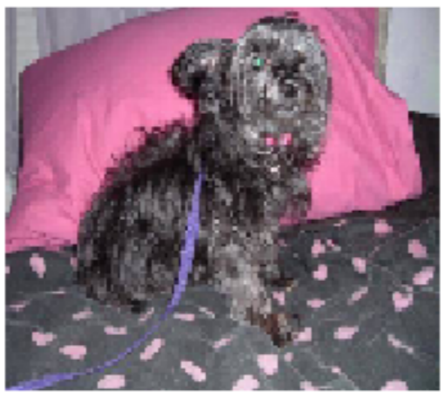
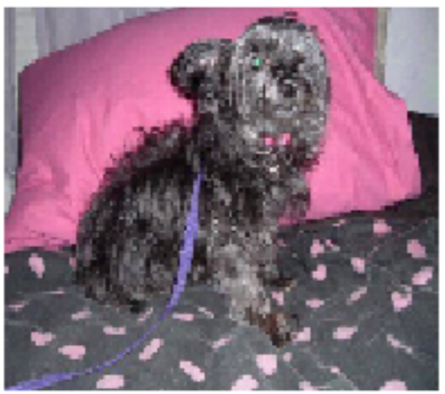
# Adversarial examples

Jan-Hendrik Plank, Jonas Dehning & Philipp Höhne

July 12, 2017

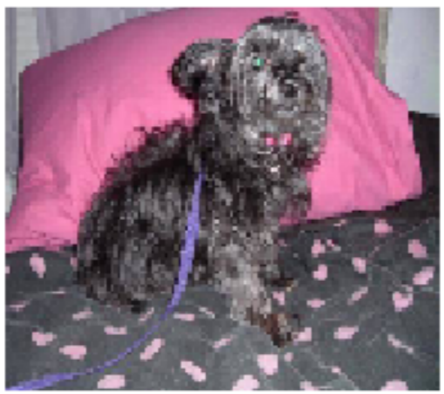
# Dog or Cat?



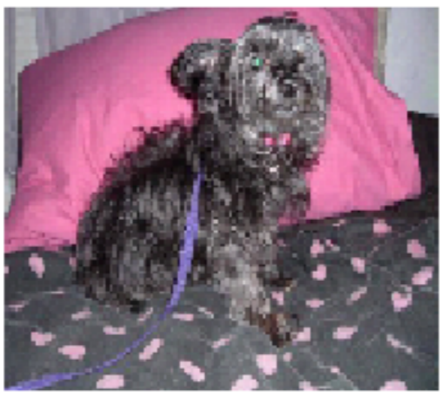


# Dog or Cat?

Prediction: dog

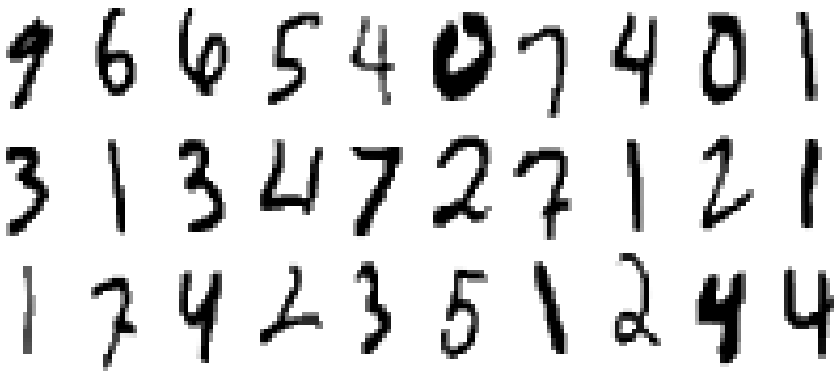


Prediction: cat



# The data sets

## MNIST



- greyscale
- 28x28 pixel

## Dogs vs. Cats

- colored
- 128x128 pixel

# Networks

- Convolutional Neural Network

| Data set | Network structure | # Layers |
|---|---|---|
| MNIST | 5 CL - 2 FC | 7 |
| Dogs vs. Cats | 11 CL - 3 FC | 14 |

# Finding adversarial examples: Gradient Method

$$\vec{\eta} = \epsilon \cdot \text{sign}\left(\nabla_{\vec{x}} J_{loss}\big|_{\vec{x}}\right)$$

