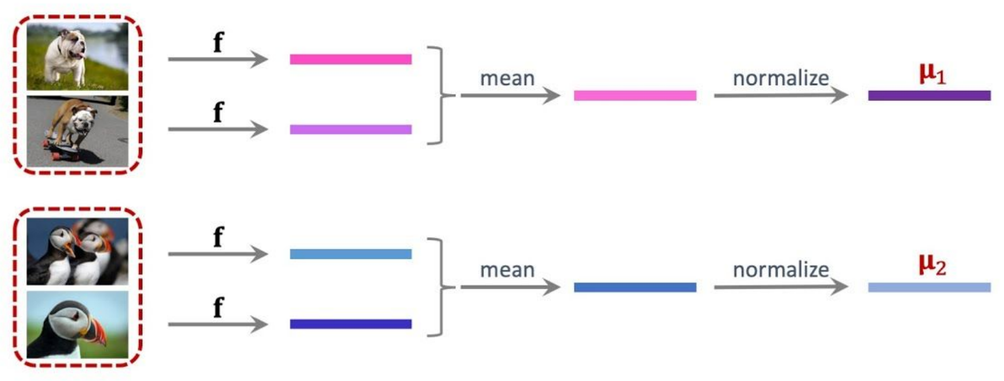
The **Trade-off** between **Universality** and **Label Efficiency** of Representations from Contrastive Learning **(Spotlight)**

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

Motivation

Few-Shot Learning:
Pretraining + Fine Tuning



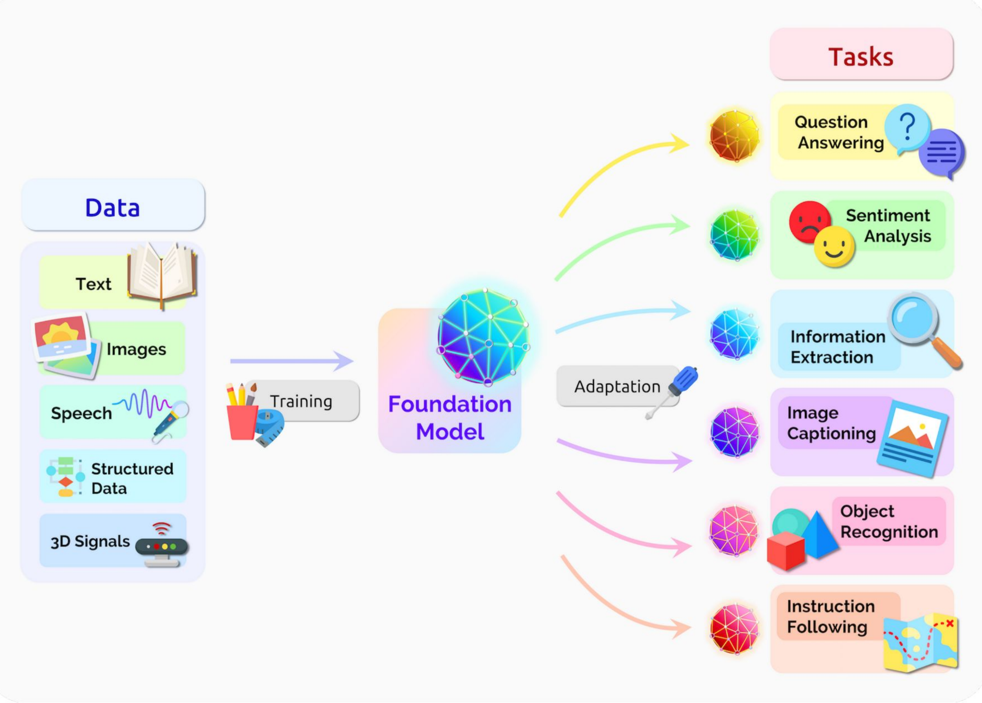
Label Efficiency

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

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)

Experiments

Model MoCo v2 (ResNet18), MoCo v3 (ViT-S), SimSiam (ResNet50).

