Accelerating deep learning with the biggest losers

Angela Jiang, Daniel Wong, Giuio Zhou, Dave Andersen, Jeff Dean*, Greg Ganger, Gauri Joshi, Michael Kaminsky^, Michael A. Kozuch†, Zachary C. Lipton, Padmanabhan Pillai†
Carnegie Mellon University; * Google AI; †Intel Labs, ^brdg.ai

Overview

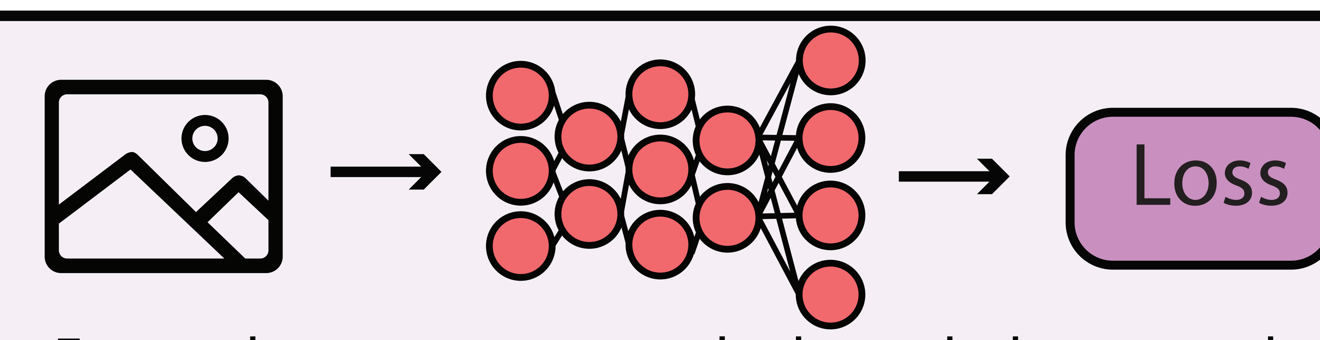
Can we speed up DNN training by backpropogating only useful examples?

Motivation
- Labeled datasets are getting larger
- Not enough time/resources to train on whole dataset (e.g., ImageNet)
- Training bottlenecked by backprop

Goal
- Speed up training by reducing the number of backprops
  - Learn from surprising examples that have more to teach the model

Approach
- Identify useful examples using inference (output of forward pass)
  - If example's output is different from target, learn from this example
- Further accelerate training by reducing the number of forwards
  - Use example's historical loss to decide if we want to backprop it

1



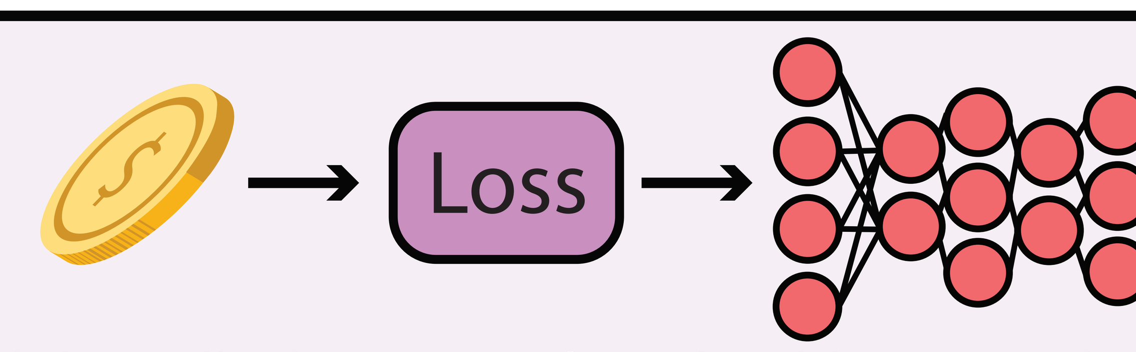
Forward propagate example through the network
(or use a historical loss from previous epoch)

