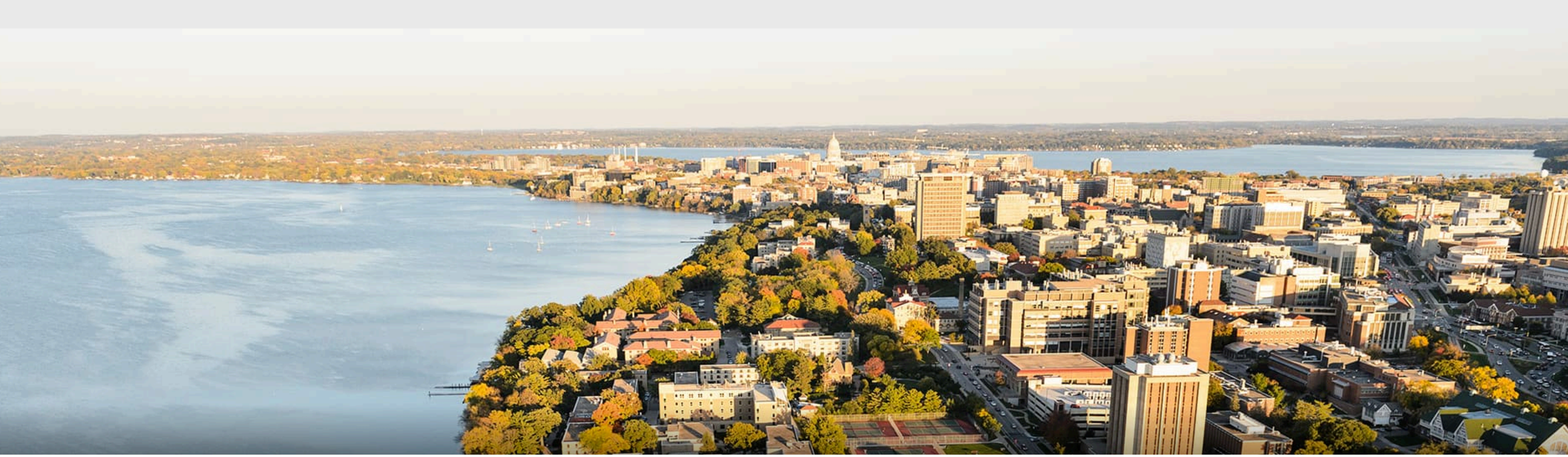


CS839 Special Topics in AI: Deep Learning
Learning with Less Supervision

Sharon Yixuan Li
University of Wisconsin-Madison

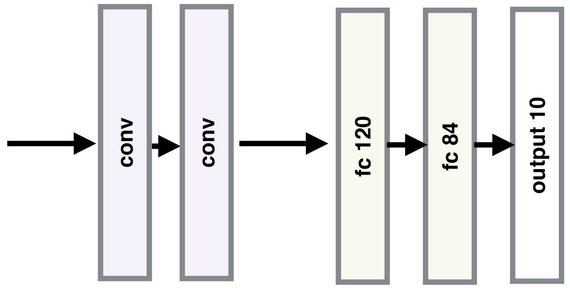
October 29, 2020

Overview

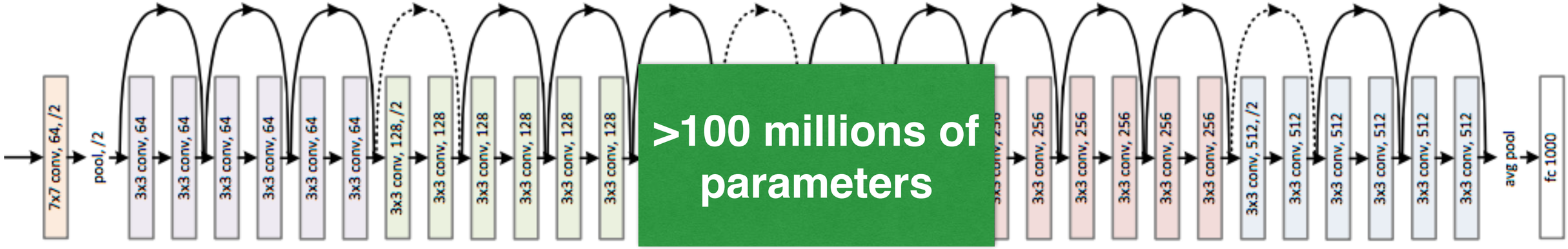
- **Weakly Supervised Learning**
  - Flickr100M
  - JFT300M (Google)
  - Instagram3B (Facebook)
- **Data augmentation**
  - Human heuristics
  - Automated data augmentation
- **Self-supervised Learning**
  - Pretext tasks (rotation, patches, colorization etc.)
  - Invariant vs. Covariant learning
  - Contrastive learning based framework (current SoTA)

Part I: Weakly Supervised Learning

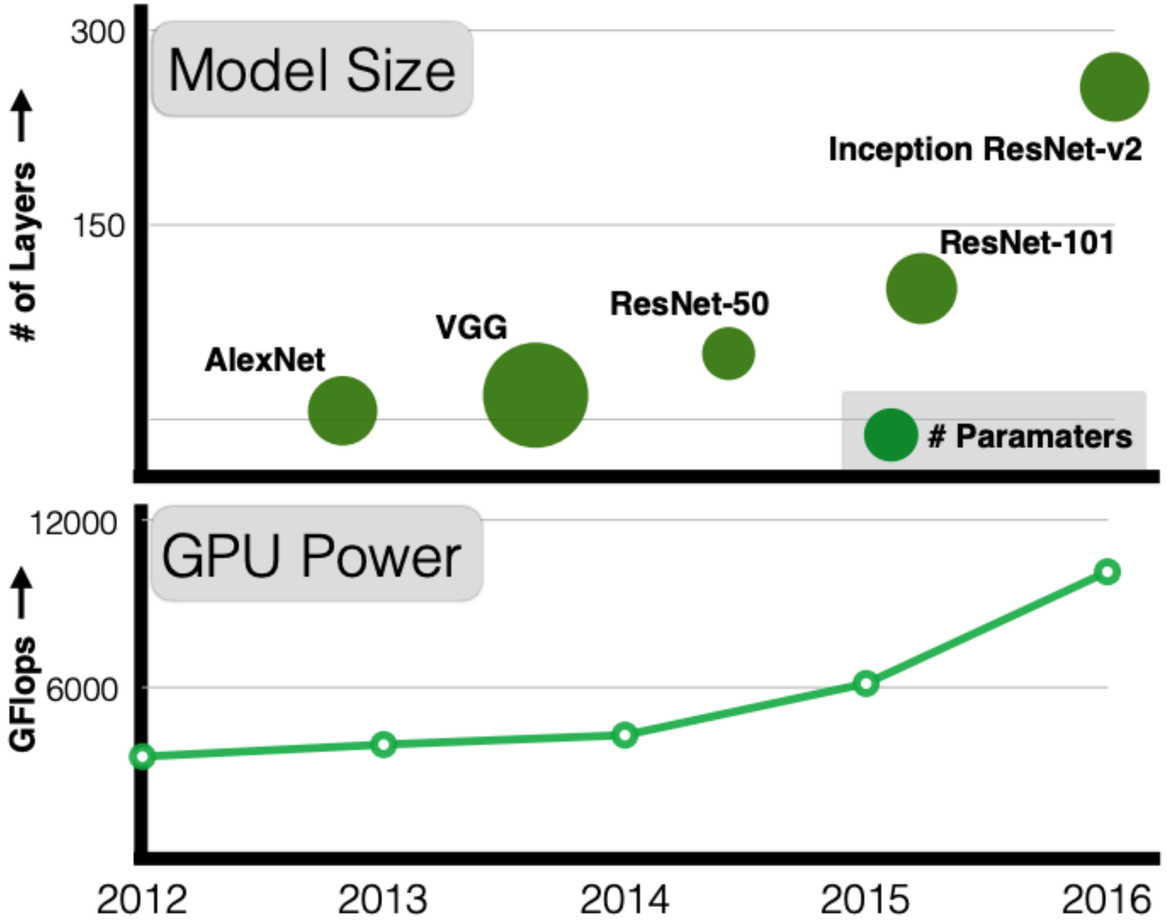
Model Complexity Keeps Increasing



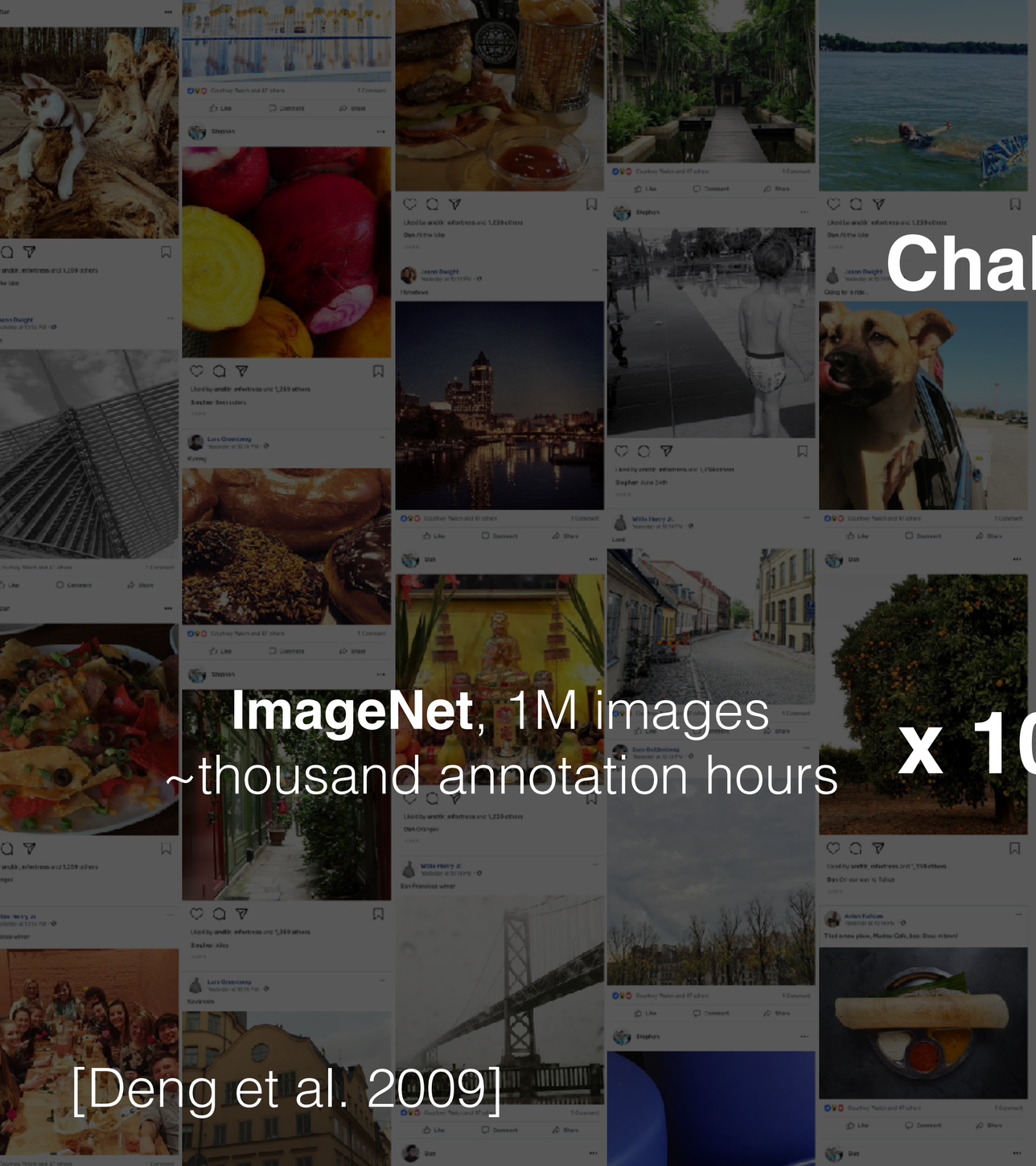
LeNet (Lecun et al. 1998)



ResNet (He et al. 2016)



[Sun et al. 2017]



Challenge: Limited labeled data

ImageNet, 1M images
~thousand annotation hours

x 1000

1B images
~million annotation hours

[Deng et al. 2009]

TRAINING AT SCALE

Levels of Supervision

TRAINING AT SCALE

Noisy Data

Flickr 100M [Joulin et al. 2015]

Learning Visual Features from Large Weakly Supervised Data

Armand Joulin* ajoulin@fb.com
Laurens van der Maaten* lvdmaaten@fb.com
Allan Jabri ajabri@fb.com
Nicolas Vasilache ntv@fb.com

Facebook AI Research
770 Broadway, New York NY 10003

Abstract

Convolutional networks trained on large supervised dataset produce visual features which form the basis for the state-of-the-art in many computer-vision problems. Further improvements of these visual features will likely require even larger manually labeled data sets, which severely limits the pace at which progress can be made. In this paper, we explore the potential of leveraging massive, weakly-labeled image collections for learning good visual features. We train convolutional networks on a dataset of 100 million Flickr photos and captions, and show that these networks produce features that perform well in a range of vision problems. We also show that the networks appropriately capture word similarity, and learn correspondences between different languages.

the veranda hotel portixol palma

plane approaching zrh avro regional jet rj

not as impressive as embankment that s for sure

student housing by lungaard tranberg architects in copenhagen

article in the local paper about all the unusual things found

this was another one with my old digital camera i like the way it looks for some things though slow and lower resolution than new cameras another problem is that it s a bit of a brick to carry and is a pain unless you re carrying a bag with some room it s nearly x x

JFT 300M [Sun et al. 2017]

Revisiting Unreasonable Effectiveness of Data in Deep Learning Era

Chen Sun[1], Abhinav Shrivastava[1,2], Saurabh Singh[1], and Abhinav Gupta[1,2]

[1]Google Research
[2]Carnegie Mellon University

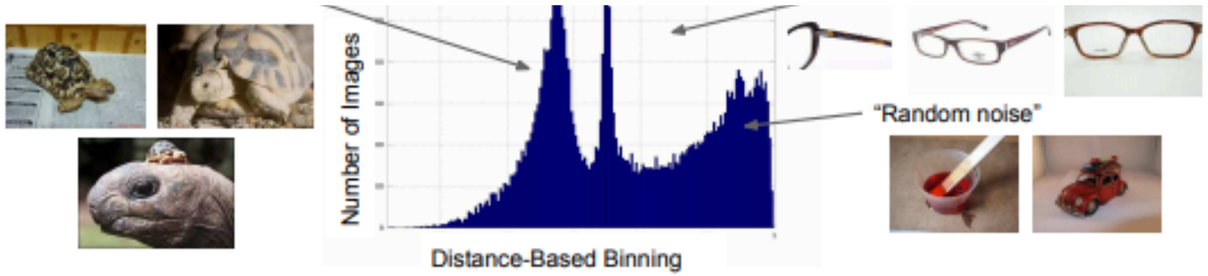


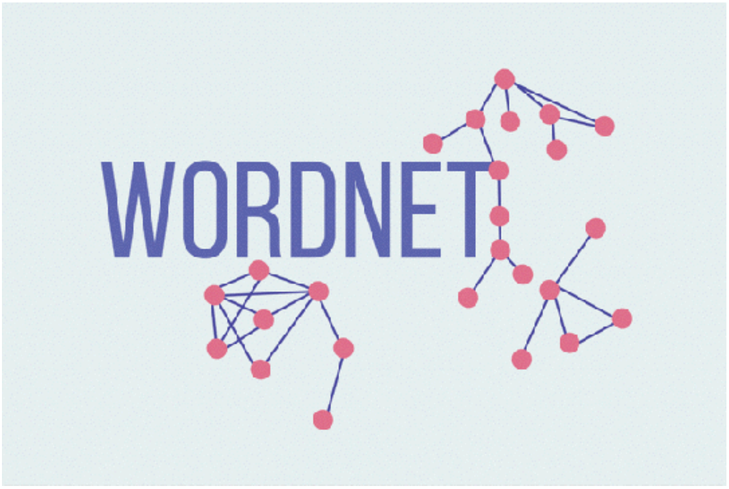
Figure 2. JFT-300M dataset can be noisy in terms of label confusion and incorrect labels. This is because labels are generated via a complex mixture of web signals, and not annotated or cleaned by humans. x-axis corresponds to the quantized distances to K-Means centroids, which are computed based on visual features.

Can we use **billions** of images with hashtags for pre-training?

