Self-supervised learning in computer vision

Ishan Misra

FAIR, Meta AI
Supervised Learning

Input → Model (ConvNet) → Prediction (Dog, Tree, Table) → Label (Update)

Input Image: A beach scene with a tree, followed by the diagram showing the process of supervised learning with a ConvNet model predicting labels for different categories.
So what is the bottleneck?

- Supervision!!
- Getting "real" labels is difficult
Can we get labels for all data?
Can we get labels for all data?

ImageNet (14 million images) needed 22 human years to label
Can we get labels for all data?

• What about complex concepts?
  • Video?
• Labelling cannot scale to the size of the data we generate
Rare concepts?

10% of the classes account for 93% of the data

Slide credit: Rob Fergus
Different Domains?

Labeled data can be hard to obtain
Other Limitations of Supervised Learning
Commercial supervised AI models

Does Object Recognition Work For Everyone? - DeVries, Misra, Wang, van der Maaten
Commercial supervised AI models

Commercial supervised AI models

What is "self" supervision?

- Obtain "labels" from the data itself by using a "semi-automatic" process
- Predict part of the data from other parts
In the context of Computer Vision
Pretext task

- Self-supervised task used for learning representations
- Often, not the "real" task (like image classification) we care about

Pretext task - Doersch et al., 2015, Unsupervised visual representation learning by context prediction
Pretext task

- Using images
- Using video
- Using video and sound

Observed data

Hidden data

Hidden property of the data
Relative Position of patches

Input: Two patches
Output: 8-way classification

Doersch et al., 2015, Unsupervised visual representation learning by context prediction
Jigsaw puzzles
(Noorozi & Favaro, 2016)

Input: nine patches
Permute using one of N permutations

Output: N-way classification

Set N << 9!
Predicting Rotations

**Input**: image rotated by [0, 90, 180, 270]

**Output**: 4-way classification

Gidaris et al., 2018, Predicting Image Rotations
What is missing from "pretext" tasks?
Or in general "proxy" tasks
The hope of generalization

- We really **hope** that the pre-training task and the transfer task are "aligned"
The hope of generalization ... ?

Pre-training
Weak or self-supervised

Pre-train data
ConvN
Jigsaw

Transfer
Linear classifiers on "fixed" features
ConvNet
Higher layers do not generalize ...

Linear classifier on VOC07

mAP = mean Average Precision (Higher is better)
Pre-trained features should ...

• Represent how images relate to one another

• Be robust to "nuisance factors" -- Invariance
  • e.g., exact location of objects, lighting, exact color
Learn features such that:

\[ f_\theta(I) = f_\theta(\text{augment}(I)) \]

Figure from Dosovitskiy et al., 2014
Learn features such that:

\[ f_\theta(I) = f_\theta(\text{augment}(I)) \]

Learned features are invariant to "nuisance factors" or data augmentation
Can it work?

\[ f_\theta(I) = f_\theta(\text{augment}(I)) \]
Trivial Solutions

\[ f_\theta(I) = f_\theta(\text{augment}(I)) \]

\[ f_\theta(I) = \text{constant} \]

Satisfies the invariance property, but not useful
Categorization of recent self-supervised methods
Many ways to avoid trivial solutions

**Similarity Maximization Objective**
- Contrastive learning
  - MoCo, PIRL, SimCLR
- Clustering
  - DeepCluster, SeLA, SwAV
- Distillation
  - BYOL, SimSiam, DINO

**Redundancy Reduction Objective**
- Redundancy Reduction
  - Barlow Twins, VICReg
Many ways to avoid trivial solutions

**Similarity Maximization Objective**
- Contrastive learning
  - MoCo, PIRL, SimCLR
- Clustering
  - DeepCluster, SeLA, SwAV
- Distillation
  - BYOL, SimSiam

**Redundancy Reduction Objective**
- Redundancy Reduction
  - Barlow Twins
Pretraining

- ImageNet without labels (1.3M images)
- ResNet-50 initialized randomly
Evaluation using Transfer Learning

Transfer to downstream task
- Train a linear classifier on frozen features
- Full finetuning of the network
Evaluation – fine-tuning vs. linear classifier vs. kNN

Fine-tune all layers

Linear classifier

kNN
Is this representation learning
OR
learning a good initialization?
The great spiral of research

Pre 2015 - Sparse encoding, RBMs, contrastive

2015 - Pretext

2018/19 - Invariance using Contrastive

2020 - Invariance using non-contrastive

2021 - Pretext tasks are cool again
Pretext-Invariant Representation Learning (PIRL)

