Self-Supervised Learning by Cross-Modal Audio-Video Clustering

Jonas Grimm

October 5, 2020

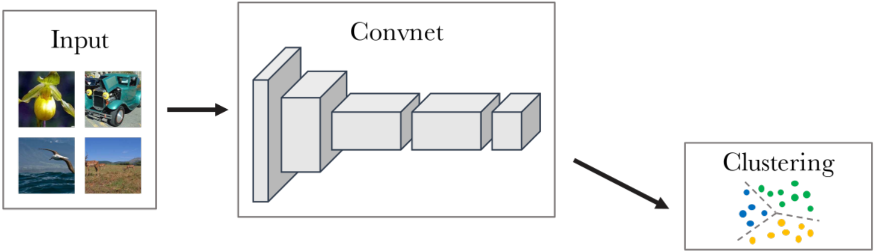
Introduction

Challenges in supervised video model learning:

- High cost of scaling up the size of manually-labeled video data sets
- Unclear definition of suitable label spaces for action recognition

Aim: Pretrain spatiotemporal models for action recognition on unlabeled data

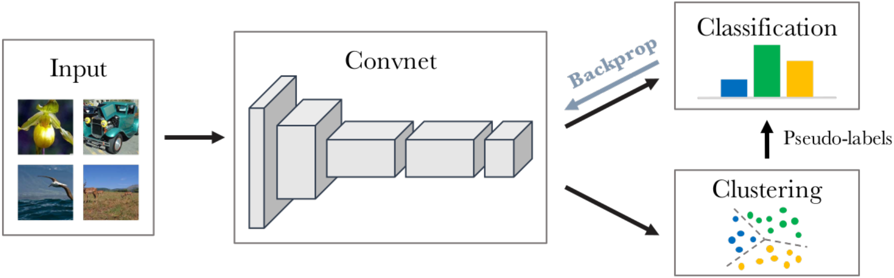
Single-Modality Deep Clustering



[Caron et al., 2018]

- First step: Cluster deep features from an encoder

Single-Modality Deep Clustering



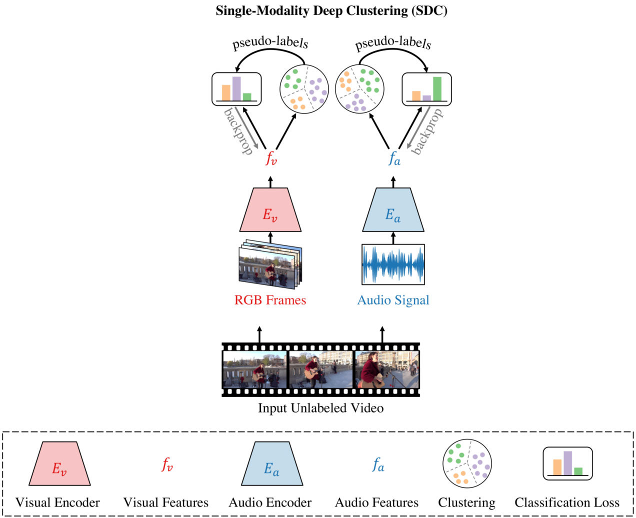
[Caron et al., 2018]

- First step: Cluster deep features from an encoder
- Second step: Update encoder using cluster assignments as labels

Multi-Modality Deep Clustering

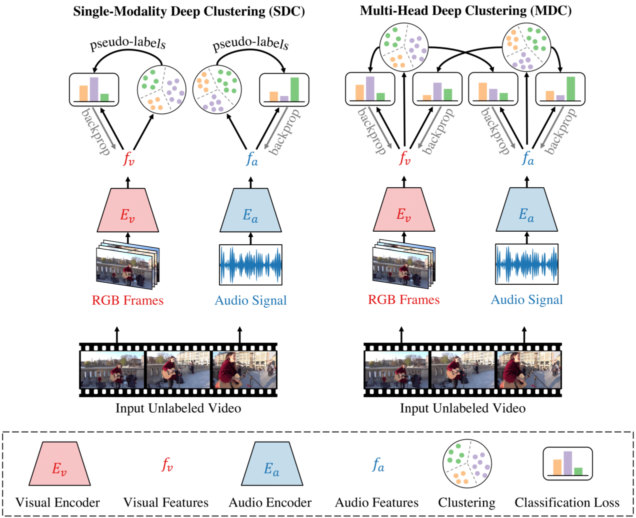
- Visual and audio modalities are highly correlated yet they contain different information
- Correlations allow predictions from one input space to the other
- Intrinsic differences make cross-model prediction an enriching self-supervised task

Single-Modality Deep Clustering on Videos



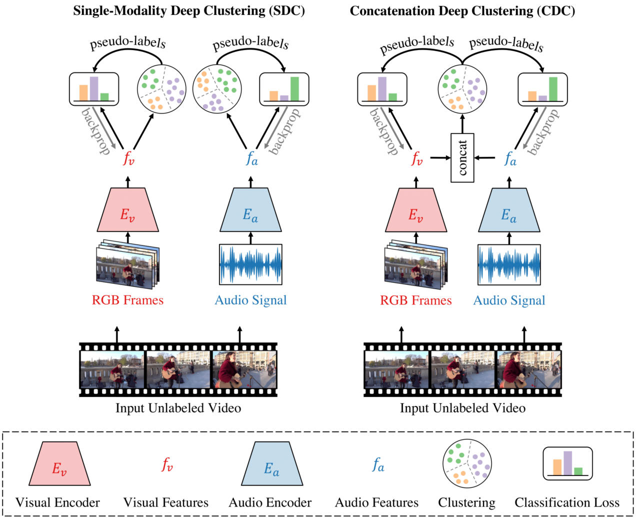
[Alwassel et al., 2019]

Multi-Modality Deep Clustering



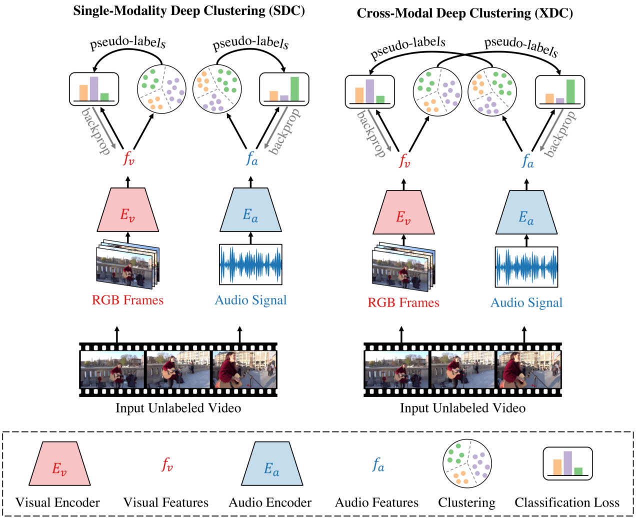
[Alwassel et al., 2019]

Multi-Modality Deep Clustering



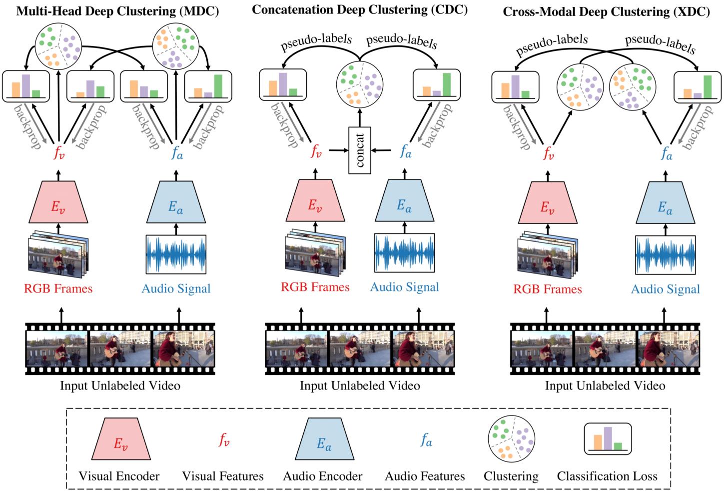
[Alwassel et al., 2019]