2

$$P(\text{Backprop}) = \text{CDF}^b(\text{Loss}, \text{Recent losses})$$

Calculate usefulness of backpropping example based on its accuracy

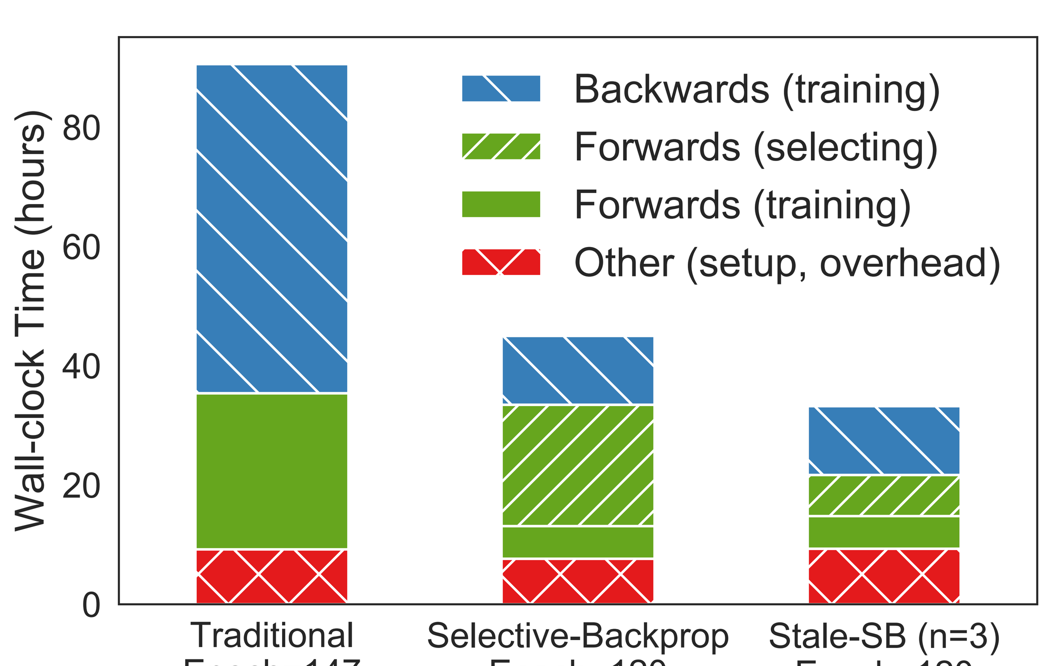
3



Probabilistically determine if we should train on the example

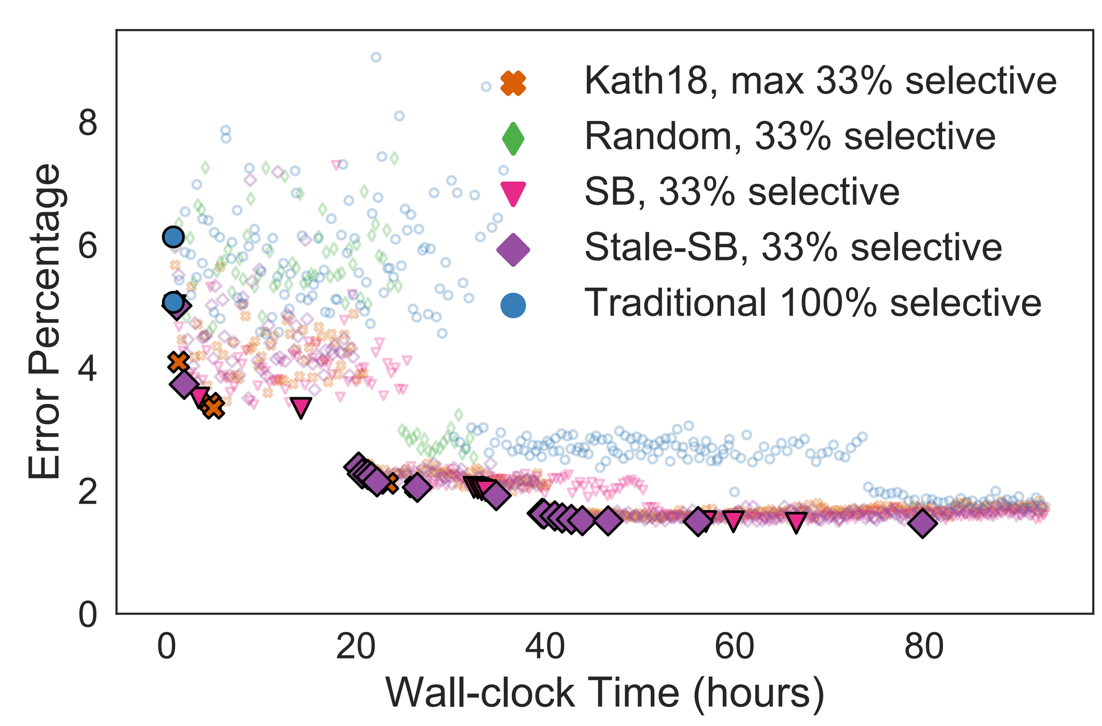
Training with Selective-Backprop (SB) and StaleSB

