Self-Training with Noisy Student: Impact of the ratio of unlabeled to labeled data on accuracy

Pierre Andurand, Tzu Yi Chuang, Kuan Yu Ko

Deep Learning, Columbia University

Introduction

- We will introduce The Self-Training Noisy Student algorithm, and test it on 2 different datasets: CIFAR10 and Street View House Numbers. The underlying algorithms chosen are Resnet20 and a Simple CNN.
- The aim of this report is to study the impact of the ratio between unlabeled data and labeled data.
- We test different levels of noise to the Student, and whether choosing soft labels, hard labels, or hard labels with a probability threshold impact the shape of the accuracy curve.

Self-Training Noisy Student

- Qiezhe Xie, Minh-Thang Luong, Eduard Hovy and Quoc V. Le (2019) Self-training with Noisy Student improves ImageNet Classification
- Authors train EfficientNet model on labeled ImageNet images and use it as Teacher to generate pseudo labels on 300M unlabeled images
- Then train a larger noised EfficientNet as a Student model on the combination of labeled and pseudo labeled images
- Iterate 3 times
- Noise used on Student only with RandAugment data augmentation, dropout and stochastic depth
- Achieve 88.4% top-1 Accuracy on ImageNet, 2.0% better than other state-of-the art model that requires 3.5B weakly labeled Instagram image

STNS Algorithm

- (a) Train Teacher model with labeled data
- (b) Infer pseudo-labels on unlabeled data
- (c) Train equal-or-larger noised Student model on concatenation of true and pseudo labeled data (noise: Data augmentation, dropout, Stochastic depth)
- (d) Use new weights and go back to step (b). Iterate 3 times

Datasets

- Datasets: CIFAR10 and SVHN
- CIFAR-10: 50,000 training, and 10,000 testing color images. Low resolution (32x32). 10 classes: airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, trucks
- SVHN: Street View House Numbers. 73,257 training images. 26,032 testing images. Low-resolution (32x32) colour images. 10 classes (digits)