$\vec{\eta}$ : noise

$\vec{x}$ : picture

$\epsilon \ll 1$

From Goodfellow et al. 2015

# Finding adversarial examples: Minimizer

Minimize over $\eta$:

$$\min_{\vec{\eta}} \left( c \cdot \sigma(\vec{\eta}) + \frac{1}{1 + \delta - p(\vec{x} + \vec{\eta})} \right)$$

$\vec{\eta}$ : noise

$\vec{x}$ : picture

$p$ : prediction

$c$ : a constant

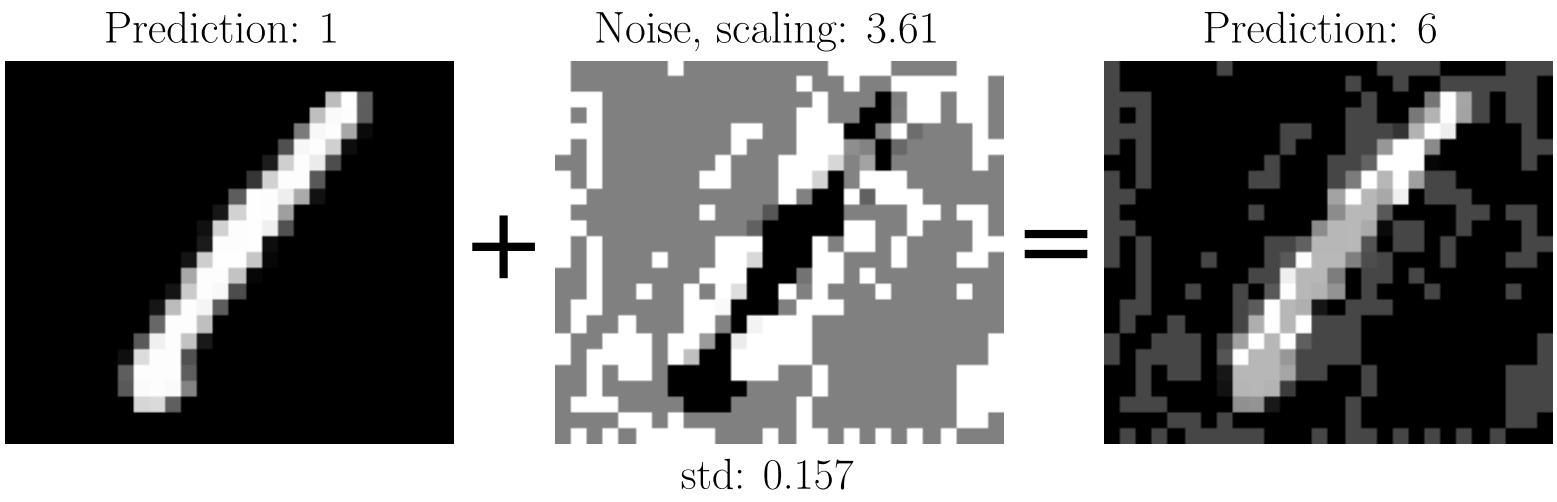
$\delta \ll 1$

with constraint:

