Machine Learning (CE 40717)

Fall 2024

Ali Sharifi-Zarchi

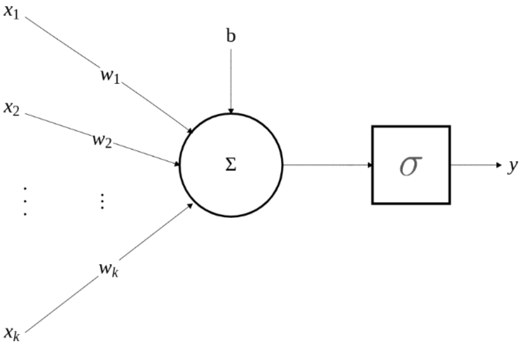
CE Department
Sharif University of Technology

November 26, 2024

Self-Supervised Learning

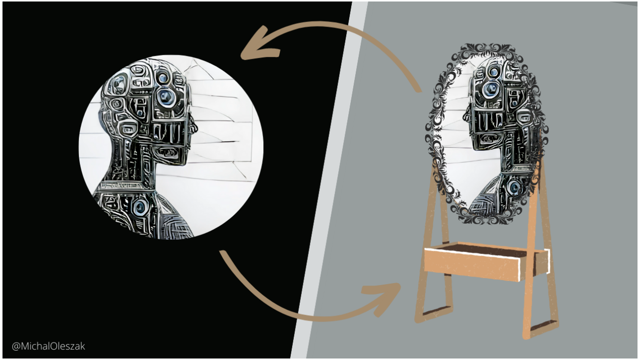
"the dark matter of intelligence" [1]

- $\{x_1, x_2, \ldots, x_k\}$: input features
- $\{w_1, w_2, \ldots, w_k\}$: feature weights
- b : bias term
- $\sigma(\cdot)$: activation function
- y : output of the neuron



[1] https://ai.meta.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/

Self-Supervised Learning



"the dark matter of intelligence"[2]

[2] https://ai.meta.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/

Why Neural Networks?

- Self-supervised learning defines a "pretext" task based on unlabeled inputs to produce descriptive and intelligible representations [Hastie et al., 2009, Goodfellow et al., 2016]
    - Learn with supervised learning objectives, e.g., classification, regression.
    - Labels of these pretext tasks are generated *automatically*
    - Can be used in other downstream tasks.

Example Workflow

- Training objective: predicting the context surrounding a word
- encourages the model to capture relationships among words
- The same SSL model representations can be used across a range of downstream tasks. e.g.
  - translating text across languages
  - summarizing
  - generating text

Motivation

- Problem: Supervised Learning is Expensive!
  - Labeling data is costly
  - SSL: Use signals that can be created automatically from data.
- Labled data is harder to find. There is much more unlabled data.
- Supervised Learning is not how "we" learn
  - Babies don't get supervision for everything they see!

Comparison

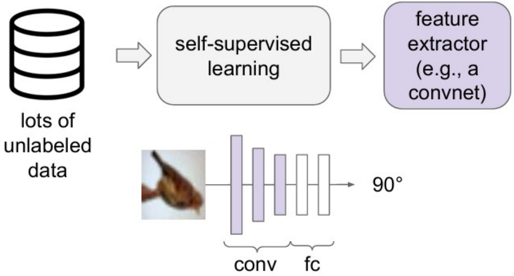
Methods that learn from data without annotations.

- **Unsupervised Learning**: Model isn't told what to predict. Older terminology, not used as much today.
- **Self-Supervised Learning**: Model is trained to predict some naturally occurring signal in the raw data rather than human annotations.
- **Semi-Supervised Learning**: Train jointly with some labeled data and (a lot) of unlabeled data

Evaluation

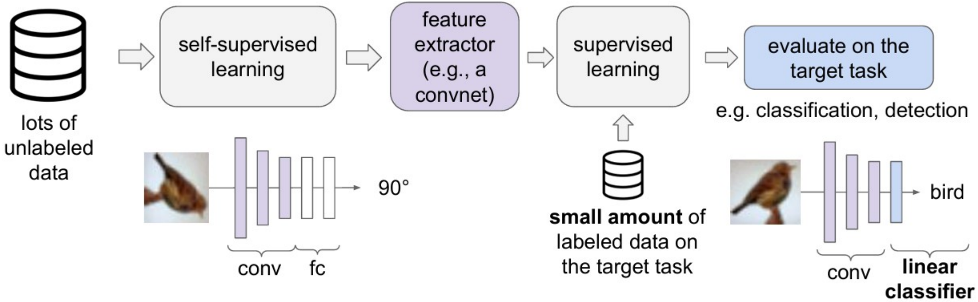
- We usually don't care about the performance of the self-supervised learning task, e.g., we don't care if the model learns to predict image rotation perfectly.
- Evaluate the learned feature encoders on downstream target tasks

Evaluation Cont.



1. Learn good feature extractors from self-supervised pretext tasks, e.g., predicting image rotations

Evaluation Cont.



