Faster DNN Training with Selective Backpropagation

Angela Jiang

Daniel L.-K. Wong, Michael Kaminsky, Michael A. Kozuch,
Padmanabhan Pillai, David G. Andersen, Gregory R. Ganger

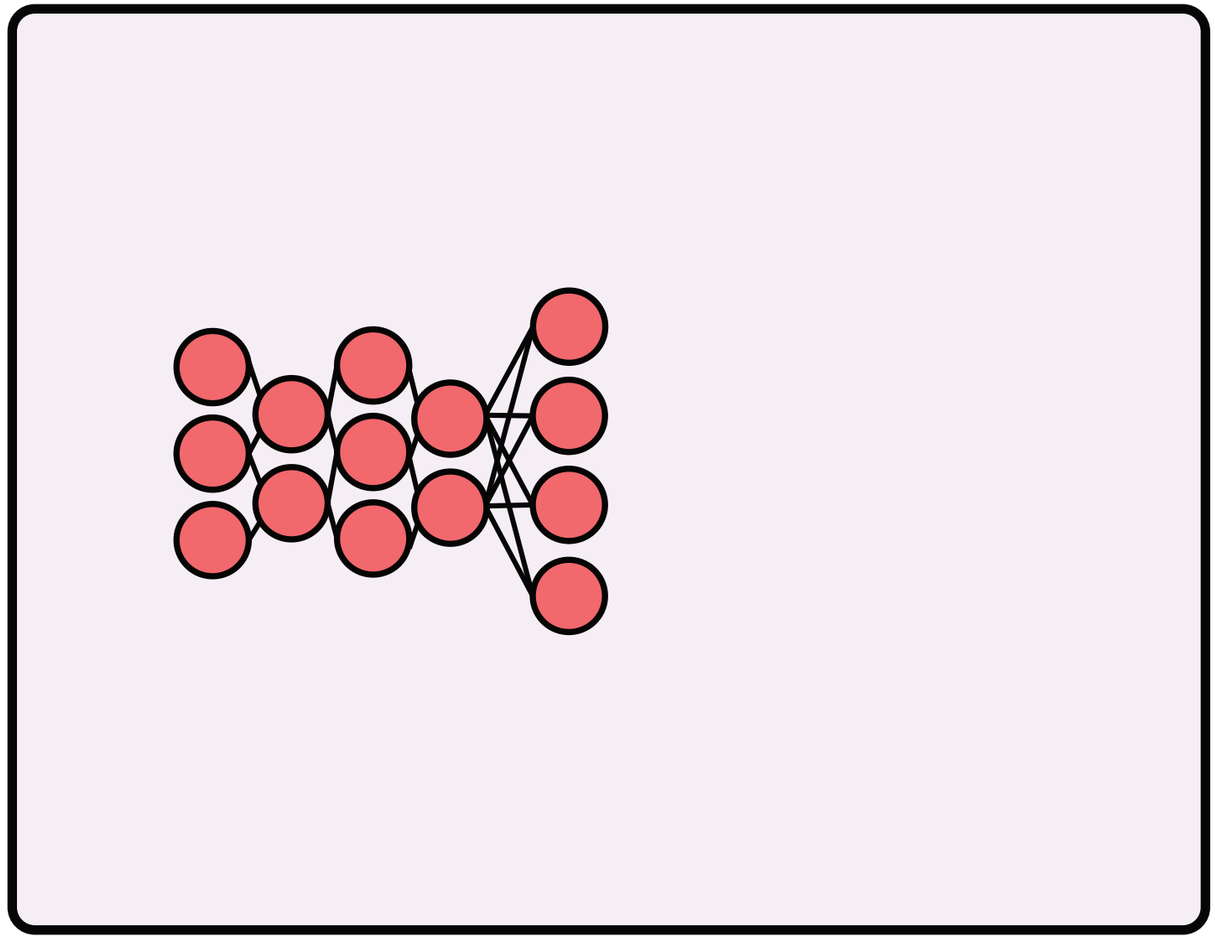
PARALLEL DATA LABORATORY

Carnegie Mellon University

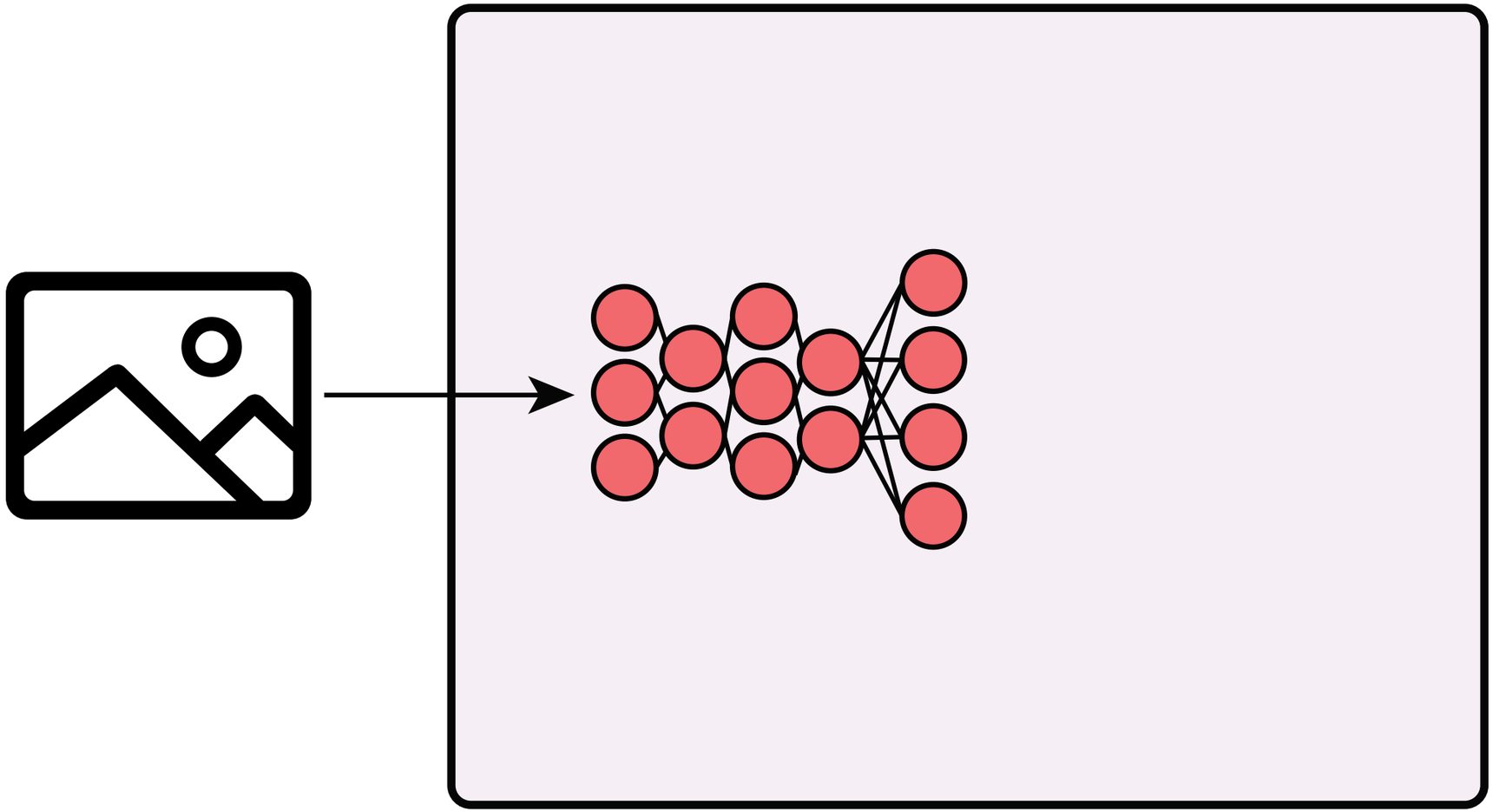
