Teacher-Student Learning

Eric He

Who is this guy?

- My name is Eric He
- Third-year student at NYU
- Grew up in Bay Area, Evergreen Valley High School
- Wanted to do marketing, now studying mathematics and data science
- Strong interest in explanatory modeling

What is Teacher-Student Learning?

- An alternative method of training neural nets
- Normal loss functions penalize deviation of student model from the ground truth label
- Teacher-Student loss penalizes deviation of student model from the teacher's predictions, or **soft labels**

	Panther	Cat	Truck	Goose
Ground Truth	1	0	0	0
Soft Labels	0.7	0.28	0.015	0.005

Dark Knowledge: A panther is more like a cat than a truck or a goose

Image source: http://awesomejelly.com/panther-pinata/

Potential of Teacher-Student Learning

gyrfalcon technology

- Data flexibility
  - May not need to expend resources to label training data
  - Customers do not need to hand over proprietary data used to train models
- Model compression
  - train a smaller model for deployment
- Model diversity
  - Different architectures learn different features, which can all be transferred onto the student model

Infrastructure

gyrfalcon technology

Machine Learning

Caffe

python™

Hardware

NVIDIA®

ubuntu®

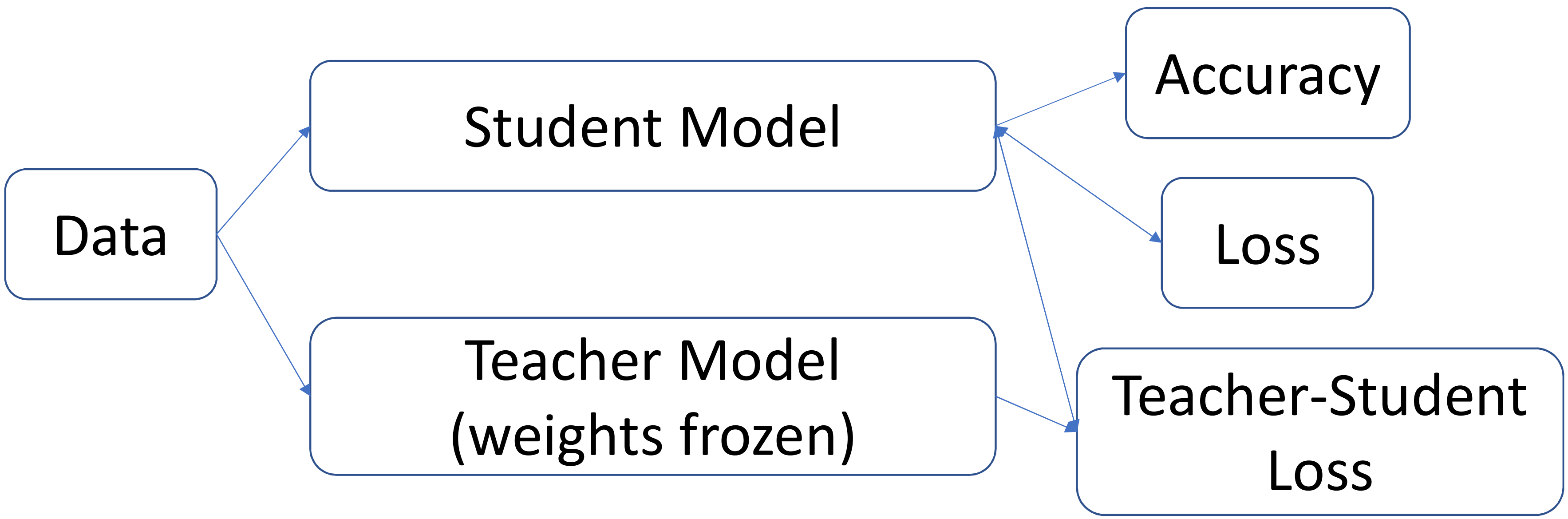
amazon web services™

Graphics

ggplot2

R

Setting up Teacher-Student Modeling



Method: Stack both teacher and student models into one .prototxt file

Problems: Highly memory intensive, both models require same data dimensions

- Batch size of 20 with AlexNet teacher and VGG16 student takes 12 GB of GPU memory

Three Performance Metrics

gyrfalcon technology

Accuracy

Proportion of predicted top1 labels in concordance with the true labels

Loss

Sum of softmax cross-entropy scores between student-generated probability distributions and true labels

Teacher-Student Loss

Sum of softmax cross-entropy scores between student and teacher-generated probability distributions

Cross Entropy Loss

$$\mathcal{L}(y, \hat{y}) = -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$$

For the purposes of visualization, Loss and Teacher-Student Loss are graphed as percentages of their maximum.

