S4L: Self-Supervised Semi-Supervised Learning

Authors: Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, Lucas Beyer

ICCV 2019

Presenter: SABAINA HAROON

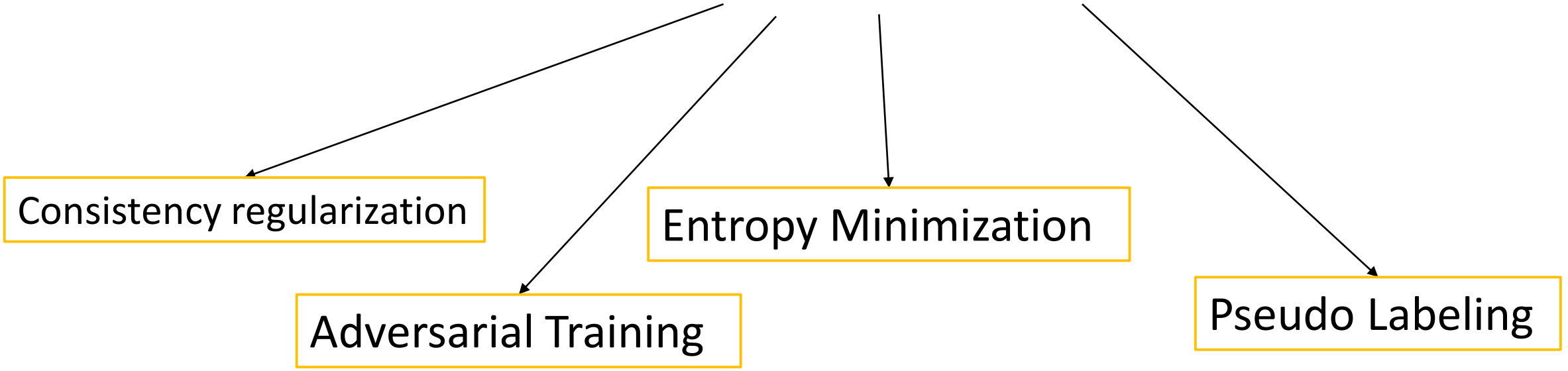
CAP-6412 SPRING 2020

Overview

- Introduction
- Baselines
- S4L
- Experiments
- MOAM
- Conclusion

Introduction

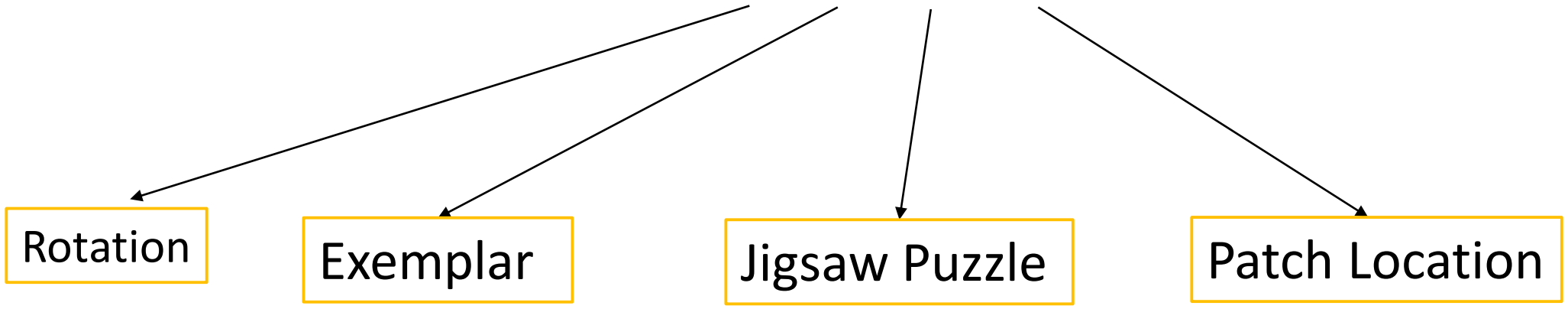
Semi-supervised Learning: supervised + unsupervised learning



Introduction

Semi-supervised Learning: supervised + unsupervised learning

Self-supervised: Learn unknown data representation from known data



Introduction

Semi-supervised Learning: supervised + unsupervised learning

Self-supervised: Learn unknown data representation from known data

Use **Self-supervised** visual representation for **Semi-supervised** learning

Motivation:

- Bridge together Self supervised learning and Semi-supervised learning
- Learn self supervised visual representations learning for semi-supervised learning
- Human brain can exploit unlabeled data
- Unlabeled data is usually cheap to collect