Labeled Datasets are Getting Larger

- ImageNet: **15 million** images
- OpenImages: **9 million** images
- Production datasets are often much larger
  - JFT: **300 million** images
  - click-through data
  - autonomous vehicle training video

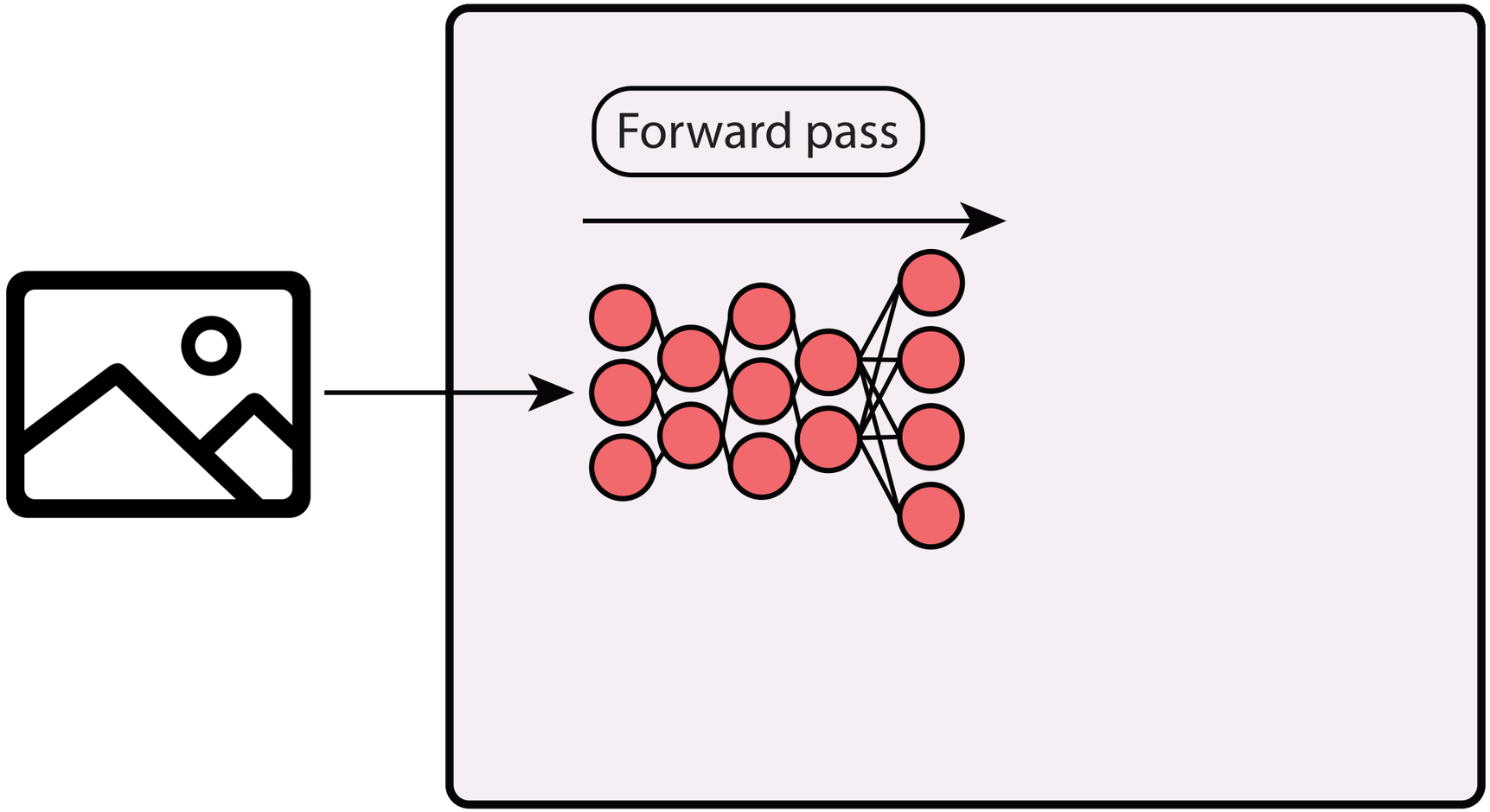
Training cycle is bottlenecked by backwards pass



