Self-training with Noisy Student improves ImageNet classification

Денисенко Наталья, БПМИ181

Мотивация

	ImageNet top-1 acc.	ImageNet-A top-1 acc.	ImageNet-C mCE	ImageNet-P mFR
Prev. SOTA	86.4%	61.0%	45.7	27.8
Ours	**88.4%**	**83.7%**	**28.3**	**12.2**

Мотивация

18	FixResNeXt-101 32x48d	86.4%	98.0%	829M	✓	Fixing the train-test resolution discrepancy
19	NoisyStudent (EfficientNet-B6)	86.4%	97.9%	43M	✓	Self-training with Noisy Student improves ImageNet classification
20	FixEfficientNet-B5	86.4%	97.9%	30M	✓	Fixing the train-test resolution discrepancy: FixEfficientNet
21	Swin-L (384 res, ImageNet-22k pretrain)	86.4%			✓	Swin Transformer: Hierarchical Vision Transformer using Shifted Windows

Мотивация

8	**NoisyStudent** (EfficientNet-L2)	88.4%	98.7%	480M	✓	Self-training with Noisy Student improves ImageNet classification
13	**NoisyStudent** (EfficientNet-B7)	86.9%	98.1%	66M	✓	Self-training with Noisy Student improves ImageNet classification

Мотивация

1	**Meta Pseudo Labels** (EfficientNet-L2)	90.2%	98.8%	480M	✓	Meta Pseudo Labels
2	**Meta Pseudo Labels** (EfficientNet-B6-Wide)	90%	98.7%	390M	✓	Meta Pseudo Labels