MNIST
- 60000 training images
- 20 x 20 x 1 (grayscale)
- 10 categories

CIFAR10
- 60000 training images
- 32 x 32 x 3 (color)
- 10 categories

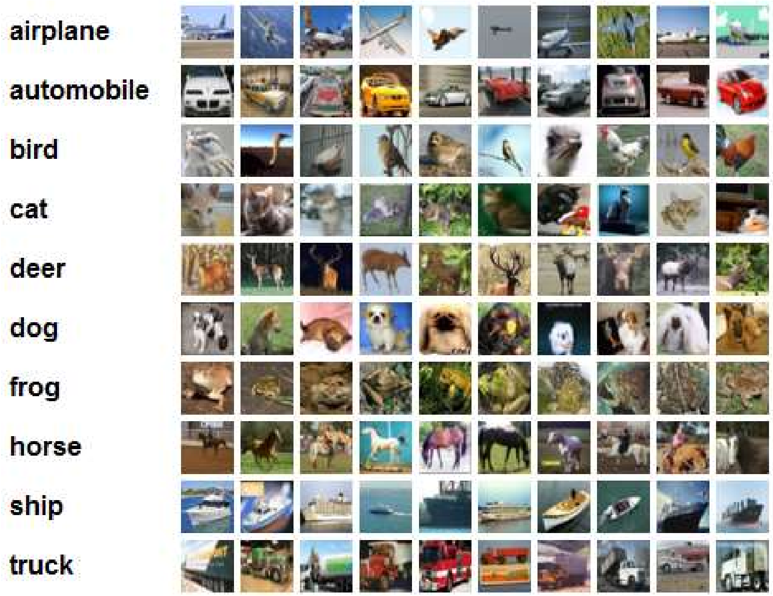


Image sources: https://github.com/cazala/mnist, http://karpathy.github.io/2011/04/27/manually-classifying-cifar10/

