tilted boundary*: networks do not fit data perfectly (or lack training samples): there are adversarial pockets of inputs that exist between the boundary of the classifier and the sub-manifold of sampled data. (+ criticism of 1 and 2)
4. *Adversarial Examples Are Not Bugs, They Are Features*: humans are limited to 3 dimensions and can’t distinguish noise patterns from one another. Networks are a pattern recognition machine more sophisticated than ourselves.
5. Link with High frequencies ?

Evasion attacks

- Happens at inference time.
- Usually find small perturbation on an input such that the confidence or the prediction changes.
- Black box (the attacker to know anything about the model) vs White box (requires access to the model).

Description	Black box attack	White box attack
Adversary Knowledge	Restricted knowledge from being able to only observe the networks output on some probed inputs.	Detailed knowledge of the network architecture and the parameters resulting from training
Attack Strategy	Based on a greedy local search generating an implicit approximation to the actual gradient w.r.t the current output by observing changes in input	Based on the gradient of the network loss function w.r.to the input

(Papernot, McDaniel, and I. Goodfellow 2016)

Figure 4: Adversarial attacks: Towards Deep Learning Models Resistant to Adversarial Attacks (2017).

What is an adversarial attack ?

$$\min_{\theta} \rho(\theta), \quad \text{where} \quad \rho(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta \in \mathcal{S}} L(\theta, x + \delta, y) \right] .$$

Figure 5: Adversarial attacks: Towards Deep Learning Models Resistant to Adversarial Attacks (2017).

It is a worst-case mindset/scenario.

SOTA Attacks

1. FGSM
2. BIM
3. Iterative Least Likely Method
4. DeepFool
5. CW (Carlini and Wagner 2017)

FGSM (Fast Gradient Sign Method)

- Introduced in I. J. Goodfellow et al. 2014.
- Main idea: compute the sign of the gradient ∇ of the loss wrt to each pixel of the input image.
- Move in the opposite direction of ∇ by a step of size ε.
- FGSM increases the cost function with the correct label, hoping that this will be enough to change the prediction.
- We obtain a perturbation of size ε in $\|.\|_\infty$.

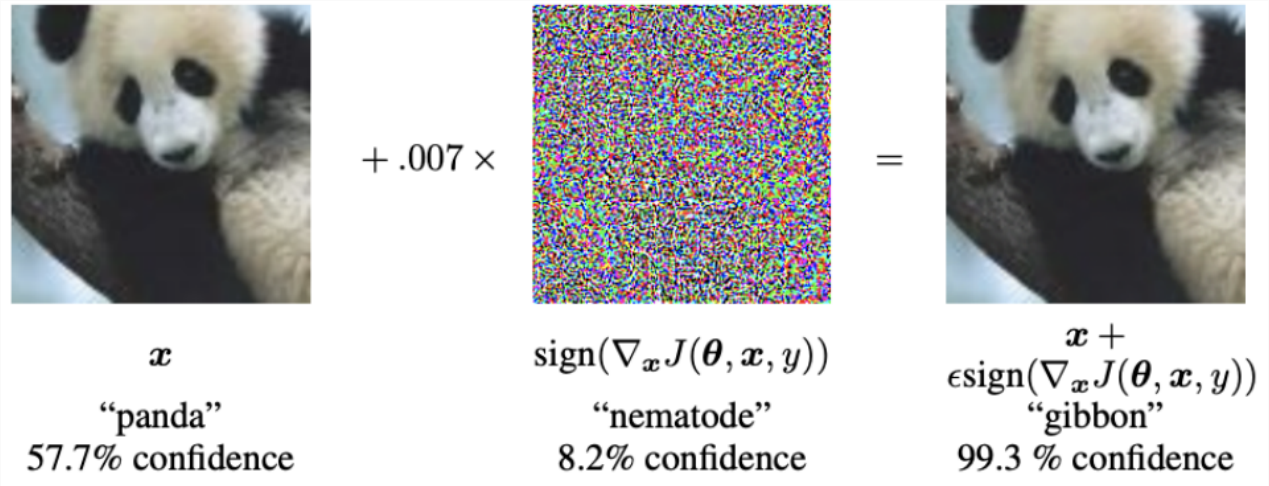
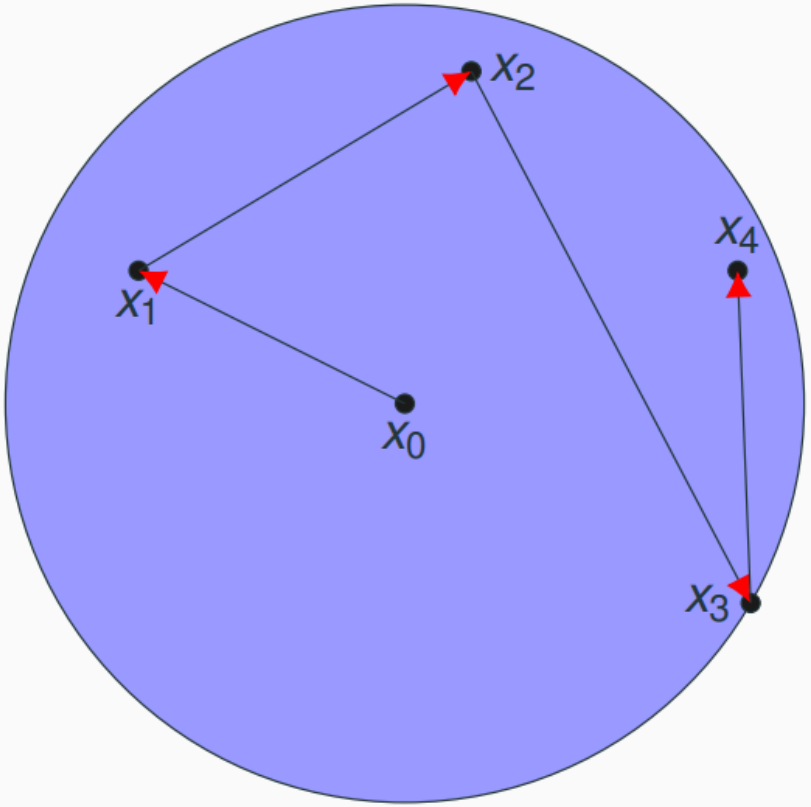


