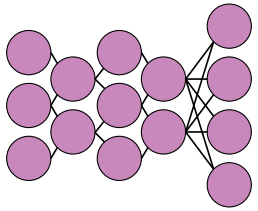
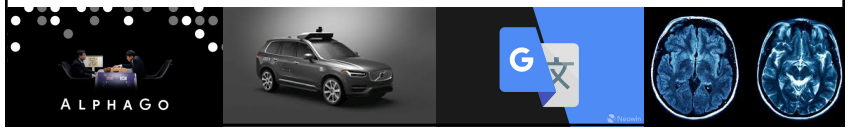
Accelerating Deep Learning by Focusing on the Biggest Losers

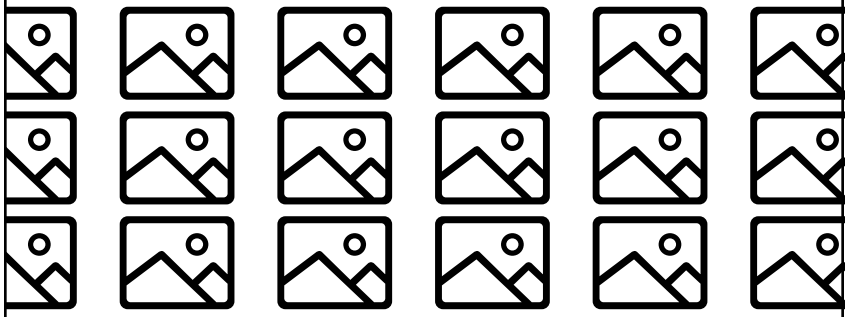
Angela H. Jiang, Daniel L.-K. Wong, Giulio Zhou,
David G. Andersen, Jeffrey Dean, Gregory R. Ganger,
Gauri Joshi, Michael Kaminsky, Michael A. Kozuch,
Zachary C. Lipton, Padmanabhan Pillai

Deep learning enables emerging applications

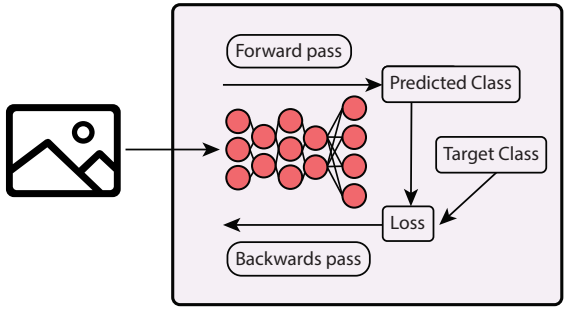




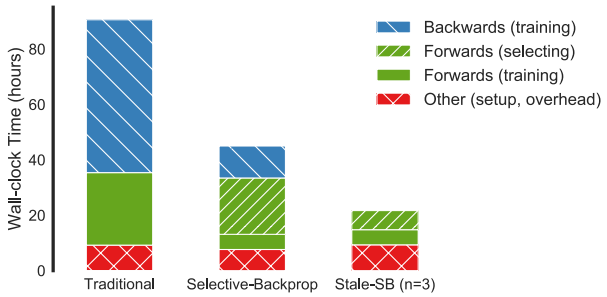
DNN training analyzes many examples



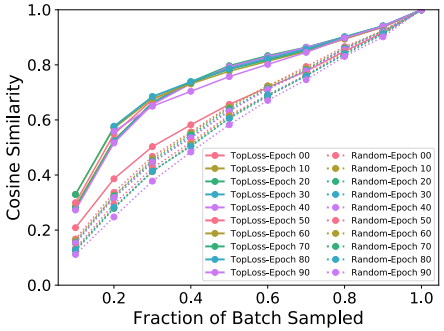
Deep learning training is slow



SelectiveBackprop targets slowest part of training

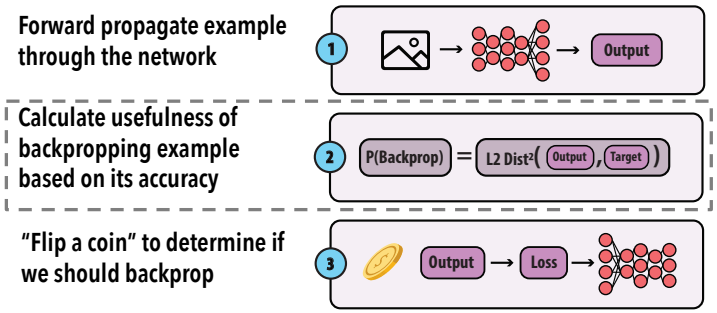


Using loss as an indicator of usefulness



Selective Backprop algorithm

Selective-Backprop approach



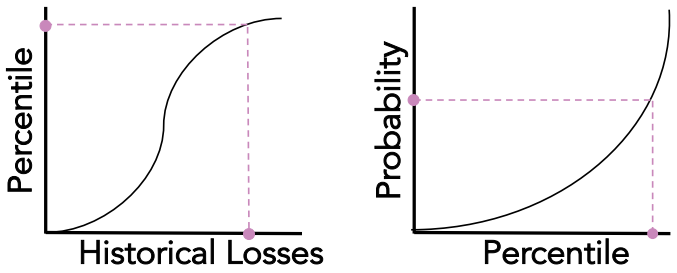
Bad idea #1:
Deciding with a hard threshold

if loss < threshold: backprop()