LeNet Structure

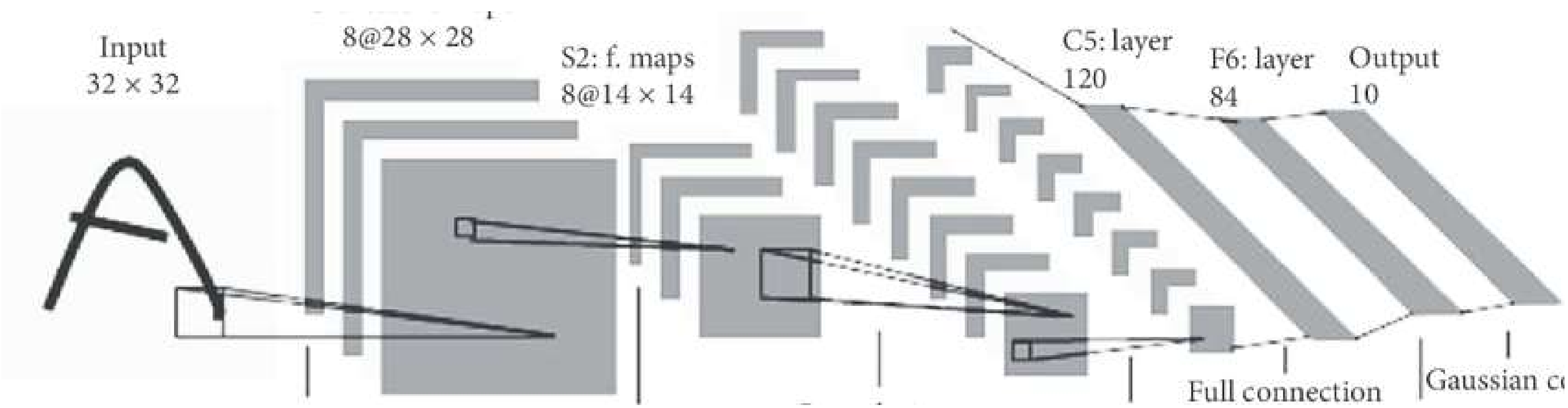
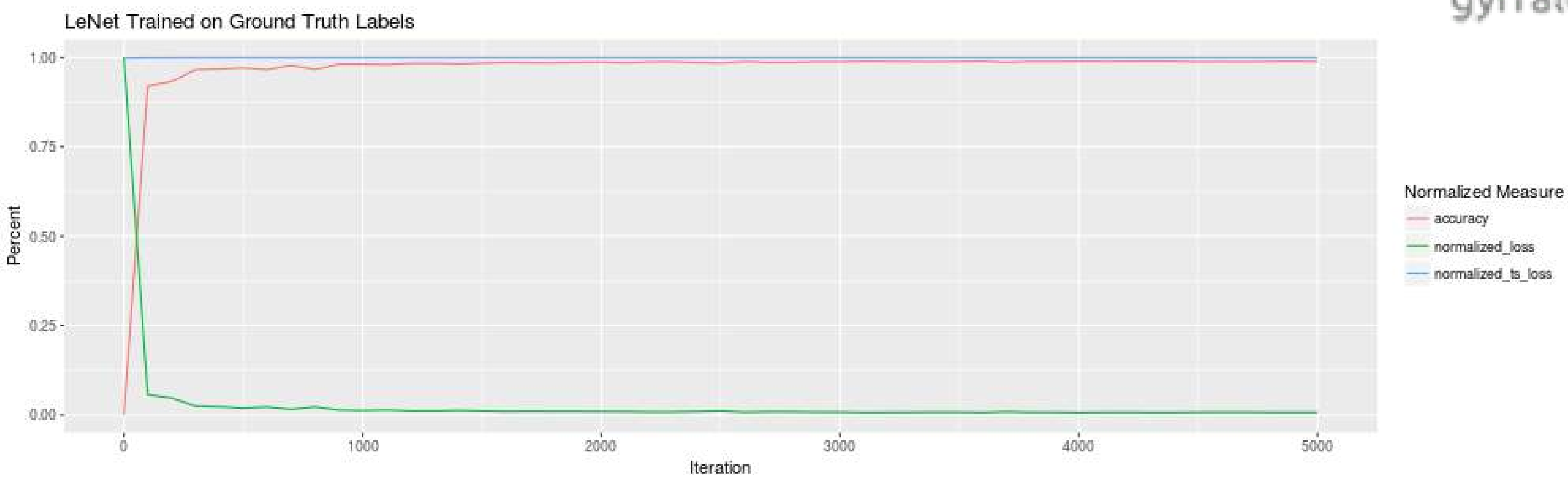
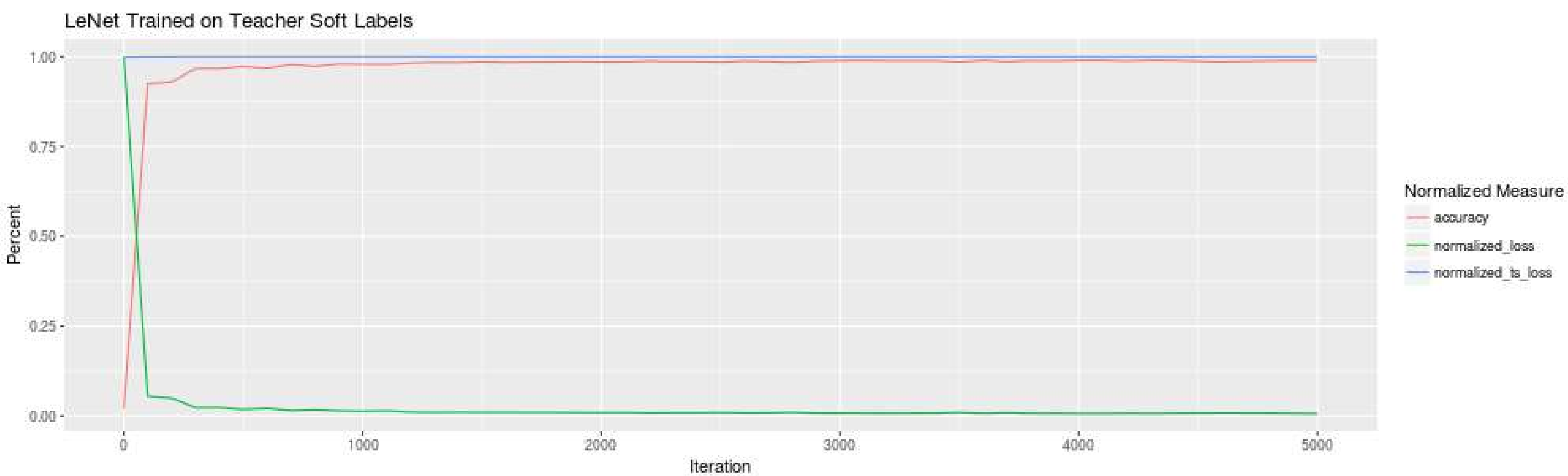