Figure 6: FGSM: Explaining and Harnessing Adversarial Examples (2014).

BIM (Basic Iterative Method)

Main idea: Apply FGSM several times while ensuring that we stay in an ε-ball around the original image w.r.t. the $\|.\|_\infty$ norm.



Iterative Least Likely Method

- Both of the previous methods are untargeted attacks.
- By changing the BIM algorithm to alter the image towards a specific target class, it yields the Iterative Gradient Sign Method.
- Now, we target the Least Likely class, to give an idea on the worst case scenario.

DeepFool

- The DeepFool algorithm searches for an adversary with the smallest possible perturbation.
- The algorithm tries to shift the image towards the closest decision boundary.

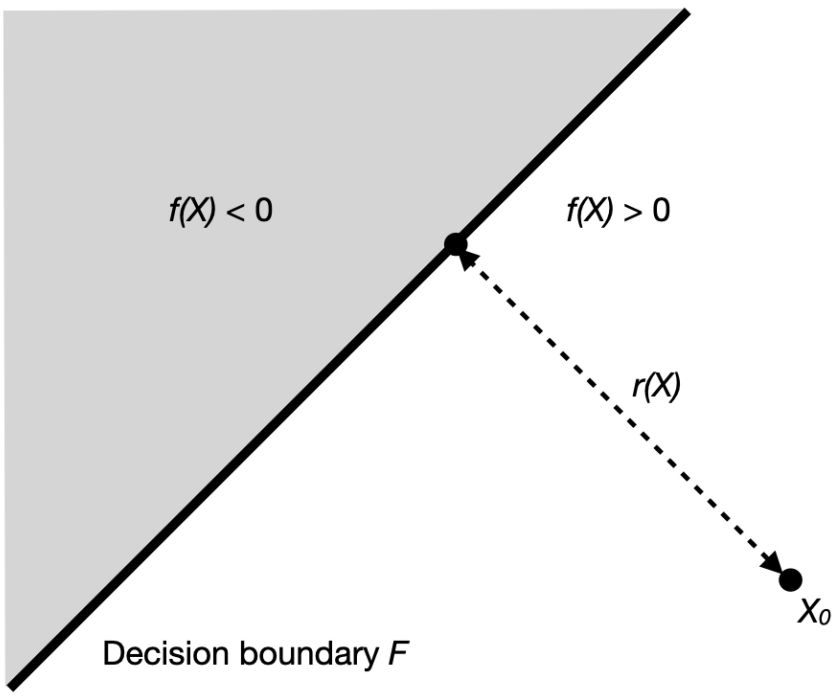


Figure 7: DeepFool for a linear, binary classifier. From Moosavi-Dezfooli et al. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks (2016).

