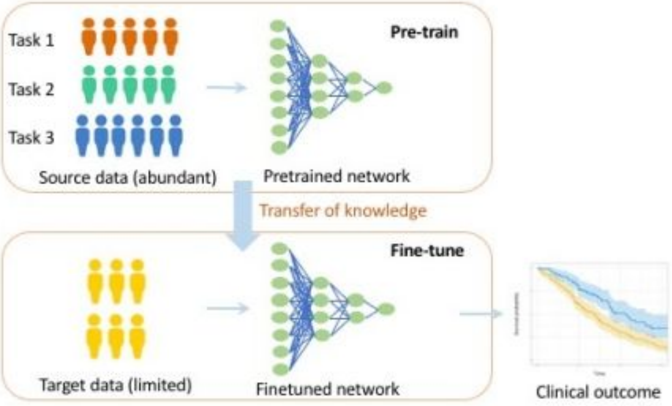
Lab 10
Transfer Learning

Haoxu Huang

Motivation

- People use prior knowledge in performing a new task instead of learning from scratch.
- Similarly, can we use prior information to train Machine Learning models faster and more sample efficiently?
- Train a model on different (but relevant) learning tasks such that it could help solve new tasks with only a few samples

Transfer Learning / Meta Learning

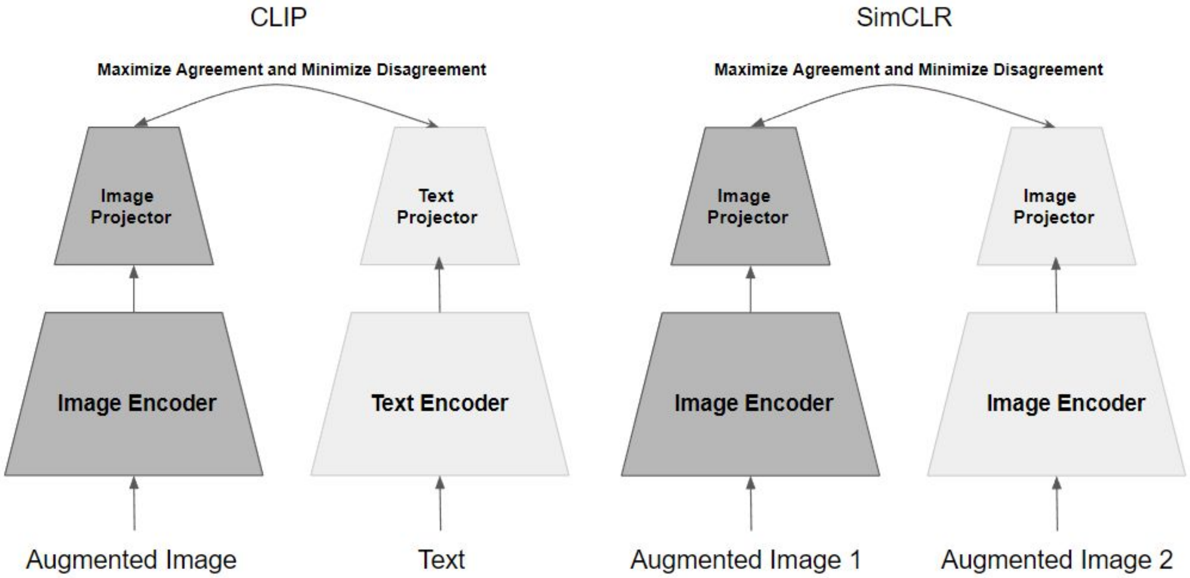


$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{e} \hat{L}(D_i^s, f_i) \text{ such that } f_i = A(D_i^s, \theta)$$