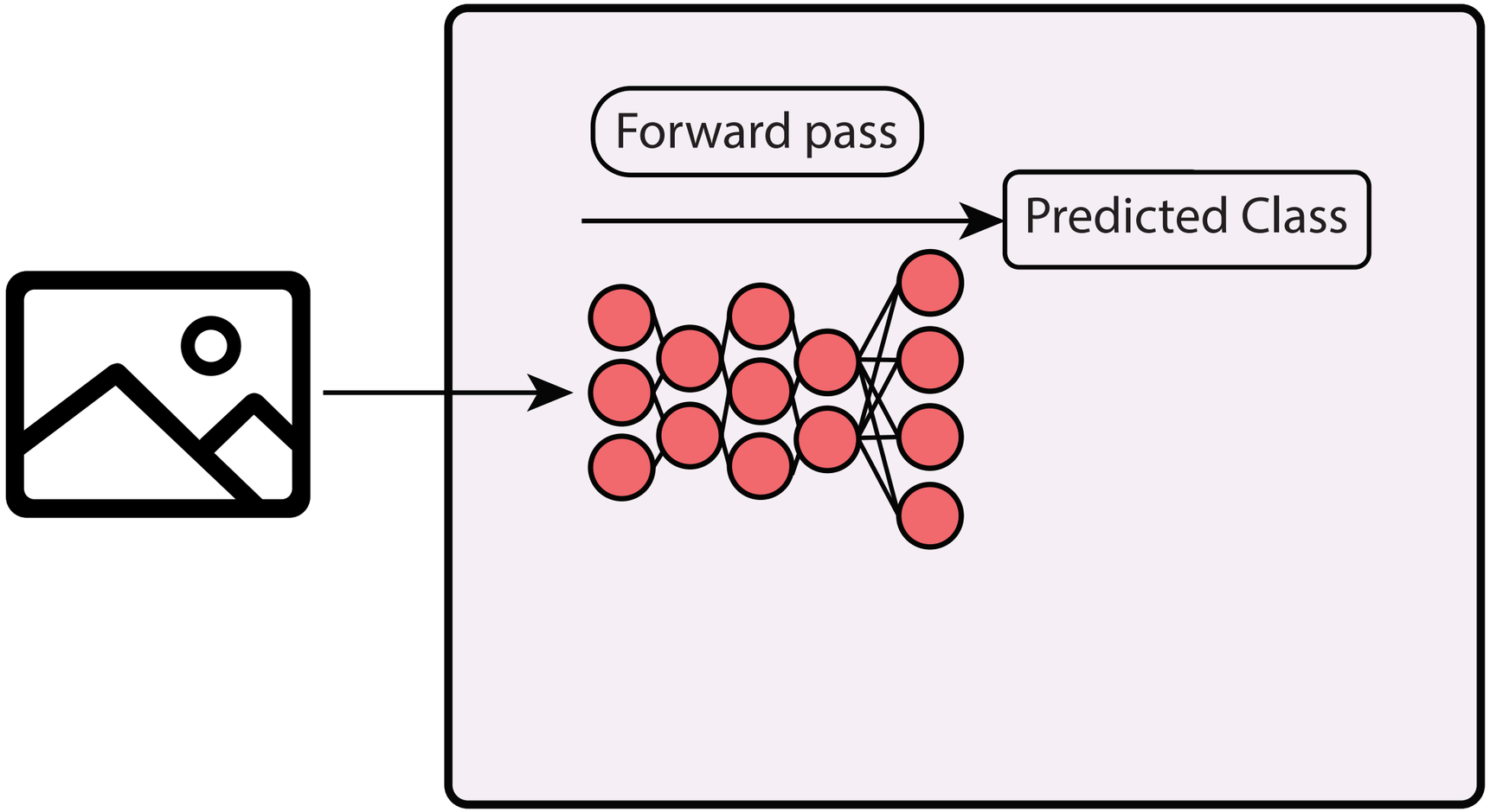
Training cycle is bottlenecked by backwards pass



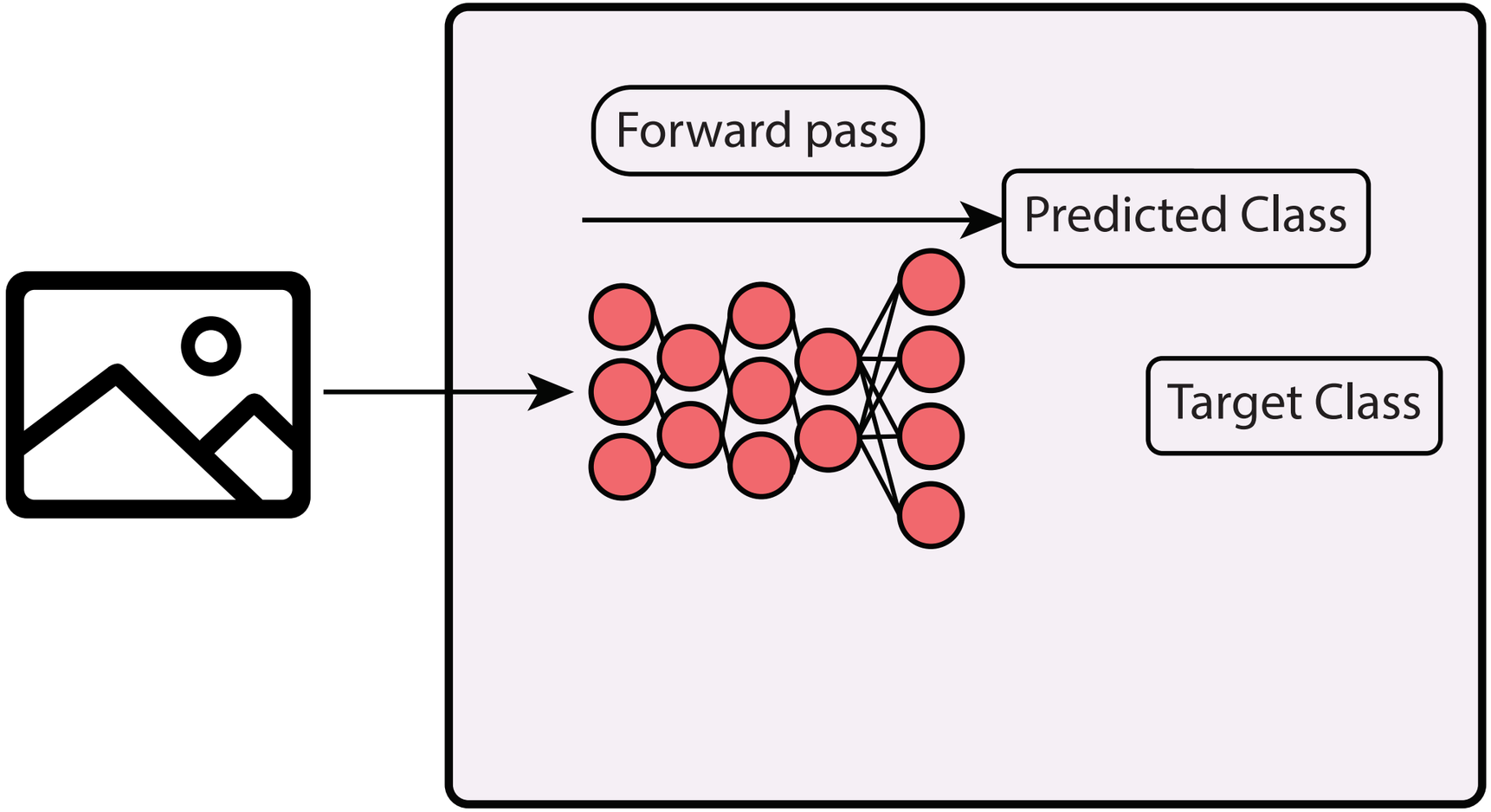
Training cycle is bottlenecked by backwards pass



