Adversarial Training and Security in ML

Lili Mou, Priyank Jaini
doublepower.mou@gmail.com

David R. Cheriton School of Computer Science
University of Waterloo

UROC 22 Sep 2017

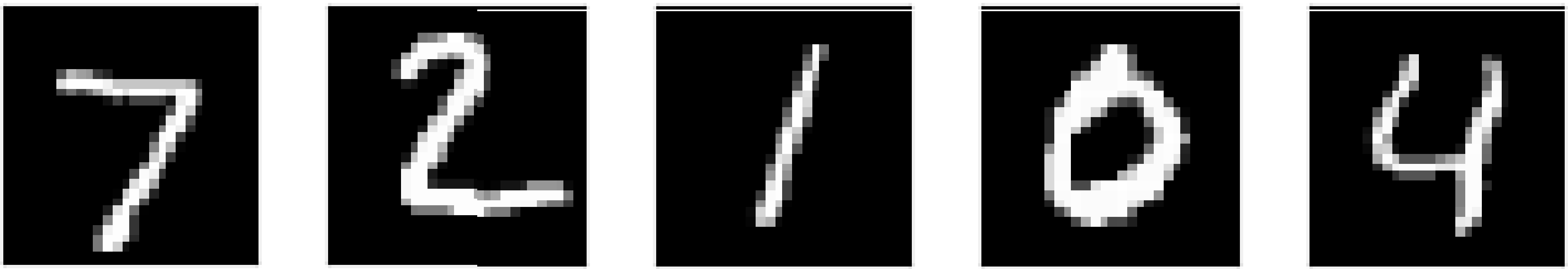
Agenda

- **Background of neural networks**
  - Miniproject: CNN and its visualization
- Adversarial samples
  - Miniproject: Crafting adversarial data
- Open research

