Are we done with ImageNet?

Lucas Beyer[1*] **Olivier J. Hénaff**[2*] **Alexander Kolesnikov**[1*] **Xiaohua Zhai**[1*] **Aäron van den Oord**[2*]
[1]Google Brain (Zürich, CH) and [2]DeepMind (London, UK)

2020 arXiv

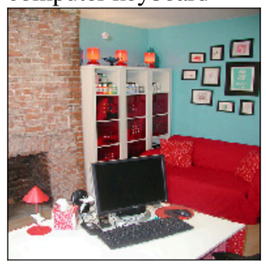
Presenter : Jason Lee

What is wrong with ImageNet?

Old label: pier
ReaL: dock; pier; speedboat; sandbar; seashore

Old label: hammer
ReaL: screwdriver; hammer; power drill; carpenter's kit

Old label: monitor
ReaL: mouse; desk; desktop computer; lamp; studio couch; monitor; computer keyboard



Old label: zucchini
ReaL: broccoli; zucchini; cucumber; orange; lemon; banana

Old label: ant
ReaL: ant; ladybug

Old label: quill
ReaL: feather boa

Old label: water jug
ReaL: water bottle

Old label: chain
ReaL: necklace

Old label: purse
ReaL: wallet

Old label: passenger car
ReaL: school bus

Old label: sunglass
ReaL: sunglass; sunglasses

Old label: sunglasses
ReaL: sunglass; sunglasses

Old label: laptop
ReaL: notebook; laptop; computer keyboard

Old label: notebook
ReaL: notebook; laptop; computer keyboard

Old label: laptop
ReaL: notebook; laptop

Single label per image

Real-world images contain multiple objects

Overly restrictive label proposals

ImageNet annotation pipeline lead to inaccuracies

Arbitrary class distinctions

ImageNet classes contain a handful of essentially duplicate pairs

Relabeling the ImageNet Validation Set

1. 각 모델마다 각 이미지에 대해 logits와 probabilities 계산
    - ImageNet validation dataset은 총 50,000장임
    - 1개의 모델은 1개의 이미지에 대해 1000개 class에 대한 logit과 probability를 가짐
    - 즉 1개의 모델은 50M logits, 50M probabilities를 가짐
2. 각 모델마다 구한 logits, probabilities 중 largest 150k만 선택하여 image-label pairs를 생성
    - 1개의 모델은 largest 150k logits, largest 150k probabilities를 가짐
    - 선택된 largest 150k logits, 150k probabilities를 이용하여 image-label pairs를 생성
    - 즉 1개의 모델은 logit을 통해 150k의 image-label pairs와 probability를 통해 150k의 image-label pairs를 가짐
3. 모든 모델에 대하여 pool
    - 모든 모델에 대하여 image-label pairs를 pool
    - collections.Counter 개념으로 보면 이해하는데 있어서 수월할 듯
4. 1개의 모델에서 1번만 등장한 image-label pairs를 제거
    - image-label pairs 중에서 한번만 등장한 경우를 제거한다는 의미와 동일
5. 모든 이미지들에 대해 각 모델별 top-1 prediction image-label pairs와 original ILSCRV-2012 image-label pairs를 추가

Instead of relabeling 50,000 images from scratch, let's relabeling in more smart way

Use pre-trained models to make label proposals

Relabeling the ImageNet Validation Set

Table 3: The models used for label proposal in this study, and their characteristics.

