Dimensional Collapse in Video Representation Learning

Master's Thesis Proposal

In 1992, Becker et al. were probably the first to propose self-supervised learning representations from signals derived from the data itself. Since then, the field has been growing steadily, with publications by Devlin et al. (2018), Feichtenhofer et al. (2019), and Brown et al. (2020) being some of the most prominent examples related to learning video representations in recent years. These approaches match or even exceed the performance of other supervised learning approaches.

It is common practice to evaluate the performance of an approach by its accuracy on a defined downstream task. This often overlooks the problem of dimensional collapse, as described by Hua et al. (2020). In addition, Mialon et al. (2022), Zbontar et al. (2021), and others address this issue to varying degrees, as it affects whether an approach can be used to its full potential. However, counteracting or even just testing for dimensional collapse has not yet become standard practice.

Inspired by these and other related publications, we hypothesize that there may be (large) differences in how effectively different approaches use their representation spaces — verifying our hypothesis is the topic of the proposed thesis.

The student working on this thesis will conduct a comprehensive literature review of related work. They will identify the most commonly used of the recently proposed approaches in video representation learning, which they will then examine for dimensional collapse. The results obtained will be evaluated and put into perspective.

In addition, we will support students who are recognized for the outstanding quality of their work in submitting their results to a scientific journal.

References

Becker et al. Self-organizing neural network that discovers surfaces in random-dot stereograms. Nature, vol. 355, pp. 161–163, 1992.

Brown et al. Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901, 2022.

Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv, 2018.

Feichtenhofer et al. SlowFast Networks for Video Recognition. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.

Hua et al. On Feature Decorrelation in Self-Supervised Learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9598–9608, 2021.

Mialon et al. Variance Covariance Regularization Enforces Pairwise Independence in Self-Supervised Representations. arXiv, 2022.

Zbontar et al. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. Proceedings of the 38th International Conference on Machine Learning, vol. 139, pp. 12310–12320, 2021.

Contact

Please include your transcript and, if available, a summary of your experience in your application. The thesis will be supervised by

M. Kwiatkowski
m.kwiatkowski@tu-berlin.de

P. Reiske
patrik.reiske@tu-berlin.de

both doctoral candidates working on video representation learning in Prof. Hellwich's lab for computer vision and remote sensing at the TU Berlin.