FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
FACULTY OF ENGINEERING

Ferienakademie 2018 - Sarntal

Adversarial examples

September 27th, 2018

Anatol Maier

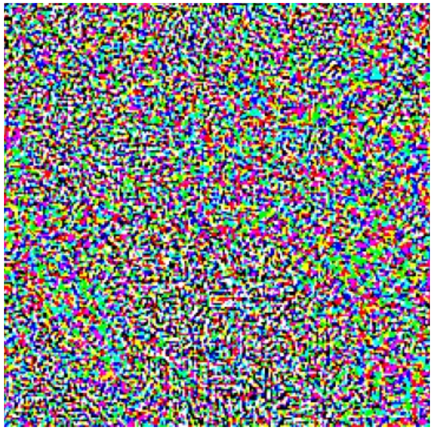
What are adversarial examples?

What are adversarial examples?

- small but intentionally worst-case perturbations applied to input data
- Perturbated input results in the model outputting an incorrect answer with high confidence

What are adversarial examples?

+ 0.07 x



=

x

η

$x + \epsilon\eta$

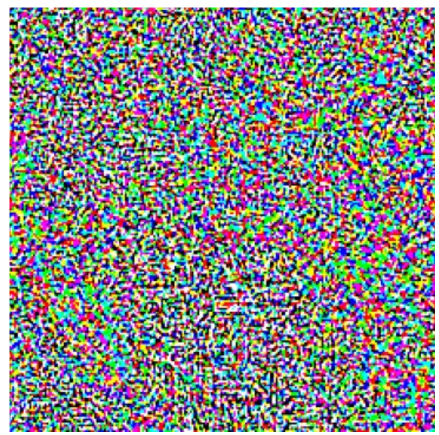
„panda“
57.7% confidence

„gibbon“
99.3% confidence

[2]

What are adversarial examples?

+ 0.07 x



=

x

η

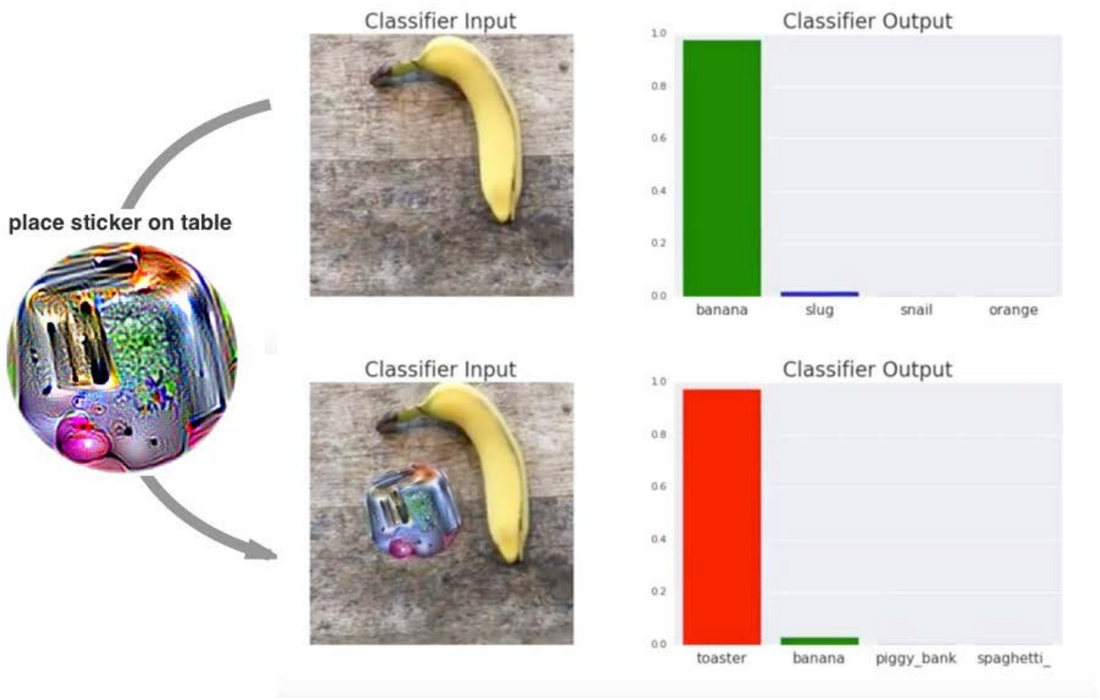
$x + \epsilon \eta$

„panda“
57.7% confidence

„gibbon“
99.3% confidence

[2]

What are adversarial examples?



Adversarial patch [3]

How to generate adversarial examples?

How to generate adversarial examples?

- Fast gradient sign method

How to generate adversarial examples?

- Fast gradient sign method

$$J(\Theta, x, y)$$

How to generate adversarial examples?

