# Self-training with Noisy Student improves ImageNet classification

**Authors:**

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, Quoc V. Le

**Presenting student:**

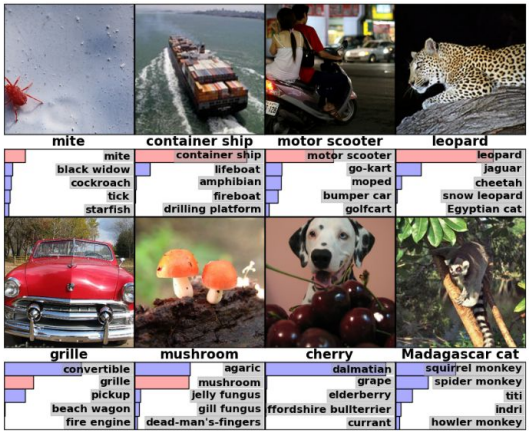
Sergio Izquierdo Barranco

# Outline

- Introduction
- Related works
- Noisy Student Training
- Experiments
- Conclusion

# 1. Introduction

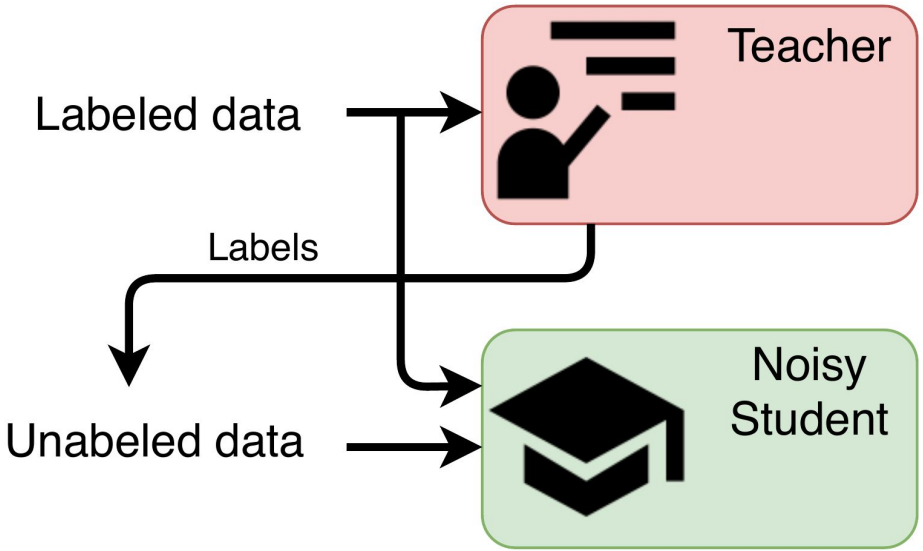
- Image classification
- Requires large amount of labeled data
  - Expensive to acquire
- How to take advantage of unlabeled data?




Krizhevsky et al. ImageNet Classification with Deep CNN

# Overview of Noisy Student Training

- Student improves the teacher
- Outperforms state-of-the-art methods: from 86.4% to **88.4**%



# 2. Related works

- Weakly labeled data
- Teacher-student approaches
  - Knowledge Distillation
  - Data Distillation

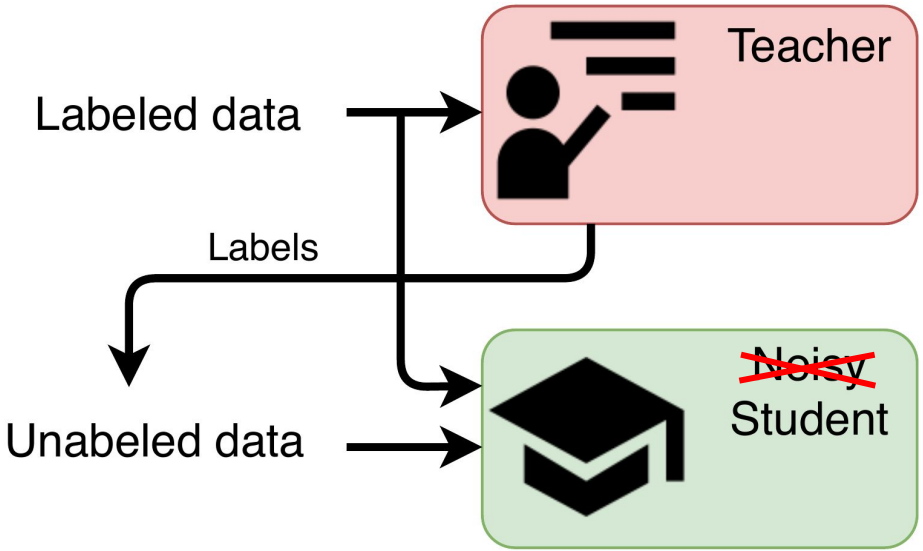
UNI FREIBURG

# Weakly Labeled Data

- Previous state-of-the-art [Mahajan et al. 2018]
- 3.5 billion Instagram images
- Labeled using hashtags

