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

- Consistency regularization
- Adversarial Training
- Entropy Minimization
- Pseudo Labeling
Introduction

Semi-supervised Learning: supervised + unsupervised learning

Self-supervised: Learn unknown data representation from known data

- Rotation
- Exemplar
- Jigsaw Puzzle
- Patch Location
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

1. Train network with labeled data

Labeled data

Unlabeled data

2. Predict labels through trained network

Pseudo labels

3. Retrain the network with predicted labels above a certain threshold and ground labels
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 $f_\theta$

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

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

- Approximates the maximal change in predictions within $\mathcal{E}_{\text{vat}}$ vicinity of unlabeled data points
- $\mathcal{E}_{\text{vat}} = |\delta|_2$ 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

\[
L_u = w_{vat} L_{vat} + w_{entmin} L_{entmin}
\]

\[
L_{entmin} = \frac{1}{|\mathcal{D}_u|} \sum_{x \in \mathcal{D}_u} \sum_{y \in Y} -f_\theta(y|x) \log f_\theta(y|x)
\]
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

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

Patch Augmentation

(Image source: Dosovitskiy et al., 2015)
Triplet loss

\[ \| f(A) - f(P) \|^2 \leq \| f(A) - f(N) \|^2 \]

\[ \| f(A) - f(P) \|^2 - \| f(A) - f(N) \|^2 + \alpha \leq 0 \]

[Schroff et al., 2015, FaceNet: A unified embedding for face recognition and clustering]
Self-Supervised Semi-Supervised Learning
**Self Supervised:**
Apply any self-supervision loss on data without labels
$$\mathcal{L}_u = \mathcal{L}_{\text{rot}} / \mathcal{L}_{\text{exemplar}}$$

**Semi Supervised:**
Apply regular supervised loss on data with labels
$$\mathcal{L} = w_{\text{sup}} \mathcal{L}_{\text{sup}} + w_{\text{rot}} \mathcal{L}_{\text{rot}}$$
Experiments
Baseline Architecture: ResNetV2

ResNetV2 - 50
- 7x7 conv, 64
- 3x3 conv, 64
- 3x3 conv, 64
- 3x3 conv, 64
- 3x3 conv, 64
- 3x3 conv, 64
- 3x3 conv, 64
- 3x3 conv, 64
- 3x3 conv, 64
- 3x3 conv, 64
- 152 layers
- 3x3 conv, 512
- 3x3 conv, 512
- 3x3 conv, 512
- fc 6

ResNetV2 - 152
- 7x7 conv, 64
- 3x3 conv, 64
- 3x3 conv, 64
- 3x3 conv, 64
- 3x3 conv, 64
- 3x3 conv, 64
- 3x3 conv, 64
- 3x3 conv, 64
- 3x3 conv, 64
- (Highlighted in green)
- 50 layers
- 3x3 conv, 128
- 3x3 conv, 128
- 3x3 conv, 128
- 152 layers
- 3x3 conv, 512
- 3x3 conv, 512
- 3x3 conv, 512
- fc 6
Dataset: ILSVRC-2012 ImageNet

- Object Scale
  - Candle
  - Oyster
  - Cannon
  - Spider Web
- Number of Instances
  - Lizard
  - Stocking
  - Mushroom
  - Strawberry
- Image Clutter
  - Compass
  - Racket
  - Minivan
  - Steel Drum
- Deformability
  - Canoe
  - Pill Bottle/Hinges Part
  - Monkey
- Amount of Texture
  - Slowdriver
  - Hatchet/Pool Table
  - Leopard
- Color Distinctiveness
  - Mug
  - Tank
  - Ant
  - Red Wine
- Shape Distinctiveness
  - Jigsaw Puzzle
  - Foreland
  - Lion
  - Bell
- Real-world Size
  - Orange
  - Laptop
  - Four-poster
  - Airliner

Images:
- Flamingo
- Cock
- Ruffed Grouse
- Quail
- Partridge
- Egyptian Cat
- Persian Cat
- Siamese Cat
- Tabby
- Lynx
- Dalmatian
- Keeshond
- Miniature Schnauzer
- Standard Schnauzer
- Giant Schnauzer
Dataset: ILSVRC-2012 ImageNet

Image classification annotations (1000 object classes)

