Audio-Visual Instance Discrimination

Pedro Morgado[1,2⋆] Nuno Vasconcelos[1] Ishan Misra[2]

[1]University of California, San Diego [2]Facebook AI Research

Abstract. We present a self-supervised approach to learn audio-visual representations from video. Our method uses contrastive learning for cross-modal discrimination of video from audio and vice versa. We show that optimizing for cross-modal discrimination, rather than within-modal discrimination, is important to learn good representations from video and audio. With this simple but powerful insight, our method achieves state-of-the-art results when finetuned on action recognition tasks.

1 Introduction

In this work, we leverage freely occurring audio to learn video representations in a self-supervised manner. A common technique [2, 10, 13, 14] is to setup a verification task that requires predicting whether an input pair of video and audio is 'correct' or not. However, these tasks use a *single* pair at a time and miss a key opportunity to reason about the data distribution at large. We propose a contrastive learning framework to learn cross-modal representations in a self-supervised manner by contrasting video representations against *multiple* audios at once (and vice versa). We leverage recent advances [8, 12, 21, 24] in contrastive learning to setup a Audio-Visual Instance Discrimination (AVID) task that learns a cross-modal similarity metric by grouping video and audio *instances* that co-occur. We show that the cross-modal discrimination task, *i.e.*, predicting which audio matches a video, is more powerful that the within-modal discrimination task, predicting which video clips are from the same video. Our technique improves upon the state-of-the-art self-supervised methods on action recognition benchmarks like UCF-101 and HMDB-51.

2 Audio-Visual Instance Discrimination (AVID)

Goal and Intuition. Consider a dataset of N samples (instances) $\mathcal{S} = \{s_i\}_{i=1}^N$ where each instance s_i is a video s_i^v with a corresponding audio s_i^a. AVID learns visual and audio representations $(\mathbf{v}_i, \mathbf{a}_i)$ from the training instances s_i where the representations are optimized for 'instance discrimination' [5, 24], *i.e.*, must be discriminative of s_i itself as opposed to other instances s_j.

To accomplish this, two neural networks extract unit norm feature vectors $\mathbf{v}_i = f_v(s_i^v)$ and $\mathbf{a}_i = f_a(s_i^a)$ from the video and audio independently. Slow

⋆ Work done during internship at Facebook AI Research.

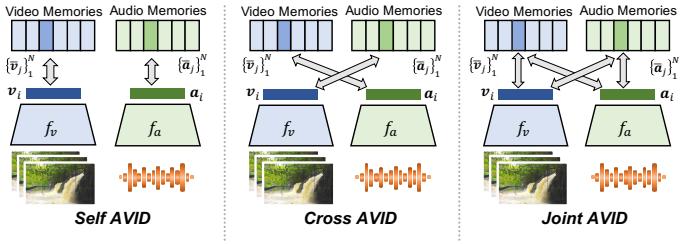

Fig. 1: Instance discrimination can be accomplished contrasting representations within the same modality (Self-AVID), across modalities (Cross-AVID) or a mixture of the two (Joint-AVID).

moving (exponential moving average) representations for both video and audio features $\{(\bar{\mathbf{v}}_i, \bar{\mathbf{a}}_i)\}_{i=1}^N$ are maintained as 'memory features' and used as targets for contrastive learning. The AVID task learns representations $(\mathbf{v}_i, \mathbf{a}_i)$ that are more similar to the memory features of the instance $(\bar{\mathbf{v}}_i, \bar{\mathbf{a}}_i)$ as opposed to memory features of other instances $(\bar{\mathbf{v}}_j, \bar{\mathbf{a}}_j)$, $j \neq i$.

Unlike previous single modality approaches [5, 24], AVID uses multiple modalities (similar to [21]), and assumes multiple forms as shown in Fig 1.

1. **Self-AVID** requires instance discrimination within the same modality.
2. **Cross-AVID** optimizes for cross-modal discrimination.
3. **Joint-AVID** combines the Self-AVID and Cross-AVID objectives.

Loss function. We use noise contrastive estimation (NCE) [7], where representations of instances s_i are contrasted to samples in a randomly sampled negative set $\mathcal{N}_i$. We build upon the implementation of [24] and refer the reader to their paper for details. The three variants of AVID depicted in Fig 1 are trained to optimize variations of the NCE loss by varying the target representations.

