e-Lab

Flattened Convolutional Neural Networks for Feedforward Acceleration

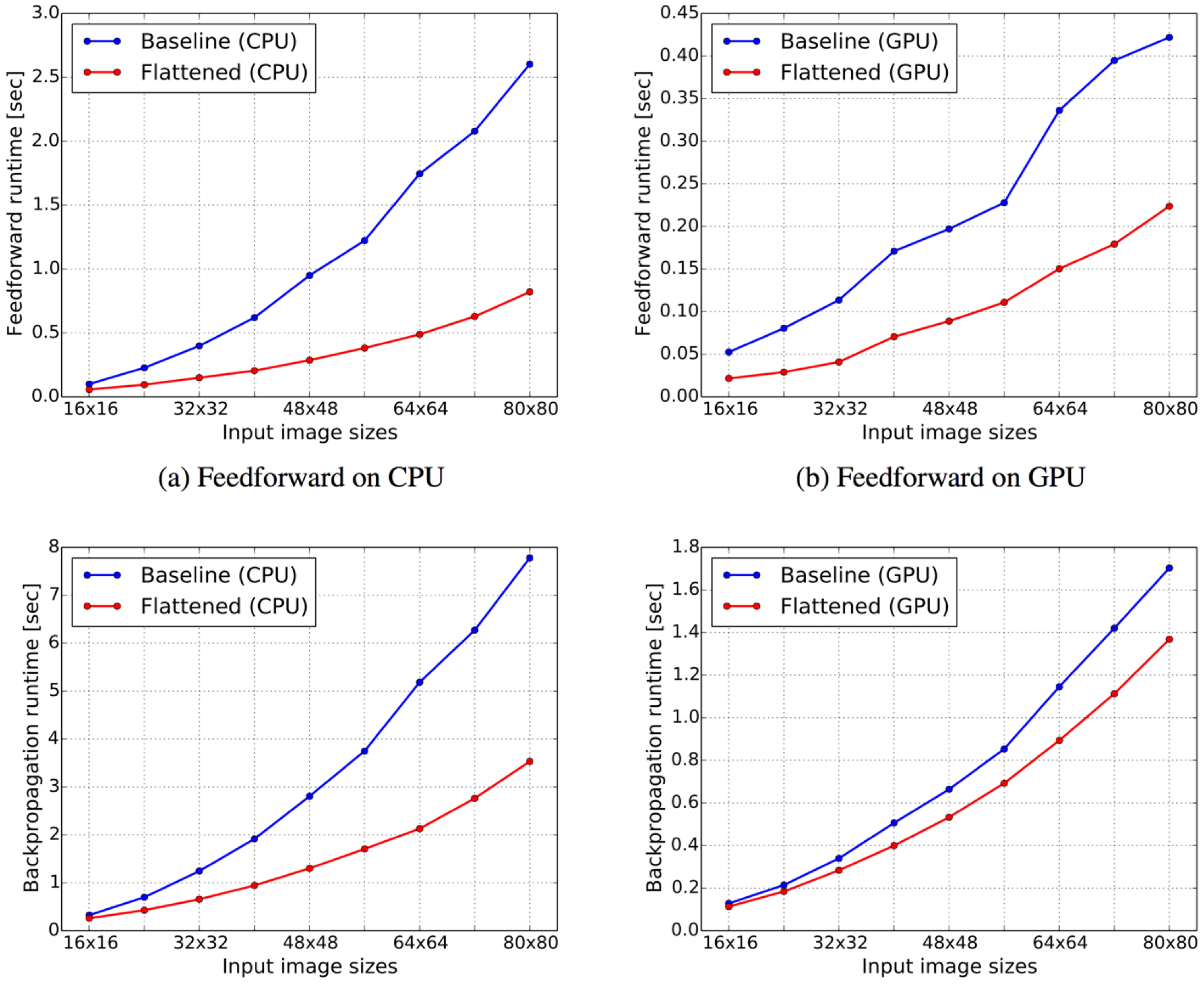
Jonghoon Jin, Aysegul Dundar and Eugenio Culurciello

PURDUE UNIVERSITY

Abstract

We present flattened convolutional neural networks that are designed for fast feedforward execution. The flattened layer, consisting of a sequence of 1D filters across all directions, can effectively substitute for the 3D filters without loss of accuracy. The flattened convolution pipelines provide around **2x speed-up** during feedforward pass with **90% parameter reduction**. Furthermore, the proposed method does not require efforts in manual tuning or post processing once the model is trained.