UNIVERSITY OF WATERLOO

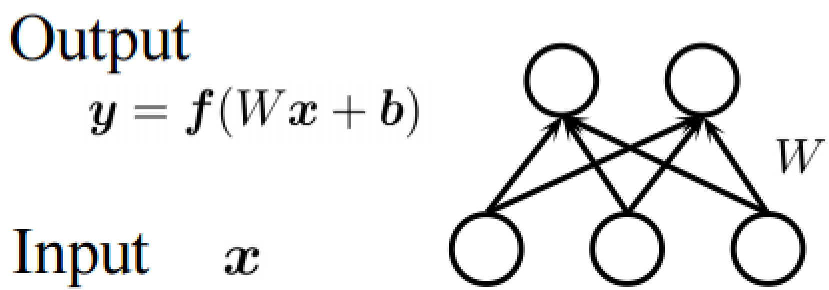
Philosophy of Deep Learning

- Consider hand-written digit recognition

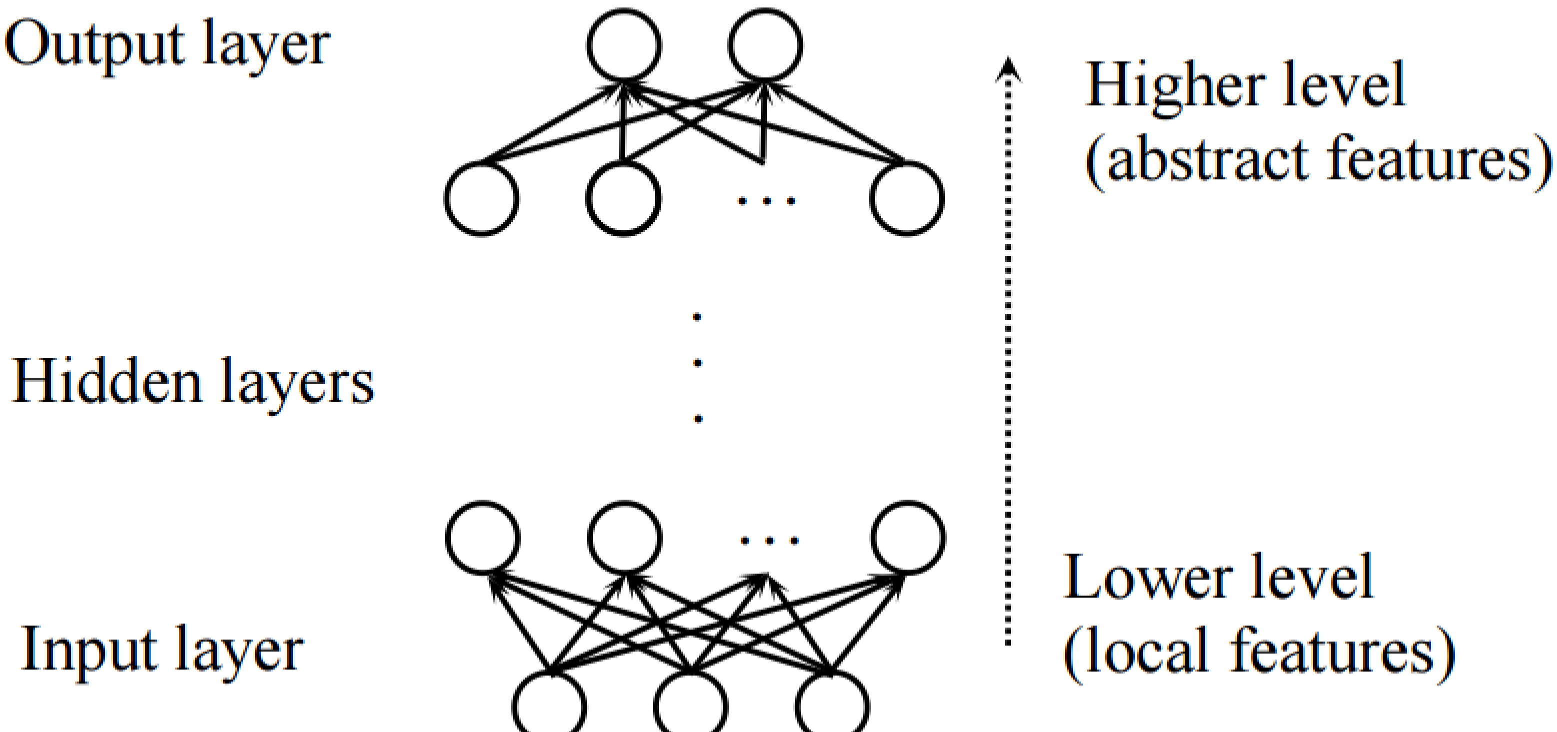


- Deep learning: End-to-end training
  - Input: raw signal (28 x 28 pixels)
  - Output: the target labels "7," "2," "1," "0," and "4."

