Harnessing small projectors & multiple views for efficient vision pretraining

Arna Ghosh*[,1,2], Kumar Krishna Agrawal*[,3], Shagun Sodhani[4], Adam Oberman+[,1,2], Blake Richards+[,1,2,5,6]

McGill University, Canada [1], Mila - Quebec AI Institute, Canada[2], UC Berkeley, USA[3], Meta FAIR, Canada[4], Montréal Neurological Institute, Canada [5], Learning in Machines and Brains Program, CIFAR, Canada[6]

Summary

Recent progress in self-supervised (SSL) visual representation learning has led to development of several frameworks, that typically rely on augmentations of images and leverage different loss formulations.

How can we improve the sample and compute efficiency of current SSL pipelines, enabling cheaper pretraining?

We build on theory and recent analytical results to design practical recommendations for competitive and efficient SSL:

➔ Promote **higher orthogonalization** among learned features

➔ Use **multiple views** of each image to optimize for invariance criterion

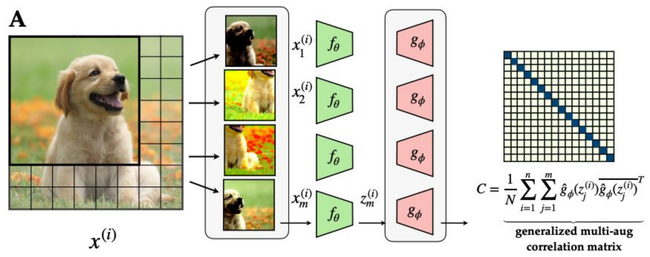

Fig 1: Proposed multi-view SSL pipeline

Methods

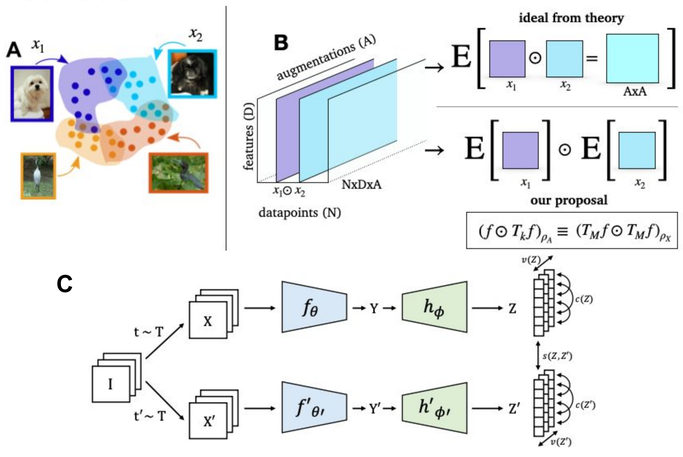

Fig 2: Design of existing SSL algorithms relies on heuristics. (A) Augmentation graphs are common in vision pretraining, providing generalizable features for downstream tasks. (B) We propose an equivalent loss function for SSL pretraining that recovers the same eigenfunctions more efficiently than existing approaches. (C) A canonical framework for non-contrastive SSL, here VICReg (Bardes et al. 2021).

Reformulated SSL objective:

$$L(F) = \sum_{i=1}^{N_k} \|T_M f_i - f_i\|^2_{L^2(\rho_X)}, \quad \textit{subject to} \quad (f_i, f_j)_{\rho_X} = \delta_{ij}$$

Results

1. Implicit bias of gradient descent on feature learning dynamics

$$\hat{L}(F) = \mathbf{E}_x\left[\sum_i \sum_j \|\overline{f_i(x))} - f_i(x_j)\|^2\right] + \beta \sum_i \sum_{i' \neq i} \mathbf{E}_x[\|f_i(x) f_{i'}(x)\|]^2$$