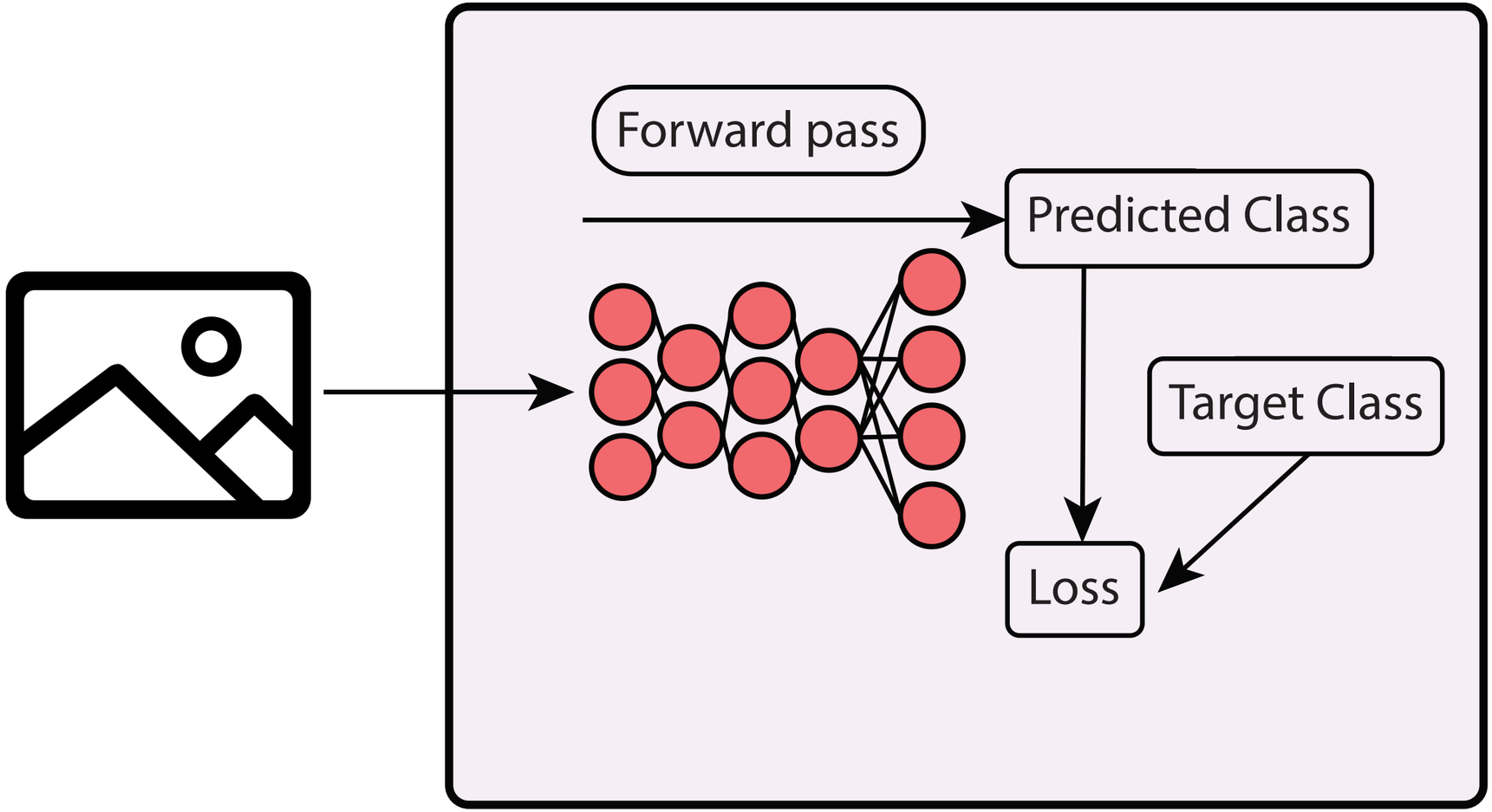
Training cycle is bottlenecked by backwards pass



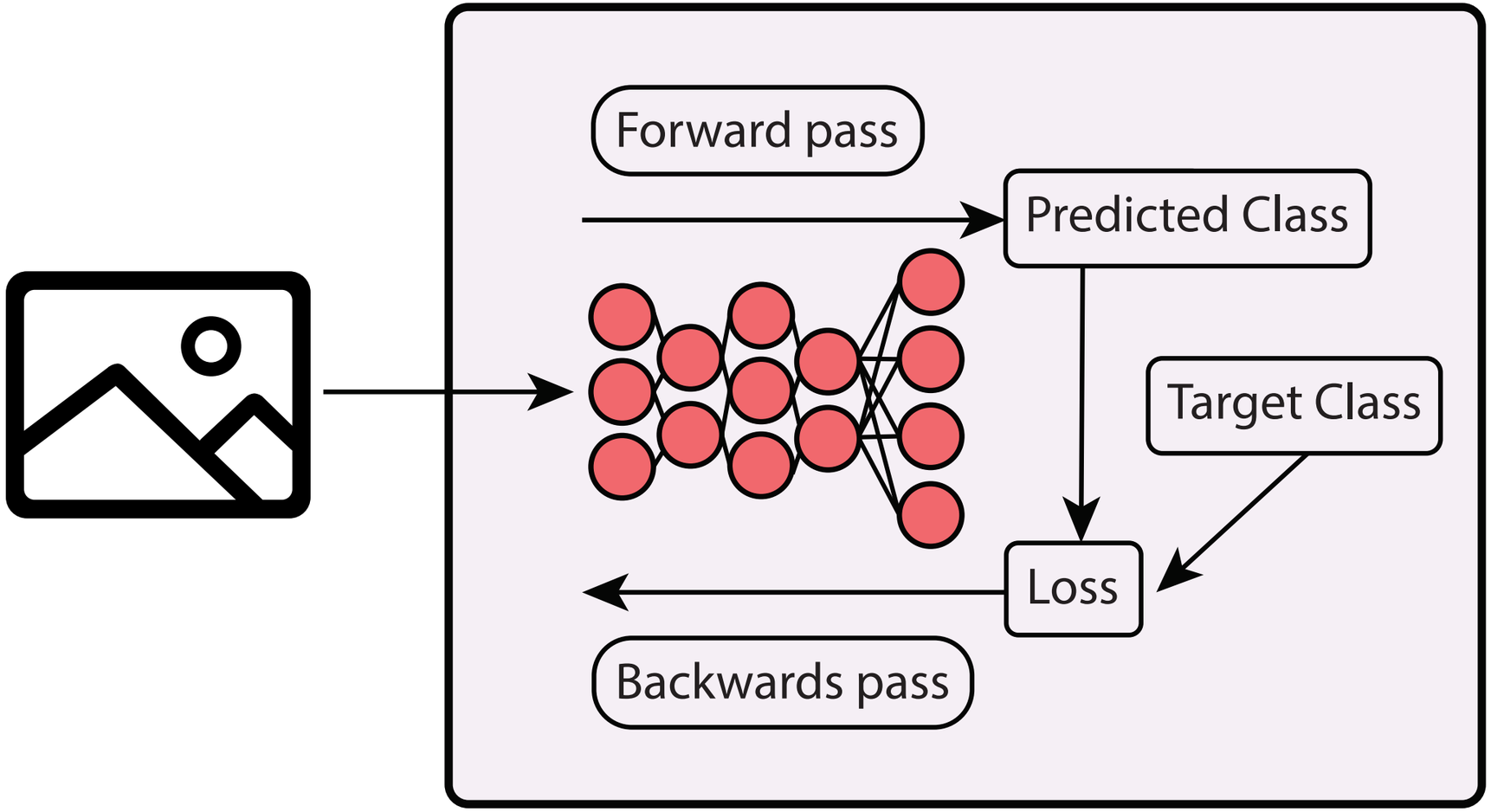
Training cycle is bottlenecked by backwards pass



