Abstract.
In the image domain, excellent representations can be learned by inducing invariance to content-preserving transformations via noise contrastive learning. In this paper, we generalize contrastive learning to a wider set of transformations, and their compositions, for which either invariance or distinctiveness is sought. We show that it is not immediately obvious how existing methods such as SimCLR can be extended to do so. Instead, we introduce a number of formal requirements that all contrastive formulations must satisfy, and propose a practical construction which satisfies these requirements. In order to maximise the reach of this analysis, we express all components of noise contrastive formulations as the choice of certain generalized transformations of the data (GDTs), including data sampling. We then consider videos as an example of data in which a large variety of generalized transformations of the data (GDTs), including data sampling.

Hierarchical sampling

Video collection

Data sampling

Time shift

Modality slicing

Augmentation

Cross-modal

Joint embedding

Test different & novel learning hypotheses

On Compositions of Transformations in Contrastive Self-Supervised Learning

Mandela Patrick*, Yuki M. Asano*, Polina Kuznetsova, Ruth Fong, João F. Henriques, Geoffrey Zweig, Andrea Vedaldi

Self-supervision = learning invariance to some transformations, variance to others.

E.g.

- Jigsaw
- RotNet
- Colorization
- Video frames shuffling

References


Korbar et al. Cooperative learning of audio and video models from self-supervised synchronization. NeurIPS 2018

Miech et al. End-to-end learning of visual representations from uncurated instructional videos. CVPR 2020

Misra et al. Shuffling or shuffling? A theory of learning with imperfect inductive bias. ICLR 2016


Gidaris et al. Unsupervised representation learning by predicting image rotations. ICLR 2018


Benaim et al. Speednet: Learning the speediness in videos. CVPR 2020