Ref.	Model	Arch. search	Self-sup.	Ext. data	Final
[3]	VGG-16	no	no	no	yes
[27]	Inception v3	no	no	no	yes
[4]	ResNet-50	no	no	no	no
[4]	ResNet-152	no	no	no	no
[28]	ResNeXt-101, 32x8d	no	no	no	no
[29]	ResNeXt-101, 32x8d, IG	no	no	yes	yes
[29]	ResNeXt-101, 32x48d, IG	no	no	yes	no
[30]	BiT-M	no	no	yes	yes
[30]	BiT-L	no	no	yes	yes
[31]	Assemble ResNet-50	yes	no	no	no
[31]	Assemble ResNet-152	yes	no	no	no
[32]	NASNet-A Large	yes	no	no	no
[32]	NASNet-A Mobile	yes	no	no	no
[33]	Once for all (Large)	yes	no	no	no
[34]	S4L MOAM	no	yes	no	no
[35]	CPC v2, fine-tuned	no	yes	no	yes
[35]	CPC v2, linear	no	yes	no	no
[36]	MoCo v2, long	no	yes	no	no
[37]	SimCLR	no	yes	no	no

Use 19 models to make label proposals

-> too many label proposals (13 label proposals per image)

-> find model subsets to reduce label proposals

Relabeling the ImageNet Validation Set

1. Gold standard를 생성
    - gold standard : 5명의 vision experts가 256장의 이미지를 라벨링한 데이터셋임
2. Model 선택
    - 위에서 각 모델별 label proposals가 존재
    - 이를 이용하여 gold standard에 대해 precision & recall을 구함
    - recall을 97% 이상 유지하면서 highest precision을 가지는 6개의 모델을 선택

Find 6 models using above methods

7.4 label proposals per image

(13 -> 7.4)

Relabeling the ImageNet Validation Set

Dog, Strawberry

Organism > animal > chordate > vertebrate > mammal > placental > carnivore > canine > dog > hunting dog > terrier > Irish terrier

Natural object > plant part > plant organ > reproductive structure > fruit > edible fruit > berry > strawberry

What images to relabel?

-> exclude images that all models agree with the original ImageNet label

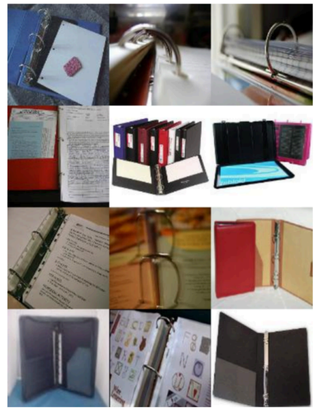
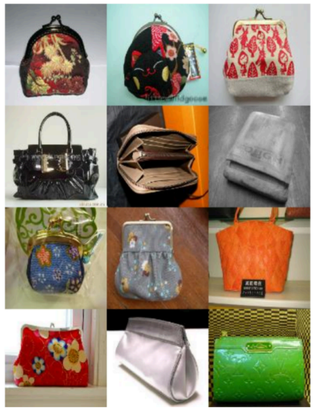
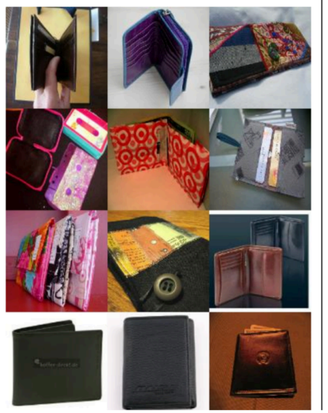
-> 24,889 images to relabel

-> split images with more than 8 label proposals into multiple labeling tasks using WordNet hierarchy

-> 37,988 labeling tasks

Relabeling the ImageNet Validation Set

For each of the 4 labels, select whether it appears in the image, then submit your selection **.**

	binder, ring-binder	mailbag, postbag	purse	wallet, billfold, notecase, pocketbook
	○ **Yes**, 95% sure this is in the image ○ **Maybe**, it could plausibly be in the image ◉ **No**, 95% sure this is not in the image	○ **Yes**, 95% sure this is in the image ○ **Maybe**, it could plausibly be in the image ◉ **No**, 95% sure this is not in the image	○ **Yes**, 95% sure this is in the image ○ **Maybe**, it could plausibly be in the image ◉ **No**, 95% sure this is not in the image	◉ **Yes**, 95% sure this is in the image ○ **Maybe**, it could plausibly be in the image ○ **No**, 95% sure this is not in the image
Example images				