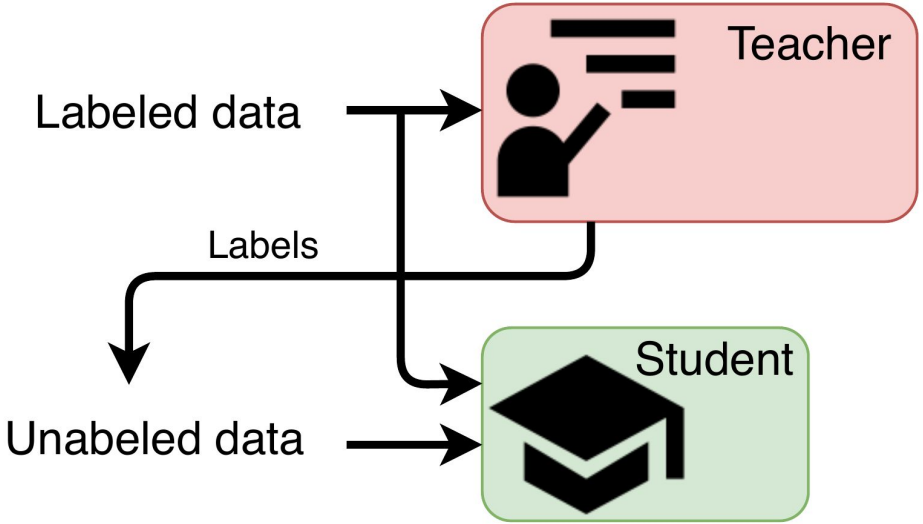
# Teacher-Student

- Noise is not used or not understood



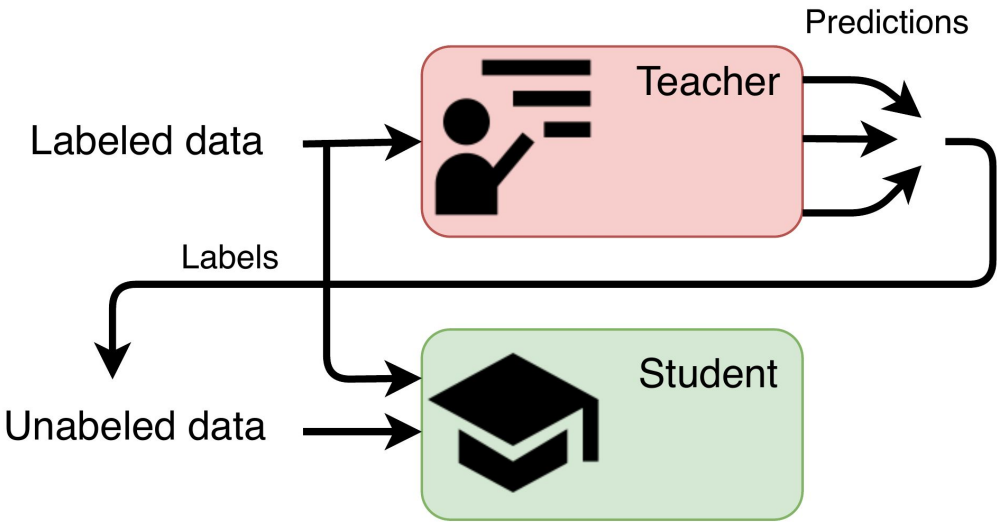
# Knowledge Distillation

- Model compression
- Student is smaller

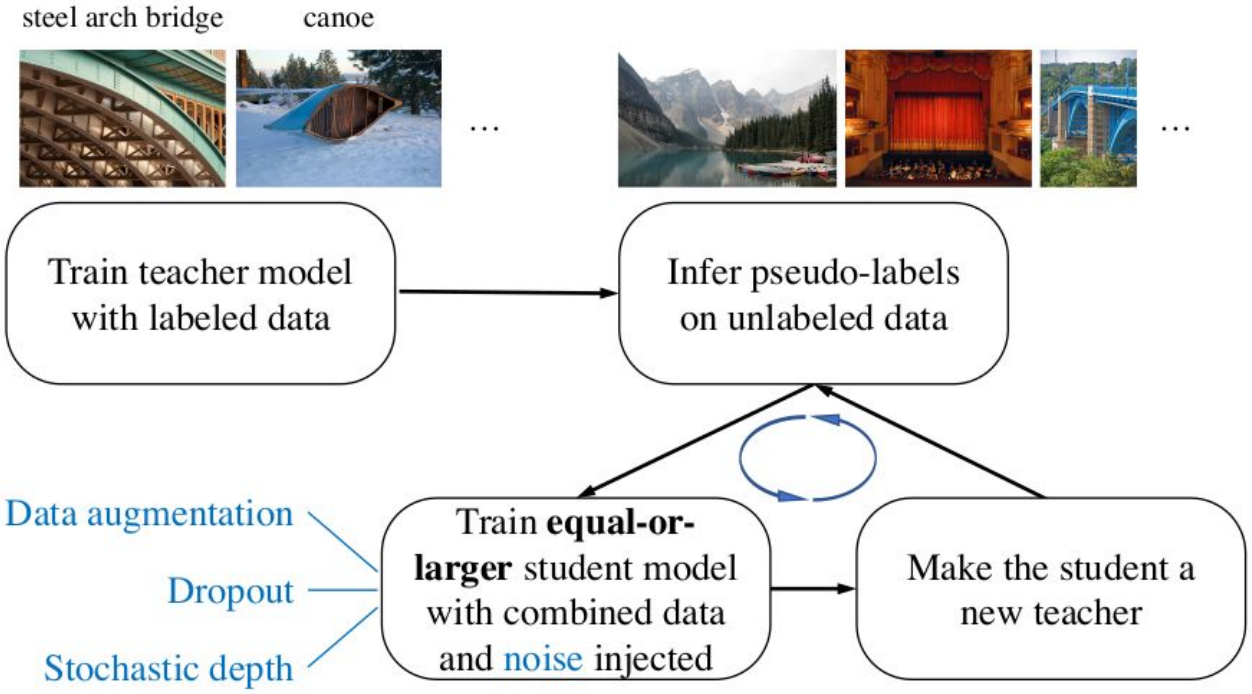


# Data Distillation [Radosavovic et al. 2017]

- Ensemble teacher predictions
- Strengthens the teacher

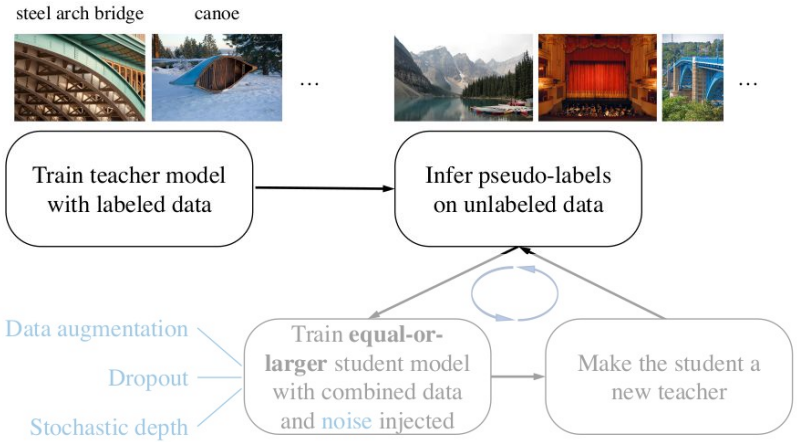