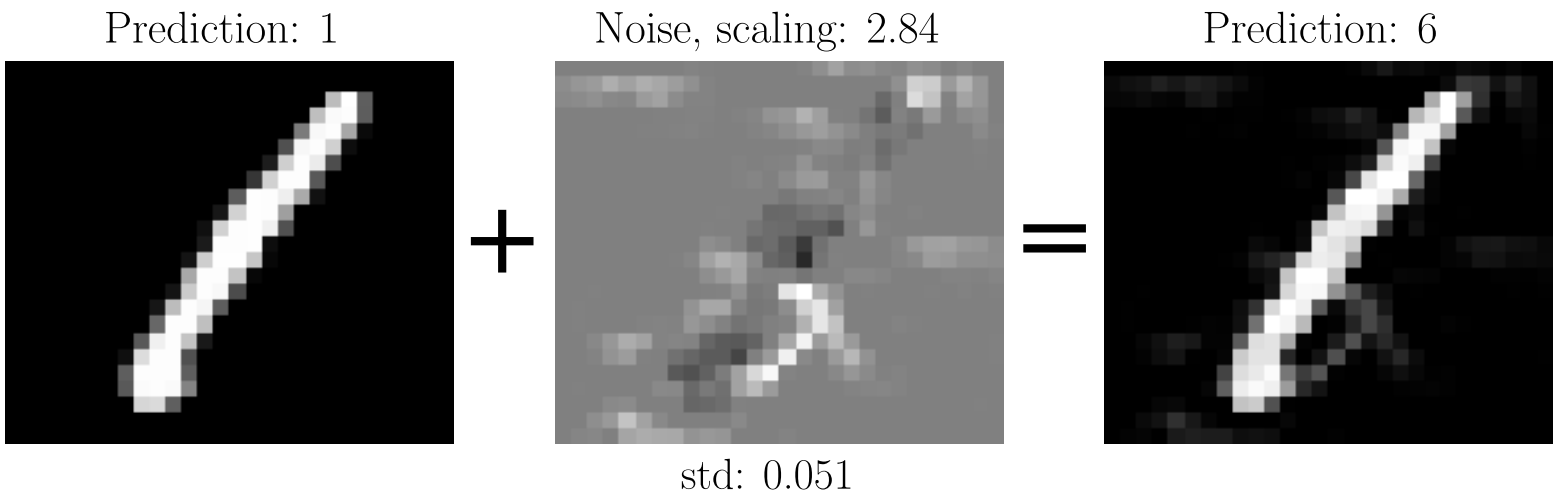
$$\vec{x} + \vec{\eta} \in [0, 1]^n$$

From Szegedy et al. 2014

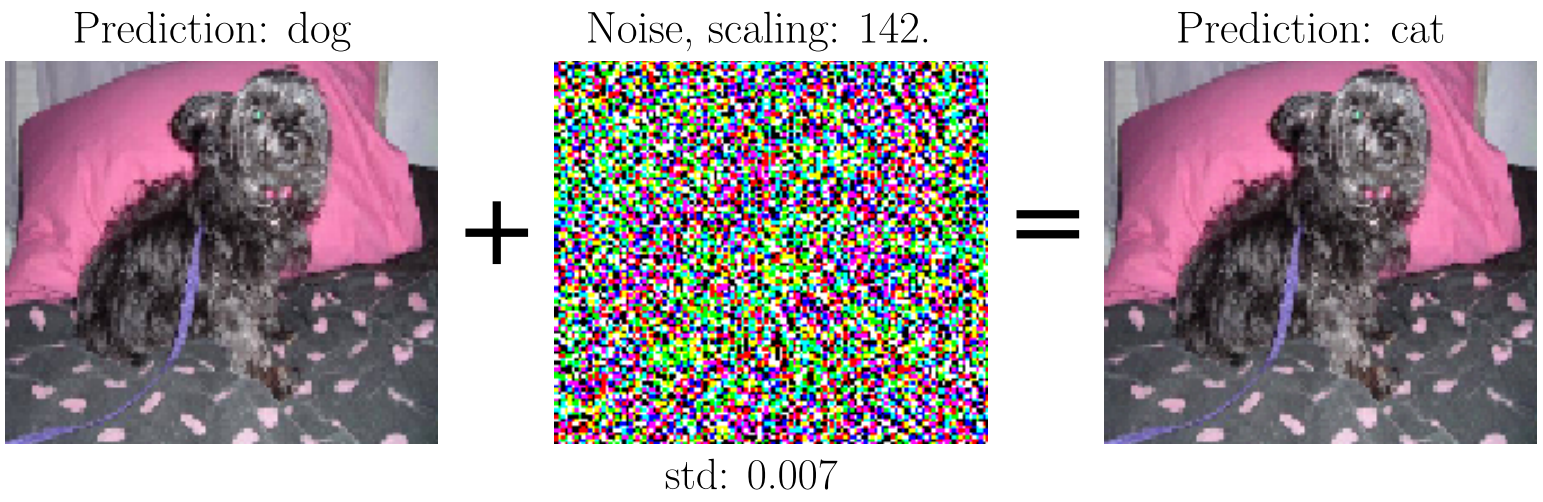
# MNIST adversarial examples (Gradient Method)



