Self-training with Noisy Student improves ImageNet classification

Болотин Арсений БПМИ182

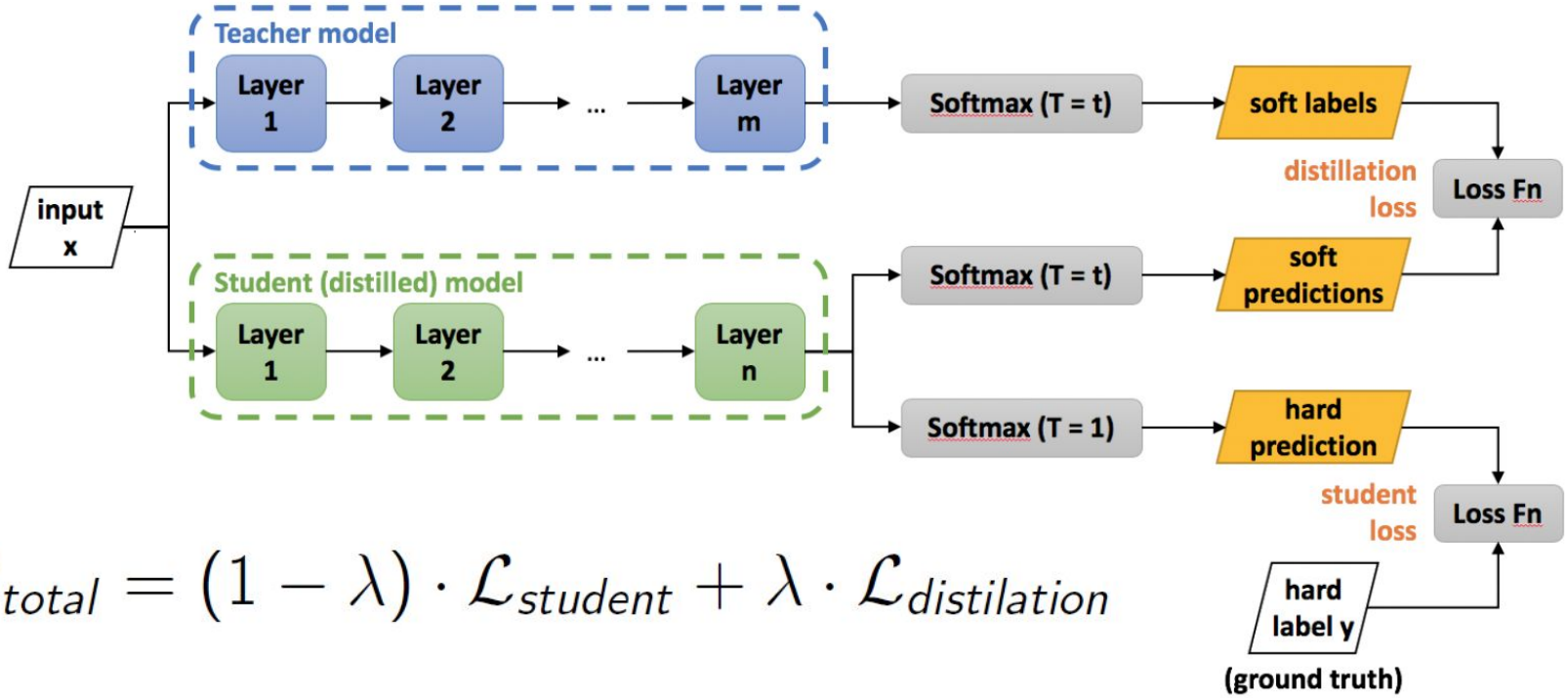
Мотивация

- Существует огромное количество неразмеченных данных - хотим использовать их как-то для обучения

- Большинство state-of-the-art подходов в различных задачах используют дополнительные данные для обучения