Some examples

criteo AI Lab

NeuroCeption

SOTA Defenses

Data augmentation with adversarial examples

criteo AI Lab

- A simple but yet effective way to defend against attacks is to add attacked images to the training set.
- It is attack specific: cumbersome process.
- Findings: FGSM adversaries don't increase robustness (for large ε): that the network overfits to these adversarial examples.

Other theoretical questions

- Standard image distribution lay on low dimension manifold (the manifold hypothesis) Fefferman et al. 2016.
- Sample complexity of adv. robust generalization can be significantly larger than that of "standard" generalization.
- Adversarially Robust Generalization Requires More Data Schmidt et al. 2018.

Defensive distillation

- Another solution proposed in Papernot et al. 2016 is based on **knowledge distillation**.
- Main idea is to transfer knowledge from a teacher model to a student model (Hinton et al. 2015).

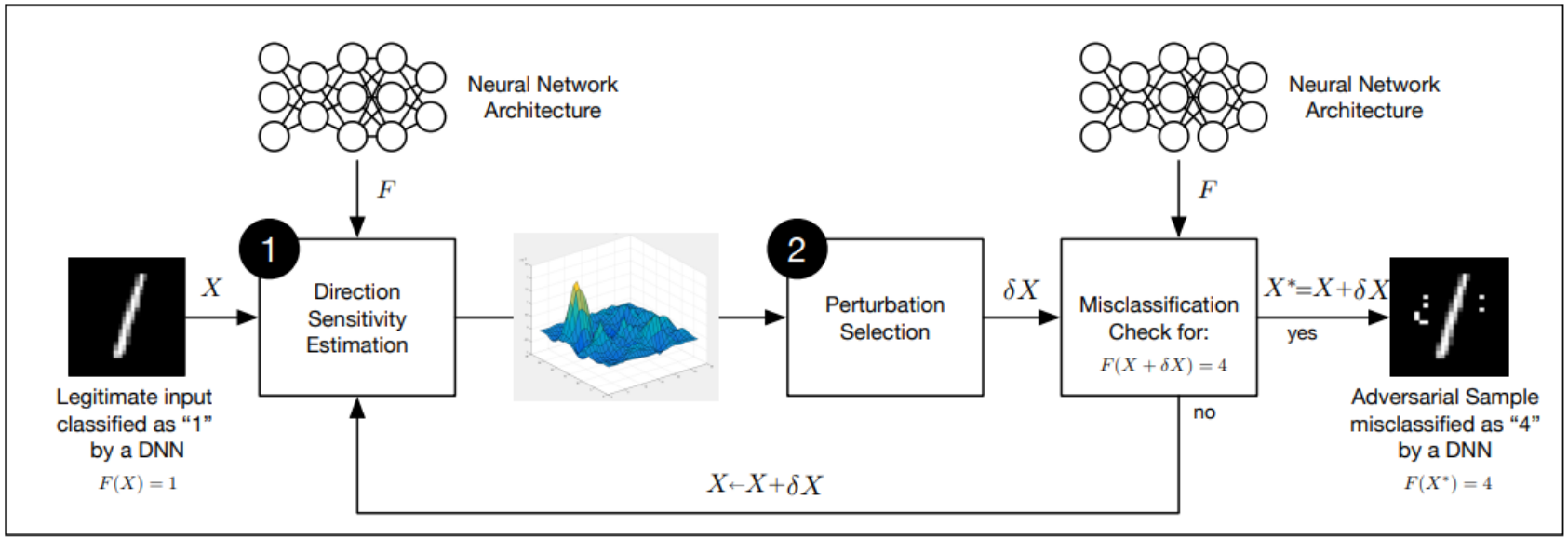


Fig. 3: **Adversarial crafting framework:** Existing algorithms for adversarial sample crafting [7], [9] are a succession of two steps: (1) *direction sensitivity estimation* and (2) *perturbation selection*. Step (1) evaluates the sensitivity of model F at the input point corresponding to sample X. Step (2) uses this knowledge to select a perturbation affecting sample X's classification. If the resulting sample $X + \delta X$ is misclassified by model F in the adversarial target class (here 4) instead of the original class (here 1), an adversarial sample X^* has been found. If not, the steps can be repeated on updated input $X \leftarrow X + \delta X$.