Each task is performed by 5 separate human annotators, using a crowdsourcing platform

Relabeling the ImageNet Validation Set

```
Dawid-Skene Algorithm

1. Using the labels given by human annotators, estimate the most likely "correct" label for each image
2. Based on the estimated correct answer for each image, compute the error rates for each worker
3. Considering the error rates for each worker, recompute the most likely "correct" label for each object
4. Go to step 2
```

Combine the 5 human assessments

Use classic method Dawid-Skene algorithm (EM algorithm)

57,553 labels for 46,837 images

discard 3,163 images that are assigned no label

Evaluation metric

ReaL accuracy

Correct if model top-1 prediction is in label set else incorrect

Strict metric

Correct if model top predictions are in label set else incorrect

Image A label a, b, c

model top-3 predictions == label set a, b, c -> correct

model top-3 predictions != label set a, b, c -> incorrect

Re-evaluating the state of the art

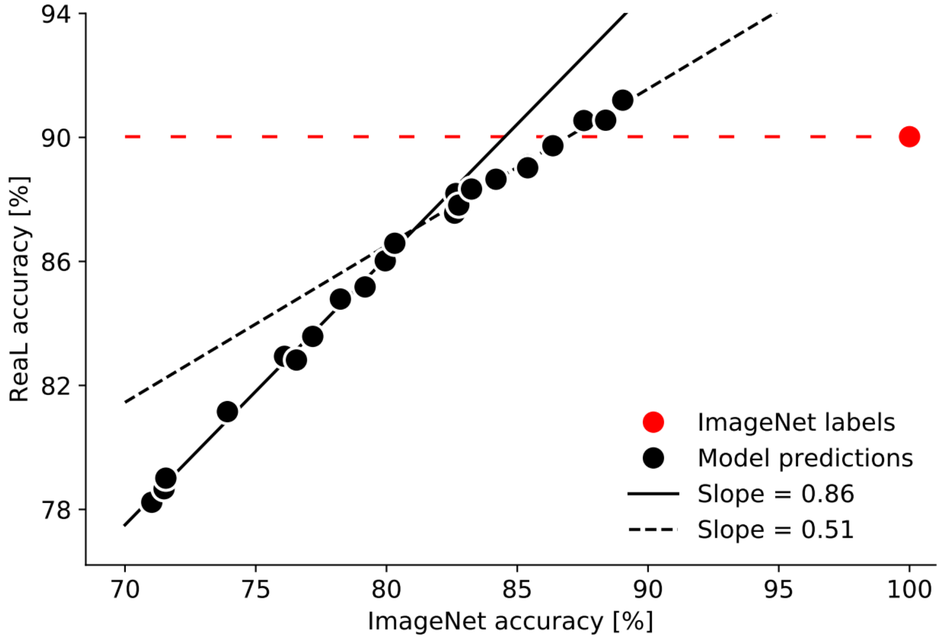


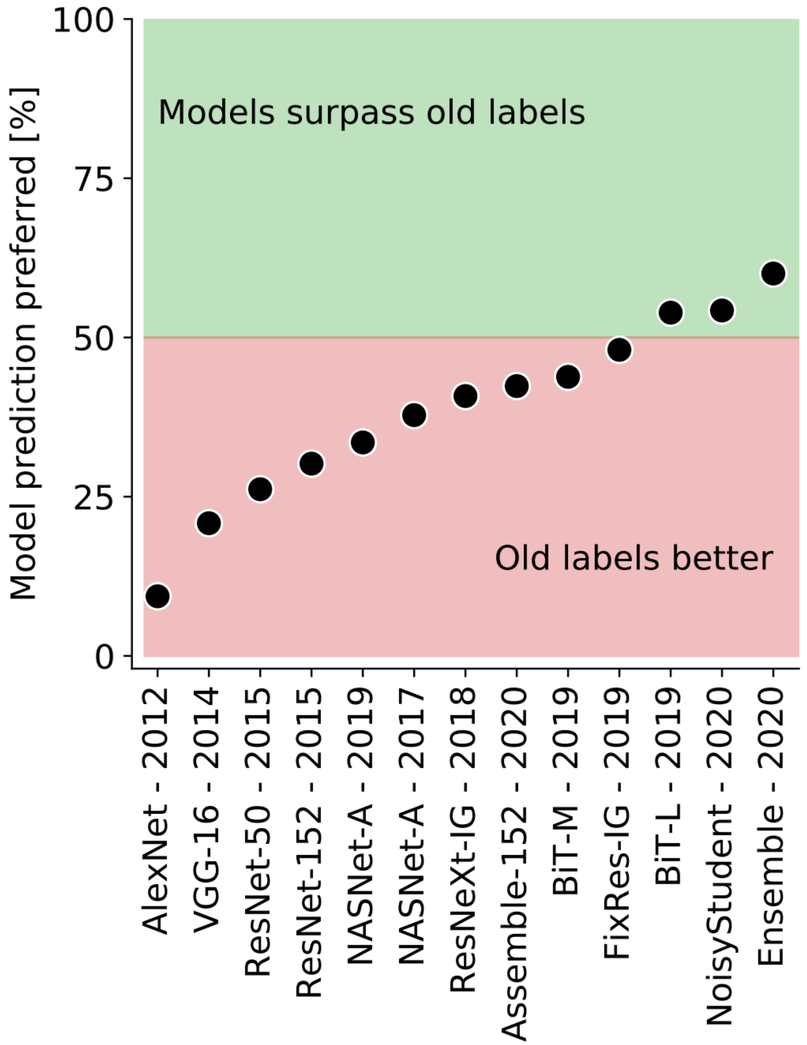
