SeLaVi: Labelling unlabelled videos from scratch with multi-modal self-supervision

Yuki M. Asano*, Mandela Patrick*, Christian Rupprecht, Andrea Vedaldi

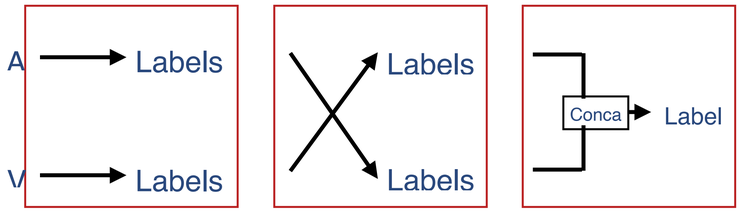
Labelling videos without any supervision.

A large part of the current success of deep learning lies in the effectiveness of data – more precisely: labelled data. Yet, labelling a dataset with human annotation continues to carry high costs, especially for videos. While in the image domain, recent methods have allowed to generate meaningful (pseudo-) labels for unlabelled datasets without supervision, this development is missing for the video domain where learning feature representations is the current focus. In this work, we a) show that unsupervised labelling of a video dataset does not come for free from strong feature encoders and b) propose a novel clustering method that allows pseudo-labelling of a video dataset without any human annotations, by leveraging the natural correspondence between the audio and visual modalities. An extensive analysis shows that the resulting clusters have high semantic overlap to ground truth human labels. We further introduce the first benchmarking results on unsupervised labelling of common video datasets Kinetics, Kinetics-Sound, VGG-Sound and AVE.

Code will be made available at https://github.com/facebookresearch/selavi

Truly multi-modal clustering.

Jointly clustering multiple modalities *properly* is not trivial.

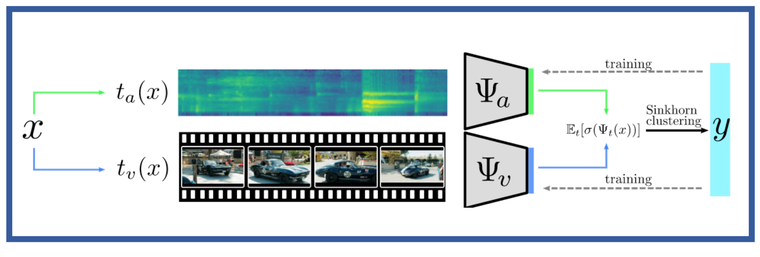


× does not use same-source information
× two different sets of clusters

× two different sets of clusters
× hard to interpret

× concatenation can just rely on stronger modality and ignore the other

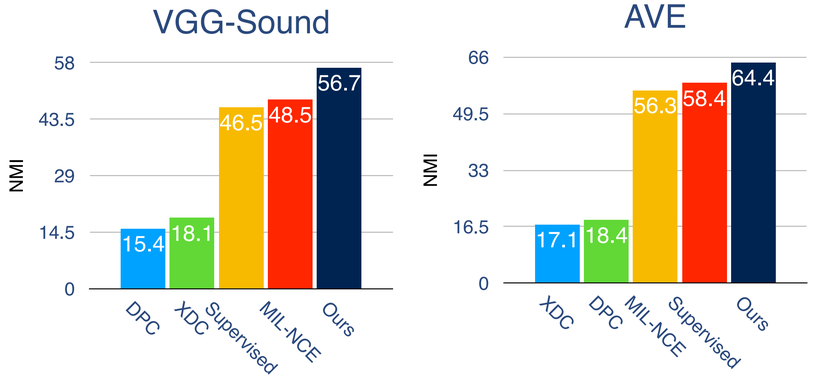
SeLaVi: Modalities as Augmentations



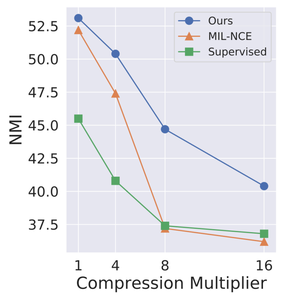
Our multi-modal clustering framework. It treats each modality as an augmentation and learns a single joint clustering with SInkhorn-Knopp optimal transport.

