Self-training with Noisy Student improves ImageNet classification

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

Presented by Minoo

Content

- Introduction
- Knowledge distillation
- Self-training and distillation methods
- Noisy student training algorithm
- Architecture
- Results
- Conclusion

Introduction

Noisy student approach:

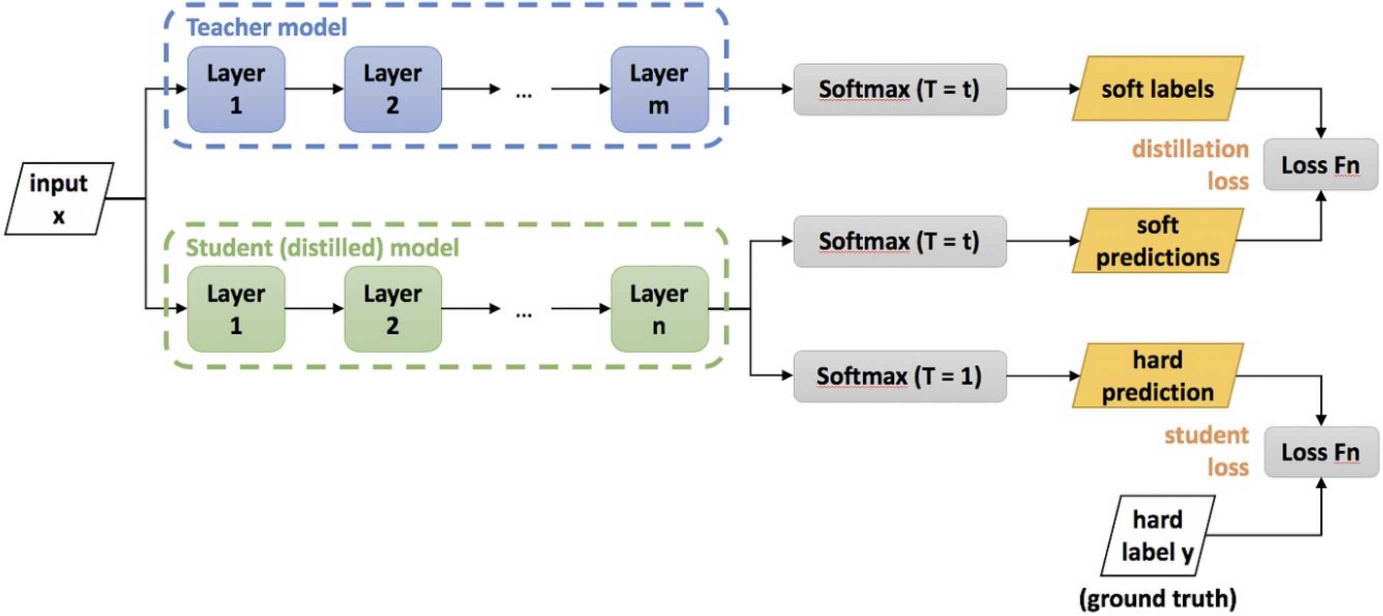
-semi supervised learning approach

-self training and knowledge distillation method

-unlabelled images (out of distribution images)

-accuracy and robustness of imagnet model improved

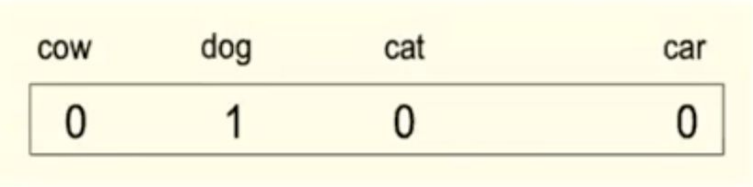
Knowledge distillation



(Image from https://nervanasystems.github.io/distiller/knowledge_distillation.html)

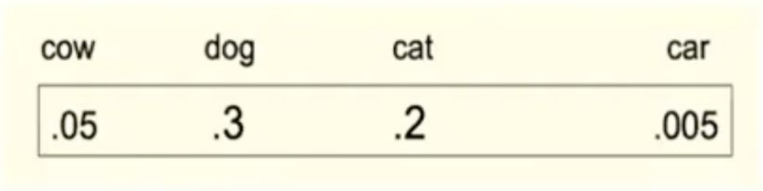
Soft vs hard labels for knowledge distillation