Дистилляция

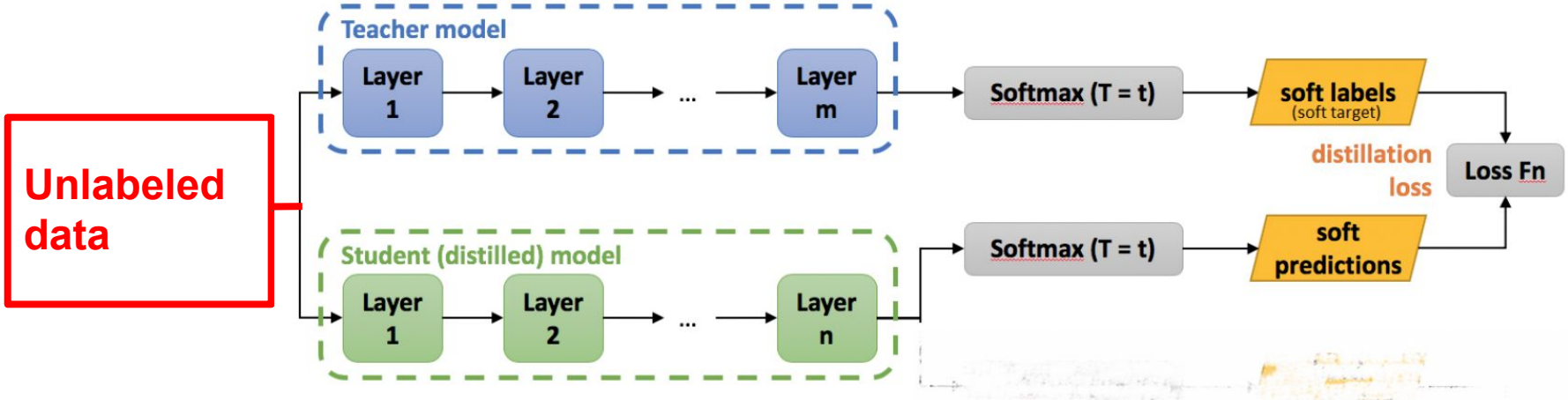
$$q_i = \frac{\exp(z_i / T)}{\sum_j \exp(z_j / T)}$$



$$\mathcal{L}_{total} = (1 - \lambda) \cdot \mathcal{L}_{student} + \lambda \cdot \mathcal{L}_{distilation}$$

Дистилляция

$$q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$



Noisy Student Training

Размеченные данные: $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$

Неразмеченные данные: $\{\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_m\}$

1. Обучаем учителя на размеченных данных $\frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f^{noised}(x_i, \theta^t))$
2. Предсказываем метки на неразмеченных данных учителем (**шум не добавляется**)