- Fast gradient sign method

$$J(\Theta, x, y)$$

$$\nabla_x J(\Theta, x, y)$$

How to generate adversarial examples?

- Fast gradient sign method

$$J(\Theta, x, y)$$

$$\nabla_x J(\Theta, x, y)$$

$$\eta = sign(\nabla_x J(\Theta, x, y))$$

How to generate adversarial examples?

– Fast gradient sign method

$$J(\Theta, x, y)$$

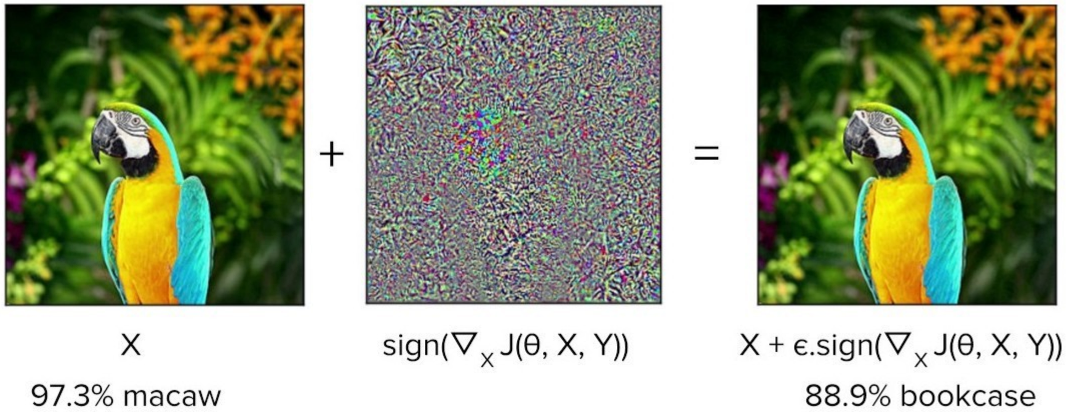
$$\nabla_x J(\Theta, x, y)$$

$$\eta = sign(\nabla_x J(\Theta, x, y))$$

$$\tilde{x} = x + \epsilon \eta$$

How to generate adversarial examples?

– Fast gradient sign method



- Most adversarial example techniques use the gradient of the model to make an attack.

- Most adversarial example techniques use the gradient of the model to make an attack.

- But what if there were no gradient?

- what if an infinitesimal modification to the image caused no change in the output of the model?

„gradient masking“ [4]

- Most adversarial example techniques use the gradient of the model to make an attack.

- But what if there were no gradient?

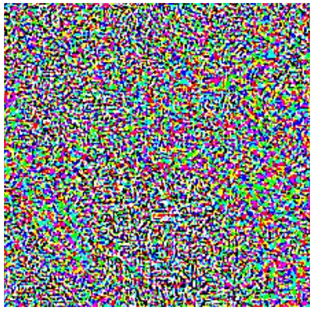
- what if an infinitesimal modification to the image caused no change in the output of the model?

„gradient masking“ [4]

Airplane = 99.9%
Cat = 0.1%

„gradient masking“ [4]

+



=

Airplane = 99.9%
Cat = 0.1%

Airplane = 99.8%
Cat = 0.2%

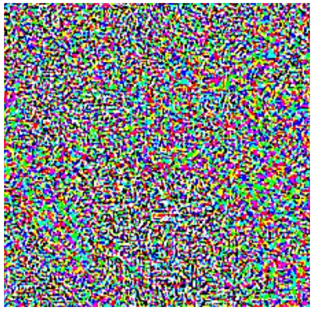
„gradient masking“ [4]

Airplane

„gradient masking“ [4]

Airplane

+



=

Airplane

„gradient masking“ [4]

„gradient masking“ [4]

- The attacker can train their own model and make adversarial examples for their model

„gradient masking“ [4]

- The attacker can train their own model and make adversarial examples for their model

- Very often, our model will misclassify these examples too.

„gradient masking“ [4]

- The attacker can train their own model and make adversarial examples for their model
- Very often, our model will misclassify these examples too.
- In the end, hiding the gradient didn’t get us anywhere.

A failed defense: „gradient masking“ [4]

- The attacker can train their own model and make adversarial examples for their model
- Very often, our model will misclassify these examples too.
- In the end, hiding the gradient didn’t get us anywhere.

Defenses against adversarial attacks

- Traditional techniques for making machine learning models more robust generally do not provide a practical defense against adversarial examples.

- So far, only two methods have provided a significant defense.