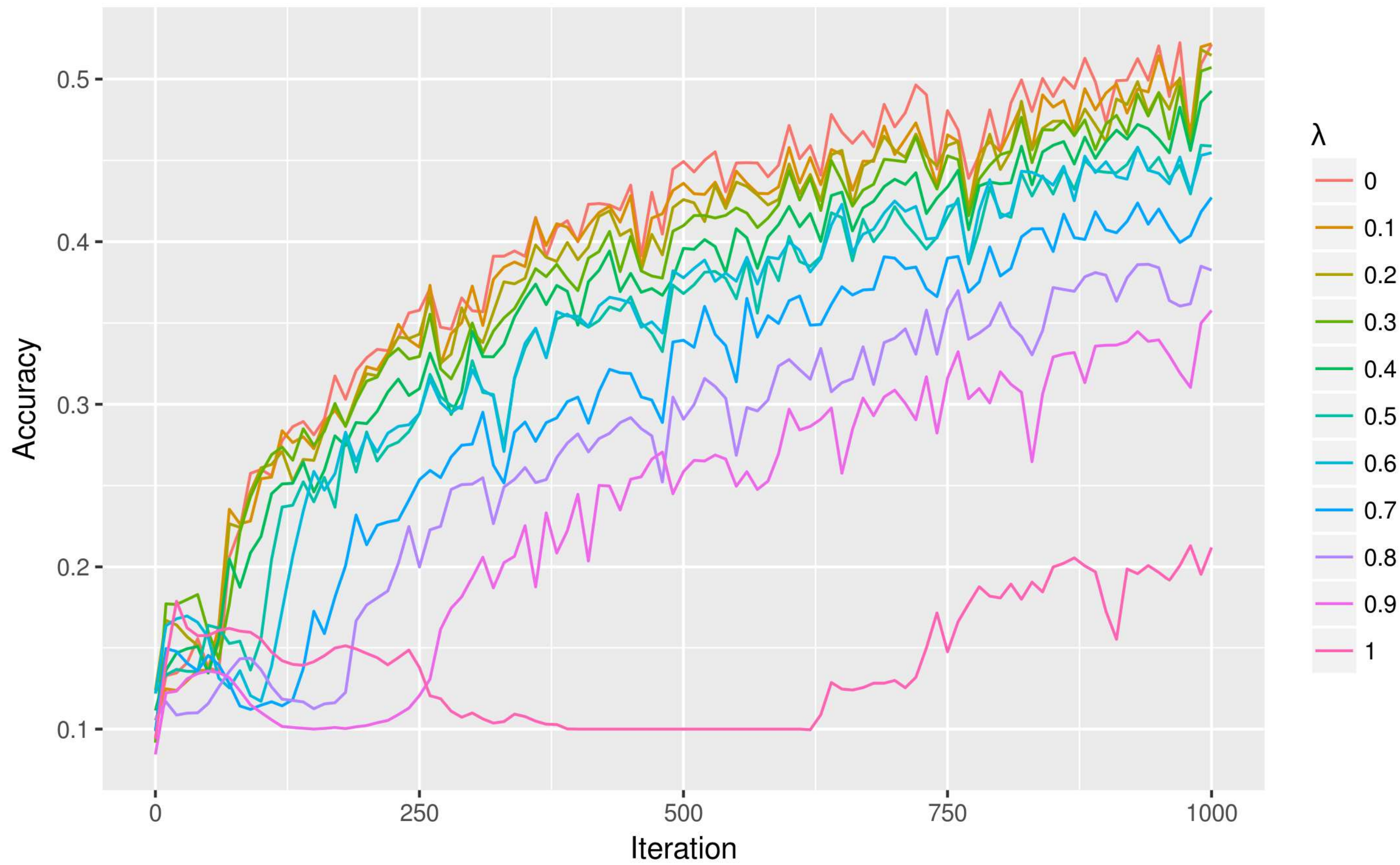
Image source: https://www.researchgate.net/figure/303857877_fig1_Figure-1-Architecture-of-CNN-by-LeCun-et-al-LeNet-5

