EXPLORING DIMENSIONAL COLLAPSE IN SELF-SUPERVISED VIDEO REPRESENTATION LEARNING

Paul Kapust **Monika Kwiatkowski** **Olaf Hellwich** **Patrik Reiske**[*]
Dept. of Computer Vision and Remote Sensing, Technische Universität Berlin, Berlin, Germany

ABSTRACT

In the field of joint embedding methods, the complete collapse to a constant feature vector is a clear indication of an immediate deficiency in the approach. Another critical concern, known as dimensional collapse, describes the utilization of a feature space only to a lower-dimensional subspace. Despite extensive efforts to address complete collapse through various preventive strategies, dimensional collapse remains largely unexplored. This paper aims to bridge this gap by extending the examination of dimensional collapse to video representation learning. Our source code is publicly available.[1]

1 INTRODUCTION AND BACKGROUND

Recent approaches in self-supervised image representation learning, like SimCLR (Chen et al., 2020), MoCo (He et al., 2020), and SwAV (Caron et al., 2021), aim to obtain expressive representations through invariance of augmented views of data samples. Thus, the shared semantic information content from different views of the same sample must be maximized. For joint embedding approaches, like the ones mentioned before, the encoder is followed by a projection head (see appendix A.1) to map the encoder's representation space to the projection head's embedding space (Chen et al., 2020). The embedding feature vectors are then subject to the learning objective. The trivial solution to this objective is constant output vectors—a complete collapse of the embedding space. State-of-the-art approaches prevent such a complete collapse with different strategies; for instance, SimCLR contrasts positive and negative views to obstruct constant vectors.

However, preventing a complete collapse of the embedding space is not synonymous with learning an embedding space with high information density. Hua et al. (2021) were the first to describe the phenomenon of dimensional collapse, where some embedding dimensions are highly correlated, resulting in the utilization of a lower-dimensional subspace within the embedding space. Some approaches, like Barlow Twins (Zbontar et al., 2021) and VICReg (Bardes et al., 2022), formulate their learning objective such that the correlation of embedding dimensions is penalized.

Since the representation vectors are used for transfer tasks rather than the embedding vectors, the utilization of the representation space is of high relevance. It can be hypothesized that a dimensional collapse in the representation space of an approach restricts its potential. Jing et al. (2022) report that using a projection head in contrastive learning enhances the performance and mitigates dimensional collapse in the encoder's representation space.

To illustrate the degree of dimensional collapse, Jing et al. (2022) apply singular value decomposition on the sample covariance matrix of feature vectors. In doing so, they visualize the (sorted) spectrum of the resulting singular values on a logarithmic scale. Li et al. (2022) apply a similar procedure to inspect dimensional collapse in non-contrastive settings. Additionally to examining the spectrum of singular values, they visualize their cumulative explained variance curve and also report the area under curve (AUC) as another measure. Both methods estimate the rank of a given feature space.

Our literature search only revealed work that addresses dimensional collapse in the context of image representation learning. To our knowledge, we are the first to examine video representation learning approaches for dimensional collapse.