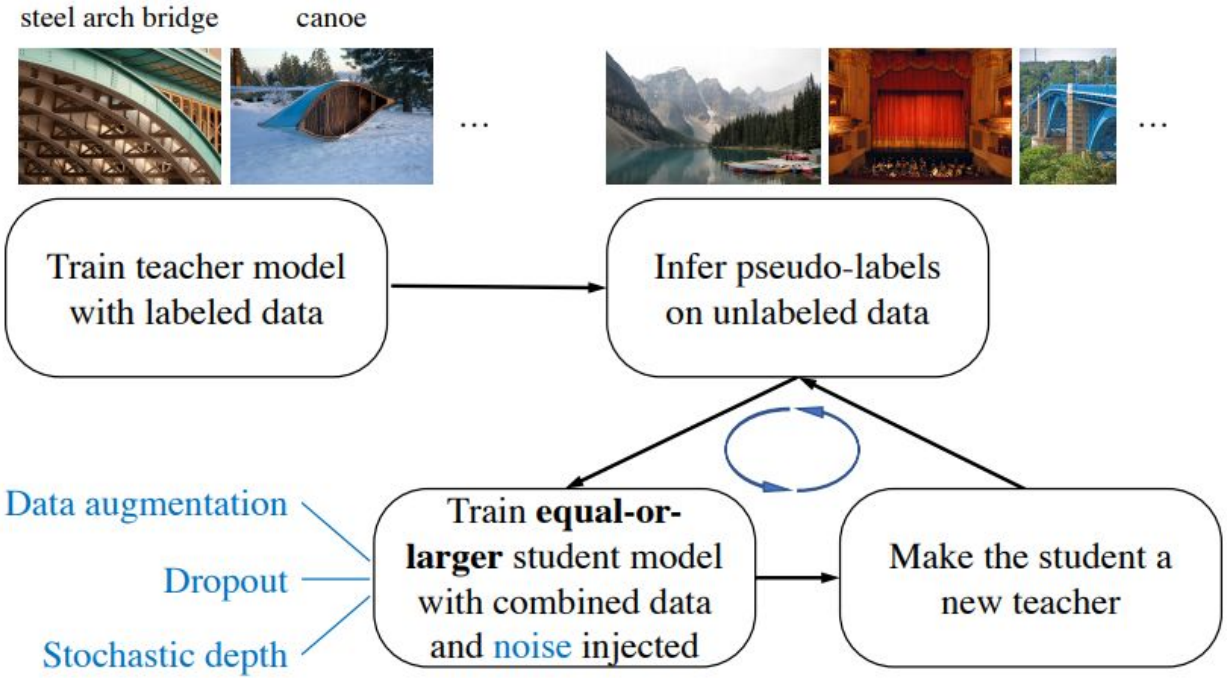
Архитектура

- Обучить модель учителя на размеченных данных
- Сгенерировать разметку на неразмеченных данных
- Обучить модель ученика на размеченных и ранее неразмеченных данных

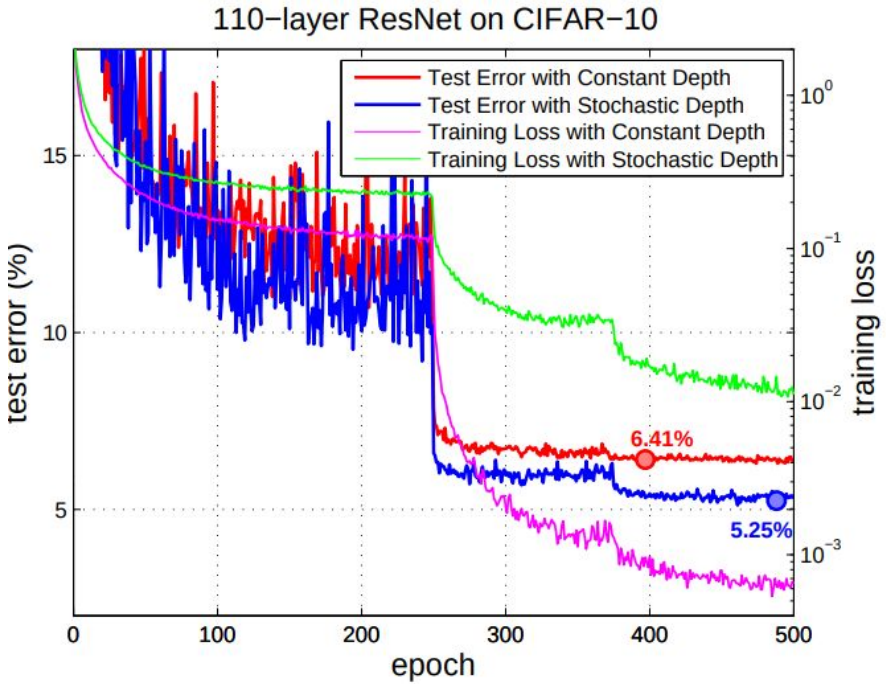
Архитектура

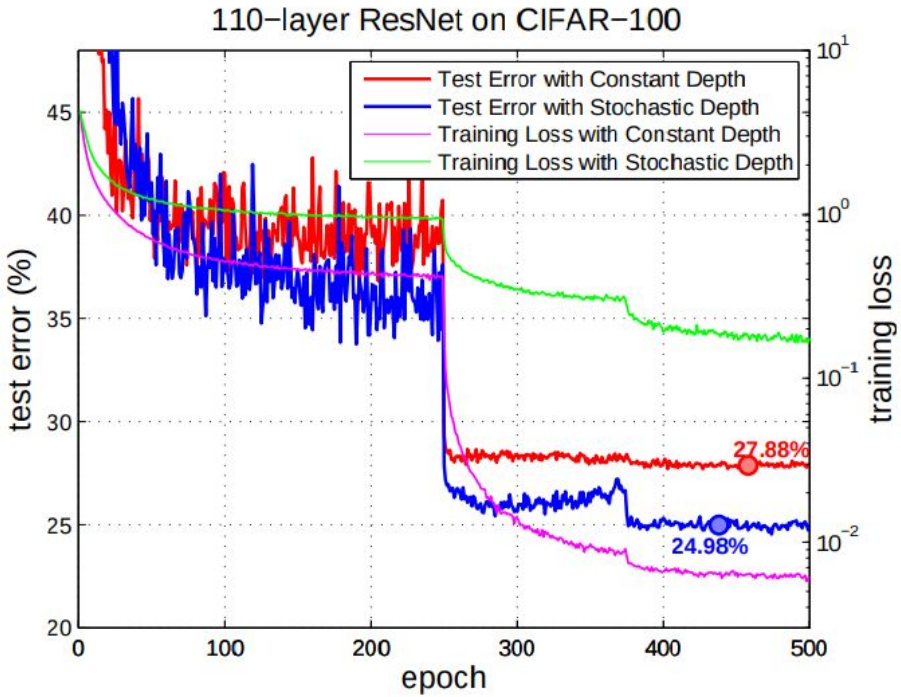
- Модель учеников будет увеличиваться
- Добавляем шум

