Investigating the Benefits of Projection Head for Representation Learning

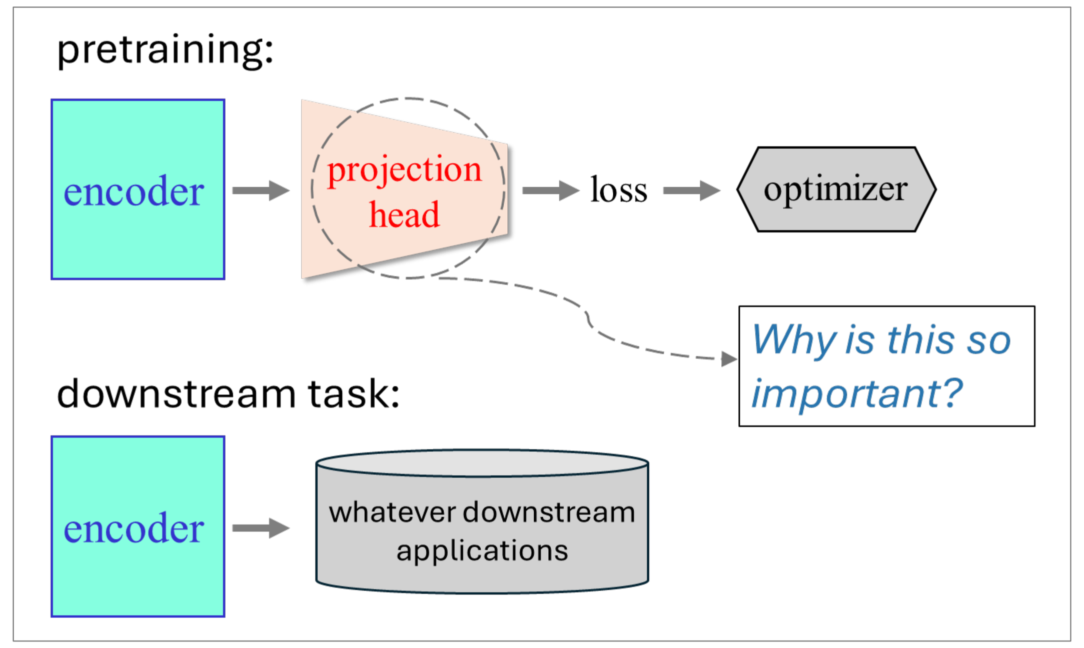
Yihao Xue, Eric Gan, Jiayi Ni, Siddharth Joshi, Baharan Mirzasoleiman

UCLA Samueli Computer Science

ICLR

Introduction

An effective technique for obtaining high-quality representations is **adding a projection head** on top of the encoder during pretraining, then **discarding** it and using the pre-projection representations for downstream tasks.



