SELAVI: SELF-LABELLING VIDEOS WITHOUT ANY ANNOTATIONS FROM SCRATCH

Yuki M. Asano[*†] **, Mandela Patrick**[*‡]**, Christian Rupprecht**[†]**, Andrea Vedaldi**[‡]
[†]Visual Geometry Group, University of Oxford [‡]Facebook AI Research

ABSTRACT

[*This work was published at NeurIPS 2020*] 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[1].

INTRODUCTION

One of the key tasks in machine learning is to convert continuous perceptual data such as images and videos into a symbolic representation, assigning discrete labels to it. This task is generally formulated as clustering (Hartigan, 1972). For images, recent contributions such as (Ji et al., 2018; Van Gansbeke et al., 2020; Caron et al., 2018; Asano et al., 2020) have obtained good results by combining clustering and representation learning. However, progress has been more limited for videos, which pose unique challenges and opportunities. Compared to images, videos are much more expensive to annotate; at the same time, they contain more information, including a temporal dimension and two modalities, aural and visual, which can be exploited for better clustering. In this paper, we are thus interested in developing methods to *cluster video datasets without manual supervision*, potentially reducing the cost and amount of manual labelling required for video data.

Just as for most tasks in machine learning, clustering can be greatly facilitated by extracting a suitable representation of the data. However, representations are usually learned by means of manually supplied labels, which we wish to avoid. Inspired by (Yan et al., 2020), we note that a solution is to consider one of the recent state-of-the-art self-supervised representation learning methods and apply an off-the-shelf clustering algorithm post-hoc. With this, we show that we can obtain very strong baselines for clustering videos.

Still, this begs the question of whether even better performance could be obtained by simultaneously learning to cluster and represent video data. Our main contribution is to answer this question affirmatively and thus to show that *good clusters do not come for free from good representations*.

In order to do so, we consider the recent method SeLa (Asano et al., 2020), which learns clusters and representations for still images by solving an optimal transport problem, and substantially improve it to work with multi-modal data. We do this in three ways. First, we relax the assumption made in (Asano et al., 2020) that clusters are equally probable; this is not the case for semantic video labels, which tend to have a highly-skewed distribution (Gu et al., 2018; Kay et al., 2017; Abu-El-Haija et al., 2016), and extend the algorithm accordingly. Second, we account for the multi-modal nature of video data, by formulating the extraction of audio and visual information from a

[*]Joint first authors.
[1]Code is available at `https://github.com/facebookresearch/selavi`

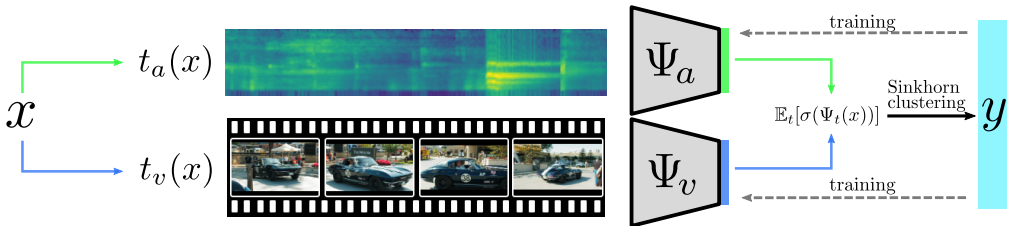

Figure 1: **Our model** views modalities as different *augmentations* and produces a multi-modal clustering of video datasets from scratch that can closely match human annotated labels.

video as a form of data augmentation, thus learning a clustering function which is invariant to such augmentations. For this to work well, we also propose a new initialization scheme that synchronizes the different modalities before clustering begins. This encourages clusters to be more abstract and thus 'semantic' and learns a redundant clustering function which can be computed robustly from either modality (this is useful when a modality is unreliable, because of noise or compression). Third, since clustering is inherently ambiguous, we propose to learn multiple clustering functions in parallel, while keeping them orthogonal, in order to cover a wider space of valid solutions.

