Unsupervised Pre-Training: Contrastive Learning

CS 330
Course Reminders

Project proposal due Monday.
(graded lightly, for your benefit)

Homework 2 due next Weds 10/25.
So Far

Few-shot learning via meta-learning

Problem: Given data from $\mathcal{T}_1, \ldots, \mathcal{T}_n$, solve new task $\mathcal{T}_\text{test}$ more quickly / proficiently / stably

Methods: black-box, optimization-based, non-parametric

What if you don’t have a lot of tasks?
What if you only have one batch of unlabeled data?
This Lecture

Unsupervised representation learning for few-shot learning

Part I: Contrastive learning

Part II (next time): Reconstruction-based methods

Relation to meta-learning.

Goals for the lecture:
- Understand contrastive learning: intuition, design choices, how to implement
- How contrastive learning relates to meta-learning
Unsupervised Pre-Training Set-Up

Goal: Get predictor for task $\mathcal{T}_j$

- Diverse unlabeled dataset $\{x_i\}$
- Pre-trained model

Unsupervised pre-training

Fine-tuning

Labeled $\mathcal{D}_{tr}^j$
Key Idea of Contrastive Learning

**Similar examples should have similar representations**

1. Select or generate examples that are semantically similar
2. Train an encoder where similar examples are closer in representation space than non-similar examples.
Key Idea of Contrastive Learning

**Similar examples should have similar representations**

Question: Why not simply minimize difference between representations?

$$\min_{\theta} \sum_{(x_i, x_j)} \| f_{\theta}(x_i) - f_{\theta}(x_j) \|^2$$

Need to both compare & contrast!
Key Idea of Contrastive Learning

**Similar examples** should have **similar representations**

Need to both compare & **contrast**!

- **Bring together** representations of similar examples.
- **Push apart** representations of differing examples.

**Key design choices:**
1. Choosing what to compare/contrast
2. Implementation of contrastive loss
Key Idea of Contrastive Learning

*Similar examples should have similar representations*

Examples with the same class label

Augmented versions of the example

Nearby image patches

Nearby video frames

(Requires labels, related to Siamese nets, ProtoNets)

(Flip & crop)

Chen, Kornblith, Norouzi, Hinton. SimCLR. ICML 2020

van den Oord, Li, Vinyals. CPC. 2018

Dog credit to Maggie & Luke
Contrastive Learning Implementation

**Similar examples** should have **similar representations**

Need to both compare & **contrast**!

V1. Triplet loss:

\[
\min_{\theta} \sum_{(x,x^+,x^-)} \max(0, \|f_\theta(x) - f_\theta(x^+)\|^2 - \|f_\theta(x) - f_\theta(x^-)\|^2 + \epsilon)
\]

Compare to Siamese networks:

Classify \((x, x')\) as same class if \(\|f(x) - f(x')\|^2\) is small.

Key difference: learns a metric space, not just a classifier

**Challenge**: need to find difficult negatives.
**Contrastive Learning Implementation**

*Similar examples* should have *similar representations*

Need to both *compare* & *contrast*!

V2. From binary to N-way classification:

$$
\mathcal{L}_{N\text{-way}}(\theta) = - \sum_z \log \frac{\exp(-d(z, z^+))}{\sum_i \exp(-d(z, z_i^-)) + \exp(-d(z, z^+))}
$$

- generalization of triplet loss to multiple negatives

Sohn. N-Pair Loss Objective. NIPS 2016
Chen, Kornblith, Norouzi, Hinton. SimCLR. ICML 2020
Contrastive Learning Implementation

SimCLR Algorithm

Unsupervised Pre-Training

1. Sample minibatch of examples $x_1, \ldots, x_N$
2. Augment each example twice to get $\tilde{x}_1, \ldots, \tilde{x}_N, \tilde{x}_{N+1}, \ldots, \tilde{x}_{2N}$
3. Embed examples with $f_\theta$ to get $\tilde{z}_1, \ldots, \tilde{z}_N, \tilde{z}_{N+1}, \ldots, \tilde{z}_{2N}$
4. Compute all pairwise distances $d(z_i, z_j) = -\frac{z_i^T z_j}{\|z_i\| \|z_j\|}$
5. Update $\theta$ w.r.t. loss $\mathcal{L}_\text{N-way}(\theta) = -\sum_i \log \frac{\exp(-d(\tilde{z}_i, \tilde{z}_{N+i}))}{\sum_{j\neq i} \exp(-d(\tilde{z}_i, \tilde{z}_j))}$

After Pre-Training: train classifier on top of representation or fine-tune entire network.