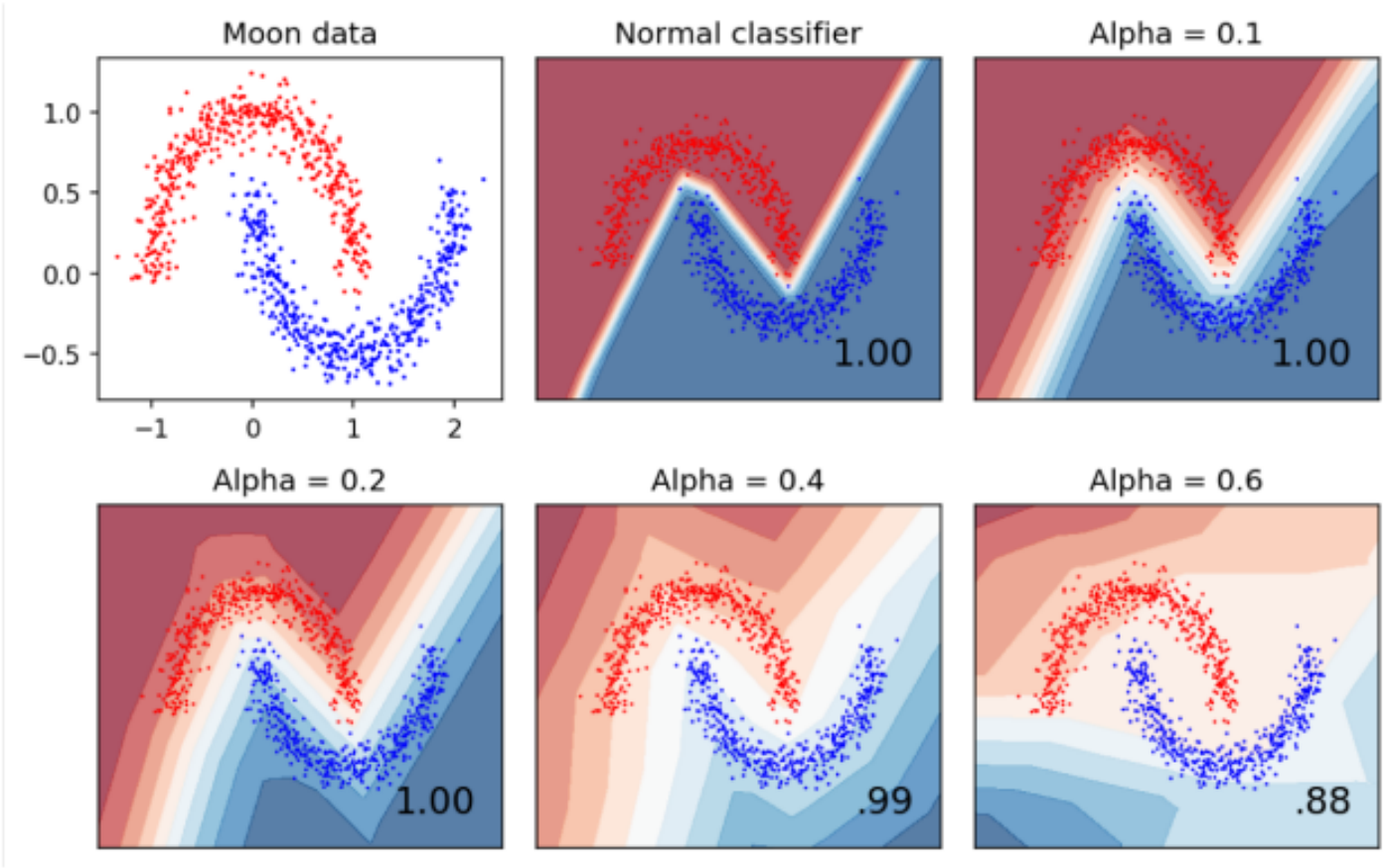
Gradient penalty

There is a connection between robustness and regularizing the gradient of the network Bietti et al. 2018.

How can we implement this regularization ?

- Clipping
- A gradient penalty
- Spectral normalization

Label smoothing



(a) **Regularization effect:** logit squeezing using ALS (different α) and a MLP classifier. Darker is more confidence.

Figure 8: Regularization effect of LS Goibert and Dohmatob 2019.