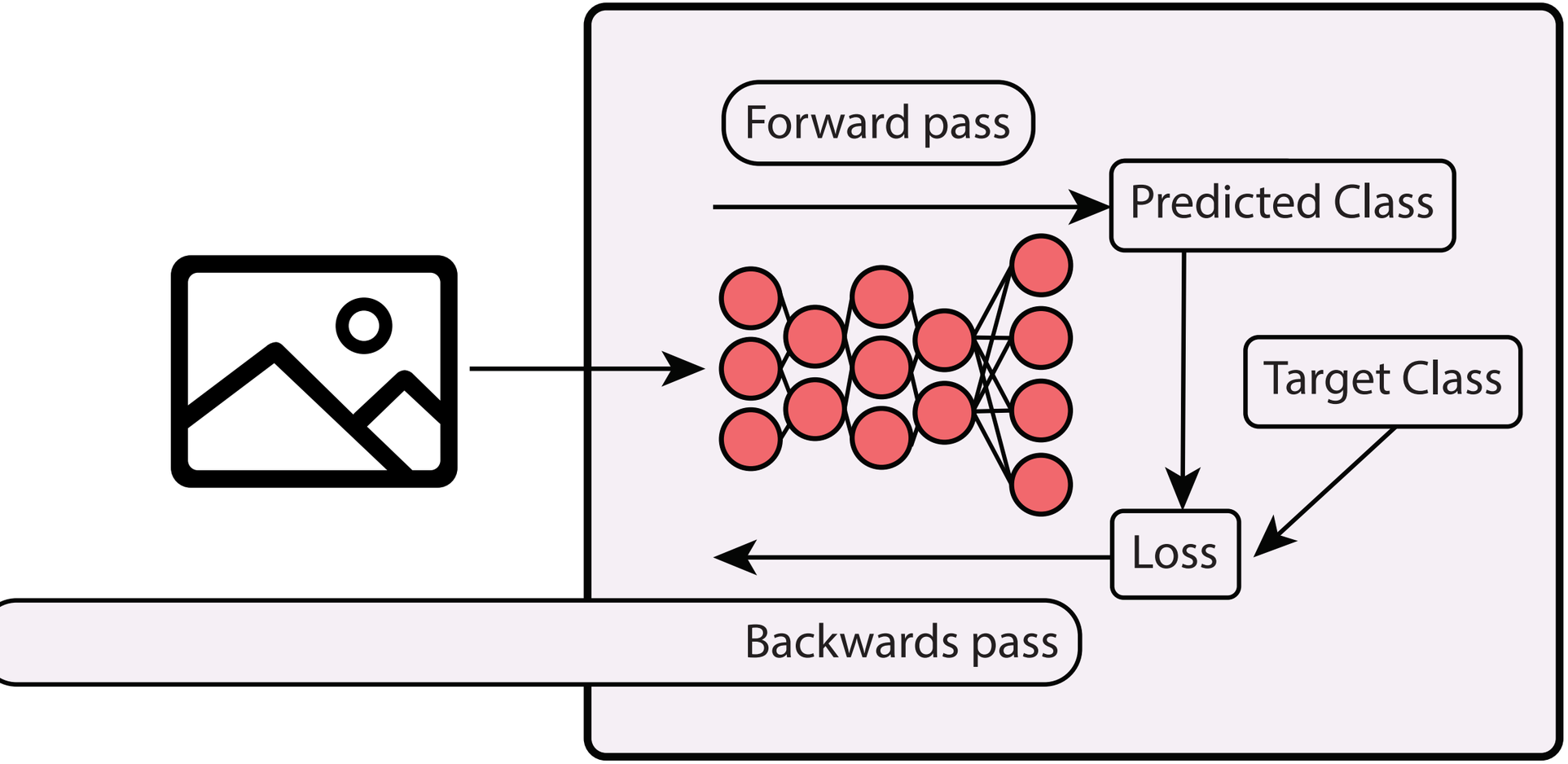
Training cycle is bottlenecked by backwards pass



Training cycle is bottlenecked by backwards pass



Training cycle is bottlenecked by backwards pass



Can we reduce the number of examples backpropped?