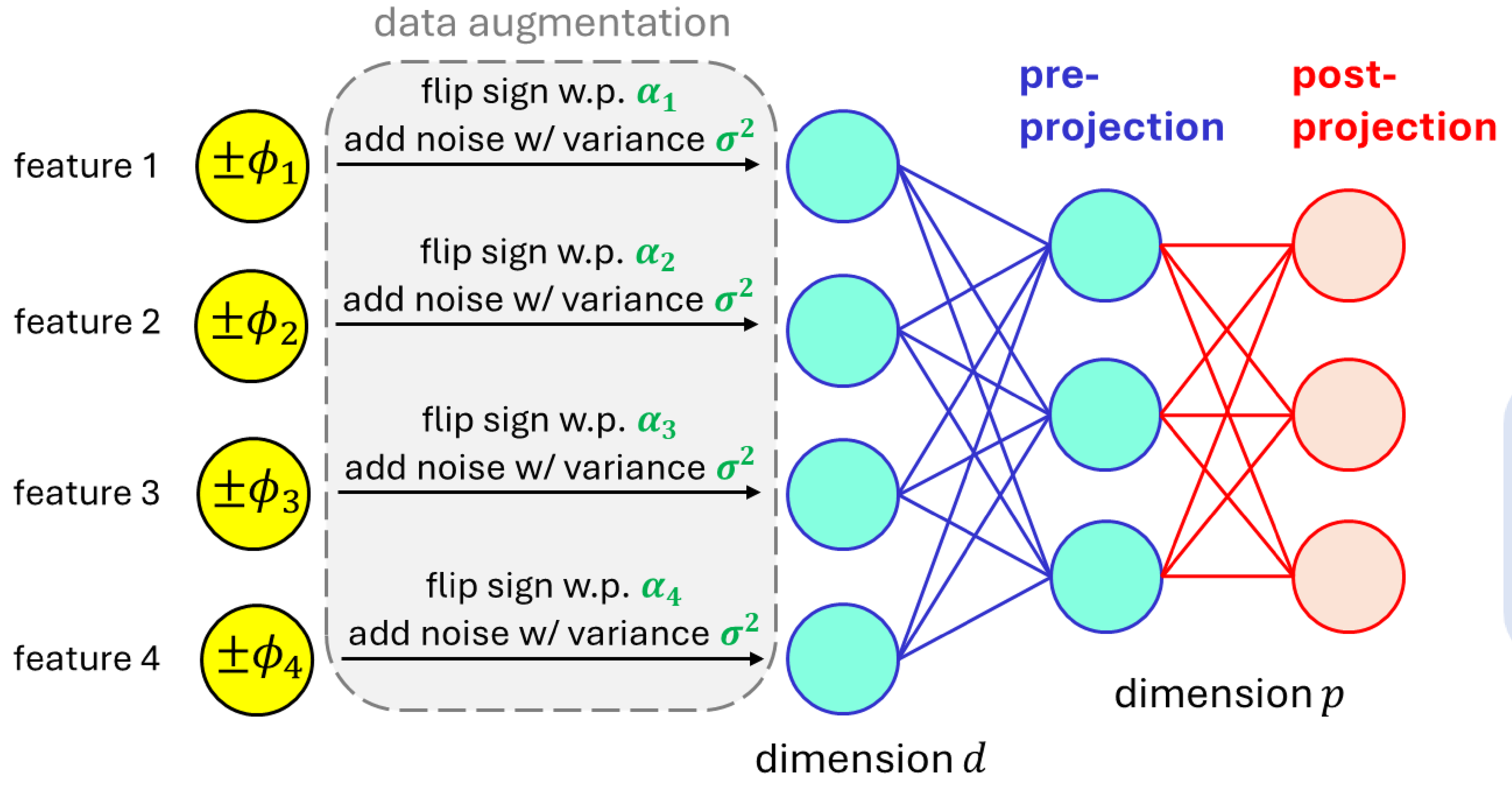
Our main conclusion: pre-projection representations represent input features more equally or capture a broader range of features compared to post-projection representations.

Self-supervised Contrastive Learning

Pretraining loss

max sim. b/w augmentations of the same examples
min sim. b/w augmentations of different examples

Simple data & linear model



Self-supervised Contrastive Learning (Cont'd)

Theorem: Define $\beta_i = \frac{(1-\alpha_i)^2\phi_i^2}{\phi_i^2+\sigma^2}$, $\gamma_i = \sqrt{\frac{(1-\alpha_i)\phi_i}{\phi_i^2+\sigma^2}}$

feature i is weighted by $\begin{cases} 0, & \text{if } \beta_i \text{ is not among the } p \text{ largest } p\ \beta'\text{s} \\ \boldsymbol{\gamma_i} \text{ pre-projection and } \boldsymbol{\gamma_i^2} \text{ post-projection} \end{cases}$

selection

weighting

Key insights

- The model selects and weights features based on the interplay between feature strengths (ϕ), noise (σ), and data augmentation (α)
- Features are weighted **more equally** pre-projection than post-projection

When is it beneficial to use pre-projection representations?

Assume that **feature i^*** is the only one useful for the downstream task. Ideally, pretraining should assign a large weight to it relative to other features. If this doesn't occur, i.e., pretraining assigns it a small weight (small γ_{i^*}), it would be better to use the pre-projection representations.

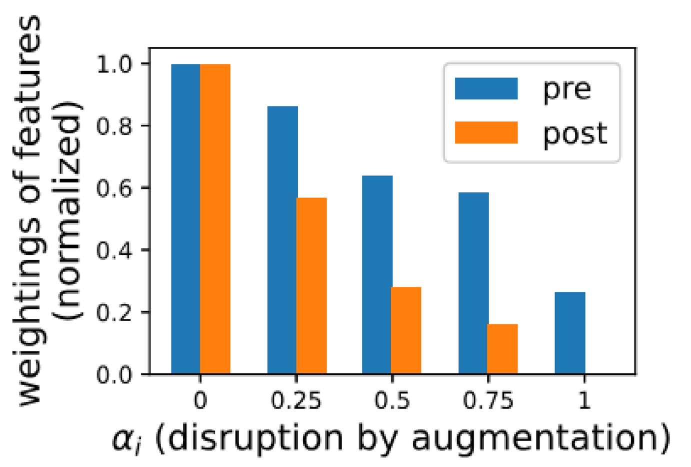
Here are some concrete scenarios considering the interaction between ϕ, σ, α