Time to train SVHN to 1.72%



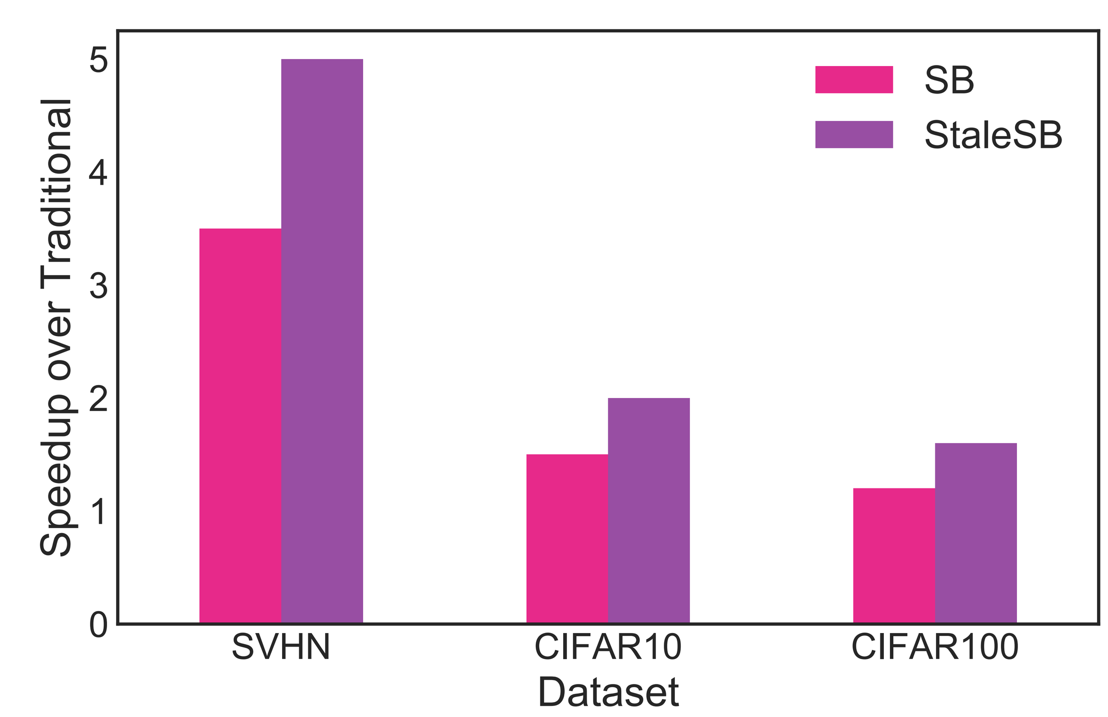
- SB reduces time spent in backwards
- StaleSB reduces time spent in forwards
  - Runs selecting passes every 3 epochs
- SB reaches same final accuracy