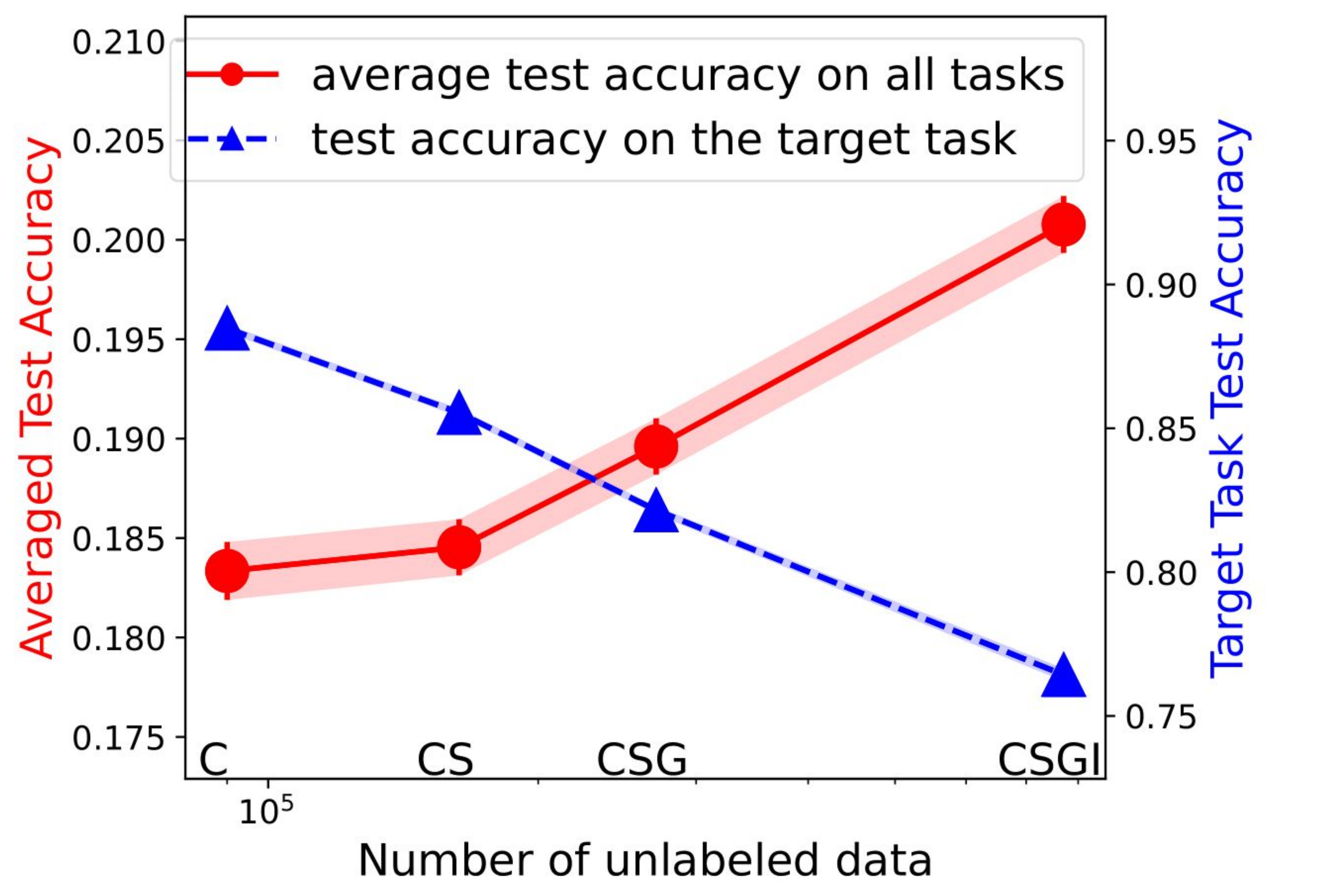
Dataset Target task CIFAR-10/Imagenet-Bird.

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 dataset.