With these technical improvements, our method for Self-Labelling Videos (SeLaVi) substantially outperforms the post-hoc approach (Yan et al., 2020), SeLa (Asano et al., 2020) applied to video frames, as well as a recent multi-modal clustering-based representation learning method, XDC (Alwassel et al., 2019). We evaluate our method by testing how well the automatically learned clusters match manually annotated labels in four different video datasets: VGG-Sound (Chen et al., 2020), AVE (Tian et al., 2018), Kinetics (Kay et al., 2017) and Kinetics-Sound (Arandjelovic & Zisserman, 2017). We show that our proposed model results in substantially better clustering performance than alternatives. For example, our method can perfectly group 32% of the videos in the VGG-Sound dataset and 55% in the AVE dataset without using any labels during training. Furthermore, we show that, while some clusters do not align with the ground truth classes, they are generally semantically meaningful (e.g. they contain similar background music) and provide an interactive cluster visualization[2].

In a nutshell, our key contributions are: **(i)** establishing video clustering benchmark results on four datasets for which labels need to be obtained in an unsupervised manner; **(ii)** developing and assessing several strong clustering baselines using state-of-the-art methods for video representation learning, and **(iii)** developing a new algorithm tailored to clustering multi-modal data resulting in state-of-the-art highly semantic labels.

METHODOLOGY

Given a dataset $D = \{\boldsymbol{x}_i\}_{i\in\{1,\dots,N\}}$ of multi-modal data $\boldsymbol{x}_i$, our goal is to learn a labelling function $y(\boldsymbol{x}) \in \{1,\dots,K\}$ without access to any ground-truth label annotations. There are two requirements that the labelling function must satisfy. First, the labels should capture, as well as possible, the semantic content of the data, in the sense of reproducing the labels that a human annotator would intuitively associate to the videos. As part of this, we wish to account for the fact that semantic classes are not all equally probable, and tend instead to follow a Zipf distribution (Abu-El-Haija et al., 2016; Kay et al., 2017). We then evaluate the quality of the discovered labels by matching them to the ones provided by human annotators, using datasets where ground-truth labels are known.

The second requirement is that the labelling method should not overly rely on a single modality. Instead, we wish to treat each modality *as equally informative* for clustering. In this way, we can learn a more robust clustering function, which can work from either modality. Furthermore, correlating of modalities has been shown to be a proxy to learn better abstractions Arandjelović & Zisserman (2018); Korbar et al. (2018); Patrick et al. (2020); Owens & Efros (2018).

While our method can work with any number of data modalities (vision, audio, depth, textual transcripts, ...), we illustrate it under the assumption of video data $\boldsymbol{x} = (a, v)$, comprising an audio

[2] https://www.robots.ox.ac.uk/~vgg/research/selavi

stream a and a visual stream v. The following two sections describe our method in detail and show how it meets our requirements.

Non-degenerate clustering via optimal transport Borrowing the notation from Asano et al. (2020), we recall the cross-entropy loss $E(q, p)$, between the labels given as one-hot vectors in q (i.e. $q(y(\boldsymbol{x})) = 1\ \forall \boldsymbol{x}$) and the softmax outputs p of a network Ψ:

$$E(p, q) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{y=1}^{K} q(y|\boldsymbol{x}_i) \log p(y|\boldsymbol{x}_i), \quad p(y|\boldsymbol{x}_i) = \operatorname{softmax} \Psi(\boldsymbol{x}_i), \tag{1}$$

where K is the number of clusters. This energy is optimized under the constraint that the marginal cluster probability $\sum_{i=1}^{N} \frac{1}{N} p(y|\boldsymbol{x}_i) = \frac{1}{K}$ is constant (meaning all clusters are a-priori equally likely). This then is a linear optimal transport problem, for which (Cuturi, 2013) provides a fast, matrix-vector multiplication based solution.

Clustering with arbitrary prior distributions A shortcoming of the algorithm just described is the assumption that all clusters are equally probable. This avoids converging to degenerate cases but is too constraining in practice since real datasets follow highly skewed distributions (Abu-El-Haija et al., 2016; Kay et al., 2017), and even in datasets that are collected to be uniform, they are not completely so (Chen et al., 2020; Kay et al., 2017; Tian et al., 2018). Furthermore, knowledge of the data distribution, for example long-tailedness, can be used as additional information (e.g. as in (Piergiovanni et al., 2020) for meta-learning) that can improve the clustering by allocating the right number of data points to each cluster. Next, we describe a mechanism to change this distribution arbitrarily.