Chen, Kornblith, Norouzi, Hinton. SimCLR. ICML 2020
## Performance of Contrastive Learning

### ImageNet Classification Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Architecture</th>
<th>Label fraction</th>
<th>1% Top 5</th>
<th>10% Top 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised baseline</td>
<td>ResNet-50</td>
<td>48.4</td>
<td>80.4</td>
<td></td>
</tr>
<tr>
<td><strong>Methods using other label-propagation:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pseudo-label</td>
<td>ResNet-50</td>
<td>51.6</td>
<td>82.4</td>
<td></td>
</tr>
<tr>
<td>VAT+Entropy Min.</td>
<td>ResNet-50</td>
<td>47.0</td>
<td>83.4</td>
<td></td>
</tr>
<tr>
<td>UDA (w. RandAug)</td>
<td>ResNet-50</td>
<td>-</td>
<td>88.5</td>
<td></td>
</tr>
<tr>
<td>FixMatch (w. RandAug)</td>
<td>ResNet-50</td>
<td>-</td>
<td>89.1</td>
<td></td>
</tr>
<tr>
<td>S4L (Rot+VAT+En. M.)</td>
<td>ResNet-50 (4×)</td>
<td>-</td>
<td>91.2</td>
<td></td>
</tr>
<tr>
<td><strong>Methods using representation learning only:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>InstDisc</td>
<td>ResNet-50</td>
<td>39.2</td>
<td>77.4</td>
<td></td>
</tr>
<tr>
<td>BigBiGAN</td>
<td>RevNet-50 (4×)</td>
<td>55.2</td>
<td>78.8</td>
<td></td>
</tr>
<tr>
<td>PIRL</td>
<td>ResNet-50</td>
<td>57.2</td>
<td>83.8</td>
<td></td>
</tr>
<tr>
<td>CPC v2</td>
<td>ResNet-161(*)</td>
<td>77.9</td>
<td>91.2</td>
<td></td>
</tr>
<tr>
<td>SimCLR (ours)</td>
<td>ResNet-50</td>
<td>75.5</td>
<td>87.8</td>
<td></td>
</tr>
<tr>
<td>SimCLR (ours)</td>
<td>ResNet-50 (2×)</td>
<td>83.0</td>
<td>91.2</td>
<td></td>
</tr>
<tr>
<td>SimCLR (ours)</td>
<td>ResNet-50 (4×)</td>
<td><strong>85.8</strong></td>
<td><strong>92.6</strong></td>
<td></td>
</tr>
</tbody>
</table>

1% labels: ~12.8 images/class

- Substantial improvements over training from scratch
- Improvements over other methods, especially in 1% label setting

*Table 7. ImageNet accuracy of models trained with few labels.*
Performance of Contrastive Learning

Effect of Batch Size & Number of Training Epochs

- Important to train for longer (~600+ epochs)
- Requires large batch size

Chen, Kornblith, Norouzi, Hinton. SimCLR. ICML 2020
Why does contrastive learning need a large batch size?

Interpretation of loss: classifying augmented example from rest of dataset

$$\mathcal{L}_{N\text{-way}}(\theta) = - \sum_i \log \frac{\exp(-d(\tilde{z}_i, \tilde{z}_{N+i}))}{\sum_{j \neq i} \exp(-d(\tilde{z}_i, \tilde{z}_j))} \leftarrow \text{summation over entire dataset}$$

Intuition: Closest $z$ will dominate the denominator, can be missed when subsampling

Mathematically?
I’m minimizing a bound on the objective.

An upper bound, right?

An upper bound, right?
Why does contrastive learning need a large batch size?

Interpretation of loss: classifying augmented example from rest of dataset

\[ \mathcal{L}_{N \text{-way}}(\theta) = - \sum_i \log \frac{\exp(-d(\tilde{z}_i, \tilde{z}_i'))}{\sum_{j \neq i} \exp(-d(\tilde{z}_i, \tilde{z}_j))} \leftarrow \text{summation over entire dataset} \]

Intuition: Closest \( z \) will dominate the denominator, can be missed when subsampling

Mathematically: Minimizing a lower-bound. 😱
Solutions to requiring a large batch size

1. **Store representations from previous batches** ("momentum contrast")
   - Good results with mini batch size of 256
   - He, Fan, Wu, Xie, Girshick. MoCo. CVPR 2020

2. **Predict representation of same image under different augmentation** ("BYOL")
   - No negatives required!
   - More resilient to batch size

![Graph showing decrease of accuracy from baseline vs. batch size](image)
Performance of contrastive learning

Contrastive methods are state-of-the-art in self-supervised pre-training for visual data.