# MNIST adversarial examples (Minimizer)

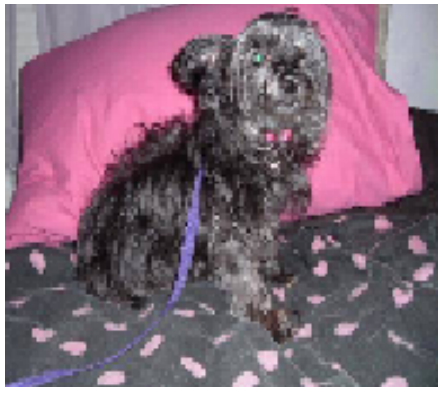


# Dogs vs. Cats adversarial examples (Gradient Method)



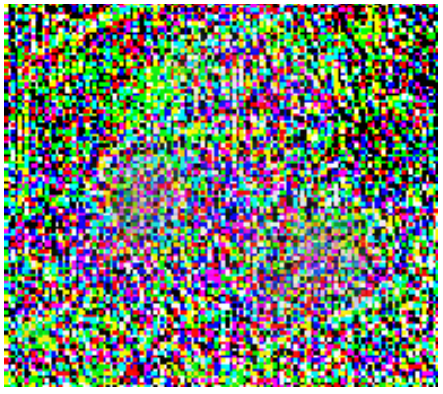
# Dogs vs. Cats adversarial examples (Minimizer)

Prediction: dog



+

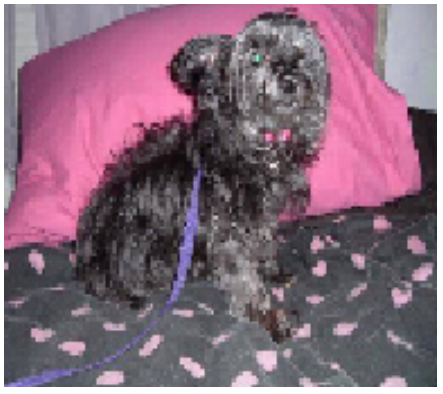
Noise, scaling: 152.