Figure 4: Comparing progress on ReaL accuracy and the original ImageNet accuracy. We measured the association between both metrics by regressing ImageNet accuracy onto ReaL accuracy for the first (solid line) and second half (dashed line) of the models in our pool.

Early models
strong linear relationship with 0.86 slope

Recent models
strong linear relationship with 0.51 slope

Some of the recent models perform better than original ImageNet label

Re-evaluating the state of the art



Images which the ImageNet label disagrees with model's prediction
(mistakes in ImageNet, correct in ReaL)

Human preference
ImageNet original label v.s. Model's prediction

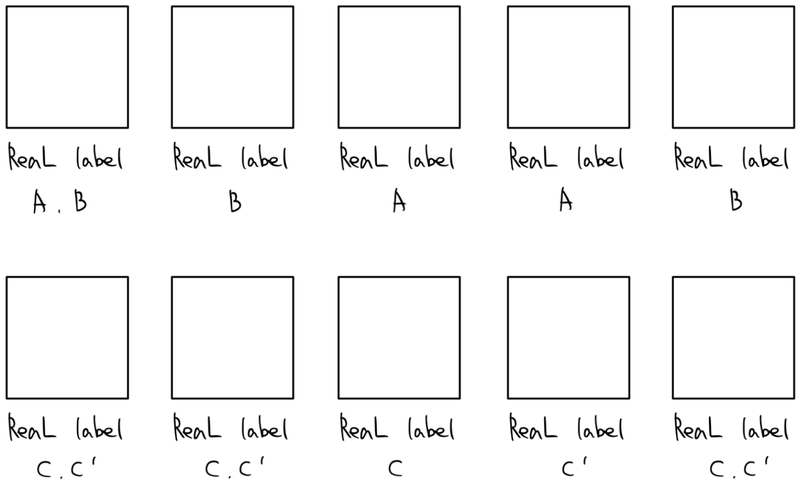
-> Recent model's prediction is more preferable to human than ImageNet label

Analysis of co-occuring classes

Approximately 29% of the images contain multiple objects or a category that corresponds to multiple synonym labels in ImageNet

Ambiguous class pair : sunglass/sunglasses, laptop/notebook, ...
Frequently appear class pair : keyboard/desk, hammer/nail, ...