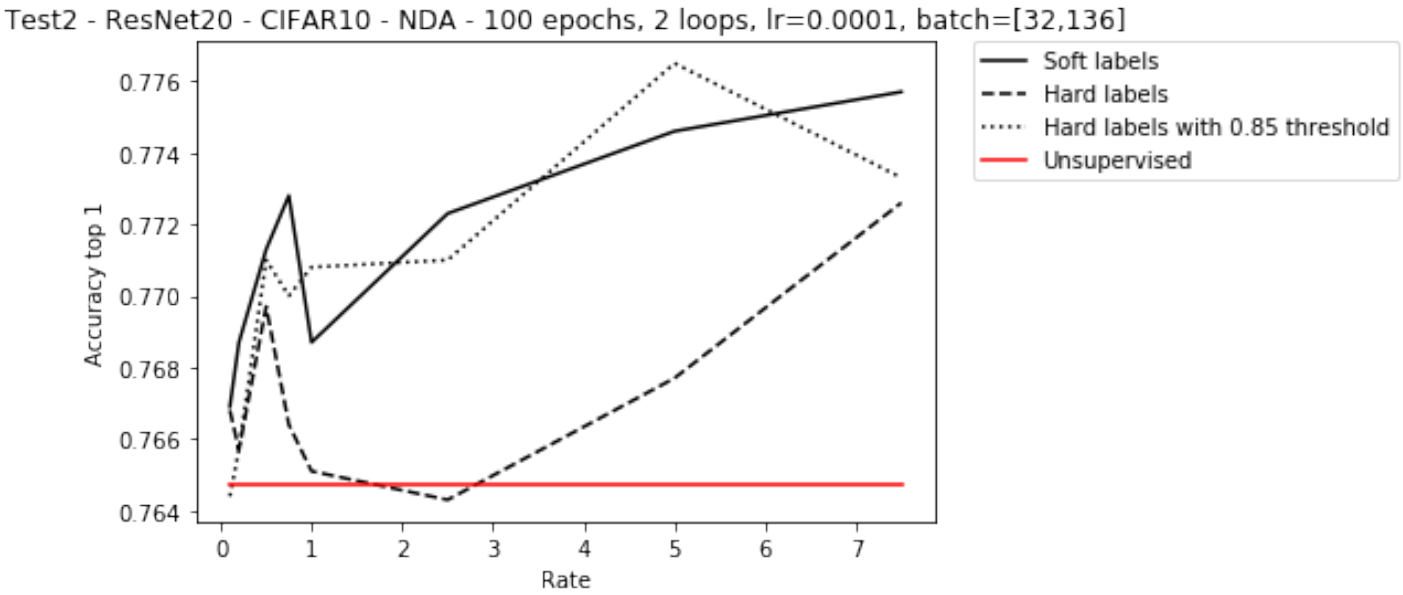
Teacher and Student models used

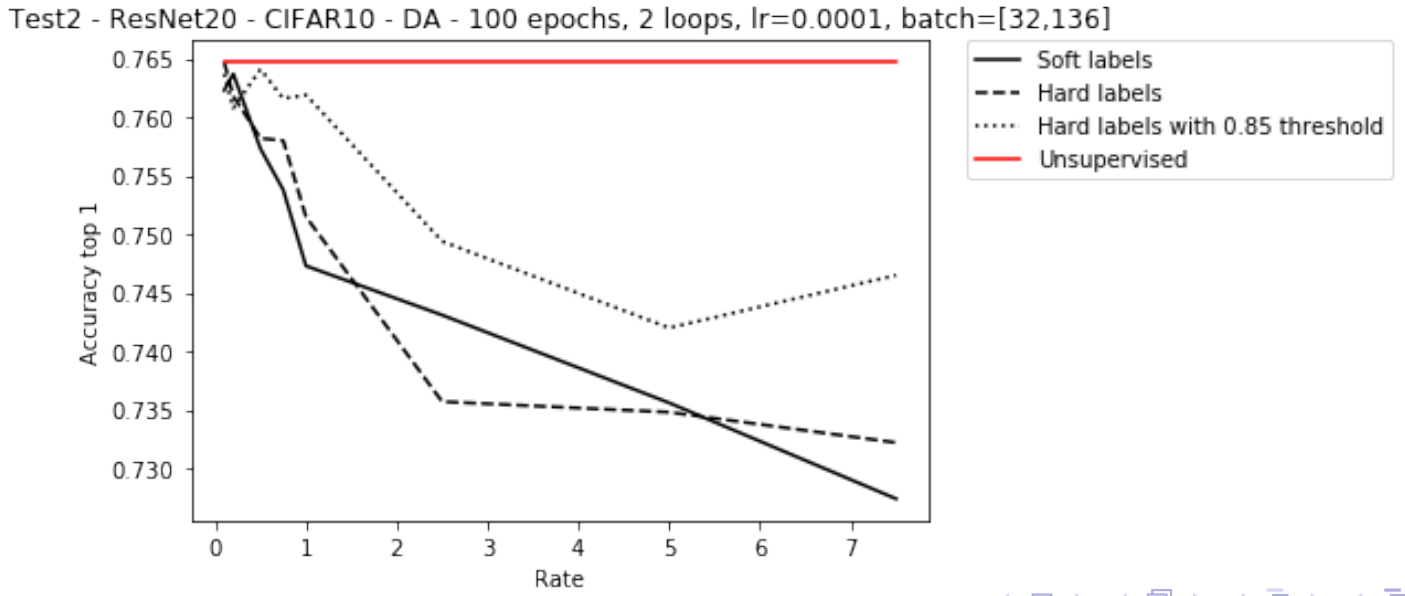
- Teacher: 2 choices: Simple CNN and ResNet20
- Student: Same as Teacher
- Noise to Student: Dropout(0.5) on penultimate layer. And also testing data augmentation on images fed to Student model