Bad idea #2:
Deciding probabilistically with absolute loss

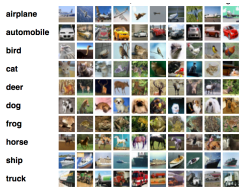
P(backprop) = clamp(loss, 0, 1)

Good idea:
Use relative probabilistic calculation



Evaluation of Selective Backprop

Datasets



CIFAR10
60,000 Training Images

CIFAR100
60,000 Training Images

SVHN
604,388 Training Images

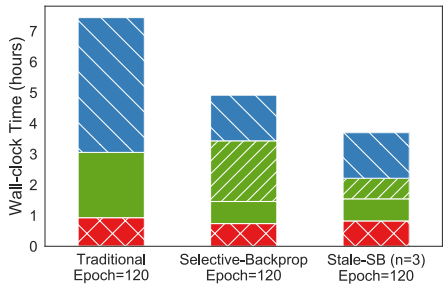
Compared approaches

Traditional
Classic SGD with no filtering

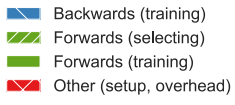
Katharopoulos18
State of the art importance sampling approach

Selective-Backprop (Us)

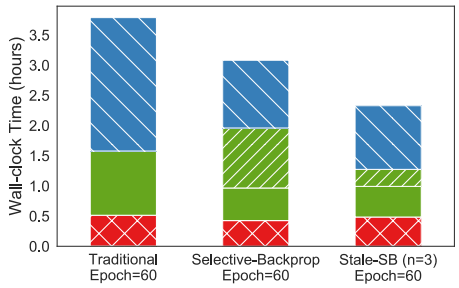
Train CIFAR10 to 4.14% (1.4x Traditional's final error)



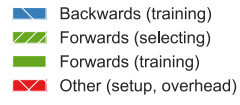
SB: 1.5X faster
StaleSB: 2X faster



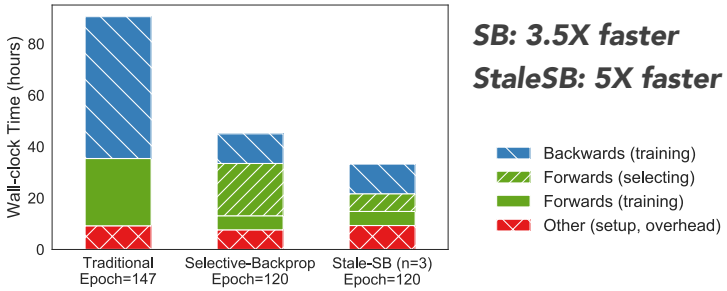
Train CIFAR100 to 25.5% (1.4x Traditional's final error)



SB: 1.2X faster
StaleSB: 1.6X faster



Train SVHN to 1.72% (1.4x Traditional's final error)



SB: 3.5X faster

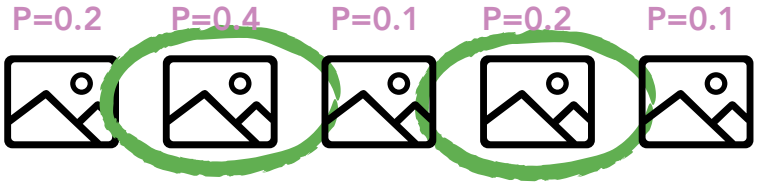
StaleSB: 5X faster

Katharopoulos18's approach

Katharopoulos18's approach

Katharopoulos18's approach

Static selectivity = 3 / 5

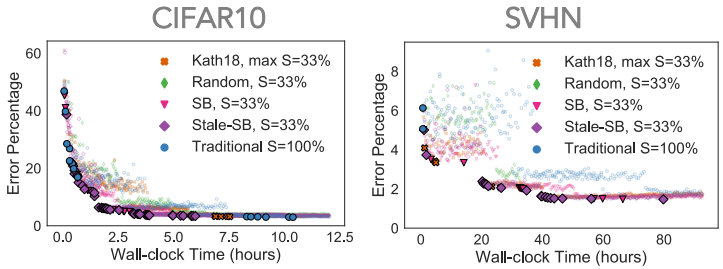


Katharopoulos18's approach

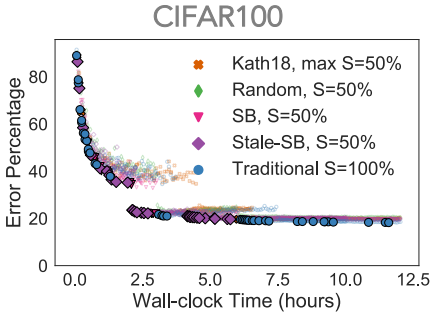
Static selectivity = 1 / 5

P=0.9 P=0.9 P=0.1 P=0.3 P=0.8

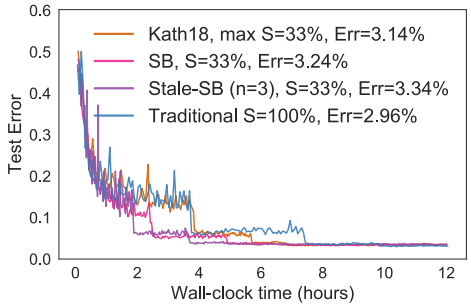
Most Pareto optimal points are SB or StaleSB