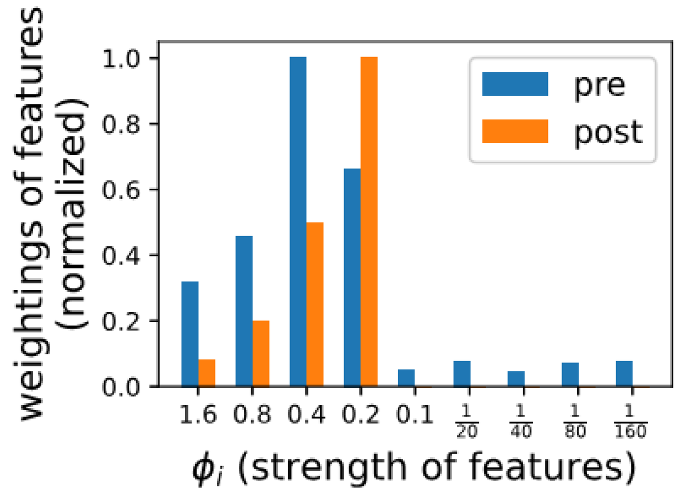
Corollary (Informal):
1. Data augmentation disrupts feature i^* too much (large $\alpha_{i^*} \to$ small γ_{i^*})
2. Feature i^* is too weak in pretraining data (very small $\phi_{i^*} \to$ small γ_{i^*})
3. Feature i^* is too strong in pretraining data (very large $\phi_{i^*} \to$ small γ_{i^*})

What about non-linear models?

Theorem (Informal): Non-linear models allow pre-projection representations to capture features that are **entirely absent** post-projection