[*] Corresponding author of this work; patrik.reiske@tu-berlin.de

[1] https://github.com/paulkapust/dimcollapse-video

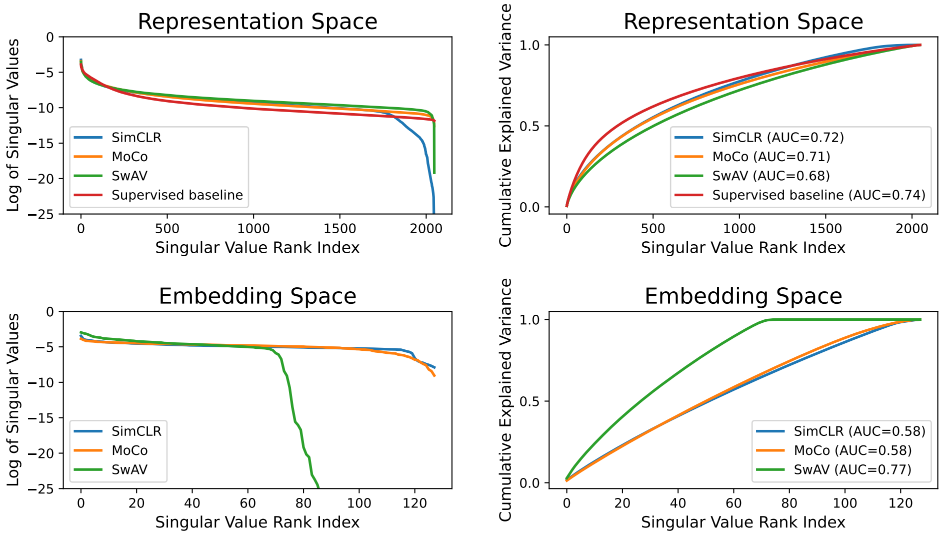

Figure 1: Singular value spectra and cumulative explained variance curves of self-supervised video representation learning approaches for both the encoders' representation spaces and the projection heads' embedding spaces. A SlowOnly-50 model learned under supervision serves as a baseline.

2 Experiments and Results

We have conducted follow-up experiments to the works of Jing et al. (2022) and Li et al. (2022) for state-of-the-art video representation learning approaches. To that end, we used publicly available, pretrained (fixed) models from the video domain,[2] specifically the adaptations of SimCLR (Chen et al., 2020), MoCo (He et al., 2020), and SwAV (Caron et al., 2021) as proposed by Feichtenhofer et al. (2021). We apply the exact methodology proposed by Jing et al. (2022) and Li et al. (2022) on the validation split of the Kinetics-400 (Kay et al., 2017) dataset, incorporating SlowOnly-50 (Feichtenhofer et al., 2019) as the designated spatiotemporal encoder architecture.

Figure 1 visualizes the (sorted) spectrum of singular values as well as the corresponding cumulative curves and AUCs for the encoders' representation spaces and projection heads' embedding spaces. There, a sudden, steep drop in the logarithm of the singular values, as well as a steep rise of the cumulative curve or a high AUC, all indicate a dimensional collapse. Our results show no evidence of (relevant) dimensional collapse in any of the representation spaces. Surprisingly, the supervised baseline encoder does not seem to acquire a representation space with a higher information density than the self-supervised approaches. We neither find indication of a (relevant) dimensional collapse in the embedding spaces of SimCLR (Chen et al., 2020) nor MoCo (He et al., 2020). However, the embedding space of SwAV (Caron et al., 2021) shows a strong indication of dimensional collapse.

In additional experiments, we discovered that this behavior does not vary in domain transfer settings but does for different approaches.[3]

3 Conclusion

Our findings support the hypothesis by Jing et al. (2022) that projection heads mitigate a dimensional collapse of the representation space. Our analysis further reveals that state-of-the-art self-supervised video representation learning approaches are at least as resilient to dimensional collapse as their supervised baseline. This strongly suggests that self-supervised methods can learn more information-dense feature spaces than their supervised counterparts.

[2] We refer interested readers to appendix A.2 for an overview of the selected models; the pretrained models and configuration details are available at `https://github.com/facebookresearch/SlowFast`

[3] Results of the experiments mentioned can be found in appendices A.3, A.4, and A.5 to this work.

URM Statement

Both the first author, Paul Kapust, and the second author, Monika Kwiatkowski, meet the URM criteria of the ICLR 2024 Tiny Papers Track.

References

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning, 2022.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations, 2020.

Christoph Feichtenhofer, Axel Pinz, and Richard P. Wildes. Spatiotemporal Residual Networks for Video Action Recognition, 2016.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast Networks for Video Recognition, 2019.

Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A Large-Scale Study on Unsupervised Spatiotemporal Representation Learning, 2021.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised Learning, 2020.

Tengda Han, Weidi Xie, and Andrew Zisserman. Video Representation Learning by Dense Predictive Coding, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning, 2020.

Tianyu Hua, Wenxiao Wang, Zihui Xue, Sucheng Ren, Yue Wang, and Hang Zhao. On Feature Decorrelation in Self-Supervised Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9598–9608, October 2021.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding Dimensional Collapse in Contrastive Self-supervised Learning, 2022.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The Kinetics Human Action Video Dataset, 2017.

H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: A large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pp. 2556–2563, 2011. doi: 10.1109/ICCV.2011.6126543.

Alexander C. Li, Alexei A. Efros, and Deepak Pathak. Understanding Collapse in Non-Contrastive Siamese Representation Learning, 2022.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild, 2012.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow Twins: Self-Supervised Learning via Redundancy Reduction, 2021.