Semi-Supervised Learning Baselines

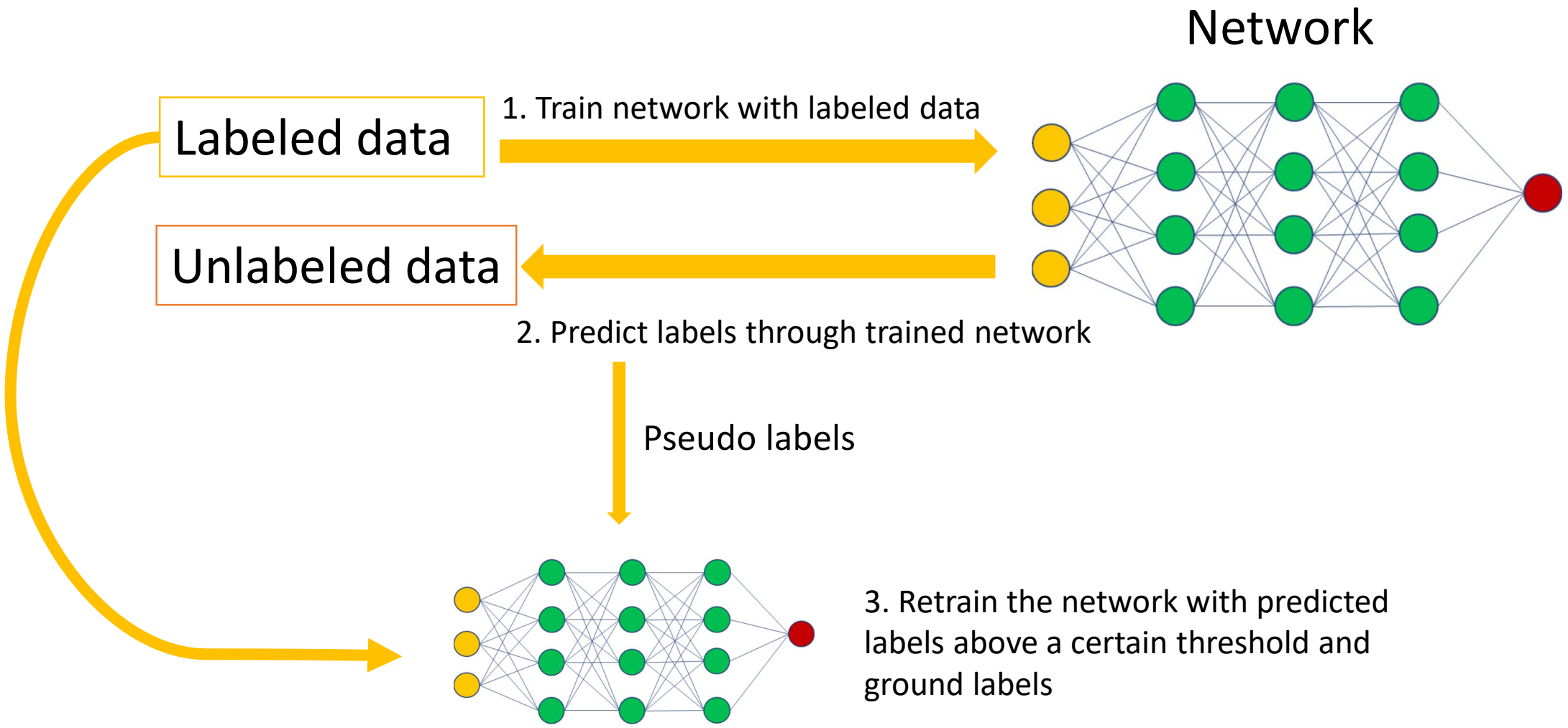


Learn a better prediction rule together from labeled and unlabeled data



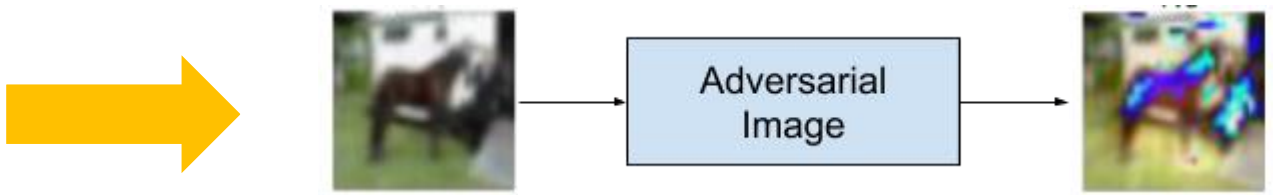
Dataset typically assumed to be sampled from the same or similar distributions.

Pseudo-Labeling

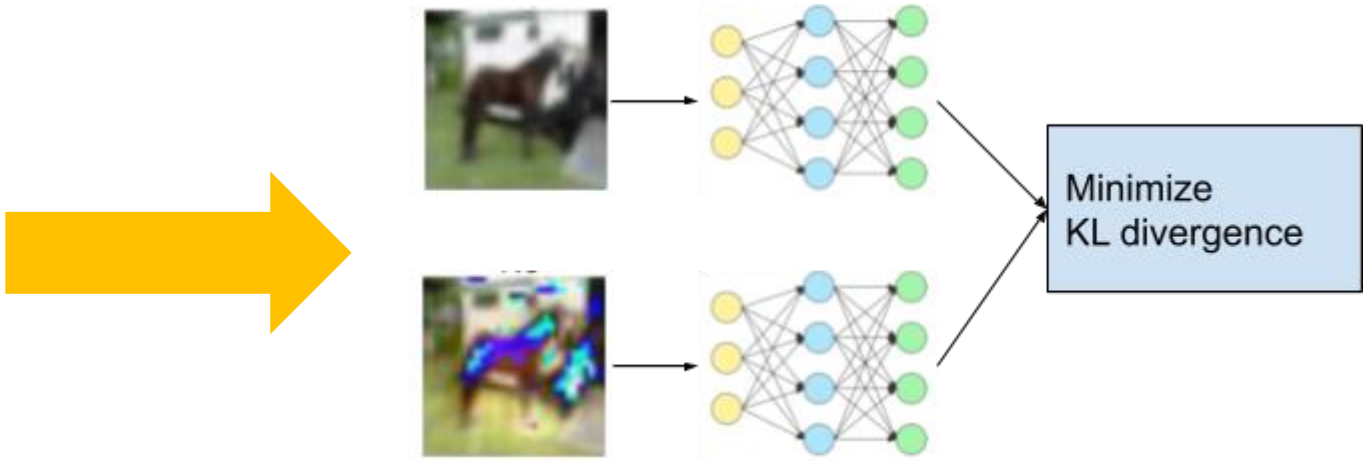


Virtual Adversarial training (VAT)

Given an image, adversarial perturbation is added in it to generate a perturbed image



Model is penalized for sensitivity with the perturbation



VAT LOSS for a model f_θ

$$\mathcal{L}_{\text{vat}} = \frac{1}{|\mathcal{D}_u|} \sum_{x \in \mathcal{D}_u} \text{KL}(f_\theta(x) \parallel f_\theta(x + \Delta x)),$$

$$\text{where } \Delta x = \underset{\delta \text{ s.t. } |\delta|_2 = \epsilon}{\arg\max} \text{KL}(f_\theta(x) \parallel f_\theta(x + \delta)).$$

- Approximates the maximal change in predictions within ϵ_{vat} vicinity of unlabeled data points
- $\epsilon_{\text{vat}} = |\delta|_2$ is a hyperparameter

Vat + Entropy minimization(EntMin) Model

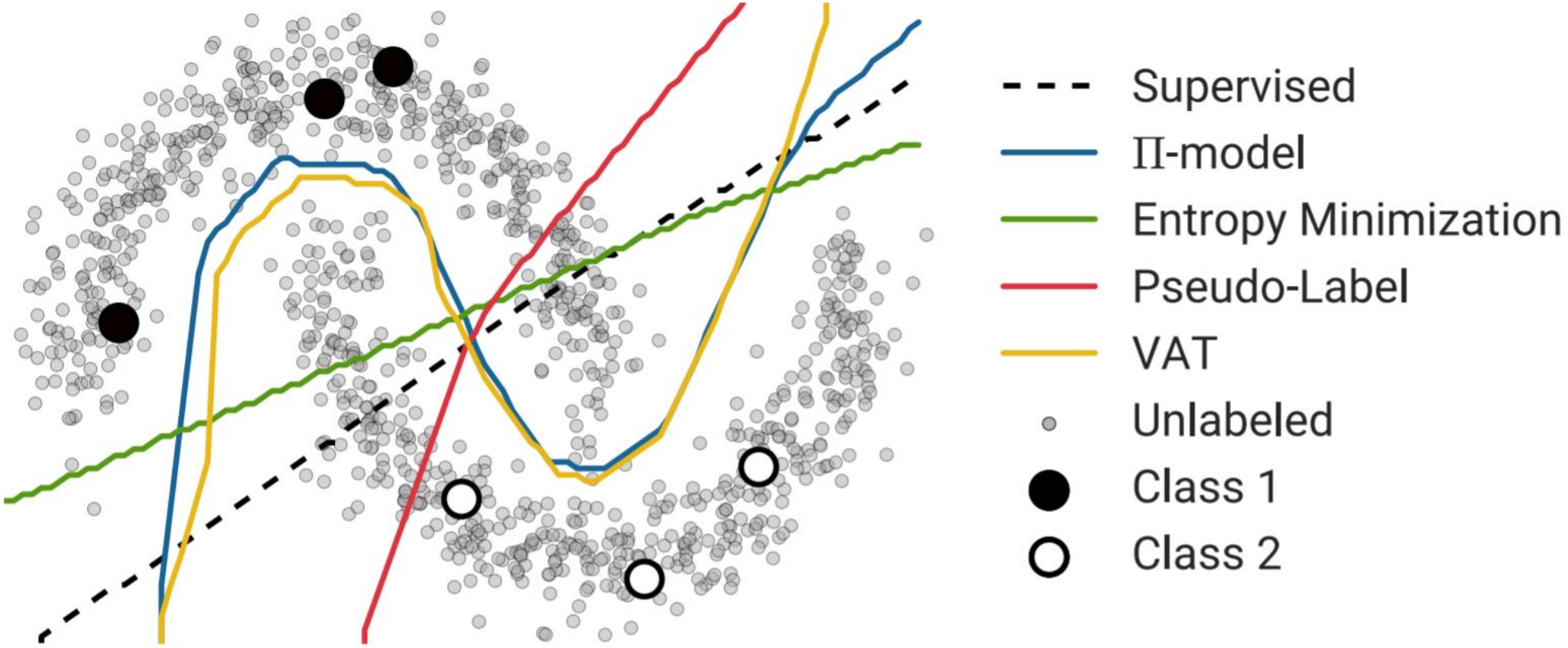
- **EntMin Assumption** : unlabeled data is based on some input class
- **Entropy** measure randomness in data distribution
- **Vat** learns vast distribution