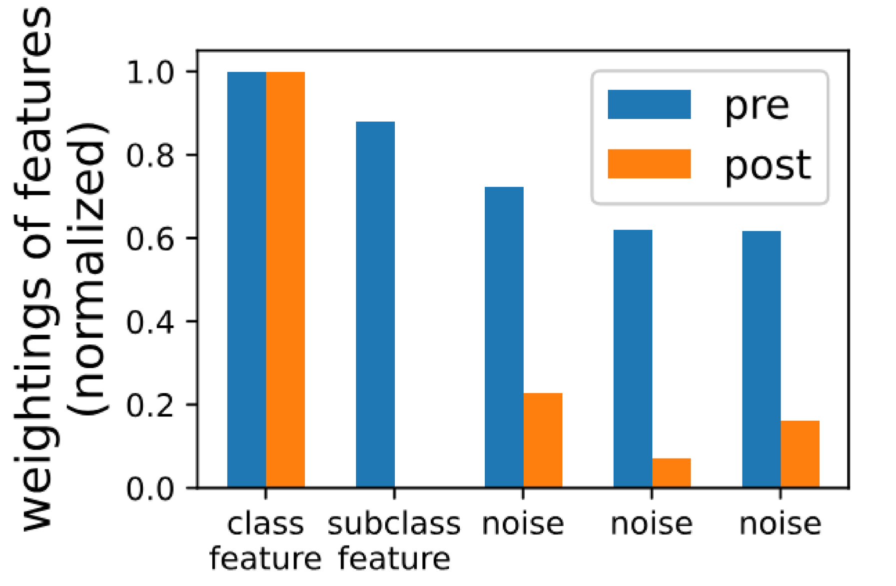
Experiments, toy example

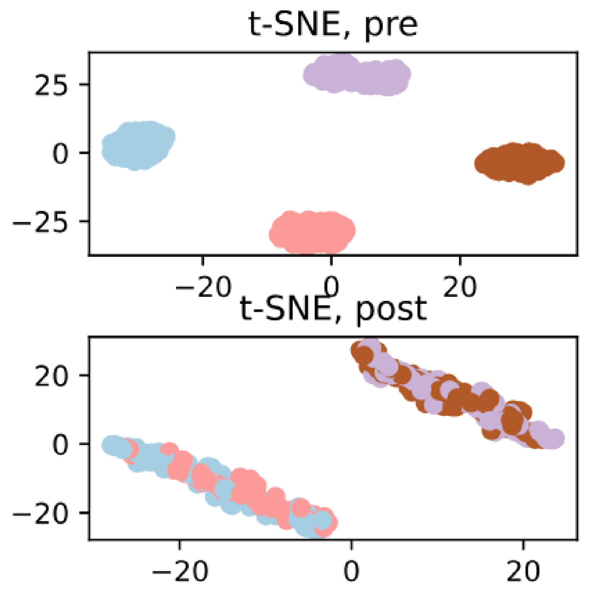




Supervised Contrastive Learning and Supervised Learning

Similar conclusions hold for both linear and non-linear models. Notably, the conclusions concerning non-linear models suggest that using pre-projection representations can mitigate **Neural/Class Collapse**, thereby enhancing *coarse-to-fine transferability*.

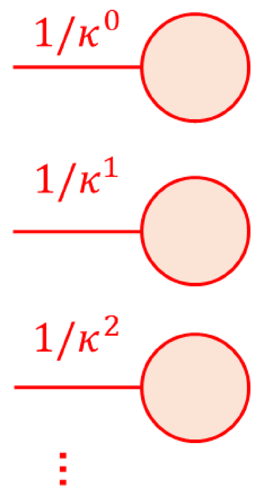




All subclasses are distinguishable.

Neural/Class Collapse: Subclasses within the same class are indistinguishable.

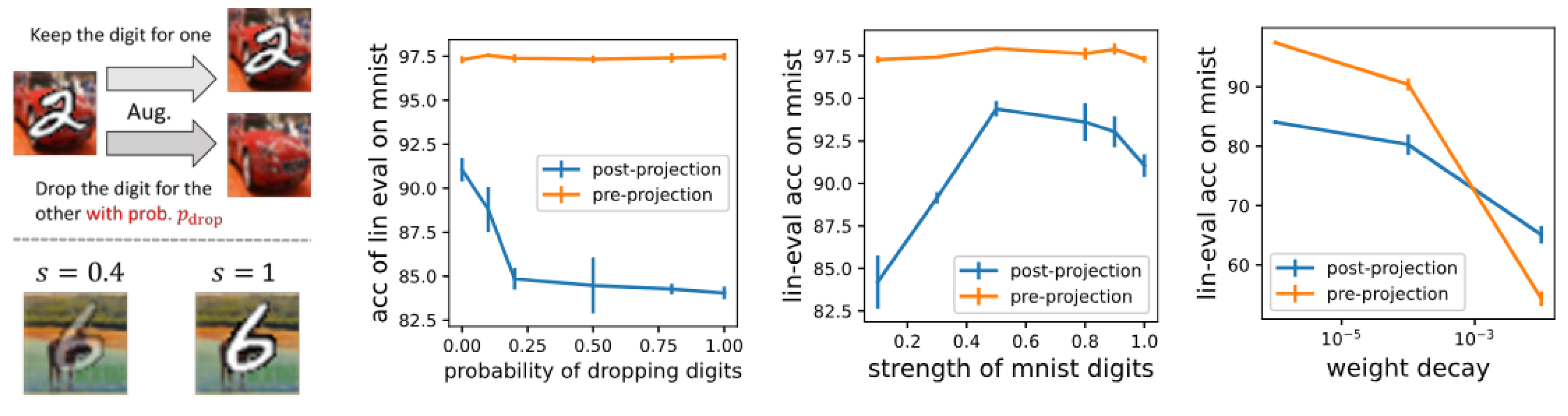
Replacing the Projection Head with a Fixed Reweighting Head



Scenario	Dataset	Alg.	Performance Measure	Performance		
				vanilla	proj	reweight
synthetic	M-on-C	SSCL	digit clf. acc.	77.0	97.3	97.3
coarse-to-fine	CIFAR100	SCL	fine-grained clf. acc.	21.8	36.0	30.2
coarse-to-fine	CIFAR100	SL	fine-grained clf. acc.	31.44	33.7	32.2
distribution shift	UrbanCars	SL	few-shot adaption acc.	82.2	86.1	87.0

The fixed reweighting head can achieve improvements that are comparable to those of the projection head across many tasks and algorithms.

Experiments, semi-synthetic



(a) An illustration (b) Effect of data aug. (c) Effect of strength (d) Effect of wd