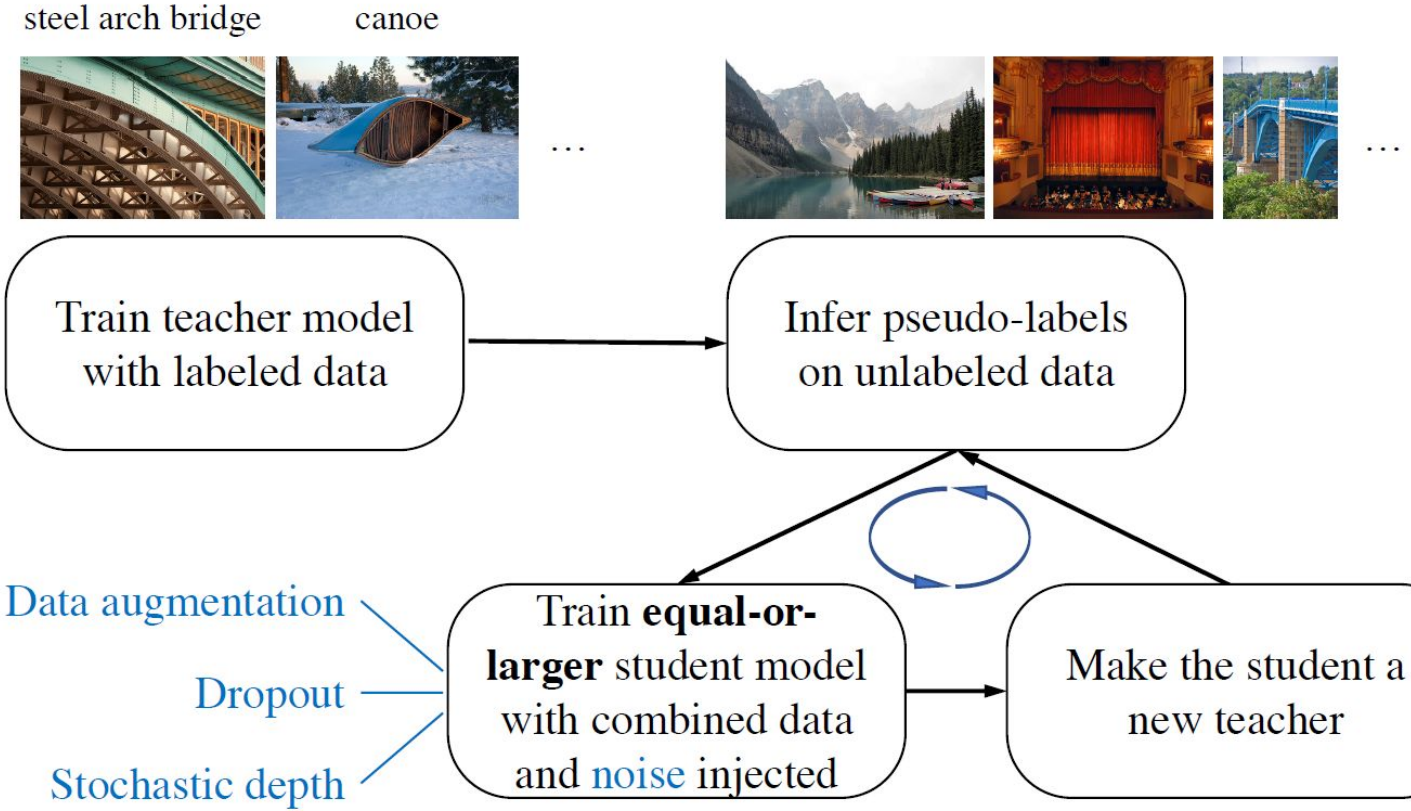
$$\tilde{y}_i = f(\tilde{x}_i, \theta_*^t), \forall i = 1 \ldots m$$

3. Обучаем студента на размеченных и неразмеченных данных

$$\frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f^{noised}(x_i, \theta^s)) + \frac{1}{m}\sum_{i=1}^{m} \ell(\tilde{y}_i, f^{noised}(\tilde{x}_i, \theta^s))$$

4. Используем студента, как учителя и возвращаемся на второй шаг

Noisy Student Training

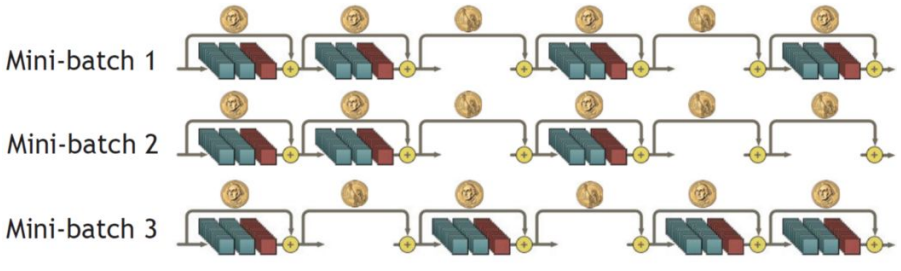


Noisy Student Training - детали

- Студент не меньше учителя
- Учитель предсказывает на 2 шаге soft labels (непрерывное распределение) или hard labels (one-hot вектор)
- Неразмеченные данные фильтруются по уверенности учителя
- Балансировка классов - у каждого класса одинаковое число объектов, достигается дублированием случайных объектов до нужного количества
- Каждый батч формируется из неразмеченных и размеченных данных

Stochastic depth

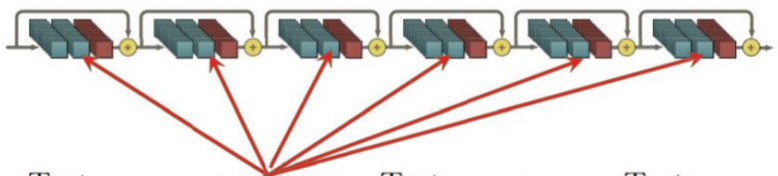
Train



$$H_\ell = \texttt{ReLU}(b_\ell f_\ell(H_{\ell-1}) + \mathrm{id}(H_{\ell-1}))$$

Bernoulli random variable

Inference



$$H_\ell^{\mathrm{Test}} = \texttt{ReLU}(p_\ell f_\ell(H_{\ell-1}^{\mathrm{Test}}; W_\ell) + H_{\ell-1}^{\mathrm{Test}})$$

All layers are on, but outputs of f_ℓ are down weighted by their corresponding survival probabilities.

Linear Decay Rule

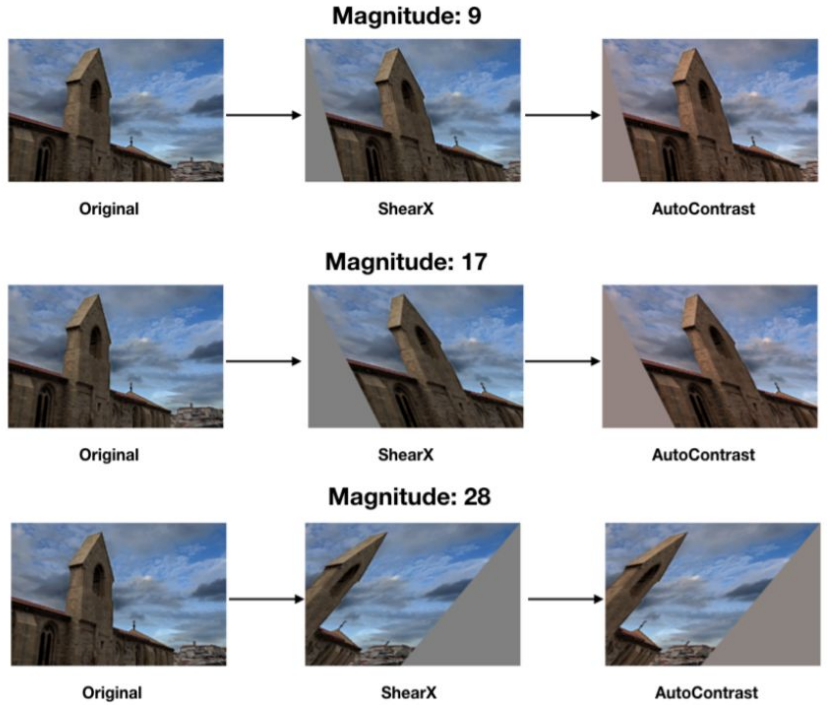
$$b_\ell \sim \texttt{Bernoulli}(p_\ell) \quad \text{with} \quad p_\ell = (1 - \frac{\ell}{L}) \times 1 + \frac{\ell}{L} \times p_L$$

RandAugment

- **N** - количество случайных преобразований

```
• identity      • autoContrast  • equalize
• rotate        • solarize      • color
• posterize     • contrast      • brightness
• sharpness     • shear-x       • shear-y
• translate-x   • translate-y
```

- **M** - целое число от 0 до 30, обозначающее величину (magnitude) преобразования
- Небольшое пространство гиперпараметров

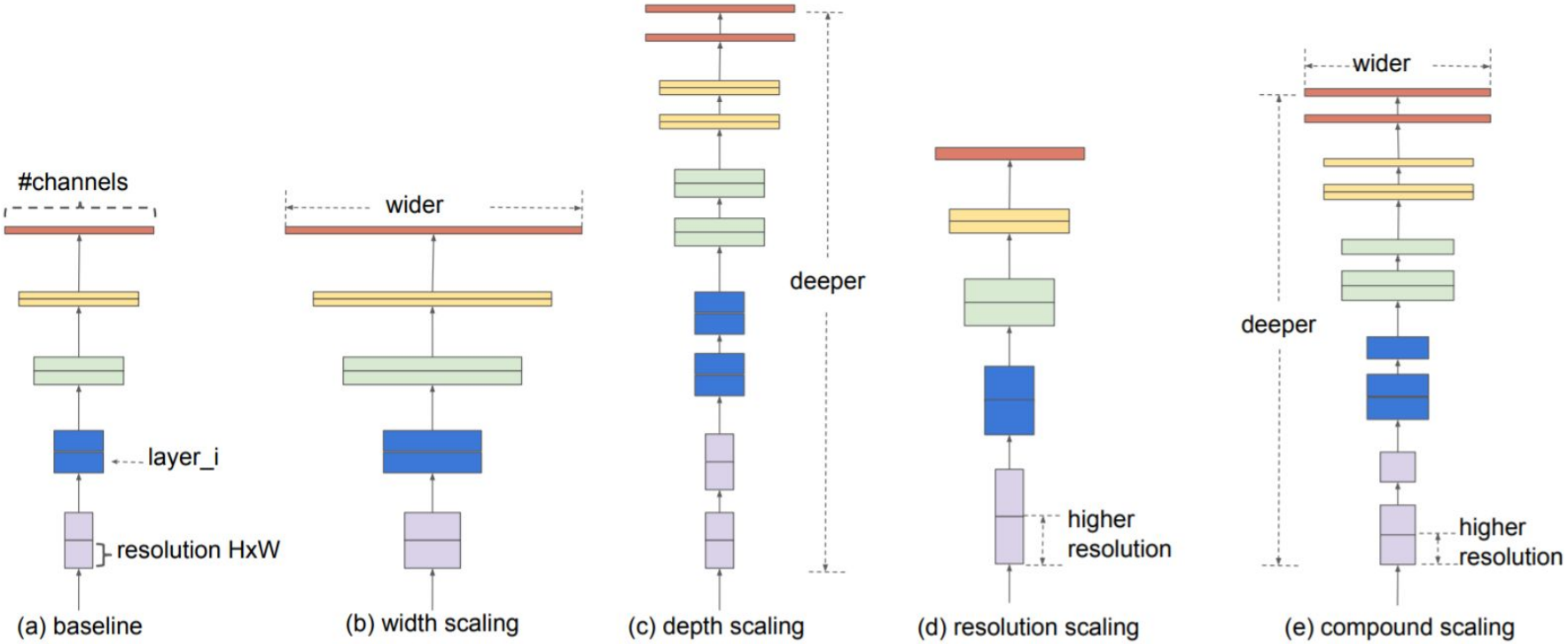


Зачем нужен шум?

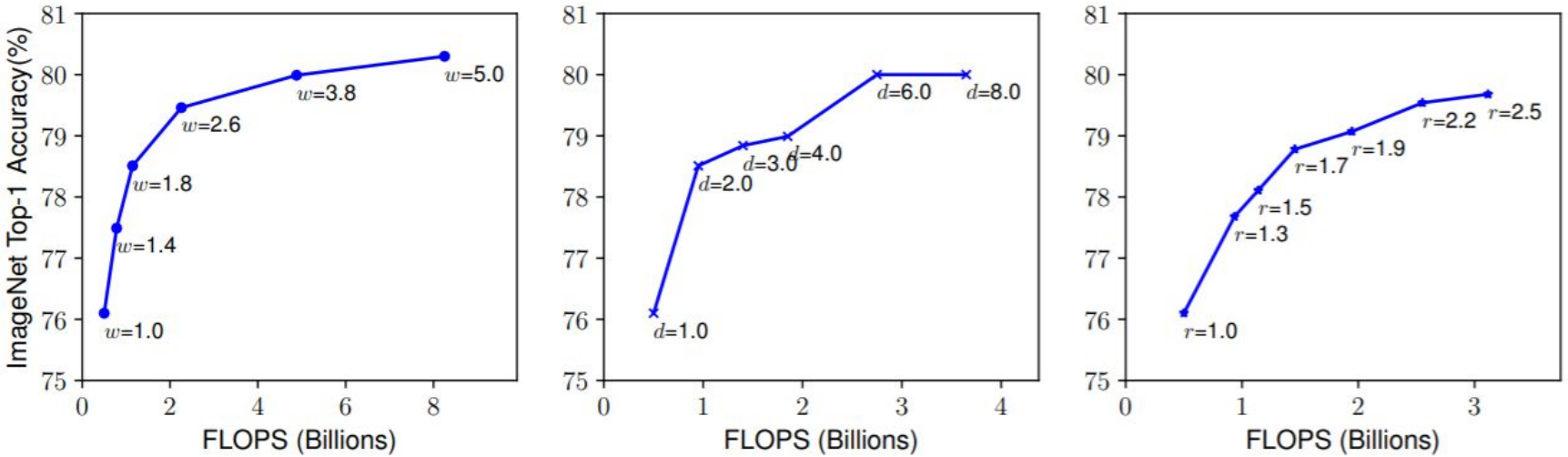
Студент должен предсказывать такие же метки на зашумленных данных, что и учитель на не зашумленных.