Hard labels



cow	dog	cat	car
0	1	0	0

Soft labels



cow	dog	cat	car
.05	.3	.2	.005

$$p_i = \frac{exp\frac{z_i}{T}}{\sum_j exp\frac{z_i}{T}}$$

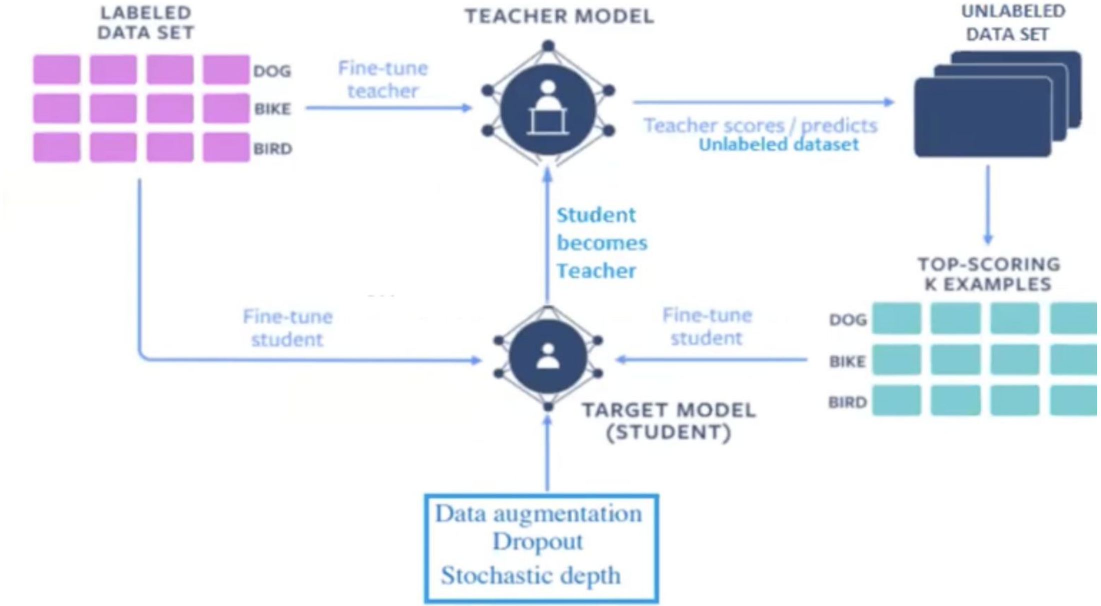
Softmax function with temperature T

For T=1, standard softmax function (Hard labels)

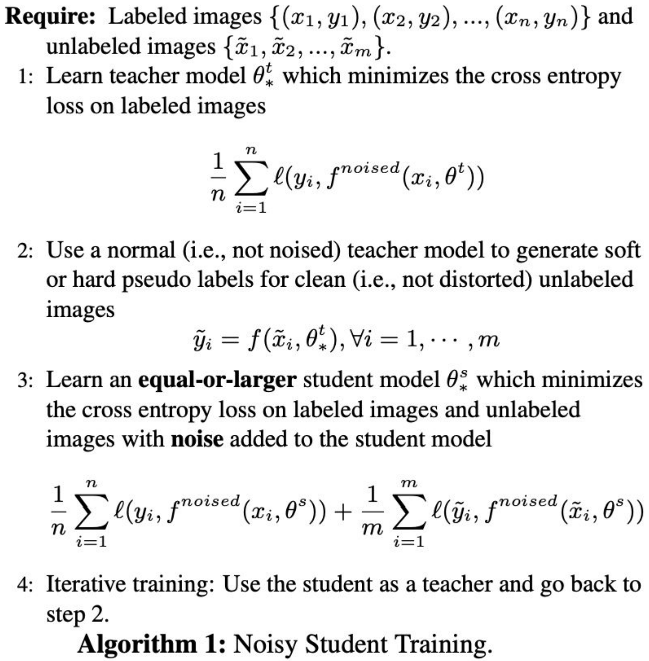
For T->inf, Soft labels

Self-training and distillation methods

1. Train a teacher model on labeled images
2. Use the teacher to generate pseudo labels on unlabeled images
3. Train a student model on the combination of labeled images and pseudo labeled images