[Mahajan et al. 2018]

Hashtags Selection

1.5K, 1B
synonyms of ImageNet labels

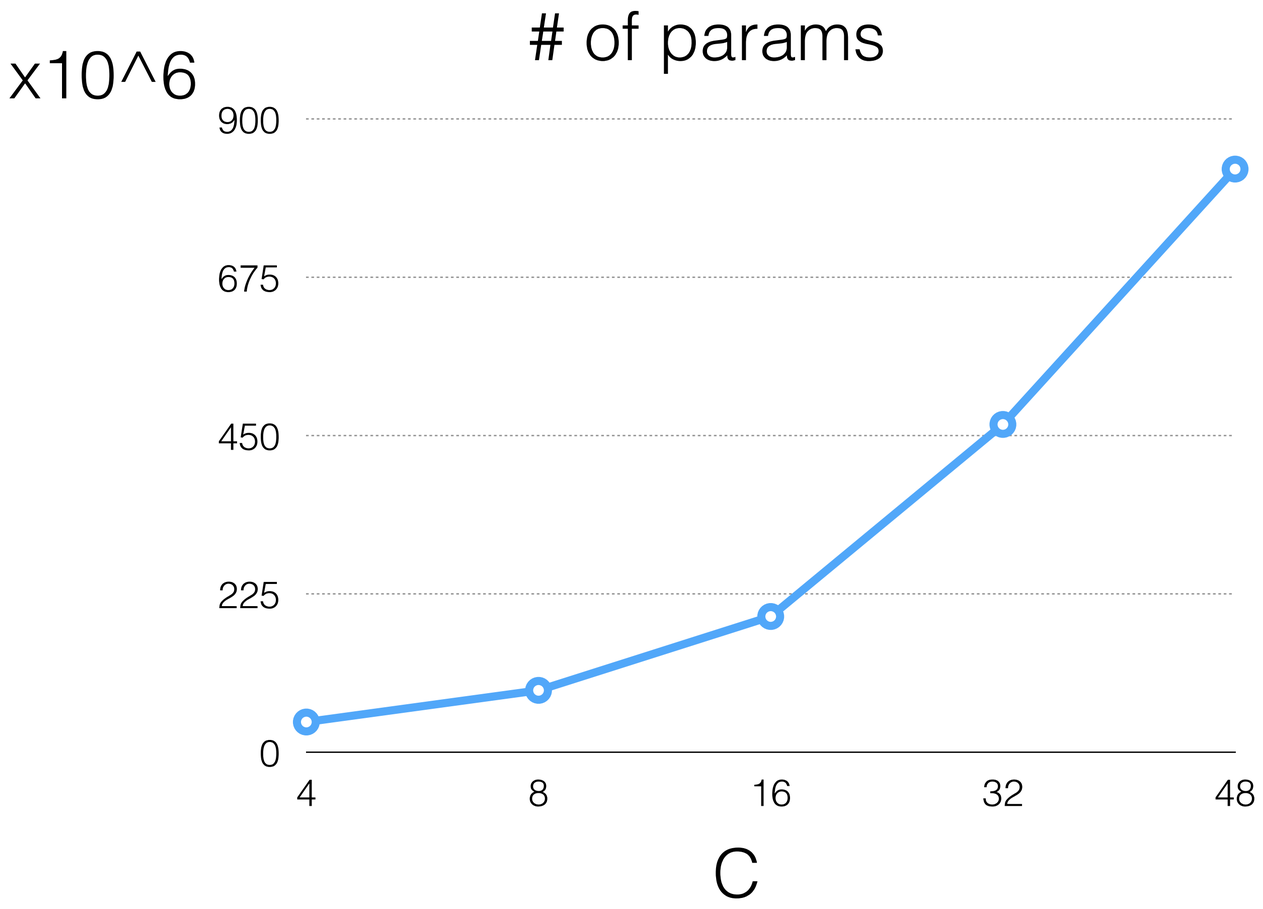


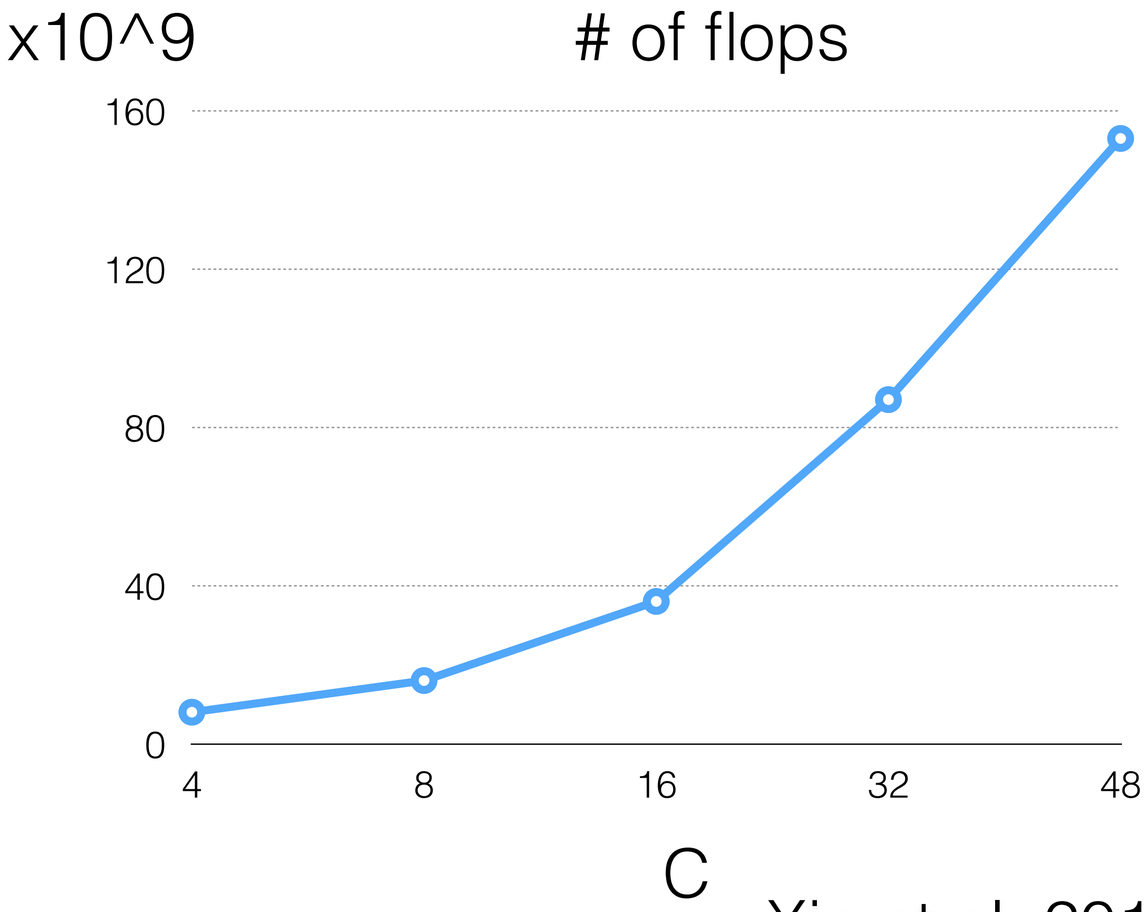
17K, 3B
synonyms of nouns in wordnet

[Mahajan et al. 2018]

Network Architecture and Capacity

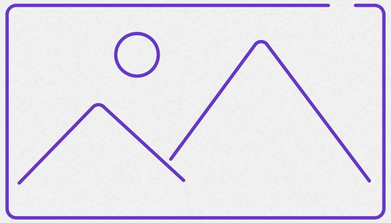
ResNeXt-101 32x**C**d



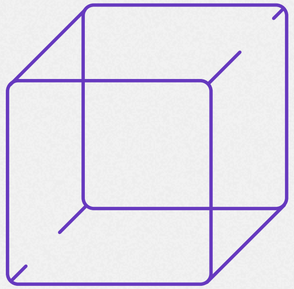


Xie et al. 2016

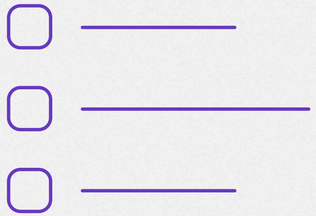
Largest Weakly Supervised Training



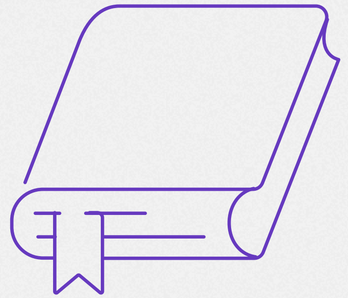
3.5B
PUBLIC INSTAGRAM
IMAGES



LARGE CAPACITY MODEL
(RESNEXT101-32X48)



17K UNIQUE LABELS

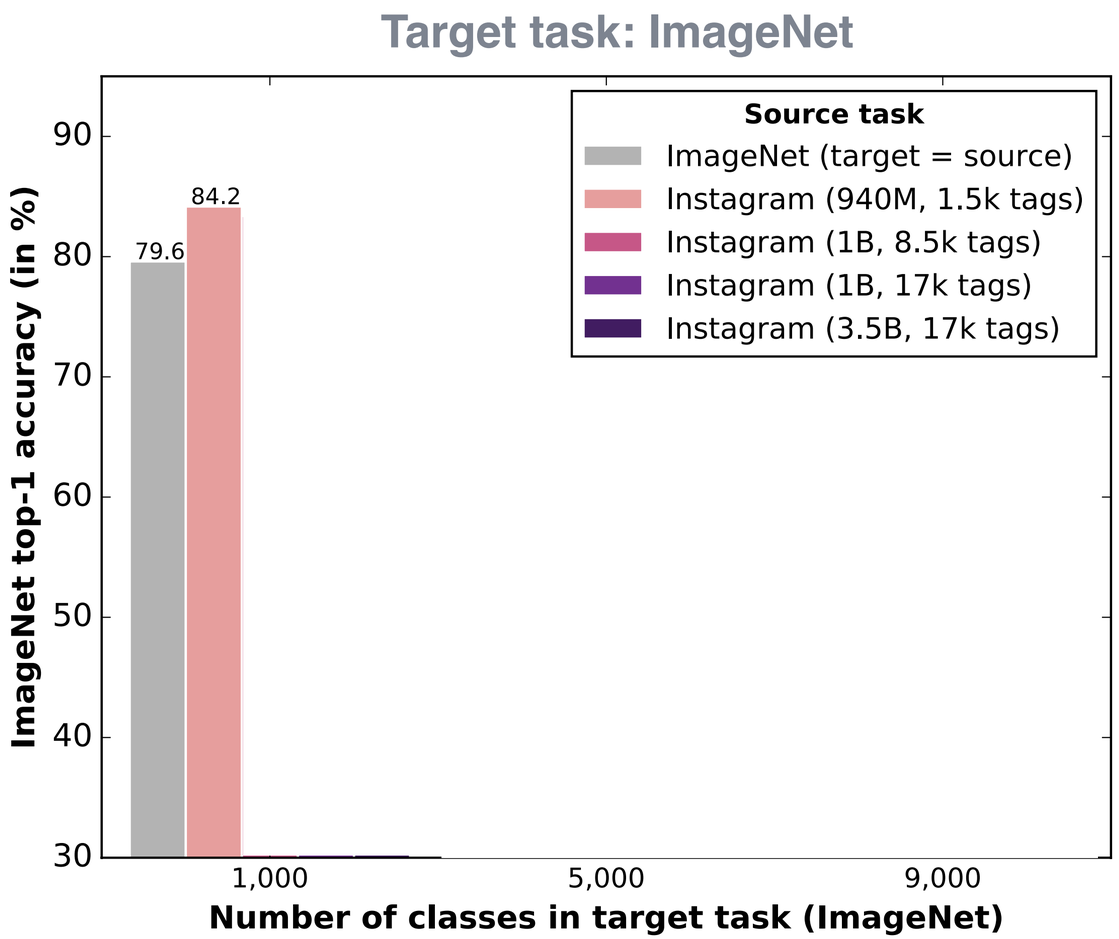


DISTRIBUTED
TRAINING
(350 GPUS)

[Mahajan et al. 2018]

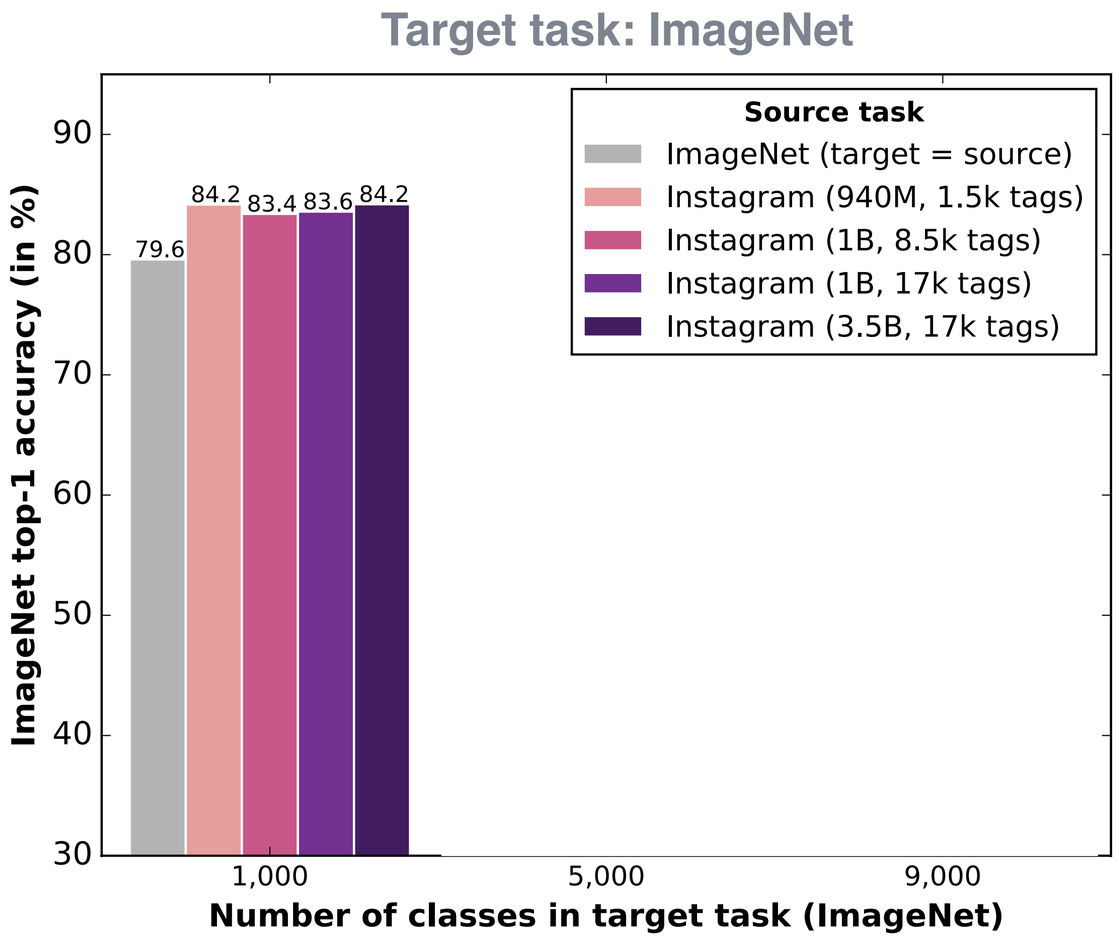
Results

Transfer Learning Performance



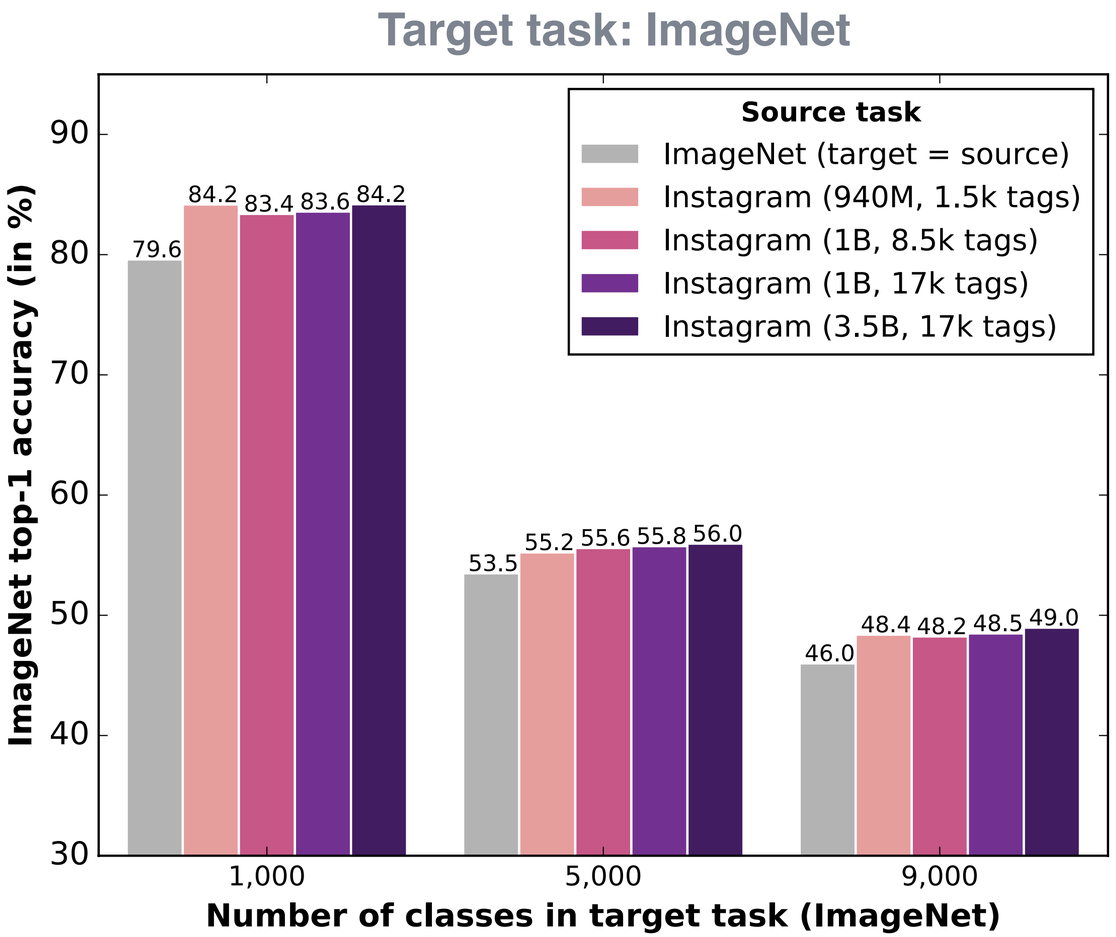
* With a bigger model, we even got 85.4% top-1 error on ImageNet-1K.

Transfer Learning Performance



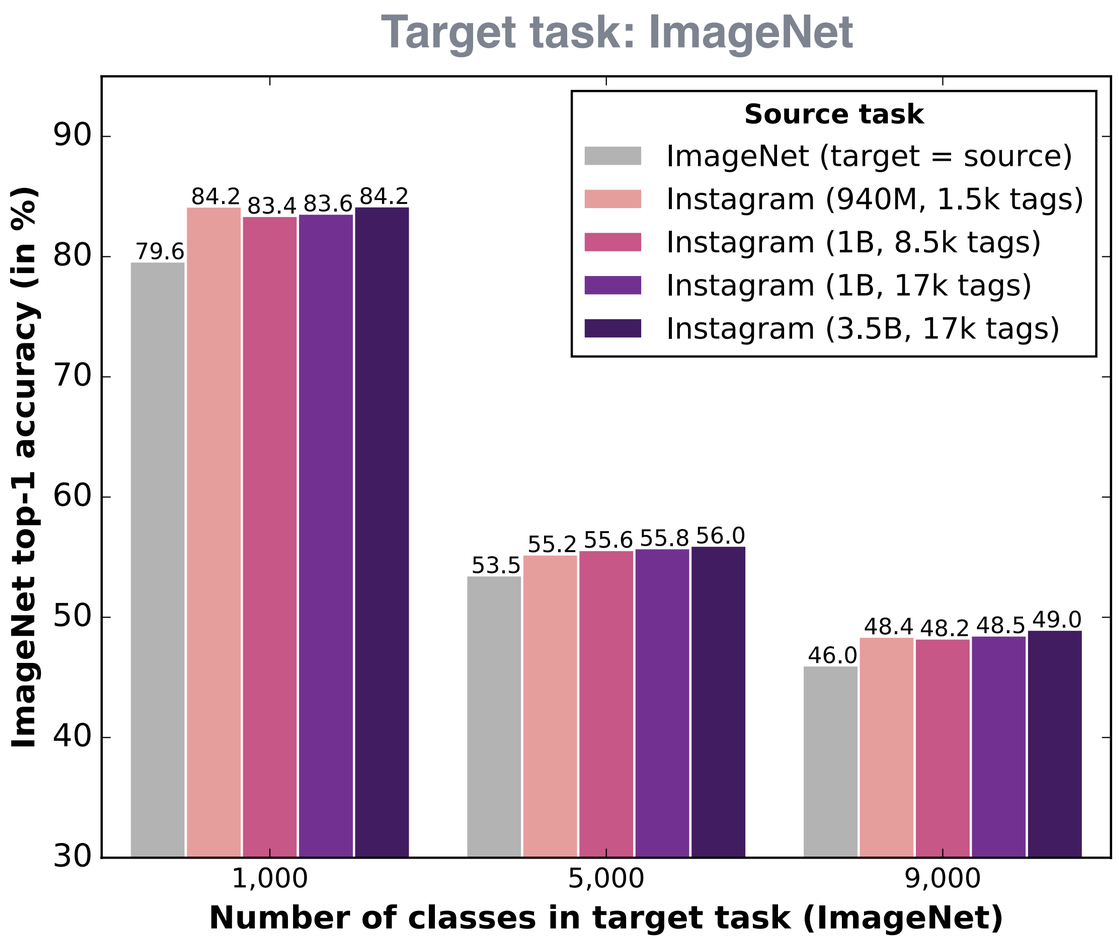
* With a bigger model, we even got 85.4% top-1 error on ImageNet-1K.

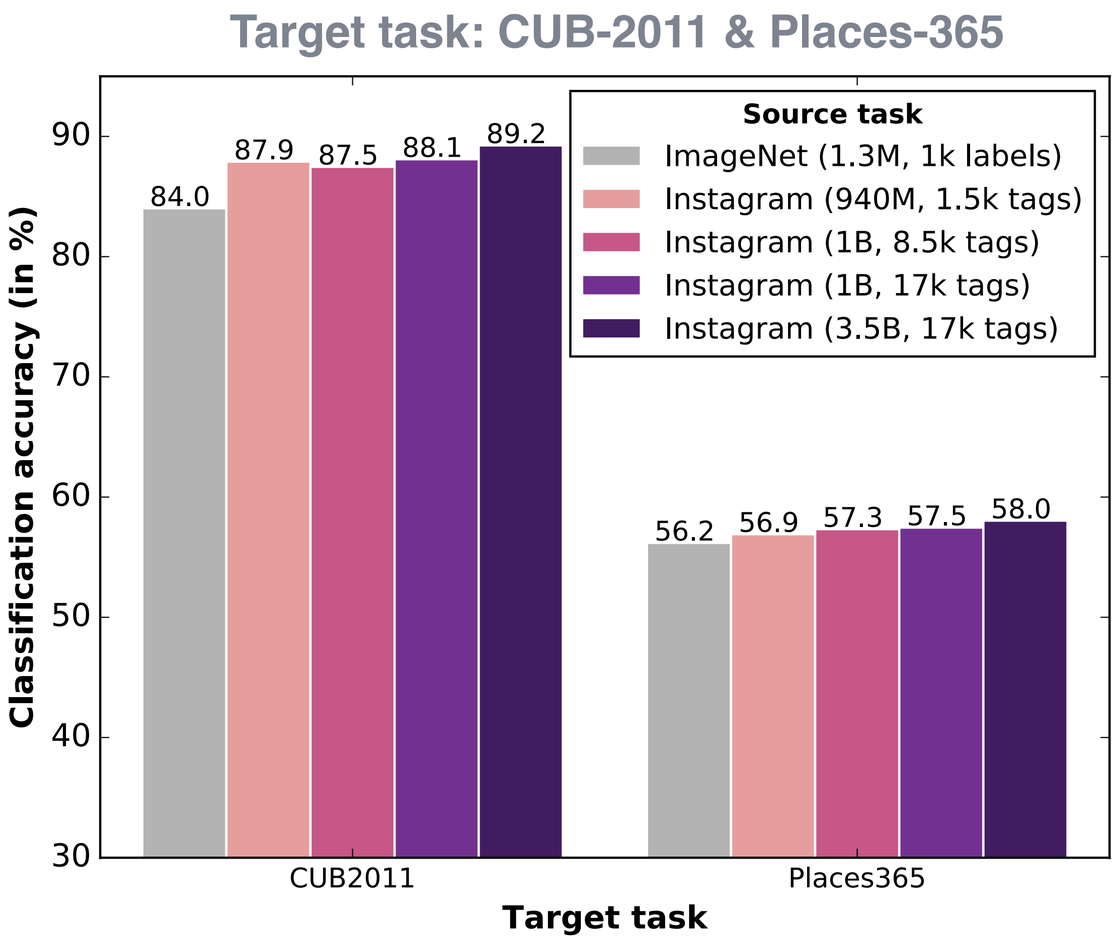
Transfer Learning Performance



* With a bigger model, we even got 85.4% top-1 error on ImageNet-1K.

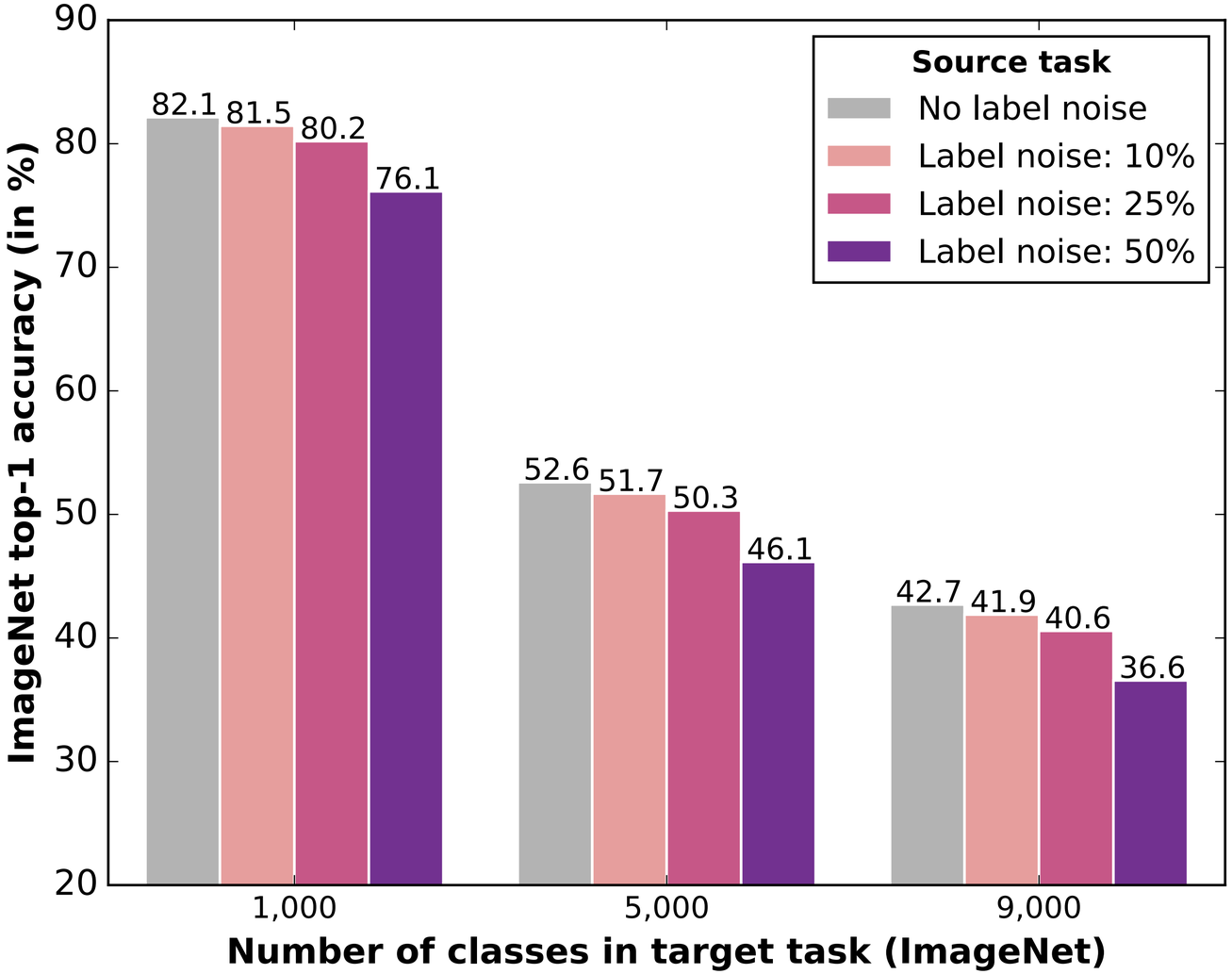
Transfer Learning Performance





* With a bigger model, we even got 85.4% top-1 error on ImageNet-1K.

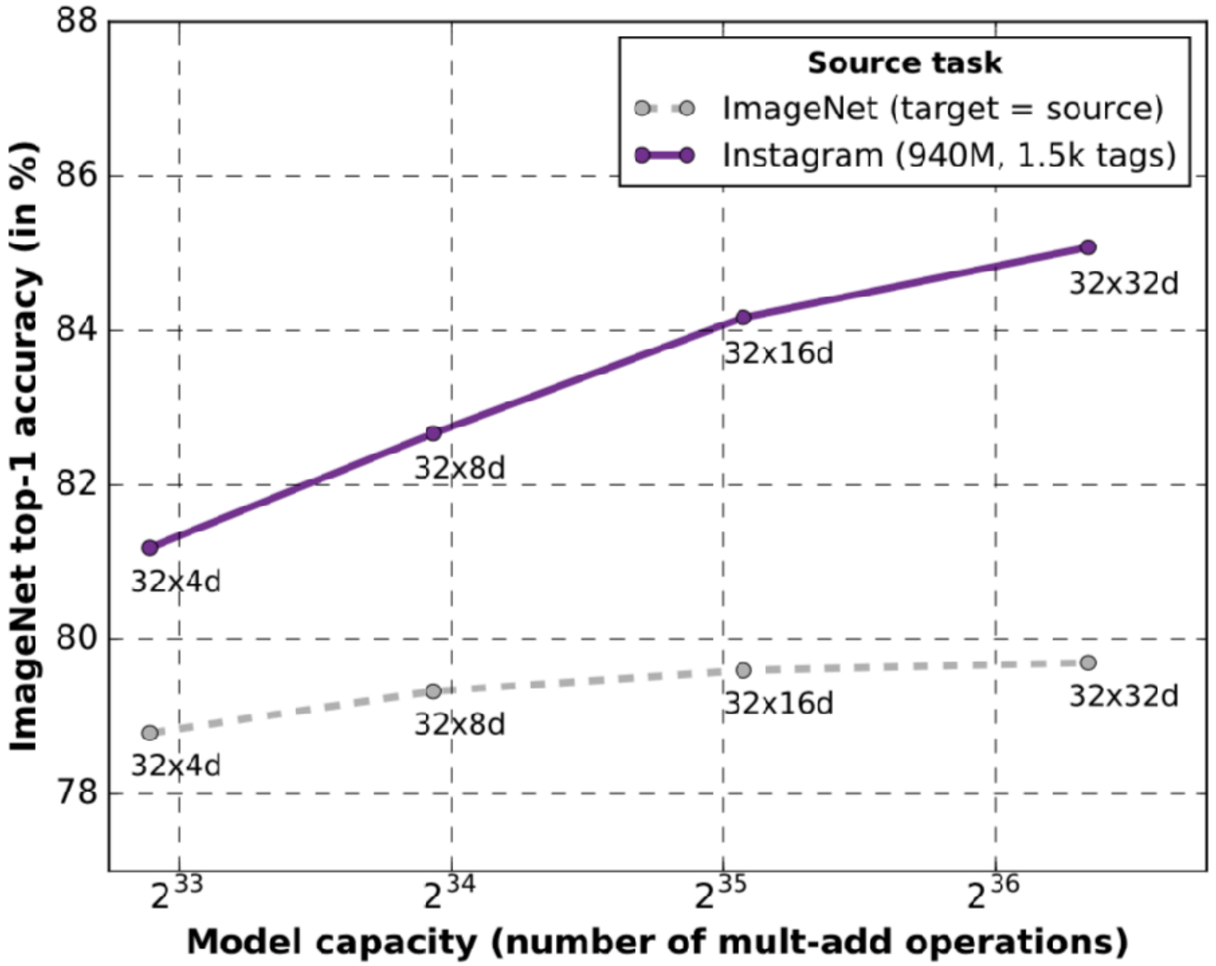
Models are surprisingly robust to label "noise"



Dataset: IG-1B-17k

Network: ResNext-101 32x16

Effect of Model Capacity



Matching hashtags to target task helps (1.5K tags)

Target task: ImageNet-1K

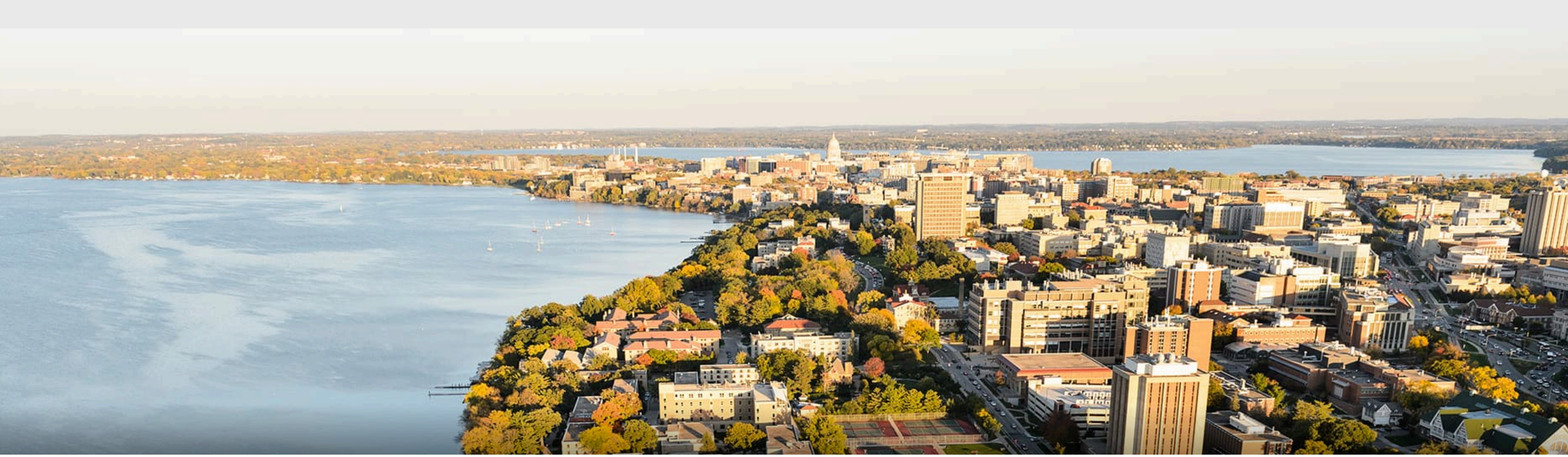
BiT Transfer [Kolesnikov et al. 2020]

Big Transfer (BiT): General Visual Representation Learning

Alexander Kolesnikov*, Lucas Beyer*, Xiaohua Zhai*,
Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby

Google Research, Brain Team
Zürich, Switzerland
{akolesnikov,lbeyer,xzhai}@google.com
{jpuigcerver,jessicayung,sylvaingelly,neilhoulsby}@google.com

Abstract. Transfer of pre-trained representations improves sample efficiency and simplifies hyperparameter tuning when training deep neural networks for vision. We revisit the paradigm of pre-training on large supervised datasets and fine-tuning the model on a target task. We scale up pre-training, and propose a simple recipe that we call Big Transfer (BiT). By combining a few carefully selected components, and transferring using a simple heuristic, we achieve strong performance on over 20 datasets. BiT performs well across a surprisingly wide range of data regimes — from 1 example per class to 1 M total examples. BiT achieves 87.5% top-1 accuracy on ILSVRC-2012, 99.4% on CIFAR-10, and 76.3% on the 19 task Visual Task Adaptation Benchmark (VTAB). On small datasets, BiT attains 76.8% on ILSVRC-2012 with 10 examples per class, and 97.0% on CIFAR-10 with 10 examples per class. We conduct detailed analysis of the main components that lead to high transfer performance.



