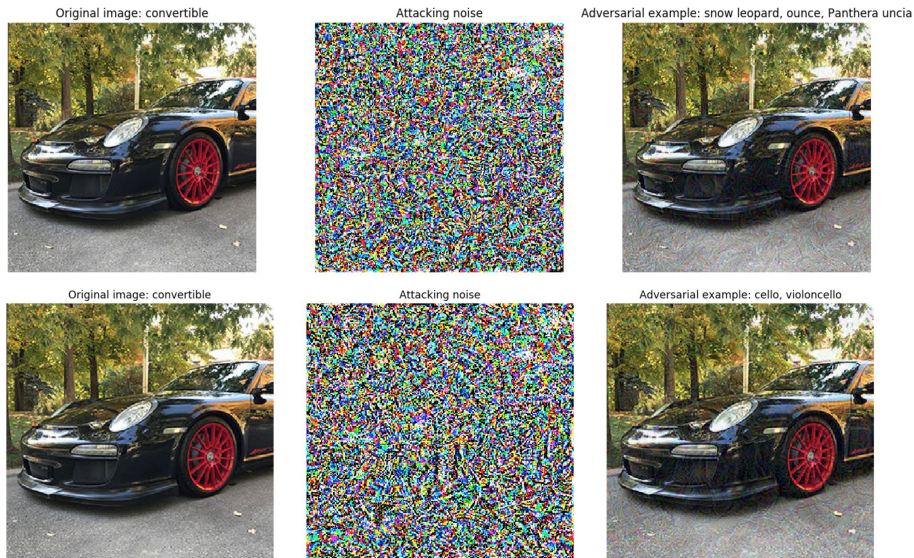


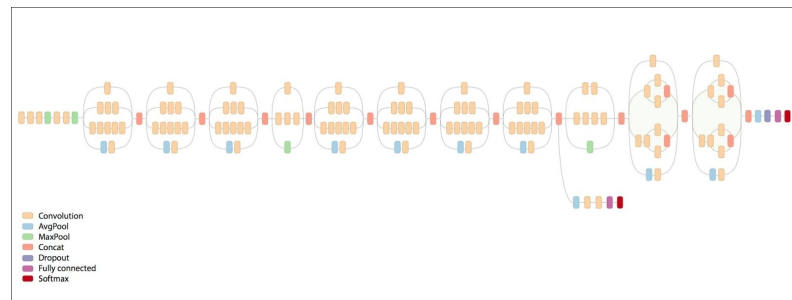
Adversarial Attacks in ML

- “Optical illusions” for machine learning models
- Sensitive decision boundaries
- Find a set of small perturbations (noise) that yield a new class prediction
- Non-targeted vs. targeted attacks
- Inception v3



Defense

- Reactive: train separate model to identify and reject adversarial input
 - ◆ Inelegant
- Proactive: adversarial examples during training
 - ◆ Incomplete
- Defensive distillation
- Gradient masking?

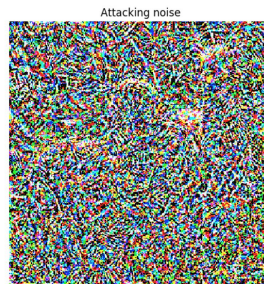


Inception v3 architecture from Google

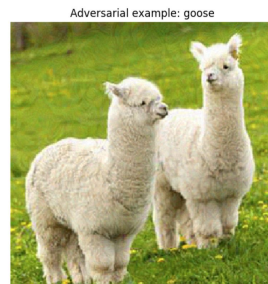
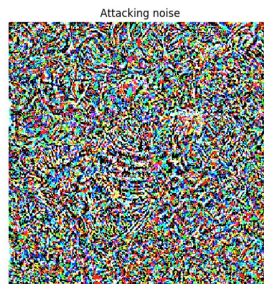
Technical Component

- Non-targeted and targeted attacks
- Black box and gradient available models
- Limit noise norm and location
- Train model on MNIST dataset, compare robustness between models trained with and without adversarial examples

$\epsilon=0.07$



$\epsilon=0.03$



Ethical Component

→ Real world scenarios

- ◆ ATM checks
- ◆ Road signs
- ◆ Disguised weapons



Evtimov et al., "Robust Physical-World Attacks on Deep Learning Models," 2017.

- What types of guarantees must we be able to give before deploying models in critical situations (e.g., autonomous vehicles being tricked by adversarial stickers on stop signs)?
- Should proactive adversarial training be required in security-critical production models?
- How deeply do we need to understand a model's behavior before we can deploy it in ethically sensitive settings?



References

Szegedy et al., “Intriguing properties of neural networks,” 2013.

Goodfellow et al., “Explaining and Harnessing Adversarial Examples,” 2014.

Kurakin et al., “Adversarial examples in the physical world,” 2016.

Papernot et al., “Practical Black-Box Attacks against Machine Learning,” 2016.

Papernot et al., “Towards the Science of Security and Privacy in Machine Learning,” 2016.

Evtimov et al., “Robust Physical-World Attacks on Deep Learning Models,” 2017.