SVHN training time vs. Error



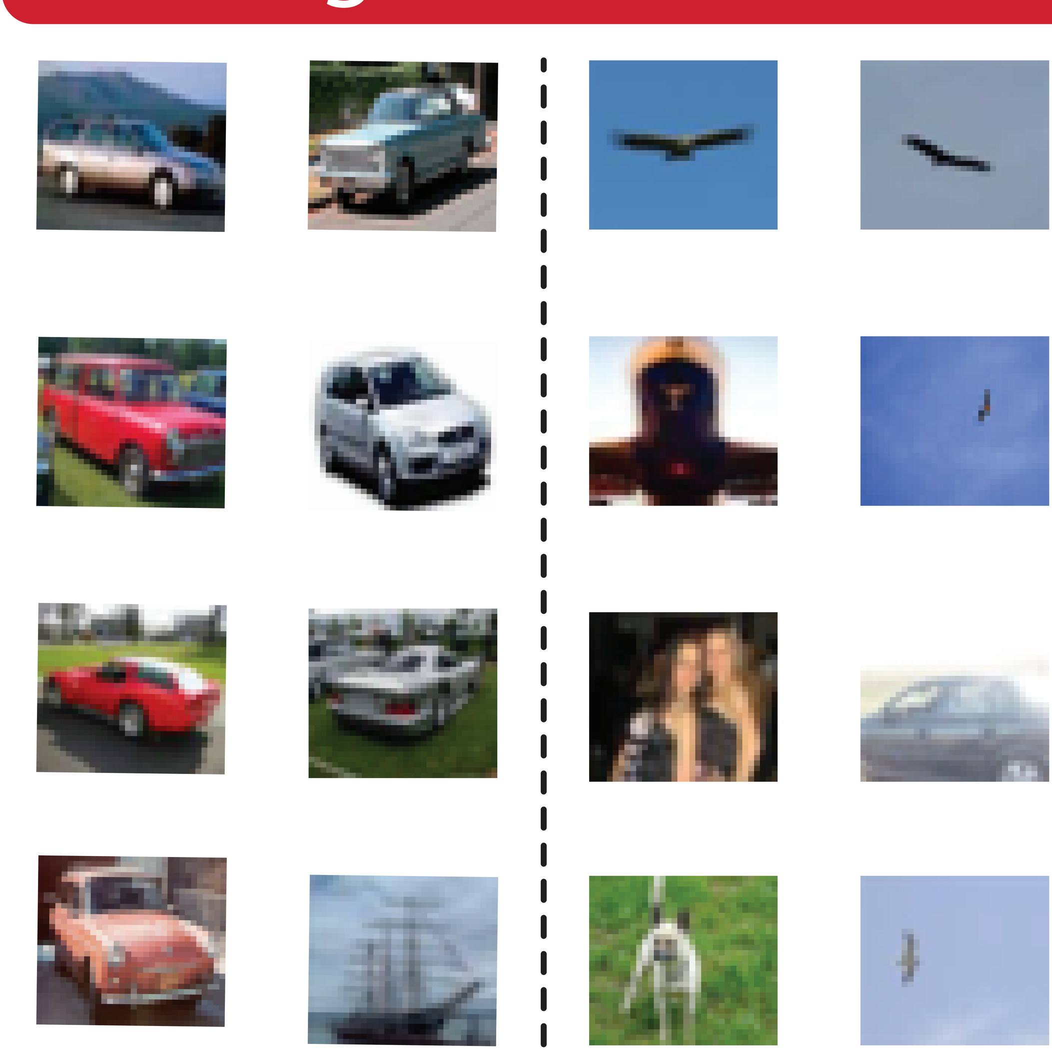
- Kath18 is a SoA sampling technique
- Pareto optimal points enlarged
- For almost all training time budgets
  - SB or Stale-SB reaches lowest err