Part II: Data Augmentation

Data Augmentation

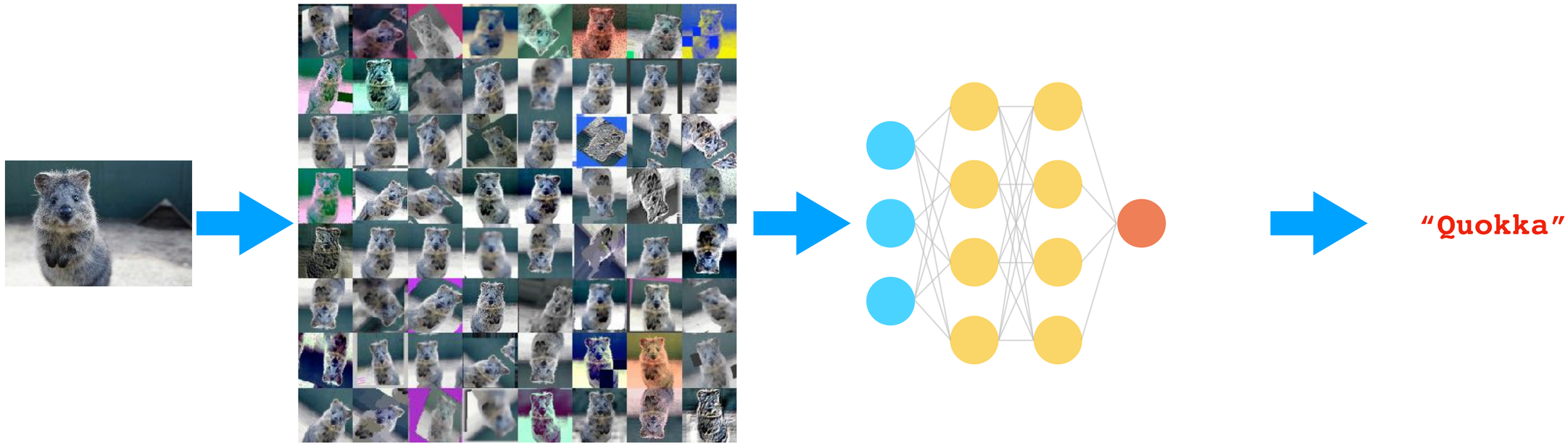
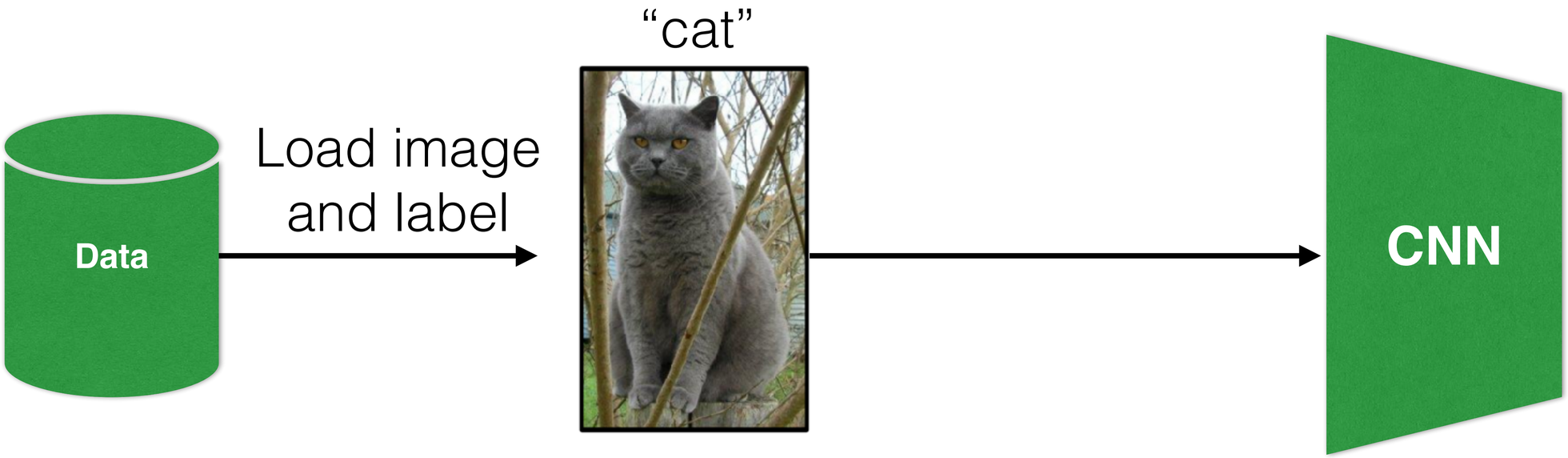
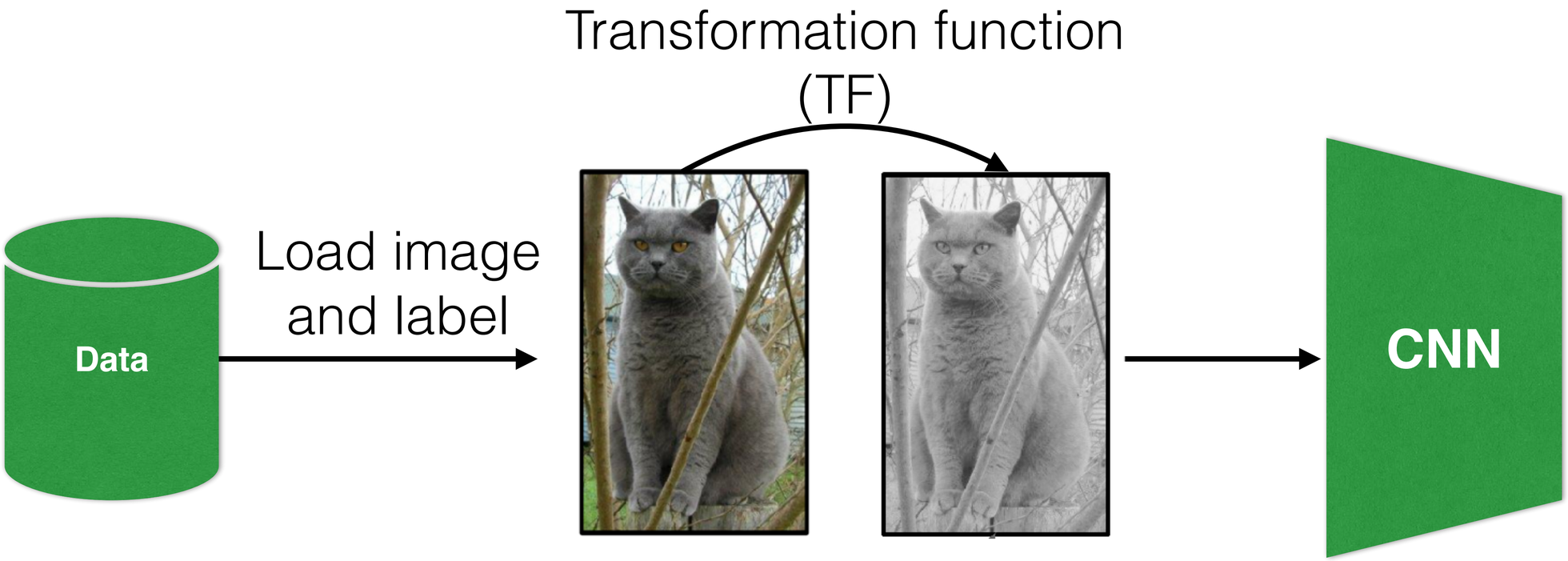


Figure credit: https://github.com/aleju/imgaug

Data Augmentation



Data Augmentation

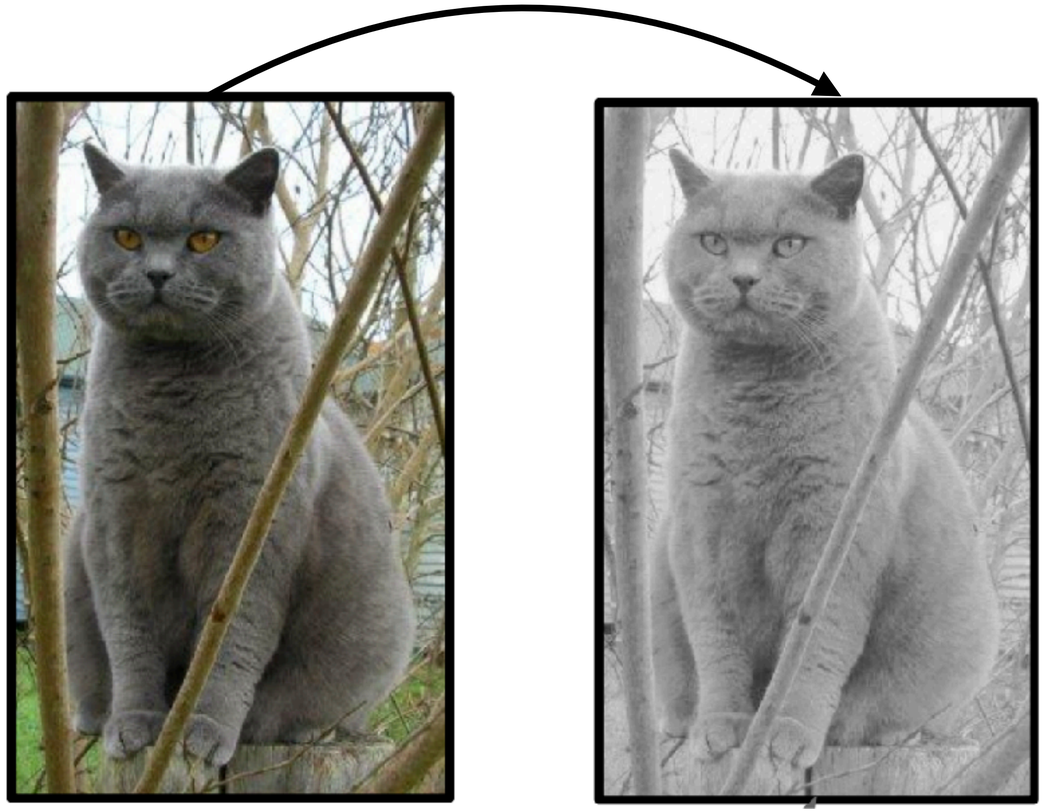


Data Augmentation

- Change the pixels without changing the labels

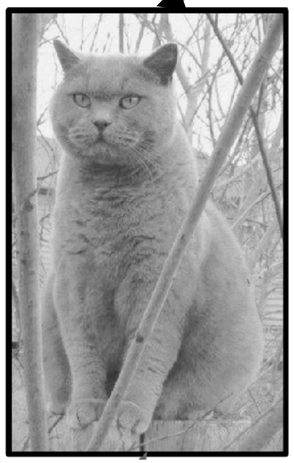
- Train on transformed data improves generalization

- VERY widely used

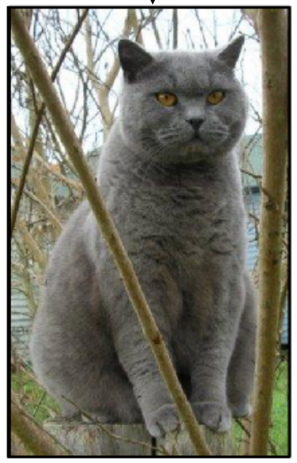


Example of Transformation Functions (TFs)

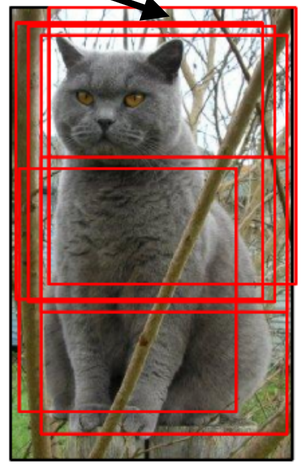
Original image



Color jitter



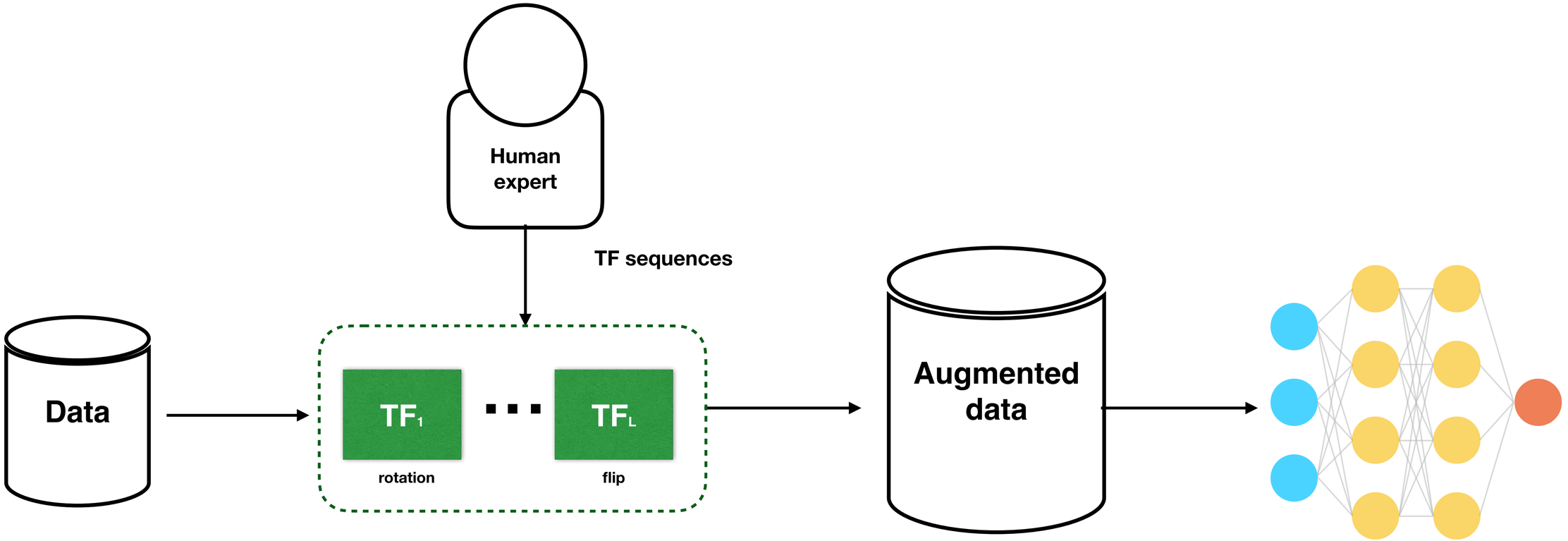
Horizontal flip



Random crop

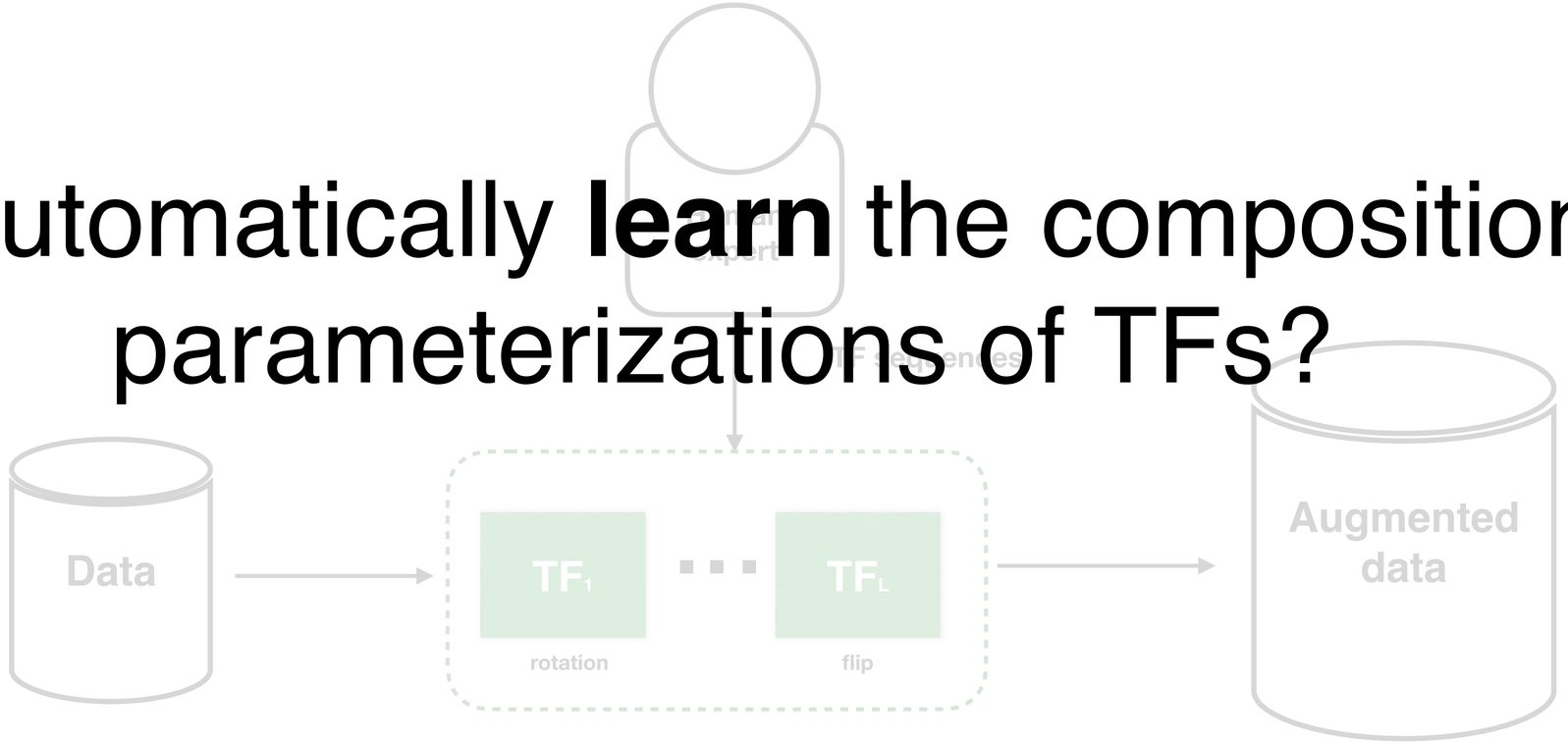
...

