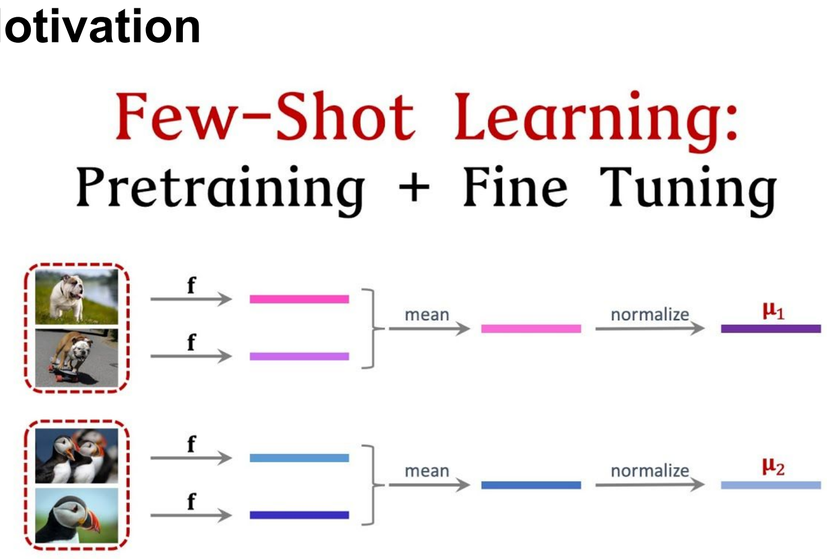
The Trade-off between Label Efficiency and Universality of Representations from Contrastive Learning

Zhenmei Shi*, Jiefeng Chen*, Kunyang Li, Jayaram Raghuram, Xi Wu, Yingyu Liang, Somesh Jha

ICML International Conference On Machine Learning

Motivation



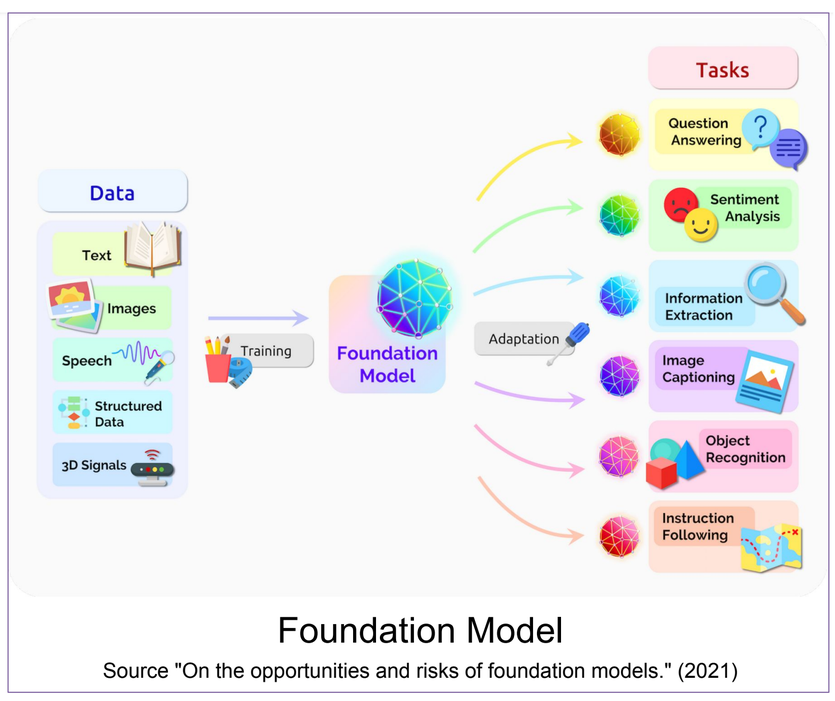
Label Efficiency

With the pre-trained representation, only a small amount of labeled data is needed to build accurate predictors for downstream target tasks.

vs

Universality

The pre-trained representation can be used for various downstream tasks.