Optimize VAT model with Entropy loss

$$\mathcal{L}_u = w_{vat}\mathcal{L}_{\text{vat}} + w_{\text{entmin}}\mathcal{L}_{\text{entmin}}$$

$$\mathcal{L}_{\text{entmin}} = \frac{1}{|\mathcal{D}_u|}\sum_{x\in\mathcal{D}_u}\sum_{y\in Y} -f_\theta(y|x)\log f_\theta(y|x)$$

Comparison of SSL methods by Oliver *et al.*



Self-Supervised Learning Baselines



Multi-class classification problem with pretext tasks to learn new representations ”



Still learns visual representations significantly inferior to that learned by fully supervised techniques

Rotation Model

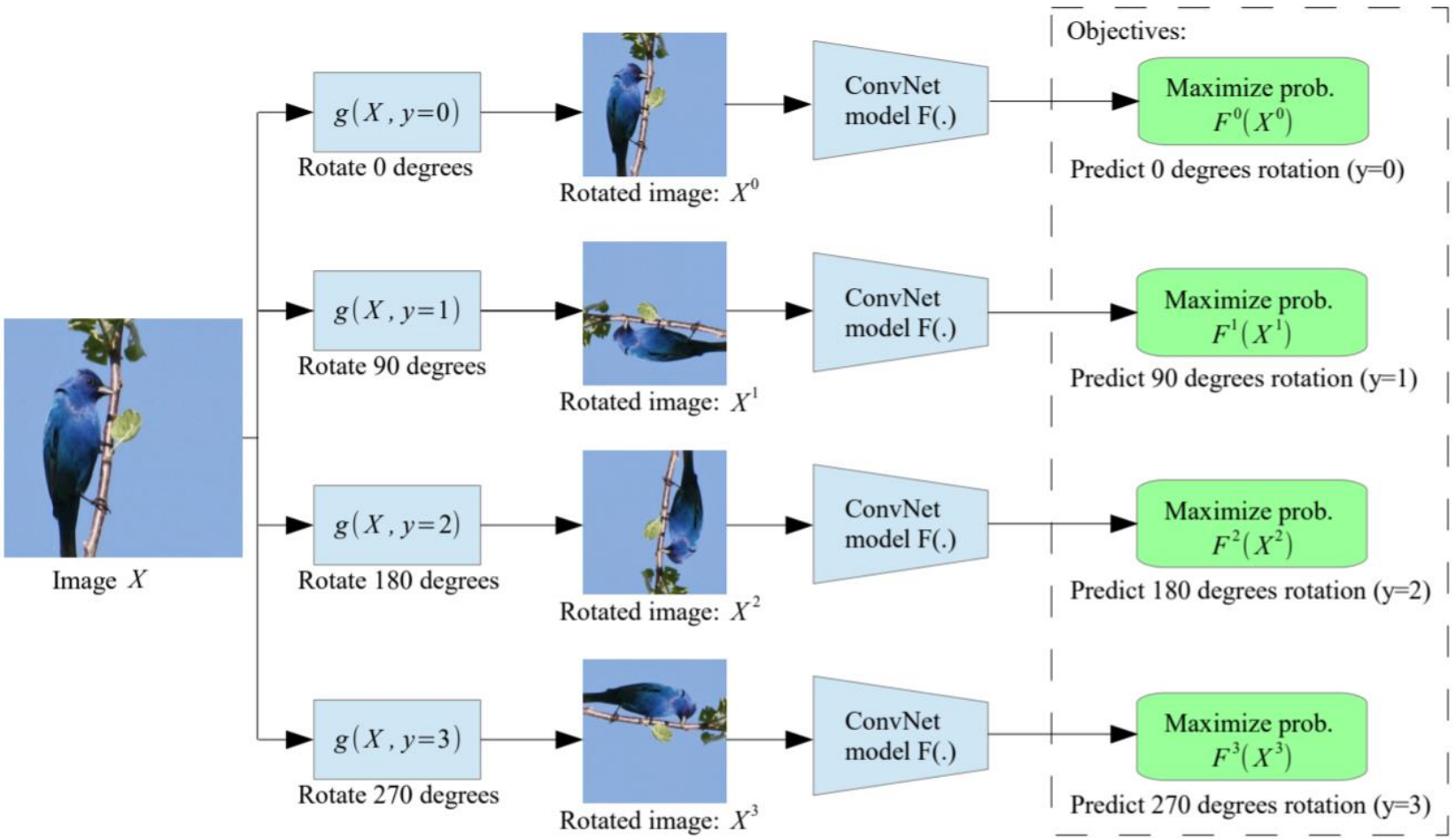
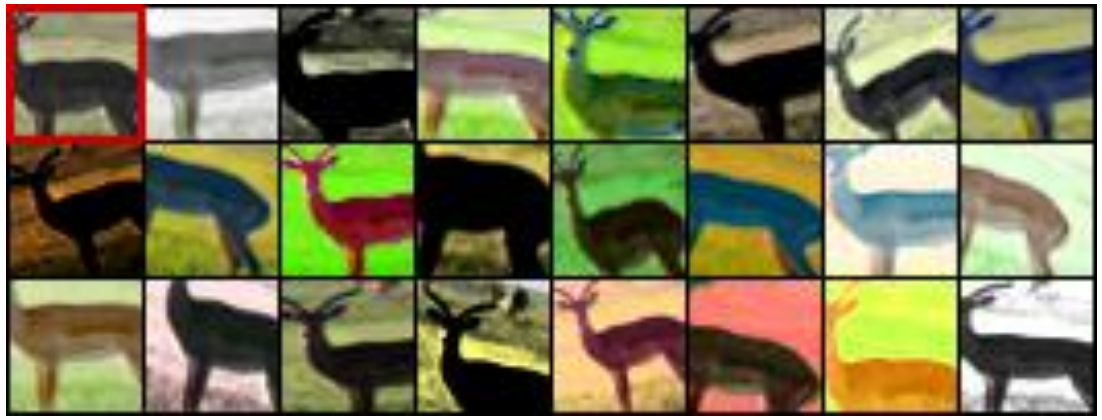