No obvious differences between training regimes

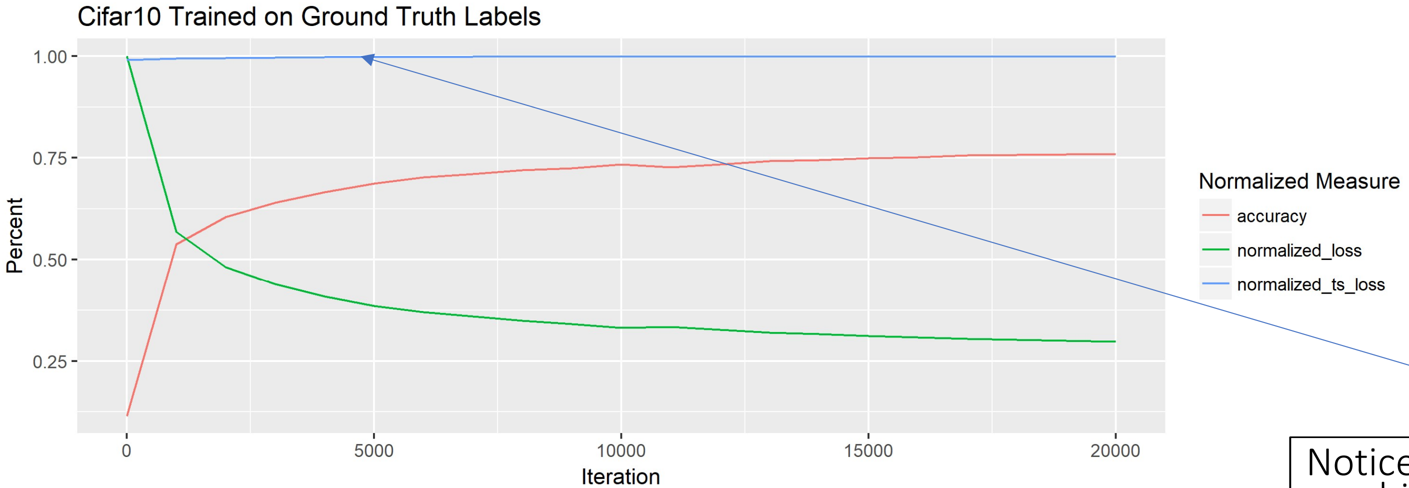




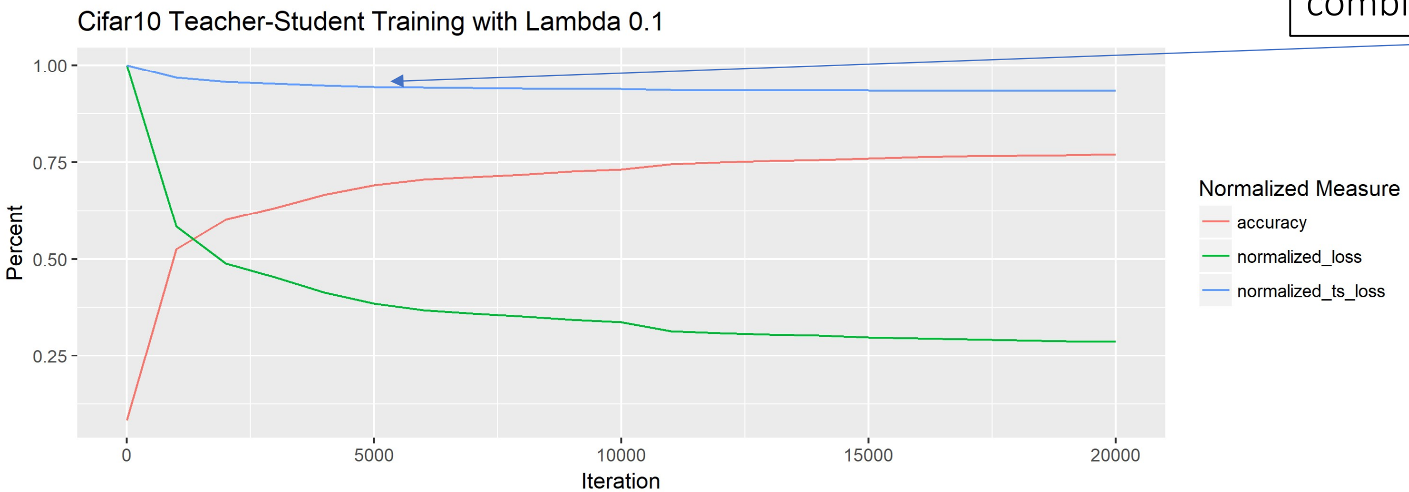
Soft labels slow training on MNIST!



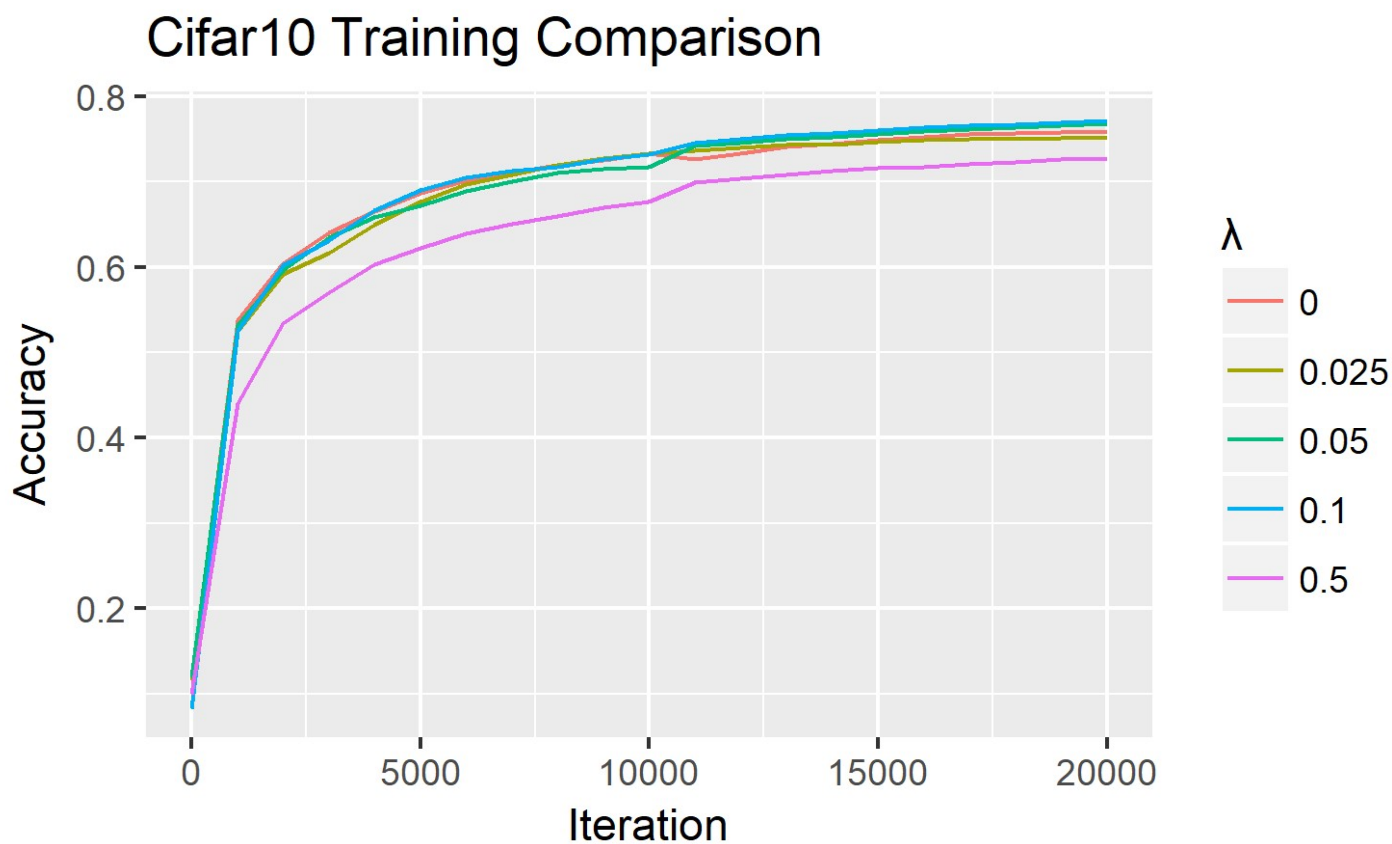
Combination Training works on CIFAR10 as well



Notice small drop in TS cross-entropy in combination training



In fact, low-weighted soft labels are faster for early epochs



Position
3
4
1
2
5

What about a smaller student model?

