Universidade de São Paulo
Escola Politécnica - Engenharia de Computação e Sistemas Digitais

Self-Supervised and Semi-Supervised Learning

Prof. Artur Jordão

Introduction

Self-Supervised and Semi-Supervised Learning

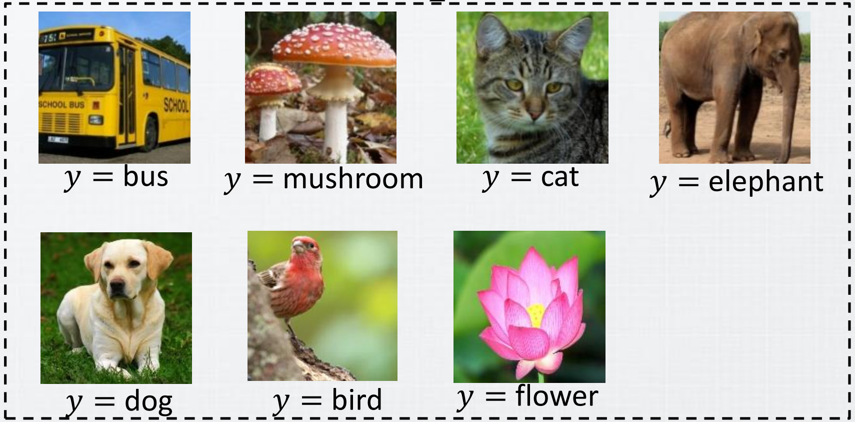
- Deep learning has driven unprecedented progress in various cognitive applications
  - However, most of them operate in a **supervised learning** scenario
- The supervised learning paradigm requires manual data labeling, which is both limited in quantity and labor-intensive
- Self-Supervised and Semi-Supervised learning (SSL) extend **supervised** learning to massive amounts of **unlabeled** data
- The SSL learning paradigm is key for training foundation models

Preliminaries

Self-Supervised and Semi-Supervised Learning

- Let $\mathcal{D}_L = \{(x_i^l, y_i^l)\}_{i=1}^{N_L}$ be a labeled dataset
- Let $\mathcal{D}_U = \{x_i^u\}_i^{N_U}$ be an unlabeled dataset
- Since unlabeled data are abundant, in practice, $N_L \ll N_U$

$\mathcal{D}_L$



$\mathcal{D}_U$

Preliminaries

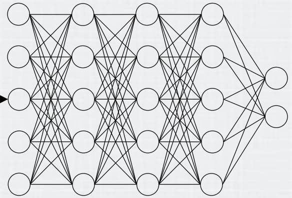
Self-Supervised and Semi-Supervised Learning

- A core idea of SSL is to use large- and web-scale unlabeled data, $\mathcal{D}_U$, to train a model to learn **meaningful representations** that can be effectively **transferred** to downstream tasks (i.e., $\mathcal{D}_L$)
    - Learn meaningful and task-agnostic latent representations

$$\mathcal{D}_U = \{x_i^u\}_i^{N_U}$$

Training on

$(\{x_i^u\}_i^{N_U}, ?)$



SSL Benchmark

Self-Supervised and Semi-Supervised Learning

- **S**elf-**S**upervised and **S**emi-**S**upervised **L**earning Benchmark (Wang. et al., 2022)
  - SSL

Wang et al. *USB: A Unified Semi-supervised Learning Benchmark for Classification*. Neural Information Processing Systems (NeurIPS) 2022

Self-Supervised Learning

Problem Definition

Self-Supervised Learning

- Self-supervised learning introduces **pseudo-label** generation, $\mathcal{P}(\cdot)$, to label data

- Given $\mathcal{D}_U = \{x_i^u\}_i^{N_U}$, the problem becomes one of automatically generating labels y_i^u
  - We can obtain y_i^u using $\mathcal{P}$: $y_i^u = \mathcal{P}(x_i^u)$

- Therefore, we can generate a (self-)**supervised** dataset ($\mathcal{D}_S$) in terms of
  - $\mathcal{D}_S = \{(x_i^u, \boldsymbol{\mathcal{P}}(\boldsymbol{x_i^u}))\}_i^{N_U}$