# 3. Noisy Student Training



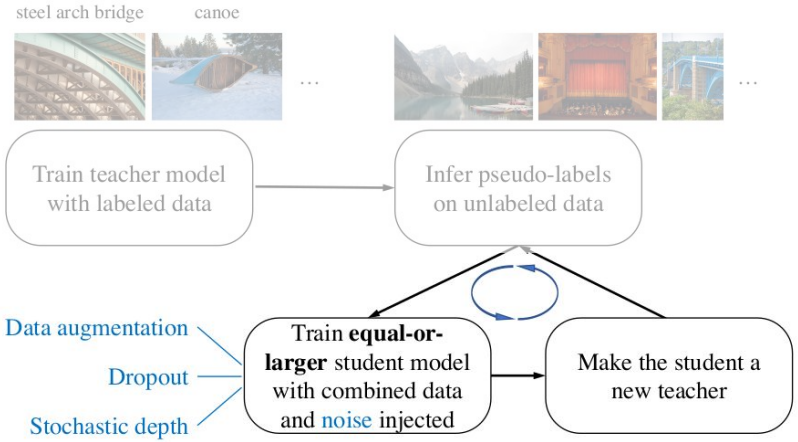
# Teacher

- Learns on labeled data and is trained until convergence
- Generates soft or hard labels for unlabeled data
- Noise-free on inference



# Student

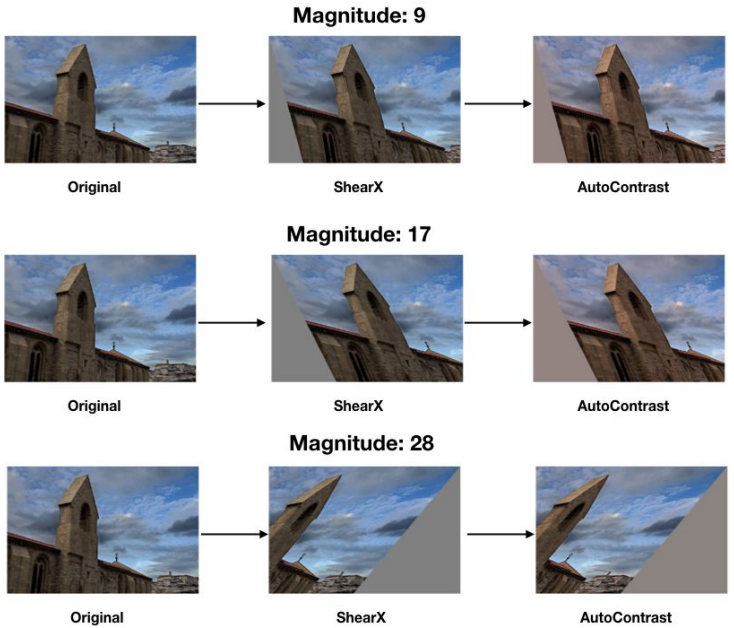
- Same or more capacity than the teacher
- Trains on both labeled and unlabeled data (with teacher labels)
- Once trained works as the teacher
- Noise (input and model noise)