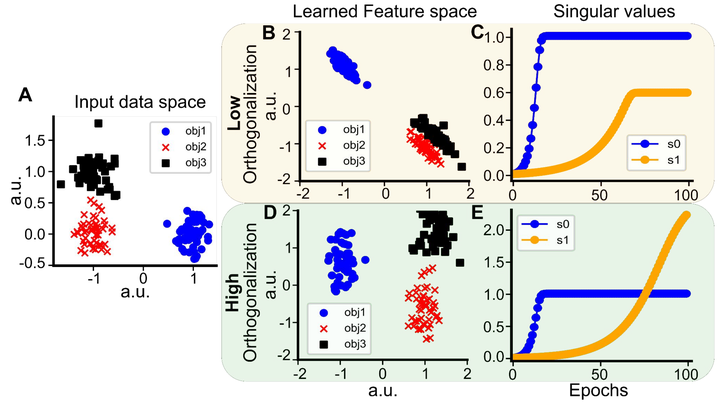

Fig 3: Understanding the feature learning dynamics of SSL. (A) 2D inputs, point clouds corresponding to different input augmentations. (B, D) Learned feature space under low and high orthogonalization constraint, respectively. (C,E) Corresponding singular values of the learned feature space.

2. Low-dimensional projectors can yield good representations

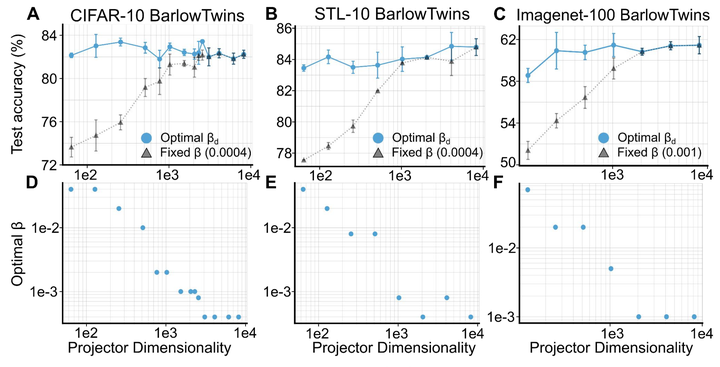

Fig 4: Higher orthogonality constraint, β, for lower projector dimensionality achieved similar performance over a wide range of projector dimensions across (A) CIFAR-10, (B) STL-10, and (C) Imagenet-100 datasets.

Start with low-dimensional projector, using β = O(1/ pdim)

3. Multiple augmentations improve performance and convergence

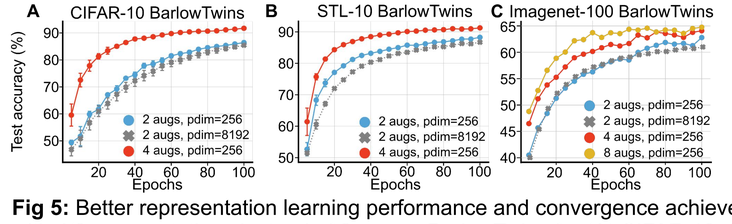

Fig 5: Better representation learning performance and convergence achieved with 4/8 augmentations instead of 2 across BarlowTwins for (A) CIFAR-10, (B) STL-10, and (C) Imagenet-100 pretraining.

4. Multi-augmentation improves sample efficiency

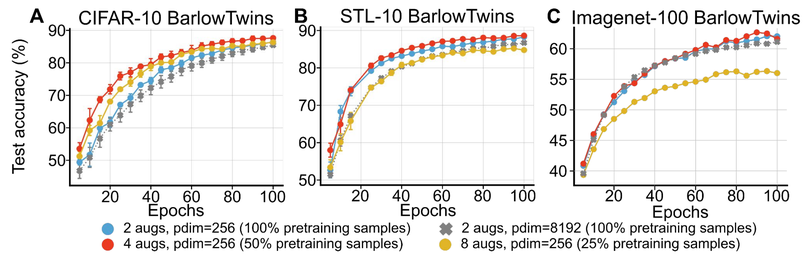

Fig 6: Similar representation learning performance achieved with significantly fewer unique samples in the pretraining dataset across BarlowTwins for (A) CIFAR-10, (B) STL-10, and (C) Imagenet-100 pretraining.