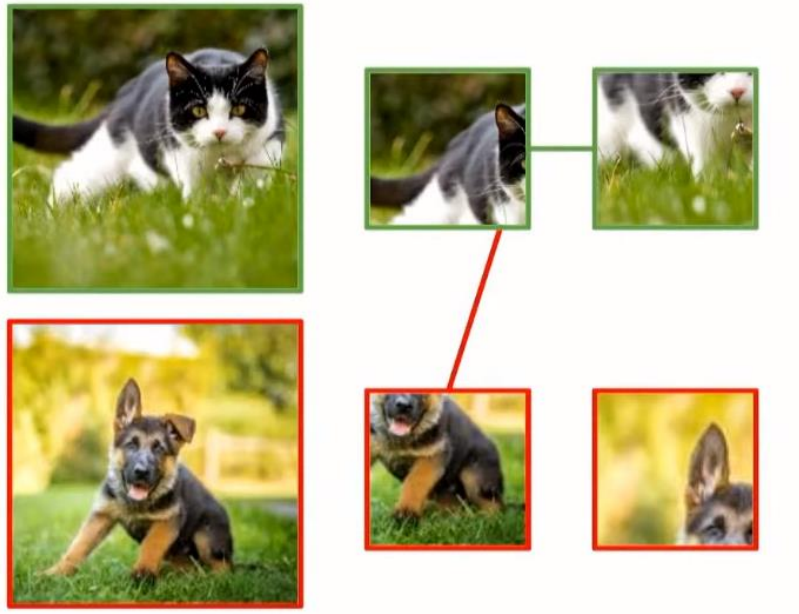
Image source: Gidaris et al. 2018)

Exemplar CNN

Hue transformation augmentation



Patch Augmentation



Learn patch distributions such that : patches from same image have similar distribution – done using triplet loss

(Image source: Dosovitskiy et al., 2015)

Triplet loss



Anchor
A



Positive
P



Anchor
A



Negative
N

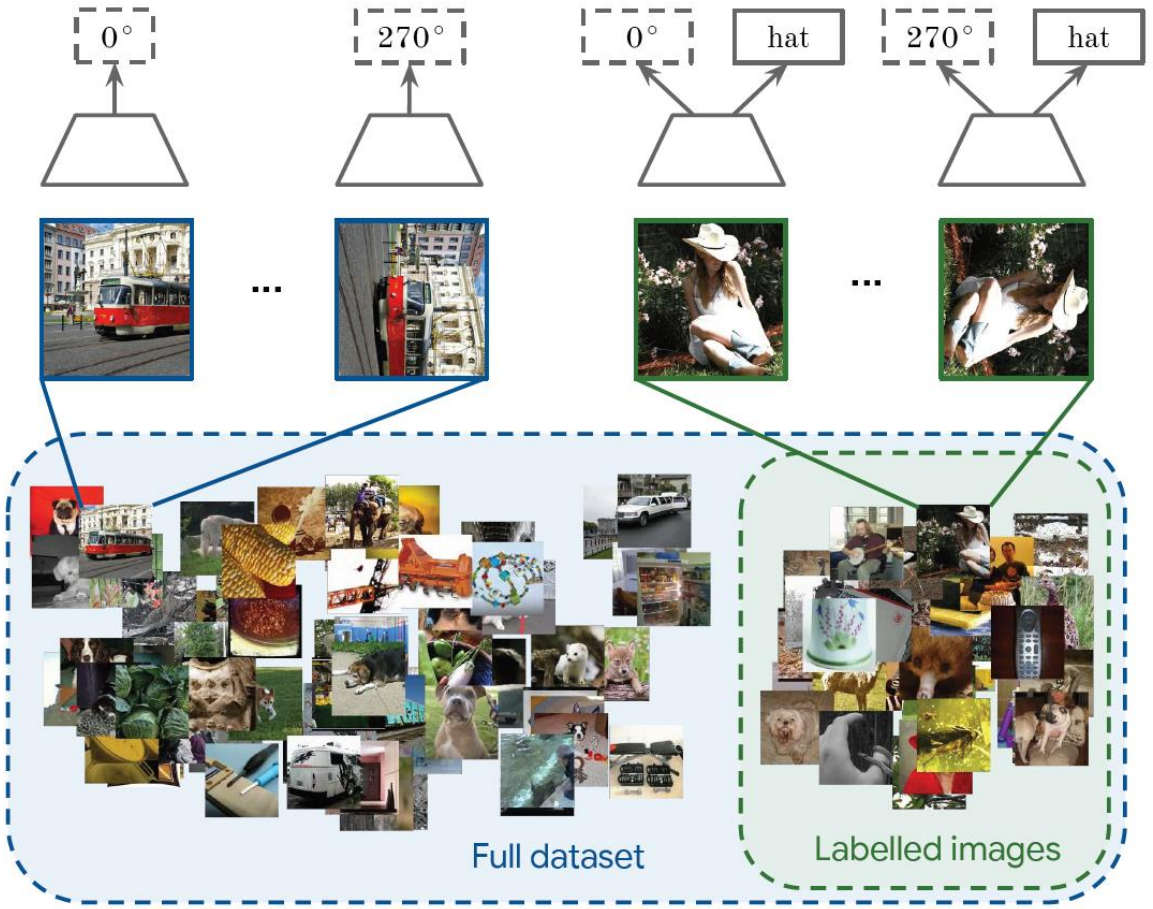
$$\|f(\boldsymbol{A}) - f(\boldsymbol{P})\|^2 \leq \|f(\boldsymbol{A}) - f(\boldsymbol{N})\|^2$$

$$\|f(\boldsymbol{A}) - f(\boldsymbol{P})\|^2 - \|f(\boldsymbol{A}) - f(\boldsymbol{N})\|^2 + \alpha \leq 0$$

[Schroff et al.,2015, FaceNet: A unified embedding for face recognition and clustering]

Self-Supervised Semi-Supervised Learning

S^4L



Self Supervised:

Apply any self-supervision loss on data without labels

$$\mathcal{L}_u = \mathcal{L}_{rot} \;/\; \mathcal{L}_{exemplar}$$

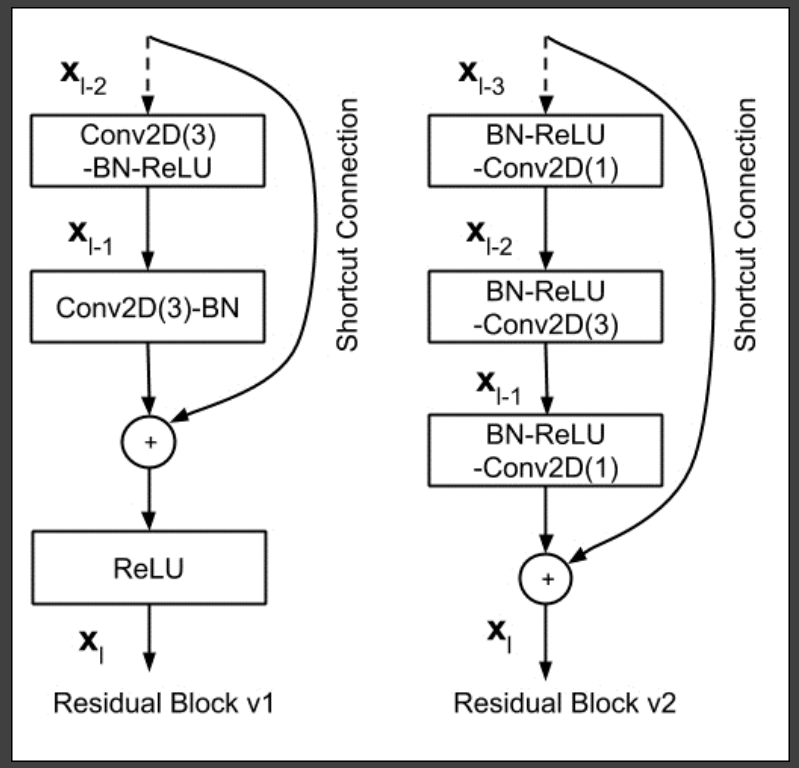
Semi Supervised:

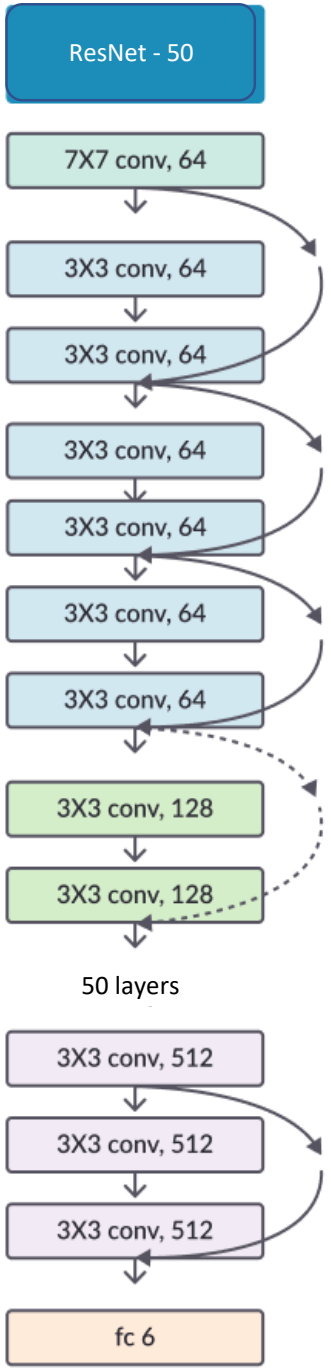
Apply regular supervised loss on data with labels

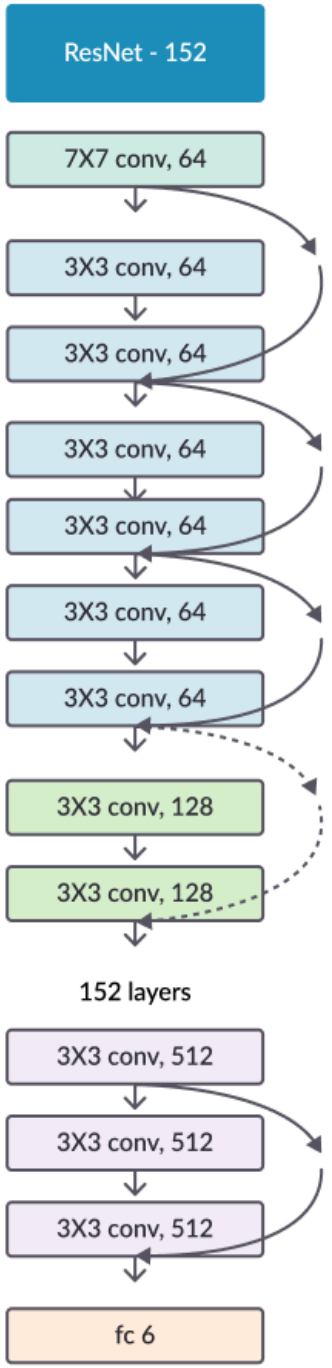
$$\mathcal{L} = w_{sup}\,\mathcal{L}_{sup} + w_{rot}\,\mathcal{L}_{rot}$$

Experiments

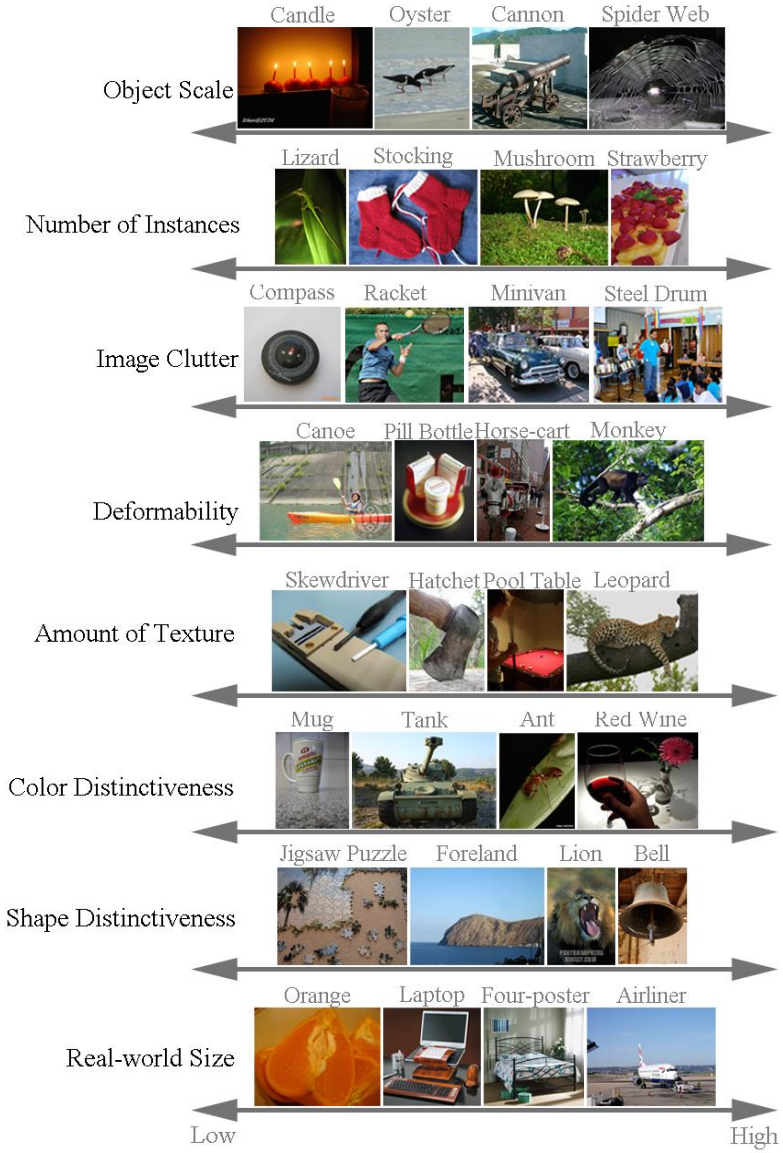
Baseline Architecture: ResNetV2







Dataset: ILSVRC-2012 ImageNet





Dataset: ILSVRC-2012 ImageNet

Image classification annotations (1000 object classes)

