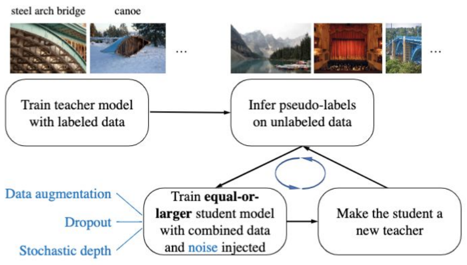
Self-training with Noisy Student improves ImageNet classification - Paper, Code (TF)

By Google Brain & CMU: Nov, 2019

Summary Notes by: Ashvini Jindal, LinkedIn

<table>
<tr><td>Contribution</td><td>● Noisy Student Training - a semi-supervised learning approach that works well even when labeled data is abundant.
● Authors show that noise helps in Student model to generalize better than Teacher model.
● Noisy Student Training boosts robustness in computer vision models.</td></tr>
<tr><td>Idea</td><td>● Paper extends the idea of self-training and distillation with equal-or-large student models and noise added to student during training.
● Idea is to learn more capacity Student model than Teacher as opposed to distillation process (smaller and without noise student model is used in distillation).
● A semi-supervised learning approach leveraging labeled as well as unlabeled data.</td></tr>
<tr><td>Training Process</td><td>1. Train an EfficientNet on labeled data and use this as teacher to generate pseudo labels (soft or one-hot) on large unlabeled data.
2. Train a larger EfficientNet as Student model (from scratch) on combination of labeled and pseudo labeled images with combined cross entropy loss.
3. Make Student as Teacher and go back to first step.

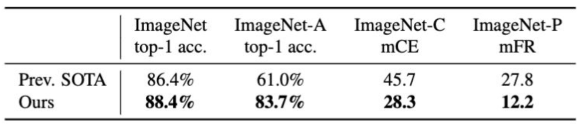
This process is repeated few times. Authors achieves best results with 3 iterations. Resultant model is called <b>EfficientNet-L2</b> (scaled EfficientNet-B7).



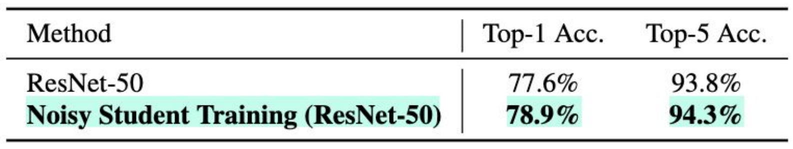
steel arch bridge
canoe
...
...
Train teacher model with labeled data
Infer pseudo-labels on unlabeled data
Data augmentation
Dropout
Stochastic depth
Train equal-or-larger student model with combined data and noise injected
Make the student a new teacher

Figure 1: Illustration of the Noisy Student Training. (All shown images are from ImageNet.)

● Two types of noise: 1) input noise - data augmentation with RandAugment. 2) model noise - dropout and stochastic depth are used.
● Specifically, the teacher produces high-quality pseudo labels by reading in clean images, while the student is required to reproduce those labels with augmented images as input.</td></tr>
<tr><td>Results</td><td>● Achieves 88.4% top-1 accuracy on ImageNet, which is 2.0% better than the SOTA model that requires 3.5B weakly labeled Instagram images. Proposed approach uses only 300M unlabeled images.
● Improves classification robustness on much harder test sets by large margins: ImageNet-A top-1 accuracy from 61.0% to 83.7%, ImageNet-C mean corruption error (mCE) from 45.7 to 28.3 and ImageNet-P mean flip rate (mFR) from 27.8 to 12.2.
●

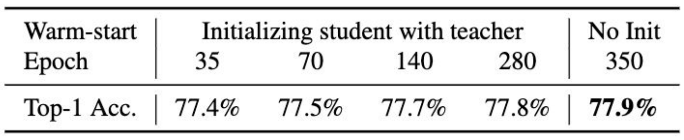

<table>
<tr><th></th><th>ImageNet top-1 acc.</th><th>ImageNet-A top-1 acc.</th><th>ImageNet-C mCE</th><th>ImageNet-P mFR</th></tr>
<tr><td>Prev. SOTA</td><td>86.4%</td><td>61.0%</td><td>45.7</td><td>27.8</td></tr>
<tr><td>Ours</td><td><b>88.4%</b></td><td><b>83.7%</b></td><td><b>28.3</b></td><td><b>12.2</b></td></tr>
</table>
Table 1: Summary of key results compared to previous state-of-the-art models [86, 55]. Lower is better for mean corruption error (mCE) and mean flip rate (mFR).</td></tr>
<tr><td>Insights</td><td>1. Among noisy student training and larger model, former contributes to majority of performance gains.
2. Vision model can generally benefit from Noisy student training even without iterative training. (Ref: Table-16)

Better Performance with:
1. Using soft probabilities than one-hot labels as pseudo labels works better. Soft pseudo labels are particularly helpful for out-of-domain unlabeled data.
2. A large amount of unlabeled data.
3. Joint training on labeled data and unlabeled data rather than two-stage process of pretraining with unlabeled data and then finetunes on labeled data (Ref: Table-13)
4. Training Student from scratch performs better than initializing with teacher (Ref: Table-15)
5. large Teacher model with better performance
6. Class balancing is crucial from better performance during Student training. Techniques: Undersampling & Duplication</td></tr>
<tr><td>Few other details</td><td>● ImageNet data is used as labeled data and JFT 300M images data is used as unlabeled set.
● Data Filtering: only images with high confidence (>0.3 prob) are used during Student model training.
● Training time of EfficientNet-L2 is five times of Efficient-B7.</td></tr>
<tr><td>Utility</td><td>● Pre-trained checkpoints can be used a feature extractor for various classification and other vision tasks (code)</td></tr>
</table>



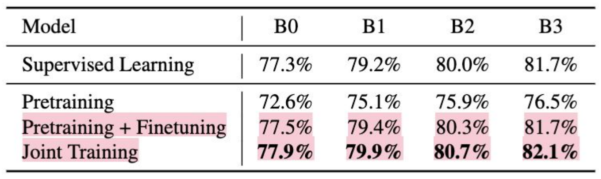
Method	Top-1 Acc.	Top-5 Acc.
ResNet-50	77.6%	93.8%
Noisy Student Training (ResNet-50)	**78.9%**	**94.3%**

Table 16: Experiments on ResNet-50.



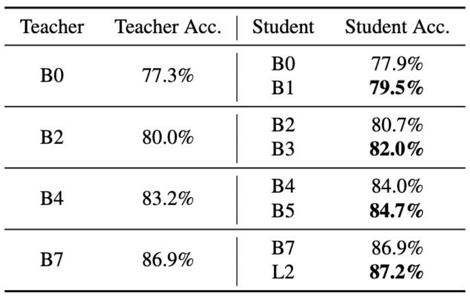
Warm-start	Initializing student with teacher				No Init
Epoch	35	70	140	280	350
Top-1 Acc.	77.4%	77.5%	77.7%	77.8%	**77.9%**

Table 15: A student initialized with the teacher still requires at least 140 epochs to perform well. The baseline model, trained with labeled data only, has an accuracy of 77.3%.



Model	B0	B1	B2	B3
Supervised Learning	77.3%	79.2%	80.0%	81.7%
Pretraining	72.6%	75.1%	75.9%	76.5%
Pretraining + Finetuning	77.5%	79.4%	80.3%	81.7%
Joint Training	**77.9%**	**79.9%**	**80.7%**	**82.1%**

Table 13: Joint training works better than pretraining and finetuning. We vary the finetuning steps and report the best results. Models are trained for 350 epochs instead of 700 epochs without iterative training.



Teacher	Teacher Acc.	Student	Student Acc.
B0	77.3%	B0	77.9%
		B1	**79.5%**
B2	80.0%	B2	80.7%
		B3	**82.0%**
B4	83.2%	B4	84.0%
		B5	**84.7%**
B7	86.9%	B7	86.9%
		L2	**87.2%**