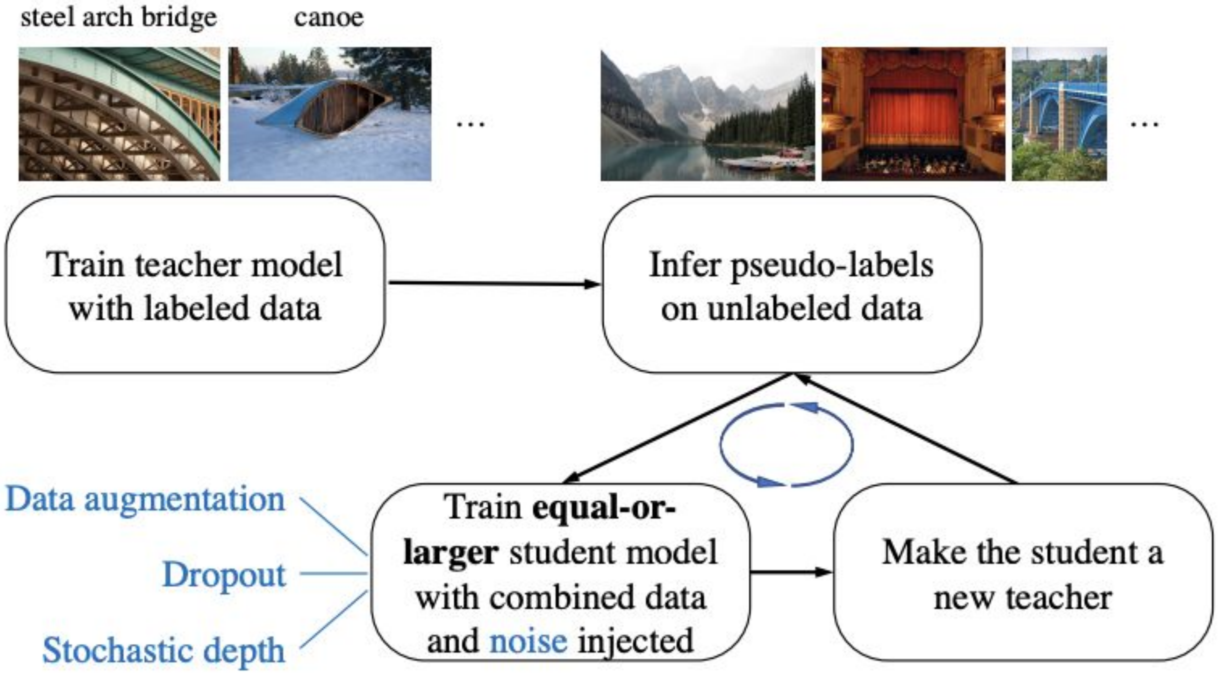
Noisy student training algorithm



```
Require: Labeled images $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ and unlabeled images $\{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_m\}$.
1: Learn teacher model $\theta^t_*$ which minimizes the cross entropy loss on labeled images
     $\frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f^{noised}(x_i, \theta^t))$
2: Use a normal (i.e., not noised) teacher model to generate soft or hard pseudo labels for clean (i.e., not distorted) unlabeled images
     $\tilde{y}_i = f(\tilde{x}_i, \theta^t_*), \forall i = 1, \cdots, m$
3: Learn an **equal-or-larger** student model $\theta^s_*$ which minimizes the cross entropy loss on labeled images and unlabeled images with **noise** added to the student model
     $\frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f^{noised}(x_i, \theta^s)) + \frac{1}{m} \sum_{i=1}^{m} \ell(\tilde{y}_i, f^{noised}(\tilde{x}_i, \theta^s))$
4: Iterative training: Use the student as a teacher and go back to step 2.
```

Algorithm 1: Noisy Student Training.

Noisy student training


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

Noise introduction techniques

1. **Data Augmentation**: Rank-based data augmentation is employed to augment images
2. **Dropout**: Dropout is applied during training. Neurons are randomly dropped out in each iteration
3. **Stochastic Depth**: Stochastic depth is used, where entire layers are skipped with a certain probability.

Data Balancing

1. In ImageNet model since the number of data in each class is the same, it is necessary to balance even with pseudo labels.
2. If a class has too few images, images are duplicated to achieve balance.
3. If a class has too many images, it will be taken in order of reliability.

Pseudo-Labels

- Soft (continuous distribution)
- Hard (one-hot)