Good feature representations ↛ good clustering

K-Means on Kinetics-supervised and self-supervised (DPC, MIL-NCE, XDC) features are not as effective as SeLaVi, that simultaneously learns to cluster and represent multi-modal video data.



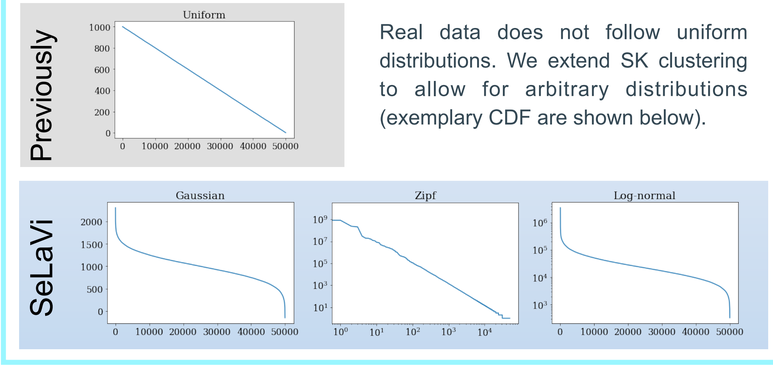
Multi-modal clustering → leverages both modalities



We evaluate clustering quality of different models when the visual modality is degraded by progressively downsampling it.

Since SeLaVi leverages both modalities equally when clustering, it is able to still cluster the data well.

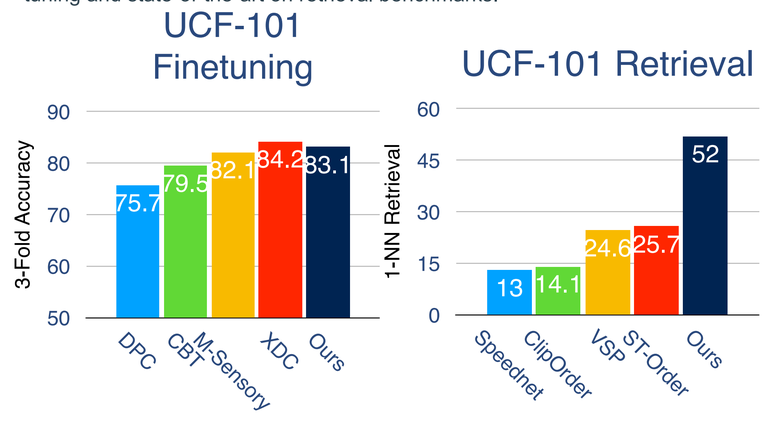
We extend SK clustering to non-uniform marginals



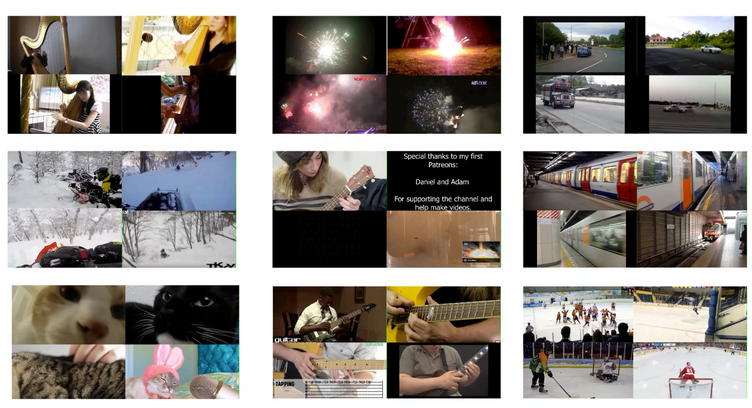
Real data does not follow uniform distributions. We extend SK clustering to allow for arbitrary distributions (exemplary CDF are shown below).

But good clustering → good representations

Visual features learned using SeLaVi are competitive on the UCF-101 fine-tuning and state-of-the-art on retrieval benchmarks.



Example clusters



Clusters discovered using SeLaVi on the VGG-Sound dataset

References

Yuki M. Asano, Christian Rupprecht and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. Proc. ICLR (2020)

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, 2018.

Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *NeurIPS*, 2018.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, 2020.

Humam Alwassel, Dhruv Mahajan, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *arXiv preprint arXiv:1911.12667*, 2019.

Interactive visualisation of all clusters: