Demystefying projection heads in contrastive learning: an expansion and shrinkage perspective

Yu Gui [1] Cong Ma [1] Yiqiao Zhong [2]

[1]Department of Statistics, University of Chicago [2]Department of Statistics, University of Wisconsin-Madison

Contrastive learning for unsupervised classification

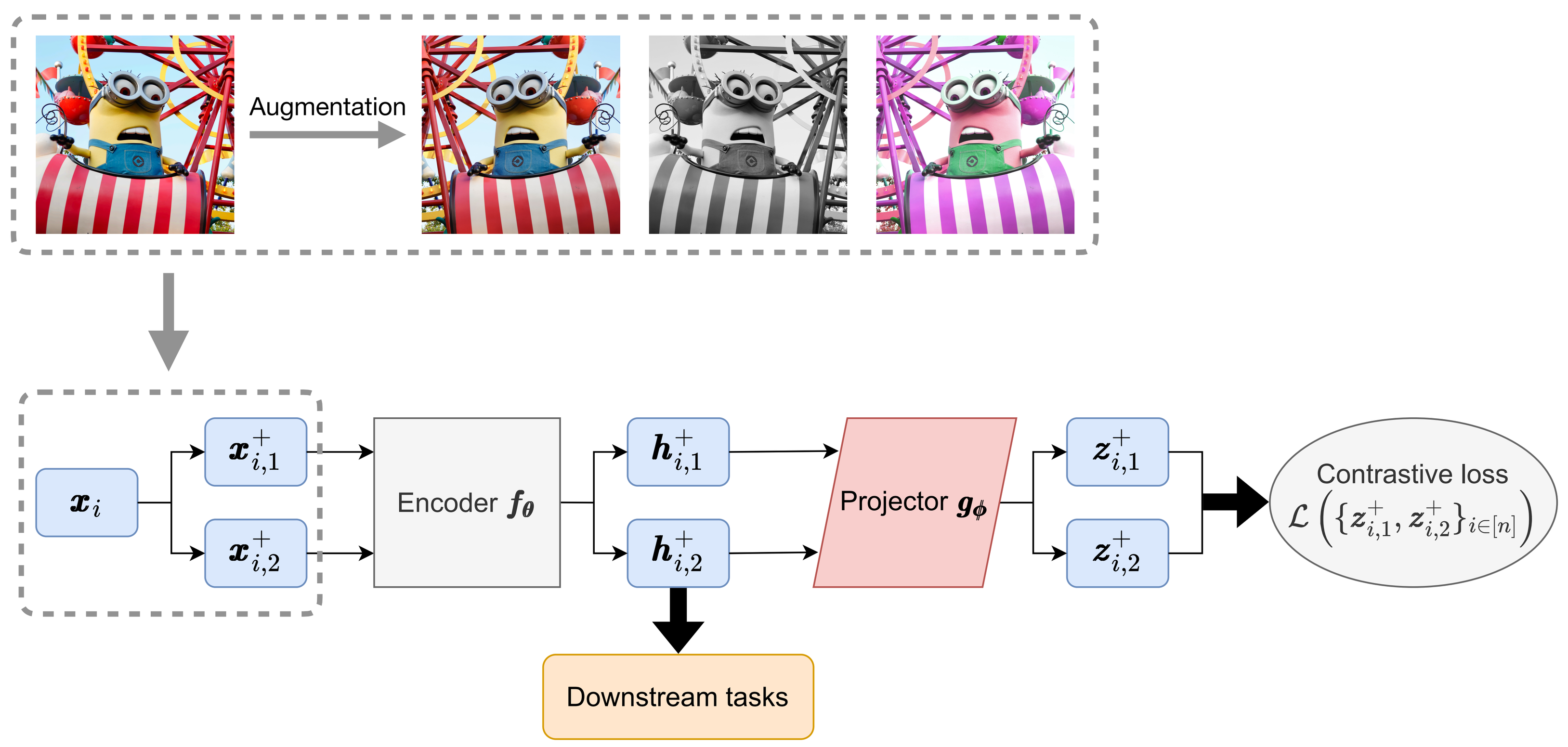

Figure 1. Encoder-projector framework.