Ishan Misra, Laurens van der Maaten
Contrastive Learning

Groups of Related and Unrelated Images
Contrastive Learning

Groups of Related and Unrelated Images

Shared network (Siamese Net)

Image Features (Embeddings)
Contrastive Learning

Related and Unrelated Images

Shared network (Siamese)

Image Features (Embeddings)

Loss Function
Embeddings from related images should be closer than embeddings from unrelated images

\[ d(\text{related images}) < d(\text{unrelated images}) \]

Hadsell et al., 2005, DrLim
Be similar

\[ L_{\text{contrastive}}(v_I, v_{I^t}) \]

Invariant to Pretext transform

Invariance to
- Data Augmentations
- Multiple views created by pretext task (Jigsaw/Rotation)
Contrastive Learning in PIRL

Dataset

Loss Function

\[ d(I) < d(I^t) \]
\[ d(I) < d(I') \]

Image Feature & Patch Features

Random Images
Semantic Features?

Accuracy on ImageNet-1K
(Linear classifier on fixed features)
Nearby patches vs. distant patches of an Image

Related (Positives) vs. Unrelated (Negative)

van der Oord et al., 2018, Henaff et al., 2019
Contrastive Predictive Coding
Patches of an image vs. patches of other images

Related (Positives)
- Wu et al., 2018, Instance Discrimination
- He et al., 2019, MoCo
- Misra & van der Maaten, 2019, PIRL
- Chen et al., 2020, SimCLR

Unrelated (Negative)
Frames of a video

Time

“Sequence” of data

Hadsell et al., 2005, DrLim
van der Oord et al., 2018, CPC

Video & Audio

AVID+CMA - Morgado et al., 2020
GDT - Patrick et al., 2020
Tracking Objects

Wang & Gupta, 2015, Unsupervised Learning of Visual Representations using Videos
3D Point Clouds

Augmentations

DepthContrast - Zhang et al., ICCV 2021
PointContrast Xie et al., CVPR 2020
Good negatives are necessary

Loss Function
Embeddings from related images should be closer than embeddings from unrelated images

\[ d(\text{Positive (Related)}) < d(\text{Negative (Unrelated)}) \]

Good negatives are very important in contrastive learning

Hadsell et al., 2005, DrLim
SimCLR

- Large batch size - e.g. in SimCLR
- Pros - Simple to implement
- Cons - Large batch size

A simple framework for contrastive learning - Chen et al., 2020
Memory Bank

- Maintain a "memory bank" -- momentum of activations
- Pros - compute efficient
- Cons - Needs large memory, not "online"

Moving average of features

Non parametric Instance Discrimination - Wu et al., 2018
MoCo

- Maintain "momentum" network - MoCo
- Pros - online
- Cons - extra memory for parameters/stored features, extra fwd pass compared to memory bank
Many ways to avoid trivial solutions

**Similarity Maximization Objective**
- Contrastive learning
  - MoCo, PIRL, SimCLR
- Clustering
  - DeepCluster, SeLA, SwAV
- Distillation
  - BYOL, SimSiam

**Redundancy Reduction Objective**
- Redundancy Reduction
  - Barlow Twins
Contrastive learning -- what does it do?

Positive Sample

Negative samples

Negative samples
Contrastive learning -- what does it do?

Negative samples

Positive Sample
Contrastive Learning $\Rightarrow$ Groups in feature space

Creates groups in the feature space
Clustering creates groups too

Creates groups in the feature space

So does clustering?!
Swapping Assignments between Views (SwAV)

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, Armand Joulin
Key Idea

• What we want
  \[ f_\theta(I) = f_\theta(\text{augment}(I)) \]

• How we do it - I and augment(I) must belong to the same "group" or cluster

• Prevent trivial solutions by controlling the clustering process
Grouping

Prototypes

Similarity of dataset sample & prototypes

(which cluster does a sample belong to?)

See also - SeLa by Asano et al., 2019
Grouping

Dataset

Prototypes
Trivial Solutions?

Embeddings

Prototypes
Equipartition constraint --
Given N samples and K prototypes, each prototype is most similar to N/K samples

Implemented using Optimal Transport (Sinkhorn-Knopp)
Soft Assignment

Embeddings

Prototypes

Codes
Prototypes

Code 1

Embeddings

Code 2
Prototypes

Backprop

Not contrastive!