<table>
<thead>
<tr>
<th>Year</th>
<th>Train images (per class)</th>
<th>Val images (per class)</th>
<th>Test images (per class)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ILSVRC2012-14</td>
<td>1,281,167 (732-1300)</td>
<td>50,000 (50)</td>
<td>100,000 (100)</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Year</th>
<th>Train images with bbox annotations (per class)</th>
<th>Train bboxes annotated (per class)</th>
<th>Val images with bbox annotations (per class)</th>
<th>Val bboxes annotated (per class)</th>
<th>Test images with bbox annotations</th>
</tr>
</thead>
<tbody>
<tr>
<td>ILSVRC2012-14</td>
<td>523,966 (91-1268)</td>
<td>593,173 (92-1418)</td>
<td>50,000 (50)</td>
<td>64,058 (50-189)</td>
<td>100,000</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>labels</th>
<th>Top-5</th>
<th>Top-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>80.43%</td>
<td>48.43%</td>
</tr>
<tr>
<td>1%</td>
<td>56.35%</td>
<td>25.35%</td>
</tr>
</tbody>
</table>

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 $\varepsilon_{vat}$ tuning

**VAT + EntMin**
- best VAT model with cross entropy minimization loss
- $\mathcal{W}_{EntMin} = \{0, 0.03, 0.1, 0.3, 1\}$
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
S⁴L Model:

- S⁴L-Rotation
- S⁴L-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

<table>
<thead>
<tr>
<th></th>
<th>labels</th>
<th>Top-5</th>
<th>Top-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOAM (proposed)</td>
<td>10%</td>
<td>88.80</td>
<td>69.73</td>
</tr>
</tbody>
</table>
2. Retraining on Pseudo Labels
Pseudo labels for full dataset generated using Rotation+VAT+EntMin model

<table>
<thead>
<tr>
<th></th>
<th>labels</th>
<th>Top-5</th>
<th>Top-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOAM (proposed)</td>
<td>10%</td>
<td>88.80</td>
<td>69.73</td>
</tr>
<tr>
<td>MOAM + pseudo label (proposed)</td>
<td>10%</td>
<td>89.96</td>
<td>71.56</td>
</tr>
</tbody>
</table>
MOAM: Mix of all Models

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

<table>
<thead>
<tr>
<th>Model</th>
<th>labels</th>
<th>Top-5</th>
<th>Top-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOAM (proposed)</td>
<td>10%</td>
<td>88.80</td>
<td>69.73</td>
</tr>
<tr>
<td>MOAM + pseudo label (proposed)</td>
<td>10%</td>
<td>89.96</td>
<td>71.56</td>
</tr>
<tr>
<td>MOAM full (proposed)</td>
<td>10%</td>
<td>91.23</td>
<td>73.21</td>
</tr>
</tbody>
</table>

Set new state-of-the-art
Experimental Results

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

- 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
<table>
<thead>
<tr>
<th>Method</th>
<th>labels</th>
<th>Top-5</th>
<th>Top-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOAM full (proposed)</td>
<td>10%</td>
<td>91.23</td>
<td>73.21</td>
</tr>
<tr>
<td>MOAM + pseudo label (proposed)</td>
<td>10%</td>
<td>89.96</td>
<td>71.56</td>
</tr>
<tr>
<td>MOAM (proposed)</td>
<td>10%</td>
<td>88.80</td>
<td>69.73</td>
</tr>
<tr>
<td>ResNet50v2 (4×wider)</td>
<td>10%</td>
<td>81.29</td>
<td>58.15</td>
</tr>
<tr>
<td>VAE + Bayesian SVM [30]</td>
<td>10%</td>
<td>64.76</td>
<td>48.41</td>
</tr>
<tr>
<td>Mean Teacher [38]</td>
<td>10%</td>
<td>90.89</td>
<td>-</td>
</tr>
<tr>
<td>†UDA [40]</td>
<td>10%</td>
<td>88.52</td>
<td>68.66</td>
</tr>
<tr>
<td>†CPCv2 [12]</td>
<td>10%</td>
<td>84.88</td>
<td>64.03</td>
</tr>
</tbody>
</table>

*Training with all labels:*

<table>
<thead>
<tr>
<th>Method</th>
<th>labels</th>
<th>Top-5</th>
<th>Top-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet50v2 (4×wider)</td>
<td>100%</td>
<td>94.10</td>
<td>78.57</td>
</tr>
<tr>
<td>MOAM (proposed)</td>
<td>100%</td>
<td>94.97</td>
<td>80.17</td>
</tr>
<tr>
<td>†UDA [40]</td>
<td>100%</td>
<td>94.45</td>
<td>79.04</td>
</tr>
<tr>
<td>†CPCv2 [12]</td>
<td>100%</td>
<td>93.35</td>
<td>-</td>
</tr>
</tbody>
</table>

\[†\] marks concurrent work.
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 1000 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