Master Thesis: Masking Principal Components for Discriminative Self-supervised Representation Learning

Alice Bizeul (alice.bizeul@inf.ethz.ch)

December 26, 2024

1 Motivation

Self-supervised representation learning (SSL) (Balestriero et al., 2023) has emerged as a cornerstone of representation learning in recent years. Models such as OpenAI's CLIP (Radford et al., 2021) demonstrate how SSL approaches can produce expressive representations applicable to a broad spectrum of downstream tasks. This paradigm relies on paired observations—whether paired views or modalities sharing the same content—to extract meaningful features.

Broadly, SSL methods fall into two categories: discriminative and generative (or reconstruction-based). Discriminative SSL (Chen et al., 2020) aims to ensure that representations of paired observations are closer in latent space than those of randomly sampled observations. In contrast, reconstruction-based SSL (He et al., 2022) involves reconstructing one observation from its pair. In multi-view settings, data augmentation techniques, such as image cropping and color jittering, are commonly used to artificially create paired observations from single ones. Among these augmentations, image cropping has proven especially impactful, driving advancements in visual learning models like Meta's DINO (Caron et al., 2021; Oquab et al., 2023) and JEPA (Assran et al., 2023).

Recent studies (Bizeul et al., 2024)[1] suggest that in the image domain, masking—conceptually similar to cropping—principal components rather than individual image pixels can generate image pairs that foster the learning of expressive features in reconstruction-based SSL. In this project, we aim to investigate whether applying a similar approach to discriminative SSL can yield comparable benefits, focusing specifically on methods like DINO, JEPA and SigLIP (Zhai et al., 2023).

2 Research Questions

Below, we outline intriguing research questions that the student can further develop and refine if time allows. With the initial research questions clearly defined and prior work demonstrating significant performance improvements in reconstruction-based SSL, the goal is for this thesis to culminate in a publication, either in an ICML 2025 workshop[2] or as a NeurIPS 2025 [3] full conference paper. The student will also benefit from existing code-bases from which to start.

- **Principal component masking as a data augmentation method**: Principal component masking in reconstruction-based self-supervised learning involves dividing the set of principal components into two disjoint subsets that capture $r\%$ and $1-r\%$ of the data's total variance, thereby creating two complementary views from a single observation. The parameter r is a hyperparameter, optimal when set to 20% across datasets (Bizeul et al., 2024) for reconstruction-based SSL. This project aims to investigate whether this approach remains optimal for discriminative SSL or if it needs adaptation to enhance downstream performance.
- **Impact on coarse downstream tasks**: For widely used discriminative SSL methods (e.g., SimCLR (Chen et al., 2020), VicReg (Bardes et al., 2021), MoCo (Chen et al., 2021), DINO (Caron et al., 2021; Oquab et al., 2023), JEPA (Assran et al., 2023), SigLIP (Zhai et al., 2023)), we aim

[1] https://alicebizeul.github.io/assets/pdf/mae.pdf
[2] https://icml.cc
[3] https://neurips.cc/Conferences/2024

to assess whether replacing standard augmentation strategies with principal component masking leads to performance improvements on coarse downstream tasks like image classification.

- **Impact on dense downstream tasks**: Discriminative SSL methods typically excel in coarse tasks such as image classification, due to their learning of representations that are invariant to data augmentations. However, they are often less effective for dense tasks like object detection or segmentation, which benefit more from representations learned via reconstruction-based SSL. This project will explore whether principal component masking can boost performance in dense tasks for discriminative SSL.
- **Comparison with multi-modal representations**: Representations learned by CLIP leverage large-scale multi-modal text-image datasets, making them highly effective for coarse downstream tasks. If the preceding research questions reveal that principal component masking positively impacts both coarse and dense tasks using image-only representations, we intend to compare its performance with CLIP representations and investigate ways to enhance CLIP representations using principal component masking.

3 Goals

- **Literature Review**: The student will familiarize themselves with the field of discriminative self-supervised learning by conducting a review of relevant methods in self-supervised representation learning. This will include understanding their advantages and limitations. The articles listed in this proposal provide a solid starting point for exploration.
- **Develop a Codebase**: Extend and adapt the existing codebase[4] to incorporate popular discriminative self-supervised learning methods.
- **Optimize Principal Component Masking**: Perform ablation studies to determine the most effective adaptations of principal component masking for discriminative SSL, aiming to optimize downstream performance on high-level tasks.
- **Train Baselines**: Train baseline models for discriminative SSL, applying their respective data augmentation techniques, and evaluate their performance on both broad and detailed tasks.
- **Analyze Results**: Evaluate and compare the performance of discriminative SSL methods using standard augmentations versus principal component masking. Draw conclusions about the robustness and effectiveness of principal component masking for visual representation learning, and conduct a detailed analysis of both positive and negative findings.
- **Write a Report**: Compile a comprehensive report that details the methodology, findings, and conclusions from the study.

4 Why you should consider this project?

- Engage in a project with a **clear motivation and well-defined goals.**
- Acquire **hands-on experience** with learning algorithms that form the foundation of representation learning and multi-modal learning, a key focus **in current AI research**.
- Contribute to a project inspired by recent advancements in reconstruction-based self-supervised learning (SSL) and **explore innovative ways to extend this approach** to other learning paradigms.
- Collaborate with a team that values knowledge sharing, supports idea exploration, and encourages **working toward a publication if the results are promising.**

If you are interested, please contact Alice Bizeul at alice.bizeul@inf.ethz.ch and include your grades, resume, and a brief introduction explaining your motivation.

[4] https://github.com/alicebizeul/pmae

References

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023.

Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, Avi Schwarzschild, Andrew Gordon Wilson, Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash, Yann LeCun, and Micah Goldblum. A cookbook of self-supervised learning, 2023. URL https://arxiv.org/abs/2304.12210.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Alice Bizeul, Thomas M. Sutter, Alain Ryser, Julius Von Kügelgen, Bernhard Schölkopf, and Julia E. Vogt. Components beat patches: Eigenvector
masking for visual representation learning. *under submission*, 2024.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986, October 2023.