Analysis of co-occuring classes



label A : $\frac{50 + 100 + 100}{3} \approx 83.3\%$

label B : $\frac{50 + 100 + 100}{3} \approx 83.3\%$

label C : $\frac{50 + 50 + 100 + 50}{4} = 62.5\%$

label C' : $\frac{50 + 50 + 100 + 50}{4} = 62.5\%$

1. 이미지의 ReaL labels에 ImageNet label이 있는 모든 케이스를 모음
2. `Unbiased oracle`이 90%보다 낮은 accuracy를 보이는 class만 선택
3. Fine-grained animal class의 경우 labeling noise가 있을 수 있기에 제외

Do models predict correct labels at random or not?

Select 253 classes using above method

Unbiased oracle model : predicts one of the ReaL labels uniformly at random

Analysis of co-occuring classes

Table 1: ImageNet classes that often co-occur with other classes. The table presents class-level accuracies for different models. Current top-performing models outperform "Ideal" model that picks a correct label for each image at random. The last column shows top co-occurring classes, the number in brackets indicates percentage of how often a certain label co-occurs.

ImageNet class	**"Oracle" model**	**Noisy Student**	**Assemble R152**	**VGG16**	**Top co-occuring classes**
desktop comp.	29.9%	**71.1%**	71.1%	60.0%	monitor (87%) ; keyboard (78%)
muzzle	73.8%	**100.0%**	90.0%	55.0%	german shepherd (5%) ; holster (5%)
convertible	73.0%	**97.6%**	95.2%	78.6%	car wheel (40%) ; grille (17%)
cucumber	74.2%	**93.2%**	88.6%	70.5%	zucchini (20%) ; bell pepper (9%)
swing	87.5%	**100.0%**	94.0%	72.0%	chain (12%) ; sweatshirt (2%)

Models exploit bias in ImageNet labeling procedure

Analysis of co-occuring classes

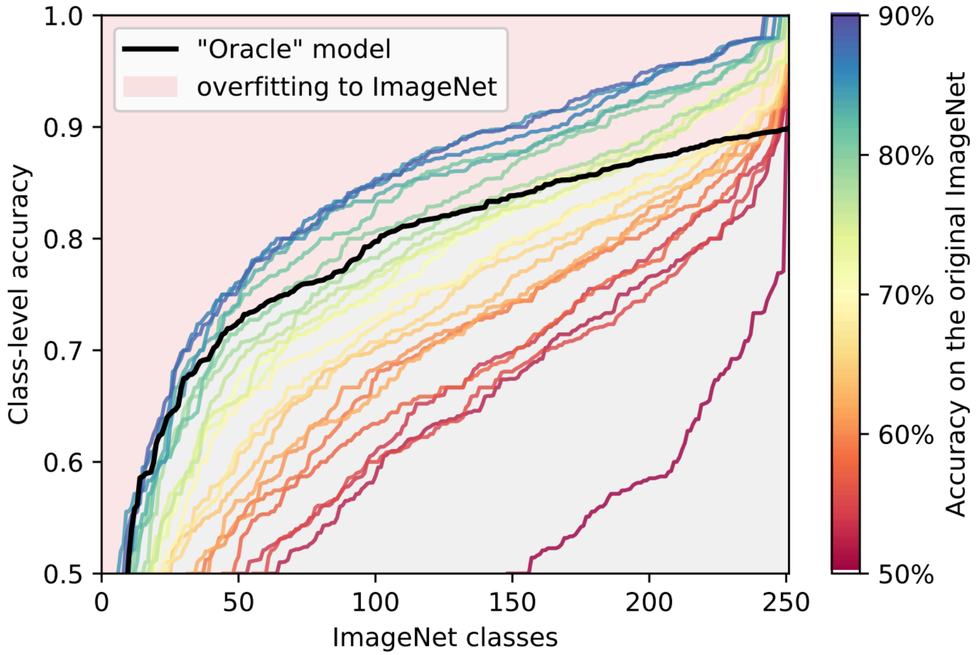


Figure 6: Each color curve corresponds to an ImageNet model and depicts sorted class-level accuracies. The black curve depicts accuracies of the "unbiased oracle". Recent top-performing models dominate the oracle curve and, thus, are overfitting to label biases present in ImageNet.