Last but not least: Distributional robustness

- We know that many problem arise from doing pure **Empirical Risk Minimization**.
- One way to circumvent this limitation is to treat the empirical distribution μ_n with skepticism and to replace it with an uncertainty set $\mathcal{U}(\mu_n)$ of distributions around μ_n.
- This gives rise to the distributionally robust obejctive Blanchet, Kang, Murthy, and Zhang 2019; Blanchet, Kang, and Murthy 2019:

$$\tilde{R}_n^{\mathcal{U}}(\theta, \epsilon) \triangleq \max_{Q \in \mathcal{U}_\epsilon(\hat{P}_n)} \mathbb{E}_{\xi \sim Q}[\ell(\xi; \theta)].$$

Minimizing this quantity w.r.t to θ yields the general program:

$$\begin{aligned}\tilde{\theta}_n &\triangleq \operatorname{argmin}_{\theta \in \Theta} \tilde{R}_n^{\mathcal{U}}(\theta, \epsilon) \\ &= \operatorname*{argmin}_{\theta \in \Theta} \max_{Q \in \mathcal{U}_\epsilon(\hat{P}_n)} \mathbb{E}_{\xi \sim Q}[\ell(\xi; \theta)].\end{aligned}$$

There is liberty on the way to construct $U_\varepsilon(\hat{P}_n)$.

$$\mathbb{E}_{\xi \sim P}[\ell(\xi; \theta)] \leq \tilde{R}_n^{\mathcal{U}}(\theta, \epsilon_n) \text{ w.h.p}$$

$$\mathbb{E}_{\xi \sim P}[\ell(\xi; \theta)] - \tilde{R}_n^{\mathcal{U}}(\theta, \epsilon_n) \to 0$$

GANs for robustness

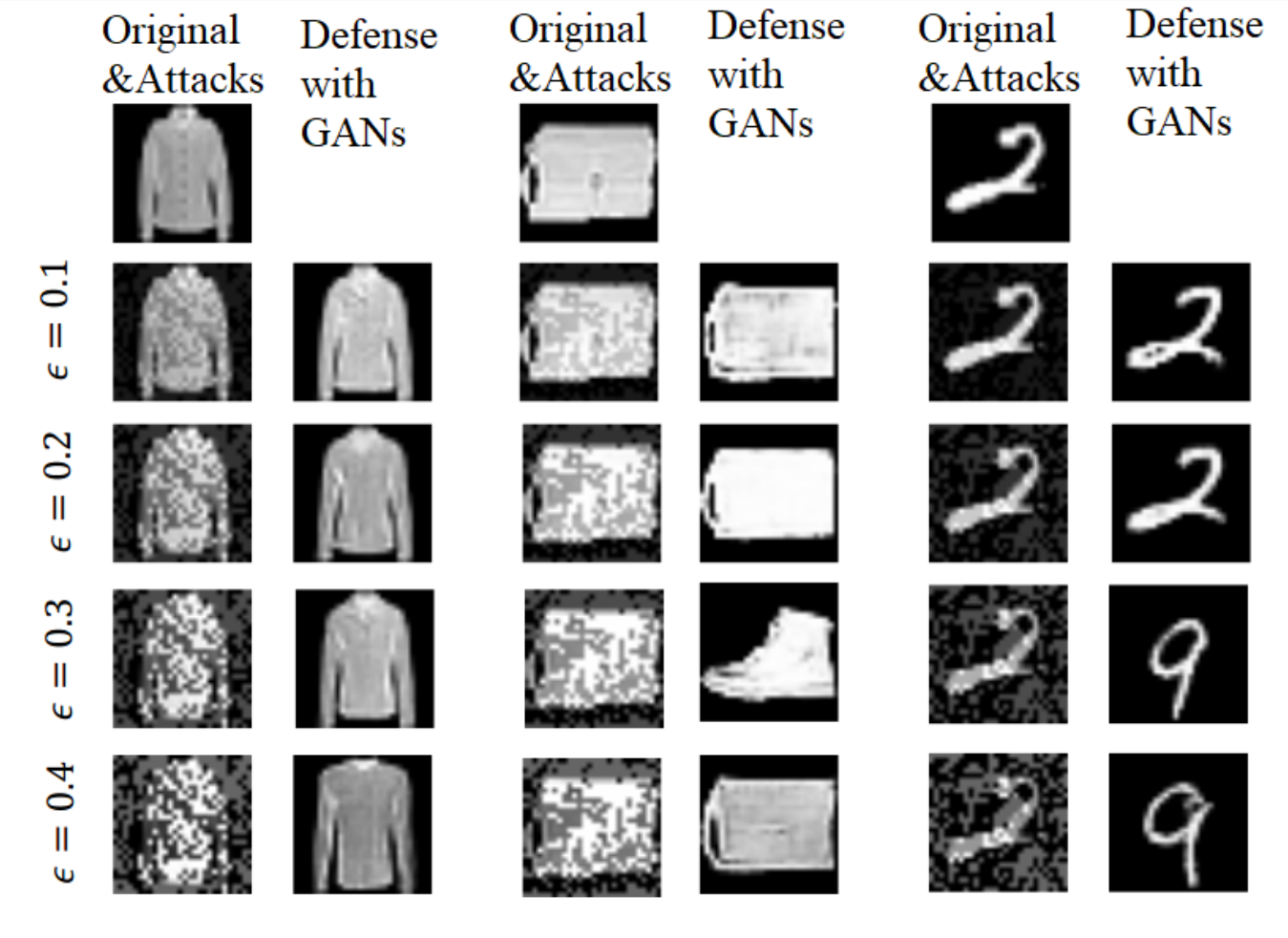


Figure 9: Defending deep nets with GANs: Samangouei et al. 2018.

