DS-GA 3001: Modern topic in Statistical Learning Theory Spring 2024

Lecture 9 — Self-Supervised Learning

Prof. Qi Lei *Scribe: Teo Zeng, Wenjun Dong*

1 Self-Supervised Learning

Self-supervised learning is an emerging paradigm in the field of machine learning, particularly deep learning, that focuses on learning useful representations of data by leveraging its inherent structure. The primary goal of self-supervised learning is to enable the learning process to occur without relying on a large amount of labeled data, as is required in supervised learning, by using the input data itself as a form of supervision.

2 Overview

we previously talked about meta representation learning

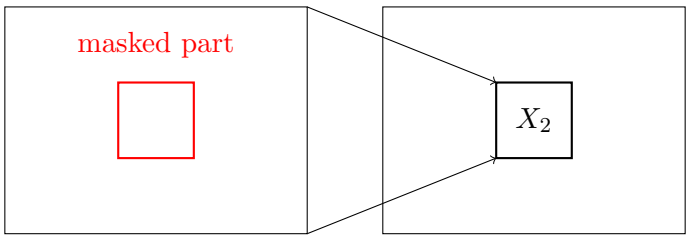
$$x \to \psi(x) \to w\psi(x)$$

The first arrow corresponds to source tasks which learn from unlabeled samples and the second corresponds to downstream tasks. Today we will talk about two types of self-supervised learning.

3 Types of Self-Supervised Learning

3.1 Type 1: Reconstruction-based self-supervised learning (SSL)

We want to predict a completed image from a part-masked image.



pretext task: $\psi : X_2 \approx \psi(X_1)$

downstream task: $Y \approx w\psi(X_1)$

We can try to let it predict part from other parts. Some examples of SSL include

- Context Encoder (Pathak et al. 2016) [7]
  learn the structure of visual data by predicting missing parts of images
- Masked Autoencoder (He et al. 2021) [5](MAE)
  Use Imagenet, trained on it and we can train it for specific tasks
- Colorization (Zhang et al. 2016) (SAV) [8]
  colorizing grayscale images using deep learning techniques

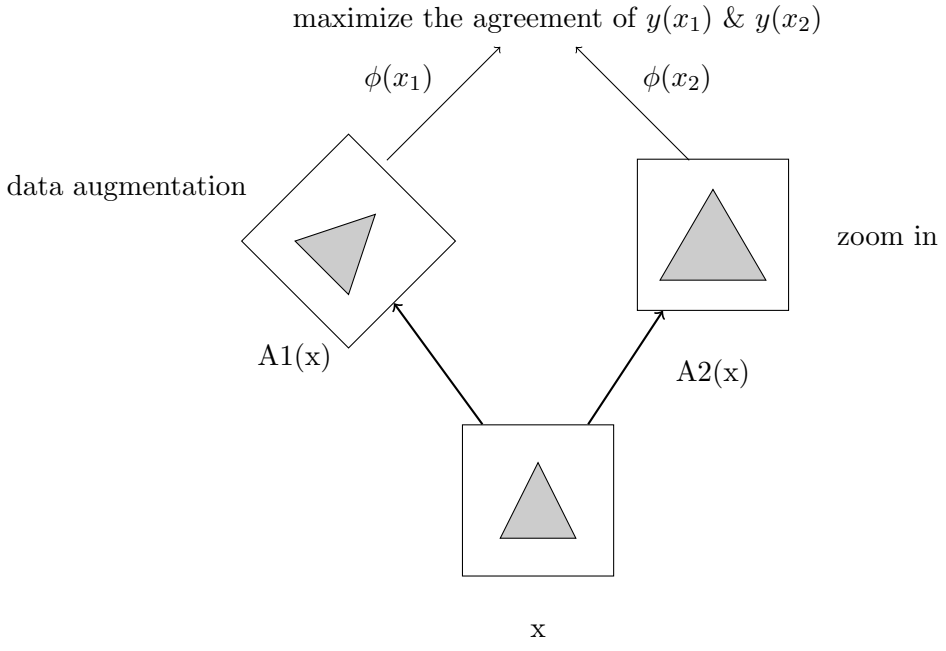
A quick [MASK] fox jumps over the [MASK] dog.
When predict masked sentences, we can use methods below.

- BERT (Devlin et al. [2018])[4] and ChatGPT (GPT-2, GPT-3, GPT-4)
  predicting the next most probable word. (a [brown] fox jump over the [lazy] dog)
- Asking the model to solve some puzzles. (Eg. Ordering the patch, predict the rotation)

Some tasks include "Ordering the patch" which involves shuffle the patches and ask the model to reorder. Or asking the model to predict the rotation of an image.

3.2 Type 2: Similarity based SSL

Enforce two views of the same data to have similar representations. By two views we meant that "two augmentations" of original data. The augmentations can be rotation, zooming in, zooming out, changing the color. The goal is to learn representations such that these two augmented versions, despite being different in appearance, are encouraged to have similar representations in the feature space.



Some examples of Similarity-Based SSL include

1. Sim-CLR (Chen et al. 2020) [2]

   The process involves creating pairs of similar (positive) and dissimilar (negative) images from a dataset, then training a neural network to identify whether a given pair of images is similar or not. Achieved through contrastive learning, where the model learns to pull similar images closer in the representation space and push dissimilar images farther apart.

2. SimSiam (Chen et al. 2021) [3]

   With Sim-CLR, these are different ways to avoid 'feature collapse' or representation 'collapse'. SimSiam focuses on a simpler approach that does not require negative pairs. Simsiam learn representations by maximizing the similarity between two different augmented views of the same image.

   Notes: SimCLR and SimSiam are different ways to avoid feature/representation collapse.