- Finally, we can train a model $\mathcal{F}$ using the **supervised paradigm on $\boldsymbol{\mathcal{D}_S}$**

Self-Supervised in Computer Vision

Self-Supervised Learning

Supervised Scenario

$y =$ bird

$y =$ dog

$y =$ mushroom

$y =$ elephant

Self-supervised Scenario

$y =$ 90º

$y =$ 270º

$y =$ 45º

$y =$ 0º

Gidaris et al. *Unsupervised Representation Learning by Predicting Image Rotations*. International Conference on Learning Representations (ICLR), 2018

Hendrycks et al. *Using Self-Supervised Learning Can Improve Model Robustness and Uncertainty*. Neural Information Processing Systems (NeurIPS), 2019

Self-Supervised in Computer Vision

Self-Supervised Learning

Supervised Scenario

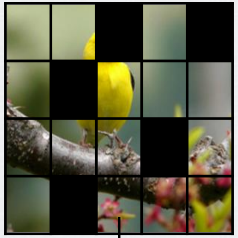
x

$\mathcal{F}(x)$

Bird

Self-supervised Scenario

x



$\mathcal{F}(x)$

He et al. *Masked Autoencoders Are Scalable Vision Learners*. Computer Vision and Pattern Recognition (CVPR), 2022

Self-Supervised for Large Language Models

Self-Supervised Learning

- Language Modeling
  - Predict the next token
- Masked Language Modeling
  - Mask out some tokens from the input sentences and then train the model to predict the masked tokens using the surrounding context
- Denoising Autoencoder
  - Take a partially corrupted input and aim to recover the original, undistorted input
- Next Sentence Prediction
  - Train the model to distinguish whether two input sentences are continuous segments from the training corpus

Loss Function and Pre-Training

Self-Supervised Learning

- When using SSL learning, we can combine supervised and unsupervised losses
- Suppose $\mathcal{L}(\cdot,\cdot)$ be a loss function (i.e., categorical cross-entropy or ℓ_2)
- Assume $\mathcal{B}_s$ and $\mathcal{B}_U$ be batches of labeled and unlabeled data
- Supervised loss $\mathcal{L}_s = \frac{1}{\mathcal{B}_S} \sum \mathcal{L}(\mathcal{F}\left(x_i^l, \theta\right), y_i^l)$
- Unsupervised loss $\mathcal{L}_U = \frac{1}{\mathcal{B}_U} \sum \mathcal{L}(\mathcal{F}(x_i^u, \theta), \boldsymbol{\mathcal{P}}(\boldsymbol{x_i^u}))$
- Total loss $\mathcal{L}_s + \mathcal{L}_U$

Loss Function and Pre-Training

Self-Supervised Learning

- Instead of learning with $\mathcal{L}_S + \mathcal{L}_U$, we can pre-train a model on unlabeled data using self-supervised learning only
  - *Pre-train then Tune* paradigm

- Then, we **fine-tune the model on labeled data**

- Pre-train using self-supervised learning can improve several aspects of model robustness (Hendricks et al., 2019)

Hendrycks et al. *Using Self-Supervised Learning Can Improve Model Robustness and Uncertainty*. Neural Information Processing Systems (NeurIPS), 2019

Semi-Supervised Learning

Problem Definition

Semi-Supervised Learning

- Semi-supervised learning employs **pre-trained models**, i.e., $\boldsymbol{\mathcal{F}_A}(\cdot)$, to generate labels

- Suppose we have a well-trained model $\mathcal{F}_A$ using the supervised paradigm on $\mathcal{D}_L$

- Given $\mathcal{D}_U = \{x_i^u\}_i^{N_U}$, the problem becomes generating labels y_i^u
  - We can obtain y_i^u using $\mathcal{F}_A : y_i^u = \mathcal{F}_A(x_i^u)$

- Therefore, we can generate a semi-supervised dataset ($\mathcal{D}_S$) in terms of
  - $\mathcal{D}_S = \{(x_i^u, \mathcal{F}_A(x_i^u))\}_i^{N_U}$

- Finally, we can train a novel model $\mathcal{F}_B$ using the **supervised paradigm on $\boldsymbol{\mathcal{D}_S}$**