UNIVERSITY OF WATERLOO



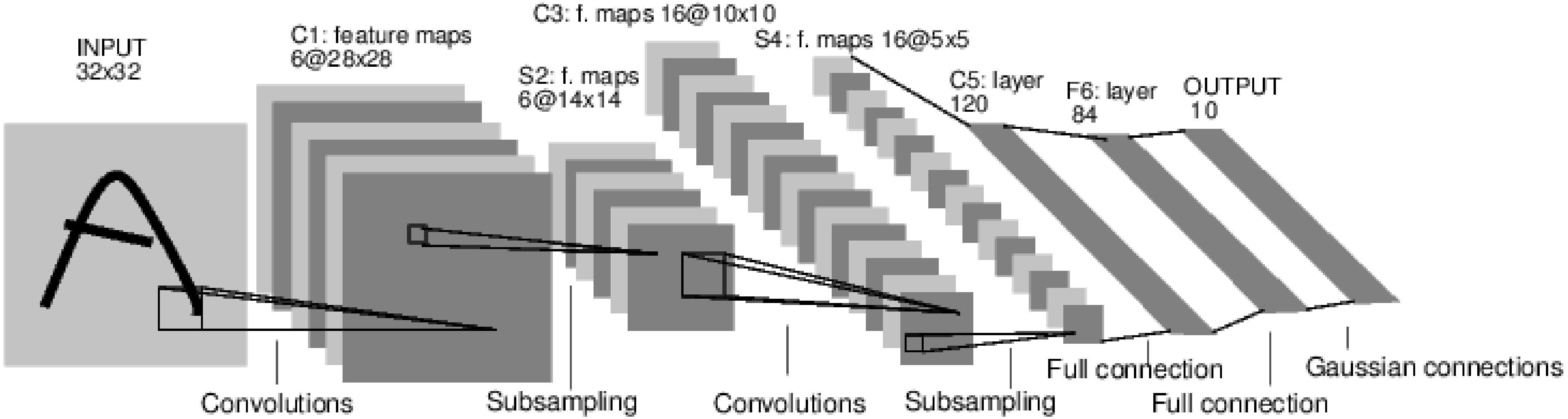
(A) A single layer of neurons



(B) A deep neural network

UNIVERSITY OF WATERLOO

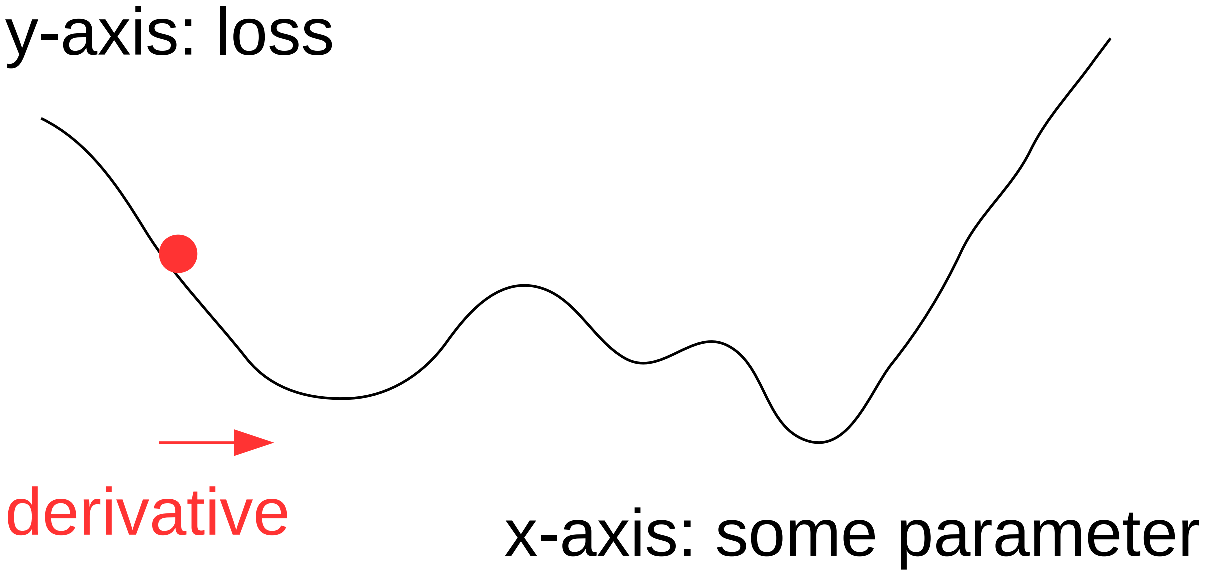
A Convolutional Neural Network



LeNet5

Training