Models exploit bias in ImageNet labeling procedure

Analysis of co-occuring classes

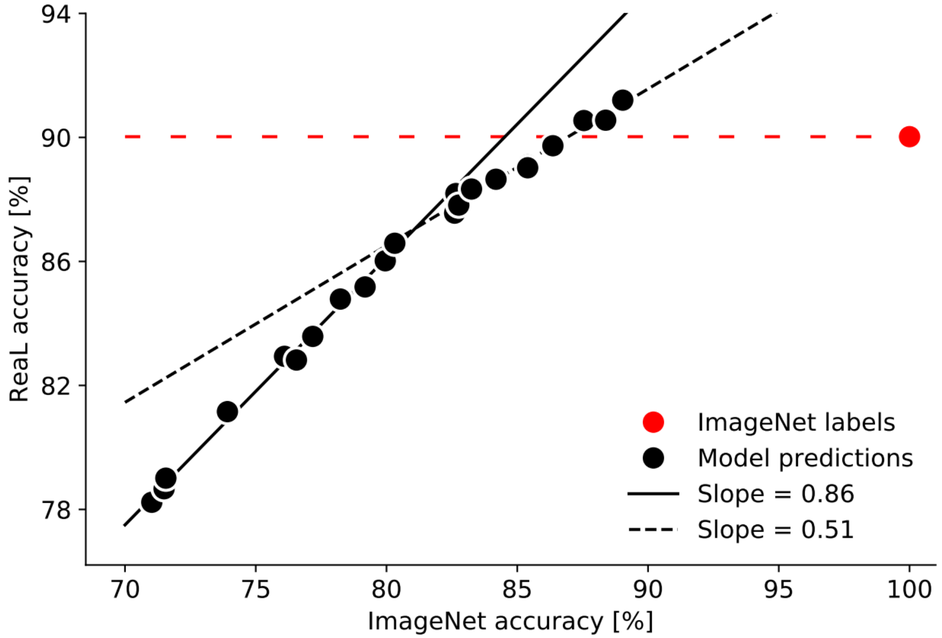


Figure 4: Comparing progress on ReaL accuracy and the original ImageNet accuracy. We measured the association between both metrics by regressing ImageNet accuracy onto ReaL accuracy for the first (solid line) and second half (dashed line) of the models in our pool.

Significant fraction of progress on ImageNet has been achieved by exploiting biases present in ImageNet labeling procedure