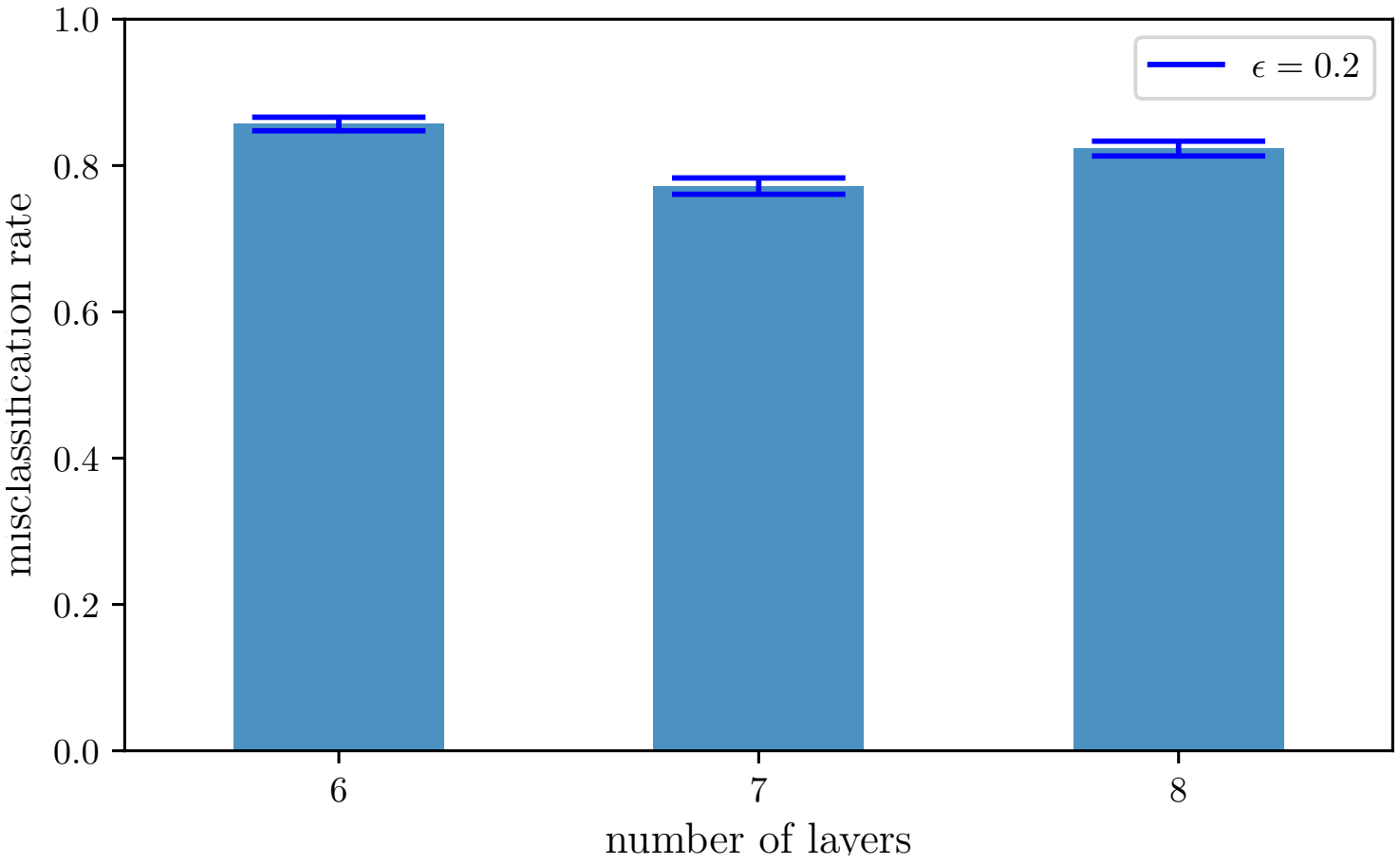
std: 0.001

=

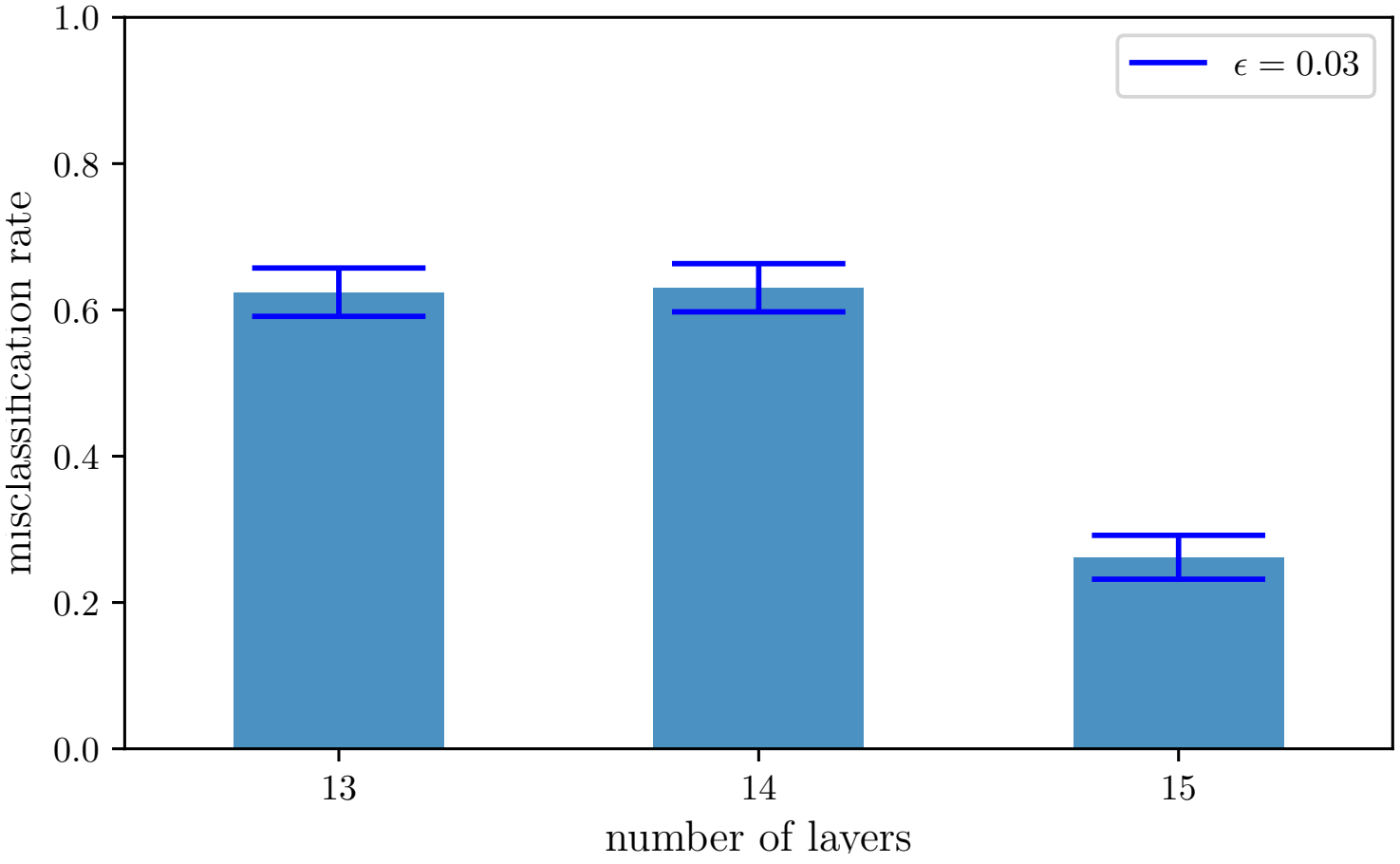