- Positive pairs: $(\boldsymbol{x}_{i,k}^+, \boldsymbol{x}_{i,l}^+)$; Negative pairs: $(\boldsymbol{x}_{i,k}^+, \boldsymbol{x}_{j,l}^+)$, $i \neq j$
- Goal: learn representations by
  1. encouraging proximity between positive pairs
  2. forcing negative pairs to be far
- Contrastive loss (cross-entropy with pseudo labels)

$$\min_{\boldsymbol{\theta},\boldsymbol{\varphi}} \sum_i \sum_{j\sim i} -\log \frac{\exp(\frac{1}{\tau}\mathrm{sim}(\boldsymbol{z}_i,\boldsymbol{z}_j))}{\sum_{k\neq i}\exp(\frac{1}{\tau}\mathrm{sim}(\boldsymbol{z}_i,\boldsymbol{z}_k))}, \qquad \mathrm{sim}(\boldsymbol{z},\boldsymbol{z}') = \langle \frac{\boldsymbol{z}}{\|\boldsymbol{z}\|}, \frac{\boldsymbol{z}'}{\|\boldsymbol{z}'\|}\rangle$$

Motivating questions

1. Effect of contrastive learning on representation and role of hyperparameters?
2. Causes of dimensional collapse in both features and embeddings?
3. Role of the projector? (removed after training in practice)

Expansion and shrinkage of the signal

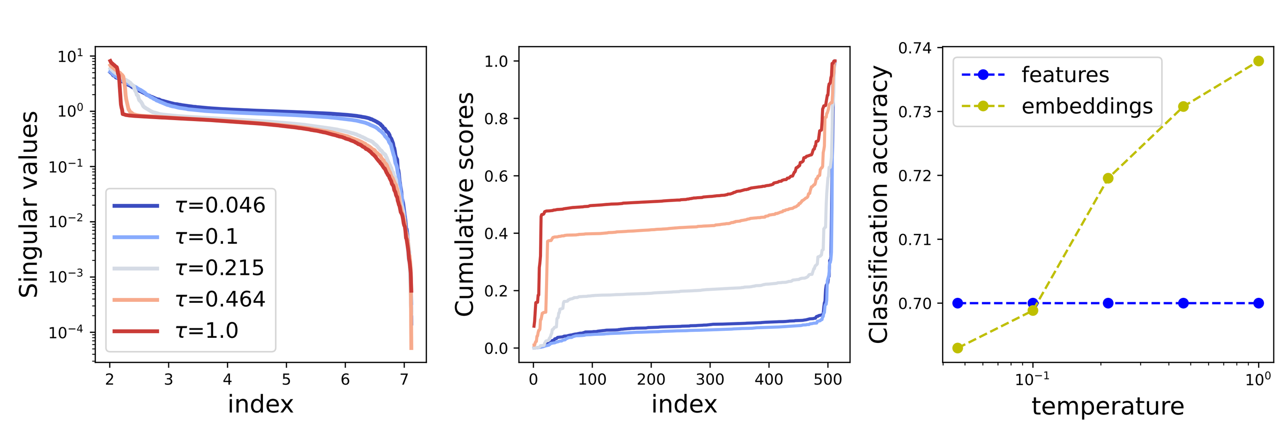

Figure 2. Results with the pretrained encoder and a one-layer linear projector.

$$\mathrm{score}_i = \sum_{j\le i} \frac{\langle \boldsymbol{v}_j, \boldsymbol{\mu}_{c_1,c_2}\rangle^2}{\|\boldsymbol{\mu}_{c_1,c_2}\|^2}$$

Feature-level GMM modeling

- 2-GMM features: $\boldsymbol{h}_i \overset{\text{i.i.d.}}{\sim} \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \mathbf{I}_p) + \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_p)$
- Augmentations: $\boldsymbol{h}_{i,1}^+, \boldsymbol{h}_{i,2}^+ \mid \boldsymbol{h}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{h}_i, \sigma_{\text{aug}}^2\mathbf{I}_p)$, $\boldsymbol{h}_i^- \overset{d}{=} \boldsymbol{h}_{j,1}^+$, $i \neq j$
- Linear projector: $\boldsymbol{z}_i = \mathbf{W}\boldsymbol{h}_i$
- Population loss:

$$\mathcal{L}(\mathbf{W}) = \frac{1}{2\tau}\cdot \frac{\mathbb{E}[\|\mathbf{W}\boldsymbol{h}_1^+ - \mathbf{W}\boldsymbol{h}_2^+\|^2]}{\left(\mathbb{E}[\|\mathbf{W}\boldsymbol{h}_1^+\|^2]\cdot\mathbb{E}[\|\mathbf{W}\boldsymbol{h}_2^+\|^2]\right)^{1/2}} + \log\left(\mathbb{E}\exp\left(-\frac{1}{2\tau}\cdot\frac{\|\mathbf{W}\boldsymbol{h}_1^+ - \mathbf{W}\boldsymbol{h}^-\|^2}{\left(\mathbb{E}[\|\mathbf{W}\boldsymbol{h}_1^+\|^2]\cdot\mathbb{E}[\|\mathbf{W}\boldsymbol{h}^-\|^2]\right)^{1/2}}\right)\right)$$

Expansion-shrinkage phase transition in GMM features

Denote $\tau^* = 2\|\boldsymbol{\mu}\|^2\left\{(1+\sigma_{\text{aug}}^2 + \|\boldsymbol{\mu}\|^2)\log(1+2\sigma_{\text{aug}}^2)\right\}^{-1}$. A three-parameter configuration $(\sigma_{\text{aug}}^2, \tau, \|\boldsymbol{\mu}\|^2)$ is said to be in the

- expansion regime if $\tau \ge \tau^*$ and shrinkage regime if $\tau < \tau^*$.