Multi-Modality Deep Clustering



[Alwassel et al., 2019]

Multi-Modality Deep Clustering

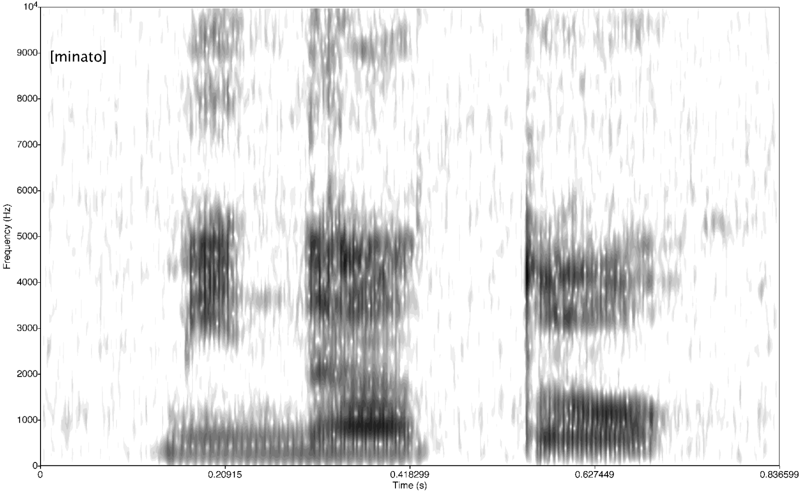


[Alwassel et al., 2019]

Experimental Setup

- Four pretraining datasets: Kinetics (action recognition), AudioSet (audio classification), IG-Kinetics (videos from social media), IG-Random (videos from social media)
- Three downstream datasets: UCF101 (action recognition), HMBD51 (action recognition), ESC50 (sound classification)
- Two baselines: training model from scratch on downstream task, supervised pretraining on large labeled dataset
- Encoders: R(2+1)D network as visual encoder, ResNet as audio encoder (spectrogram image as input)

Spectrogram



https://de.wikipedia.org/wiki/Spektrogramm#/media/Datei:Spectrogram_-minato-.png

XDC performes best

Pretraining data set: Kinetics

Method	UCF101	HMDB51	ESC50
Scratch	54.5	24.1	54.3
Superv	90.9	58.0	82.3
SDC	61.8	31.4	66.5
MDC	68.4	37.1	70.3
CDC	<u>72.9</u>	<u>37.5</u>	<u>74.8</u>
XDC	**74.2**	**39.0**	**78.0**

- Exploiting multi-modalities increases performance compared to single-modality clustering
- Self-supervision purely by the signal from the other modality yields strongest results

Analyzing number of *k*-means clusters

Pretraining Dataset	Downstream Dataset	k				
		64	128	256	512	1024
Kinetics (240K videos)	UCF101	73.8	73.1	**74.2**	<u>74.0</u>	72.6
	HMDB51	36.5	**39.0**	<u>38.3</u>	37.7	37.7
	ESC50	**78.0**	<u>76.3</u>	75.0	74.5	71.5
AudioSet-240K (240K videos)	UCF101	**77.4**	<u>77.2</u>	76.7	77.1	75.3
	HMDB51	41.3	**42.6**	<u>41.6</u>	40.6	40.7
	ESC50	**78.5**	<u>77.8</u>	77.3	76.8	73.5
AudioSet (2M videos)	UCF101	84.1	84.3	**84.9**	<u>84.4</u>	84.2
	HMDB51	47.4	47.6	**48.8**	<u>48.5</u>	48.4
	ESC50	84.8	**85.8**	<u>85.0</u>	84.5	83.0

- Best value for *k* not sensitive to number of semantic labels in downstream data set
- Best value for *k* increases with increasing pretraining data set size

