Modern Topics in Statistical Learning Theory Spring 2023

Lecture 12 — Self Supervised Learning

Prof. Qi Lei *Scribe: Md Salman Rahman, Jingtong Su*

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 Recap from the Previous Lecture

In the previous lecture, we discussed various topics related to self-supervised learning, including:

1. Data augmentation: A technique used to create new training examples by applying transformations to existing data, thus increasing the diversity and size of the dataset.
2. Semi-supervised learning: A learning approach that combines a small amount of labeled data with a large amount of unlabeled data during training.
3. Pretraining and downstream tasks: The process of training a model on a large dataset (usually unsupervised or self-supervised) and then fine-tuning it on a smaller, task-specific labeled dataset.

3 Types of Self-Supervised Learning

There are two main types of self-supervised learning:

3.1 Type I: Reconstruction-based SSL

- Pretext tasks involve reconstructing or predicting parts of the input data.
- Downstream tasks leverage the learned representations for specific supervised tasks.

Examples:

- Context Encoder (Pathak et al. [2016]): Predicts missing parts of an image.
- Masked Autoencoder (He et al. [2022]): Reconstructs partially masked inputs.

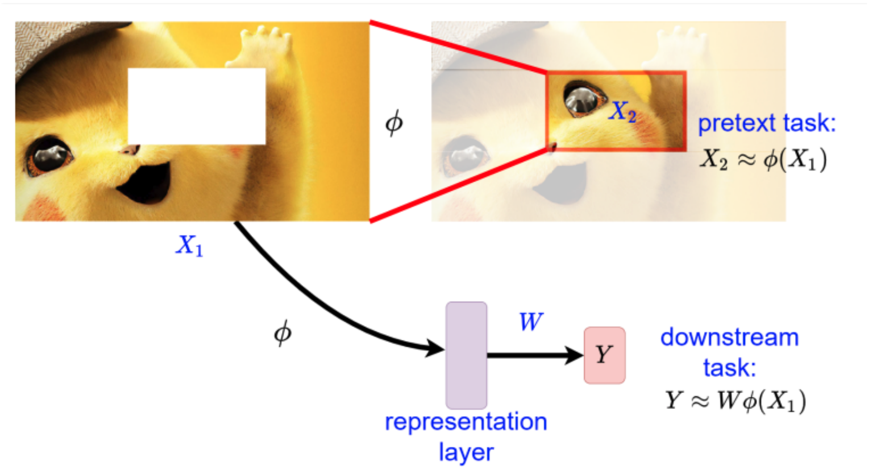

Figure 1: Type I: Reconstruction-based SSL. The figure shows the process of using X_1 for the pretext task X_2, and then using the learned weights W for the downstream task Y.

- Colorization (Zhang et al. [2016]): Predicts color information for grayscale images.
- BERT (Devlin et al. [2018]) and ChatGPT: Predict masked words or next words in a sentence.

3.1.1 Randomly Masked Example

A randomly masked example would be: "A quick [MASK] fox jumps over the [MASK] dog."

The model's prediction would be: "A quick brown fox jumps over the lazy dog."

3.1.2 Setup: Reconstruction-based SSL

In the reconstruction-based SSL setup, the goal is to learn useful representations from data by solving pretext tasks that involve reconstructing or predicting parts of the input data. These learned representations are then used in downstream tasks for specific supervised learning problems shown in Figure 1. The process involved:

- Label Y with k classes.
- Unmasked image X_1 and masked image X_2.
- Key intuition: Pretext tasks should help us reduce irrelevant features/forget information that is not necessary to predict Y.

3.1.3 Ideal Scenario

In the ideal scenario, the learned representations capture the relevant features of the data that help in predicting the downstream task labels. This can be formulated as conditional independence between the unmasked input X_1 and the masked input X_2 given the label Y: $X_1 \perp X_2|Y$. The process involved:

- Label Y with k classes.
- Unmasked image X_1 and masked image X_2.
- Ideal scenario: $X_1 \to Y \to X_2 \Leftrightarrow X_1 \perp X_2|Y$

3.1.4 A Thought Experiment

- Conditional independence: $X_1 \perp X_2|Y$
- 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 2