Обучение

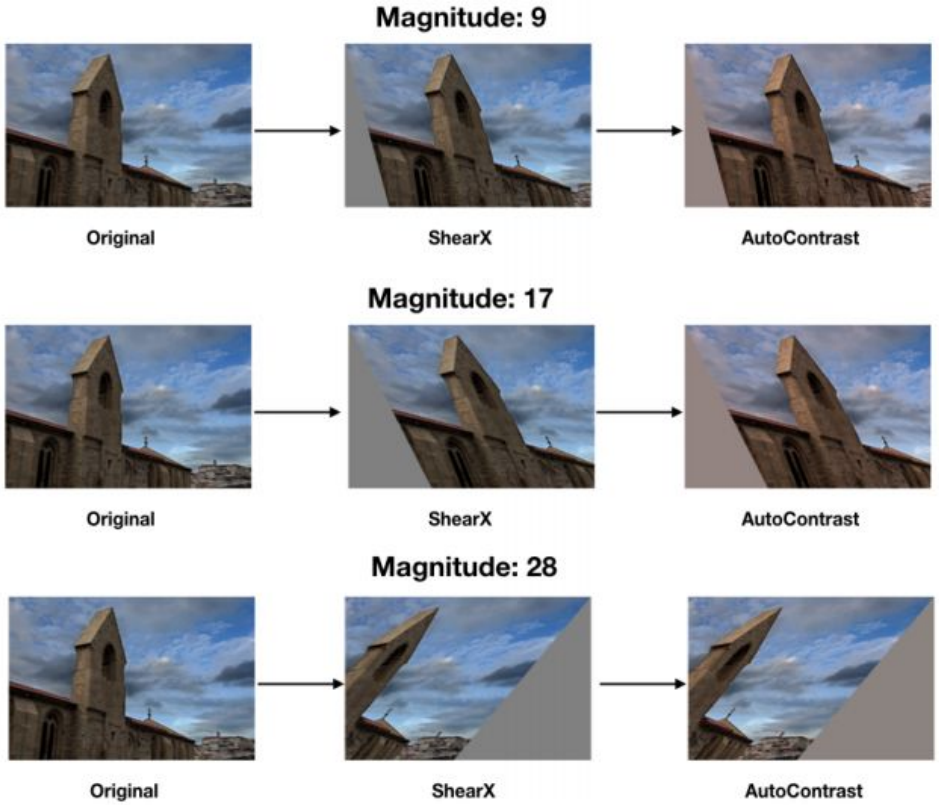

steel arch bridge
canoe
...
...
Train teacher model with labeled data
Infer pseudo-labels on unlabeled data
Data augmentation
Dropout
Stochastic depth
Train **equal-or-larger** student model with combined data and noise injected
Make the student a new teacher