In the algorithm above, changing the label prior amounts to choosing a different cluster marginal r in the polytope $U(r, c)$. The difficulty is that r is only known up to an arbitrary permutation of the clusters, as we do not know a-priori which clusters are more frequent and which ones less so. To understand how this issue can be addressed, we need to explicitly write out the energy optimised by the Sinkhorn-Knopp (SK) algorithm (Cuturi, 2013) to solve the optimal transport problem. This energy is:

$$\min_{Q \in U(r,c)} \langle Q, -\log P \rangle + \frac{1}{\lambda} \operatorname{KL}(Q \| rc^\top), \tag{2}$$

where λ is a fixed parameter. Let $r' = Rr$ where R is a permutation matrix matching clusters to marginals. We then seek to optimize the same quantity w.r.t. R, obtaining the optimal permutation as $R^* = \operatorname{argmin}_R E(R)$ where

$$E(R) = \langle Q, -\log P \rangle + \frac{1}{\lambda} \operatorname{KL}(Q \| Rrc^\top) = \text{const} + \sum_y -q(y)\ [R \log r]_y. \tag{3}$$

While there is a combinatorial number of permutation matrices, we show that minimizing Eq. (3) can be done by first sorting classes y in order of increasing $q(y)$, so that $y > y' \Rightarrow q(y) > q(y')$, and then finding the permutation that R that also sorts $[R \log r]_y$ in increasing order.[3] We conclude that R cannot be optimal unless it sorts all pairs. After this step, the SK algorithm can be applied using the optimal permutation R^*, without any significant cost (as solving for R is equivalent to sorting $\mathcal{O}(K \log K)$ with $K \ll N$). The advantage is that it allows to choose any marginal distribution, even highly unbalanced ones which are likely to be a better match for real world image and video classes than a uniform distribution.

Multi-modal single labelling Next, we tackle our second requirement of extracting as much information as possible from multi-modal data. In principle, all we require to use the clustering

[3]To see why this is optimal, and ignoring ties for simplicity, let R be any permutation and construct a permutation $\bar{R}$ by applying R and then by further swapping two labels $y > y'$. We can relate the energy of R and $\bar{R}$ as:

$$\begin{aligned} E(R) &= E(\bar{R}) + q(y)[\bar{R} \log r]_y + q(y')[\bar{R} \log r]_{y'} - q(y)[\bar{R} \log r]_{y'} - q(y')[\bar{R} \log r]_y \\ &= E(\bar{R}) + (q(y) - q(y'))\ ([\bar{R} \log r]_y - [\bar{R} \log r]_{y'}). \end{aligned} \tag{4}$$

Since the first factor is positive by assumption, this equation shows that the modified permutation $\bar{R}$ has a lower energy than R if, and only if, $[\bar{R} \log r]_y > [\bar{R} \log r]_{y'}$, which means that $\bar{R}$ sorts the pair in increasing order.

Table 1: **Unsupervised labelling of datasets.** We compare labels from our method to labels that are obtained with k-means on the representations from a supervised and various unsupervised methods on two datasets.

(a) **VGG-Sound.**

Method	**NMI**	**ARI**	**Acc.**	$\langle \mathbf{H} \rangle$	$\langle \mathbf{p}_{\max} \rangle$
Random	10.2	4.0	2.2	4.9	3.5
Supervised	46.5	15.6	24.3	2.9	30.8
XDC	18.1	1.2	4.5	4.41	7.4
MIL-NCE	48.5	12.5	22.0	2.6	32.9
SeLaVi	**55.9**	**21.6**	**31.0**	**2.5**	**36.3**