1. Learn good feature extractors from self-supervised pretext tasks, e.g., predicting image rotations

2. Attach a shallow network on the feature extractor; train the shallow network on the target task with small amount of labeled data

Example

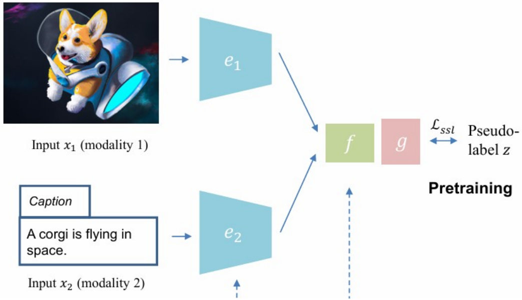
- Pretext task: predict rotations
- Hypothesis: a model could recognize the correct rotation of an object only if it has the "visual commonsense" of what the object should look like unperturbed.
- The model learns to predict which rotation is applied (4-way classification)
- (This slide will be ellaborated on and expanded with diagrams)

Idea

- Many papers would pretrain on (unlabeled) ImageNet, then evaluate on ImageNet!
- Don't learn from isolated images – take images together with some **context**
- Video: Image together with adjacent video frames
- Sound: Image with audio track from video
- 3D: Image with depth map or point cloud
- Language: Image with natural-language text

Why Language ?

- Semantic density: Just a few words give rich information
- Universality: Language can describe any concept
- Scalability: Non-experts can easily caption images; data can also be collected from the web at scale



Adapted from

Common Components

- Dataset :
  - supervised: $D_m = \{(x_1^1, \cdots, x_M^1, y^1), \cdots, (x_1^n, \cdots, x_M^n, y^n)\}$
  - self-supervised: $D_m = \{(x_1^1, \cdots, x_M^1), \cdots, (x_1^n, \cdots, x_M^n)\}$
- The psudo-label or signal generated for SSL can be denoted as $z = P(x_1, \cdot, x_M)$.
- Modality Encoder(s): $c = e_k(x_k^i; \theta_k)$ for each modality k.
- Fusion Module: f_ψ to integrate the encoded information of different modalities
- Pretext task head (like a predictive head) : g_γ and some SSL loss $\mathscr{L}_{SSL}$

Architectures

- There many veriations and structures

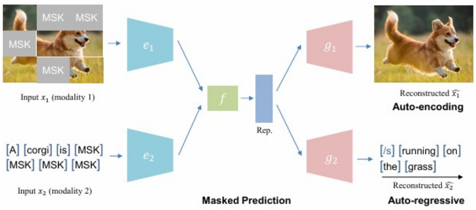


Figure 1: Figure 1 masked prediction frameworks

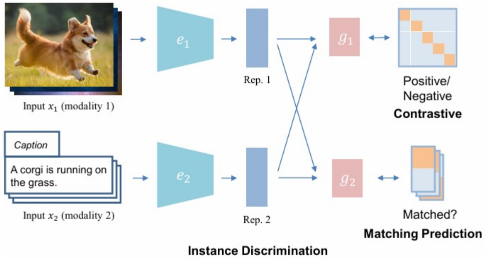


Figure 2: Figure 2 instance discrimination objectives

Contrastive Loss

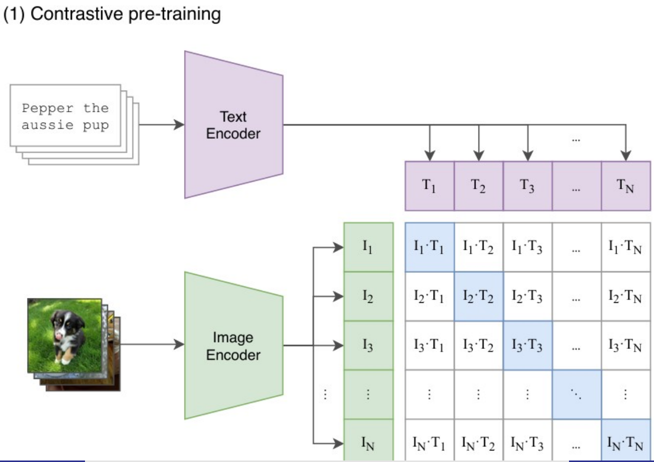
- (Put Contrastive Loss Slides here)

CLIP

- Connecting text and images
- Contrastive Language–Image Pre-training
- CLIP $\implies$ a shared representation(embedding) between two modalities (text and images) by training on a large dataset of image-text pairs.

CLIP Cont.

- Image Encoder: a Vision Transformer (ViT) or a ResNet.
- Text Encoder: A Transformer model



CLIP Goals

- CLIP was designed to mitigate a number of major problems:
- Costly datasets: Deep learning needs a lot of data, manually labeled datasets are expensive to construct.
  - CLIP learns from text–image pairs that are already publicly available on the internet
- Narrow: An ImageNet model excels at predicting the 1000 ImageNet categories but requires additional data and fine-tuning for other tasks.
  - CLIP can be adapted to perform a wide variety of visual classification tasks without needing additional training examples.

Zero-Shot Classification

- (Put Zero-Shot and Applications Slides here)

1. Introduction
2. Multimodal and CLIP
3. References

Contributions

These slides are authored by:

- Hooman Zolfaghari

[1] R. Ramakrishnan, "Deep learning course at carnegie mellon university." https://deeplearning.cs.cmu.edu/F23/index.html, 2023. Accessed: 2024–09-04.

[2] E. Mousavi and K. Alishahi, "Deep learning course at sharif university of technology." https://https://dnncourse.github.io/lectures, 2023. Accessed: 2024–09-04.

[3] D. Smilkov and S. Carter, "A neural network playground." playground.tensorflow.org, 2022. Accessed: 2024-10-14.

[4] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[5] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques for Building Intelligent Systems*. O'Reilly Media, 2019.