Analyzing pretraining data type and size

Method	Pretraining Dataset	Pretraining Size	UCF101	HMDB51	ESC50
Scratch	None	0	54.5	24.1	54.3
Superv	ImageNet	1.2M	79.9	44.5	NA
Superv	Kinetics	240K	90.9	58.0	82.3
Superv	AudioSet-240K	240K	76.6	40.8	78.3
Superv	AudioSet	2M	84.0	53.5	**90.3**
XDC	Kinetics	240K	74.2	39.0	78.0
XDC	AudioSet-240K	240K	77.4	42.6	78.5
XDC	AudioSet	2M	84.9	48.8	85.8
XDC	IG-Random	65M	88.8	61.2	86.3
XDC	IG-Kinetics	65M	**91.5**	**63.1**	84.8

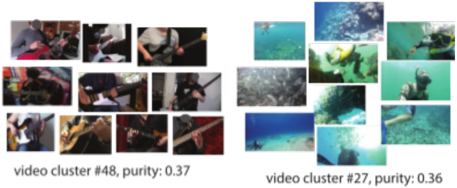
- XDC outperforms supervised pretraining when trained on large data set
- Supervised pretraining is influenced by taxonomy more than by size
- XDC is less sensitive to the data type

Comparing full finetuning vs learning linear classifier

Method	Pretraining Dataset	UCF101 fc	UCF101 all	HMBD51 fc	HMBD51 all	ESC50 fc	ESC50 all
Random	None	6.0	54.5	7.5	24.1	61.3	54.3
Superv	ImageNet	74.5	79.9	42.8	44.5	NA	NA
Superv	Kinetics	**89.7**	90.9	**61.5**	58.0	79.5	82.3
Superv	AudioSet	80.2	84.0	51.6	53.5	**88.5**	**90.3**
XDC	IG-Random	80.7	88.8	49.9	61.2	84.5	86.3
XDC	IG-Kinetics	85.3	**91.5**	56.0	**63.1**	84.3	84.8

- Performance of most pretrained models decreases if used as fixed feature extractor compared to fully finetuning on downstream data set
- Relative performance of XDC stays generally the same, making XDC useful both as a fixed feature extractor and as pretraining initialization
- Supervised pretraining followed by fc-only finetuning performs well when pretraining and downstream task are very similar
- XDC is taxonomy-independent

XDC video clusters



#	Kinetics concepts
1	playing bass guitar (0.37), playing guitar (0.16), tapping guitar (0.15)
4	swim backstroke (0.21), swim breast s. (0.16), swim butterfly s. (0.10)
5	golf putting (0.18), golf chipping (0.10), golf driving (0.05)
9	windsurfing (0.12), jetskiing (0.10), water skiing (0.09)
10	cooking chicken (0.11), barbequing (0.07), frying vegetables (0.06)
63	pull ups (0.01), gymnastics tumbling (0.01), punching bag (0.01)
74	capoeira (0.01), riding elephant (0.01), feeding goats (0.01)

[Alwassel et al., 2019]

State-of-the-Art Self-Supervised Learning Comparison

	Pretraining		Evaluation	
Method	Architecture	Dataset	UCF101	HMDB51
ClipOrder	R(2+1)D-18	UCF101	72.4	30.9
MotionPred	C3D	Kinetics	61.2	33.4
RotNet3D	3D-ResNet18	Kinetics	62.9	33.7
ST-Puzzle	3D-ResNet18	Kinetics	65.8	33.7
DPC	3D-ResNet34	Kinetics	75.7	35.7
AVTS	MC3-18	Kinetics	84.1	52.5
AVTS	R(2+1)D-18	Kinetics	86.2	52.3
XDC	R(2+1)D-18	Kinetics	86.8	52.6
AVTS	MC3-18	AudioSet	87.7	57.3
AVTS	R(2+1)D-18	AudioSet	86.8	52.6
XDC	R(2+1)D-18	AudioSet	93.0	63.7
XDC	R(2+1)D-18	IG-Random	94.6	66.5
XDC	R(2+1)D-18	IG-Kinetics	**95.5**	**68.9**
Fully supervised	R(2+1)D-18	ImageNet	82.8	46.7
Fully supervised	R(2+1)D-18	Kinetics	93.1	63.6