How to tell if an example will teach us

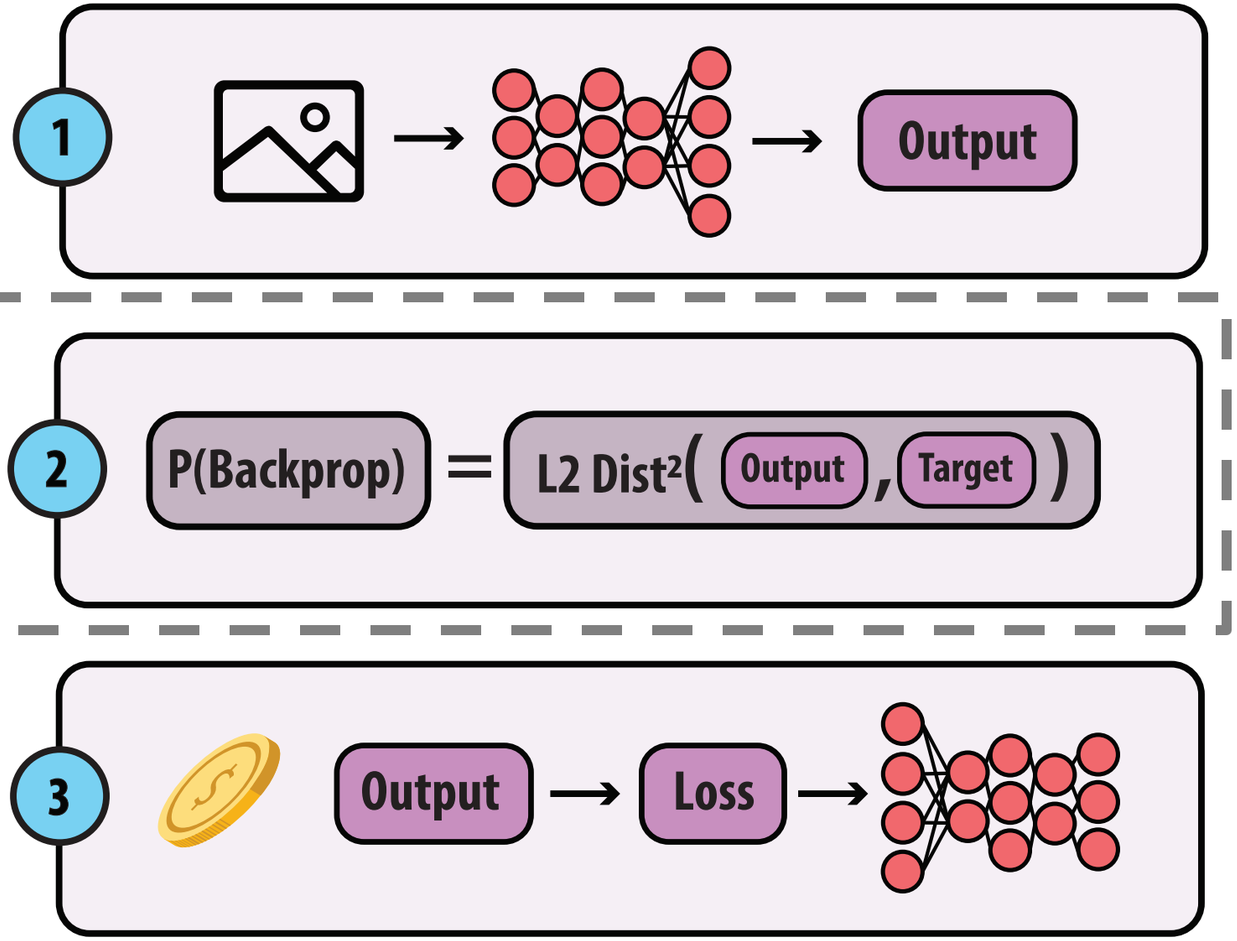
- Frame decimation, difference detectors, etc
  - Done before training => not specific to the trained model
  - Never gets to train on filtered examples
- Active learning techniques
  - Doesn't take advantage of ground truth
- Compare output of inference to ground truth (e.g., loss)
  - Fast
  - Specific to *current state* of *this* model
  - Can use all images, but change the sampling distribution