A Appendix

A.1 Embedding architecture

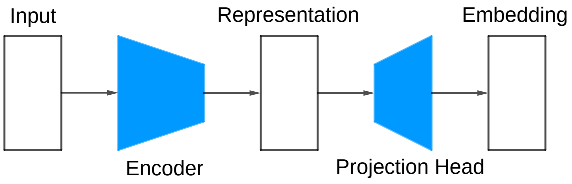

Figure 2: Sketch of an embedding architecture.

The sketch of an embedding architecture depicted in figure 2 above illustrates the relation between the encoder's representation space and the projection head's embedding space. During pretraining a multilayer perceptron, termed projection head, is used to project the representations to a embedding space. Based on these projections the model is fit using some self-supervised learning objective. This concept was first proposed by Chen et al. (2020). Once the encoder is pretrained it can be transfer learned to solve some other task, termed downstream task. This is usually done using human-generated labels for supervision.

A.2 Overview of the Pretraining Setup

For our experiments we have used the publicly available,[4] pretrained (fixed) models of the video domain adaptions (Feichtenhofer et al., 2021) of SimCLR (Chen et al., 2020), MoCo (He et al., 2020), SwAV (Caron et al., 2021), and BYOL (Grill et al., 2020). All of these models use SlowOnly (Feichtenhofer et al., 2016; 2019) as spatiotemporal encoder architecture, employ 8x8 sampling and $\rho = 2$ temporally separated clips, on the Kinetics-400 (Kay et al., 2017) dataset for pretraining over 200 epochs. SimCLR (Chen et al., 2020), MoCo (He et al., 2020), and SwAV (Caron et al., 2021) were used in both the main part of this work as well as its appendix, while we used BYOL (Grill et al., 2020) only in the additional experiments. In these additional experiments we also used a publicly available,[5] pretrained (fixed) DPC (Han et al., 2019) model. DPC uses a video-domain adaption of ResNet-34 (He et al., 2015) as encoder architecture, different than the models mentioned before.

Further details on the pretraining setups can be found in the linked repositories.

A.3 Dimensional Collapse in Domain Transfer Settings

In additional experiments we tested the video domain adoptions (Feichtenhofer et al., 2021) of SimCLR (Chen et al., 2020), MoCo (He et al., 2020), and SwAV (Caron et al., 2021) for dimensional collapse (Hua et al., 2021) in a domain transfer setting: the publicly available, pretrained (fixed) models were originally fit to the Kinetics-400 (Kay et al., 2017) dataset and have here been separately tested on both the HMBD51 (Kuehne et al., 2011) and the UFC101 (Soomro et al., 2012) dataset.

Figures 3, 4, and 5 visualize the (sorted) spectrum of singular values as well as the corresponding cumulative curves and AUCs for the encoders' representation spaces and projection heads' embedding spaces for the different approaches, as figure 1 did in the main part of this work. We find that the methods tested again do not show any (relevant) dimensional collapse in this setting either. This strongly indicates that the phenomenon of dimensional collapse (Hua et al., 2021) is not directly linked to domain transfer, and is also another testament for the methods' ability to generalize in such a setting.