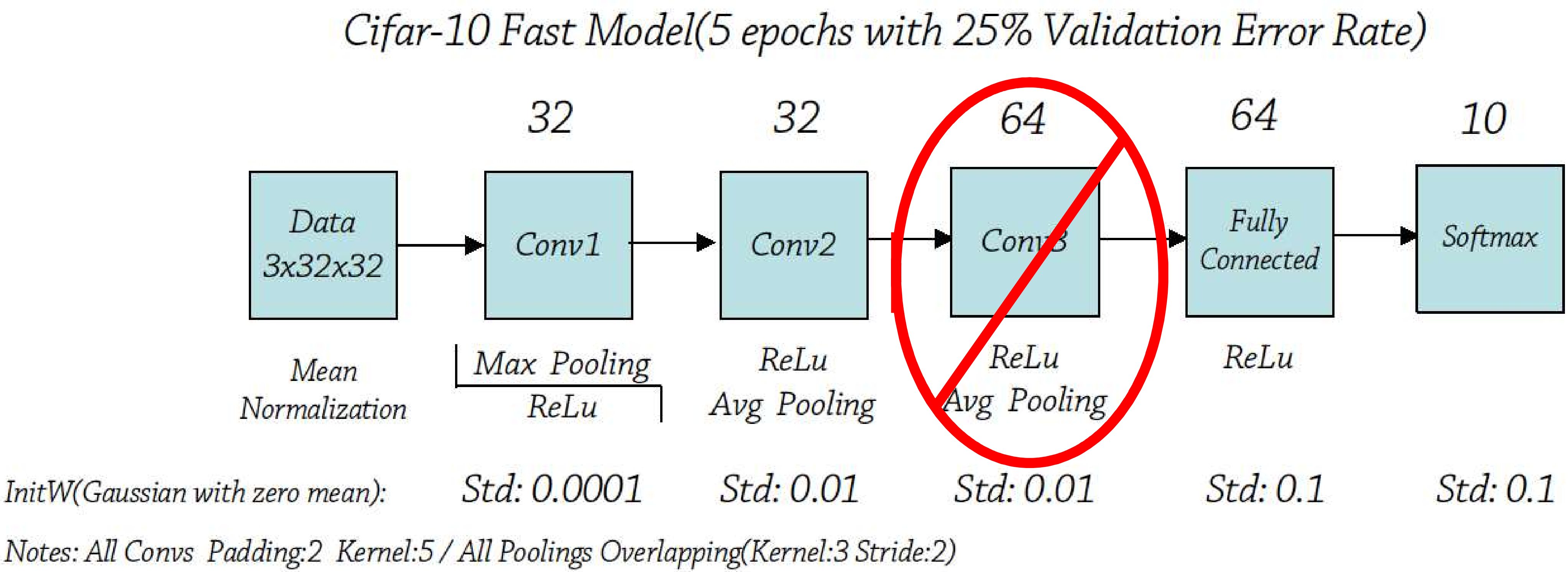
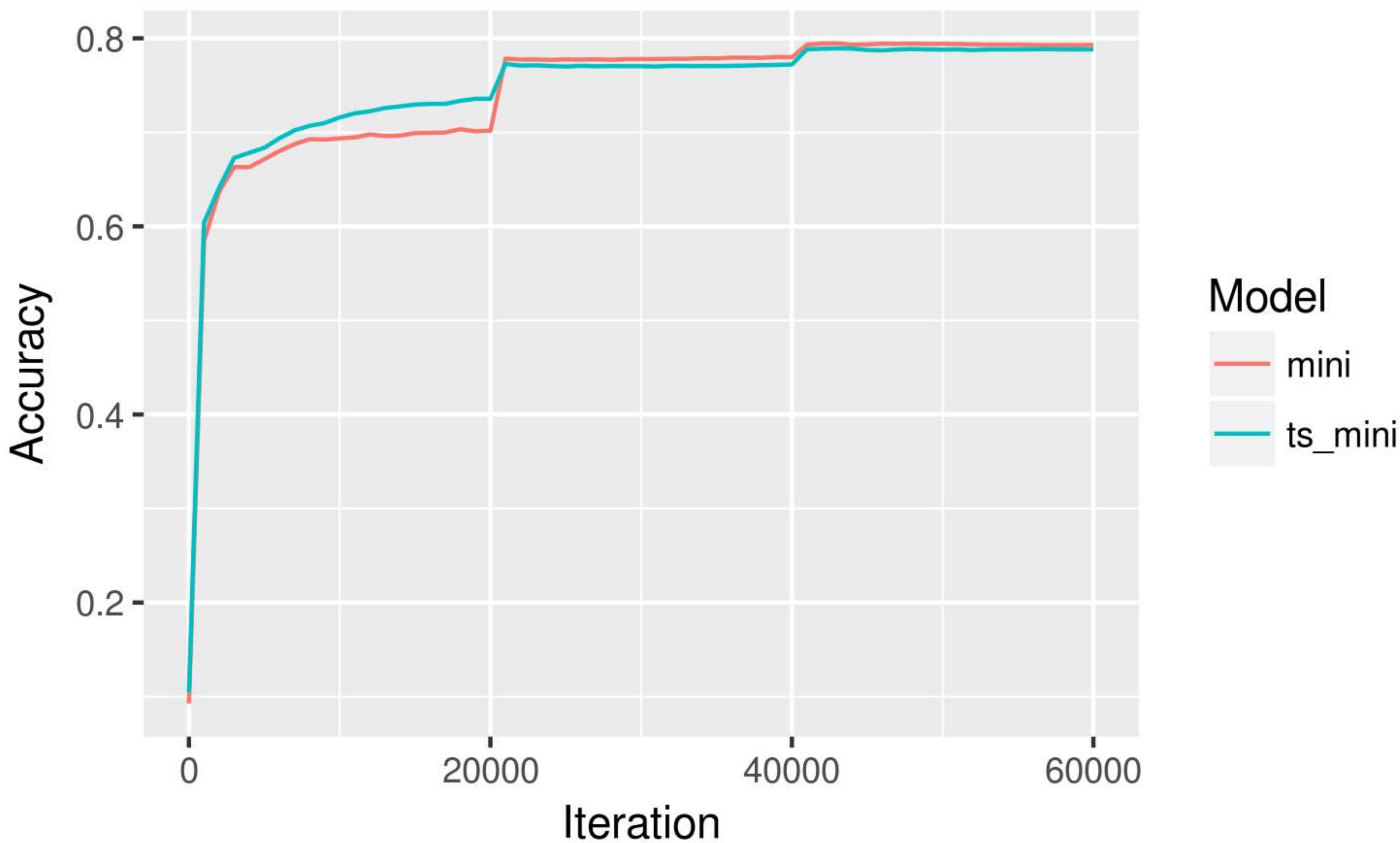


Image source: http://www.cnblogs.com/yymn/p/4589133.html

Training is faster for early epochs, but plateaus earlier



Conclusions

- Teacher-Student training IS possible!
- In combination with hard labels, soft labels appear to be a training speedup for larger datasets, slowdown for smaller datasets
- Student model is unable to make close approximation of teacher decision surface
  - Student-teacher cross-entropy remains large despite classification accuracies being about the same
  - Student and teacher models do not learn features the same way!

Looking forward

- Bigger datasets; ImageNet
  - Two of three months of the internship was spent trying and failing to make VGG16 model converge.
  - Teacher-Student training question remains unresolved for big datasets, where its potential is largest
- Best temperature for teacher soft labels?
  - A lower temperature raises the top1 probability of soft label towards 1 and decreases other class label probabilities towards 0
    - Makes teacher soft label more like hard label
  - A higher temperature makes all class probabilities closer to each other
    - Teacher models with low top1 accuracy but high top5 accuracy will be more informative
- Better loss function?
  - Cross-entropy is combination of entropy and KL-divergence
  - KL-divergence measures “distance” between two probability distributions
  - Paper-recommended loss function uses only KL-divergence for TS training

Questions?