Model / Unlabeled Set Size	1.3M	130M
EfficientNet-B5	83.3%	84.0%
Noisy Student Training (B5)	**83.9%**	**85.1%**
student w/o Aug	83.6%	84.6%
student w/o Aug, SD, Dropout	83.2%	84.3%
teacher w. Aug, SD, Dropout	83.7%	84.4%

EfficientNets

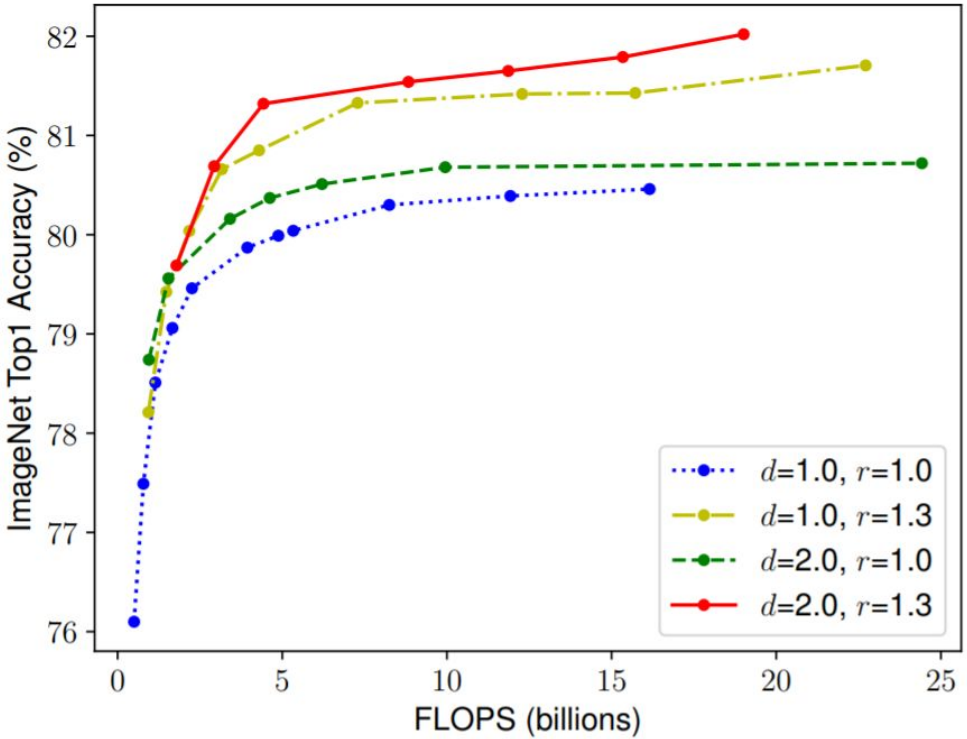


EfficientNets



Здесь и далее w, d, r во сколько раз увеличили ширину, глубину, разрешение соответственно по сравнению с baseline

EfficientNets



Compound Model Scaling

Depth: $d = \alpha^\phi$

Width: $w = \beta^\phi$

Resolution: $r = \gamma^\phi$

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

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

Если мы хотим использовать в 2^ϕраза больше ресурсов, то увеличим глубину в d раз, ширину в w раз и разрешение в r раз