$$f^T = A(D^T, \hat{\theta})$$

Credit: https://med.stanford.edu/gevaertlab/MetaLearning.html

Self-Supervised/Multi-Modal Pre-training

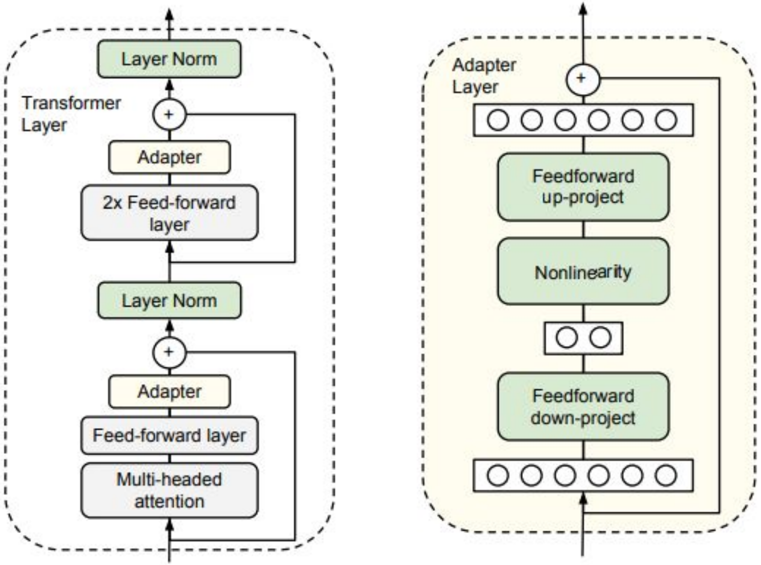


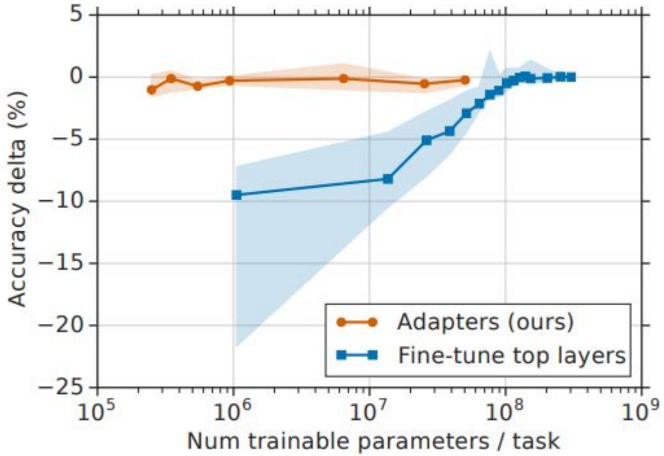
Loss: $$-\sum_{i \in \mathcal{B}} \log \frac{\exp(\tau u_i v_i)}{\sum_{j \in \mathcal{B}} \exp(\tau u_i v_j)}$$

Alec Radford et.al. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning, 2021

Ting Chen et.al. A simple framework for contrastive learning of visual representations. In International Conference on Machine Learning, 2020.

Parameter-Efficient Fine-Tuning





Neil Houlsby, et. al. Parameter-efficient transfer learning for NLP. In International Conference on Machine Learning, 2019.

When does Transfer Learning Fail?

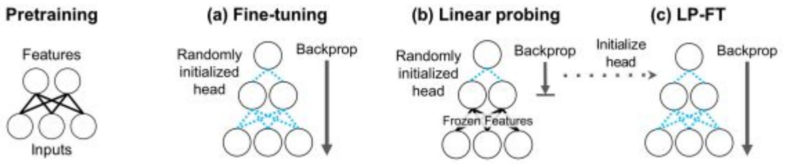
1. Out-of-Distribution Dataset

2. Discrepancy on Data Augmentations

3. Size of Downstream Dataset

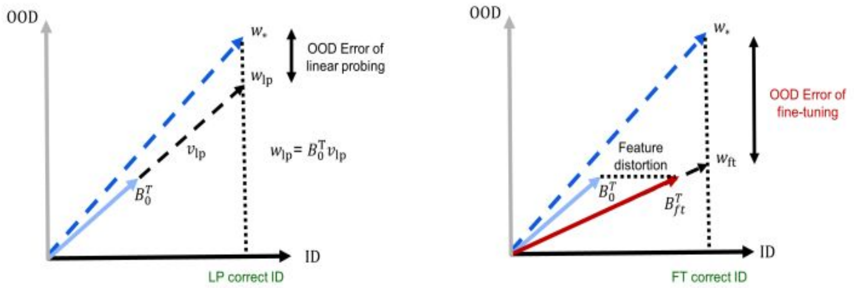
Barret Zoph, et.al. Rethinking pre-training and self-training. In Advances in Neural Information Processing Systems, 2020.

Linear Probing After Fine-Tuning



ID test	85.1%	82.9%	**85.7%**
OOD test	59.3%	66.2%	**68.9%**

Average accuracies (10 distribution shifts)



(a) Toy example (Linear probing)

(b) Toy example (fine-tuning)

Ananya Kumar et.al. Fine-tuning can distort pretrained features and underperform out-of-distribution. In International Conference on Learning Representations, 2022