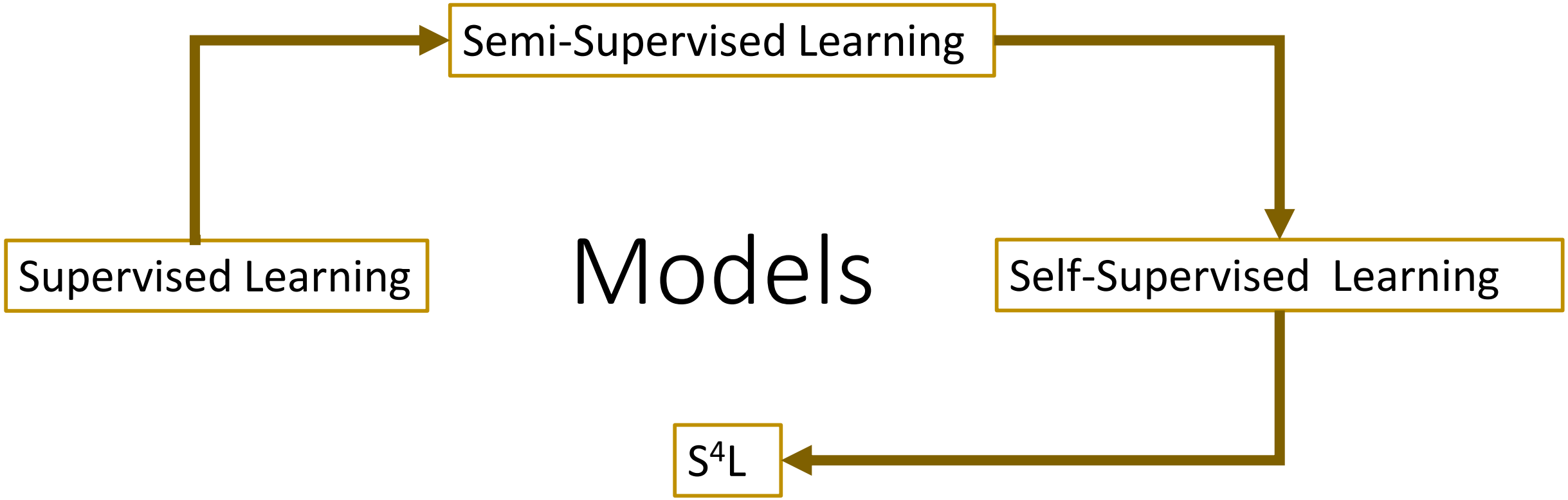
Year	Train images (per class)	Val images (per class)	Test images (per class)
ILSVRC2012-14	1,281,167 (732-1300)	50,000 (50)	100,000 (100)

Additional annotations for single-object localization (1000 object classes)

Year	Train images with bbox annotations (per class)	Train bboxes annotated (per class)	Val images with bbox annotations (per class)	Val bboxes annotated (per class)	Test images with bbox annotations
ILSVRC2012-14	523,966 (91-1268)	593,173 (92-1418)	50,000 (50)	64,058 (50-189)	100,000

Dataset: ILSVRC-2012 ImageNet

- Models tuned with Train images set divided into 1,231,121 training set and 50,046 validation set
- Retrained Best model with full 1.28M training dataset
- Model performance calculated using 50,000 validation images
- 10% Class balanced labels available (128k labeled images)



Models trained on 10% and 1% class balanced label data

Plain Supervised Learning Model:

- Trained a plane supervised model , ResNet50v2, on available labeled dataset

labels	Top-5	Top-1
10%	80.43%	48.43%
1%	56.35%	25.35%

Aim to provide stronger baseline in future : smaller gap between labeled and unlabeled data

Semi Supervised Learning Model:

Pseudo-Label

- plain supervised baseline predicts for unlabeled data
- Retrained with **full dataset**
  - Model ResNet50v2
  - Learning rate: 0.1,Weight decay: 10^{-4}

Semi Supervised Learning Model:

VAT

- First verified on CIFAR-10 with 4000 labels: 86% top1 accuracy
- Used same hyperparameters from previous models
- Additional ϵ_{vat} tuning

VAT + EntMin

- best VAT model with cross entropy minimization loss
- $w_{\text{EntMin}} = \{0, 0.03, 0.1, 0.3, 1\}$

Self Supervised Learning Model:

- Used **exemplar** and **rotation** learning models following same setting as : *Revisiting Self-Supervised Visual Representation Learning*
- Standard width 4x
- Self-sup + Linear
- Self-sup + Fine-tune

S^4L Model:

- S^4L-Rotation
- S^4L-Exemplar
- Same protocol from semi-supervised baseline, only tuned learning rate , weight decay for newly introduced loss
- Self-supervised loss on both label/unlabeled images
- Not sensitive to random seed / split of the dataset

MOAM: *Mix of all Models*

1. Rotation+VAT+EntMin

Optimized model for S^4L-Rotation, VAT and EntMin losses



	labels	Top-5	Top-1
MOAM (*proposed*)	10%	88.80	69.73

MOAM: *Mix of all Models*

2. Retraining on Pseudo Labels

Pseudo labels for full dataset generated using Rotation+VAT+EntMin model

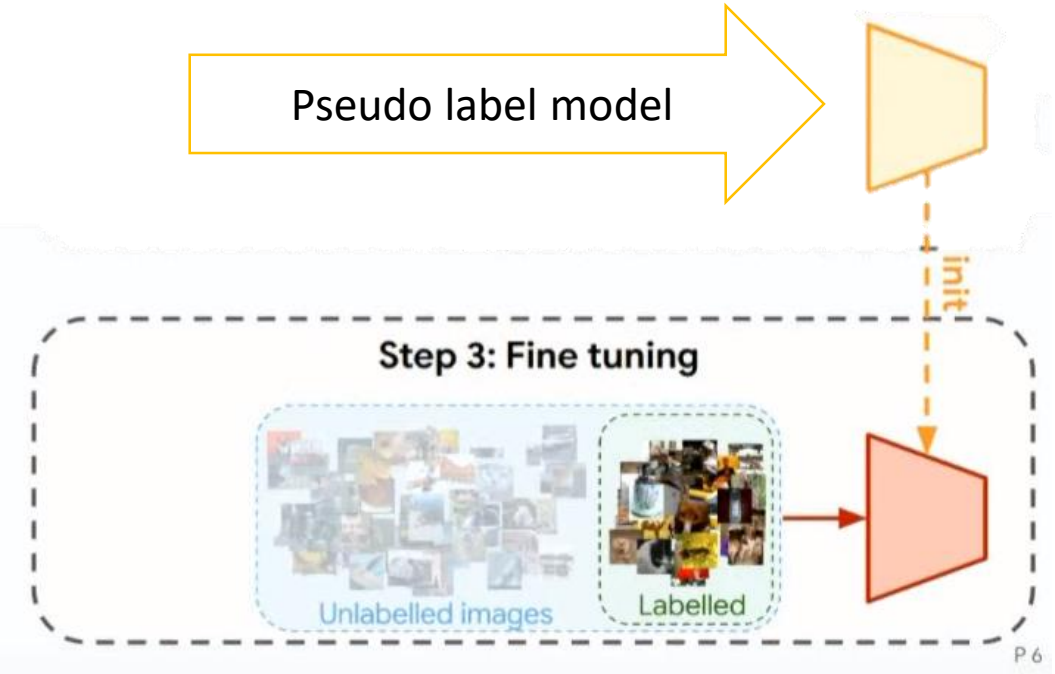


	labels	Top-5	Top-1
MOAM (*proposed*)	10%	88.80	69.73
MOAM + pseudo label (*proposed*)	10%	89.96	71.56

MOAM: *Mix of all Models*

3. Fine tuning the model

Fine tuned the model from pseudo labels on original 10% labels only



	labels	Top-5	Top-1
MOAM (*proposed*)	10%	88.80	69.73
MOAM + pseudo label (*proposed*)	10%	89.96	71.56
MOAM full (*proposed*)	10%	**91.23**	**73.21**

Set new state-of-the-art

Experimental Results

ILSVRC-2012 labels: (i.e. images per class)	10 % (128)	1 % (13)
Supervised Baseline (Section 4.1)	80.43	48.43
Pseudolabels [19]	82.41	51.56
VAT [22]	82.78	44.05
VAT + Entropy Minimization [10]	83.39	46.96
Self-sup. Rotation [16] + Linear	39.75	25.98
Self-sup. Exemplar [16] + Linear	32.32	21.33
Self-sup. Rotation [16] + Fine-tune	78.53	45.11
Self-sup. Exemplar [16] + Fine-tune	81.01	44.90
S^4L-Rotation	83.82	53.37
S^4L-Exemplar	83.72	47.02