Оптимальные константы ищутся небольшим grid search при $\phi = 1$ (эквивалентно увеличению ресурсов в 2 раза)

EfficientNet-B0 - baseline

Stage i	Operator $\hat{\mathcal{F}}_i$	Resolution $\hat{H}_i \times \hat{W}_i$	#Channels $\hat{C}_i$	#Layers $\hat{L}_i$
1	Conv3x3	224×224	32	1
2	MBConv1, k3x3	112×112	16	1
3	MBConv6, k3x3	112×112	24	2
4	MBConv6, k5x5	56×56	40	2
5	MBConv6, k3x3	28×28	80	3
6	MBConv6, k5x5	14×14	112	3
7	MBConv6, k5x5	14×14	192	4
8	MBConv6, k3x3	7×7	320	1
9	Conv1x1 & Pooling & FC	7×7	1280	1

Константы для Compound Model Scaling: $\alpha = 1.2, \beta = 1.1, \gamma = 1.15$

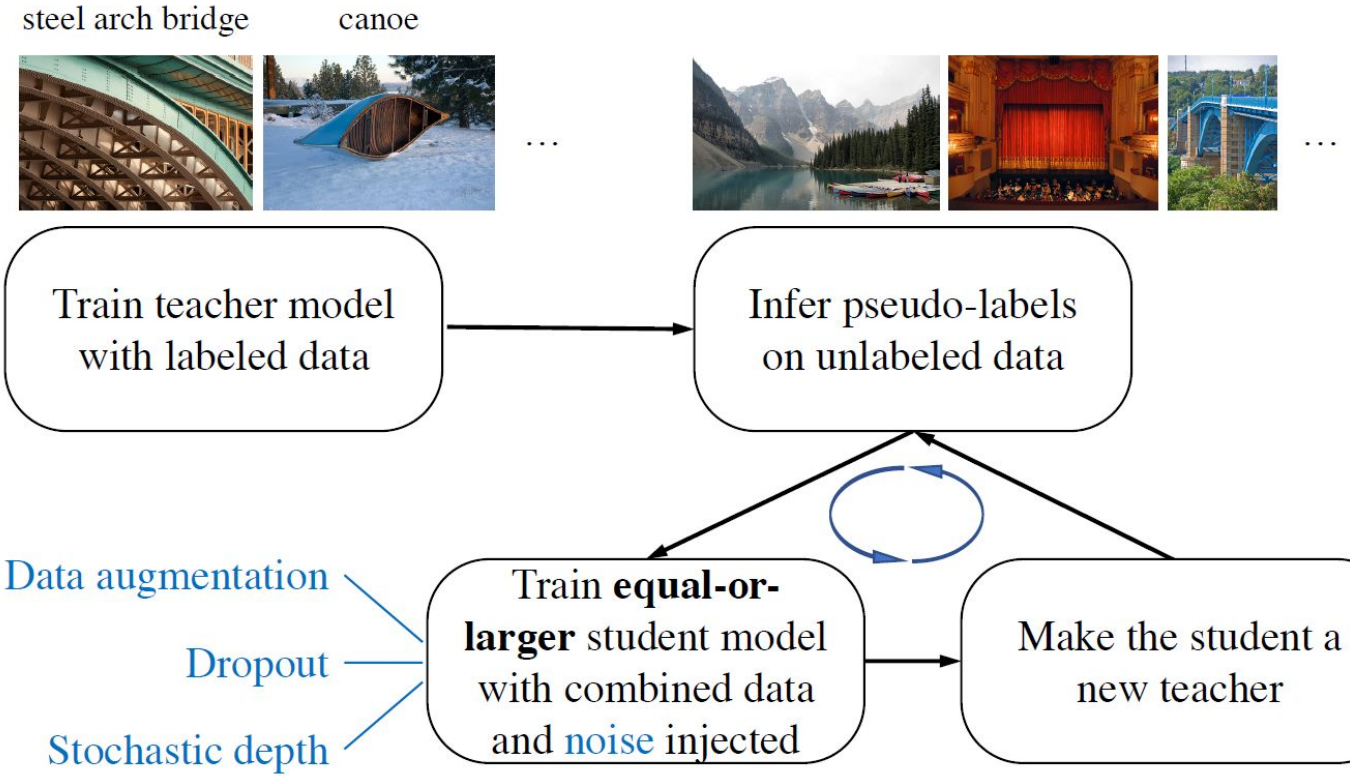
EfficientNets

- **EfficientNet-B1** - **EfficientNet-B7** получаются при помощи Compound Model Scaling с коэффициентами 1-7 соответственно