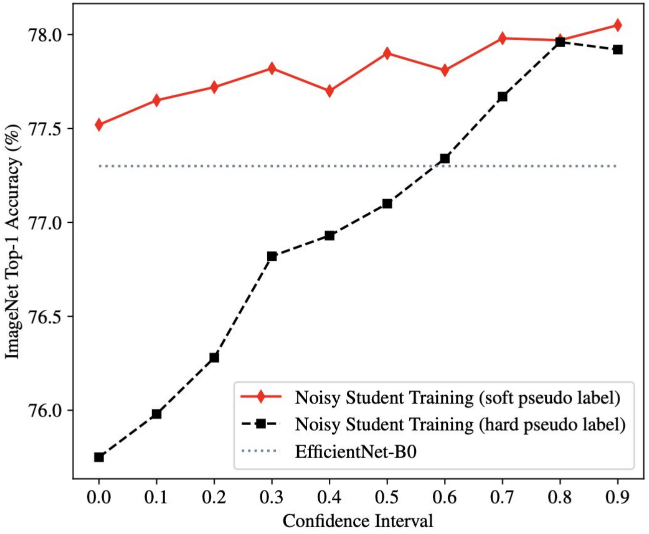
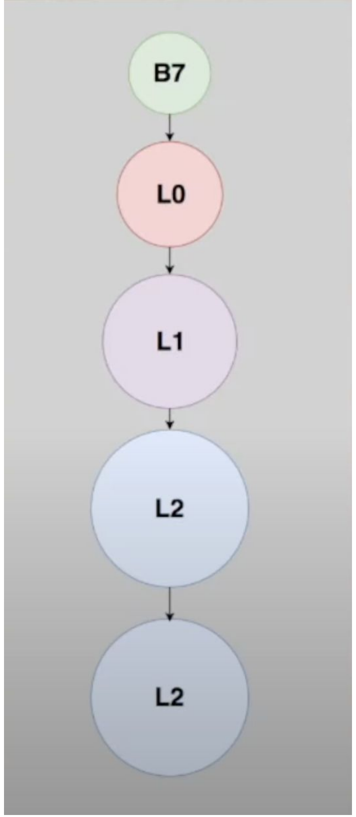
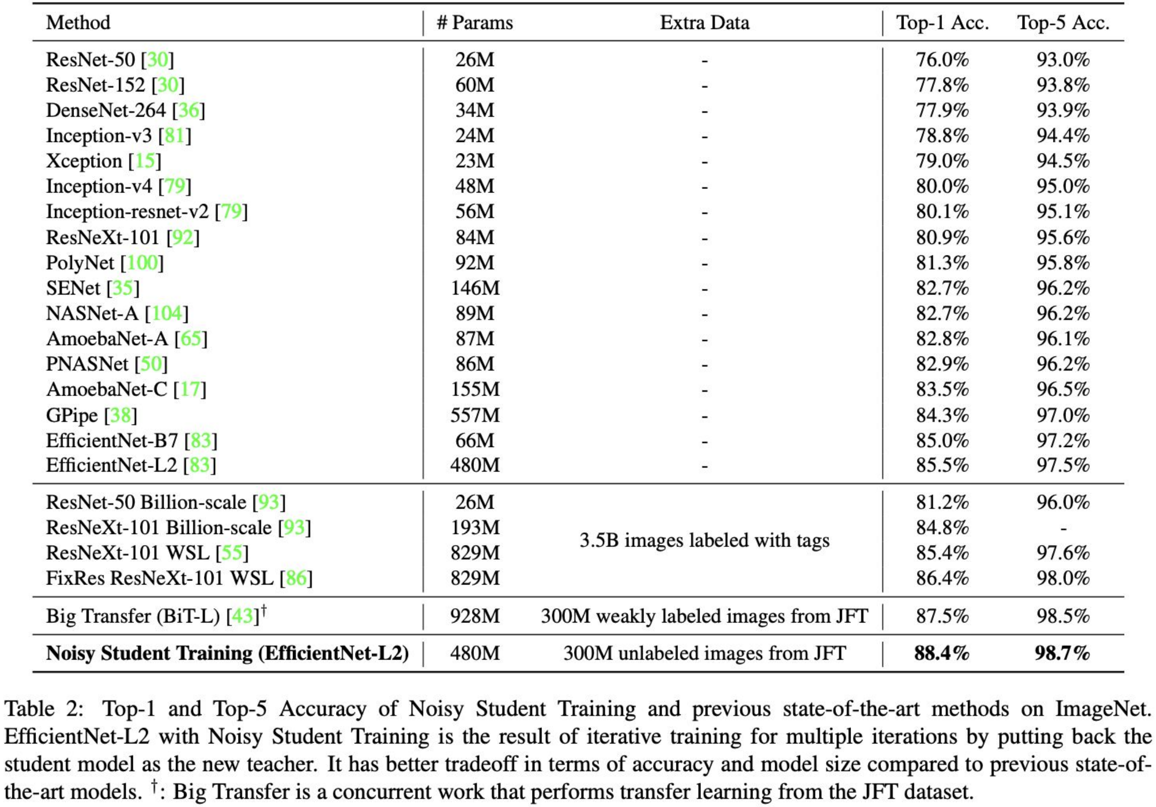


Figure 5: Soft pseudo labels lead to better performance for low confidence data (out-of-domain data). Each dot at p represents a Noisy Student Training model trained with 1.3M ImageNet labeled images and 1.3M unlabeled images with confidence scores in $[p, p + 0.1]$.