In a low-data regime, using diverse & multiple augmentations can be as effective as acquiring more unique samples.

Moving the Pareto frontier by leveraging multiple augmentations

Runtime-performance Pareto frontier: Increasing the number of pretraining samples *increases runtime* but also yields better features, i.e. *lower error rate* on downstream tasks.

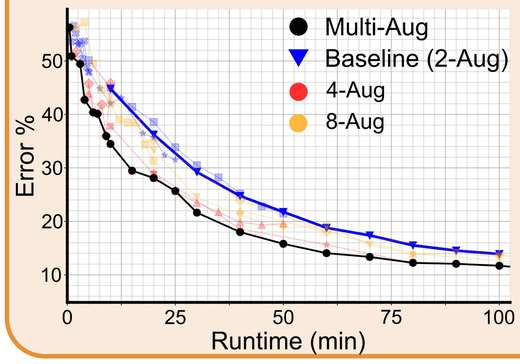

Fig 7: Using more than 2 augmentations (4, 8 or more) with a fraction of the samples in the pretraining dataset improves overall Pareto frontier. Pretraining runtime can be improved up to ~2 times by using multiple augmentations, compared to the standard 2-augmentation SSL pipeline.

Open Problems

- Can we improve **theoretical understanding** of feature learning in Joint Embedding Predictive Architecture (JEPA) pipelines?
- Beyond autoregressive losses: How do we improve **pretraining efficiency** of autoregressive and JEPA models for vision and other modalities?

References

Agrawal, et al. 2022 "α-ReQ : Assessing Representation Quality in Self-Supervised Learning by measuring eigenspectrum decay".

Bardes, et al. 2022 "VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning".

Simon, et al. 2023 " On the stepwise nature of self-supervised learning".

Zbontar, et al. 2021 "Barlow Twins: Self-Supervised Learning via Redundancy Reduction."

Zhai et al. 2024 "Understanding Augmentation-based Self-Supervised Representation Learning via RKHS Approximation and Regression".

Acknowledgements

1. Mila compute cluster

2. Vanier Canada Graduate scholarship (AG); Healthy Brains, Healthy Lives (AG & BAR)

3. NSERC, Grant No. RGPIN-2020-05105 and RGPAS2020-00031 (BAR); Arthur B. McDonald Fellowship: 566355-2022)

4. Canada CIFAR AI Chair program (AO & BAR).

Harnessing small projectors & multiple views for efficient vision pretraining

Arna Ghosh*[,1,2], Kumar Krishna Agrawal*[,3], Shagun Sodhani[4], Adam Oberman+[,1,2], Blake Richards+[,1,2,5,6]

McGill University, Canada [1], Mila - Quebec AI Institute, Canada[2], UC Berkeley, USA[3], Meta FAIR, Canada[4], Montréal Neurological Institute, Canada [5], Learning in Machines and Brains Program, CIFAR, Canada[6]

Summary

Recent progress in self-supervised (SSL) visual representation learning has led to the development of several different proposed frameworks that rely on augmentations of images but use different loss functions. We build on theory and recent analytical results to design practical recommendations for competitive and efficient SSL, by demonstrating the following:

- The idealized loss in SSL frameworks can be reformulated to a functionally equivalent loss that is **more efficient to compute**.
- Due to the implicit bias of using gradient descent to minimize reformulated loss function, **a stronger orthogonalization constraint** with a reduced projector dimensionality is necessary to yield good representations.
- The linear readout performance when training a ResNet-backbone on CIFAR, STL and Imagenet datasets can be improved with **multiple augmentations**, thereby improving the dynamics of feature learning.
- Data augmentations allow **reducing the pretraining dataset size** by up to 2× while maintaining downstream accuracy simply by using more data augmentations.