# Student Input Noise

- RandAugment
- Prediction consistency

RandAugment

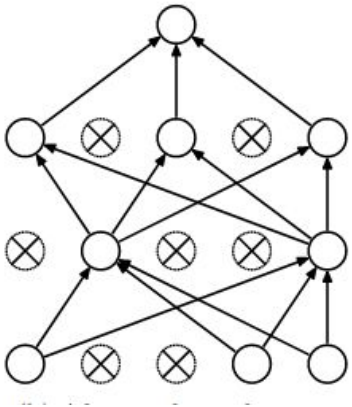



[Cubuk et al. 2020]

# Student Model Noise

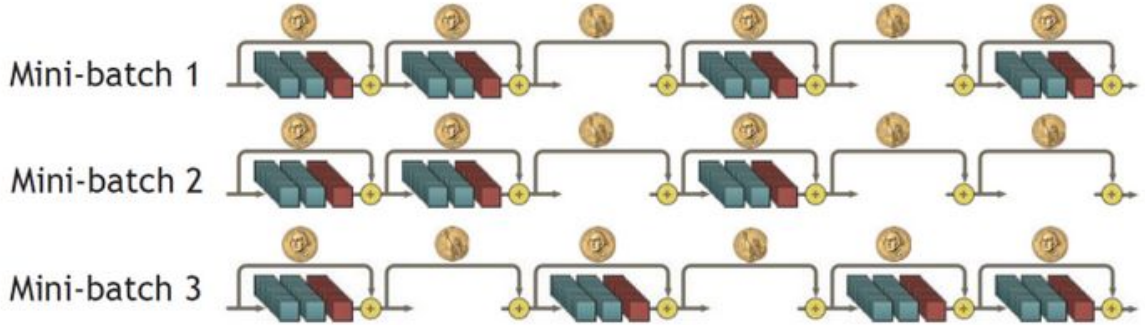
- Dropout and Stochastic Depth
- Weakens the student

Dropout



[Srivastava et al. 2014]

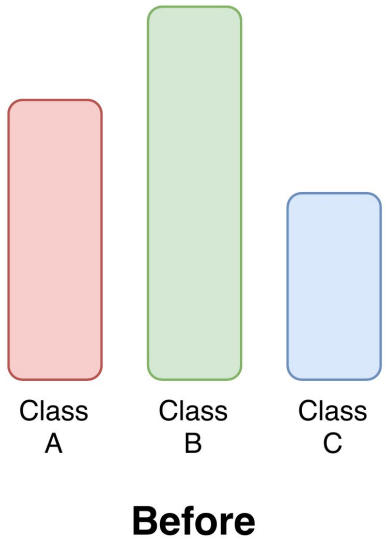
Stochastic Depth




[Huang et al. 2016]