Beyond adversarial robustness

Robustness & accuracy

- Can we get both robustness and accuracy ?
- We could think that a robust model will also generalize better.
- Counter-example found by Tsipras et al. 2018, where the authors exhibit a dataset where you cannot be both accurate and robust at the same time.

Theorem

On the above dataset, any classifier that attains at least $1 - \delta$ *standard accuracy has robust accuracy at most* $\frac{p\delta}{a-p}$ *against an* $\|.\|_\infty$*-bounded adversary.*

No free lunch theorem ?

- Understanding the tradeoff between accuracy and robustness is a very active line of research.
- See for instance the strong "no free lunch" theorem from Dohmatob 2018 "on a very broad class of data distributions, any classifier with even a bit of accuracy is vulnerable to adversarial attacks".

Just for fun

Generating images with robust network

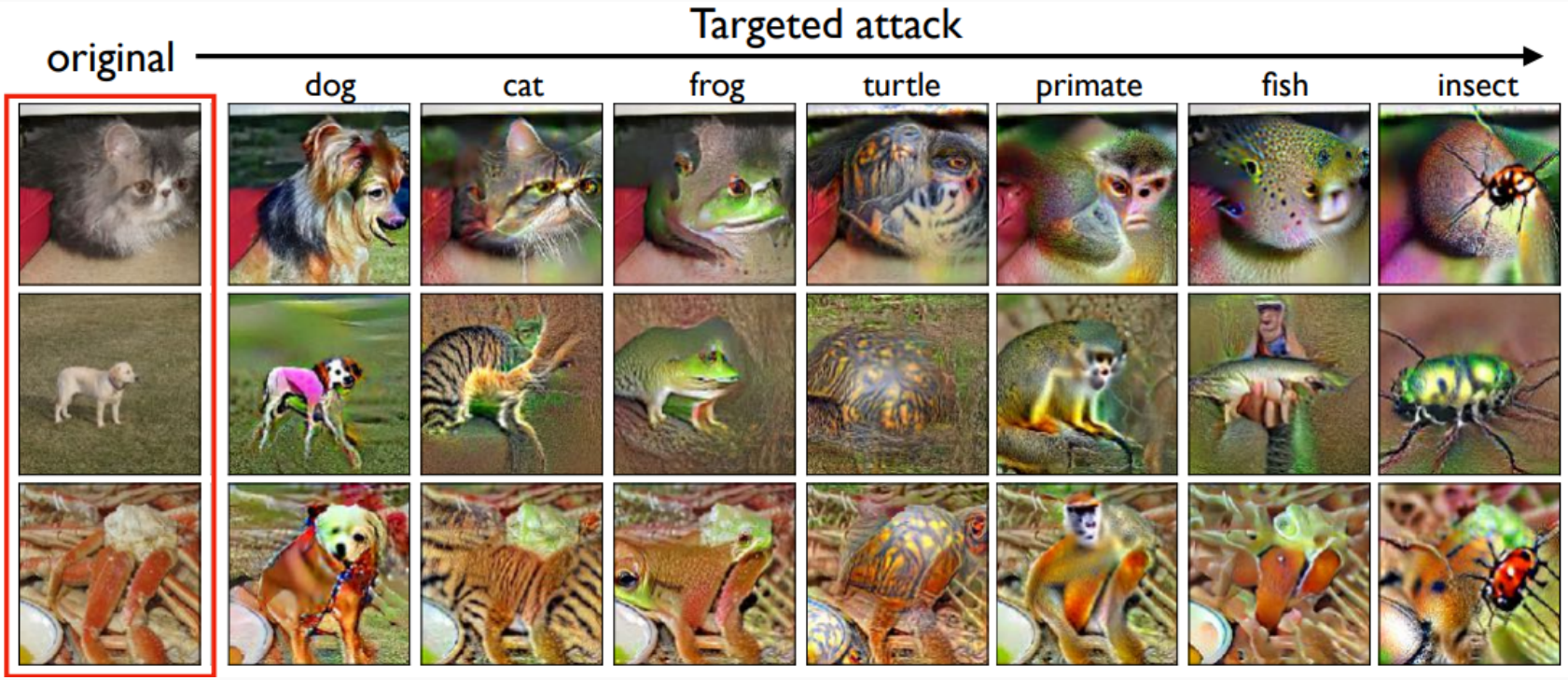


Figure 10: Santurkar et al. 2019

Generating images with robust network (2)