Experiment details

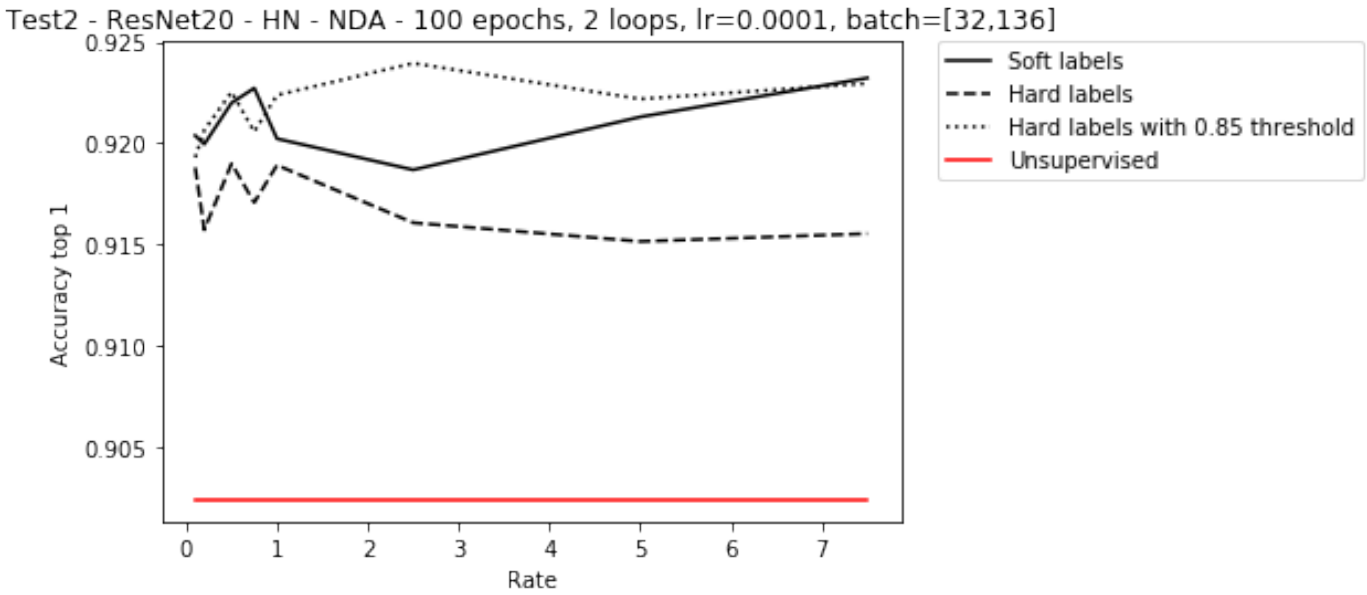
- Sample of 5000 balanced images
- Train on cross-entropy loss
- 100 epochs for step (a) Teacher training. Learning rate of 0.001 for first 80 epochs, and 0.0001 for following 20 epochs
- Learning rate of 0.0001 for Student training
- Mini-batches of 32 for teacher and 136 student training (the paper used 14x larger batches for unlabeled data than labeled data for first 2 iterations, and 28x for the third)
- Unlabeled/Labeled data ratio tested: [0.1, 0.2, 0.5, 0.75, 1, 2.5, 5, 7.5]
- Tests with soft-labels, hard-labels, and hard-labels with probability thresholds of 0.6, 0.75 and 0.85
- 2 iterations of teacher-student