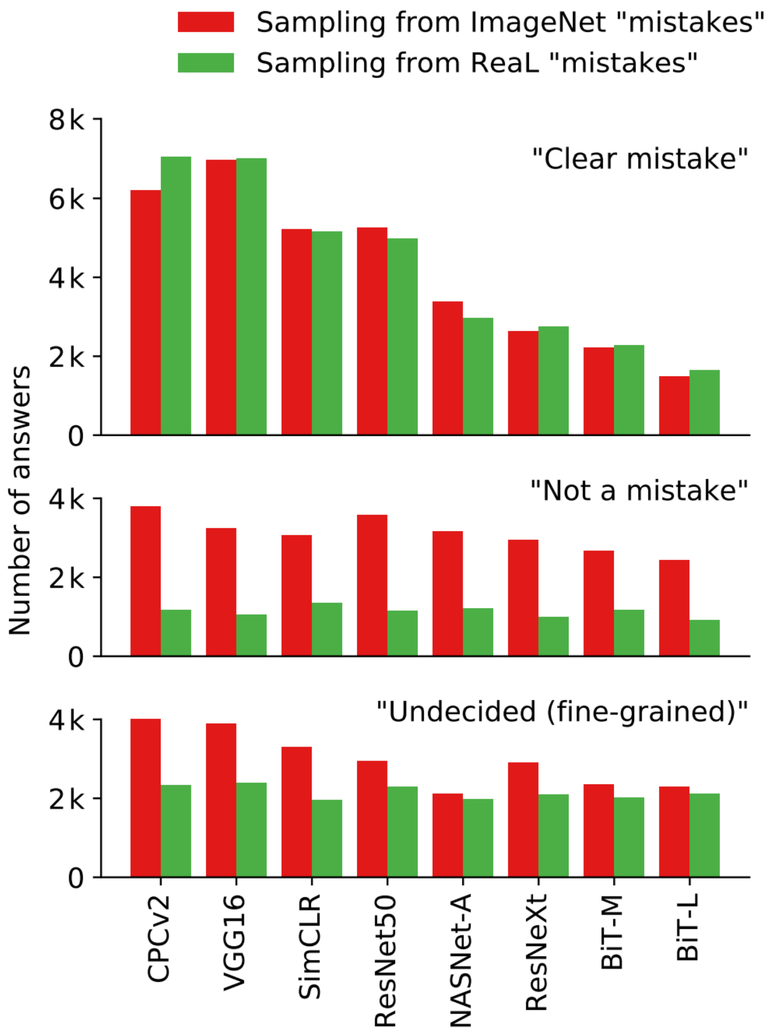
This explains why performance gain on ImageNet partially transfer to ReaL accuracy

Analyzing the remaining mistakes

Highest performing model : error rate 11% on ImageNet, 9% on ReaL

What is the nature of these mistakes?

Analyzing the remaining mistakes



For images and model predictions that are incorrect, raters identify the reason for the prediction being considered a mistake or not

ReaL labels have comparable quality to ImageNet ones for fine-grained classes
Significantly reduced the noise in the rest

-> Meaningful benchmark

Improving ImageNet training

Table 2: Top-1 accuracy (in percentage) on ImageNet with our proposed sigmoid loss and clean label set. Median accuracy from three runs is reported for all the methods. Either sigmoid loss or clean label set leads to consistent improvements over baseline. Using both achieves the best performance. The improvement of our proposed method is more pronounced with longer training schedules.

	Model	ImageNet accuracy			ReaL accuracy		
		90 epochs	270 epochs	900 epochs	90 epochs	270 epochs	900 epochs
ResNet-50	Baseline	76.0	76.9 **(+0.9)**	75.9 **(-0.1)**	82.5	82.9 **(+0.4)**	81.6 **(-0.9)**
	+ Sigmoid	76.3 **(+0.3)**	77.8 **(+1.8)**	76.9 **(+0.9)**	83.0 **(+0.5)**	83.9 **(+1.4)**	82.7 **(+0.2)**
	+ Clean	76.4 **(+0.4)**	77.8 **(+1.8)**	77.4 **(+1.4)**	82.8 **(+0.3)**	83.7 **(+1.2)**	83.3 **(+0.8)**
	+ Both	76.6 **(+0.6)**	78.2 **(+2.2)**	**78.5** **(+2.5)**	83.1 **(+0.6)**	**84.3** **(+1.8)**	84.1 **(+1.6)**
ResNet-152	Baseline	78.0	78.3 **(+0.3)**	77.1 **(-0.9)**	84.1	83.8 **(-0.3)**	82.3 **(-1.8)**
	+ Sigmoid	78.5 **(+0.5)**	78.7 **(+0.7)**	77.4 **(-0.6)**	84.6 **(+0.5)**	84.3 **(+0.2)**	82.7 **(-1.4)**
	+ Clean	78.6 **(+0.6)**	79.6 **(+1.6)**	79.0 **(+1.0)**	84.4 **(+0.3)**	85.0 **(+0.9)**	84.4 **(+0.3)**
	+ Both	78.7 **(+0.7)**	**79.8** **(+1.8)**	79.3 **(+1.3)**	84.6 **(+0.5)**	**85.2** **(+1.1)**	84.5 **(+0.4)**

How to deal with label-noise in ImageNet train dataset?