Trade-off

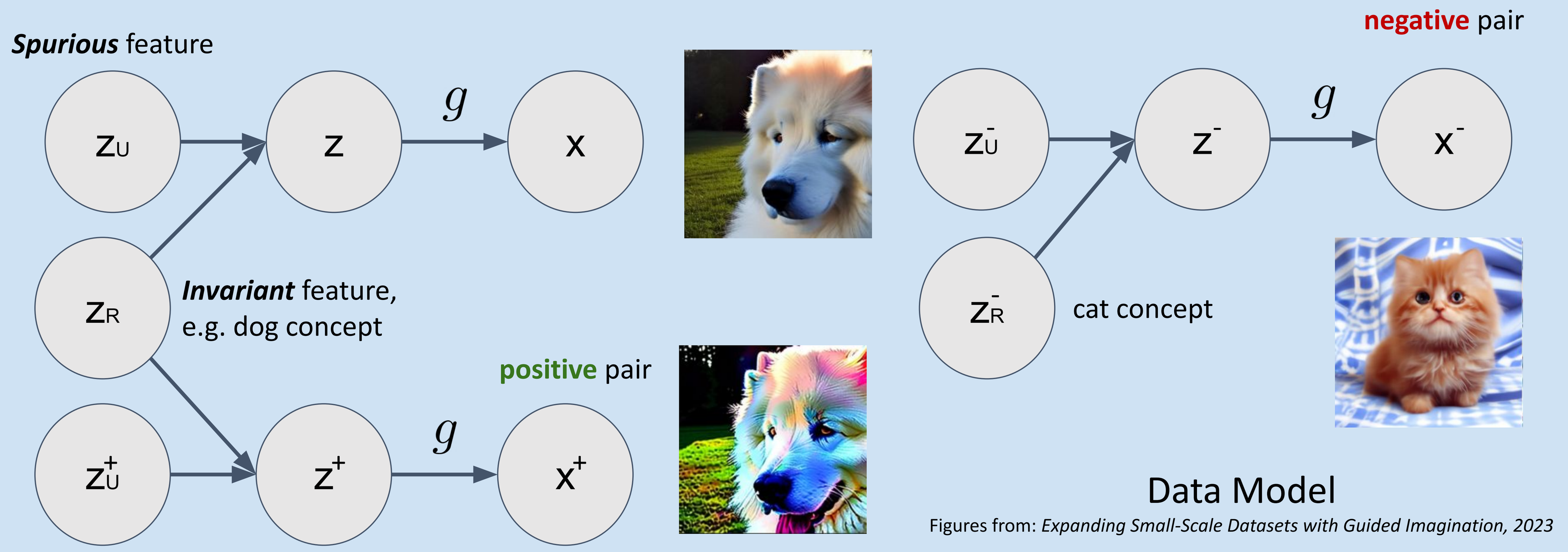
When pre-training dataset combined with more diverse data, the target task test accuracy decreases, while averaged test accuracy increases. As more diverse unlabeled data included, more labeled data from the target task is needed to achieve a comparably good target task test accuracy.



Problem Setup

Hidden representation data model

- Hidden representation space $z \in \mathcal{Z} \subseteq \mathbb{R}^d$ over distribution $\mathcal{D}_z$
- *Invariant* feature R, *Spurious* feature U, $R \cup U = [d]$, $R \cap U = \emptyset$
- $x = g(z)$, g is a generative function; y depends on z as well



Data Model

Figures from: *Expanding Small-Scale Datasets with Guided Imagination, 2023*

Contrastive learning and PCA

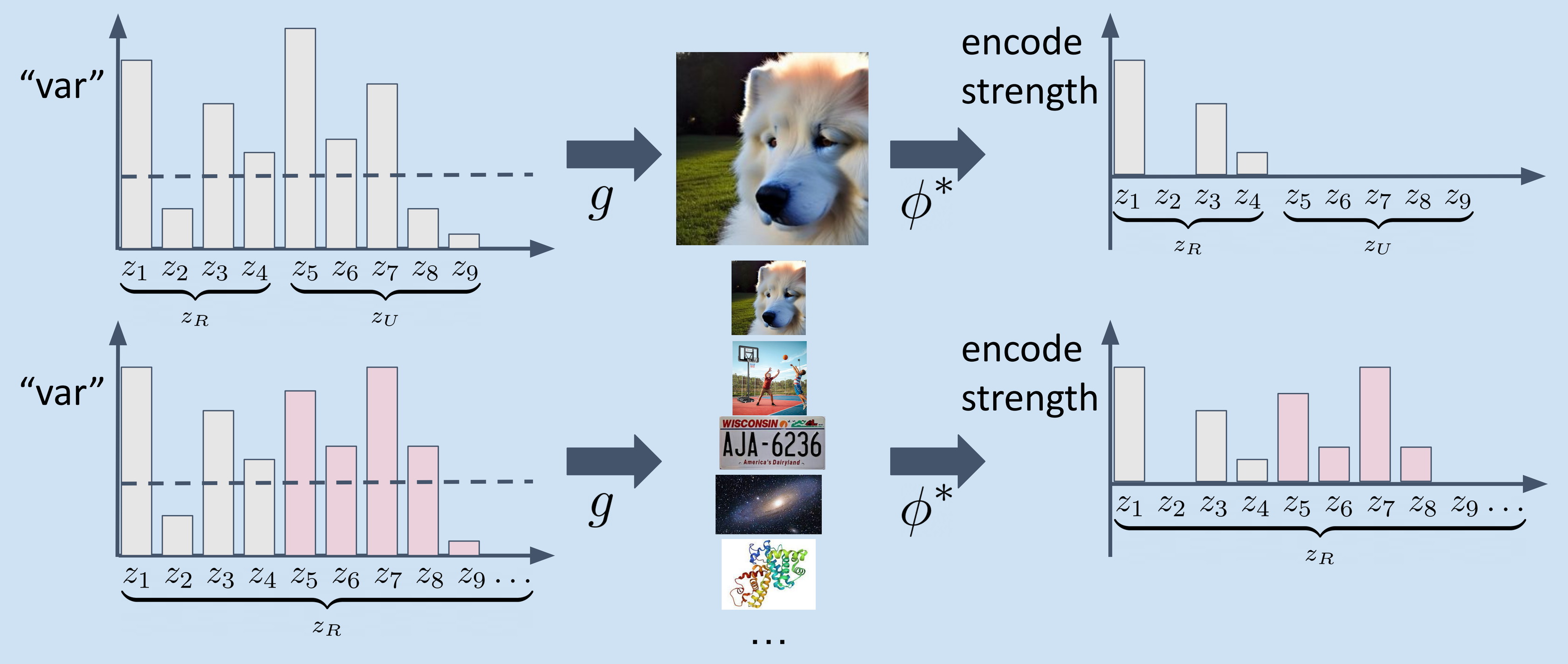
- $\phi \in \Phi$ hypothesis class of representation functions, e.g, ResNet, ViT
- **Contrastive Loss** $\min_{\phi \in \Phi} \mathbb{E}_{(x,x^+,x^-)\sim\mathcal{D}_{pre}} \left[\ell\left(\phi(x)^\top(\phi(x^+) - \phi(x^-))\right)\right]$
- In SimCLR, we have multiple negative pairs and $\ell(t) = \log(1+\exp(-t))$
- **PCA** on $\phi(x)$ $\min_{\phi\in\Phi} -\mathbb{E}_{x\sim\mathcal{D}}[\|\phi(x) - \mathbb{E}_{x'\sim\mathcal{D}}[\phi(x')]\|^2] = -\mathbb{E}_{x\sim\mathcal{D}}[\|\phi(x) - \phi_0\|^2]$
- $\phi_{z_R} := \mathbb{E}[\phi(x) \mid z_R] = \mathbb{E}[\phi(g(z)) \mid z_R]$