Foundation Model

Source "On the opportunities and risks of foundation models." (2021)

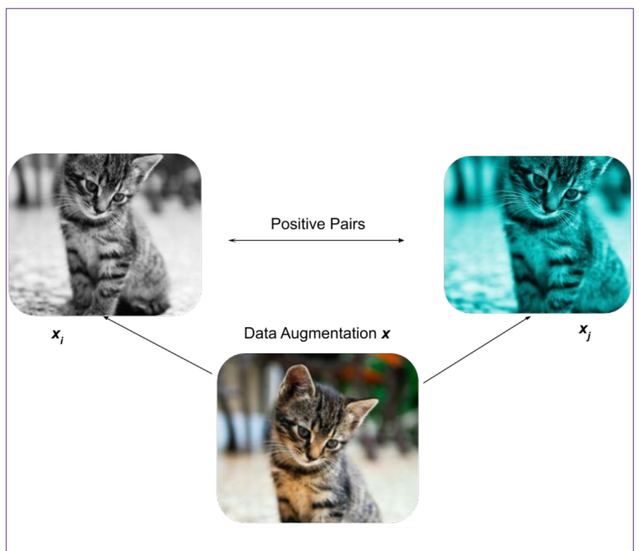
Take-Home Message

Pre-training on diverse data allows learning diverse features but can down-weight those for a target task, thus having worse prediction performance.

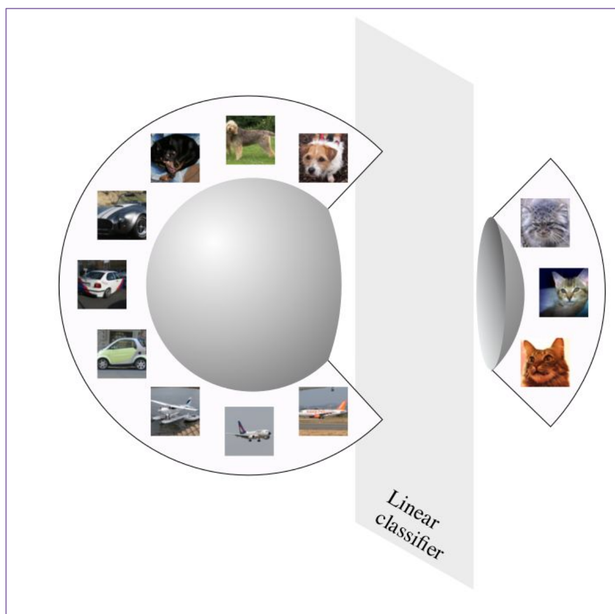
Key Intuition

The contrastively learned representation encodes frequent data features that are not affected by the transformations.

1. Output representation will not encode *Spurious* feature in the input which is changed by the transformations.
2. More common *Invariant* features will have a higher impact on the learned representation.
3. Then imply the trade-off between two properties.



Augmentation in Contrastive Learning



Linear Probing

Source "Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere." (2020)

Experiments

Model

MoCo v2, SimSiam.

Dataset

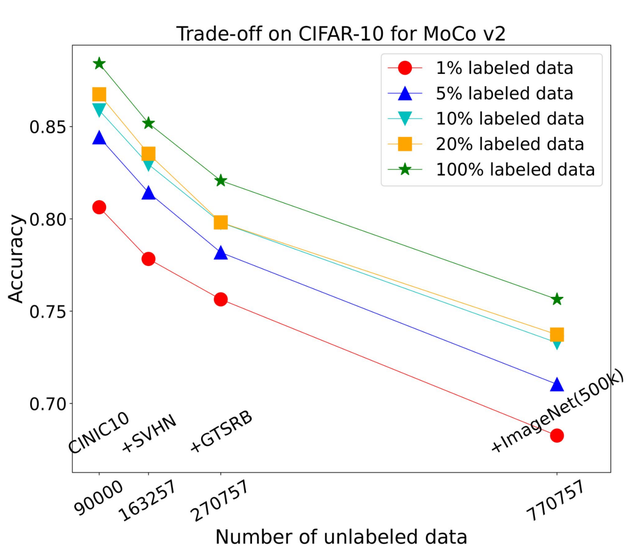
- Target task CIFAR-10: CINIC-10 target relevant.
- Target task MNIST: EMNIST-Digits&Letters target relevant.

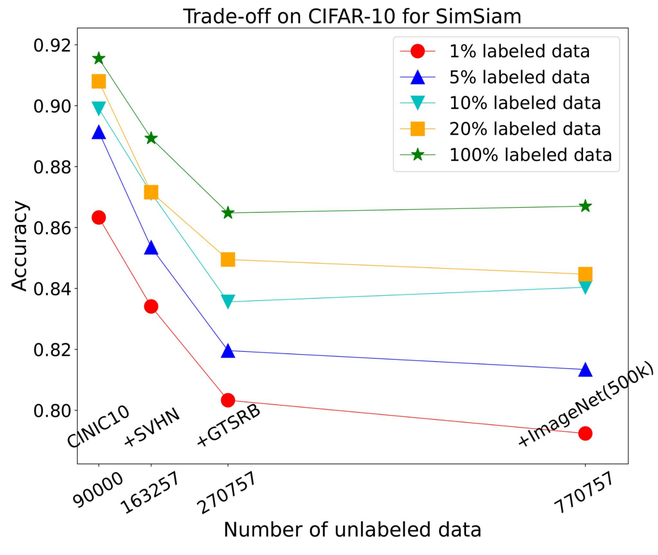
Evaluation & Methods

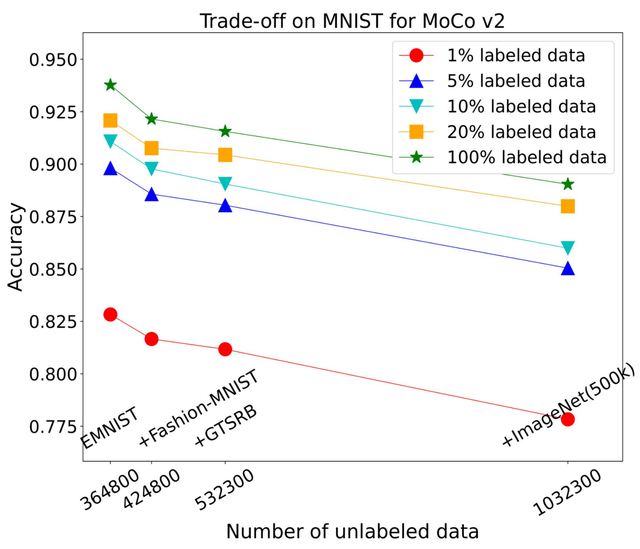
Pre-train ResNet18 using SGD. Then fix the pre-trained feature extractor, and train a linear classifier on 1%, 5%, 10%, 20%, 100% of the labeled data from the downstream task using SGD. Report test accuracy.

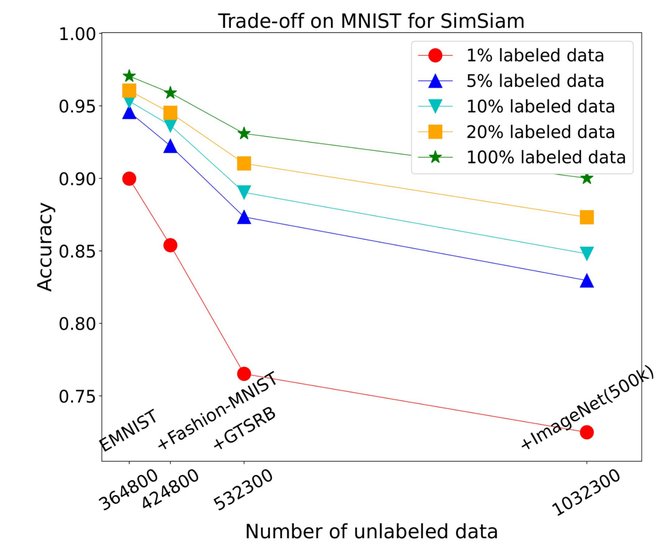
Results

When pre-training dataset combined with more diverse data, the test accuracy for the specific downstream task decreases. As more diverse unlabeled data included, more labeled data from the target task is needed to achieve a comparably good prediction accuracy.









Theoretical Analysis

Contrastive Learning

Pre-training objective function:

$$\min_{\phi \in \Phi} \mathbb{E}_{(x,x^+)} \left[\ell \left(\phi(x)^\top (\phi(x^+) - \mathbb{E}_{x^-} \phi(x^-)) \right) \right]$$

Downstream predictor objective function, f is a linear classifier:

$$\min_{f \in \mathcal{F}_\phi} \mathbb{E}_{(x,y)\sim \mathcal{D}_t} \left[\ell_c(f(\phi(x)), y) \right]$$

Hidden Representation Data Model

- First sampling the hidden representation with spurious features and invariant features, and then generate input.
- Label depends on a subset of invariant features.
- *Positive* pairs share the same invariant features, while *Negative* pairs are i.i.d. with each other.

Proposition of Optimal Representation (Informal)

An *optimal* representation for the contrastive loss has a closed-form. The representation ignores all spurious features and keeps weighted invariant features, where the weights depend on the frequency of feature being in data distribution.

Proposition of Complexity (Informal)

With a proper ground-truth labeling function, the estimated Rademacher complexity of pre-training on multi-tasks is larger than just pre-training on the target task.