Prediction: cat



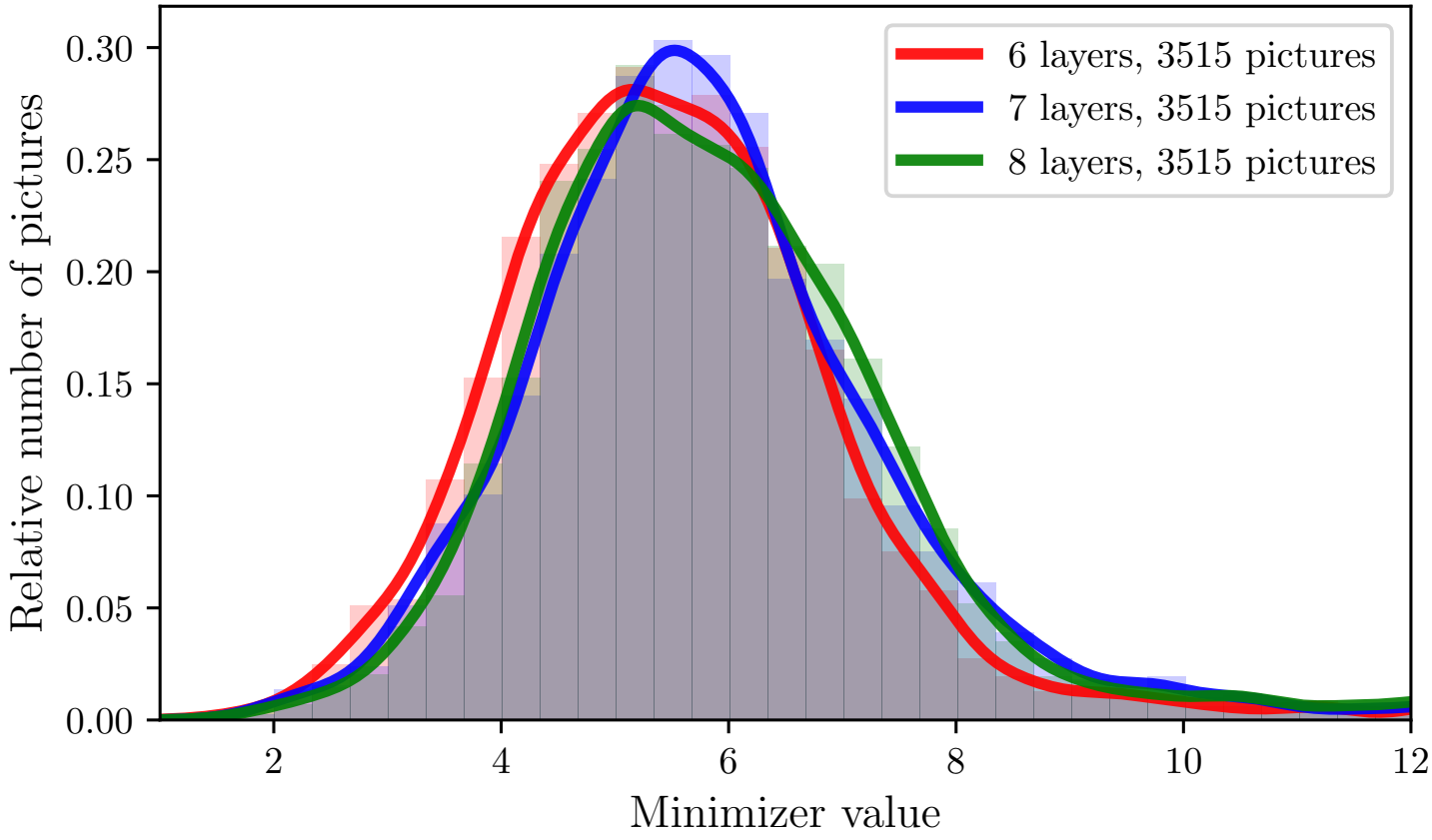
# MNIST misclassification rate for gradient method