Embeddings

Code 1

Code 2
## Key Results

<table>
<thead>
<tr>
<th></th>
<th>Linear Classifier (Fixed Features)</th>
<th>Detection (Fine-tuned)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ImageNet</td>
<td>Places</td>
</tr>
<tr>
<td>Supervised</td>
<td>76.5</td>
<td>53.2</td>
</tr>
<tr>
<td>Prior self-supervised</td>
<td>71.1 ((-5.4))</td>
<td>52.1</td>
</tr>
<tr>
<td>SwAV</td>
<td>75.3 ((-1.2))</td>
<td><strong>56.7</strong></td>
</tr>
</tbody>
</table>
Pretrained on ImageNet without labels
Curated pretraining data

ImageNet data is "curated"
- All images belong to 1000 classes
- All images contain a prominent object
- Very limited clutter
Curated pretraining data

Pretraining on non-ImageNet data hurts performance
Real world data

Images have
- different distributions (cartoon images, memes)
- no single prominent object
SEER: Learning from uncurated images

Train on 1.3 billion random images
Images are NOT filtered in any way

SEER - Goyal et al., 2021
SEER: Improves performance on benchmarks

SEER - Goyal et al., 2021

SEER - Self-supervised vision model on 1 billion random internet images. **No Labels/metadata.**
SEER: AI that works for everyone

**Spices** (Nepal)
Supervised - cleaning equipment, kitchen sink, shower
SEER - spices, medication, bowls

**Stove** (China)
Supervised - lock on front door, power switches, cooking utensils
SEER - cooking utensils, stove

SEER - Goyal et al., 2021
Audio Visual Instance Discrimination with Cross Modal Agreement (AVID + CMA)

Pedro Morgado, Nuno Vasconcelos, Ishan Misra

https://github.com/facebookresearch/AVID-CMA
Contrastive (Audio Video Instance Discrimination)

Positives

\[ d(\text{blue}, \text{blue}) < d(\text{blue}, \text{green}) \]
\[ d(\text{blue}, \text{blue}) < d(\text{blue}, \text{purple}) \]

Audio & Video (same sample)

Negatives

Relate to other video/audio using negatives
Grouping using Audio-visual Agreements (CMA)

Video Similarity ($v^T v$)

Audio Similarity ($a^T a$)

Reference

Positives

Visual Negatives

Audio Negatives

Positive Set

Negative Set

Positives

Negatives

$d(\text{blue} \text{ blue}) < d(\text{blue} \text{ green})$

$d(\text{green} \text{ blue}) < d(\text{green} \text{ purple})$

Videos that are similar in audio & video features
Grouping using Audio-visual Agreements (CMA)

Video Similarity ($v_i v_j$)

Audio Similarity ($a_i a_j$)

Reference

Positives

Visual Negatives

Audio Negatives

Positive Set

Negative Set

Example 1
Dancing

Example 2
Playing Violin

Example 3
Moving Train

Dancing

Playing Violin

Moving Train

Playing Guitar

Fire Truck Station

Fishing with background music

Playing Accordion

Moving Boat
Many ways to avoid trivial solutions

**Similarity Maximization Objective**
- Contrastive learning
  - MoCo, PIRL, SimCLR
- Clustering
  - DeepCluster, SeLA, SwAV
- Distillation
  - BYOL, SimSiam, DINO

**Redundancy Reduction Objective**
- Redundancy Reduction
  - Barlow Twins
Distillation

• What we want \[ f_\theta(I) = f_\theta(\text{augment}(I)) \]

• How we do it \[ f_\theta^{\text{student}}(I) = f_\theta^{\text{teacher}}(\text{augment}(I)) \]

• Prevent trivial solutions by asymmetry
  • Asymmetric **learning rule** between student teacher
  • Asymmetric **architecture** between student teacher
BYOL

- **What we want**
  \[
  f_\theta(I) = f_\theta(\text{augment}(I))
  \]

- **How we do it**
  \[
  f_{\text{student}}(I) = f_{\text{teacher}}(\text{augment}(I))
  \]

Diagram:
- Student encoder
- Teacher encoder
- Image
- BYOL - Grill et al., 2020
SimSiam

• What we want

\[ f_\theta(I) = f_\theta(\text{augment}(I)) \]
DINO - Distillation with No Labels

Mathilde Caron, Hugo Touvron, Ishan Misra, Herve Jegou, Julien Marial, Piotr Bojanowski, Armand Joulin

https://github.com/facebookresearch/dino
DINO - Main idea

Image

encode

Softmax

similarity

mom. encode

moving average

grad

Softmax
Type of encoder - Vision Transformer

Vision Transformer (ViT)

Transformer Encoder

MLP Head

Class
Bird
Ball
Car...