(b) **AVE.**

Method	**NMI**	**ARI**	**Acc.**	$\langle \mathbf{H} \rangle$	$\langle \mathbf{p}_{\max} \rangle$
Random	9.2	1.3	9.3	2.9	12.6
Supervised	58.4	34.8	50.5	1.1	60.6
XDC	17.1	6.0	16.4	2.6	19.1
MIL-NCE	56.3	30.3	42.6	**1.2**	57.1
SeLaVi	**66.2**	**47.4**	**57.9**	**1.1**	**59.3**

formulation Eq. (1) with multi-modal data $\boldsymbol{x} = (a, v)$ is to design a corresponding multi-modal representation $\Psi(\boldsymbol{x}) = \Psi(a, v)$. However, we argue for *multi-modal single labelling* instead. By this, we mean that we wish to cluster data one modality at a time, but in a way that is modality agnostic. Formally, we introduce modality splicing transformations (Patrick et al., 2020) $t_a(\boldsymbol{x}) = a$ and $t_v(\boldsymbol{x}) = v$ and use these as data augmentations. Recall that augmentations are random transformations t such as rotating an image or distorting an audio track that one believes should leave the label/cluster invariant. We thus require our activations used for clustering to be an average over augmentations by replacing matrix $\log P$ with

$$[\log P]_{yi} = \mathbb{E}_t[\log \operatorname{softmax}_y \Psi(t\boldsymbol{x}_i)]. \tag{5}$$

If we consider splicing as part of the augmentations, we can learn clusters that are invariant to standard augmentations as well as the choice of modality. In practice, to account for modality splicing, we define and learn a pair $\Psi = (\Psi_a, \Psi_v)$ of representations, one per modality, resulting in the same clusters ($\Psi_a(t_a(\boldsymbol{x})) \approx \Psi_v(t_v(\boldsymbol{x}))$). This is illustrated in Figure 1.

Decorrelated clustering heads. Conceptually, there is no single 'correct' way of clustering a dataset: for example, we may cluster videos of animals by their species, or whether they are taken indoor or outdoor. In order to alleviate this potential issue, inspired by (Asano et al., 2020; Ji et al., 2018), we simply learn multiple labelling functions y, using multiple classification heads for the network. We improve this scheme as follows. In each round of clustering, we generate two random augmentations of the data. Then, the applications of SK to half of the heads (at random) see the first version, and the other half the second version, thus increasing the variance of the resulting clusters. This increases the cost of the algorithm by only a small amount — as more time is used for training instead of clustering. As we show in the main paper, this substantially increases clustering performance, so we apply this per default to our runs.

Results

Table 1 shows the quality of the labels obtained automatically by our algorithm. We find that for the datasets VGG-Sound and AVE, our method achieves state-of-the-art clustering performance with high accuracies of 55.9% and 57.9%, even surpassing the one of the strongest video feature encoder at present, the manually-supervised R(2+1)D-18 network. This result echoes the findings in the image domain (Van Gansbeke et al., 2020) where plain k-means on representations is found to be less effective compared to learning clusters. For ablations and results on Kinetics-400 and Kinetics-Sound, we refer to the main paper.

Conclusion

In this work, we have shown how self-supervised clustering with optimal transport can be used to obtain strong semantic labels for video datasets, without any supervision. In this extended abstract, we show strong performance when comapred to the manual annotations for both the VGG-Sound and the AVE dataset using various metrics.

REFERENCES

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. YouTube-8M: A large-scale video classification benchmark. arXiv preprint arXiv:1609.08675, 2016.

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.

Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In Proc. ICCV, 2017.

Relja Arandjelović and Andrew Zisserman. Objects that sound. In ECCV, 2018.

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

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

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. ICASSP), May 2020.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In NeurIPS, pp. 2292–2300, 2013.

Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In CVPR, pp. 6047–6056, 2018.

John A Hartigan. Direct clustering of a data matrix. Journal of the american statistical association, 67(337): 123–129, 1972.

Xu Ji, João F. Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation, 2018.

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. CoRR, abs/1705.06950, 2017.

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

Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In ECCV, 2018.

Mandela Patrick, Yuki M. Asano, Ruth Fong, João F. Henriques, Geoffrey Zweig, and Andrea Vedaldi. Multi-modal self-supervision from generalized data transformations, 2020.

AJ Piergiovanni, Anelia Angelova, and Michael S. Ryoo. Evolving losses for unsupervised video representation learning, 2020.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In ECCV, 2018.

Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In European Conference on Computer Vision (ECCV), 2020.

Xueting Yan, Ishan Misra, Abhinav Gupta, Deepti Ghadiyaram, and Dhruv Mahajan. ClusterFit: Improving Generalization of Visual Representations. In CVPR, 2020.