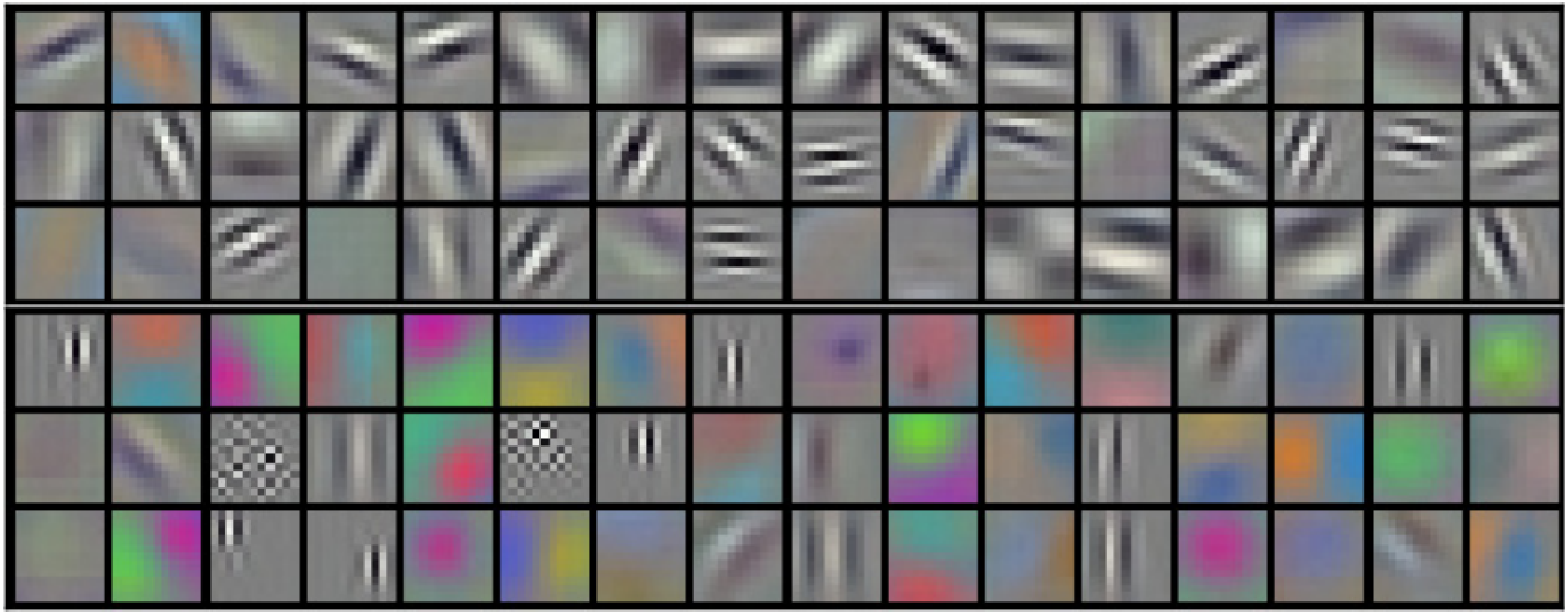
- How do we learn weights?
  - Backpropagation
  - Compute the partial derivative of a "loss" w.r.t. each parameter
  - Take a small step towards the derivative



UNIVERSITY OF WATERLOO

What are these features?