Acceleration



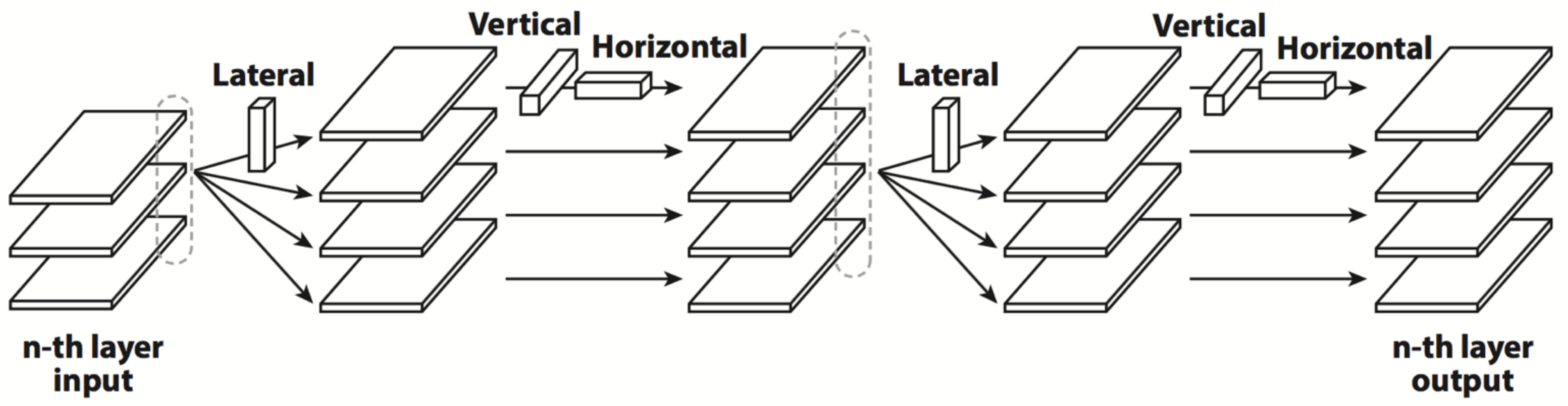
(a) Feedforward on CPU

(b) Feedforward on GPU

(c) Backpropagation on CPU

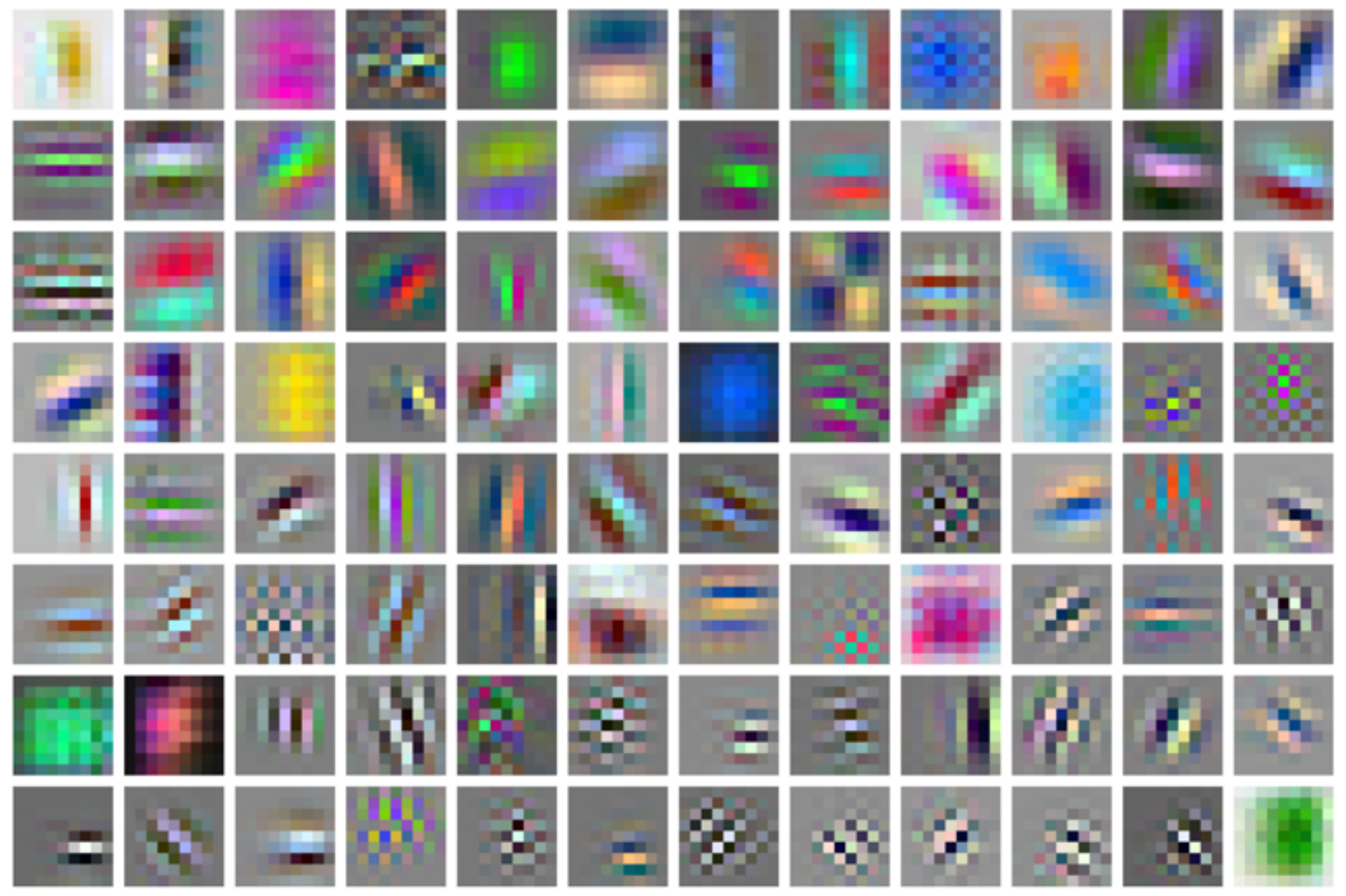
(d) Backpropagation on GPU