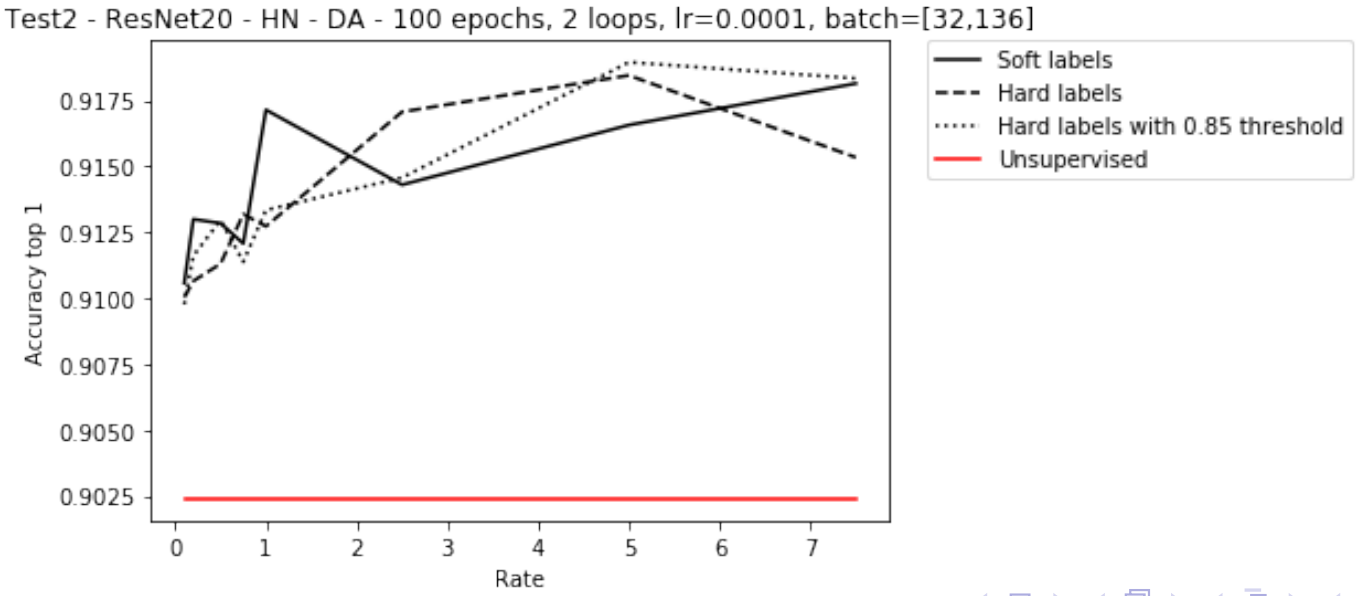
Results: ResNet20 with CIFAR10 - NDA and DA