Improves ~20% over previously supervised baseline

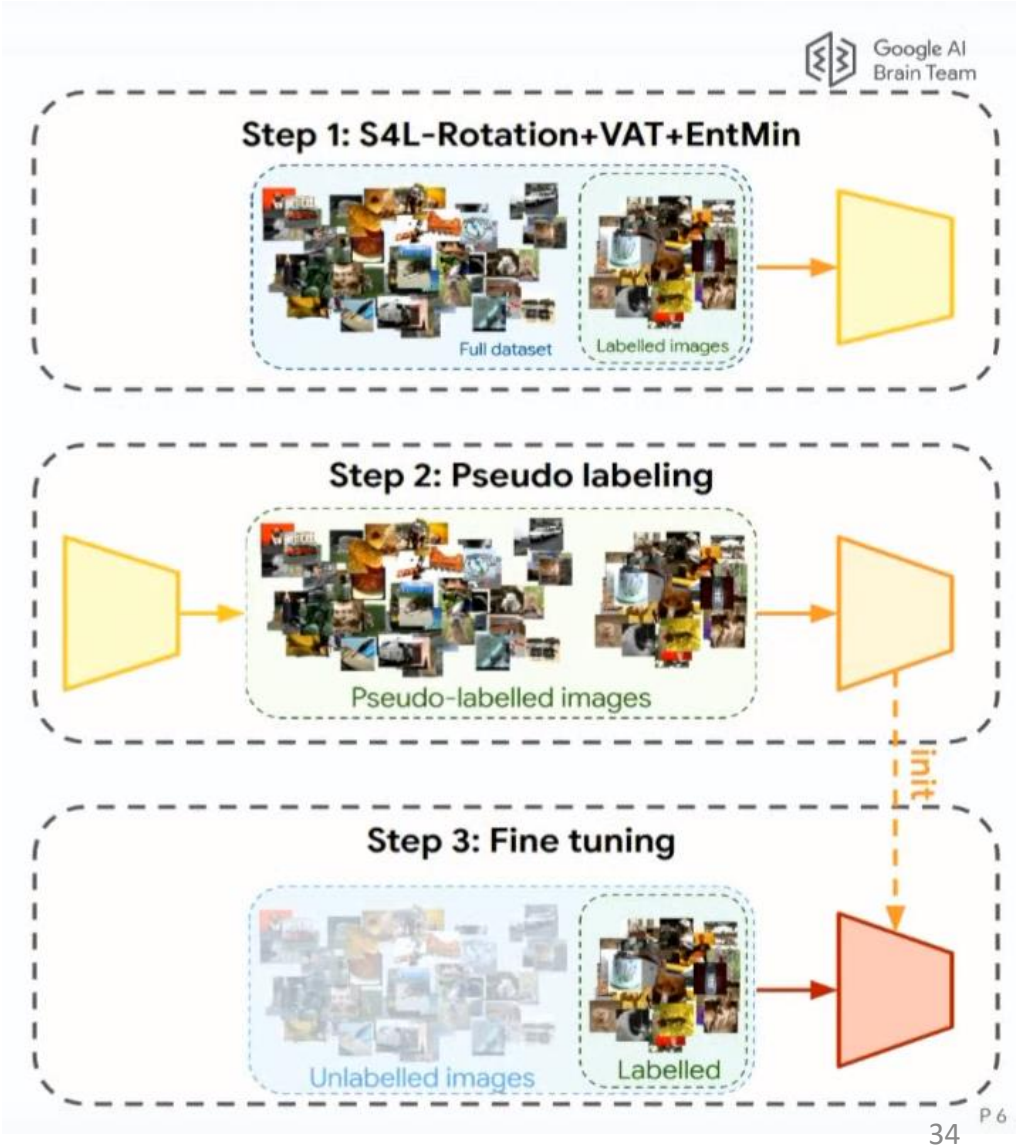
Reported standard SSL technique first time for ILSVR-2012 ImageNet dataset

First trained with self supervision then linear/fine tune with labels

Proposed S^4L method gives competitive semi-supervised learning methods

MOAM

	labels	Top-5	Top-1
MOAM full (*proposed*)	10%	**91.23**	**73.21**
MOAM + pseudo label (*proposed*)	10%	89.96	71.56
MOAM (*proposed*)	10%	88.80	69.73
ResNet50v2 (4×wider)	10%	81.29	58.15
VAE + Bayesian SVM [30]	10%	64.76	48.41
Mean Teacher [38]	10%	90.89	-
†UDA [40]	10%	88.52†	68.66†
†CPCv2 [12]	10%	84.88†	64.03†
Training with all labels:		*†marks concurrent work.*	
ResNet50v2 (4×wider)	100%	94.10	78.57
MOAM (*proposed*)	100%	**94.97**	**80.17**
†UDA [40]	100%	94.45†	79.04†
†CPCv2 [12]	100%	93.35†	-

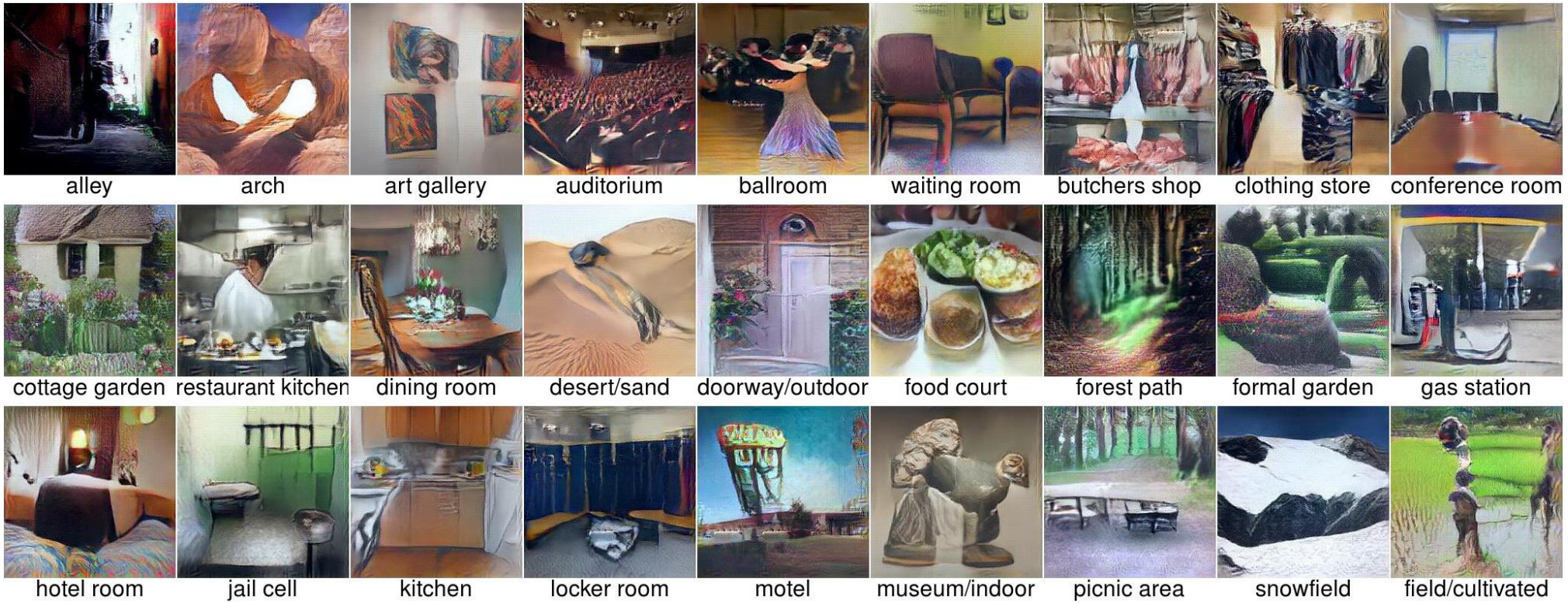


Transfer of Learned Representation

- S^4L Rotation model as a fixed feature extractor
- Train a logistic Regression model on Place205
- Compared to pure self supervised Rotation learning
- MOAM (full) transfers better than baseline 83.35% vs 83.1%