Selective-Backprop (SB) approach

Forward propagate example through the network

Calculate usefulness of backpropping example based on its accuracy

"Flip a coin" to determine if we should backprop



Selective-Backprop's Usefulness Metric

- More surprising => More likely to backprop
- Fast
- Self-Regulating

Selective-Backprop's Usefulness Metric

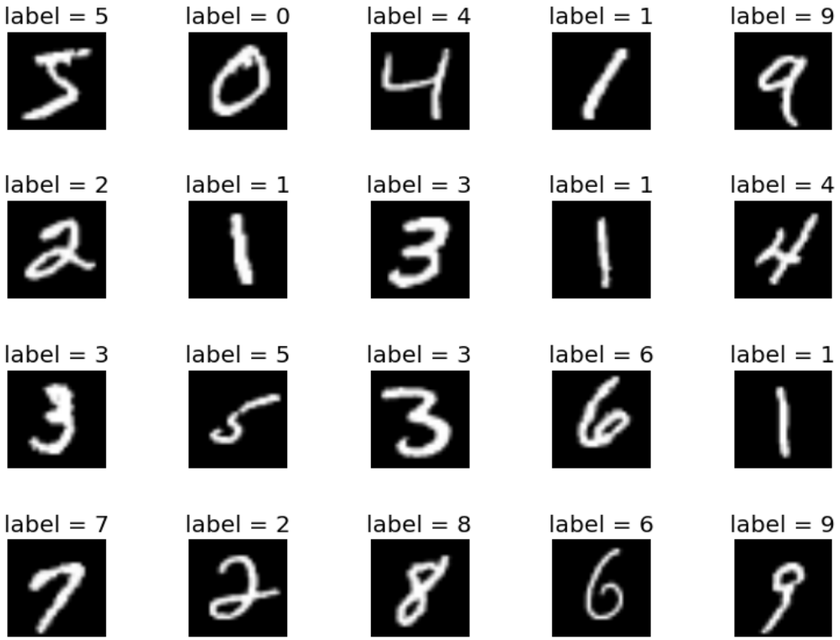
$$\text{Target} = [0, 0, 1]$$

$$\text{Output} = [0.1, 0.3, 0.6]$$

$$\text{L2 Dist}^2(\text{Output}, \text{Target}) = 0.509$$

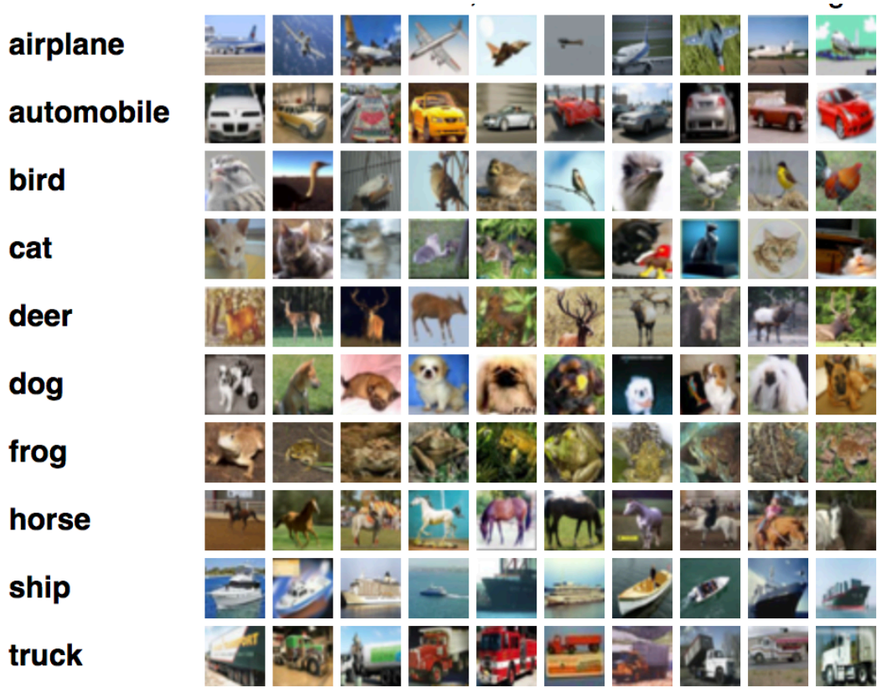
$$P(\text{Backprop}) = 0.509$$

Datasets



MNIST