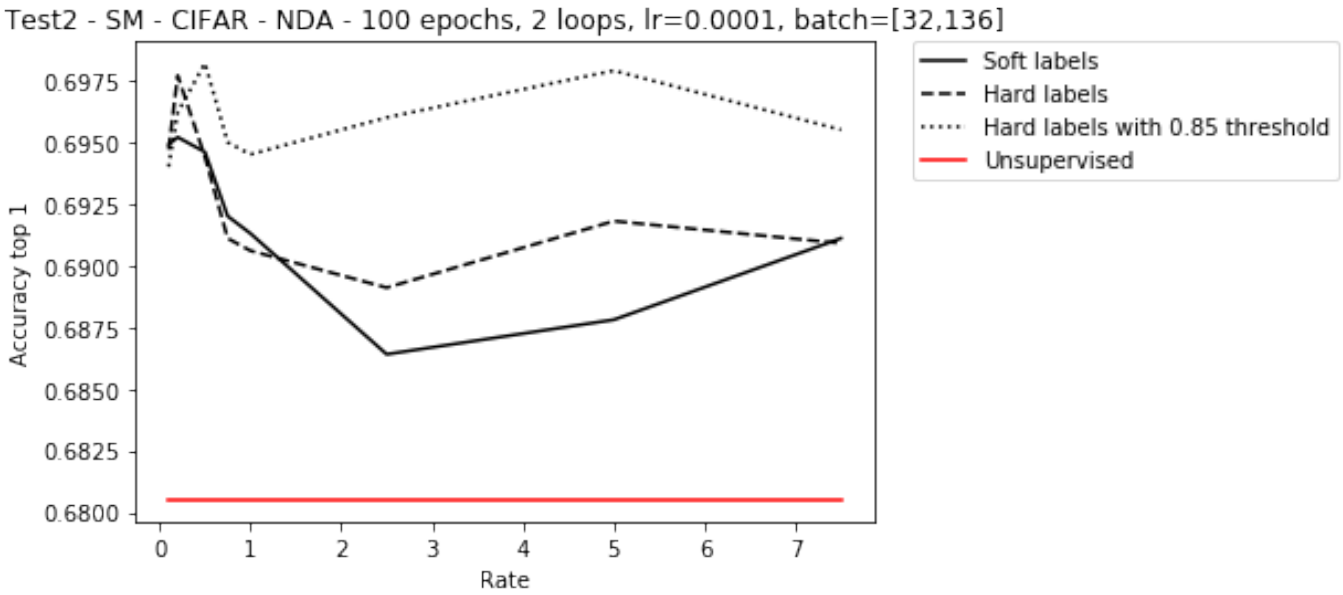


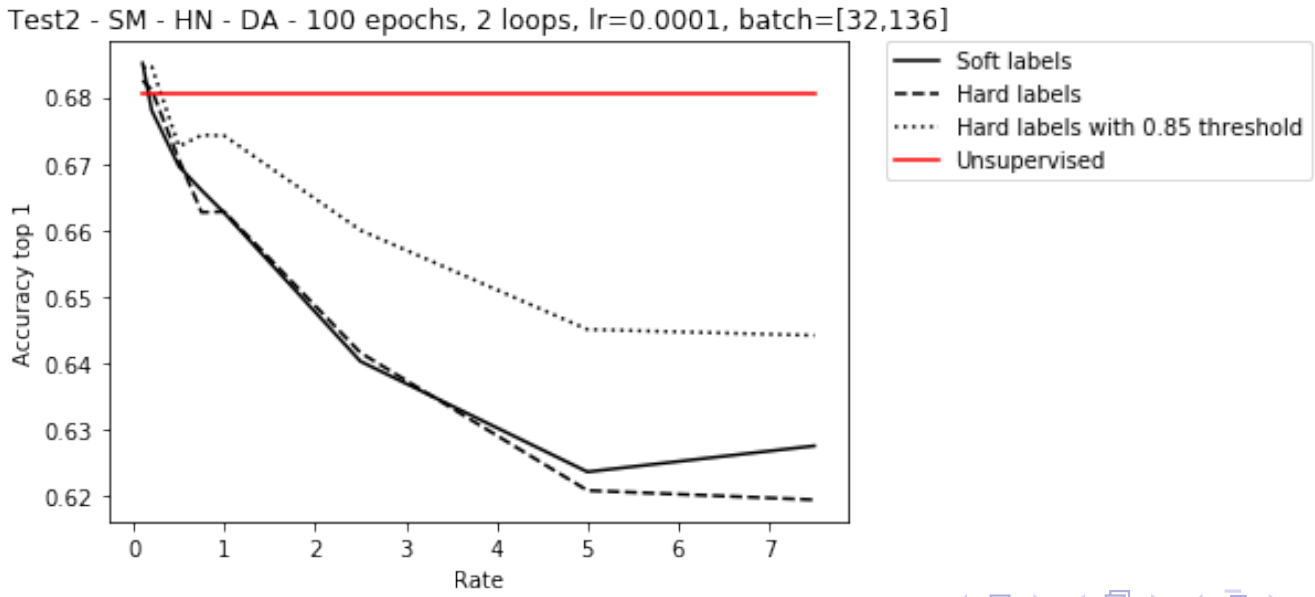
Results: ResNet20 with House Numbers - NDA and DA