Transfer of Learned Representation

Place205



- Around 2.5 million images.
- 205 different scene types

Transfer of Learned Representation

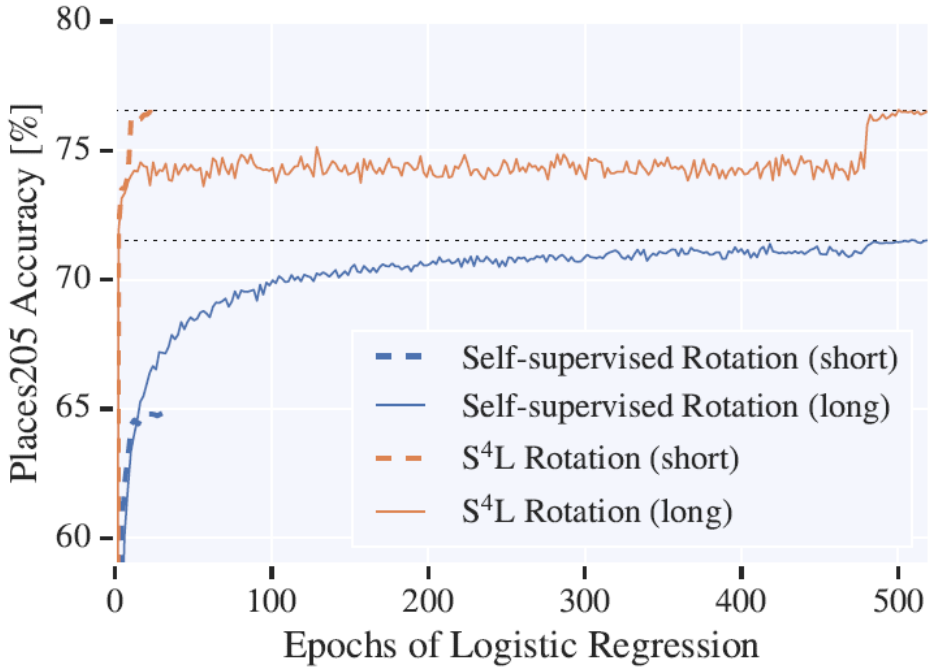


Figure 2. Places205 learning curves of logistic regression on top of the features learned by pre-training a self-supervised versus S^4L-Rotation model on ILSVRC-2012. The significantly faster convergence ("long" training schedule vs. "short" one) suggests that more easily separable features are learned.

Tiny validation set enough for tuning

Current:

- Use full validation set for tuning and evaluating model

Ours:

- different data for tuning and testing

Previous Claim:

- Model selection using small validation set not viable, gap in accuracies

Our Claim:

- Small validation dataset is enough!
- Proved by training on 5000 and 1000 validation images
- **Small variance in accuracy now**

Tiny validation set enough for tuning

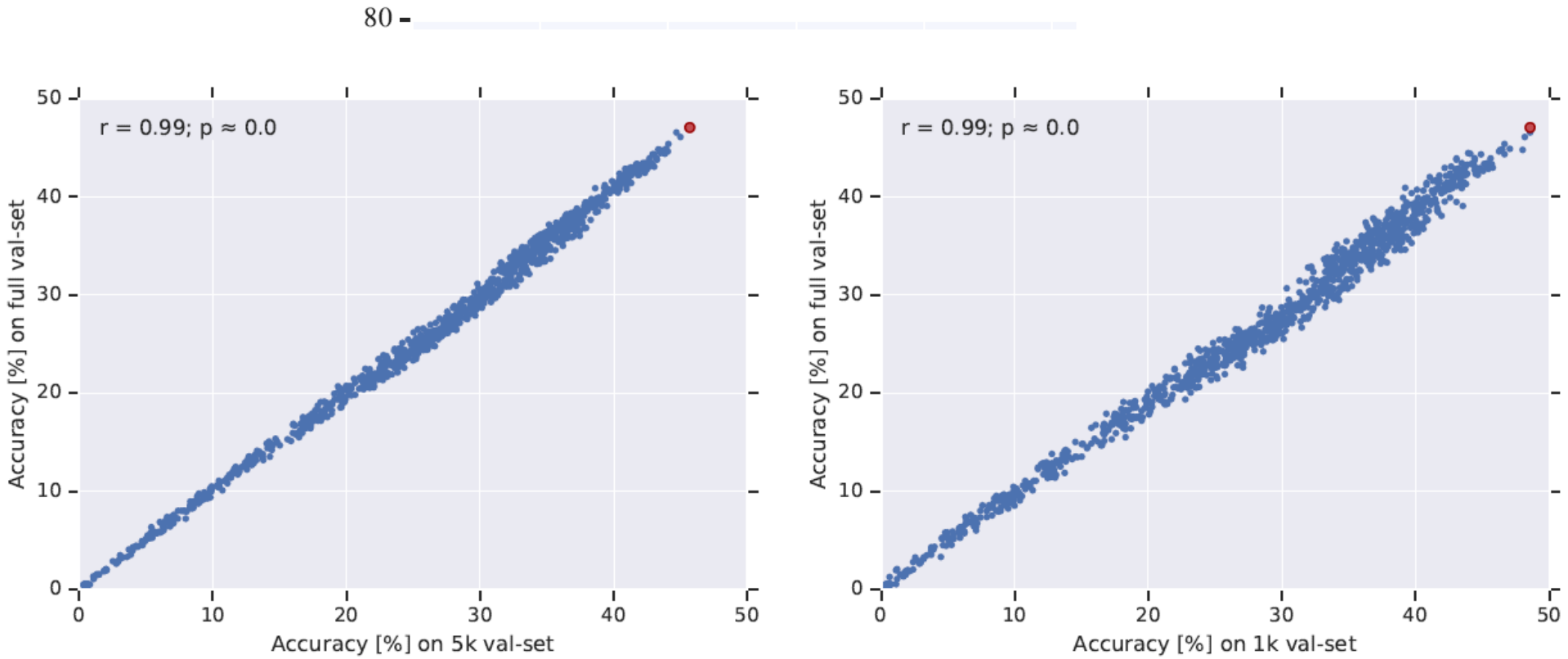


Figure 3. Correlation between validation score on a (custom) validation set of 1000, 5000, and 50 046 images on ILSVRC-2012. Each point corresponds to a *trained model* during a sweep for plain supervised baseline for the 1 % labeled case. The best model according to the validation set of 1 000 is marked in red. As can be seen, evaluating our models even with only a single validation image per class is robust, and in particular selecting an optimal model with this validation set works as well as with the full validation set.

Conclusion

- Bridged gap between semi and self supervised learning
- Any self supervision model extended to semi-supervised learning
- S^4L methods do not compete but complement existing methods
- State-of-the-art performance by MOAM

Questions