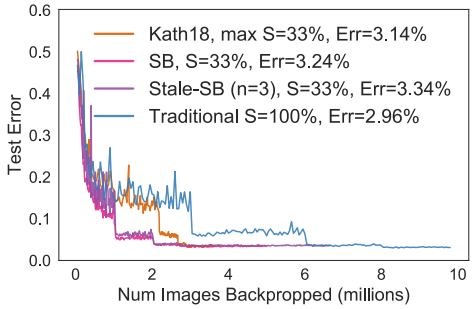
Most Pareto optimal points are SB or StaleSB



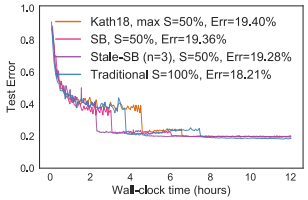
SB reduces wall-clock time for many error rates



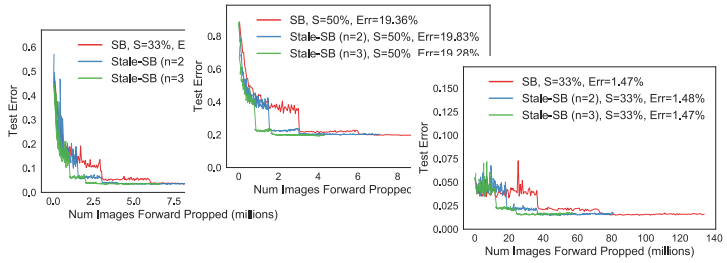
SB reduces iterations for a many target errors



SB saves training time on CIFAR100



StaleSB saves forward passes



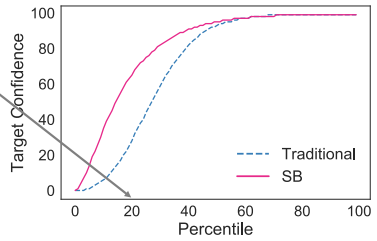
Why does SB perform well?

SB on CIFAR10 focuses on hard examples

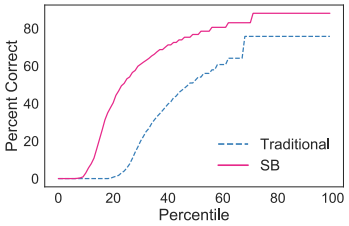
3% examples correct w/ Traditional

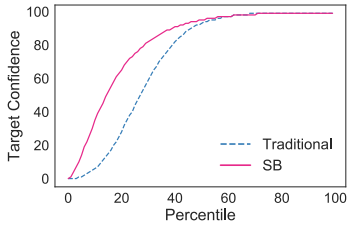
29% examples correct w/ SB

Target confidence = 0.3

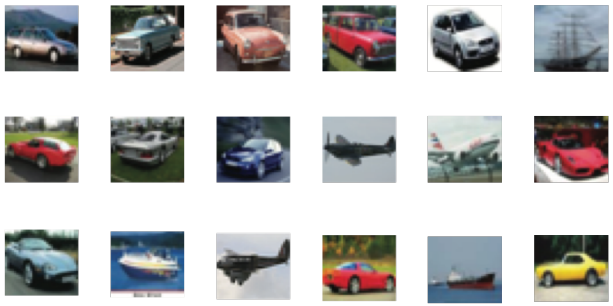


SB on CIFAR10 focuses on hard examples





Which images are easy?

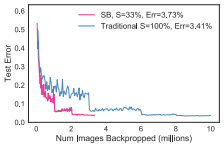


Which images are hard?

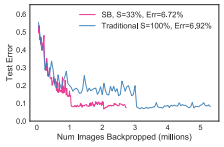
Is SB robust to label error?

SB is robust to modest amounts of error

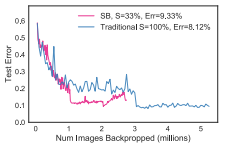
0.1% Randomized



10% Randomized



20% Randomized



Selective-Backprop accelerates training
Reduces time spent in the backwards pass by prioritizing high-loss examples

SelectiveBackprop outperforms static approaches
Trains up to 3.5x faster compared to standard SGD
Trains 1.02-1.8X faster than state-of-the-art importance sampling approach

Stale-SB further accelerates training
Trains on average 26% faster compared to SB

www.github.com/angelajiang/SelectiveBackprop