# Implementation Details

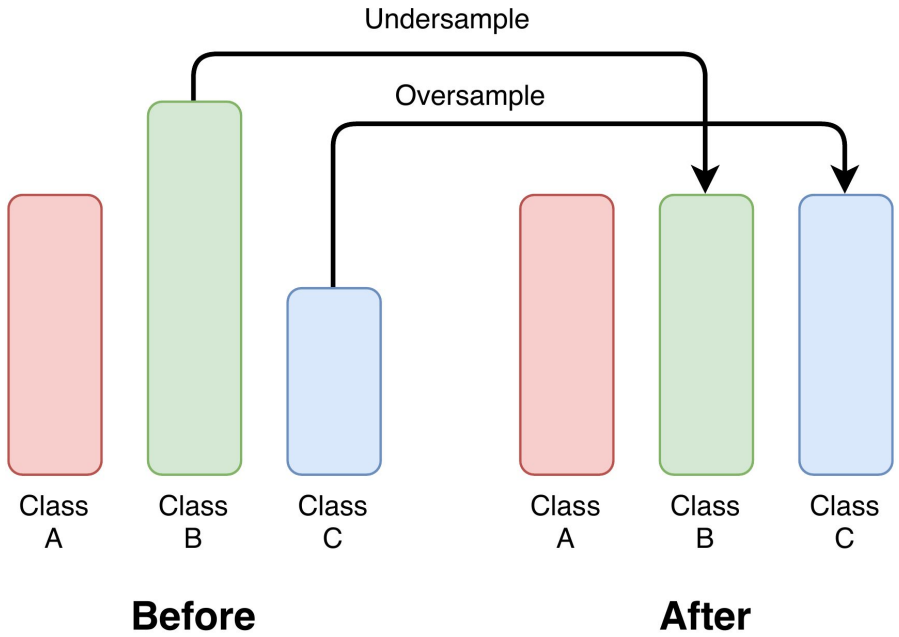
- Data balancing



UNI FREIBURG

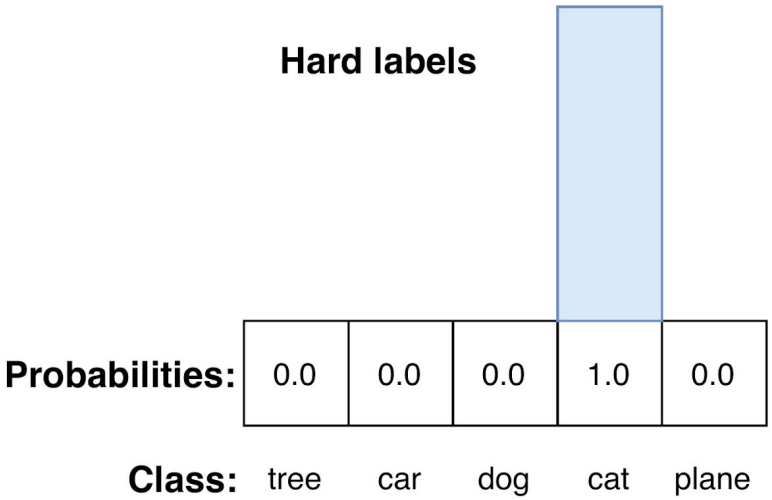
# Implementation Details

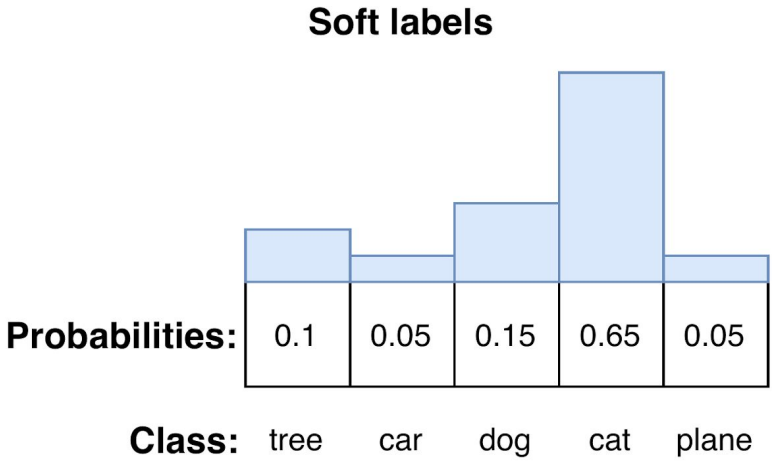
- Data balancing



# Implementation Details

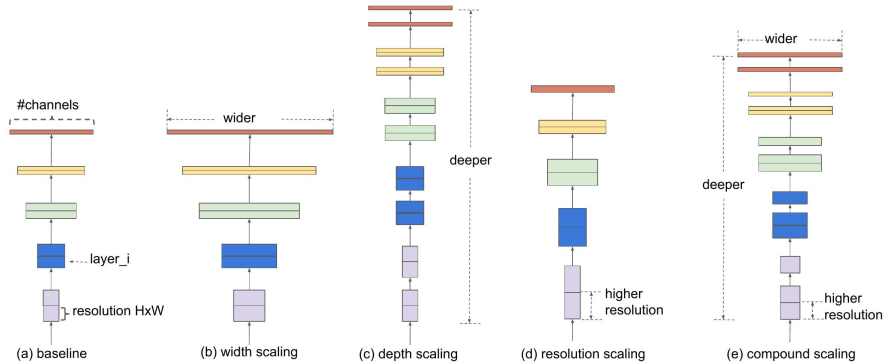
- Use of soft labels



# Implementation Details

- Model architecture: EfficientNet [Mingxing Tan and Quoc V. Le 2019]
- Uniformly scales depth/width/resolution
- EfficientNet B0, B1, …, B7 and **L2**

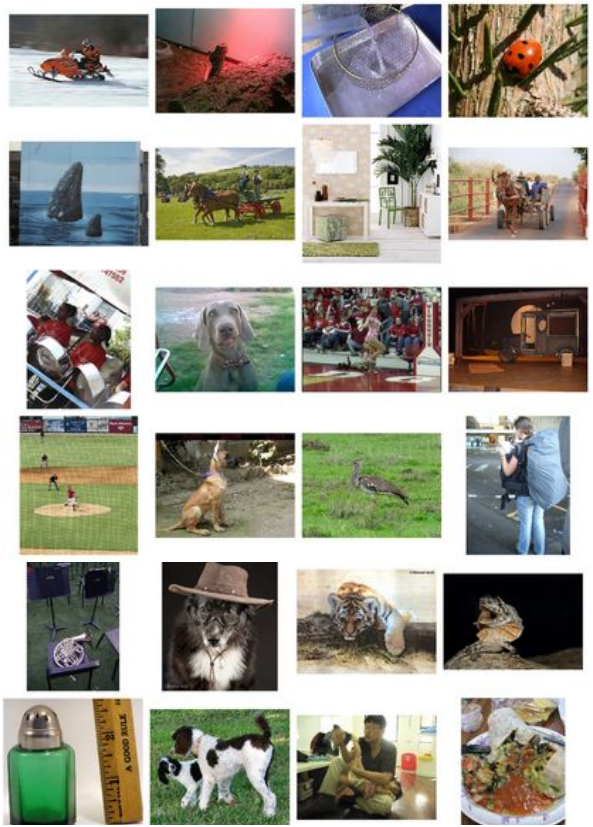



[Mingxing Tan and Quoc V. Le 2019]

$$\text{depth: } d = \alpha^\phi$$
$$\text{width: } w = \beta^\phi$$
$$\text{resolution: } r = \gamma^\phi$$
$$\text{s.t. } \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$$
$$\alpha \geq 1, \beta \geq 1, \gamma \geq 1$$