Architecture



Results



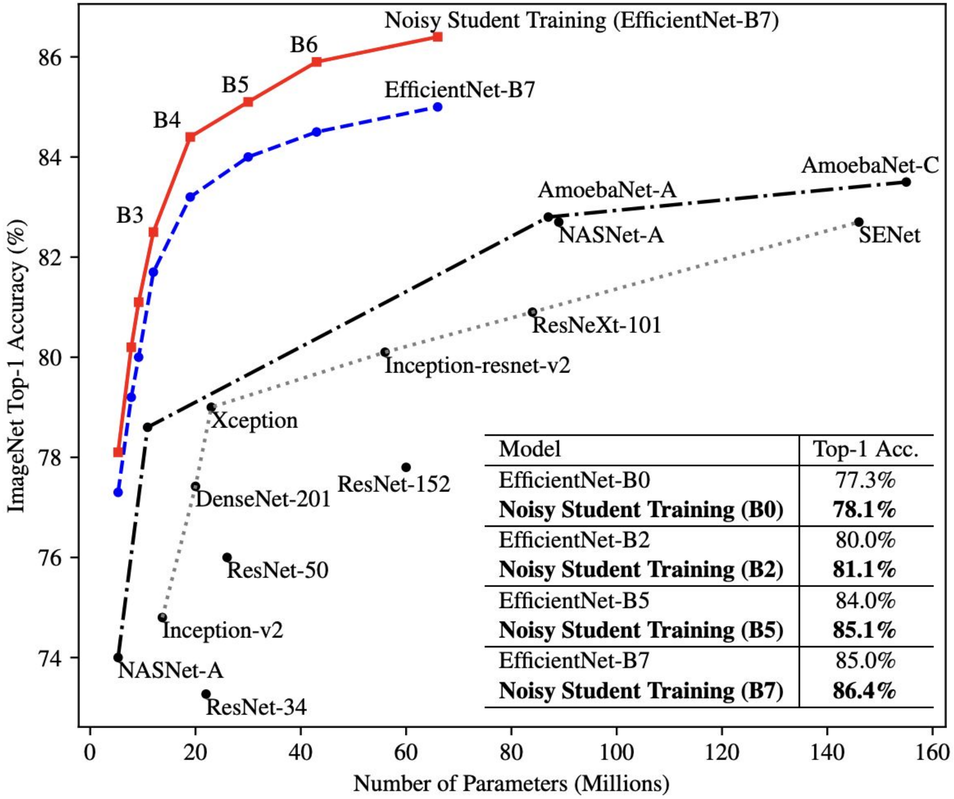
Method	# Params	Extra Data	Top-1 Acc.	Top-5 Acc.
ResNet-50 [30]	26M	-	76.0%	93.0%
ResNet-152 [30]	60M	-	77.8%	93.8%
DenseNet-264 [36]	34M	-	77.9%	93.9%
Inception-v3 [81]	24M	-	78.8%	94.4%
Xception [15]	23M	-	79.0%	94.5%
Inception-v4 [79]	48M	-	80.0%	95.0%
Inception-resnet-v2 [79]	56M	-	80.1%	95.1%
ResNeXt-101 [92]	84M	-	80.9%	95.6%
PolyNet [100]	92M	-	81.3%	95.8%
SENet [35]	146M	-	82.7%	96.2%
NASNet-A [104]	89M	-	82.7%	96.2%
AmoebaNet-A [65]	87M	-	82.8%	96.1%
PNASNet [50]	86M	-	82.9%	96.2%
AmoebaNet-C [17]	155M	-	83.5%	96.5%
GPipe [38]	557M	-	84.3%	97.0%
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%**

Table 2: Top-1 and Top-5 Accuracy of Noisy Student Training and previous state-of-the-art methods on ImageNet. EfficientNet-L2 with Noisy Student Training is the result of iterative training for multiple iterations by putting back the student model as the new teacher. It has better tradeoff in terms of accuracy and model size compared to previous state-of-the-art models. †: Big Transfer is a concurrent work that performs transfer learning from the JFT dataset.

Results

- Noisy Student Training for EfficientNet B0-B7 without Iterative Training.

- Noisy Student Training leads to consistent improvement of arou 0.8% for all model sizes



