Self-supervised learning

M.C. Escher, *Drawing Hands* (1948) – via A. Efros
Motivation

• Overcoming reliance on *supervised pre-training*
Self-supervised vs. unsupervised learning

- The terms are sometimes used interchangeably in the literature, but self-supervised learning is a particular kind of unsupervised learning

- **Self-supervised learning**: the learner “makes up” labels from the data and then solves a supervised task

- **Unsupervised learning**: any kind of learning without labels
  - Clustering and quantization
  - Dimensionality reduction, manifold learning
  - Density estimation
  - Learning to sample
Types of self-supervised learning

**Data prediction**
- Data $x$ → Network → Data $x'$

**Transformation prediction**
- Data $T(x)$ → Network → $T$

**Contrastive learning**
- Data $T(x)$ → Network → Similar/dissimilar
- Data $T'(x)$ → Network → Similar/dissimilar
Self-supervised learning: Outline

• Data prediction
  • Colorization
• Transformation prediction
  • Context prediction, jigsaw puzzle solving, rotation prediction
• Deep clustering and instance prediction
• Contrastive learning
  • PIRL, MoCo, SimCLR, SWaV
• Self-supervision beyond still images
  • Audio, video, language
Self-Supervision as data prediction

- Colorization
- Inpainting
- Cross-channel encoding
- Future prediction
Colorization

R. Zhang, P. Isola, and A. Efros, Colorful Image Colorization, ECCV 2016
Colorization: Architecture

At each spatial location, predict probability distribution over 313 quantized (a,b) values

Colorization: Results

R. Zhang, P. Isola, and A. Efros, Colorful Image Colorization, ECCV 2016
Failure Cases

Source: A. Efros, R. Zhang
Inherent Ambiguity

Grayscale

Source: A. Efros, R. Zhang
Inherent Ambiguity

Prediction

Ground Truth

Source: A. Efros, R. Zhang
Biases

Source: A. Efros, R. Zhang
Biases

Source: A. Efros, R. Zhang
Self-supervised learning: Outline

• Data prediction
  • Colorization

• Transformation prediction
Self-supervision by transformation prediction

- Context prediction
- Jigsaw puzzle solving
- Rotation prediction
Context prediction

- **Pretext task**: randomly sample a patch and one of 8 neighbors
- Guess the spatial relationship between the patches

**Question 1:**

A: Bottom right

**Question 2:**

A: Top center

Context prediction: Semantics from a non-semantic task

Source: A. Efros
Context prediction: Details

Prevent “cheating”: sample patches with gaps, pre-process to overcome chromatic aberration

Context prediction: Results

- Use learned weights in R-CNN model to perform detection on PASCAL VOC 2007
- Unsupervised pre-training is 5% mAP better than training from scratch, but still 8% below pre-training with ImageNet label supervision

Jigsaw puzzle solving

Crop out tiles

Shuffle

Pretext task: reassemble

Claim: jigsaw solving is easier than context prediction, trains faster, transfers better

Jigsaw puzzle solving: Details

Predetermined set of 1000 permutations (out of 362,880 possible)

Rotation prediction

- Pretext task: recognize image rotation (0, 90, 180, 270 degrees)
Rotation prediction

During training, feed in all four rotated versions of an image in the same mini-batch

S. Gidaris, P. Singh, and N. Komodakis. Unsupervised representation learning by predicting image rotations. ICLR 2018
### Rotation prediction: PASCAL VOC Transfer results

<table>
<thead>
<tr>
<th>Method</th>
<th>Classification</th>
<th>Detection (mAP)</th>
<th>Segmentation (mIoU)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised (ImageNet)</td>
<td>79.9</td>
<td>56.8</td>
<td>48.0</td>
</tr>
<tr>
<td>Colorization</td>
<td>65.6</td>
<td>46.9</td>
<td>35.6</td>
</tr>
<tr>
<td>Context</td>
<td>65.3</td>
<td>51.1</td>
<td></td>
</tr>
<tr>
<td>Jigsaw</td>
<td>67.6</td>
<td>53.2</td>
<td>37.6</td>
</tr>
<tr>
<td>Rotation</td>
<td>73.0</td>
<td>54.4</td>
<td>39.1</td>
</tr>
</tbody>
</table>
Self-supervised learning: Outline

• Data prediction
  • Colorization
• Transformation prediction
  • Context prediction, jigsaw puzzle solving, rotation prediction
• Deep clustering and instance prediction
Deep Clustering

- Cluster the features to obtain pseudo-labels
- Use pseudo-label prediction as pretext task to train the network
- Re-cluster the features, iterate

# Deep Clustering: PASCAL VOC Transfer results

<table>
<thead>
<tr>
<th>Method</th>
<th>Classification</th>
<th>Detection (mAP)</th>
<th>Segmentation (mIoU)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised (ImageNet)</td>
<td>79.9</td>
<td>56.8</td>
<td>48.0</td>
</tr>
<tr>
<td>Colorization</td>
<td>65.6</td>
<td>46.9</td>
<td>35.6</td>
</tr>
<tr>
<td>Context</td>
<td>65.3</td>
<td>51.1</td>
<td></td>
</tr>
<tr>
<td>Jigsaw</td>
<td>67.6</td>
<td>53.2</td>
<td>37.6</td>
</tr>
<tr>
<td>Rotation</td>
<td>73.0</td>
<td>54.4</td>
<td>39.1</td>
</tr>
<tr>
<td>DeepCluster</td>
<td>73.7</td>
<td>55.4</td>
<td>45.1</td>
</tr>
</tbody>
</table>

