Learning Visual Representations

Neural Network → Feature representation → Classification, Clustering, Segmentation, Detection, Image captioning, Etc.
Generative Models

• Autoencoder

• Variational Autoencoder (VAE)
Discriminative Models (Supervised Learning)

Input image

Convolutional layer

ReLU layer

Pooling layer

Fully connected layer

Output vector

Feature representation

Classification
Supervised Representation Learning

- Train neural networks for image classification
- Use internal features in the network as feature representations
- Applications

Supervised Representation Learning

Image clustering

t-Distributed Stochastic Neighbor Embedding (t-SNE)


Supervised Representation Learning

• Training with classification loss functions
  • E.g., cross-entropy loss

• Can we have better loss functions for representation learning?

• Deep metric learning
  • Learning distance metrics with neural networks
Distance metrics

• **L1 distance**
  \[ D(x, y) = \sum_{i=1}^{N} |x_i - y_i| \]

• **L2 distance**
  \[ D(x, y) = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2} \]

• **Cosine distance**
  \[ D(x, y) = 1 - \frac{x \cdot y}{\|x\| \|y\|} \]
  Cosine similarity
Deep Metric Learning

Learning the distance metric is equivalent to learning the feature representation

\[ D(x_1, x_2) = D(f(x_1), f(x_2)) \]

L2 distance

\[ D(x_1, x_2) = \| f(x_1) - f(x_2) \|_2 \]
Contrastive Loss

• Use positive pairs and negative pairs

\[ f(x_1) \rightarrow x_1 \rightarrow \text{Neural Network} \]

\[ f(x_2) \rightarrow x_2 \rightarrow \text{Neural Network} \]

Positive pair \[ f(x_1) f(x_2) \] should be close

\[ D(x_1, x_2) \] small

Negative pair \[ f(x_1) f(x_2) \] should be far

\[ D(x_1, x_2) \] large

Learning a Similarity Metric Discriminatively, with Application to Face Verification. Chopra et al., CVPR, 2005.
Contrastive Loss

- **Training data**
  \[
  \{(x_i, x_j, y_{ij})\} \quad y_{ij} = \begin{cases} 
  1 & \text{if positive pair} \\
  0 & \text{if negative pair}
  \end{cases}
  \]

(a) Contrastive embedding

\[
J = \frac{1}{m} \sum_{(i,j)}^{m/2} y_{i,j} D_{i,j}^2 + (1 - y_{i,j}) [\alpha - D_{i,j}]_+^2
\]

\[
[x]_+ = \max(0, x)
\]

**Learning a Similarity Metric Discriminatively, with Application to Face Verification.** Chopra et al., CVPR, 2005.
Contrastive Loss

• Compute Gradient

\[ J = \frac{1}{m} \sum_{(i,j)}^{m/2} y_{i,j} D_{i,j}^2 + (1 - y_{i,j}) [\alpha - D_{i,j}]^2_+ \]

\[ \frac{\partial J}{\partial D_{i,j}} = \frac{2}{m} \left( y_{i,j} D_{i,j} - (1 - y_{i,j}) [\alpha - D_{i,j}]_+ \right) \]

\[ D_{i,j} = \| f(x_i) - f(x_j) \|_2 \]

\[ \frac{\partial D_{i,j}}{\partial f(x_i)} = \frac{f(x_i) - f(x_j)}{\| f(x_i) - f(x_j) \|} \]

\[ X_i \rightarrow \text{Neural Network} \rightarrow f(x_i) \]

\[ \frac{\partial J}{\partial f(x_i)} \]

Gradients
**Triplet Loss**

- Use a triplet (anchor, positive, negative)

\[
J = \frac{3}{2m} \sum_{i}^{m/3} \left[ D_{ia,ip}^2 - D_{ia,in}^2 + \alpha \right]_+
\]

\[
D_{ia,ip} = \| f(x^a_i) - f(x^p_i) \| \quad \quad D_{ia,in} = \| f(x^a_i) - f(x^n_i) \|
\]

Lifted Structured Loss

• Consider all positive pairs and negative pairs in a mini-batch

\[ J = \frac{1}{2|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} \max (0, J_{i,j})^2 \]

\[ J_{i,j} = \max \left( \max_{(i,k) \in \mathcal{N}} \alpha - D_{i,k}, \max_{(j,l) \in \mathcal{N}} \alpha - D_{j,l} \right) + D_{i,j} \]