Heuristic Data Augmentation



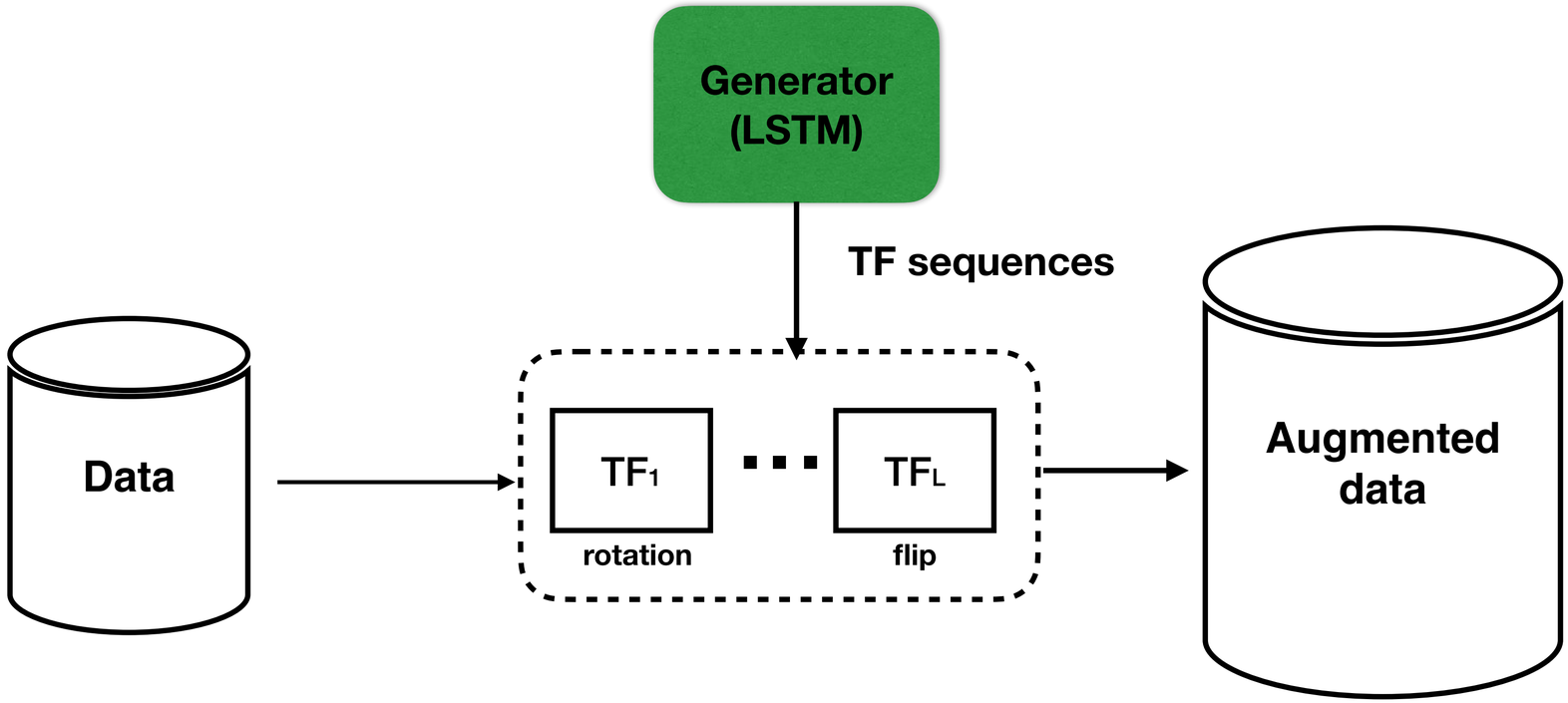
Heuristic Data Augmentation

How to automatically **learn** the compositions and parameterizations of TFs?



TANDA

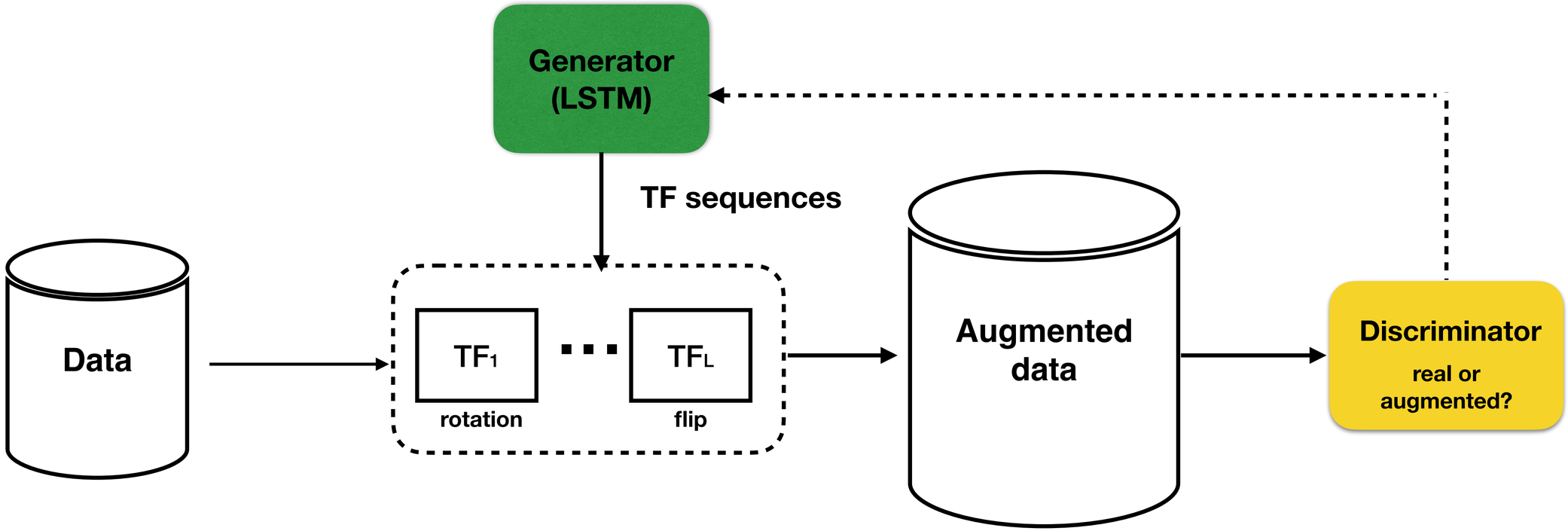
Transformation **A**dversarial **N**etworks for **D**ata **A**ugmentations



[Ratner et al. 2017]

TANDA

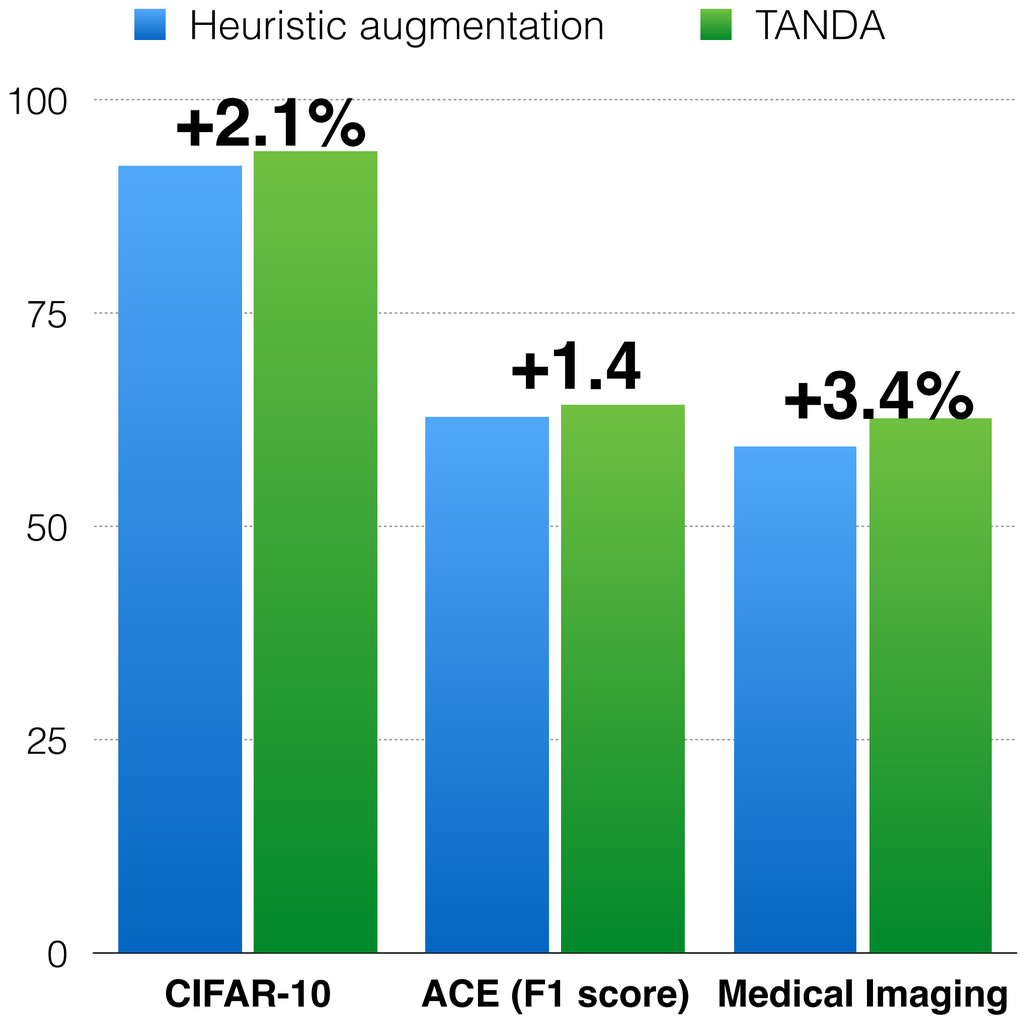
Transformation **A**dversarial **N**etworks for **D**ata **A**ugmentations



[Ratner et al. 2017]

TANDA

Transformation **A**dversarial **N**etworks for **D**ata **A**ugmentations



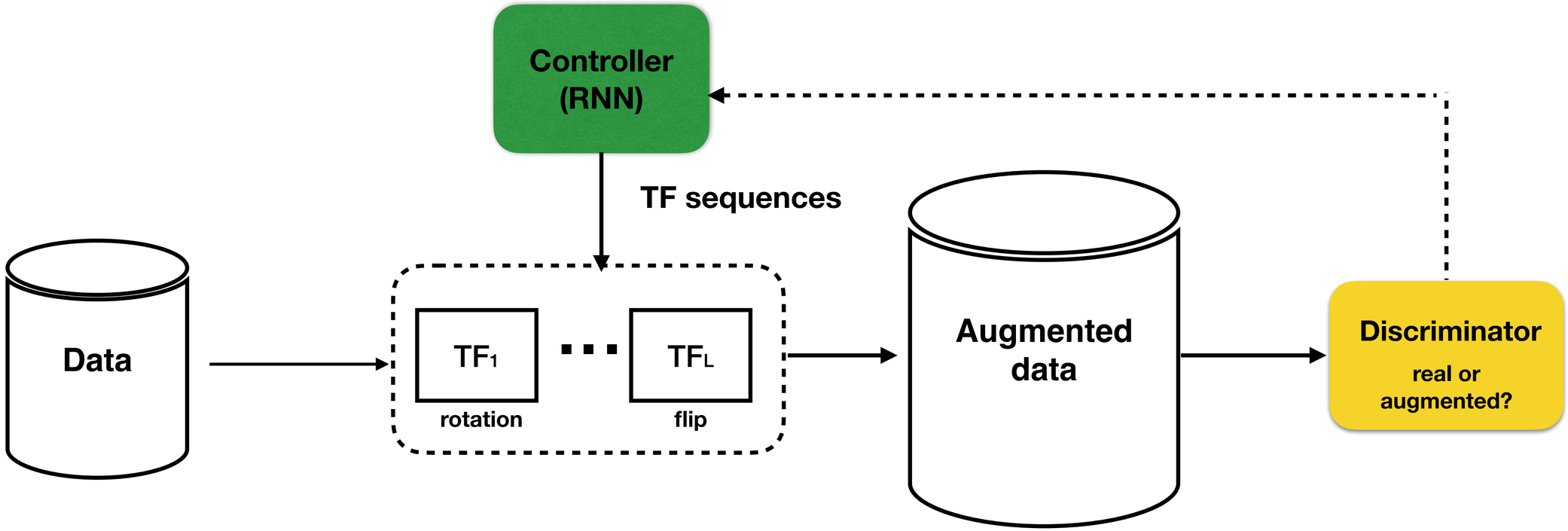
Generated MNIST samples

[Ratner et al. 2017]

AutoAugment

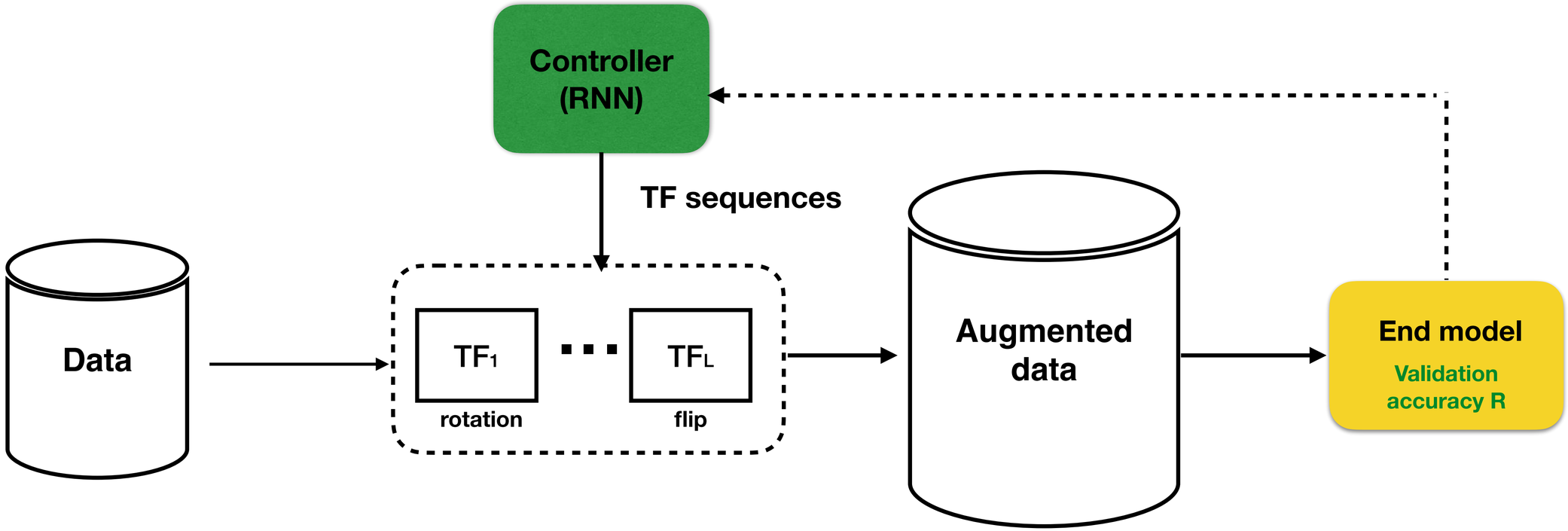
[Cubuk et al. 2018]

AutoAugment



[Cubuk et al. 2018]

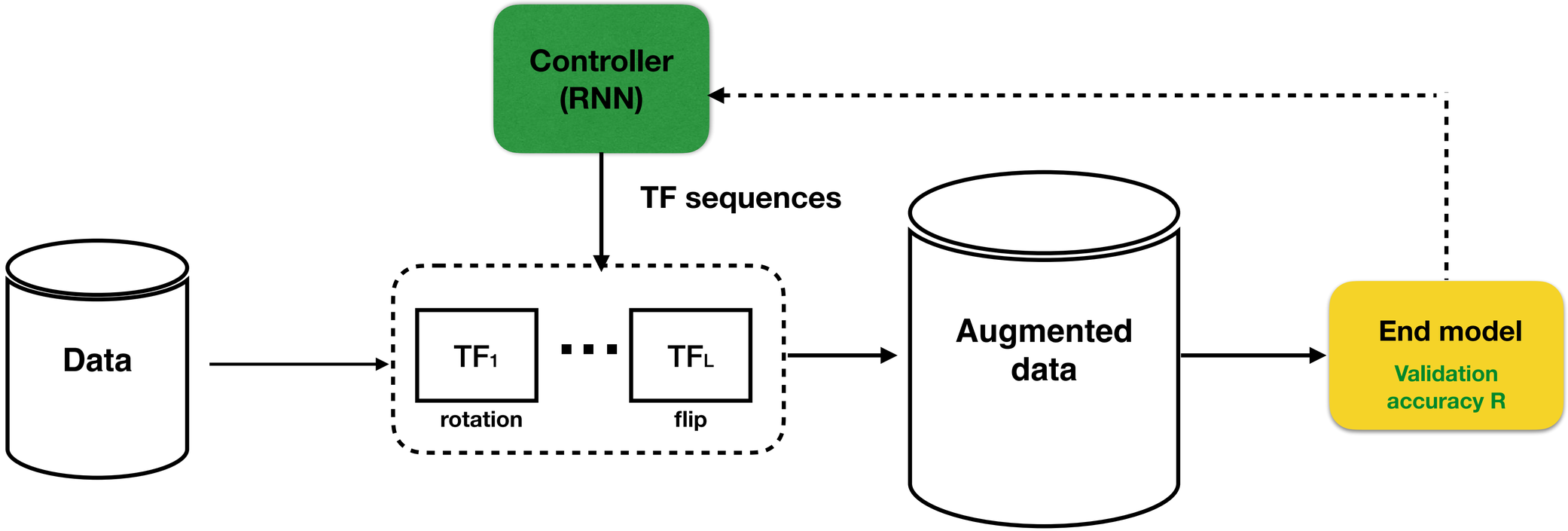
AutoAugment



State-of-the-art performance on various benchmarks, however the computational cost is very high.