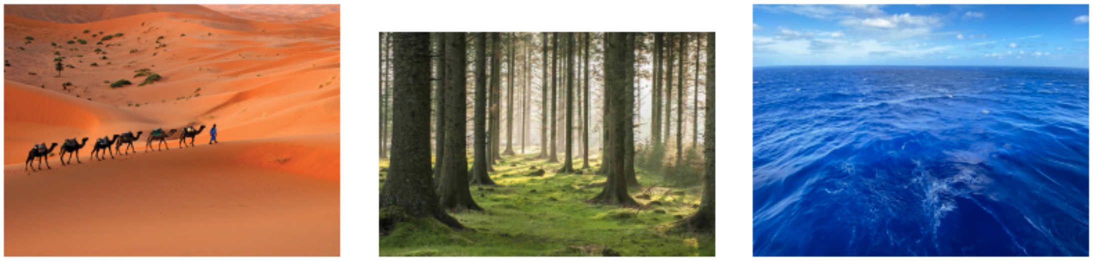

Figure 2: Image Colorization

- Image inpainting: The pretext task involves predicting missing or occluded parts of an image to reconstruct the entire image shown in Figure 3.

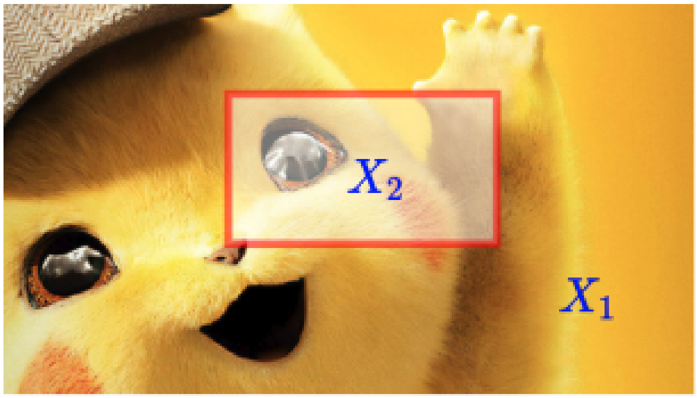

Figure 3: Image Painting

3.2 Type II: Similarity-based SSL

- Enforces similar representations for different views or augmentations of the same data point.
- Maximizes the agreement between representations of different views or augmentations.

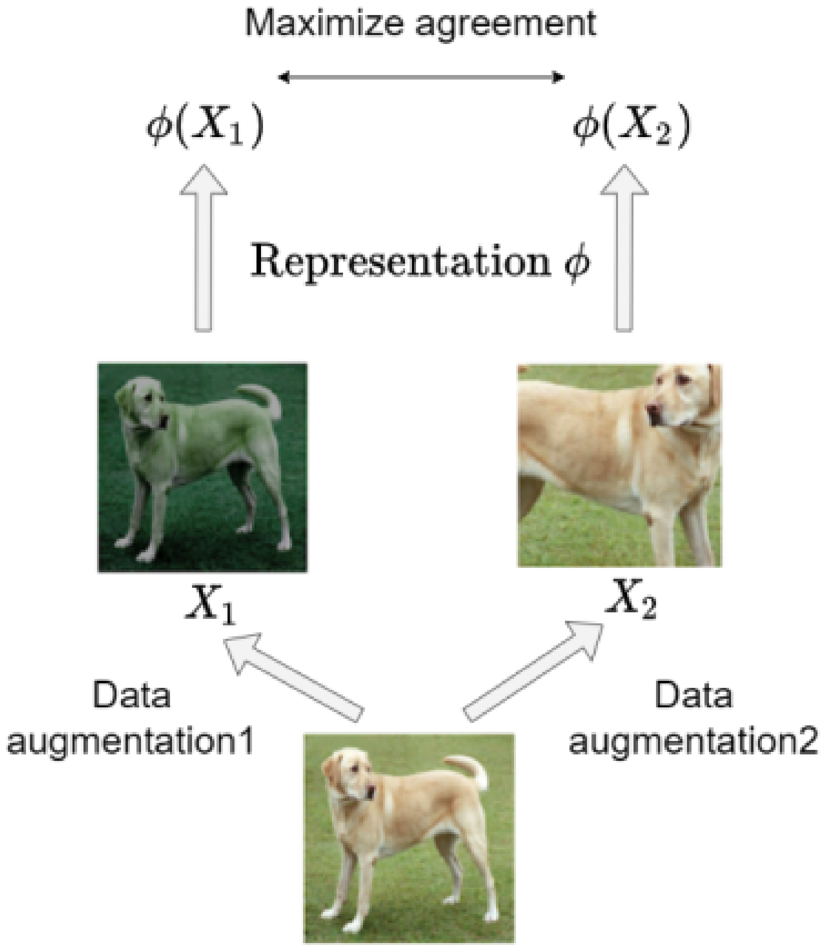

Figure 4: Type II: Similarity-based SSL

Examples:

- SimSiam (Chen and He [2021]): Maximizes similarity between two augmented views of the same image.
- CLIP (Radford et al. [2021]): Aligns image and text representations in a shared embedding space.
- SimCLR (Chen et al. [2020]): Maximizes similarity between representations of augmented image pairs while minimizing similarity to other images in the batch.

To prevent representation collapse, various techniques are employed, such as contrastive learning, negative sampling, or the use of stop-gradient operations.

4 The power of Conditional Independence (CI)

We proceed with a theoretical work, Lee et al. [2021], where the authors unified the two types of SSL losses and provided theoretical insights into the effectiveness of these algorithms given proper conditional independence. To start with, we define the concepts of pretext task and downstream task mentioned above in detail as follows.

Pretext Task: Learn a representation $\psi(x_1)$ close to $\psi^* := \arg\min_{g \in \mathcal{H}} \mathbb{E}||X_2 - g(X_1)||^2$, where $\mathcal{H}$ can vary for different settings. An example is all deep neural networks with a specific structure.

Downstream task: Perform linear regression on Y with $\psi(X_1)$, i.e. $f(x_1) := (W^*)^\top \psi(x_1)$, where $W^* \leftarrow \arg\min_W \mathbb{E}_{X_1,Y}[||Y - W^\top \psi(x_1)||^2]$.

Our goal is to obtain good generalization with a pre-training stage on the pretext task with unlabeled data only and finetuning on the downstream task with a small number of labeled data. In this section, we introduce a way to ensure the success of such a procedure by the Conditional Independence (CI) between X_1 and X_2 given Y:

$$P(X_1, X_2|Y) = P(X_1|Y)P(X_2|Y).$$

Before we step into the details, let us first get a good intuition about CI. We illustrate two examples here. (1): One thought experiment we have mentioned above. We want to classify desert, forest, and sea images. Denote X_1, X_2, and Y to be the input image, color channel, and downstream label, respectively. Given knowledge of the label Y, one can possibly predict the background X_2 without knowing much about X_1. In other words, X_2 is approximately independent of X_1 conditional on the label Y. (2): Consider an image inpainting task where we are given some background information X_1 and want to recover the front of some hidden area X_2. With proper label information Y (*e.g.,* a "building" with a detailed description, or a "mountain" with trees or with snow covering it) will ensure that variation in X_2 given Y is small, which is equivalent in a mathematical language that X_1 is approximate conditionally independent of X_2 given Y.

Near conditionally independence basically states that knowing the information of Y, there will be a small variance predicting X_2 without even referring to the information of X_1. It can be shown that with exact CI, the perfect solution W^* of the downstream task would achieve 0 approximation error when predicting Y.

The main insight from the CI point of view is, with approximate CI as in the above examples, a method that predicts X_2 from X_1 will inadvertently implicitly encode and learn to predict Y from X_1 as an intermediate step, and then predict X_2 from Y, thus confirming the success of this pretext-downstream pretrain-finetune procedure.

However, the strict CI condition does not hold in general. We need to come up with a way to characterize the extent that this statement holds. Thus, we propose the approximate conditional independence as follows. A conceptual illustration is given in Figure 5.

$$\epsilon_{CI} = \mathbb{E}_{X_1}||\mathbb{E}[X_2|X_1] - \mathbb{E}_Y[\mathbb{E}[X_2|Y]|X_1]||^2.$$