Hard negative

Distance for the negative pair

Distance for the positive pair

Relaxed loss

\[ \tilde{J}_{i,j} = \log \left( \sum_{(i,k) \in \mathcal{N}} \exp\{\alpha - D_{i,k}\} + \sum_{(j,l) \in \mathcal{N}} \exp\{\alpha - D_{j,l}\} \right) + D_{i,j} \]

Multi-class N-pair Loss

• Use a positive pair and N-1 negative ones and

\[
\mathcal{L}_{N\text{-pair}}(\mathbf{x}, \mathbf{x}^+, \{\mathbf{x}_i^-\}_{i=1}^{N-1}) = \log \left( 1 + \sum_{i=1}^{N-1} \exp(f(\mathbf{x})^T f(\mathbf{x}_i^-) - f(\mathbf{x})^T f(\mathbf{x}^+)) \right)
\]

\[
= -\log \frac{\exp(f(\mathbf{x})^T f(\mathbf{x}^+))}{\exp(f(\mathbf{x})^T f(\mathbf{x}^+)) + \sum_{i=1}^{N-1} \exp(f(\mathbf{x})^T f(\mathbf{x}_i^-))}
\]

Softmax for multi-class classification

InfoNCE (Noise Contrastive Estimation) Loss

• Similar to multi-class N-pair Loss

\[ \mathcal{L}_q = -\log \frac{\exp(q \cdot k_+ / \tau)}{\sum_{i=0}^{K} \exp(q \cdot k_i / \tau)} \]

Query q

Positive k+  (K+1)-way softmax classification

Negatives ki Motivated from identifying targets from noisy data
Supervised Representation Learning

• Use class labels to specify positive pairs and negative pairs

• Loss functions
  • Contrastive loss
  • Triplet loss
  • Lifted structured loss
  • N-pair loss
  • InfoNCE

• Consider more relationships in a mini-batch is better
Unsupervised/Self-supervised Representation Learning

• Pretext tasks
  • Tasks designed for feature learning
  • Not the final tasks

• Positive pairs from different views of the same image

Learning Representations by Maximizing Mutual Information Across Views. Bachman et al., NeurIPS, 2019
Unsupervised/Self-supervised Representation Learning

- Pretext task: context prediction

Unsupervised/Self-supervised Representation Learning

• Pretext task: rotation prediction

Unsupervised Representation Learning by Predicting Image Rotations. Gidaris, et al., ICLR, 2018
Unsupervised/Self-supervised Representation Learning

• Pretext task: colorization

Unsupervised/Self-supervised Representation Learning

- Pretext task: inpainting

Unsupervised/Self-supervised Representation Learning

- Pretext task: clustering

Deep Clustering for Unsupervised Learning of Visual Features. Caron et al., ECCV, 2018
SimCLR

• A simple framework for contrastive learning of visual representations

\[
\ell_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j) / \tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k) / \tau)}
\]

SimCLR

- Transformations

(a) Original  (b) Crop and resize  (c) Crop, resize (and flip)  (d) Color distort. (drop)  (e) Color distort. (jitter)

(f) Rotate $\{90^\circ, 180^\circ, 270^\circ\}$  (g) Cutout  (h) Gaussian noise  (i) Gaussian blur  (j) Sobel filtering

SimCLR

• After training, keep the encoder network
  \[ h_i = f(\tilde{x}_i) = \text{ResNet}(\tilde{x}_i) \]

• Linear evaluation protocol for classification
  • A linear classifier is trained on top of the frozen base network

SimCLR


ImageNet top-1 accuracy

2x, 4x: more channels in ResNet
SimCLR

https://github.com/google-research/simclr
Summary: Visual Representation Learning

• Generative models
  • Autoencoder
  • VAE

• Discriminative models
  • Supervised learning
    • Training with image classification
    • Deep metric learning
  • Unsupervised/self-supervised learning
    • Use pretext tasks
    • Metric learning loss functions
Further Reading

• Learning a Similarity Metric Discriminatively, with Application to Face Verification, 2005 http://yann.lecun.com/exdb/publis/pdf/chopra-05.pdf


• Improved Deep Metric Learning with Multi-class N-pair Loss Objective, 2016 https://papers.nips.cc/paper/2016/file/6b180037abbebea991d8b1232f8a8ca9-Paper.pdf