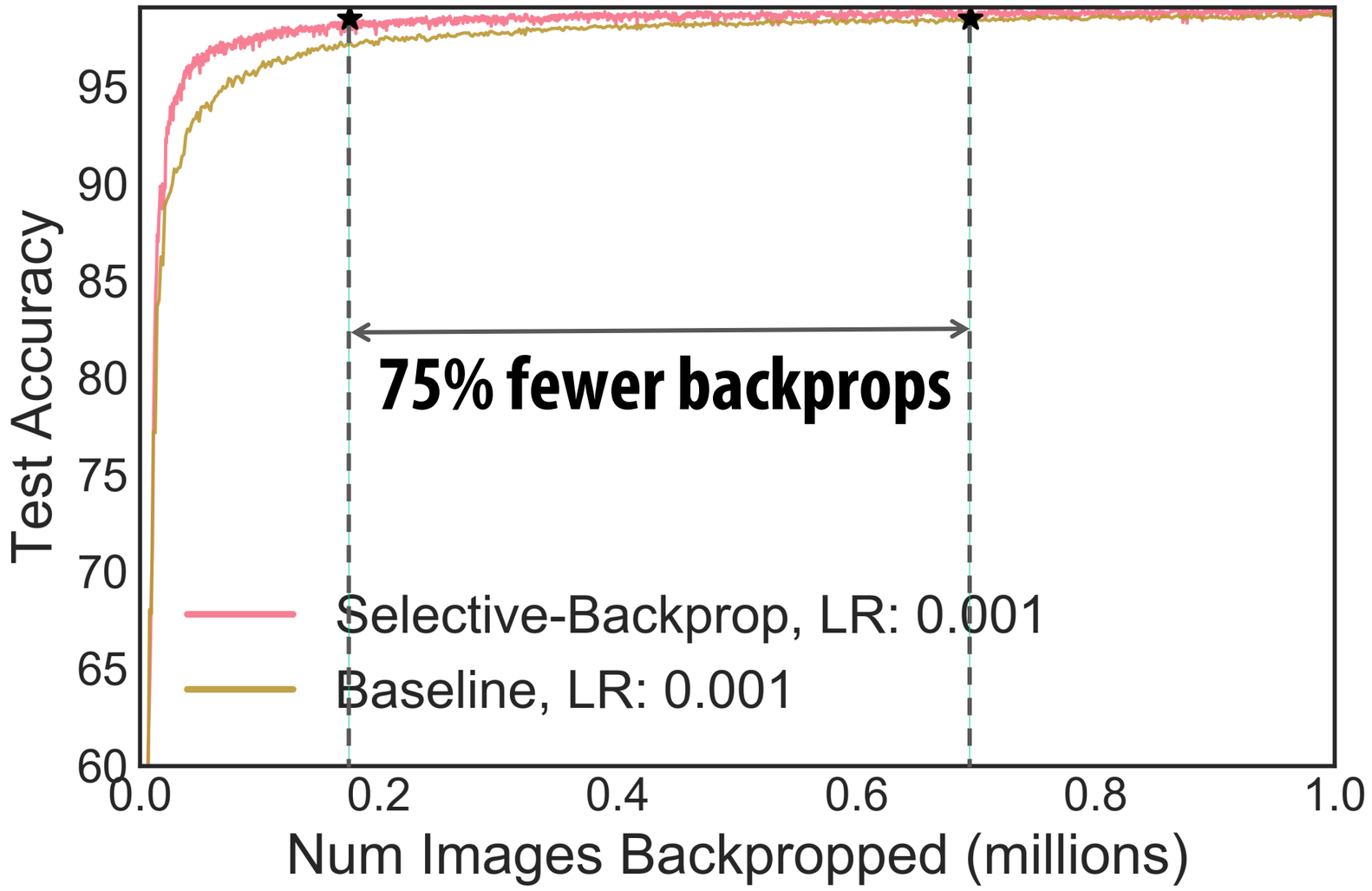
50,000 Training Images



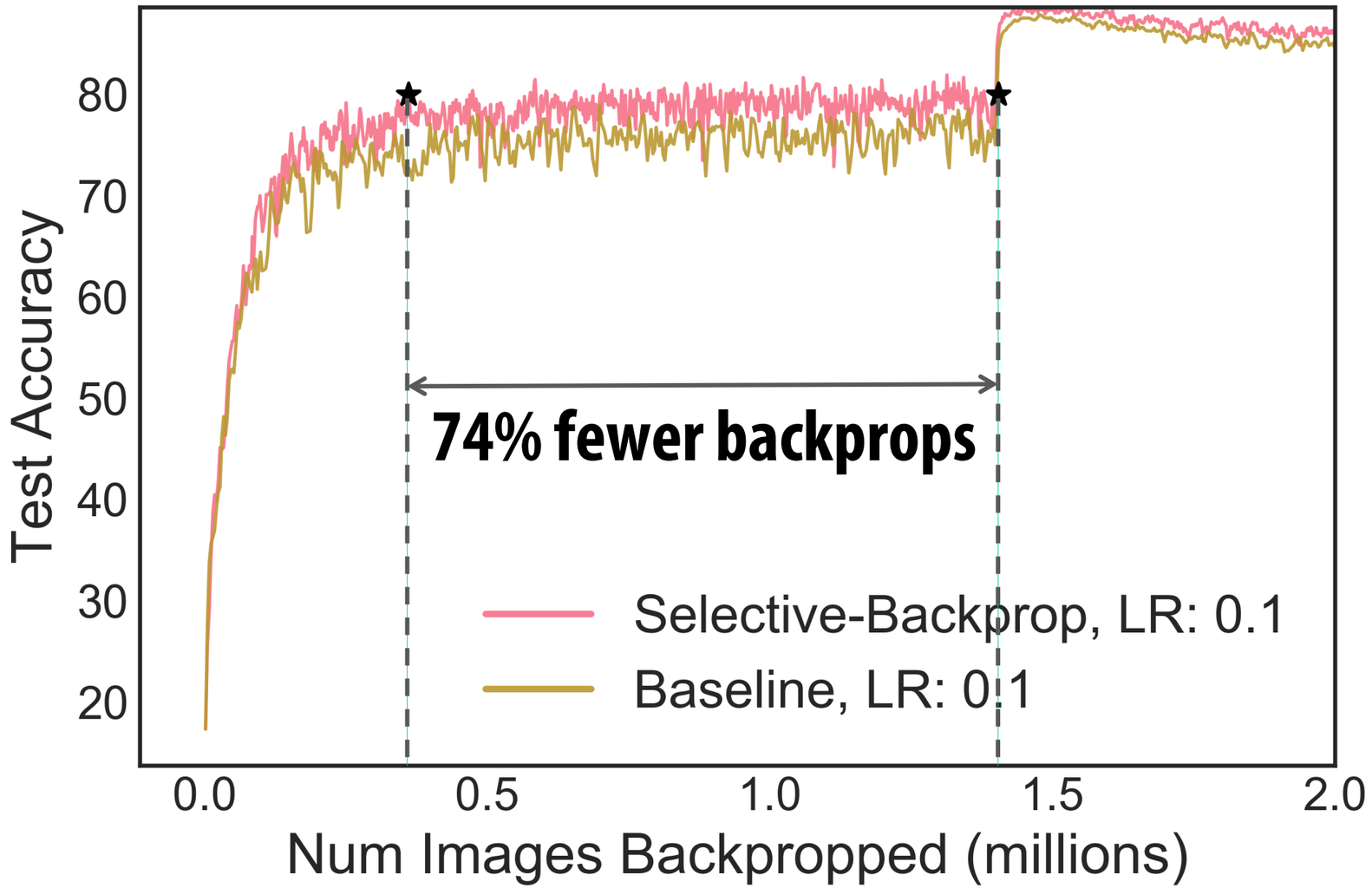
CIFAR10

60,000 Training Images

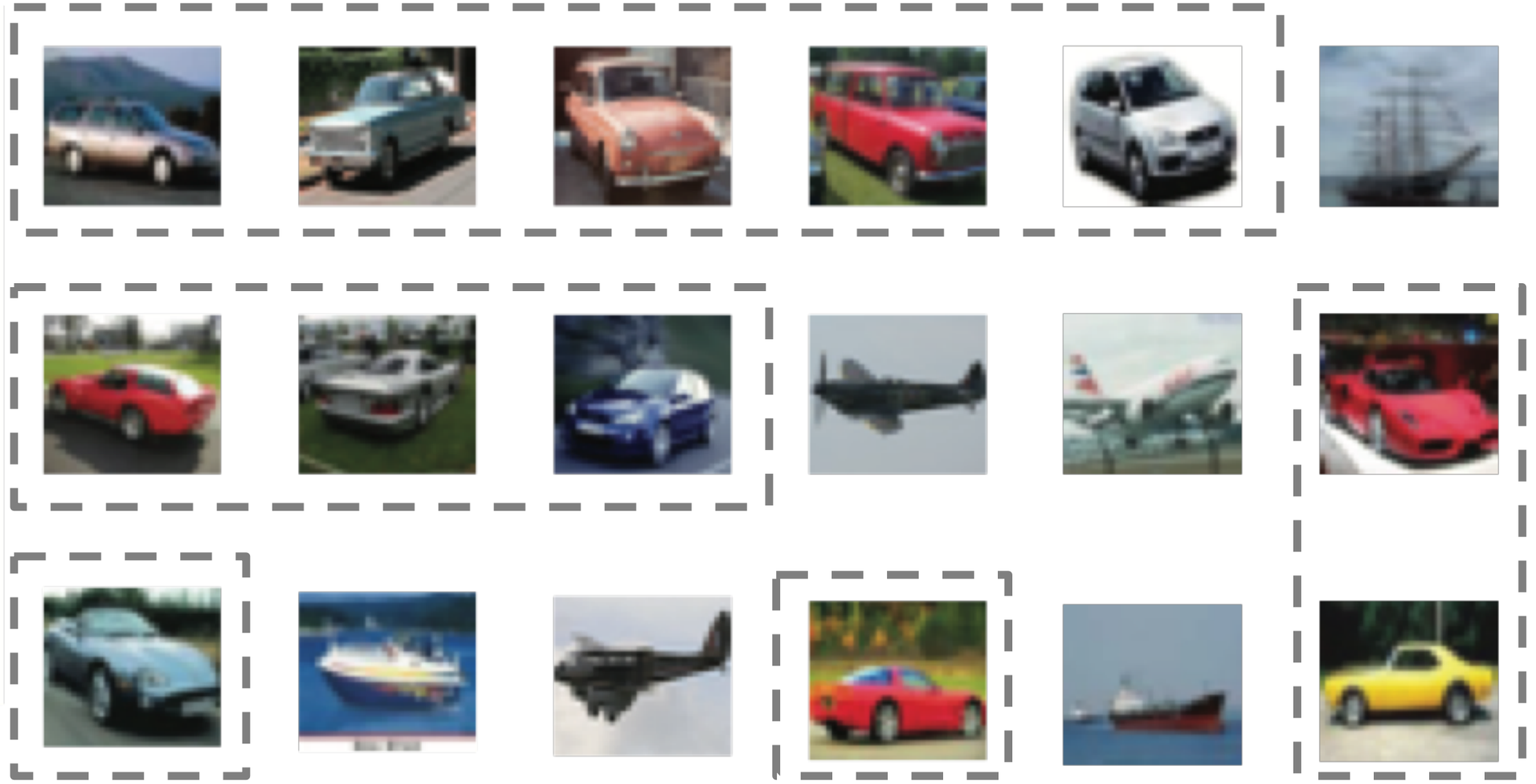
SB on MNIST gets 98.5% acc. w/ 75% fewer backprops



SB on CIFAR10 gets 80% acc. w/ 74% fewer backprops



Which images are easy?



Which images are hard?

Summary

- **Selective-Backprop** only trains on **surprising** examples
- Achieves target accuracies with fewer backprops
  - Achieves 98.5% on MNIST with **75% fewer backprops**
  - Achieves 80% on CIFAR10 with **74% fewer backprops**
- Up next: Reduce wall-clock time of training using Selective-Backprop