Patch + Position Embedding
* Extra learnable [class] embedding

Linear Projection of Flattened Patches

Vision Transformer - Dosovitskiy et al., 2020
No pooling!

Feature maps in CNN

Feature maps in ViT

Source: https://towardsdatascience.com/vision-transformers-or-convolutional-neural-networks-both-de1a2c3c62e4
Segmentation emerges!

Visualize the “CLS” token attention.

Note that the CLS token or the network are not supervised
Segmentation across different heads
Many ways to avoid trivial solutions

Similarity Maximization Objective
• Contrastive learning
  • MoCo, PIRL, SimCLR
• Clustering
  • DeepCluster, SeLA, SwAV
• Distillation
  • BYOL, SimSiam

Redundancy Reduction Objective
• Redundancy Reduction
  • Barlow Twins, VICReg
Barlow Twins: Self-supervised Learning via Redundancy Reduction

Jure Zbontar*, Li Jing*, Ishan Misra, Yann LeCun, Stéphane Deny

https://github.com/facebookresearch/barlowtwins
Horace Barlow's Efficient Coding Hypothesis

• Inspired by Information Theory
• Neurons communicate via "spiking codes"
• Spiking codes aim to reduce redundancy between neurons
Redundancy Reduction

• N neurons produce a representation: N dimensional feature
• Each neuron should satisfy
  • Invariance -- be invariant under different data augmentation
  • Independent of other neurons -- reduce redundancy

• VERY roughly speaking

$$f_{\theta}(I)[i] = f_{\theta}(\text{augment}(I))[i]$$

$$f_{\theta}(I)[i] \neq f_{\theta}(\text{augment}(I))[j]$$
Barlow Twins

Images \( X \) are distorted by \( T \sim T \), then passed through a neural network \( f_\theta \) to produce representations \( Z^A \) and \( Z^B \). Barlow Twins - Zbontar et al., 2020
Barlow Twins - Invariance

Barlow Twins - Zbontar et al., 2020
Barlow Twins - Redundancy Reduction

Barlow Twins - Zbontar et al., 2020
Barlow Twins
Barlow Twins - Loss

Barlow Twins - Zbontar et al., 2020
Barlow Twins - Loss

Barlow Twins - Zbontar et al., 2020
Barlow Twins Objective Function

\[
C_{ij} \triangleq \frac{\sum_b z^A_{b,i} z^B_{b,j}}{\sqrt{\sum_b (z^A_{b,i})^2} \sqrt{\sum_b (z^B_{b,j})^2}}
\]

\[
\mathcal{L}_{BT} \triangleq \sum_i (1 - C_{ii})^2 + \lambda \sum_i \sum_{j \neq i} C_{ij}^2
\]

- \text{invariance term}
- \text{redundancy reduction term}
Trivial Solutions?

Center $Z^A$ and $Z^B$ before computing cross-correlation.
Prevents trivial solutions without

- "Negatives" like in contrastive learning
Prevents trivial solutions without Asymmetric Learning

SwAV - Caron et al., 2020
BYOL - Grill et al., 2020
SimSiam - Chen & He, 2020

Redundant
Barlow Twins
The great spiral of research

Pre 2015 - Sparse encoding, RBMs, contrastive

2015 - Pretext

2018/19 - Invariance using Contrastive

2020 - Invariance using non-contrastive

2021 - Pretext tasks are cool again
BeIT: BERT Pre-Training of Image Transformers

Hangbo Bao, Li Dong, Furu Wei
BelIT

Original Image → Patches → Masked Patches
BelIT: Masked Prediction Problem
Masked Prediction: Vision & NLP

Masked Patches → Model → Patches

Masked Sentence → Model → A sunny day

A _ day
He et al.,
Masked Autoencoders Are Scalable Vision Learners

Xie et al.,
A Simple Framework for Masked Image Modeling
How to evaluate?