- **EfficientNet-L2** получается из EfficientNet-B7 увеличением глубины и ширины, но уменьшением разрешения.

Architecture Name	w	d	Train Res.	Test Res.	# Params
EfficientNet-B7	2.0	3.1	600	600	66M
EfficientNet-L2	4.3	5.3	475	800	480M

Noisy Student Training



ImageNet - данные

- Размеченные данные - ImageNet(train ~1M, test 100K)
- Неразмеченные - JFT(~300M, метки игнорируются)
- EfficientNet-B0, обученный на ImageNet, предсказывает метки на JFT
- Из JFT отбираются изображения с уверенностью модели хотя бы 0.3 и не более 130К для каждого класса (~81M)
- Балансировка классов: для всех классов, где меньше 130К изображений случайные дублируются(130M)

ImageNet - шум

- Stochastic depth: Linear decay rule(P_L = 0.8)
- Dropout: последний слой (dropout rate = 0.5)
- RandAugment: N = 2, M = 27

ImageNet - итеративное обучение

- Учитель на 1 шаге EfficientNet-B7
- Студенты:

Iteration	Model	Batch Size Ratio	Top-1 Acc.
1	EfficientNet-L2	14:1	87.6%
2	EfficientNet-L2	14:1	88.1%
3	EfficientNet-L2	28:1	88.4%

ImageNet - результаты

Method	# Params	Extra Data	Top-1 Acc.	Top-5 Acc.
EfficientNet-B7 [83]	66M	-	85.0%	97.2%
EfficientNet-L2 [83]	480M	-	85.5%	97.5%
ResNet-50 Billion-scale [93]	26M	3.5B images labeled with tags	81.2%	96.0%
ResNeXt-101 Billion-scale [93]	193M		84.8%	-
ResNeXt-101 WSL [55]	829M		85.4%	97.6%
FixRes ResNeXt-101 WSL [86]	829M		86.4%	98.0%
Big Transfer (BiT-L) [43]†	928M	300M weakly labeled images from JFT	87.5%	98.5%
Noisy Student Training (EfficientNet-L2)	480M	300M unlabeled images from JFT	**88.4%**	**98.7%**

ImageNet-A

sea lion lighthouse

submarine canoe

dragonfly bullfrog

starfish wreck

hummingbird bald eagle

basketball parking meter

ImageNet-C

snow leopard electric ray

swing mosquito net

toaster pill bottle

gown ski

parking meter vacuum

cannon television

ImageNet-P

plate rack refrigerator

racing car car wheel

plate rack medicine chest

racing car fire engine

plate rack medicine chest

racing car car wheel

Итоги

- EfficientNets - класс масштабируемых сверточных сетей в зависимости от ресурсов
- Self-training with Noisy Student даёт почти + 3% top-1 accuracy на ImageNet
- Модель с использованием self-training with Noisy Student более устойчива к шуму, поворотам, сдвигам …
- Одна итерация алгоритма похожа на дистилляцию.
  - Отличия: шум, студент не меньше учителя

Источники

- Self-training with Noisy Student: https://arxiv.org/abs/1911.04252
- Distillation:https://github.com/bayesgroup/HSE_ML_research_seminar/blob/master/2020-2021/182/14_Elenik_distillation.pdf
- Stochastic Depth:
  - https://arxiv.org/abs/1603.09382
  - towardsdatascience.com - красивые картинки
- RandAugment: https://arxiv.org/abs/1909.13719
- EfficientNet: https://arxiv.org/abs/1905.11946
- MBConv: https://paperswithcode.com/method/inverted-residual-block