- Visualizing weights

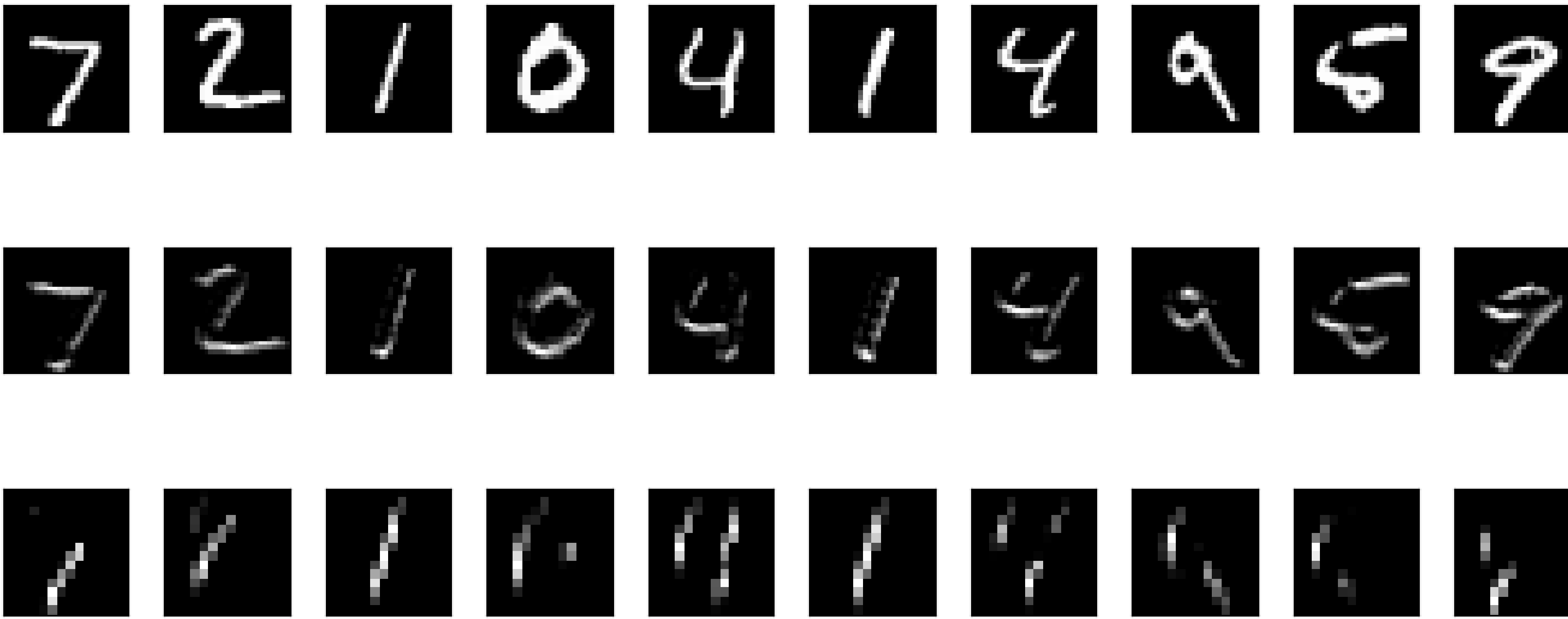


Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *NIPS*. 2012.

UNIVERSITY OF WATERLOO

What are these features?

- Visualizing the activation functions



Mini-Project

- Code

  https://www.dropbox.com/s/iafhbi87mtk67gk/Adversarial.zip?dl=0

- Cached data

  https://www.dropbox.com/s/m2qn92q5b8ky4mo/adv_data.zip?dl=0

- Slides available at

  http://sei.pku.edu.cn/~moull12

Agenda

- Background of neural networks
  - Miniproject: CNN and its visualization
- **Adversarial samples**
  - Miniproject: Crafting adversarial data
- Open research

Think of the Training Process

- Loss: $L = f(x; w)$
- Training objective: minimize L
- Compute: $\nabla_w f(x; w)$

Think of the Training Process

- Loss: $L = f(x; w)$
- Training objective: minimize L
- Compute: $\nabla_w f(x; w)$