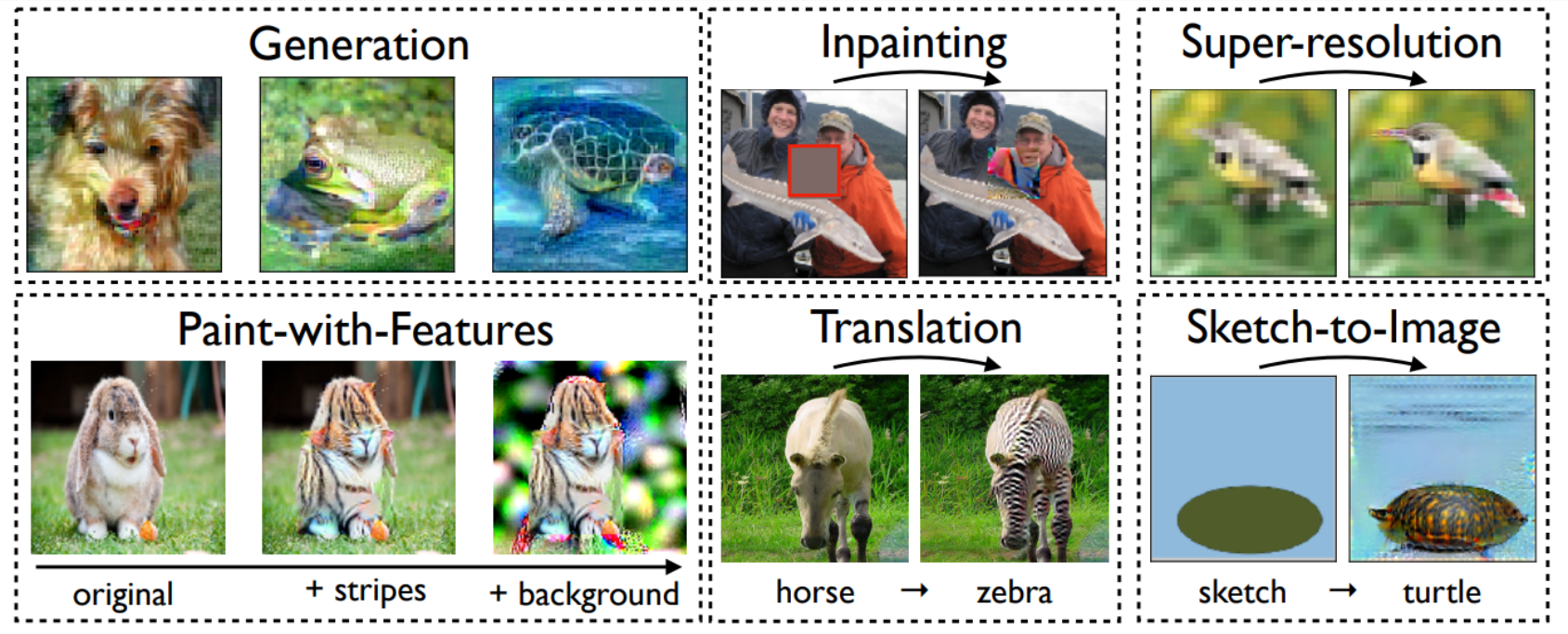


Figure 11: Santurkar et al. 2019

Interesting information

Robustness from the Madry lab

Robustness package: one can

- Train and evaluate standard and robust models on a variety of datasets/architectures.
- Import pre-trained robust models.

Adversarial Robustness Toolbox

criteo AI Lab

- Adversarial Robustness Toolbox (ART) is a Python library for Machine Learning Security.
- ART provides tools that enable developers and researchers to evaluate, defend, certify and verify Machine Learning models and applications against the adversarial threats of Evasion, Poisoning, Extraction, and Inference.