Speedup over Traditional

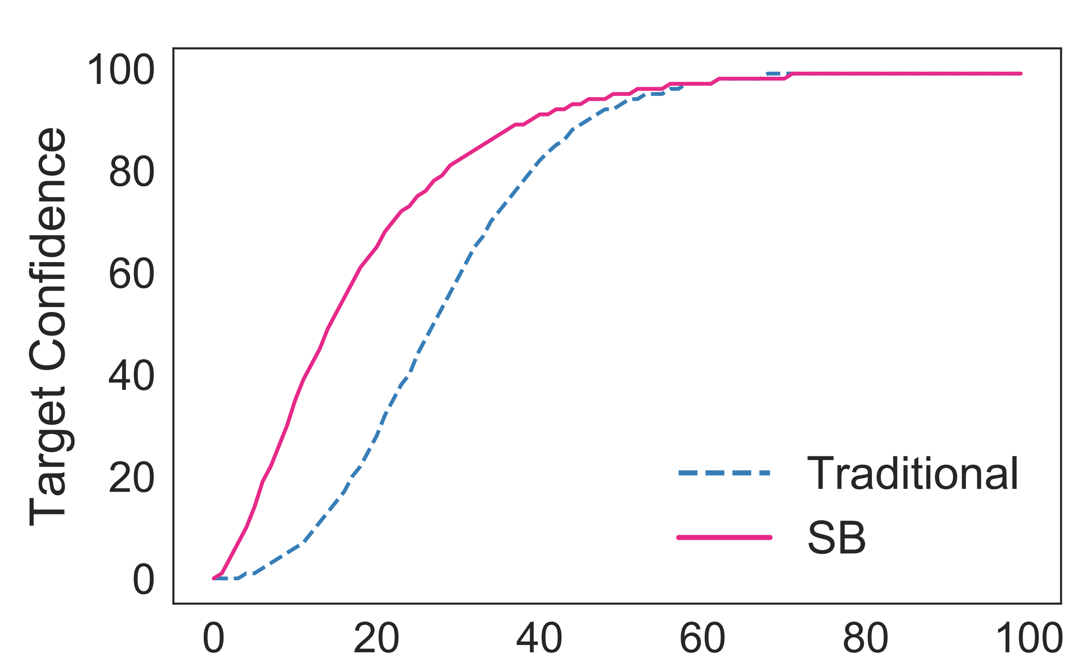


- Traditional does not filter examples
- Compare time to reach 1.4x of best acc
- SB accelerates training by up to 5x
- StaleSB accelerates SB by avg of 26%

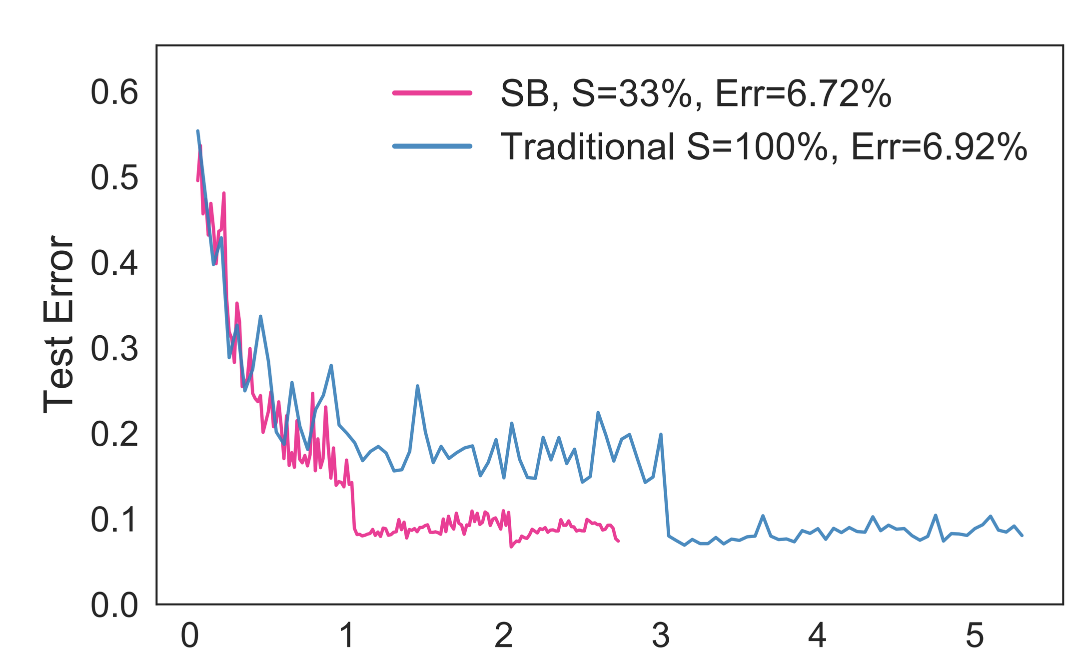
Diving into CIFAR10



Easy Examples Hard Examples



- Y-axis is conf in our pred of correct class
  - On test examples after 10 epochs
- SB improves conf of hard examples
  - W/out sacrificing acc of easier examples



- Training CIFAR10 w/ 10% randomized labels
- SB accelerates training despite label err
- SB tolerates modest amounts of label err
- SVHN known to have label error too