Methods

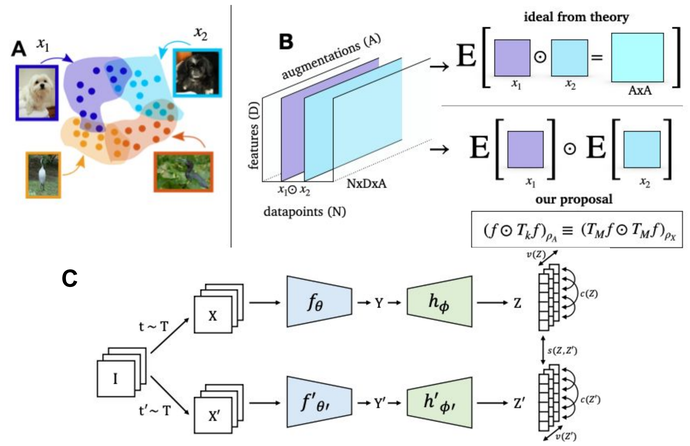

Fig 1: Design of existing SSL algorithms relies on heuristics. (A) Augmentation graphs are common in vision pretraining, providing generalizable features for downstream tasks. (B) We propose an equivalent loss function for SSL pretraining that recovers the same eigenfunctions more efficiently than existing approaches. (C) A canonical framework for non-contrastive SSL, here VICReg (Bardes et al. 2021)

Reformulated SSL objective:

$$L(F) = \sum_{i=1}^{N_k} \|T_M f_i - f_i\|^2_{L^2(\rho_X)}, \quad \textit{subject to} \quad (f_i, f_j)_{\rho_X} = \delta_{ij}$$

Results

1. Implicit bias of gradient descent on feature learning dynamics

$$\hat{L}(F) = \mathbf{E}_x \left[\sum_i \sum_j \|\overline{f_i(x))} - f_i(x_j)\|^2 \right] + \beta \sum_i \sum_{i' \neq i} \mathbf{E}_x [\|f_i(x) f_{i'}(x)\|]^2$$

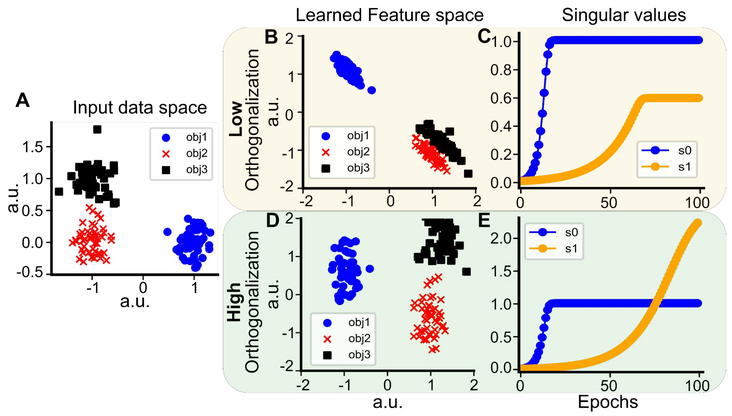

Fig 2: Understanding the feature learning dynamics of SSL. (A) 2D inputs, with each point indicating an augmented version of an input. (B, D) Learned feature space under low and high orthogonalization constraint, respectively. (C,E) Corresponding singular values of the learned feature space.

2. Low-dimensional projectors can yield good representations

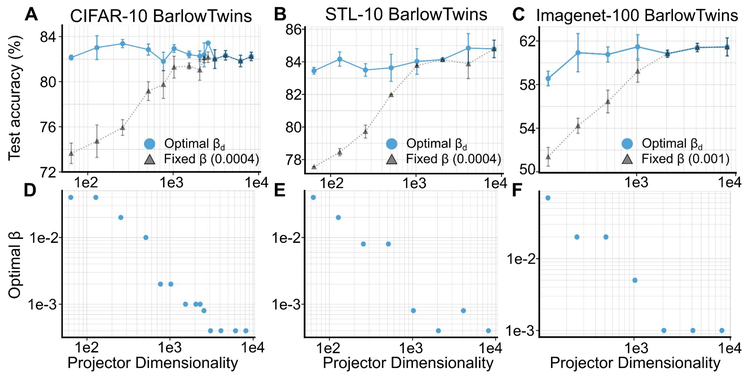

Fig 3: Higher orthogonality constraint, β, for lower projector dimensionality achieved similar performance over a wide range of projector dimensions across (A) CIFAR-10, (B) STL-10, and (C) Imagenet-100 datasets.