A survey on robustness

criteo AI Lab

Articles	Attacks	Applications
Fredrikson et al. [26]	Model Inversion	Biomedical Imaging, biometric identification
Tramèr et al. [73]	Extraction of target machine learning models using APIs	Attacks extend to multiclass classifications & neural networks
Anteniese et al. [11]	Meta-classifier to hack other classifiers	Speech Recognition
Biggio et al. [19], [20]	Poisoning based attacks:	Crafted training data for Support vector Machines
Dalvi et al. [24] Biggio et al. [16], [15]	Adversarial Classification, Pattern recognition	Email Spam detection, fraud detection, intrusion detection, biometric identification
Papernot et al. [60], [57]	Adversarial samples crafting, adversarial sample transferability	digit recognition, black-box attacks against classifiers hosted by Amazon and Google
Hitaj et al. [34]	GAN under collaborative learning	Classification
Goodfellow et al. [30]	Generative Adversarial Network	Classifiers, Malware Detection
Shokri et al. [67]	Membership inference attack	Attack on classification models trained by commercial "ML as a service" providers such as Google and Amazon
Moosavi et al. [52] Carlini et al. [22] Li et al. [45]	Adversarial perturbations: and sample generation: Poisoning based attack	Image classification intrusion detection Collaborative filtering systems

Table 2. Overview of Attacks and Applications

Figure 12: Chakraborty et al. 2018

Conclusion

Three commandments of Secure/Safe ML

1. You shall not train on data you don't fully trust (because of data poisoning).
2. You shall not let anyone use your model (or observe its outputs) unless you completely trust them (because of model stealing and black box attacks).
3. You shall not fully trust the predictions of your model (because of adversarial examples)

References

References

- Bietti, Alberto et al. (2018). “On regularization and robustness of deep neural networks”. In.
- Blanchet, Jose, Yang Kang, and Karthyek Murthy (2019). “Robust Wasserstein profile inference and applications to machine learning”. In: Journal of Applied Probability 56.3, pp. 830–857.
- Blanchet, Jose, Yang Kang, Karthyek Murthy, and Fan Zhang (2019). “Data-driven optimal transport cost selection for distributionally robust optimization”. In: 2019 winter simulation conference (WSC). IEEE, pp. 3740–3751.

References ii

- Carlini, Nicholas and David Wagner (2017). “Towards evaluating the robustness of neural networks”. In: 2017 ieee symposium on security and privacy (sp). IEEE, pp. 39–57.
- Chakraborty, Anirban et al. (2018). “Adversarial attacks and defences: A survey”. In: arXiv preprint arXiv:1810.00069.
- Chen, Xinyun et al. (2017). “Targeted backdoor attacks on deep learning systems using data poisoning”. In: arXiv preprint arXiv:1712.05526.
- Dohmatob, Elvis (2018). “Limitations of adversarial robustness: strong no free lunch theorem”. In: arXiv preprint arXiv:1810.04065.
- Fefferman, Charles et al. (2016). “Testing the manifold hypothesis”. In: Journal of the American Mathematical Society 29.4, pp. 983–1049.
- Goibert, Morgane and Elvis Dohmatob (2019). “Adversarial robustness via label-smoothing”. In: arXiv preprint arXiv:1906.11567.
- Guo, Chuan et al. (2017). “On calibration of modern neural networks”. In: International Conference on Machine Learning. PMLR, pp. 1321–1330.

References iii

- Koh, Pang Wei et al. (2021). “Stronger data poisoning attacks break data sanitization defenses”. In: Machine Learning, pp. 1–47.
- Miller, David J et al. (2020). “Adversarial learning targeting deep neural network classification: A comprehensive review of defenses against attacks”. In: Proceedings of the IEEE 108.3, pp. 402–433.
- Papernot, Nicolas et al. (2016). “Distillation as a defense to adversarial perturbations against deep neural networks”. In: 2016 IEEE Symposium on Security and Privacy (SP). IEEE, pp. 582–597.
- Qin, Yao et al. (2021). “Improving calibration through the relationship with adversarial robustness”. In: Advances in Neural Information Processing Systems 34.
- Samangouei, Pouya et al. (2018). “Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models”. In: International Conference on Learning Representations.

References iv

- Santurkar, Shibani et al. (2019). “Image synthesis with a single (robust) classifier”. In: Advances in Neural Information Processing Systems 32.
- Schmidt, Ludwig et al. (2018). “Adversarially robust generalization requires more data”. In: Advances in neural information processing systems 31.
- Su, Jiawei et al. (2019). “One pixel attack for fooling deep neural networks”. In: IEEE Transactions on Evolutionary Computation 23.5, pp. 828–841.
- Tsipras, Dimitris et al. (2018). “Robustness may be at odds with accuracy”. In: arXiv preprint arXiv:1805.12152.