Flattened ConvolutionLayer



1D convolutions over channels (Lateral) and in space (Vertical / Horizontal)

Reconstructed Filters



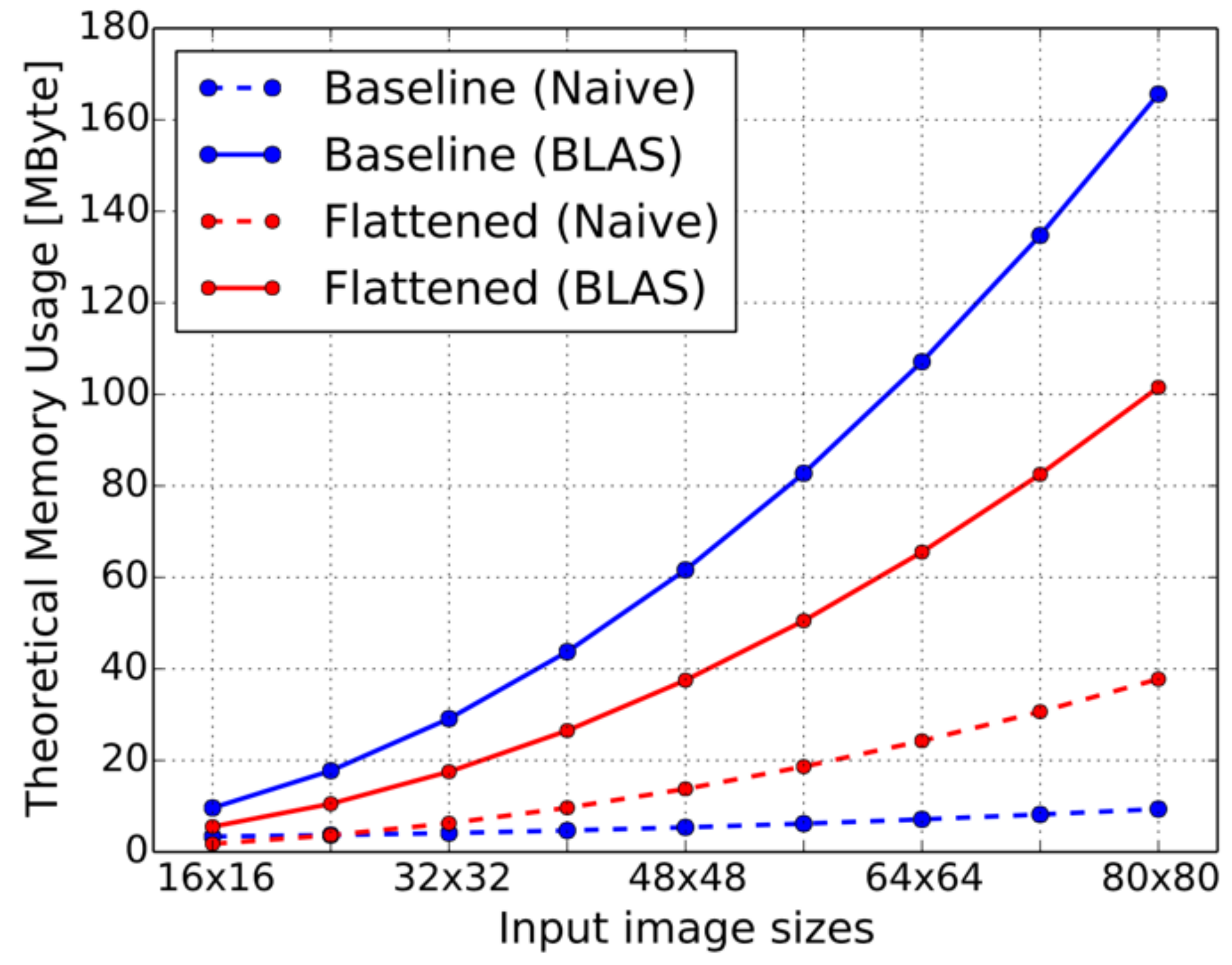
Sparse and sharp
edge filters in 1st layer