3. CLIP (Redford et al. 2021)[1] Two views are images and its caption.

4. CCA Canonical-correlation analysis (Compared with PCA)

   CCA try to get PCA by analyzing correlation of 2 groups of data.

Key intuition: pretext task should help us reduce irrelevant features information / forget information not relevant

3.3 Ideal Scenario

The ideal scenario is

$$x_1 \perp x_2|Y \Leftarrow P(X_1 = x_1|Y = y)P(X_2 = x_2|Y = y) = P(X_1 = x_1, X_2 = x_2|Y = y)$$

There, x_1 and x_2 are independent.

A thought experiment:

3.3.1 Image colorization for photos of desert, forest, and sea

The pretext task involves predicting color information for grayscale images of different environments such as deserts, forests, and seas shown in Figure 1

$$Y \in \{\text{desert, forest, sea}\}$$

$$X_1 \text{ is grayscale, } X_2 \text{ is color channel}$$

- Given Y is desert, we don't weed to refer to X_1 to predict X_2. X_1 and X_2 are independent.
- Knowing the information of Y, there will be small variance in predicting X_2 without even refering to the information of X_1

(a) Desert (b) Forest (c) Sea

Figure 1: Various classification

3.3.2 Image inpainting task

The goal was to fill in missing or damaged parts of an image in a visually plausible way.

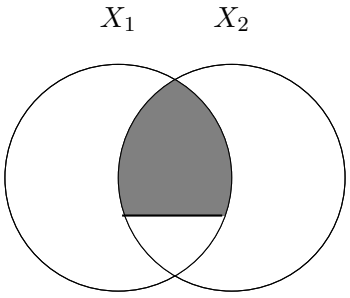
Figure 2: Image Painting

Given the part is mountain, mountains are similar. We can simply predict that is mountain.

3.4 Charcterize the approximate conditional independece

$$CI = P(X_1, X_2|Y) = P(X_1|Y)P(X_2|Y)$$
$$\epsilon_{CI} := \mathbf{E_{x_1}}\|\mathbf{E_{x_2}}[x_2|x_1] - \mathbf{E_Y}[\mathbf{E}[x_2|Y]|x_1]\|^2$$

Remark: If CI, ϵ_{CI}=0:



The gray interaction part represents Y and the white part of interaction represents ϵ_{CI}

The test error is bounded by the sum of the estimation error and the approximation error:

$$\text{test error} \leq \frac{k}{n_{\mathcal{L}}} + \epsilon_{\text{CI}}$$

where k is the cardinality of Y, $n_{\mathcal{L}}$ is estimation error and ϵ_{CI} is approximation error.

In traditional learning theory with supervised learning

$$\text{test error} \leq \frac{C(w\psi)}{n_{\mathcal{L}}}$$

k and ϵ_{CI} are tight in some circumstances.

3.5 Conceptual message:

We want to take $X_u \perp X_v$ to have smaller dependence given Y.

For example, the context encoder(Pathak et al. 2016)[7] is to predict the surrounding area of the bounding box.

And in masked autoencoder, an image was divided into many grids, and predict random grids from adjacent parts.

3.5.1 Text domain

In text domain, if we want to do sentiment analysis

$$Y \in \{\text{happy, unhappy}\}$$

I like this movie. The CG is awesome

Say like is x_2, then given Y we can predict that x_2 is more likely to be 'like'.

3.5.2 Reconstruction-based SSL

In reconstruction-based SSL, formally

$$L_{\text{mask}} = \mathbf{E}\|x_2 - \psi(x_1)\|^2,\ \sigma_x(\mathbf{E}[X2 \cdot Y]) > 0$$

This equation request $\sigma_k(\mathbf{E}(X_2Y)) > 0$. There, $(\mathbf{E}(X_2Y)) > 0$ is linear correlation between X_2 and Y.

To handle nonlinear dependence of X_2 and Y,

$$L_{\text{sim}} = E\|\eta(X_2) - \psi(X_1)\|^2$$

$$\underset{\eta,\psi}{\operatorname{argmin}} \left[\mathbb{E}\|\eta(X_2)\|^2+\mathbb{E}\|\eta(X_1)\|^2 - 2\mathbb{E}\left[\eta(X_2)^\top\psi(X_1)\right]\right]$$

if we normalize $\eta(X_2)$ and $\psi(X_1)$

$$\underset{\eta,\psi normalized}{\operatorname{argmax}} \mathbb{E}\left[\eta(X_2)^\top\psi(X_1)\right]$$

This is exactly CCA as similarity based SSL

Notes:

- representation error $\leq \tilde{\epsilon}_{\mathrm{CI}}$

$$\tilde{\epsilon}_{\mathrm{CI}} := \max_{\|g\|=1} \mathbb{E}_{x_1}\left[\mathbb{E}\left[g(X_2)^|X_1)\right]\right] = \mathbb{E}\left[\mathbb{E}\left[g(X_2)|Y\right]|X_1\right]^2$$

- need k-th maximal correlation between X_2 and $Y > 0$.
  Similarity is based on SSL. (Lee, Lei, saunshi,Zhuo. 2021)[6]

References

[1] Chris Hallacy Aditya Ramesh Gabriel Goh Sandhini Agarwal Girish Sastry Amanda Askell Pamela Mishkin Jack Clark Gretchen Krueger Ilya Sutskever Alec Radford, Jong Wook Kim. Learning transferable visual models from natural language supervision, 2021.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[3] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2021.

[6] Jason D. Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. Predicting what you already know helps: Provable self-supervised learning, 2021.

[7] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting, 2016.

[8] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 649–666. Springer, 2016.