Limitations and Confirmation Bias

Semi-Supervised Learning

- Semi-supervised learning requires additional computational cost to label $\mathcal{D}_U$
  - We need to forward $\mathcal{D}_U$ through the pre-trained model
  - If $\mathcal{D}_U$ is a web-scale dataset, the forward pass could become computationally prohibitive

- The main problem of SSL is how to generate accurate pseudo labels

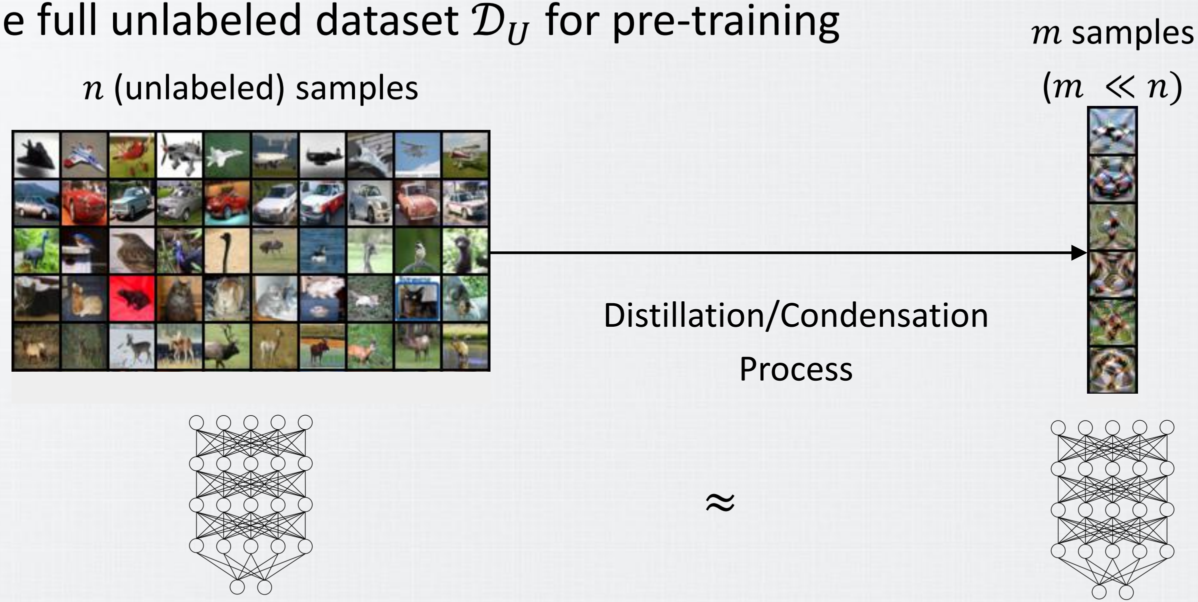
- Overfitting to incorrect pseudo-labels predicted by the network is known as *confirmation bias* (Li et al., 2024)

Li et al. *SemiReward: A General Reward Model For Semi-supervised Learning*. International Conference on Learning Representations (ICLR), 2024

Self-Supervised Dataset Distillation

Semi-Supervised Learning

- Lee et al. (2024) proposed the *self-supervised dataset distillation* problem
- The central idea is to **accelerate** the pre-training of a model by utilizing the **distilled dataset** in place of the full unlabeled dataset $\mathcal{D}_U$ for pre-training



Lee et al. *Self-Supervised Dataset Distillation for Transfer Learning*. International Conference on Learning Representations (ICLR), 2024

Bibliography

Bibliography

- Hendrycks et al., *Using Self-Supervised Learning Can Improve Model Robustness and Uncertainty*. Neural Information Processing Systems (NeurIPS), 2019

- Chen et al. *Big Self-Supervised Models are Strong Semi-Supervised Learners.* Neural Information Processing Systems (NeurIPS), 2020

NEURAL INFORMATION PROCESSING SYSTEMS

Bibliography

- Lee et al. *Self-supervised Dataset Distillation for Transfer Learning*. International Conference on Learning Representations (ICLR), 2024
- Li et al. *SemiReward: A General Reward Model For Semi-supervised Learning*. International Conference on Learning Representations (ICLR), 2024

ICLR
International Conference On Learning Representations