Classification Accuracy

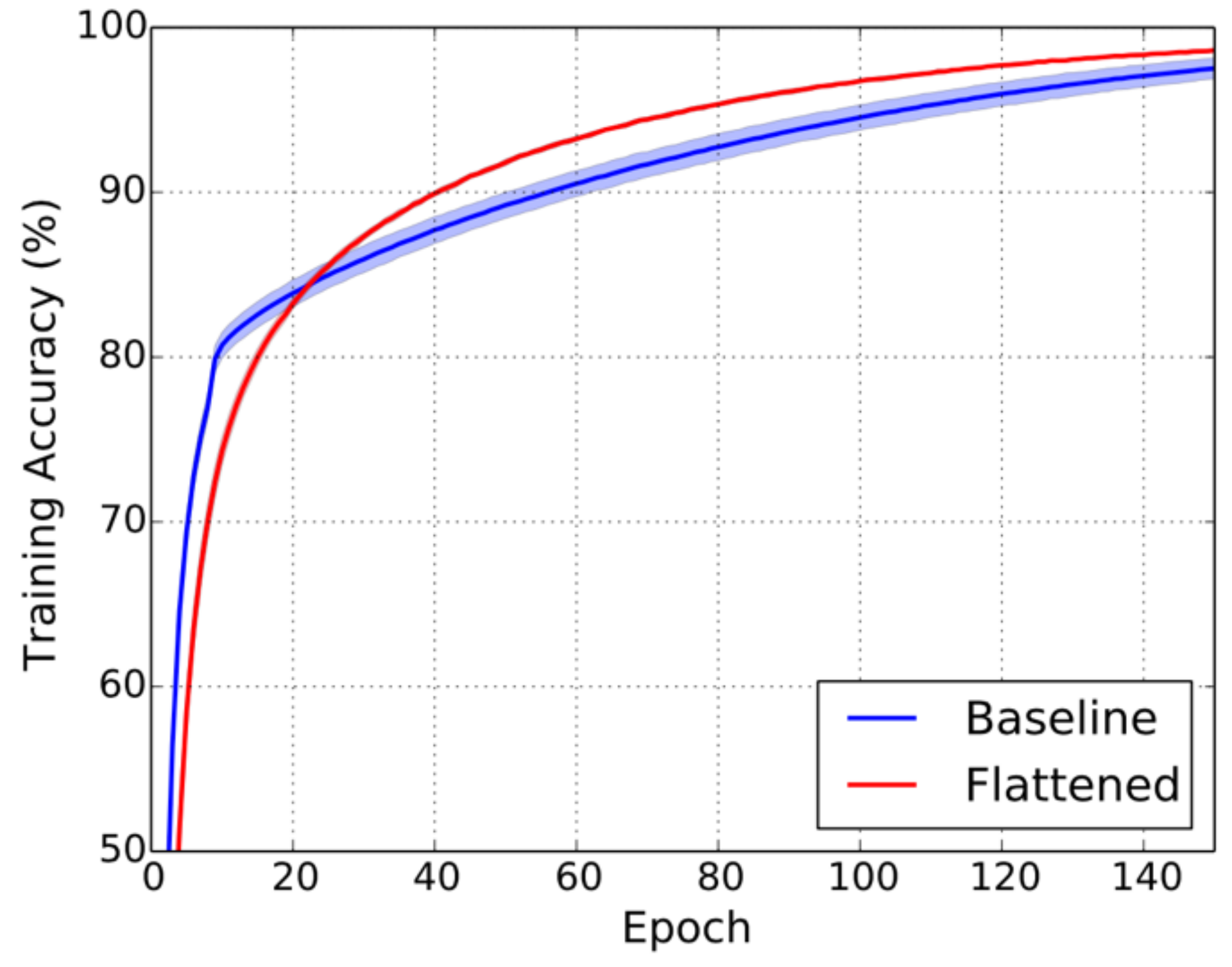
Dataset	Model Type	Test Accuracy
CIFAR-10	Baseline Model Flattened Model	86.42% 87.04%
CIFAR-100	Baseline Model Flattened Model	60.08% 60.92%
MNIST	Baseline Model Flattened Model	99.62% 99.56%

Comparable performance
as vanilla CNNs

Memory Usage



Convergence Rate


Training Accuracy (%)
Epoch
Baseline
Flattened