State-of-the-Art Self-Supervised Learning Comparison

Method	ESC50
Piczak ConvNet	64.5
SoundNet	74.2
L3-Net	79.3
AVTS	82.3
ConvRBM	**86.5**
XDC (AudioSet)	84.8
XDC (IG-Random)	85.4

Method	DCASE
RNH	77
Ensemble	78
SoundNet	88
L3-Net	93
AVTS	94
XDC (AudioSet)	**95**
XDC (IG-Random)	**95**

Conclusion

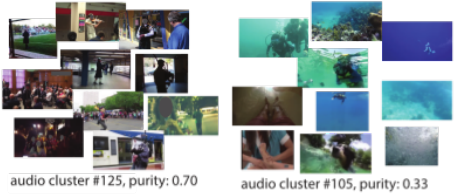
- Deep Clustering is a promising self-supervised method
- Exploiting multi-modalities enriches the self-supervised task
- Pure supervision by different modality yields strongest results
- XDC model even outperformed large-scale fully supervised pretraining

Literature

Alwassel, H., Mahajan, D., Torresani, L., Ghanem, B., and Tran, D. (2019).
Self-supervised learning by cross-modal audio-video clustering.
arXiv preprint arXiv:1911.12667.

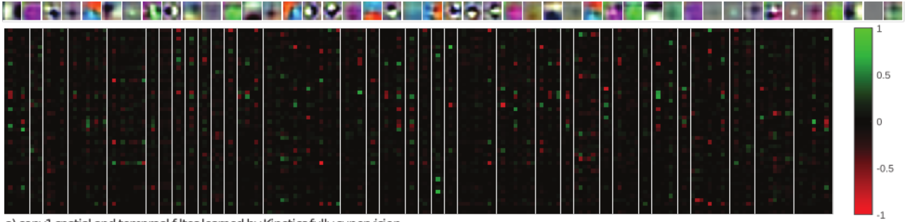
Caron, M., Bojanowski, P., Joulin, A., and Douze, M. (2018).
Deep clustering for unsupervised learning of visual features.
In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149.

XDC audio clusters

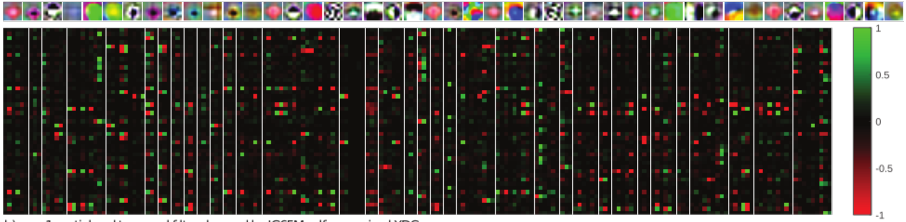


#	Kinetics concepts
1	play bagpipes (0.70), play harmonica (0.04), play violin (0.03)
2	scuba diving (0.33), snorkeling (0.27), feeding fish (0.11)
4	pass football (0.17), play kickball (0.06), catch/throw softball (0.05)
8	play cello (0.15), play trombone (0.11), play accordion (0.09)
10	moving lawn (0.14), driving tractor (0.09), motorcycling (0.06)
127	abseiling (0.01), grooming horse (0.01), milking cow (0.01)
128	washing feet (0.01), motorcycling (0.01), headbanging (0.01)

[Alwassel et al., 2019]



a) conv1 spatial and temproal f ltes learned by Kinetics fully supervision.



b) conv1 spatial and temporal f lters learned by IG65M self-supervised XDC.

[Alwassel et al., 2019]

Avoiding trivial solution

Any method that jointly learns a discriminative classifier and labels is prone to trivial solutions:

- Empty clusters: All inputs assigned to single cluster. Solution: Reassign empty clusters
- Trivial parametrization: Different cluster sizes lead to a imbalanced class distribution. Solution: Sample images on uniform distribution over classes

Despite those challenges, deep clustering achieved impressive results and outperformed previous state-of-the-art methods