Theorem 1

Consider minimizer $\mathbf{W}^*$ of certain first-order approximation $\widetilde{\mathcal{L}}(\mathbf{W})$.

- When $\tau \ge \tau^*$ (expansion regime) , $\mathbf{W}^* = \sum_j \sigma_j^* \boldsymbol{u}_j^* \boldsymbol{v}_j^{*\top}$ satisfies

$$\sigma_2^* = \cdots = \sigma_p^* = 0, \qquad \langle \boldsymbol{v}_1^*, \boldsymbol{\mu}\rangle^2 = \|\boldsymbol{\mu}\|^2 \quad \text{i.e., perfect alignment}$$

- When $\tau < \tau^*$ (shrinkage regime),

$$\text{if } \sigma_{\text{aug}}^2 \to 0, \qquad \text{then } \max_j |\sigma_j \langle \boldsymbol{v}_j^*, \boldsymbol{\mu}\rangle| \to 0 \quad \text{i.e., compress if correlated}$$

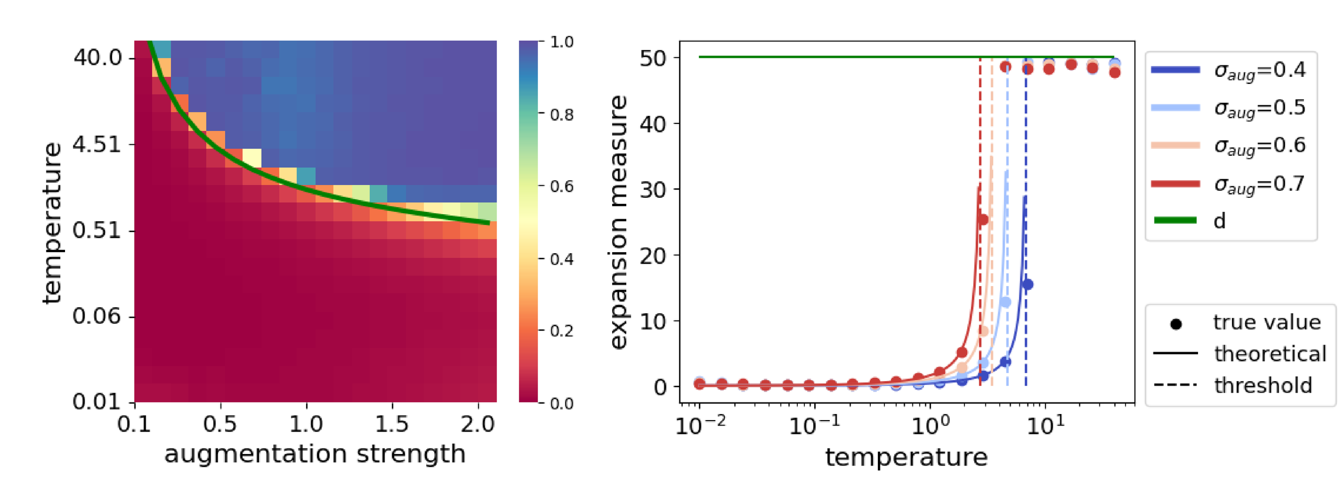

Figure 3. Expansion measure $\widetilde{t}(\mathbf{W}) = \|\mathbf{W}\boldsymbol{\mu}\|^2/(\|\mathbf{W}\|_F^2\|\boldsymbol{\mu}\|^2)$.

