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

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

Motivation

Few-Shot Learning: Pretraining + Fine Tuning

Label Efficiency

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

Universality

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

Problem Setup

Hidden representation data model

- Hidden representation space $z \in \mathbb{R}^D$ over distribution $D_z$
- Invariant feature $z_I$. Spurious feature $z_S$. $\mathbb{R} \cup U = \{\}$. $\mathbb{R} \cap U = \{\}$
- $x = g(z)$. $g$ is a generative function. $g$ depends on $z$ as well.

Label Efficiency

Contrastive learning and PCA

- $\phi \in \mathcal{H}$ hypothesis class of representation functions, e.g., ResNet, ViT
- Contrastive Loss $\ell_{C} = \min_{x \neq y} \mathbb{E}[\phi(x) - \phi(y)]^2$
- In SimCLR, we have multiple negative pairs and $f(t) = \log(1 + e^{\langle x, y \rangle})$

Theoretical Analysis

What features are learned by contrastive learning?

Theorem (Contrastive Learning is Generalized Nonlinear PCA)

If $f(t) = -t$, Contrastive Learning is equivalent to PCA on $\phi_{NC}$.

Moreover, if $f$ is linear function, it is equivalent to linear PCA on $\phi_{NC}$.

Theorem (Encode Invariant Feature; Remove Spurious Feature)

(1) $\phi$ does not encode spurious feature: $\phi^* \circ g(z) \perp z_I$.
(2) $\phi$ only encodes invariant feature whose “variance” large enough, and encoding strength increases when “variance” becomes larger.

Experiments

Model

- MoCo v2 (ResNet18), MoCo v3 (ViT-S), SimSiam (ResNet50).
- Dataset: Target task CIFAR-10/Imagenet-Bird.

Evaluation & Methods

From left to right, incrementally add to pre-training: CINIC-10 (C), SVHN (S), GTSRB (G), and ImageNet32 (I). Then fix the pre-trained feature extractor, and train a linear classifier on labeled data from the downstream task. Report target task test accuracy and averaged test accuracy over all pre-training datasets.

Trade-off

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

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