Defenses against adversarial attacks

- Adversarial training

Defenses against adversarial attacks

- Adversarial training
  - Use adversarial examples for training

Defenses against adversarial attacks

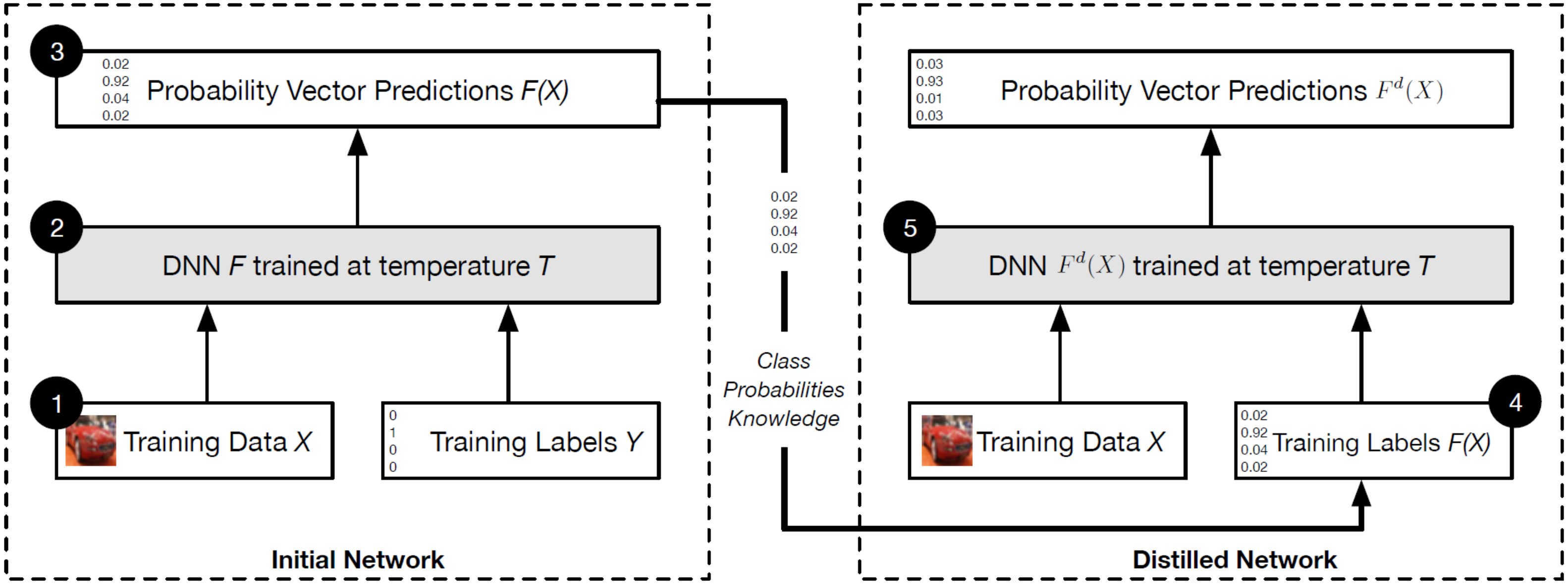
- defensive distillation

Defenses against adversarial attacks

- defensive distillation
  - train the model to output probabilities of different classes, rather than hard decisions about which class to output.
  - The probabilities are supplied by an earlier model, trained on the same task using hard class labels.

Defenses against adversarial attacks

- defensive distillation



[5]

Why is it hard to defend against adversarial examples?

Why is it hard to defend against adversarial examples?

- It’s difficult to construct a theoretical model of the adversarial example crafting process.

Why is it hard to defend against adversarial examples?

- It’s difficult to construct a theoretical model of the adversarial example crafting process.
- Adversarial examples require machine learning models to produce good outputs *for every possible input*.

Why is it hard to defend against adversarial examples?

- It's difficult to construct a theoretical model of the adversarial example crafting process.
- Adversarial examples require machine learning models to produce good outputs *for every possible input*.
- So far current strategies fail because they're not *adaptive*.
- Designing a defense that can protect against a powerful, adaptive attacker is an research area

Conclusion

- Adversarial examples show that many modern machine learning algorithms can be broken in surprising ways.

- These failures of machine learning demonstrate that even simple algorithms can behave very differently from what their designers intend

References

[1] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. and Fergus, R., 2013. **Intriguing properties of neural networks**

[2] Goodfellow, I.J., Shlens, J. and Szegedy, C., **Explaining and harnessing adversarial examples** (2014)

[3] T. B. Brown, D. Mané, A. Roy, et al. **“Adversarial Patch”**. In: ArXiv e-prints (Dec. 2017). arXiv: 1712.09665 [cs.CV].

[4] Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., Swami, A., **Practical Black-Box Attacks against Machine Learning** In: ArXiv e-prints (Feb. 2016). arXiv:1602.02697 [cs.CR]

References

[5] Papernot N., McDaniel P., Wu X., Jha S., Swami A., **Distillation as a Defense to AdversarialPerturbations against Deep Neural Networks,** in ArXiv e-prints (Mar. 2016) arXiv:1511.04508 [cs.CR]