Добавление шума

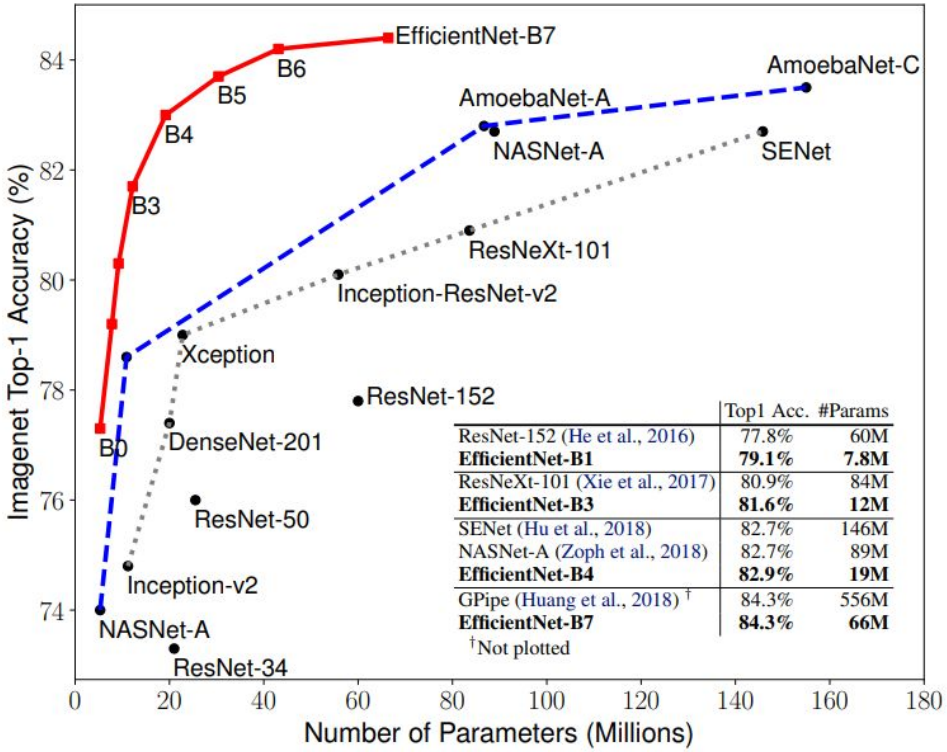




Добавление шума



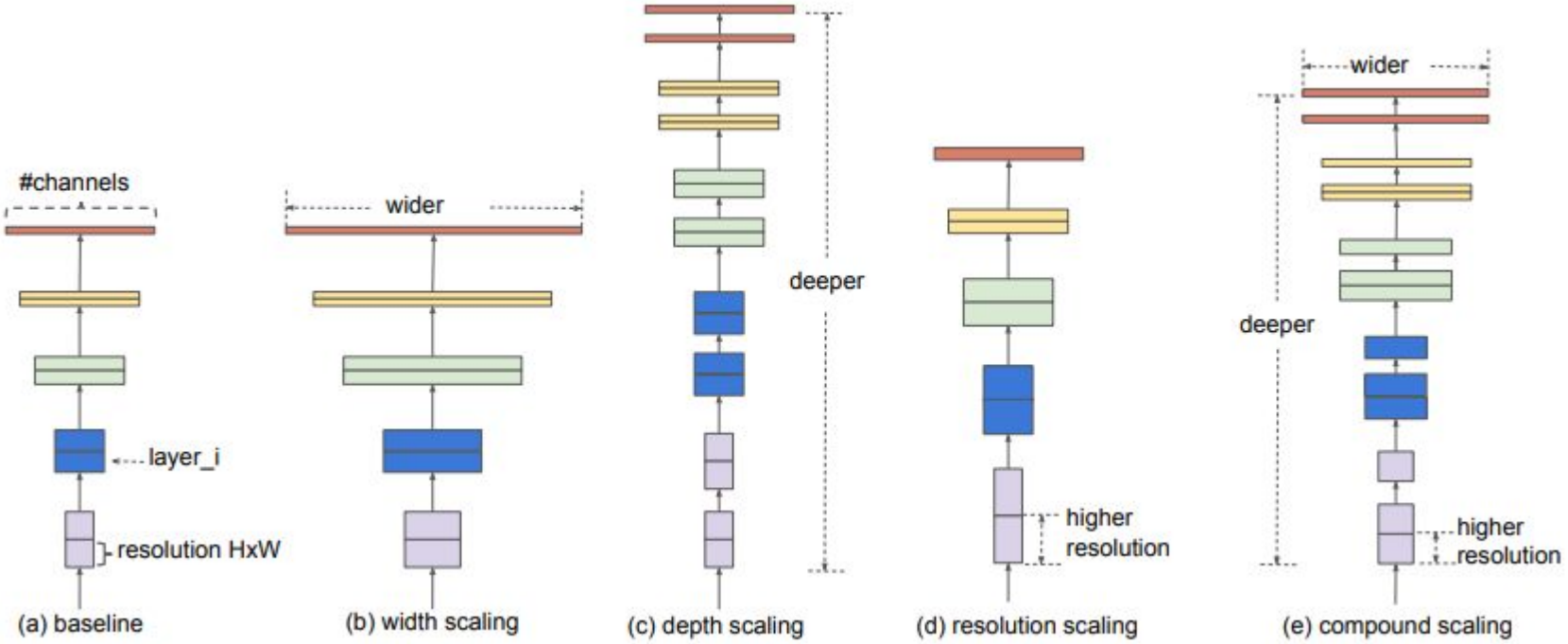
Что же мы улучшаем



	Top1 Acc.	#Params
ResNet-152 (He et al., 2016)	77.8%	60M
EfficientNet-B1	**79.1%**	**7.8M**
ResNeXt-101 (Xie et al., 2017)	80.9%	84M
EfficientNet-B3	**81.6%**	**12M**
SENet (Hu et al., 2018)	82.7%	146M
NASNet-A (Zoph et al., 2018)	82.7%	89M
EfficientNet-B4	**82.9%**	**19M**
GPipe (Huang et al., 2018) †	84.3%	556M
EfficientNet-B7	**84.3%**	**66M**

†Not plotted

Что же мы улучшаем



Немного формул

$$\mathcal{N} = \bigodot_{i=1\ldots s} \mathcal{F}_i^{L_i}\left(X_{\langle H_i, W_i, C_i\rangle}\right)$$

$\mathcal{F}_i(X_i)$ - оператор который определяет один слой

$\mathcal{F}_i^{L_i}$ - повторяем данный слой L_i раз на этапе i

$\langle H_i, W_i, C_i\rangle$ размеры входного тензора

Немного формул

$$Accuracy(\mathcal{N}(d, w, r))$$

$$\mathcal{N}(d, w, r) = \bigodot_{i=1\ldots s} \hat{\mathcal{F}}_i^{d\cdot\hat{L}_i}\left(X_{\langle r\cdot\hat{H}_i, r\cdot\hat{W}_i, w\cdot\hat{C}_i\rangle}\right)$$

Немного формул

$$d = \alpha^{\phi}$$

$$w = \beta^{\phi}$$

$$r = \gamma^{\phi}$$

$$\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$$

$$\alpha \geq 1, \beta \geq 1, \gamma \geq 1$$

Немного формул

$$ACC(m) \times [FLOPS(m)/T]^{w}$$

Вернёмся к шумным студентам


sea lion lighthouse **submarine** canoe


snow leopard electric ray **swing** mosquito net


plate rack refrigerator **racing car** car wheel


dragonfly bullfrog **starfish** wreck


toaster pill bottle **gown** ski


plate rack medicine chest **racing car** fire engine


hummingbird bald eagle **basketball** parking meter


parking meter vacuum **cannon** television


plate rack medicine chest **racing car** car wheel

Заключение

1. Модель учителя необходимо делать сложнее
2. Для лучших результатов надо брать большое число неразмеченных изображений
3. Мягкая псевдо разметка работает лучше для данных вне классов
4. Модель ученика должна быть сложнее чем модель учителя
5. Для маленьких моделей важно не забывать про балансировку элементов в классах
6. Необходимо использовать joint-training на размеченной и неразмеченной выборках
7. Должна быть значимая разница между размерами неразмеченной и размеченной подвыборок
8. Обучать учеников с нуля порой выигрывает перед инициализацией через учителя