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%**

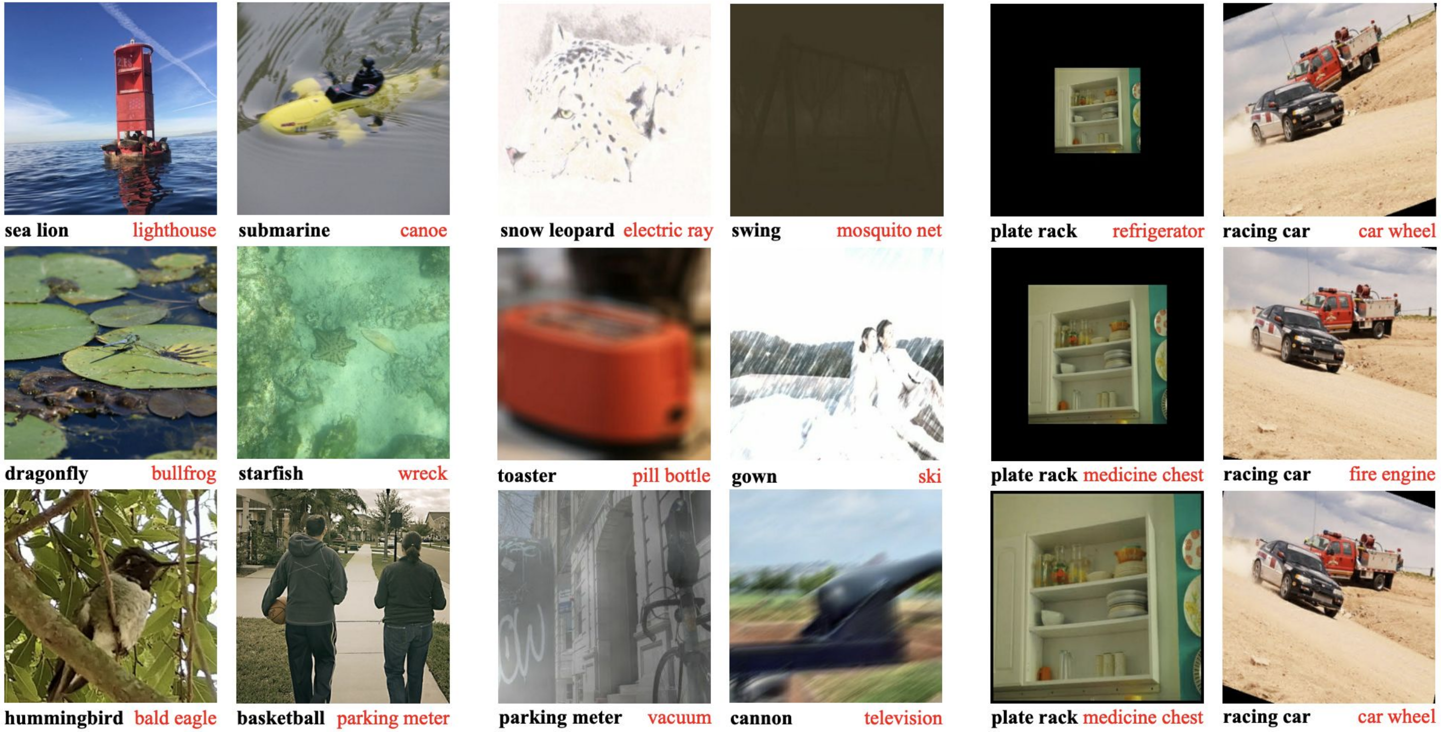
Robustness

The authors evaluated the best model that achieved an 88.4% top-one accuracy on three robustness test sets:

- **ImageNet A**: which is focused on adversarial images. This set consists of difficult images that cause significant drops in accuracy for state-of-the-art models, essentially simulating adversarial attacks with hard-to-detect images.
- **ImageNet C**: The second set involves general distortions like blurring or fogging.
- **ImageNet P**: involves rotations and scaling.

These three sets essentially test the model's resilience against adversarial attacks, blurring, and scaling/rotation challenges.