3. Multiple augmentations improve performance and convergence

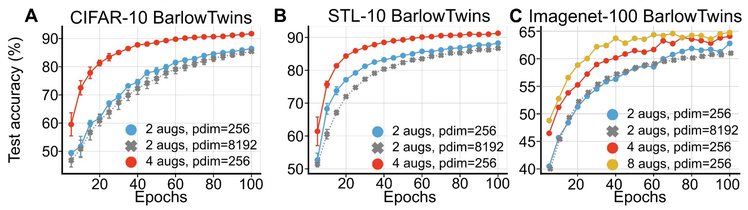

Fig 4: Better representation learning performance and convergence achieved with 4 augmentations instead of 2 across BarlowTwins for (A) CIFAR-10, (B) STL-10, and (C) Imagenet-100 pretraining.

4. Multi-augmentation improves sample efficiency

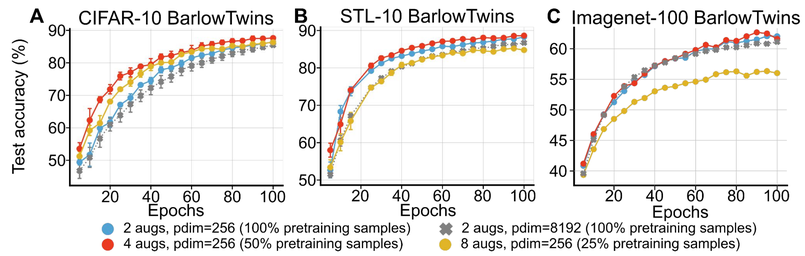

Fig 5: Similar representation learning performance achieved with significantly fewer unique samples in the pretraining dataset across BarlowTwins for (A) CIFAR-10, (B) STL-10, and (C) Imagenet-100 pretraining.

5. A specific target error level can be achieved with either a larger pretraining dataset or more augmentations

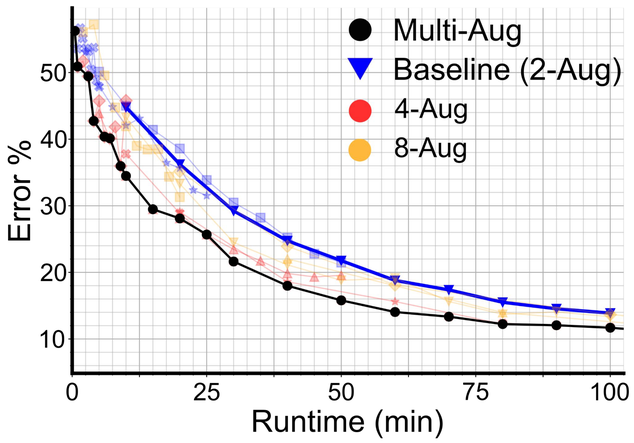

Fig 6: Using more than 2 augmentations with a fraction of the dataset improves overall Pareto frontier, sped runtime up to ~ 2×.

Open Problems

H.

References

Agrawal, et al. 2022 "α-ReQ : Assessing Representation Quality in Self-Supervised Learning by measuring eigenspectrum decay"

Bardes, et al. 2022 "VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning"

Simon, et al. 2023 " On the stepwise nature of self-supervised learning"

Zbontar, et al. 2021 "Barlow Twins: Self-Supervised Learning via Redundancy Reduction."

Zhai et al. 2024 "Understanding Augmentation-based Self-Supervised Representation Learning via RKHS Approximation and Regression

Acknowledgements

1. Mila compute cluster
2. Vanier Canada Graduate scholarship (AG); Healthy Brains, Healthy Lives (AG & BAR)
3. NSERC, Grant No. RGPIN-2020-05105 and RGPAS2020-00031 (BAR); Arthur B. McDonald Fellowship: 566355-2022)
4. Canada CIFAR AI Chair program (AO & BAR).