Thesis Topic: Self-supervised Video Similarity Search with Pseudo Label Generation via Agglomerative Clustering

Background

Recently, there has been significant interest in video representation learning, which is a crucial task in computer vision. Since annotating video datasets is very expensive self-supervised video representation learning has drawn a lot of attention because it doesn't require external data annotations.

Many of the state-of-the-art methods are based on contrastive instance discrimination (e.g. MoCo[1] and SimCLR [2]), which trains a network to recognize two augmented views of the same instance (called a query and a positive) while discriminating against a pool of other random instances (called negatives) from the dataset. However, this method requires a large batch of negatives during training. It has been studied that only 5% of the negatives are necessary and beneficial [3].

A pseudo label generation method has been proposed by [4] to run agglomerative clustering on the unlabeled image data and then use the cluster results to perform semi-hard negatives mining. We propose to adopt this method and extend it from image understanding to video understanding.

Objectives

The goal is to build a self-supervised video representation learning framework which learns embeddings that group semantically similar videos together while pushing dissimilar ones apart. We propose a simplistic pseudo label generation stage via clustering prior to the training. The obtained pseudo labels can guide the negatives sampling process during the training and further close the gap between self-supervised and supervised learning.

Methods

We obtain the pseudo-labels by performing agglomerative clustering on the video data using a pre-trained neural network prior to the contrastive training process. Unlike the traditional instance discrimination methodology, we sample positives from the same category of the query based on the pseudo labels, and steer away from the same-class negatives as well as the unnecessary easy negatives during the negatives sampling process.We select 3D-ResNets [5] as backbones of the encoder network, where the original 2D convolution kernels are expanded to 3D to extract spatiotemporal representations in videos.

Similar to simCLR, we adopt composition of multiple data augmentation techniques to create a richer set of data, and concatenate a learnable non-linear projection head at the end of the 3D-ResNet to reduce the output dimensionality.

We evaluate our proposed model on a downstream video retrieval task against multiple baseline models (e.g. IIC[6], CMC[7], DSM[8], etc.). The video retrieval is conducted by K nearest neighbor search with cosine similarity as distance metric, and evaluated on UCF101 and HMDB-51.

References

[1]] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2019. Momentum contrast for unsupervised visual representation learning. arXiv preprint arXiv:1911.05722 (2019).

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709 (2020).

[3] T. (T. Cai, D. J. Schwab, J. Frankle, and A. S. Morcos, "Are all negatives created equal in contrastive instance discrimination?" arXiv preprint arXiv:2010.06682 (2020).

[4] K. Koreitem1, F. Shkurti1, T. Manderson2, and W.-D. Chang2, "One-Shot Informed Robotic Visual Search in the Wild." arXiv preprint arXiv:2003.10010 (2020).

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. 770–778.

[6] Li Tao, Xueting Wang, and Tochihiko Yamasaki. Self-supervised Video Learning Using Inter-intra Contrastive Framework. arXiv preprint arXiv:2008.02531 (2020)

[7] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2019. Contrastive multiview coding. arXiv preprint arXiv:1906.05849 (2019).

[8] Jinpeng wang, Yuting Gao, Ke Li, Xinyang Jiang, Xiaowei Guo and Rongrong Ji. Enhancing Unsupervised Video Representation Learning by Decoupling the Scene and the Motion. arXiv preprint arXiv:2009.05757 (2020).

[9] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012).

[10] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. 2011. HMDB: a large video database for human motion recognition. In 2011 International Conference on Computer Vision. IEEE, 2556–2563.