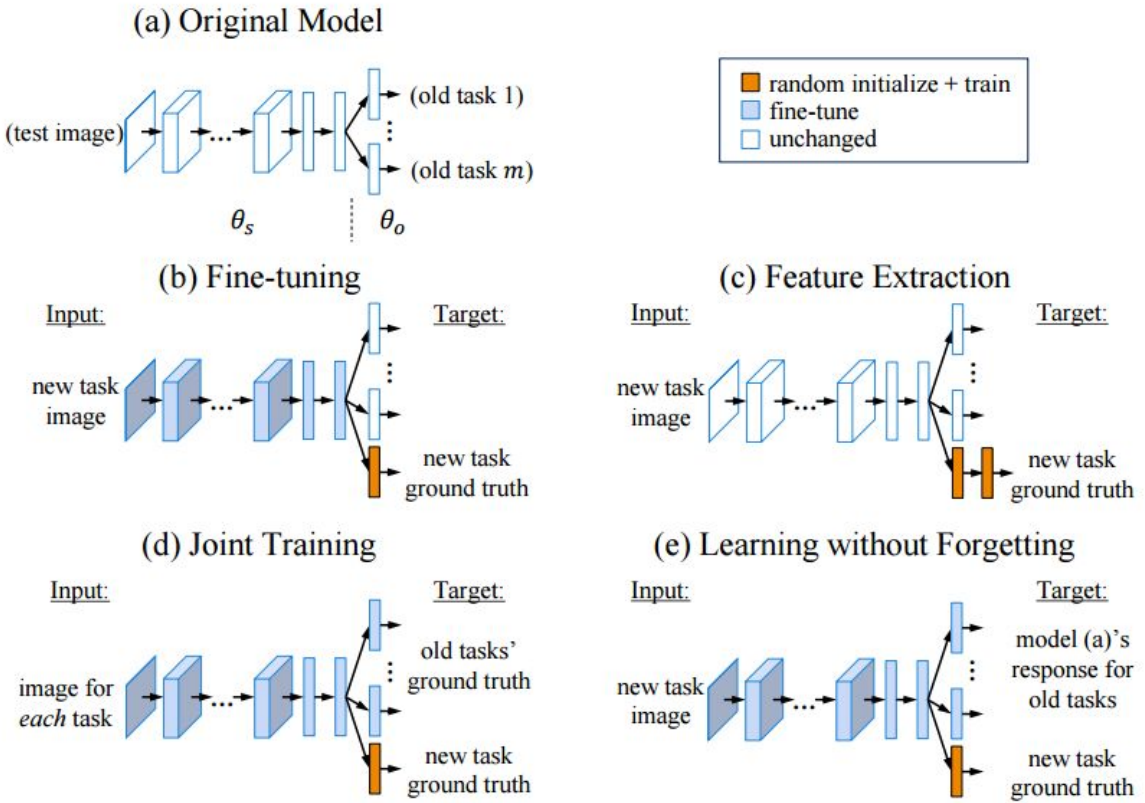
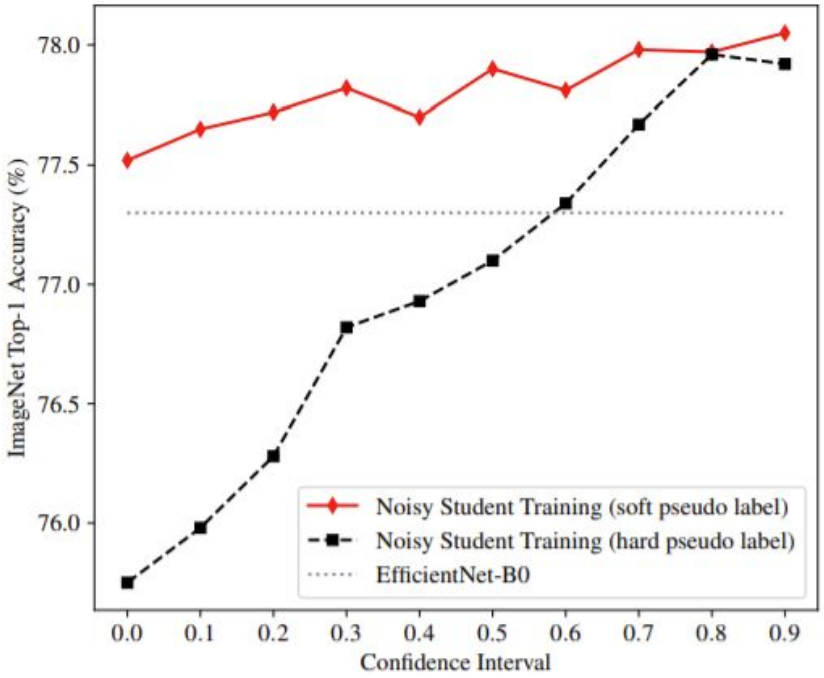