# 4. Experiments

- Datasets
  - ImageNet2012 ILSVRC (public)
  - JFT dataset (not public)
  - YFCC100M (public)
- Metric: Accuracy
- Iterative process: B7 - L2 - L2 - L2
- Just L2 takes 6 days of training on TPU



[ImageNet 2015]

UNI FREIBURG

# Results

| Method | # Params | Extra Data | Top-1 Acc. | Top-5 Acc. |
|---|---|---|---|---|
| ResNet-50 | 26M | - | 76.0% | 93.0% |
| ResNet-152 | 60M | - | 77.8% | 93.8% |
| DenseNet-264 | 34M | - | 77.9% | 93.9% |
| Inception-v3 | 24M | - | 78.8% | 94.4% |
| Xception | 23M | - | 79.0% | 94.5% |
| Inception-v4 | 48M | - | 80.0% | 95.0% |
| Inception-resnet-v2 | 56M | - | 80.1% | 95.1% |
| ResNeXt-101 | 84M | - | 80.9% | 95.6% |
| PolyNet | 92M | - | 81.3% | 95.8% |
| SENet | 146M | - | 82.7% | 96.2% |
| NASNet-A | 89M | - | 82.7% | 96.2% |
| AmoebaNet-A | 87M | - | 82.8% | 96.1% |
| PNASNet | 86M | - | 82.9% | 96.2% |
| AmoebaNet-C | 155M | - | 83.5% | 96.5% |
| GPipe | 557M | - | 84.3% | 97.0% |
| EfficientNet-B7 | 66M | - | 85.0% | 97.2% |
| EfficientNet-L2 | 480M | - | 85.5% | 97.5% |
| ResNet-50 Billion-scale | 26M | 3.5B images labeled with tags | 81.2% | 96.0% |
| ResNeXt-101 Billion-scale | 193M | | 84.8% | - |
| ResNeXt-101 WSL | 829M | | 85.4% | 97.6% |
| FixRes ResNeXt-101 WSL | 829M | | 86.4% | 98.0% |
| Big Transfer (BiT-L) † | 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%** |

UNI FREIBURG

# Results

| Method | # Params | Extra Data | Top-1 Acc. | Top-5 Acc. |
|---|---|---|---|---|
| ResNet-50 | 26M | - | 76.0% | 93.0% |
| ResNet-152 | 60M | - | 77.8% | 93.8% |
| DenseNet-264 | 34M | - | 77.9% | 93.9% |
| Inception-v3 | 24M | - | 78.8% | 94.4% |
| Xception | 23M | - | 79.0% | 94.5% |
| Inception-v4 | 48M | - | 80.0% | 95.0% |
| Inception-resnet-v2 | 56M | - | 80.1% | 95.1% |
| ResNeXt-101 | 84M | - | 80.9% | 95.6% |
| PolyNet | 92M | - | 81.3% | 95.8% |
| SENet | 146M | - | 82.7% | 96.2% |
| NASNet-A | 89M | - | 82.7% | 96.2% |
| AmoebaNet-A | 87M | - | 82.8% | 96.1% |
| PNASNet | 86M | - | 82.9% | 96.2% |
| AmoebaNet-C | 155M | - | 83.5% | 96.5% |
| GPipe | 557M | - | 84.3% | 97.0% |
| EfficientNet-B7 | 66M | - | 85.0% | 97.2% |
| EfficientNet-L2 | 480M | - | 85.5% | 97.5% |
| ResNet-50 Billion-scale | 26M | 3.5B images labeled with tags | 81.2% | 96.0% |
| ResNeXt-101 Billion-scale | 193M | | 84.8% | - |
| ResNeXt-101 WSL | 829M | | 85.4% | 97.6% |
| FixRes ResNeXt-101 WSL | 829M | | 86.4% | 98.0% |
| Big Transfer (BiT-L) † | 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%** |