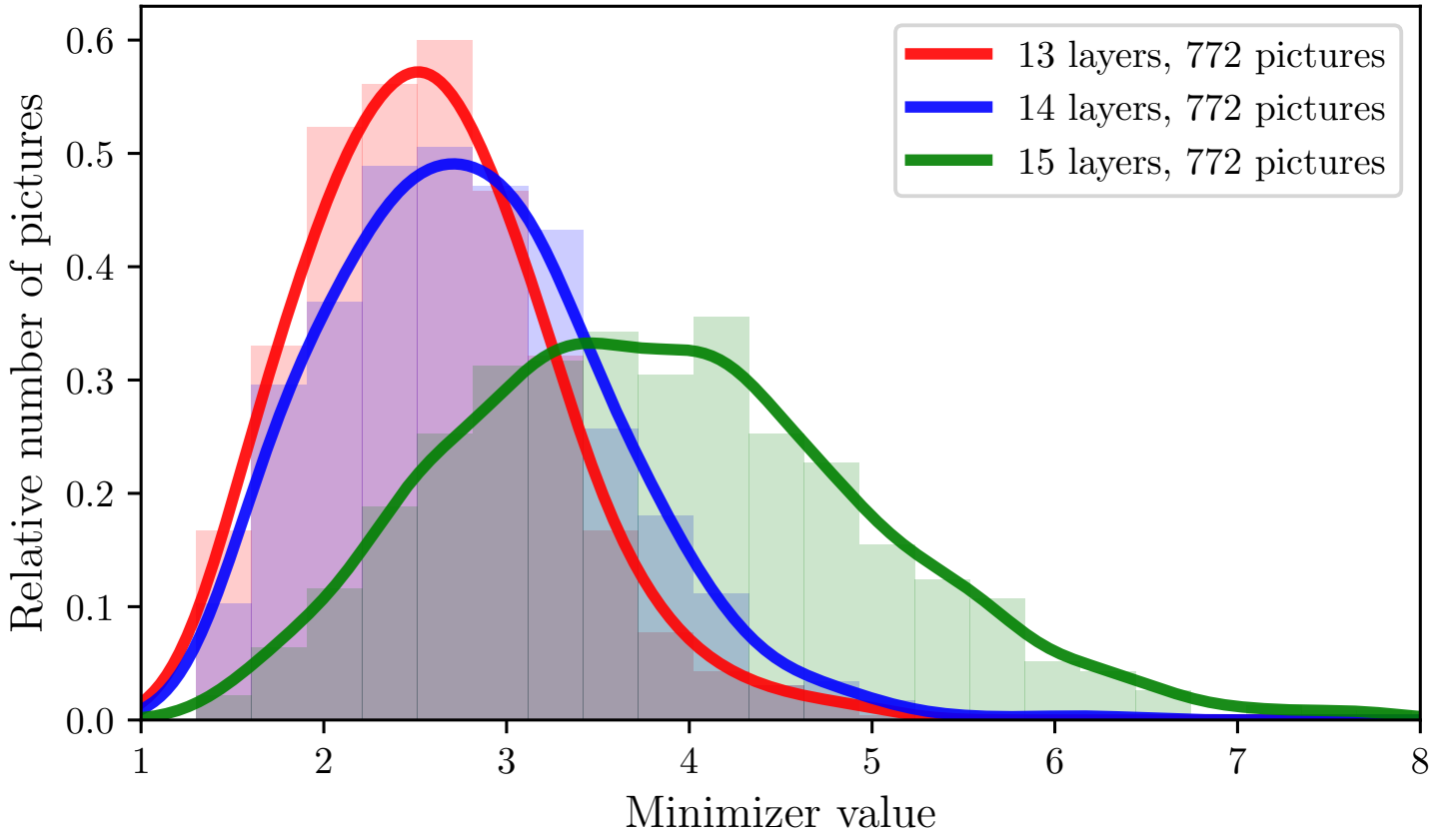
# Dogs vs. Cats misclassification rate for gradient method



# MNIST robustness Minimizer



# Dogs vs. Cats robustness Minimizer



# Conclusion

- Minimizer: smaller noise
- Gradient: computationally cheap

- MNIST:
  - robust
  - no dependence on number of layers found
- Dogs vs. Cats:
  - vulnerable
  - deeper networks seemingly less vulnerable