Fig. 2. Illustration for our method (e) and methods we compare to (b-d). Images and labels used in training are shown. Data for different tasks are used in alternation in joint training.

Model	# Params	Top-1 Acc.	Top-5 Acc.
EfficientNet-B0	5.3M	77.3%	93.4%
Noisy Student Training (B0)		78.1%	94.2%
Noisy Student Training (B0, L2)		**78.8%**	**94.5%**
EfficientNet-B1	7.8M	79.2%	94.4%
Noisy Student Training (B1)		80.2%	95.2%
Noisy Student Training (B1, L2)		**81.5%**	**95.8%**
EfficientNet-B2	9.2M	80.0%	94.9%
Noisy Student Training (B2)		81.1%	95.5%
Noisy Student Training (B2, L2)		**82.4%**	**96.3%**
EfficientNet-B3	12M	81.7%	95.7%
Noisy Student Training (B3)		82.5%	96.4%
Noisy Student Training (B3, L2)		**84.1%**	**96.9%**
EfficientNet-B4	19M	83.2%	96.4%
Noisy Student Training (B4)		84.4%	97.0%
Noisy Student Training (B4, L2)		**85.3%**	**97.5%**
EfficientNet-B5	30M	84.0%	96.8%
Noisy Student Training (B5)		85.1%	97.3%
Noisy Student Training (B5, L2)		**86.1%**	**97.8%**
EfficientNet-B6	43M	84.5%	97.0%
Noisy Student Training (B6)		85.9%	97.6%
Noisy Student Training (B6, L2)		**86.4%**	**97.9%**
EfficientNet-B7	66M	85.0%	97.2%
Noisy Student Training (B7)		86.4%	97.9%
Noisy Student Training (B7, L2)		**86.9%**	**98.1%**

Data	1/128	1/64	1/32	1/16	1/4	1
Top-1 Acc.	83.4%	83.3%	83.7%	83.9%	83.8%	**84.0%**



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%**

Model	B0	B1	B2	B3
Supervised Learning	77.3%	79.2%	80.0%	81.7%
Noisy Student Training	**77.9%**	**79.9%**	**80.7%**	82.1%
w/o Data Balancing	77.6%	79.6%	80.6%	82.1%

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%**

Teacher (Acc.)	Batch Size Ratio	Top-1 Acc.
B4 (83.2)	1:1	84.0%
	3:1	84.0%
L2 (87.0)	1:1	86.7%
	3:1	**87.4%**
L2 (87.4)	3:1	87.4%
	6:1	**87.9%**

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%**

- https://arxiv.org/abs/1603.09382
- https://arxiv.org/abs/1911.04252
- https://arxiv.org/pdf/1905.11946
- https://arxiv.org/abs/1807.11626