UNI FREIBURG

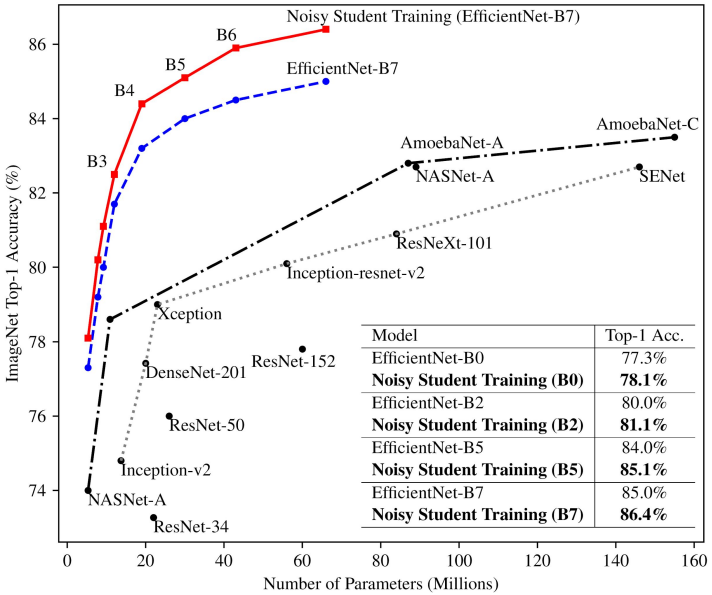
# Results

| Method | # Params | Extra Data | Top-1 Acc. | Top-5 Acc. |
|---|---|---|---|---|
| ResNet-50 | 26M | - | 76.0% | 93.0% |
| ResNet-152 | 60M | - | 77.8% | 93.8% |
| DenseNet-264 | 34M | - | 77.9% | 93.9% |
| Inception-v3 | 24M | - | 78.8% | 94.4% |
| Xception | 23M | - | 79.0% | 94.5% |
| Inception-v4 | 48M | - | 80.0% | 95.0% |
| Inception-resnet-v2 | 56M | - | 80.1% | 95.1% |
| ResNeXt-101 | 84M | - | 80.9% | 95.6% |
| PolyNet | 92M | - | 81.3% | 95.8% |
| SENet | 146M | - | 82.7% | 96.2% |
| NASNet-A | 89M | - | 82.7% | 96.2% |
| AmoebaNet-A | 87M | - | 82.8% | 96.1% |
| PNASNet | 86M | - | 82.9% | 96.2% |
| AmoebaNet-C | 155M | - | 83.5% | 96.5% |
| GPipe | 557M | - | 84.3% | 97.0% |
| EfficientNet-B7 | 66M | - | 85.0% | 97.2% |
| EfficientNet-L2 | 480M | - | 85.5% | 97.5% |
| ResNet-50 Billion-scale | 26M | 3.5B images labeled with tags | 81.2% | 96.0% |
| ResNeXt-101 Billion-scale | 193M | | 84.8% | - |
| ResNeXt-101 WSL | 829M | | 85.4% | 97.6% |
| FixRes ResNeXt-101 WSL | 829M | | 86.4% | 98.0% |
| Big Transfer (BiT-L) † | 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%** |

UNI FREIBURG

# Model Size Study

- Noisy Student can improve other models
- Even without iterative training



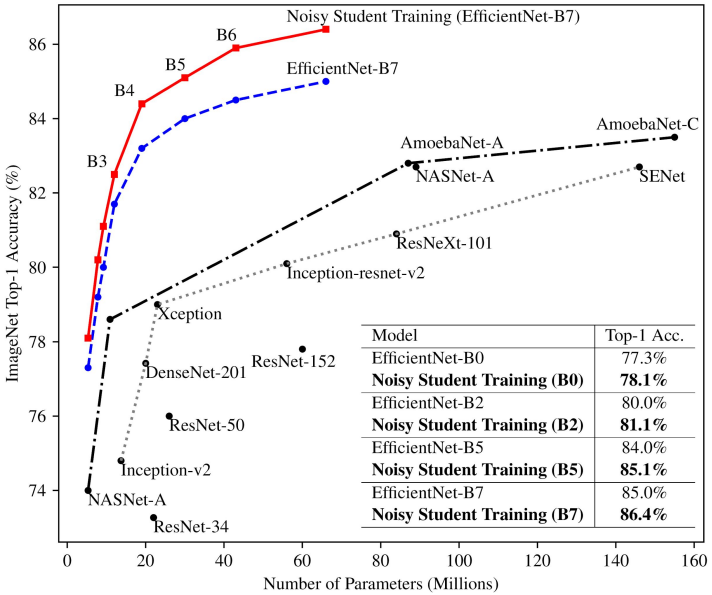

| Model | Top-1 Acc. |
|---|---|
| EfficientNet-B0 | 77.3% |
| **Noisy Student Training (B0)** | **78.1%** |
| EfficientNet-B2 | 80.0% |
| **Noisy Student Training (B2)** | **81.1%** |
| EfficientNet-B5 | 84.0% |
| **Noisy Student Training (B5)** | **85.1%** |
| EfficientNet-B7 | 85.0% |
| **Noisy Student Training (B7)** | **86.4%** |

# Model Size Study

- Noisy Student can improve other models
- Even without iterative training

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

Noisy Student on ResNet-50




| Model | Top-1 Acc. |
|---|---|
| EfficientNet-B0 | 77.3% |
| **Noisy Student Training (B0)** | **78.1%** |
| EfficientNet-B2 | 80.0% |
| **Noisy Student Training (B2)** | **81.1%** |
| EfficientNet-B5 | 84.0% |
| **Noisy Student Training (B5)** | **85.1%** |
| EfficientNet-B7 | 85.0% |
| **Noisy Student Training (B7)** | **86.4%** |