[4] https://github.com/facebookresearch/SlowFast

[5] https://github.com/TengdaHan/DPC

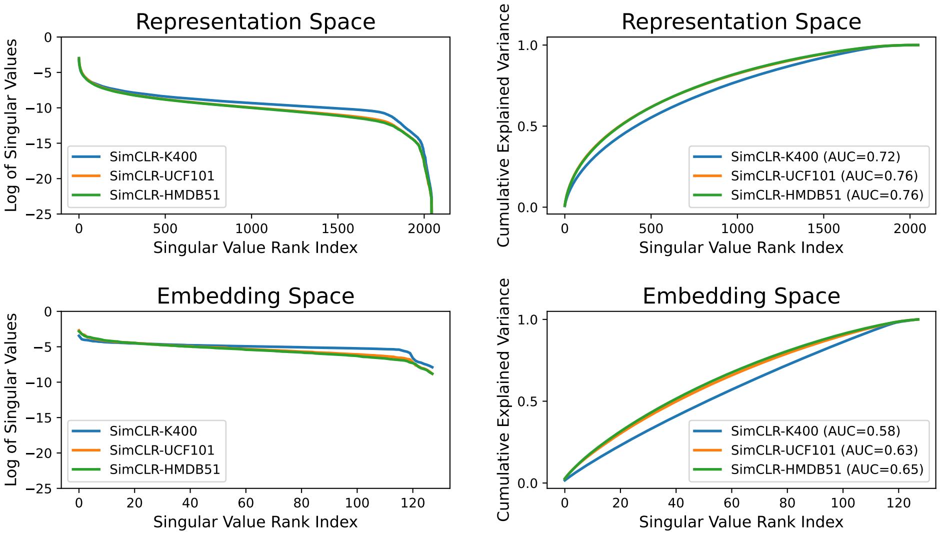

Figure 3: Singular value spectra and cumulative explained variance curves of the video domain adaption of SimCLR for both the encoder's representation space and the projection head's embedding space in a domain transfer setting, Kinetics-400 to HMBD51 and UCF101 respectively.

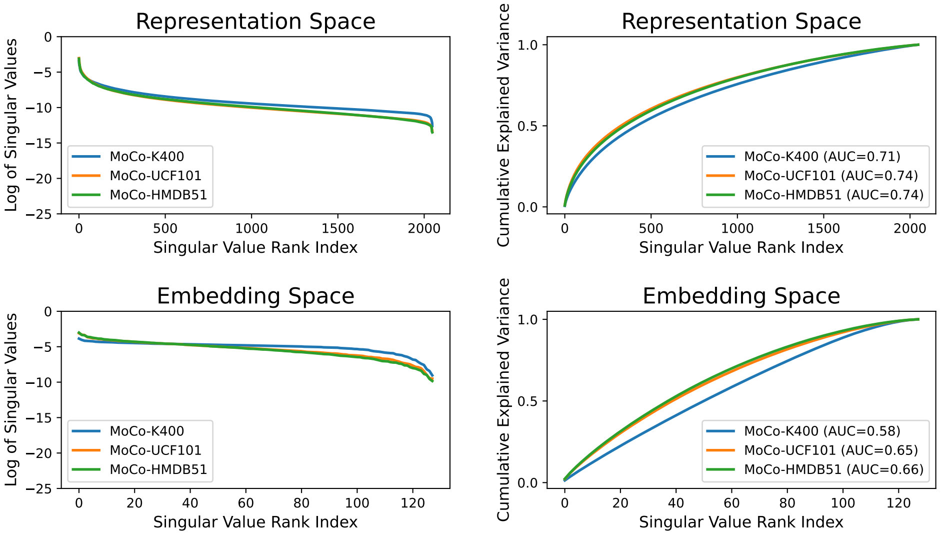

Figure 4: Singular value spectra and cumulative explained variance curves of the video domain adaption of MoCo for both the encoder's representation space and the projection head's embedding space in a domain transfer setting, Kinetics-400 to HMBD51 and UCF101 respectively.

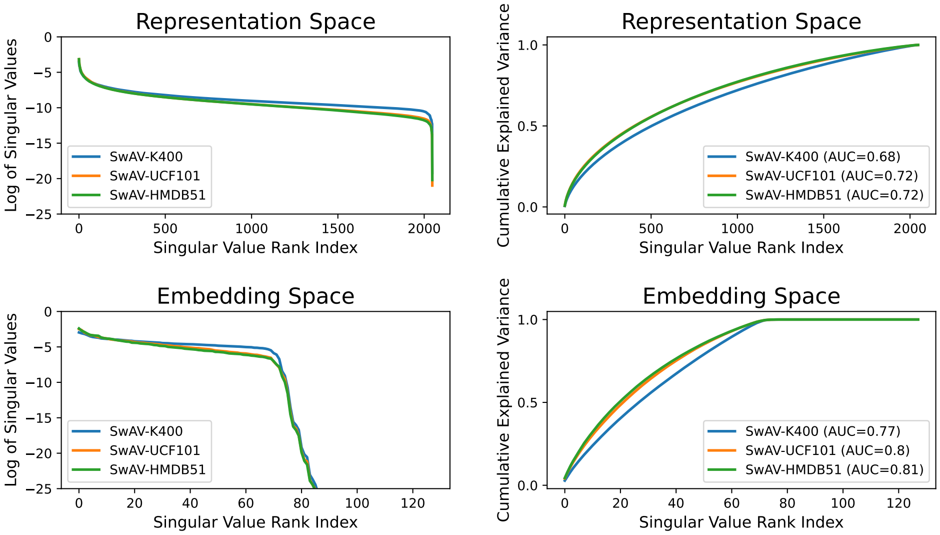

Figure 5: Singular value spectra and cumulative explained variance curves of the video domain adaption of SwAV for both the encoder's representation space and the projection head's embedding space in a domain transfer setting, Kinetics-400 to HMBD51 and UCF101 respectively.