Robustness results

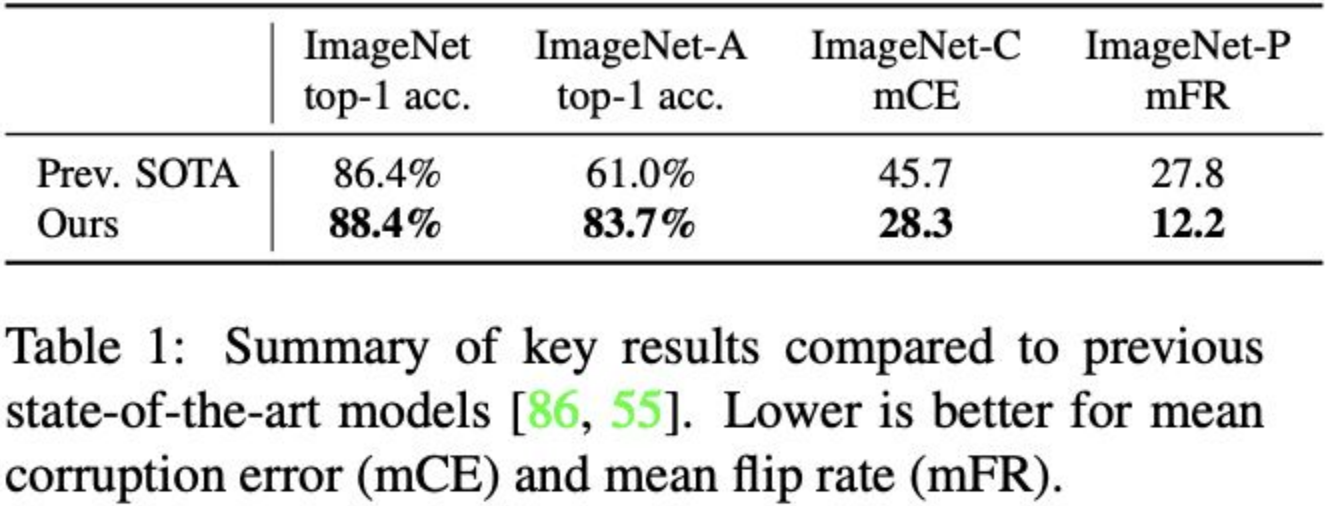


(a) ImageNet-A

(b) ImageNet-C

(c) ImageNet-P

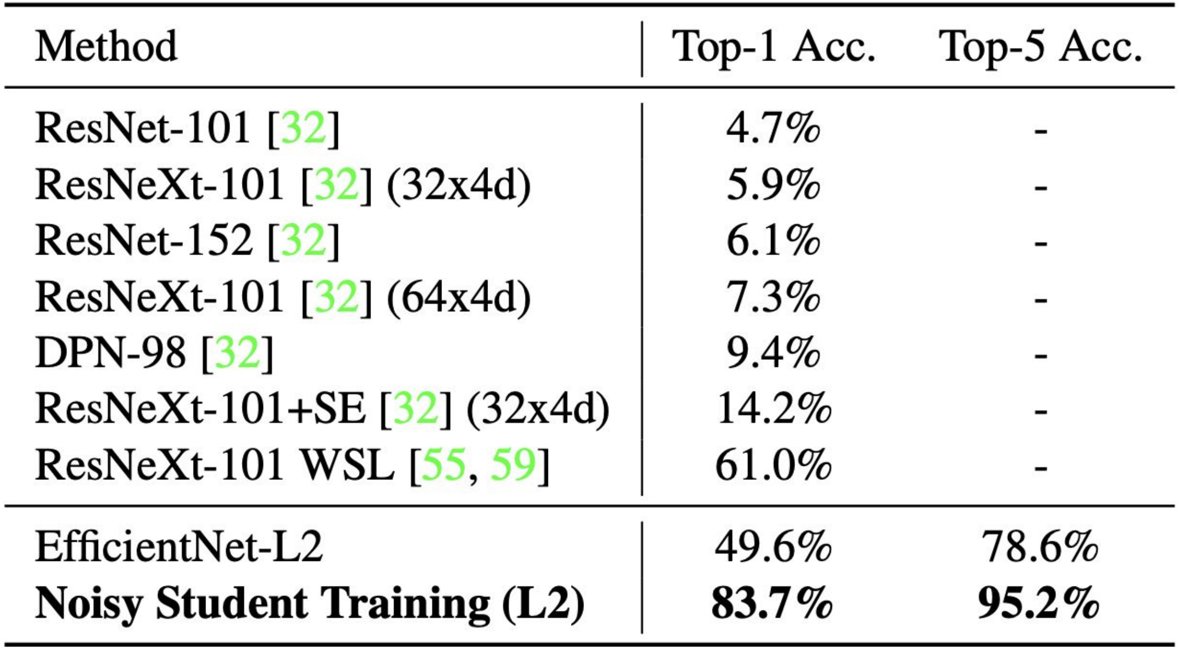
Robustness results



	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**

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).