M. Caron, P. Bojanowski, A. Joulin, and M. Douze. Deep clustering for unsupervised learning of visual features. ECCV 2018
Instance prediction

- Key idea: make each instance into its own class

Self-supervised learning: Outline

- Data prediction
  - Colorization
- Transformation prediction
  - Context prediction, jigsaw puzzle solving, rotation prediction
- Deep clustering and instance prediction
- Contrastive learning
  - PIRL, MoCo, SimCLR, SWaV
Contrastive methods

- Encourage representations of transformed versions of the same image to be the same and different images to be different
Contrastive methods

- Encourage representations of transformed versions of the same image to be the same and different images to be different
Contrastive loss formulation

• Given: query point $x$, positive samples $x^+$, negative samples $x^-$
  • Positives are typically transformed versions of $x$, negatives are random examples from the same mini-batch or memory bank
  • Key idea: learn representation to make $x$ similar to $x^+$, dissimilar from $x^-$ (similarity is measured by dot product of normalized features)
  • Intuitively, contrastive loss for $x$, $x^+$ is the loss of a softmax classifier that tries to classify $x$ as $x^+$:
    \[
    l(x, x^+) = -\log \frac{\exp(f(x)^T f(x^+)/\tau)}{\exp(f(x)^T f(x^+)/\tau) + \sum_{j=1}^{N} \exp(f(x)^T f(x_j^-)/\tau)}
    \]
  • $\tau$ is the temperature hyperparameter (determines how concentrated the softmax is)
Momentum contrast

- Use instance discrimination as pretext task, transform query and key by random augmentations, use queue encoded by a momentum encoder instead of memory bank.
Pretext-invariant representation learning (PIRL)

- Key idea: instead of predicting the transformation of the input, learn a representation *invariant* to the transformation

I. Misra and L. van der Maaten. Self-Supervised Learning of Pretext-Invariant Representations. CVPR 2020
Pretext-invariant representation learning (PIRL)

I. Misra and L. van der Maaten. Self-Supervised Learning of Pretext-Invariant Representations. CVPR 2020
### Table 1: Object detection on VOC07+12 using Faster R-CNN.

Detection AP on the VOC07 test set after finetuning Faster R-CNN models (keeping BatchNorm fixed) with a ResNet-50 backbone pre-trained using self-supervised learning on ImageNet. Results for supervised ImageNet pre-training are presented for reference. Numbers with * are adopted from the corresponding papers. Method with † finetunes BatchNorm. PIRL significantly outperforms supervised pre-training without extra pre-training data or changes in the network architecture. Additional results in Table 6.
SimCLR

- Form two views of the input by composing data augmentations
  - Cropping and resizing, color distortion, blur

SimCLR

- Form two views of the input by composing data augmentations
  - Cropping and resizing, color distortion, blur
- No memory bank, large mini-batch size (on cloud TPU)
- Introduce nonlinear transformation between representation and contrastive loss (or, use representation a few layers below the contrastive loss)

SimCLR: Evaluation

No detection evaluation

Swapping Assignments Between Views (SWaV)

- Predict cluster assignment of one “view” (transformed version of input image) from representation of another “view”
  - Prototypes or cluster centers are learned online within mini-batch
  - Once again, data augmentation strategy matters

M. Caron et al. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments, arXiv 2020
SWaV: Results

The Gelato Bet

Made at Berkeley on September 23, 2014 between Alyosha Efros and Jitendra Malik:

“If, by the first day of autumn (Sept 23) of 2015, a method will exist that can match or beat the performance of R-CNN on Pascal VOC detection, without the use of any extra, human annotations (e.g. ImageNet) as pre-training, Mr. Malik promises to buy Mr. Efros one (1) gelato (2 scoops: one chocolate, one vanilla).”

https://people.eecs.berkeley.edu/~efros/gelato_bet.html
Self-supervised learning: Outline

- Data prediction
  - Colorization
- Transformation prediction
  - Context prediction, jigsaw puzzle solving, rotation prediction
- Deep clustering and instance prediction
- Contrastive learning
  - PIRL, MoCo, SimCLR, SWaV
- Self-supervision beyond still images
  - Video, audio, language
Learning from audio

Ego-motion features

D. Jayaraman and K. Grauman. Learning image representations tied to ego-motion. ICCV 2015
Future prediction

Future prediction

C. Finn and S. Levine. Deep Visual Foresight for Planning Robot Motion. ICRA 2017
Self-supervised learning in NLP (coming up)

- word2vec, GloVe, BERT, ELMO, GPT, …

Figure source
For further reading

https://github.com/jason718/awesome-self-supervised-learning