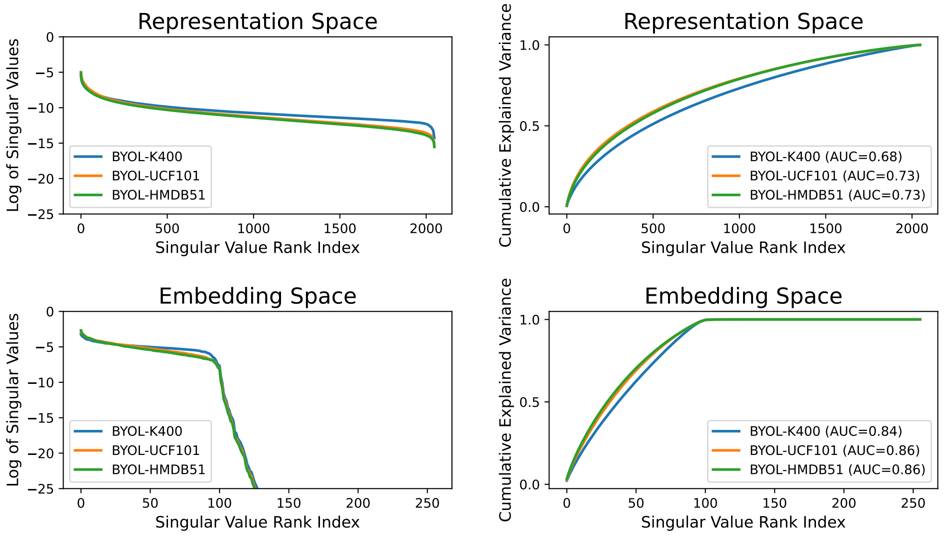

Figure 6: Singular value spectra and cumulative explained variance curves of the video domain adaption of BYOL for both the encoder's representation space and the projection head's embedding space in a domain transfer setting, Kinetics-400 to HMBD51 and UCF101 respectively.

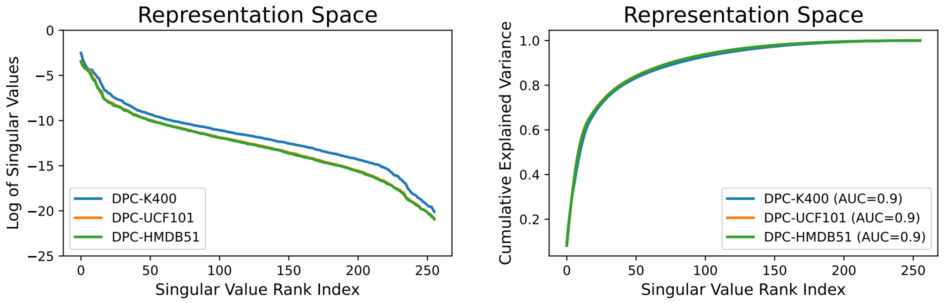

Figure 7: Singular value spectra and cumulative explained variance curves of DPC for the representation space in a domain transfer setting, Kinetics-400 to HMBD51 and UCF101 respectively.

A.4 Video-Based BYOL

The analysis of the video domain adaption (Feichtenhofer et al., 2021) of BYOL (Grill et al., 2020) was not included in figure 1 for readability, because its embedding space is higher-dimensional.

Figure 6 shows that the sensibility of video-based BYOL (Grill et al., 2020; Feichtenhofer et al., 2021) for dimensional collapse (Hua et al., 2021) is analogous to video-based SwAV (Caron et al., 2021; Feichtenhofer et al., 2021) in figure 1. There we find strong indication of a dimensional collapse (Hua et al., 2021) in the embedding space but not for the representation space in both the training domain and domain transfer settings.

A.5 DPC

The analysis of the DPC (Han et al., 2019) was not included in figure 1 for readability, because its representation space is lower-dimensional. Also, DPC does not make use of a projection head and is (other than the others) a generative approach.

Figure 7 shows that the representation space of DPC (Han et al., 2019) suffers from dimensional collapse (Hua et al., 2021) in both the training domain and domain transfer settings.