Theoretical Analysis

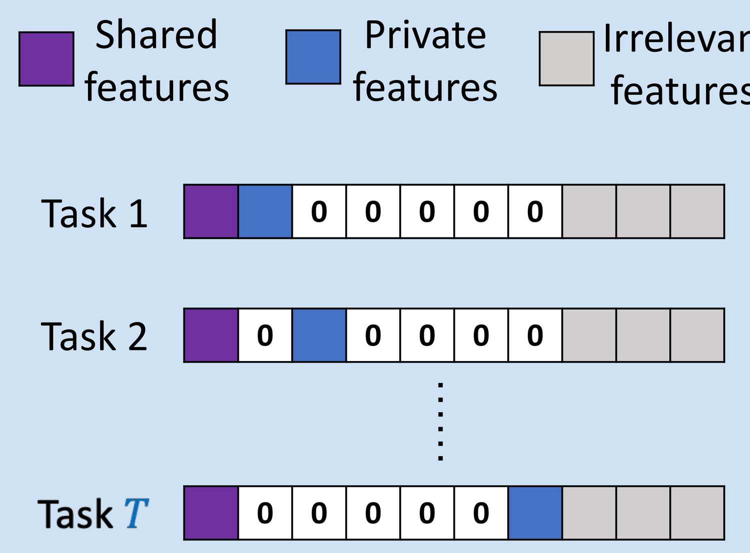
What features are learned by contrastive learning?

Theorem (Contrastive Learning is Generalized Nonlinear PCA)
If $\ell(t) = -t$, Contrastive Learning is equivalent to PCA on ϕ_{z_R}.
Moreover, if ϕ is linear function, it is equivalent to linear PCA on ϕ_{z_R}.

Theorem (Encode *Invariant* Feature; Remove *Spurious* Feature)
If $\ell(t)$ is convex, decrease, lower-bound, and $z_R \to x$ is one-to-one, with regular assumption, the optimal representation ϕ^* satisfies:
(1) ϕ^* does not encode *spurious* feature: $\phi^* \circ g(z) \perp z_U$
(2) ϕ^* only encodes *invariant* feature whose "variance" large enough, and encoding strength increases when "variance" becomes larger.



Trade-off comes from feature weighting



- Input: linearly generated from features
- Label: linear on shared/private features
- Pre-train a linear representation and then learn linear classifiers
- Best representation: weight shared/private features equally
- Pre-trained on Task 1:
  - Recover features for Task 1 but not for others
  - Good prediction on Task 1 but not on others
- Pre-trained on mixture of all tasks:
  - Recover all shared/private features
  - Up-weights the shared features by $O(\sqrt{T})$
  - $O(\sqrt{T})$ worse on Task 1 but better on average

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.