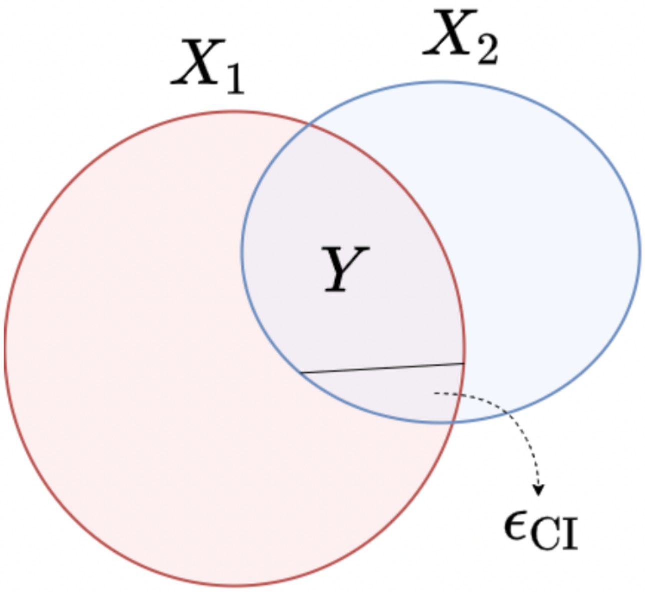

Figure 5: Conceptual illustration of ϵ_{CI}, which can be viewed as quantification of extra shared features between X_1 and X_2 are not captured by Y.

With proper sub-Gaussian assumptions on the formulation of data distribution, one can prove that

$$\text{Test Error} \lesssim \frac{k}{n_L} + \epsilon_{CI}.$$

Where k is the cardinality of Y, n_L is the number of (labeled) samples of the downstream task. Constants are hidden within the $\lesssim$ operator. Omitting technical details, it tells us:

- With ERM, only $n_L \asymp$ Complexity of Function Class labeled samples are required in the downstream finetuning stage to attain good generalization.
- Both terms are tight in some scenarios.

Given the theorem, we obtain a theoretical advice that we want to make X_1 and X_2 to have small dependence given Y.

Some real-world X_1 and X_2 formulation:

- Context Encoder: a fixed area of an image is picked as X_2, and the rest serves as X_1.
- Masked AutoEncoder: instead of fixing the position of X_2, during each epoch, random small patches of an image are picked as X_2, and the rest serves as X_1.
- LLM: in BERT and GPT training, words in a sentence are randomly picked to be masked as X_2, and the whole sentence or the preceding words are used as X_1.

5 Unifying reconstruction- and similarity-based SSL

Here we briefly introduce some high-level ideas of loss unification in SSL. For detailed analysis and full theorems, refer to Lee et al. [2021].

5.1 Unification Procedure

We start with the reconstruction-based SSL, where we mask X_2 out and try to reconstruct it from X_1. The loss is

$$L_{\text{mask}} = \mathbb{E}[||X_2 - \phi(X_1)]||^2$$

which requires $\sigma_k(\mathbb{E}[X_2 Y]) > 0$[1] Thus, to handle the non-linear dependence of X_2 and Y, instead of learning X_2, we seek help from the generalized alternating conditional expectation (ACE) algorithm that optimizes the following:

$$\min_{\phi,\eta} L_{\text{ACE}}(\phi, \eta) : \mathbb{E}[||\eta(X_2) - \phi(X_1)||]^2.$$

Here $\eta(\cdot)$ can be any (deterministic) function.

Under norm constraint with ϕ, η, the ACE loss is equivalent to the so-called non-linear canonical correlation analysis (CCA) loss:

$$\max_{\phi,\eta} L_{\text{CCA}} = \mathbb{E}[||\eta(X_2)^\top \phi(X_1)||]^2$$

The norm-constrained above CCA loss, plus normalization on ϕ, η and proper stop-gradient configuration, leads us to the famous SimSiam loss function, a representative similarity-based SSL algorithm. Viewing X_1 and X_2 in the ACE loss as images and description texts, we have also unifed the CLIP loss into this regime.

5.2 Theoretical Implication

Under L_{CCA}, we have the following theorem:

$$\text{Representation Error} \lesssim \min\{\frac{\tilde{\epsilon}_{CI}}{\beta}, \frac{\alpha}{1 - \tilde{\epsilon}_{CI}}\},$$

where β is the k-th maximal correlation between X_2 and Y, α is the Bayes error of predicting Y with X_1, and $\tilde{\epsilon}_{CI} = \max_{||g||=1} \mathbb{E}_{X_1}[\mathbb{E}[g(X_2)|X_1] - \mathbb{E}_Y[\mathbb{E}[g(X_2)|Y]|X_1]]^2$. Note that here we also require the maximal correlations to be greater than zero.

[1] σ_k is the maximal correlation coefficient.

References

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

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, 2022.

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.

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.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

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.

Jason D Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. Predicting what you already know helps: Provable self-supervised learning. *Advances in Neural Information Processing Systems*, 34: 309–323, 2021.