Empirical evidence for feature-level modeling

1. Linear separable features after *a few* epochs
2. Contrastive loss decomposition at each epoch t,

$$\mathcal{L}(\boldsymbol{\theta}^{(t)}, \boldsymbol{\varphi}^{(t)}) = \min_{\boldsymbol{\varphi}} \mathcal{L}(\boldsymbol{\theta}^{(t)}, \boldsymbol{\varphi}) + \mathcal{L}^\perp(\boldsymbol{\theta}^{(t)}, \boldsymbol{\varphi}^{(t)}),$$

which satisfies $\mathcal{L}^\perp(\boldsymbol{\theta}^{(t)}, \boldsymbol{\varphi}^{(t)}) \ll \min_{\boldsymbol{\varphi}} \mathcal{L}(\boldsymbol{\theta}^{(t)}, \boldsymbol{\varphi})$ and

$$\|\widetilde{\boldsymbol{\varphi}}^{(t)} - \boldsymbol{\varphi}^{(t)}\| \ll \|\boldsymbol{\varphi}^{(t)}\|, \qquad \widetilde{\boldsymbol{\varphi}}^{(t)} = \mathrm{argmin}_{\boldsymbol{\varphi}} \mathcal{L}(\boldsymbol{\theta}^{(t)}, \boldsymbol{\varphi})$$

Effect of projectors on generalization accuracy

- **Motivating question**: *how does expansion/shrinkage affect generalization in downstream tasks?*
- Consider the linear projection from the following class:

$$\mathcal{W} = \left\{\mathbf{W}_\eta = \mathbf{I}_p + \eta\cdot\rho^{-1}\boldsymbol{\mu}\boldsymbol{\mu}^\top : \eta > -1\right\}, \text{ where } \rho = \|\boldsymbol{\mu}\|^2$$

Invariance of generalization error in low-dimensional regime

- ℓ_2-regularized logistic regression for non-separable data

$$\ell_n(\gamma, \boldsymbol{\beta}; \lambda_n) = \mathbb{E}_n\left\{\log\left[1 + e^{-y\left(\gamma + \boldsymbol{z}^\top\boldsymbol{\beta}\right)}\right]\right\} + \lambda_n\|\boldsymbol{\beta}\|^2$$

- Classification error $\mathtt{Err}(\gamma, \boldsymbol{\beta}; \eta, \lambda_n) = \mathbb{P}(\gamma + y'\langle \boldsymbol{z}', \boldsymbol{\beta}\rangle < 0)$

Proposition

Let $(\widehat{\gamma}, \widehat{\boldsymbol{\beta}})$ be the minimizer of $\ell_n(\gamma, \boldsymbol{\beta}; \lambda_n)$.

1. If $\lambda_n = a\cdot b_n > 0$ with constant $a > 0$ and $0 < b_n \ll \sqrt{n}$, then the test error

$$\mathtt{Err}\left(\widehat{\gamma}, \widehat{\boldsymbol{\beta}}; \eta, \lambda_n\right) = \Phi(-\|\boldsymbol{\mu}\|) + O_{\mathbb{P}}(b_n n^{-1/2}).$$

2. If $\lambda_n = a\sqrt{n}$ with constant $a > 0$, then $\mathtt{Err}\left(\widehat{\gamma}, \widehat{\boldsymbol{\beta}}; \eta, \lambda_n\right)$ is decreasing in η.

Decreasing generalization error in high-dimensional regime

- Implicit bias in *overparametrized* models: GD for logistic regression converges to max-margin classifier for separable data

$$\max_{\boldsymbol{\beta}} \quad \min_{i\le n} y_i\langle \boldsymbol{z}_i, \boldsymbol{\beta}\rangle \qquad \text{subject to} \quad \|\boldsymbol{\beta}\| \le 1$$

- Classification error $\mathtt{Err}(\widehat{\boldsymbol{\beta}}; \eta) = \mathbb{P}(y'\langle \boldsymbol{z}', \widehat{\boldsymbol{\beta}}\rangle < 0)$
- A linear layer $\boldsymbol{z}_i = \mathbf{W}\boldsymbol{h}_i$ can be interpreted as *reparametrization*

Theorem 2

Suppose $n/p \to \delta > 0$. There exists threshold $\delta^*(\rho) > 0$ such that

- (separability) if $\delta < \delta^*$, there exists a unique solution $\widehat{\boldsymbol{\beta}}$ with the margin

$$\widehat{\kappa} = \min_{i\le n} y_i\langle \boldsymbol{z}_i, \widehat{\boldsymbol{\beta}}\rangle \xrightarrow{p} \kappa^*(\|\boldsymbol{\mu}\|, \eta) > 0$$

and conversely data are not separable w.h.p. if $\delta > \delta^*$.

- (monotone error) if $\delta < \delta^*$, the asymptotic error $\mathtt{Err}^*(\eta)$, namely

$$\mathtt{Err}\left(\widehat{\boldsymbol{\beta}}; \eta\right) \xrightarrow{p} \mathtt{Err}^*(\eta),$$

is decreasing in η.

References

[1] Yu Gui, Cong Ma, and Yiqiao Zhong. Demystefying projection heads in contrastive learning: an expansion and shrinkage perspective. *In preparation*, 2023.

yugui@uchicago.edu