ImageNet Top 1 Accuracy w/ Self-Supervised Pre-Training

Plot source: paperswithcode.com
Contrastive learning beyond augmentations

We don’t have good engineered augmentations for many applications!

1. **Learn** the augmentations in adversarial manner (but perturbations bounded to $\ell_1$ sphere)
   Tamkin, Wu, Goodman. Viewmaker Networks. ICLR 2021

   —> competitive with SimCLR on image data
   —> good results on speech & sensor data

2. **Time-contrastive learning** on videos effective for robotics pre-training
   Nair, Rajeswaran, Kumar, Finn, Gupta. R3M. CoRL 2022.

Given 20 demos (<10 min of supervision)

- 60% success
- 40% success
Contrastive learning beyond augmentations

We don’t have good engineered augmentations for many applications!

1. **Learn** the augmentations in adversarial manner (but perturbations bounded to $\ell_1$ sphere)
   Tamkin, Wu, Goodman. Viewmaker Networks. ICLR 2021

2. **Time-contrastive learning** on videos effective for robotics pre-training
   Nair, Rajeswaran, Kumar, Finn, Gupta. R3M. CoRL 2022.

3. **Image-text** contrastive pre-training produces robust zero-shot models
Summary of Contrastive Learning

Pros:
+ General, effective framework
+ No generative modeling required
+ Can incorporate domain knowledge through augmentations

Challenges:
- Negatives can be hard to select
- Often requires large batch size
- Most successful with augmentations
This Lecture

Unsupervised representation learning for few-shot learning

Part I: Contrastive learning
Part II (next time): Reconstruction-based methods

Relation to meta-learning.
Contrastive Learning as Meta-Learning

Meta-learning algorithm

1. Given unlabeled dataset \( \{x_i\} \).
2. Create image class \( y_i \) from each datapoint via data augmentation \( \mathcal{D}_i := \{\tilde{x}_i, \tilde{x}_i', \ldots \} \)
3. Run your favorite meta-learning algorithm.

Differences:
- SimCLR samples **one task** per minibatch; meta-learning usually samples **multiple**
- SimCLR compares **all pairs** of samples; meta-learning compares query examples only to support examples & not to other query examples.
Contras S

**Contrastive Learning as Meta-Learning**

**Meta-learning algorithm**

1. Given unlabeled dataset \( \{ x_i \} \).

2. Create image class \( y_i \) from each datapoint via data augmentation \( \mathcal{D}_i := \{ \tilde{x}_i, \tilde{x}_i', \ldots \} \)

3. Run your favorite meta-learning algorithm.

**Contrastive vs. meta-learning representations, transfer from ImageNet**

<table>
<thead>
<tr>
<th></th>
<th>Flowers102</th>
<th>DTD</th>
<th>VOC2007</th>
<th>Aircraft</th>
<th>Food101</th>
<th>SUN397</th>
<th>CIFAR-10</th>
<th>CIFAR-100</th>
</tr>
</thead>
<tbody>
<tr>
<td>SimCLR</td>
<td>92.4</td>
<td>72.7</td>
<td>66.0</td>
<td>83.7</td>
<td>86.3</td>
<td>57.4</td>
<td>94.8</td>
<td>79.1</td>
</tr>
<tr>
<td>ProtoNet</td>
<td>92.7</td>
<td>71.5</td>
<td>64.7</td>
<td>83.9</td>
<td>86.2</td>
<td>56.4</td>
<td>96.0</td>
<td>79.1</td>
</tr>
<tr>
<td>R2-D2</td>
<td><strong>94.5</strong></td>
<td><strong>73.8</strong></td>
<td><strong>69.9</strong></td>
<td><strong>86.2</strong></td>
<td><strong>86.9</strong></td>
<td><strong>59.7</strong></td>
<td><strong>96.7</strong></td>
<td><strong>82.8</strong></td>
</tr>
</tbody>
</table>

Representations transfer similarly well.
Lecture Outline

Unsupervised representation learning for few-shot learning

Part I: Contrastive learning

Part II (next time): Reconstruction-based methods

Relation to meta-learning.

Goals for the lecture:
- Understand contrastive learning: intuition, design choices, how to implement
- How contrastive learning relates to meta-learning
Course Reminders

Project proposal due Monday.
(graded lightly, for your benefit)

Homework 2 due next Weds 10/25.