[Cubuk et al. 2018]

RandAugment



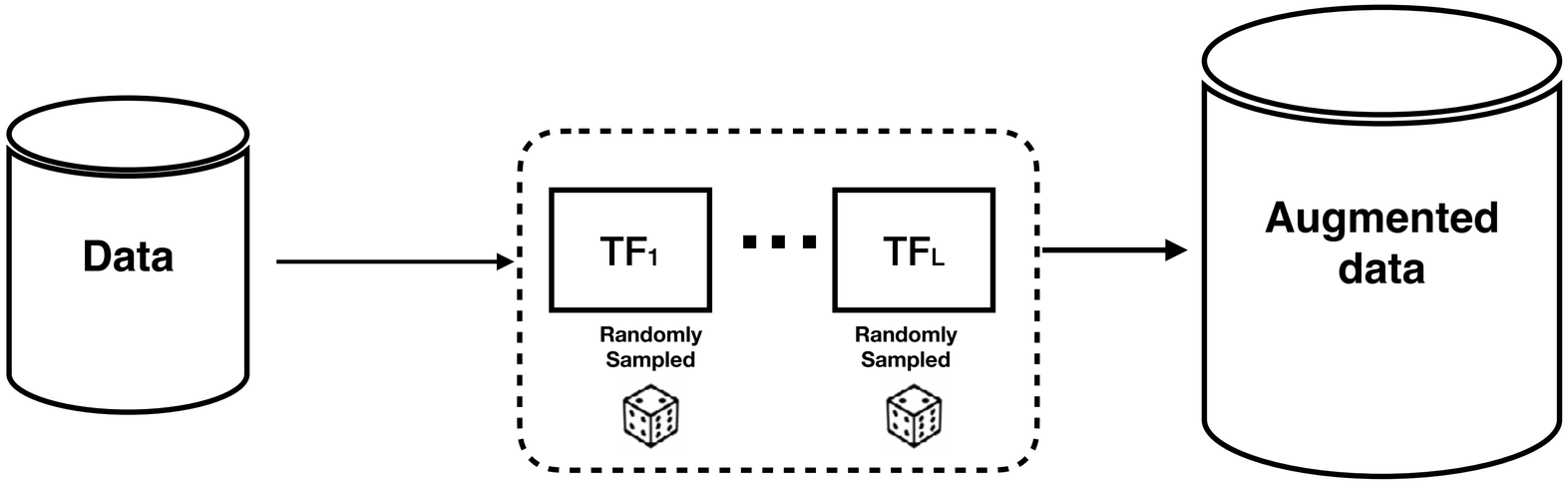
[Cubuk et al. 2019]

RandAugment

(1) random sampling over the transformation functions

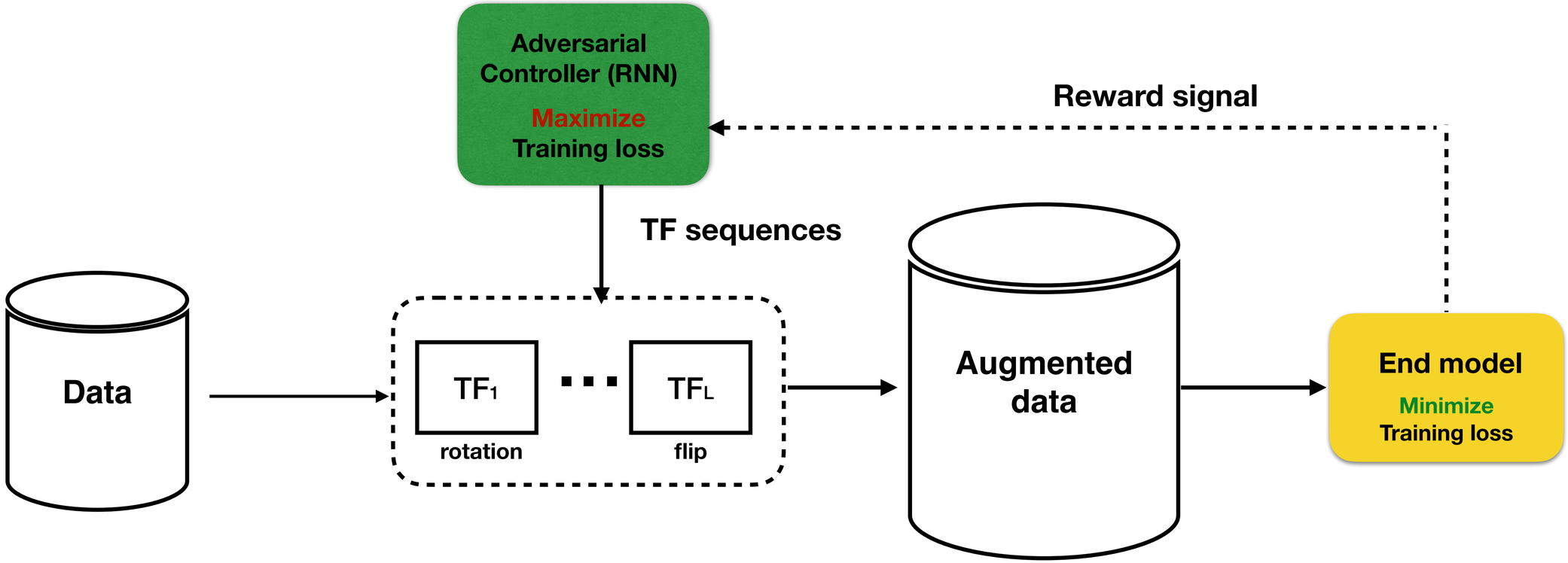
(2) grid search over the parameters of each transformation

Outperform AutoAugment



[Cubuk et al. 2019]

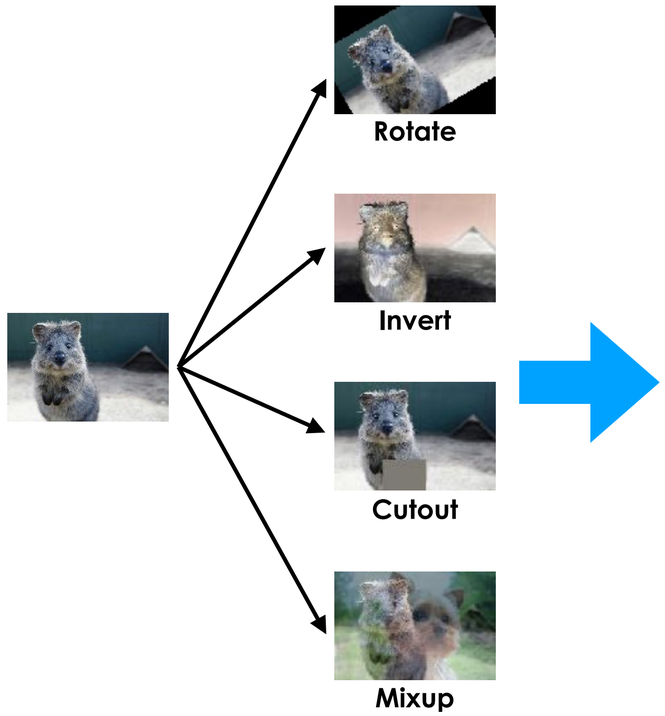
Adversarial AutoAugment



12x reduction in computing cost on ImageNet, compared to AutoAugment.
Top-1 error **1.36%** on CIFAR-10 (new sota).

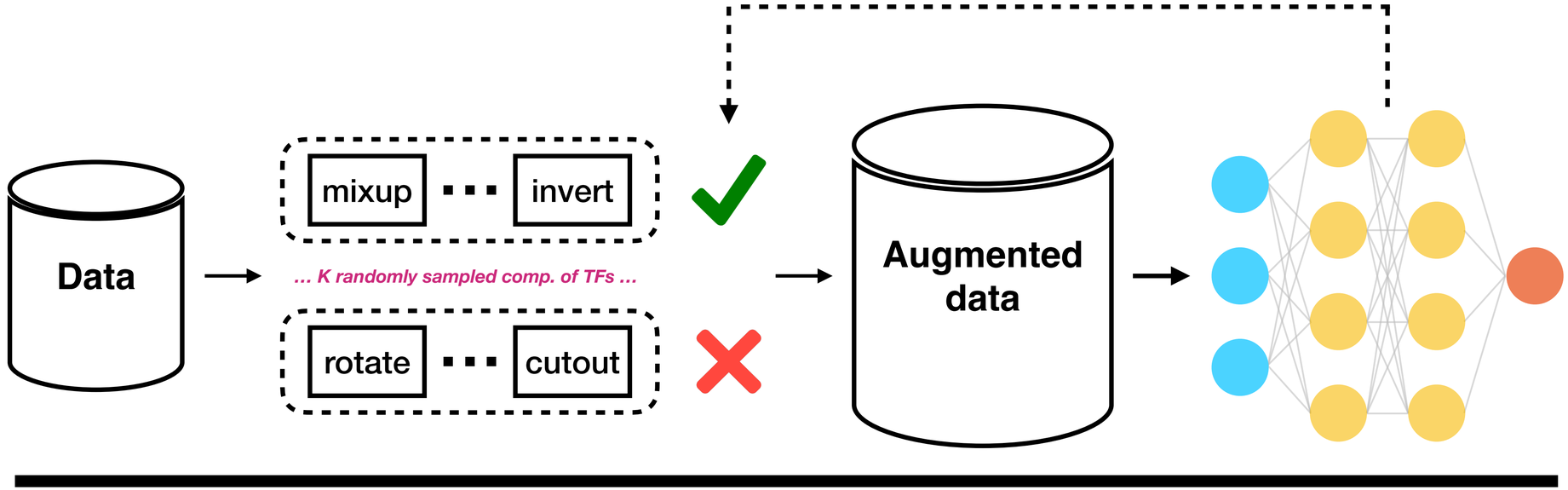
[Zhang et al. 2019]

Uncertainty-based sampling augmentation



Users provide **transformation functions (TFs)**

Model selects the TFs that provides **the most information** during training
—**No policy learning required**



[Wu et al. 2020]

Empirical results: State of the art quality

Improved the existing methods across domains

- SoTA on CIFAR-10, CIFAR-100, and SVHN
  - **84.54%** on CIFAR-100 using Wide-ResNet-28-10
    - outperforming RandAugment (Cubuk et al.'19) by **1.24%**
- Improved 0.28 pts. in accuracy on text classification problem

CIFAR-10

CIFAR-100

SVHN

Check out the blog post series!

Automating the Art of Data Augmentation

Part I Overview

Series edited by Sharon Li and Chris Ré. Referencing work by many other members of Hazy Research.

Posted on February 26, 2020

[Automating the Art of Data Augmentation (Part I: Overview)]

[Automating the Art of Data Augmentation (Part II: Practical Methods)]

[Automating the Art of Data Augmentation (Part III: Theory)]

[Automating the Art of Data Augmentation (Part IV: New Direction)]

Part III: Self-supervised Learning

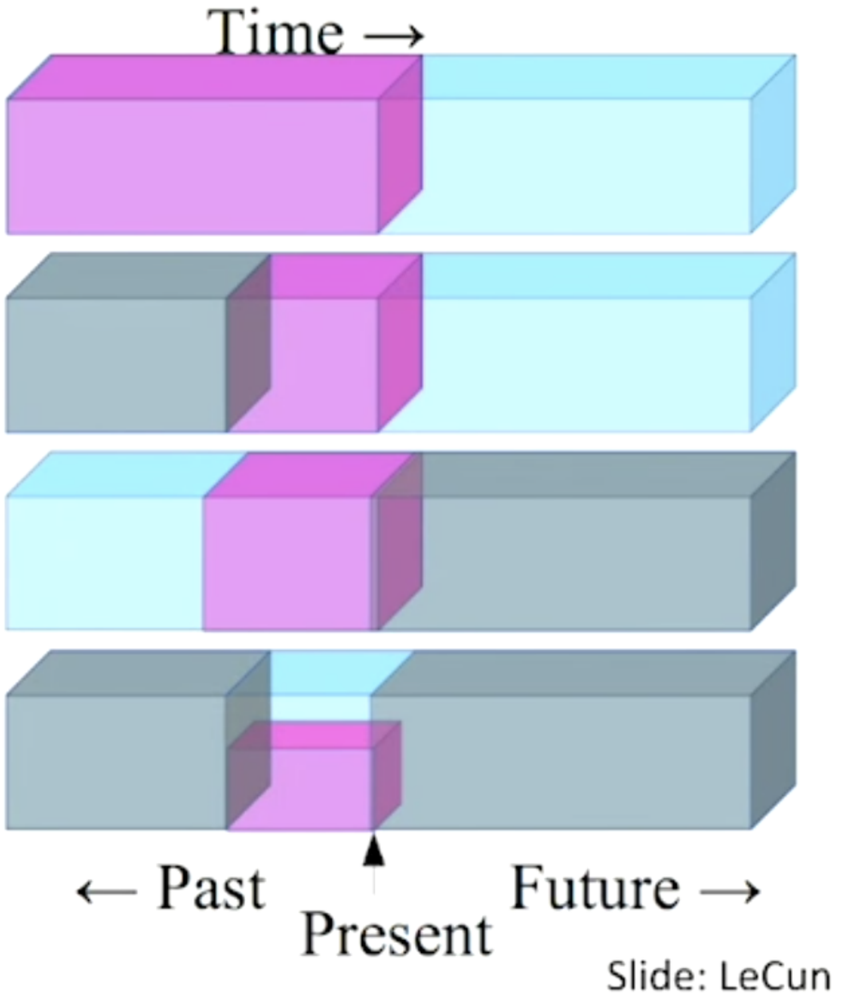
- **"Pure" Reinforcement Learning (cherry)**
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**
- **Supervised Learning (icing)**
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - **10→10,000 bits per sample**
- **Self-Supervised Learning (cake génoise)**
  - The machine predicts any part of its input for any observed part.

Source: Yann LeCun's talk

What if we can get labels for free for unlabelled data and train unsupervised dataset in a supervised manner?

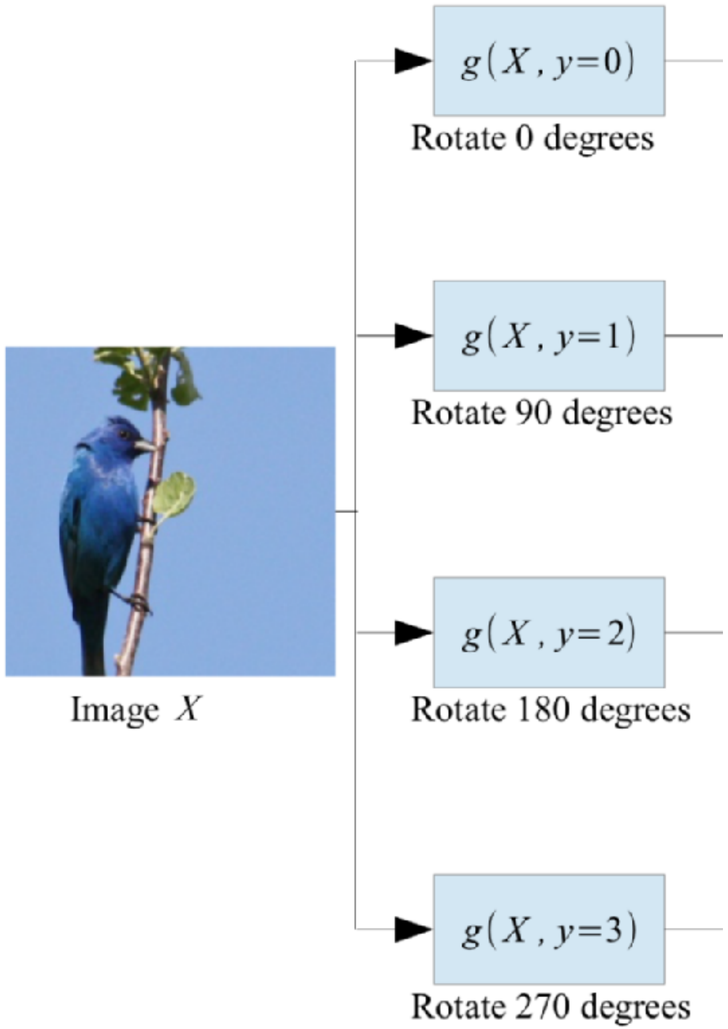
Pretext Tasks

- Predict any part of the input from any other part.
- Predict the future from the past.
- Predict the future from the recent past.
- Predict the past from the present.
- Predict the top from the bottom.
- Predict the occluded from the visible
- Pretend there is a part of the input you don't know and predict that.