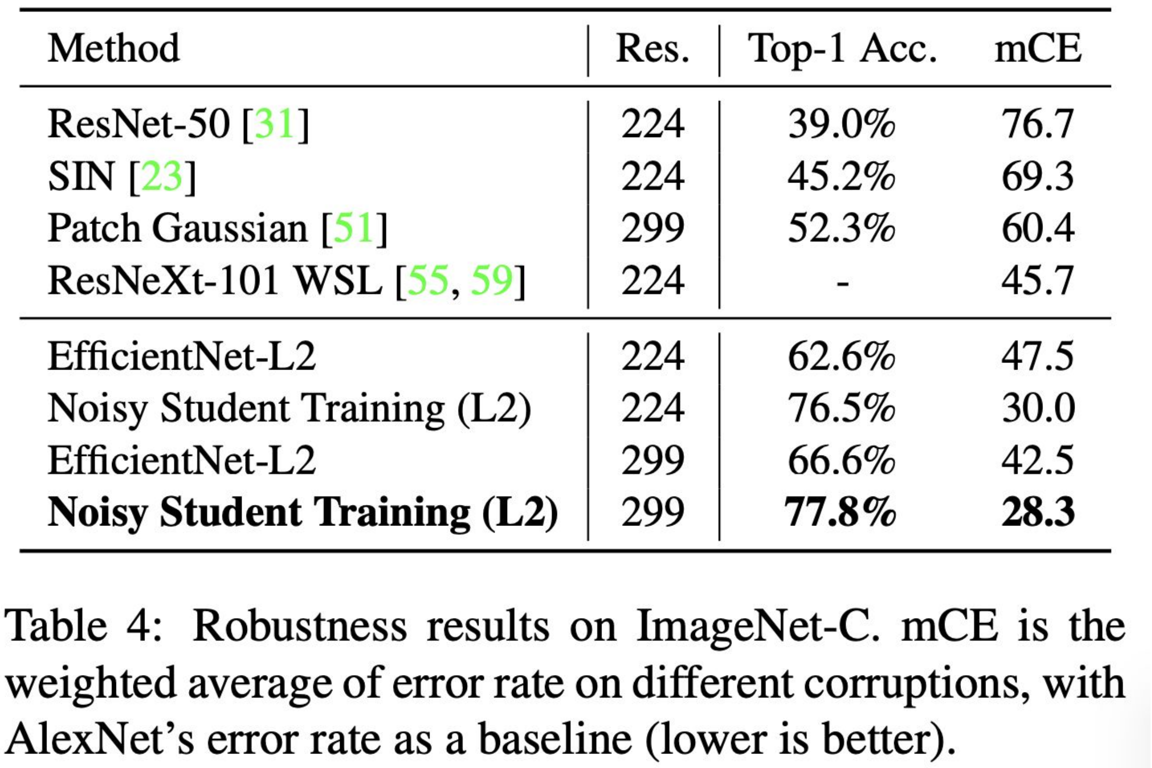
Robustness results



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%**

Table 3: Robustness results on ImageNet-A.

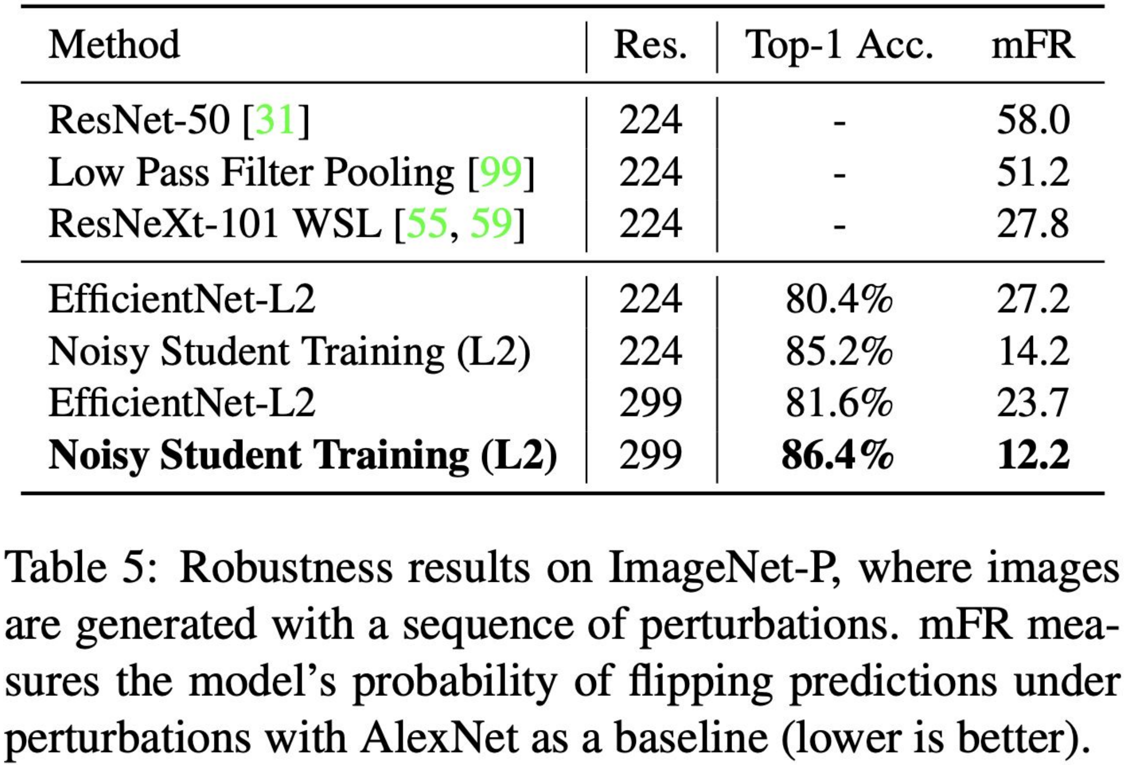
Robustness results



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**

Table 4: Robustness results on ImageNet-C. mCE is the weighted average of error rate on different corruptions, with AlexNet's error rate as a baseline (lower is better).

Robustness results



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**

Table 5: Robustness results on ImageNet-P, where images are generated with a sequence of perturbations. mFR measures the model's probability of flipping predictions under perturbations with AlexNet as a baseline (lower is better).

Conclusion

- The paper’s approach have resulted in increased accuracy and model robustness compared to previous studies that relied on weakly supervised learning.
- The new method doesn't require intentionally increasing the amount of data, yet it still enhances the model's performance.

Thank You

References

- Xie, Q., Luong, M. T., Hovy, E., & Le, Q. V. (2020). Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10687-10698).