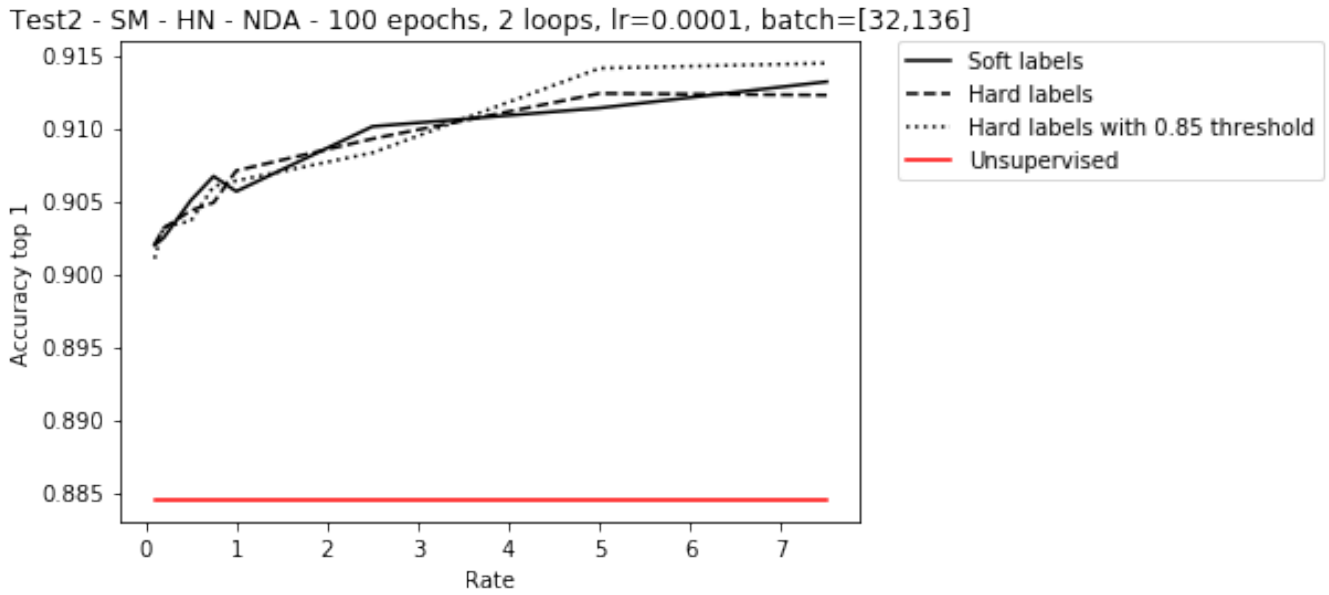


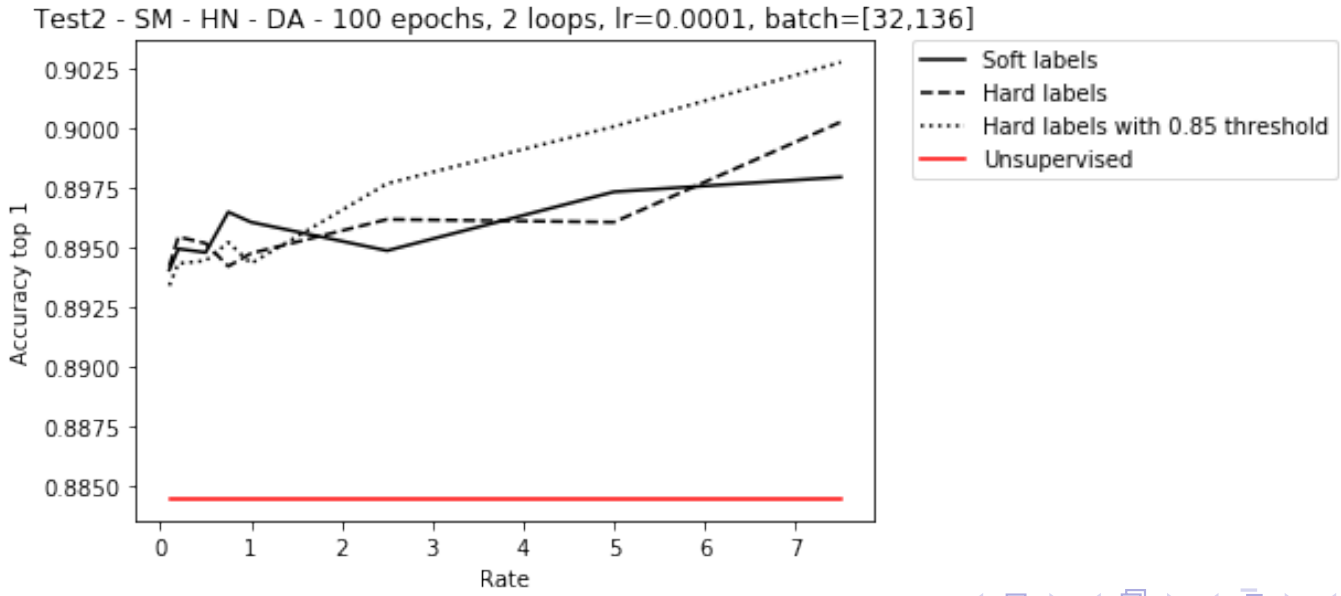
Results: Simple CNN with CIFAR10 - NDA and DA





Results: Simple CNN with House Numbers - NDA and DA





Conclusion

- No clear consistent pattern in shape of curve accuracy=f(rate)
- Shape dependent on dataset, algorithm, and hyperparameters
- Data Augmentation consistently gave worse results
- Better accuracy than with fully supervised model