- Fine-tune all layers
- Linear classifier
- kNN
## Full finetuning

<table>
<thead>
<tr>
<th>Models</th>
<th>Model Size</th>
<th>Image Size</th>
<th>ImageNet</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training from scratch (i.e., random initialization)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ViT$_{384}$-B (<a href="#">Dosovitskiy et al., 2020</a>)</td>
<td>86M</td>
<td>$384^2$</td>
<td>77.9</td>
</tr>
<tr>
<td>ViT$_{384}$-L (<a href="#">Dosovitskiy et al., 2020</a>)</td>
<td>307M</td>
<td>$384^2$</td>
<td>76.5</td>
</tr>
<tr>
<td>DeiT-B (<a href="#">Touvron et al., 2020</a>)</td>
<td>86M</td>
<td>$224^2$</td>
<td>81.8</td>
</tr>
<tr>
<td>DeiT$_{384}$-B (<a href="#">Touvron et al., 2020</a>)</td>
<td>86M</td>
<td>$384^2$</td>
<td>83.1</td>
</tr>
<tr>
<td><strong>Supervised Pre-Training on ImageNet-22K (using labeled data)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ViT$_{384}$-B (<a href="#">Dosovitskiy et al., 2020</a>)</td>
<td>86M</td>
<td>$384^2$</td>
<td>84.0</td>
</tr>
<tr>
<td>ViT$_{384}$-L (<a href="#">Dosovitskiy et al., 2020</a>)</td>
<td>307M</td>
<td>$384^2$</td>
<td>85.2</td>
</tr>
<tr>
<td><strong>Self-Supervised Pre-Training on ImageNet-1K (without labeled data)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>iGPT-1.36B† (<a href="#">Chen et al., 2020a</a>)</td>
<td>1.36B</td>
<td>$224^2$</td>
<td>66.5</td>
</tr>
<tr>
<td>ViT$_{384}$-B-JFT300M‡ (<a href="#">Dosovitskiy et al., 2020</a>)</td>
<td>86M</td>
<td>$384^2$</td>
<td>79.9</td>
</tr>
<tr>
<td>DINO-B (<a href="#">Caron et al., 2021</a>)</td>
<td>86M</td>
<td>$224^2$</td>
<td>82.8</td>
</tr>
<tr>
<td>BEiT-B (ours)</td>
<td>86M</td>
<td>$224^2$</td>
<td>83.2</td>
</tr>
<tr>
<td>BEiT$_{384}$-B (ours)</td>
<td>86M</td>
<td>$384^2$</td>
<td>84.6</td>
</tr>
<tr>
<td>BEiT-L (ours)</td>
<td>307M</td>
<td>$224^2$</td>
<td>85.2</td>
</tr>
<tr>
<td>BEiT$_{384}$-L (ours)</td>
<td>307M</td>
<td>$384^2$</td>
<td><strong>86.3</strong></td>
</tr>
</tbody>
</table>
## Full finetuning - Segmentation

<table>
<thead>
<tr>
<th>Models</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised Pre-Training on ImageNet</td>
<td>45.3</td>
</tr>
<tr>
<td>DINO (Caron et al., 2021)</td>
<td>44.1</td>
</tr>
<tr>
<td>BEiT (ours)</td>
<td>45.6</td>
</tr>
<tr>
<td>BEiT + Intermediate Fine-Tuning (ours)</td>
<td><strong>47.7</strong></td>
</tr>
</tbody>
</table>
Masked Siamese Networks for Label-Efficient Learning

Mido Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski
Florian Bardes, Pascal Vincent, Armand Joulin, Mike Rabbat, Nicolas Ballas
Masked Siamese Networks

original

anchor view  patchify & mask

f₀

EMA

Self-distillation Loss

target view  patchify

f₀

f₀

prediction

p_
target

H(p, p_
)
Masked Siamese Networks

Original

Random Mask

Focal Mask
Label-efficient learning

Low-Shot Evaluation on ImageNet-1k

Evaluation on 1% ImageNet-1k

- Frozen Features
- Fine-Tuned
Robust representations

<table>
<thead>
<tr>
<th>Model</th>
<th>IN-A (top-1 ↑)</th>
<th>IN-R (top-1 ↑)</th>
<th>IN-Sketch (top-1 ↑)</th>
<th>IN-C (mCE ↓)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised ResNet50</td>
<td>0.04</td>
<td>36.11</td>
<td>24.2</td>
<td>76.7</td>
</tr>
<tr>
<td>MAE ViT-B/16 [22]</td>
<td>35.9</td>
<td>48.3</td>
<td>34.5</td>
<td>51.7</td>
</tr>
<tr>
<td>MSN ViT-B/16</td>
<td><strong>37.5</strong></td>
<td><strong>50.0</strong></td>
<td><strong>36.3</strong></td>
<td><strong>46.6</strong></td>
</tr>
</tbody>
</table>
Reconstructing images

Bordes et al.,
High Fidelity Visualization of What Your Self-Supervised Representation Knows About
arXiv, 2022.
Reconstructing images

Bordes et al.,
High Fidelity Visualization of What Your Self-Supervised Representation Knows About
arXiv, 2022.
Reconstructing images

Bordes et al.,
High Fidelity Visualization of What Your Self-Supervised Representation Knows About
arXiv, 2022.