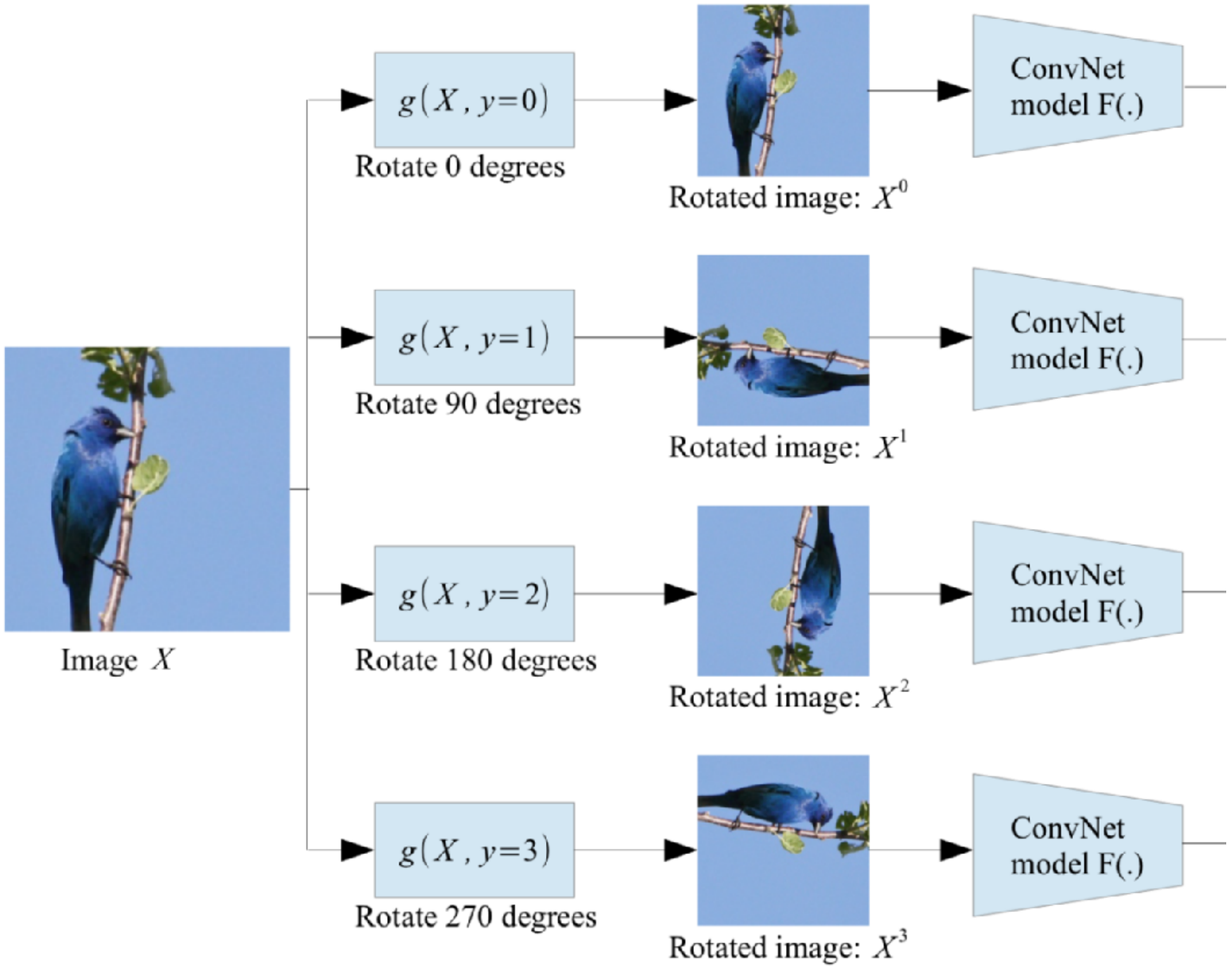
Slide: LeCun

Rotation



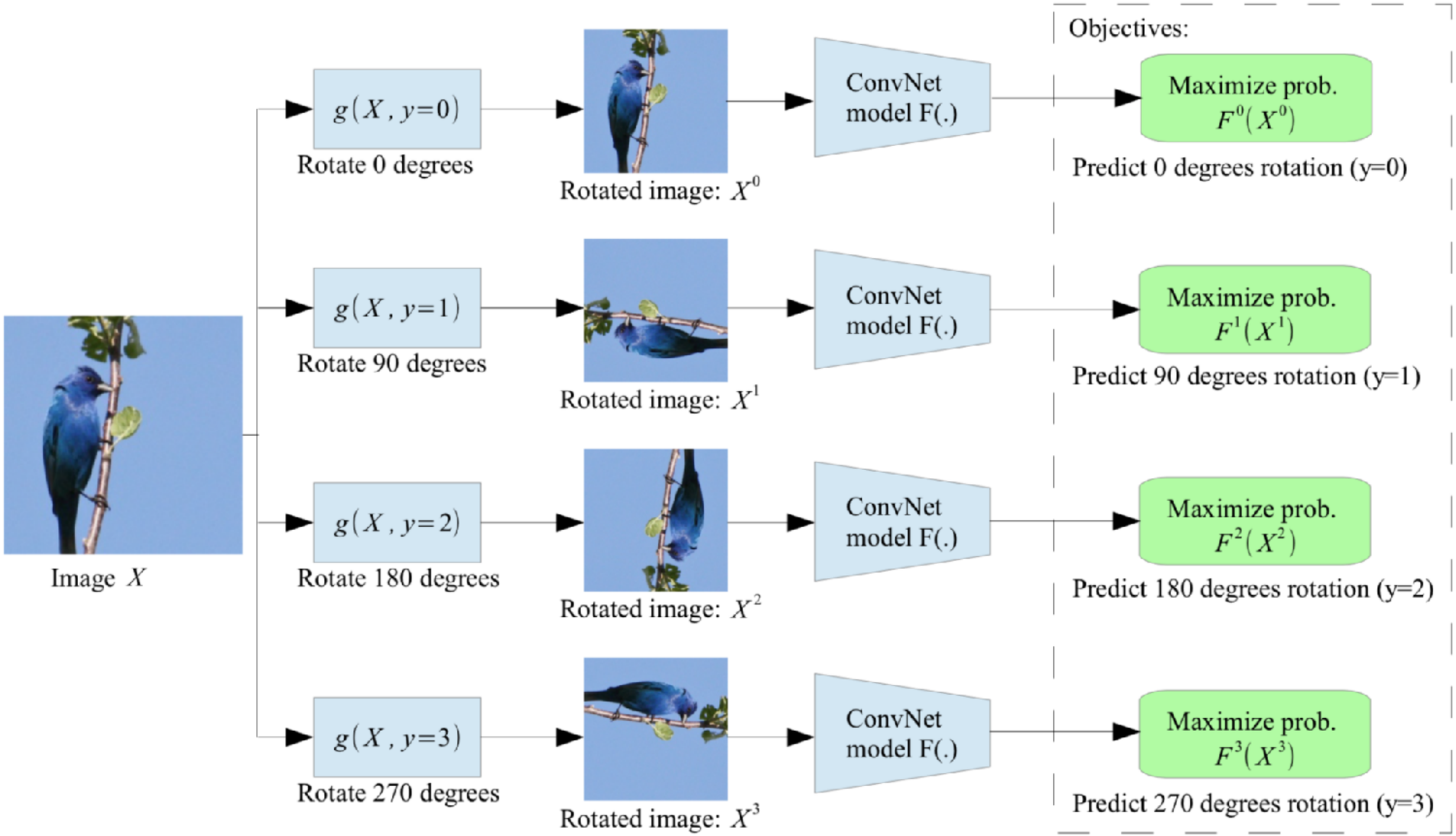
[Gidaris et al. 2018]

Rotation



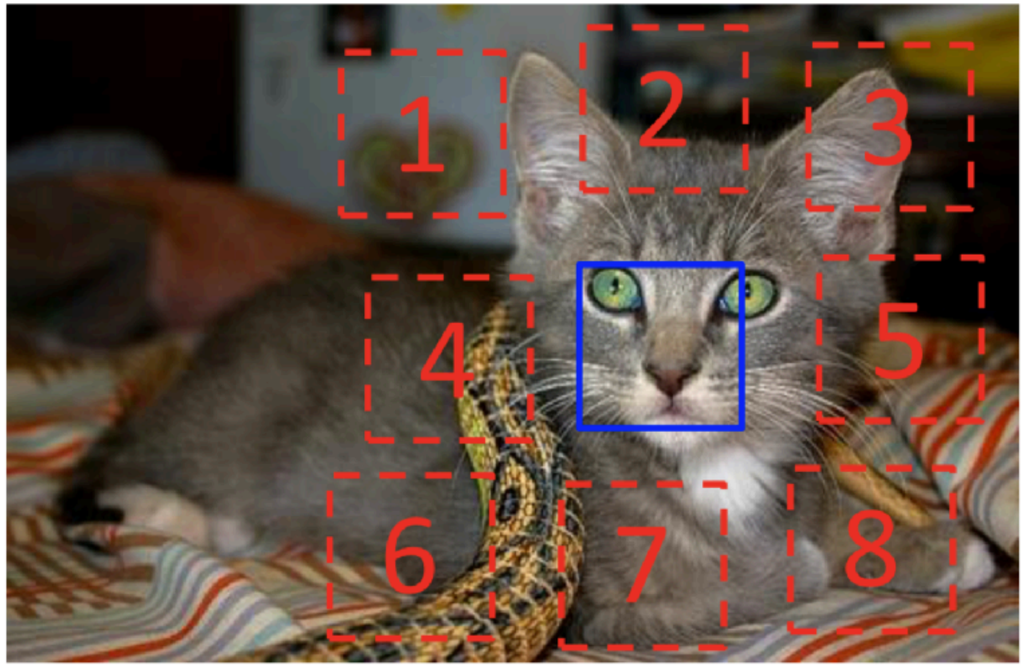
Gidaris et al. 2018

Rotation



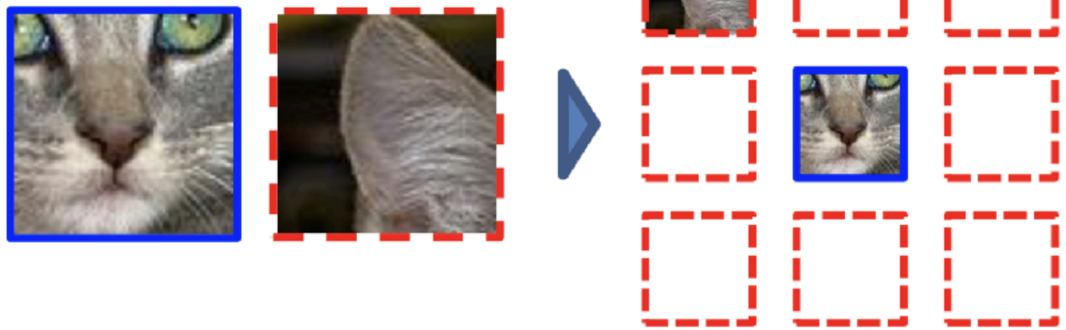
Gidaris et al. 2018

Patches



X = (,); Y = 3

Example:



Question 1:

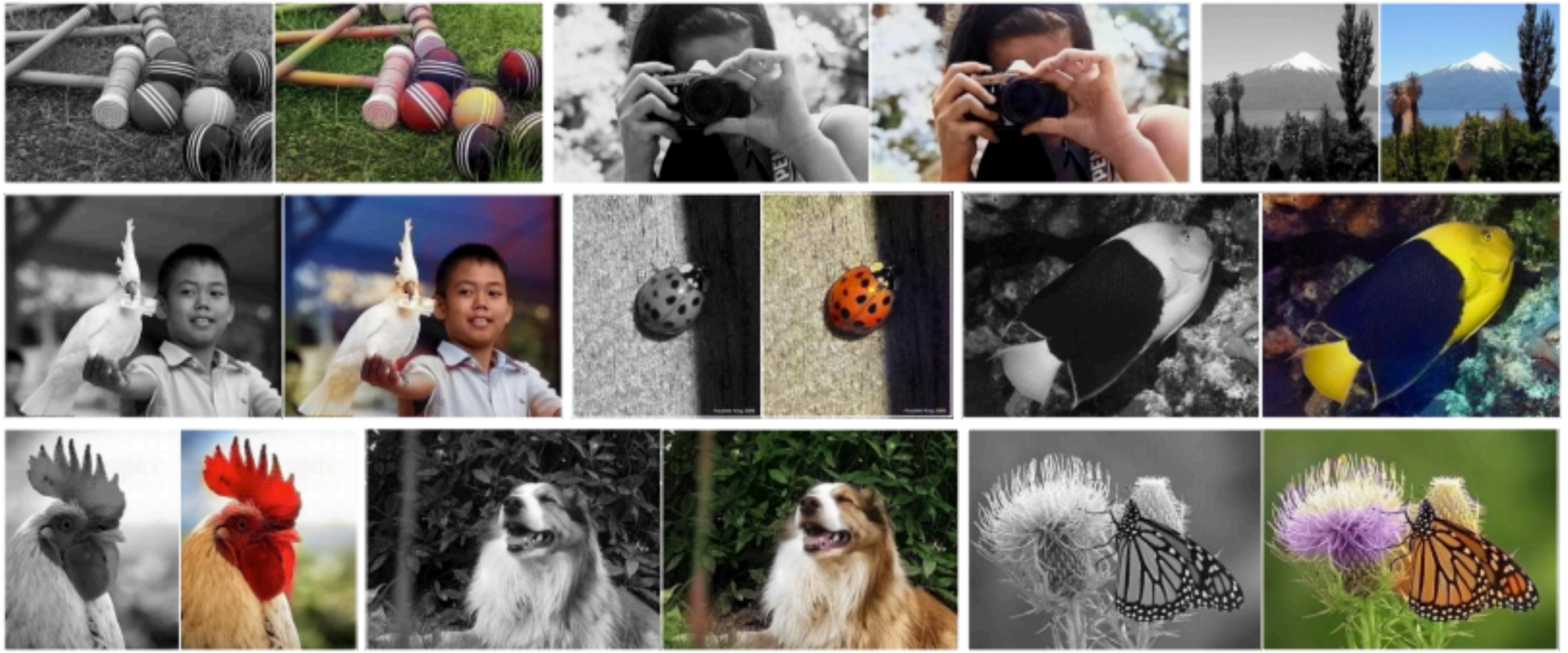
?

Question 2:

?

[Doersch et al., 2015]

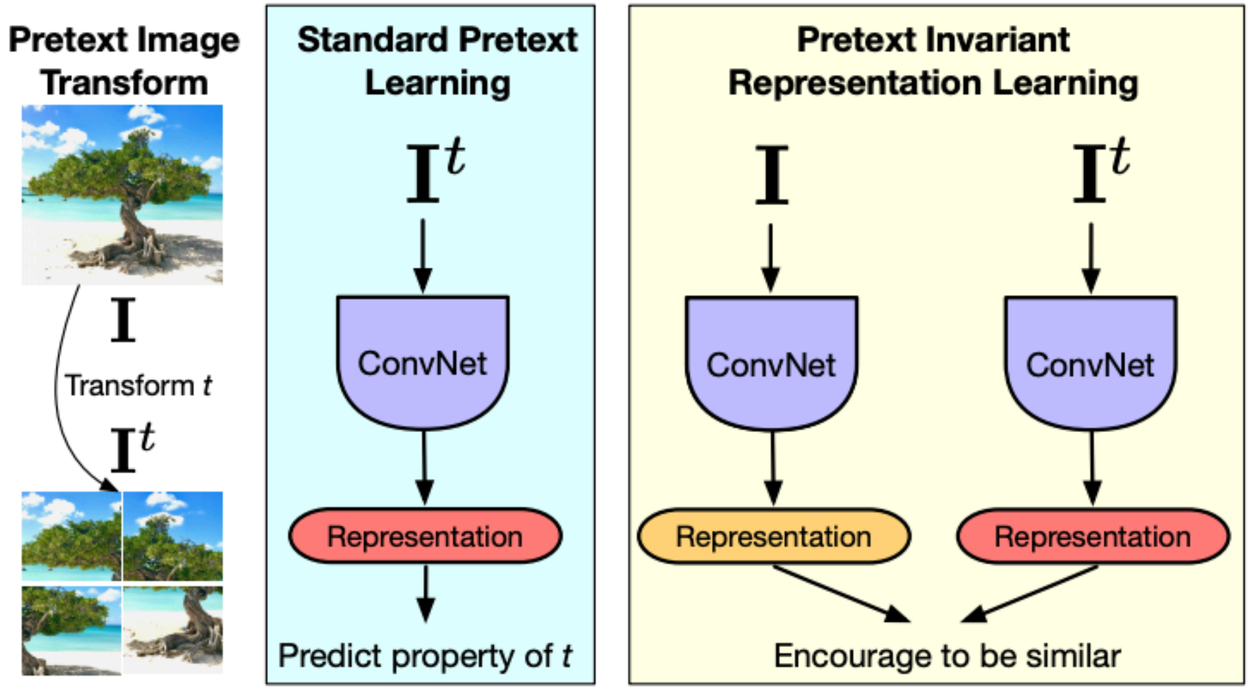
Colorization



[Zhang et al. 2016]

http://richzhang.github.io/colorization/

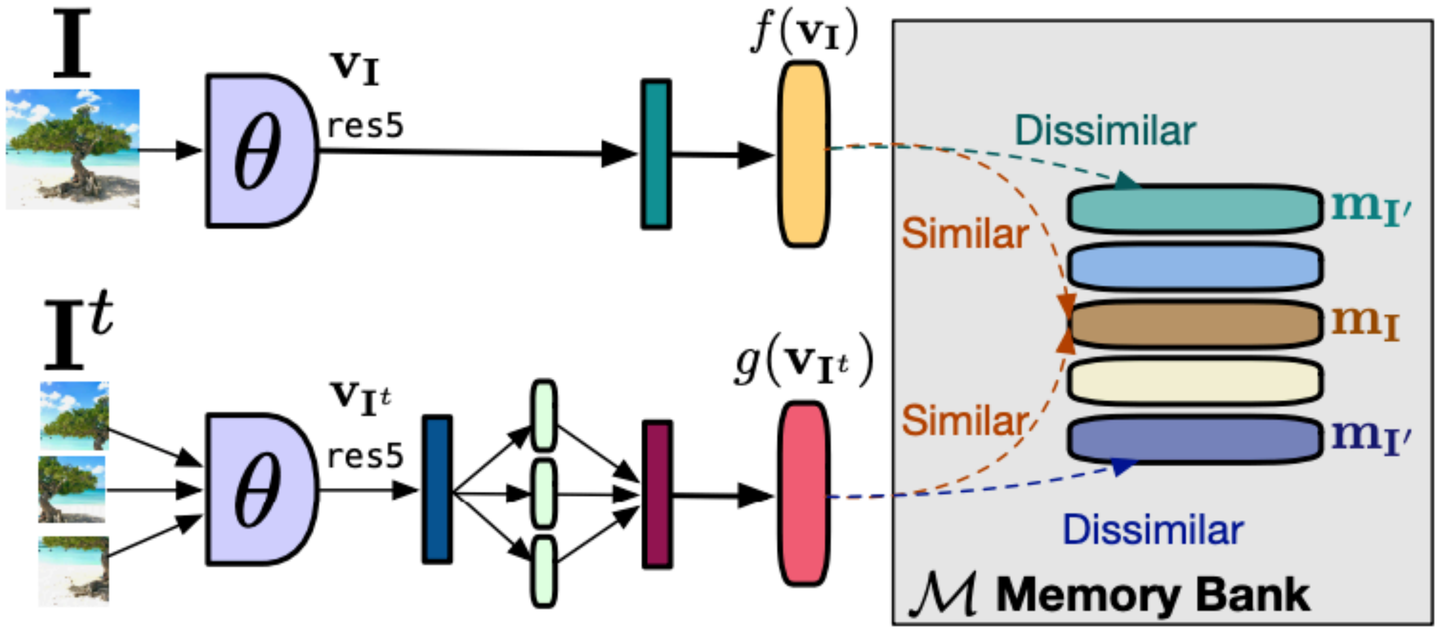
Pretext Invariant Representation Learning (PIRL) [Misra et al. 2019]



$$\ell_{co}(\theta;\mathcal{D}) = \mathbb{E}_{t\sim p(\mathcal{T})}\left[\frac{1}{|\mathcal{D}|}\sum_{\mathbf{I}\in\mathcal{D}} L_{co}\left(\mathbf{v}_{\mathbf{I}}, z(t)\right)\right]$$

$$\ell_{inv}(\theta;\mathcal{D}) = \mathbb{E}_{t\sim p(\mathcal{T})}\left[\frac{1}{|\mathcal{D}|}\sum_{\mathbf{I}\in\mathcal{D}} L\left(\mathbf{v}_{\mathbf{I}}, \mathbf{v}_{\mathbf{I}^t}\right)\right]$$