UNI FREIBURG

# Robustness

- ImageNet-A (hard images)

| Method | Top-1 Acc. | Top-5 Acc. |
|---|---|---|
| ResNet-101 [32] | 4.7% | - |
| ResNeXt-101 [32] (32x4d) | 5.9% | - |
| ResNet-152 [32] | 6.1% | - |
| ResNeXt-101 [32] (64x4d) | 7.3% | - |
| DPN-98 [32] | 9.4% | - |
| ResNeXt-101+SE [32] (32x4d) | 14.2% | - |
| ResNeXt-101 WSL [55, 59] | 61.0% | - |
| EfficientNet-L2 | 49.6% | 78.6% |
| **Noisy Student Training (L2)** | **83.7%** | **95.2%** |

UNI FREIBURG

# Robustness

- ImageNet-C (images with corruptions)

| Method | Res. | Top-1 Acc. | mCE |
|---|---|---|---|
| ResNet-50 [31] | 224 | 39.0% | 76.7 |
| SIN [23] | 224 | 45.2% | 69.3 |
| Patch Gaussian [51] | 299 | 52.3% | 60.4 |
| ResNeXt-101 WSL [55, 59] | 224 | - | 45.7 |
| EfficientNet-L2 | 224 | 62.6% | 47.5 |
| Noisy Student Training (L2) | 224 | 76.5% | 30.0 |
| EfficientNet-L2 | 299 | 66.6% | 42.5 |
| **Noisy Student Training (L2)** | 299 | **77.8%** | **28.3** |

UNI FREIBURG

# Robustness

- ImageNet-P (images with perturbations)

| Method | Res. | Top-1 Acc. | mFR |
|---|---|---|---|
| ResNet-50 [31] | 224 | - | 58.0 |
| Low Pass Filter Pooling [99] | 224 | - | 51.2 |
| ResNeXt-101 WSL [55, 59] | 224 | - | 27.8 |
| EfficientNet-L2 | 224 | 80.4% | 27.2 |
| Noisy Student Training (L2) | 224 | 85.2% | 14.2 |
| EfficientNet-L2 | 299 | 81.6% | 23.7 |
| **Noisy Student Training (L2)** | 299 | **86.4%** | **12.2** |

UNI FREIBURG

# Ablations

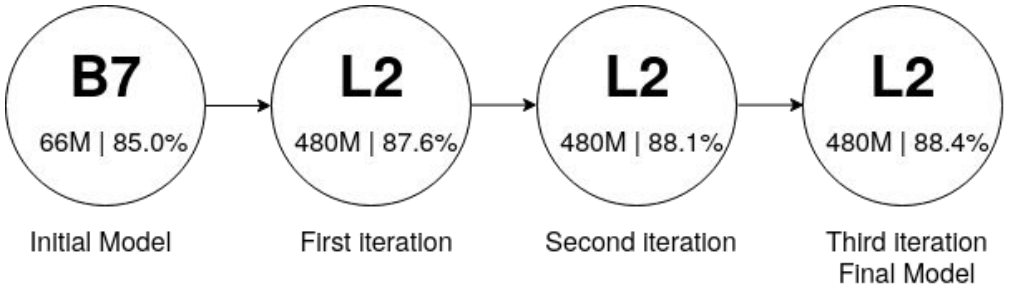
- Importance of Noise in the student
- Student should not exactly copy the teacher

| 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% |

UNI FREIBURG

# Ablations

- Importance of Iterative Training



UNI FREIBURG

# Ablations

- Larger teacher leads to better results

UNI FREIBURG

# Ablations

- Larger teacher leads to better results
- The more unlabeled data the best

UNI FREIBURG

# Ablations

- Larger teacher leads to better results
- The more unlabeled data the best
- Soft labels are preferred

UNI FREIBURG

# Ablations

- Larger teacher leads to better results
- The more unlabeled data the best
- Soft labels are preferred
- Larger student leads to better results

UNI FREIBURG

# Ablations

- Larger teacher leads to better results
- The more unlabeled data the best
- Soft labels are preferred
- Larger student leads to better results
- Balancing the data can be useful

UNI FREIBURG

# Ablations

- Larger teacher leads to better results
- The more unlabeled data the best
- Soft labels are preferred
- Larger student leads to better results
- Balancing the data can be useful
- Jointly training on labeled and unlabeled gives better results

UNI FREIBURG

# Ablations

- Larger teacher leads to better results
- The more unlabeled data the best
- Soft labels are preferred
- Larger student leads to better results
- Balancing the data can be useful
- Jointly training on labeled and unlabeled gives better results
- A large ratio between unlabeled and labeled is preferred

UNI FREIBURG

# Ablations

- Larger teacher leads to better results
- The more unlabeled data the best
- Soft labels are preferred
- Larger student leads to better results
- Balancing the data can be useful
- Jointly training on labeled and unlabeled gives better results
- A large ratio between unlabeled and labeled is preferred
- Training the student from scratch can be beneficial

# 5. Conclusion

- Significant improvement using unlabeled data
  - SOTA of ImageNet with 88.4%
- Applying Noisy Student improves performance also for smaller models or different architectures
- Significant increase in robustness over similar methods

# Thank you!