- What if we compute: $\nabla_x f(x; w)$

UNIVERSITY OF WATERLOO

Adversarial Samples from Random

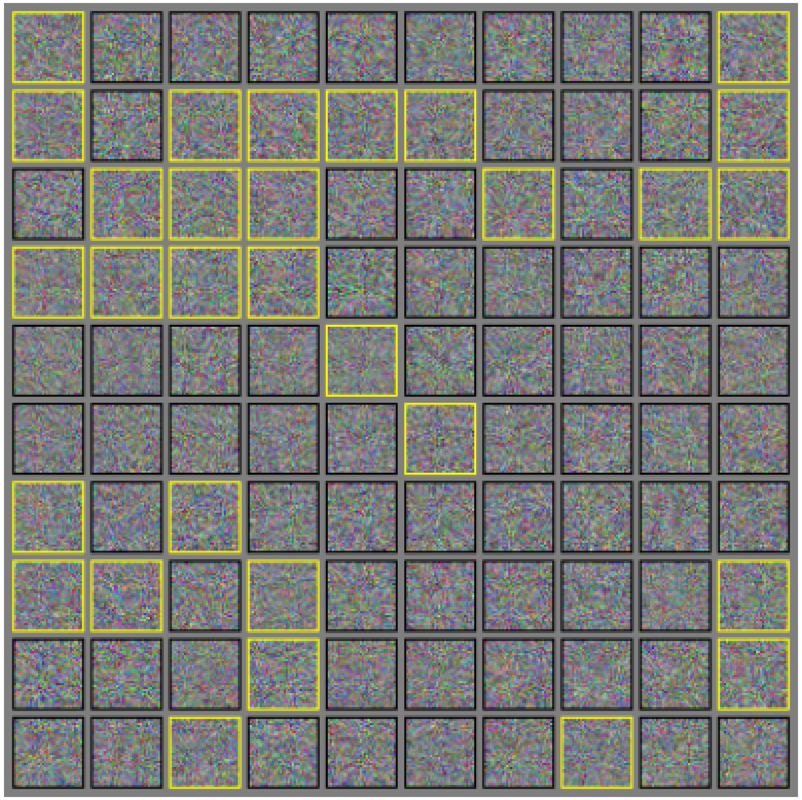
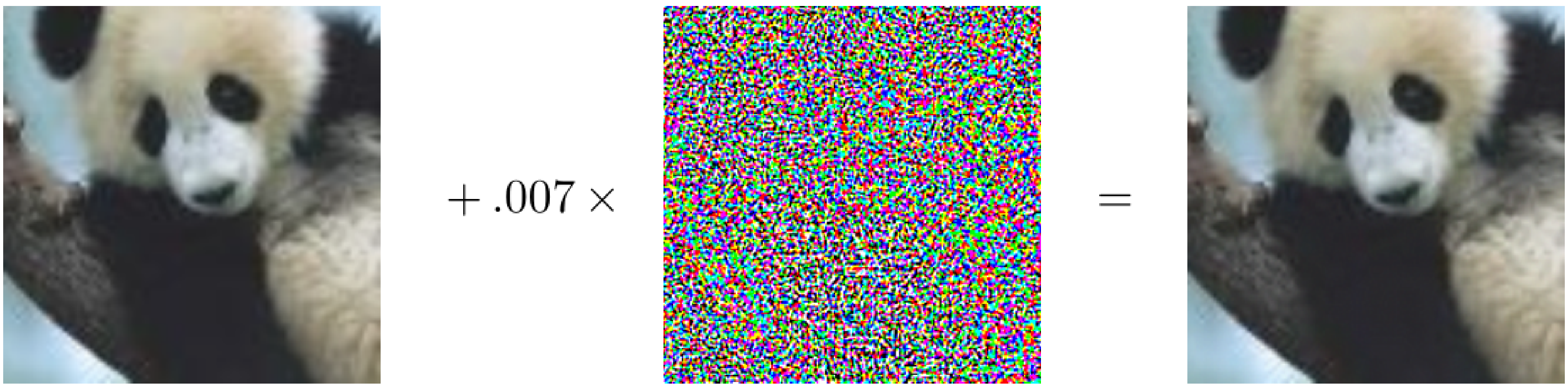


Figure 5: Randomly generated fooling images for a convolutional network trained on CIFAR-10. These examples were generated by drawing a sample from an isotropic Gaussian, then taking a gradient sign step in the direction that increases the probability of the "airplane" class. Yellow boxes indicate samples that successfully fool the model into believing an airplane is present with at least 50% confidence. "Airplane" is the hardest class to construct fooling images for on CIFAR-10, so this figure represents the worst case in terms of success rate.

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *ICLR*, 2015.

Adversarial Samples from Real Data



$\boldsymbol{x}$ $+ .007 \times$ $\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$ $=$ $\boldsymbol{x} + \epsilon \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"panda"
57.7% confidence