Pretext Invariant Representation Learning (PIRL) [Misra et al. 2019]

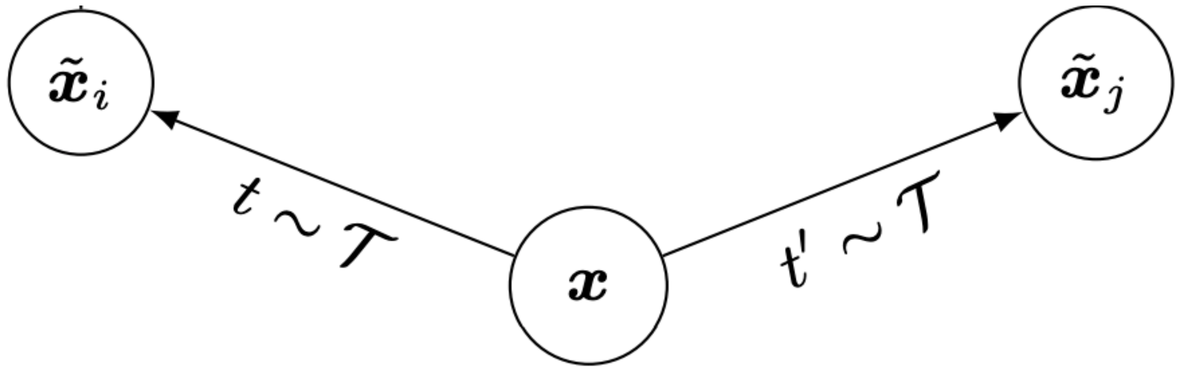


$$h(\mathbf{v}_\mathbf{I}, \mathbf{v}_{\mathbf{I}^t}) = \frac{\exp\left(\frac{s(\mathbf{v}_\mathbf{I}, \mathbf{v}_{\mathbf{I}^t})}{\tau}\right)}{\exp\left(\frac{s(\mathbf{v}_\mathbf{I}, \mathbf{v}_{\mathbf{I}^t})}{\tau}\right) + \sum_{\mathbf{I}' \in \mathcal{D}_N} \exp\left(\frac{s(\mathbf{v}_{\mathbf{I}^t}, \mathbf{v}_{\mathbf{I}'})}{\tau}\right)}.$$

Positive pair

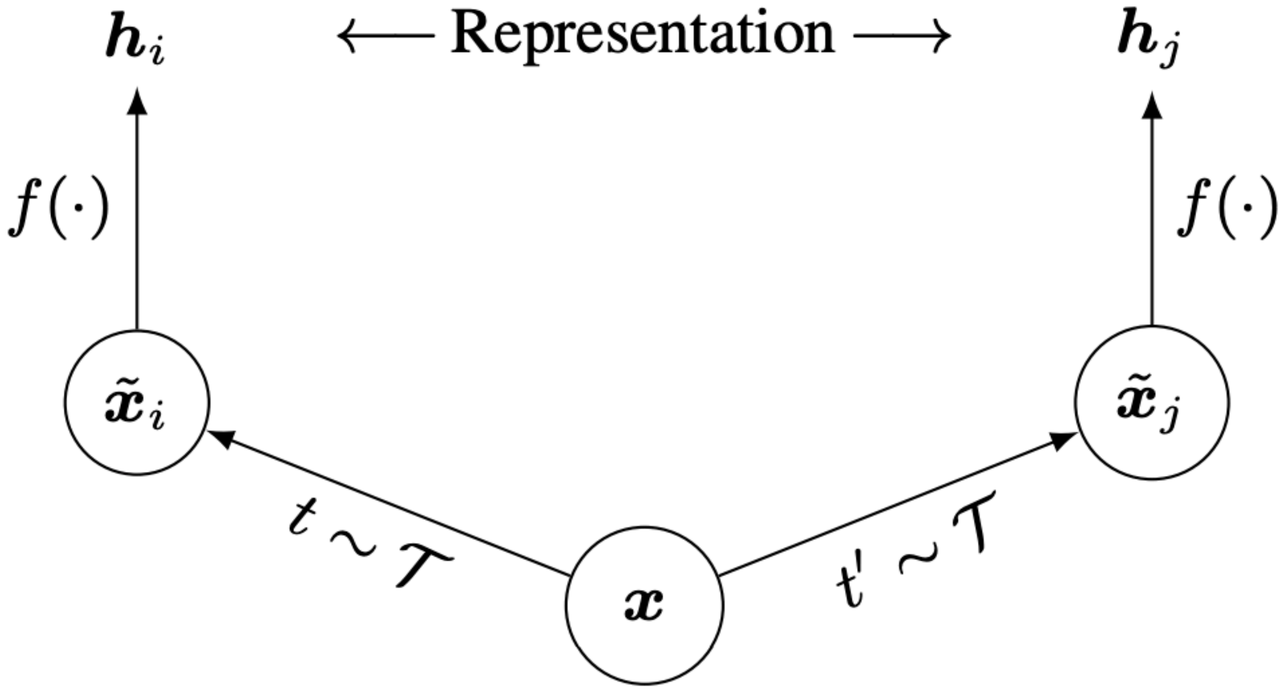
Negative pairs

SimCLR



[Chen et al. 2020]

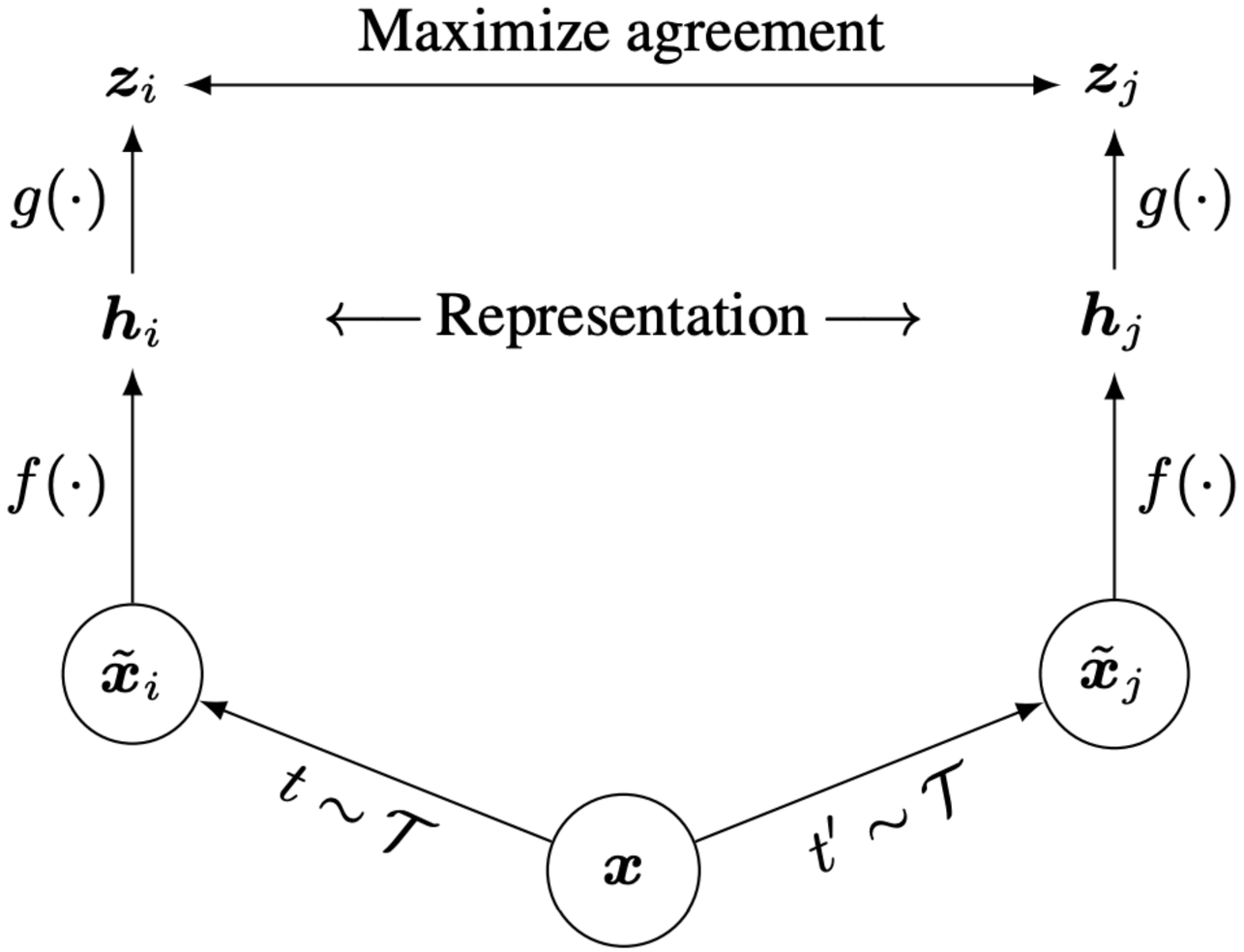
SimCLR



$$\boldsymbol{h}_i = f(\tilde{\boldsymbol{x}}_i) = \text{ResNet}(\tilde{\boldsymbol{x}}_i)$$

[Chen et al. 2020]

SimCLR



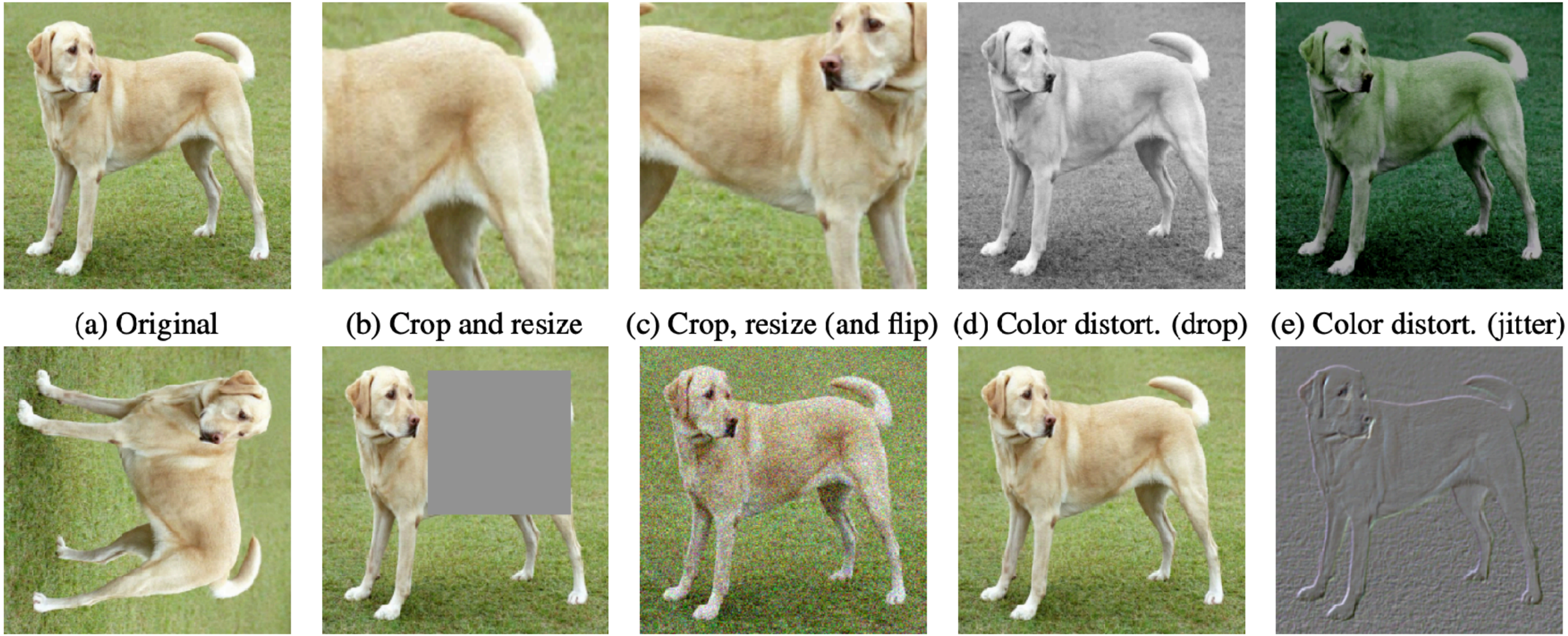
$$\boldsymbol{h}_i = f(\tilde{\boldsymbol{x}}_i) = \text{ResNet}(\tilde{\boldsymbol{x}}_i)$$

$$\boldsymbol{z}_i = g(\boldsymbol{h}_i) = W^{(2)}\sigma(W^{(1)}\boldsymbol{h}_i)$$

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

[Chen et al. 2020]

Data Augmentation is the key

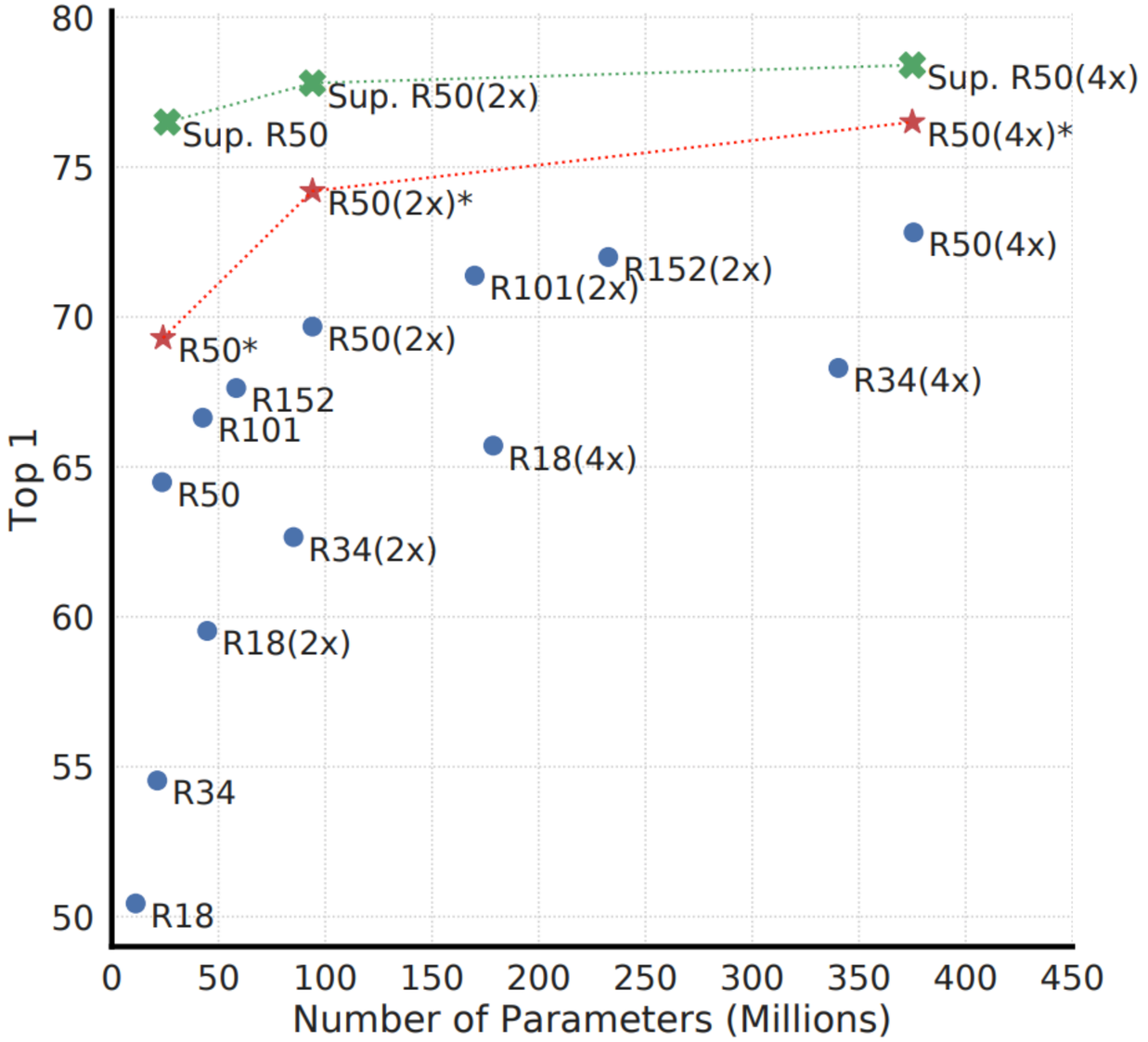


(a) Original (b) Crop and resize (c) Crop, resize (and flip) (d) Color distort. (drop) (e) Color distort. (jitter)

(f) Rotate $\{90°, 180°, 270°\}$ (g) Cutout (h) Gaussian noise (i) Gaussian blur (j) Sobel filtering

[Chen et al. 2020]

Unsupervised learning benefits more from bigger models



[Chen et al. 2020]

Summary

- **Weakly Supervised Learning**
  - Flickr100M
  - JFT300M (Google)
  - Instagram3B (Facebook)
- **Data augmentation**
  - Human heuristics
  - Automated data augmentation
- **Unsupervised Learning**
  - Pretext tasks (rotation, patches, colorization etc.)
  - Invariant vs. Covariant learning
  - Contrastive learning based framework (current SoTA)

Questions?