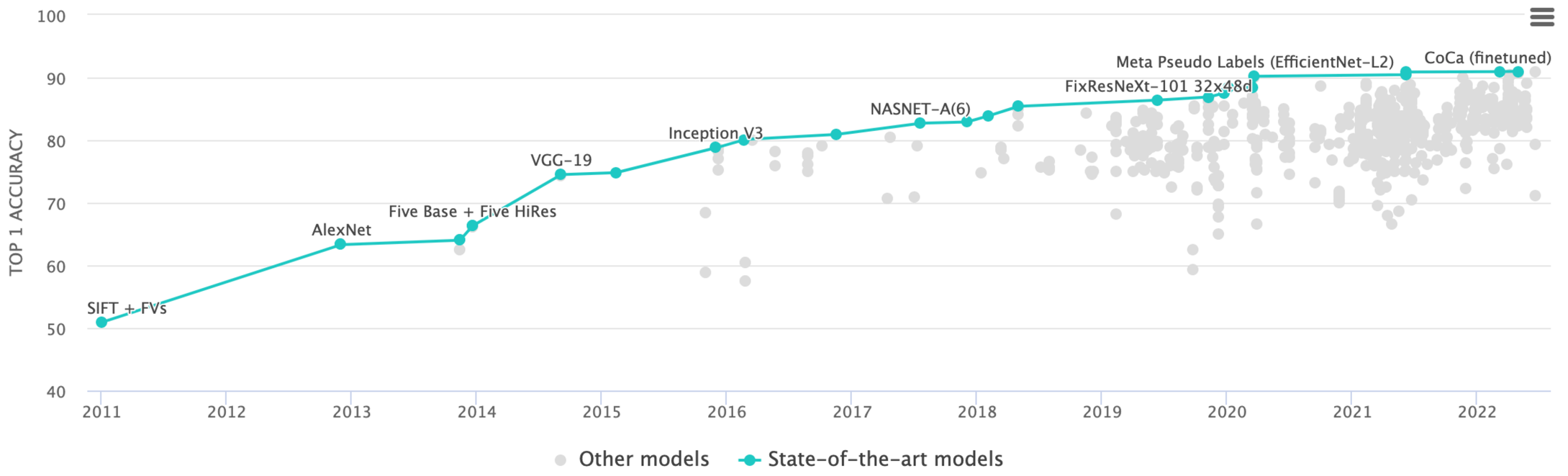
Use sigmoid cross-entropy loss instead of softmax cross-entropy loss

Label cleansing using good model

Image Classification on ImageNet

Leaderboard Dataset

View Top 1 Accuracy by Date for All models



Rank	Model	Top 1 Accuracy	Top 5 Accuracy	Number of params	GFLOPs	Extra Training Data	Paper	Code	Result	Year	Tags
1	**CoCa** (finetuned)	91.0%		2100M		✓	CoCa: Contrastive Captioners are Image-Text Foundation Models			2022	Transformer, JFT-3B, ALIGN
2	**Model soups** (BASIC-L)	90.98%		2440M		✓	Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time			2022	Conv+Transformer, JFT-3B, ALIGN
9	**Meta Pseudo Labels** (EfficientNet-L2)	90.2%	98.8%	480M		✓	Meta Pseudo Labels			2020	EfficientNet, JFT-300M
10	**DaViT-H**	90.2%		362M		✓	DaViT: Dual Attention Vision Transformers			2022	Transformer
11	**Florence-CoSwin-H**	90.05%	99.02%	893M		✓	Florence: A New Foundation Model for Computer Vision			2021	Transformer, FLD-900M
12	**Meta Pseudo Labels** (EfficientNet-B6-Wide)	90%	98.7%	390M		✓	Meta Pseudo Labels			2020	EfficientNet, JFT-300M