"nematode"
8.2% confidence

"gibbon"
99.3 % confidence

Approach

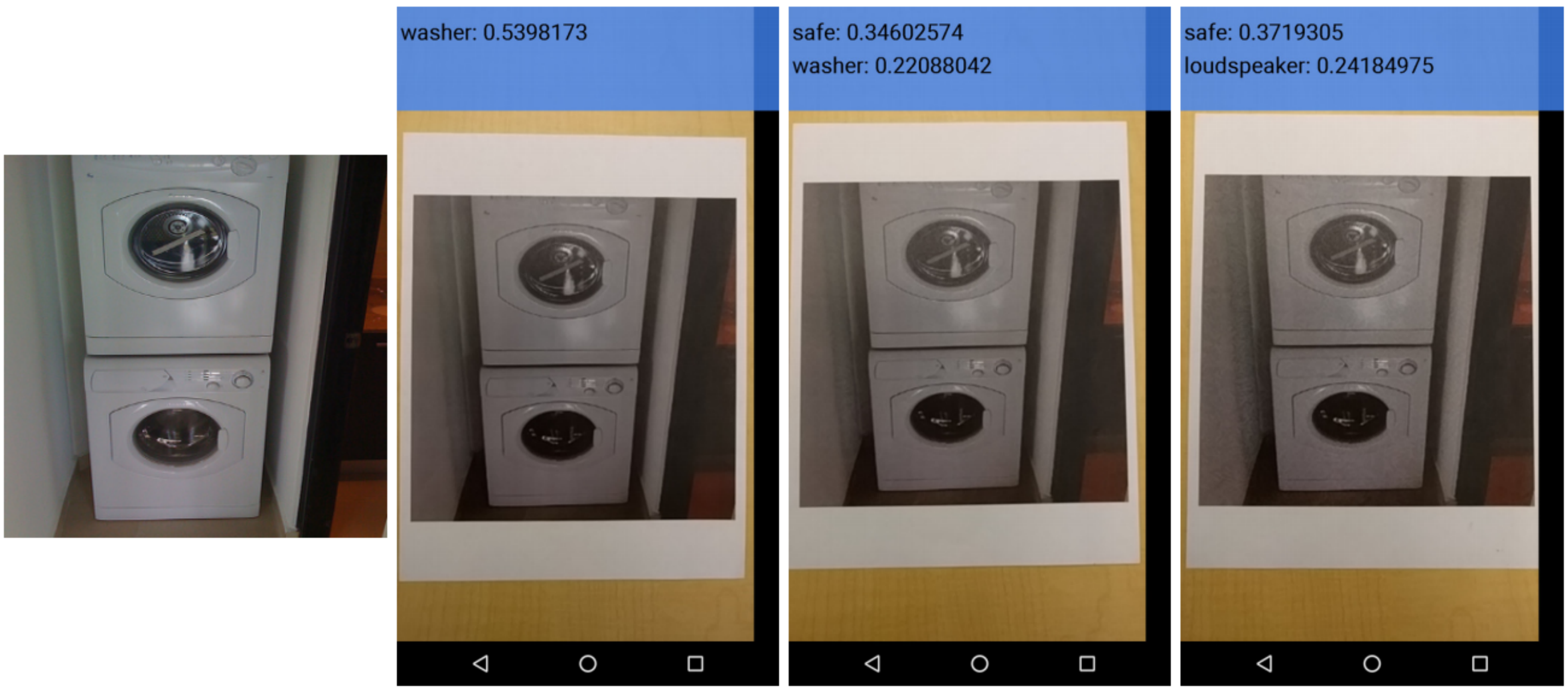
$$X_{\mathrm{adv}} = X - \epsilon \cdot \mathrm{sign}(\nabla_X J(X, y_{\mathrm{target}}))$$

- y_{target} : whatever target you want
- Take the sign of the partial derivative
  - Alternatively, we can truncate the gradiate
  - So that the adversarial image is not too far away
- Can interate several times if necessary

UNIVERSITY OF WATERLOO

Ubiquity of Adverarial Samples

- A same adversarial sample works:
  - For different networks (models)
  - Even after further perturbation with noise



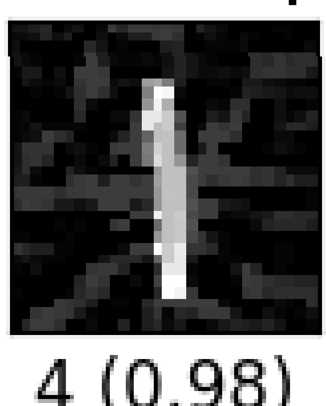
Kurakin, Alexey, Ian Goodfellow, and Samy Bengio. "Adversarial examples in the physical world." *ICLR*, 2017

UNIVERSITY OF WATERLOO

Mini-Project

- Generate Adversarial Samples by Yourselves
- Results

- Interpretation, predicted as "4" w.p. 98%



UNIVERSITY OF WATERLOO

Agenda

- Background of neural networks
  - Miniproject: CNN and its visualization
- Adversarial samples
  - Miniproject: Crafting adversarial data
- **Open research**

UNIVERSITY OF WATERLOO

Open Topics

- Test the robustness of NNs
- Further confirm the ubiquity of adv samples
- Crafting more deceptive adversarial samples
- Training more robust machine learning models

UNIVERSITY OF WATERLOO

Thanks!

QA