$$\mathcal{L}_{\text{Self-AVID}}(\mathbf{v}_i, \mathbf{a}_i) = \mathcal{L}_{\text{NCE}}(\mathbf{v}_i; \bar{\mathbf{v}}_i, \mathcal{N}_i) + \mathcal{L}_{\text{NCE}}(\mathbf{a}_i; \bar{\mathbf{a}}_i, \mathcal{N}_i) \quad (1)$$

$$\mathcal{L}_{\text{Cross-AVID}}(\mathbf{v}_i, \mathbf{a}_i) = \mathcal{L}_{\text{NCE}}(\mathbf{v}_i; \bar{\mathbf{a}}_i, \mathcal{N}_i) + \mathcal{L}_{\text{NCE}}(\mathbf{a}_i; \bar{\mathbf{v}}_i, \mathcal{N}_i) \quad (2)$$

$$\mathcal{L}_{\text{Joint-AVID}}(\mathbf{v}_i, \mathbf{a}_i) = \mathcal{L}_{\text{Self-AVID}}(\mathbf{v}_i, \mathbf{a}_i) + \mathcal{L}_{\text{Cross-AVID}}(\mathbf{v}_i, \mathbf{a}_i) \quad (3)$$

3 Experiments

It is not immediately obvious what are the the relative advantages of the AVID variants described above. We now analyze them and show that, surprisingly, the seemingly minor differences between them translate to significant differences in performance. Models are trained using a random subset of Audioset dataset [6] containing 100K videos. The video model is a smaller version of the R(2+1)D models proposed in [22] with 9 layers. The audio network is a 9 layer 2D ConvNet with batch normalization. In both cases, output activations are max-pooled, projected into a 128-dimensional feature using a MLP [4] and normalized. We will provide full training details and release the code and models. We evaluate learned features by training linear classifiers on fixed features. Visual features are

Method	block1	block2	block3	block4	Best
Cross-AVID	**19.80**	**26.98**	**34.81**	**39.95**	**39.95**
Self-AVID	17.10	22.28	27.23	32.08	32.08
Joint-AVID	18.65	23.60	29.47	33.04	33.04

(a) Top-1 Accuracy of linear probing on Kinetics.

	block1	block2	block3	block4	Best
Cross-AVID	**67.25**	**73.15**	**74.80**	**75.05**	**75.05**
Self-AVID	66.92	72.64	71.45	71.61	72.64
Joint-AVID	65.45	68.65	71.77	68.41	71.77

(b) Top-1 Accuracy of linear probing on ESC.

Table 1: Transfer performance of representations learned by AVID variants.

Method	Backbone	Input	UCF	HMDB
		Pre-training DB: Kinetics		
DPC [9]	3D ResNet-34	25×128^2	75.7	35.7
CBT [20]	S3D Inception	16×112^2	79.5	44.6
L3* [2]	R(2+1)D-18	16×224^2	74.4	47.8
AVTS [10]	MC3-VGGish-9	25×224^2	85.8	56.9
XDC [1]	R(2+1)D-18	8×224^2	74.2	39.0
		32×224^2	84.2	47.1
Cross-AVID	R(2+1)D-18	8×224^2	82.3	49.1
		32×224^2	**86.9**	**59.9**
		Pre-training DB: Audioset		
L3* [2]	R(2+1)D-18	16×224^2	82.3	51.6
Multisensory [13]	3D-Resnet-18	64×224^2	82.1	–
AVTS [10]	MC3-VGGish-9	25×224^2	89.0	61.6
XDC [1]	R(2+1)D-18	8×224^2	84.9	48.8
		32×224^2	**91.2**	61.0
Cross-AVID	R(2+1)D-18	8×224^2	88.3	57.5
		32×224^2	91.0	**64.1**

(a) Action recognition

Method	ESC	DCASE
	Pre-training DB: None	
RandomForest [16]	44.3	–
ConvNet [15]	64.5	–
ConvRBM [17]	86.5	–
	Pre-training DB: Flickr-SoundNet	
SoundNet [3]	74.2	88
L3 [2]	79.3	93
	Pre-training DB: Kinetics	
AVTS [10]	76.7	91
XDC [1]	**78.5**	–
Cross-AVID	77.6	**93**
	Pre-training DB: Audioset	
AVTS [10]	80.6	93
XDC [1]	85.8	–
Cross-AVID	**89.2**	**96**

(b) Sound classification

Table 2: Top-1 accuracy on UCF, HMDB, ESC and DCASE validation data. Methods are organized by pre-training dataset. Our AVID model achieves state-of-the-art performance in most cases.

evaluated on the Kinetics dataset [23] for action recognition, and audio features on the ESC-50 [16] dataset.

Cross-modal *vs*. within-modal instance discrimination: The performance of the three AVID variants are shown in Tab 1. We observe that Self-AVID is consistently outperformed by Cross-AVID on both visual and audio tasks. Self-AVID uses within-modality instance discrimination which is an "easier" (self-referential) pretext task and can be partially solved by matching low-level statistics. This hypothesis is supported by the fact that Joint-AVID, which combines the objectives of both Cross-AVID and Self-AVID, also performs worse than Cross-AVID. Cross-AVID uses a "harder" cross-modal instance discrimination task where the video features are required to match to the corresponding audio and vice-versa. As a result, it generalizes better to downstream tasks.

Comparison to prior work We now compare Cross-AVID to recent self-supervised methods. We use the 18-layer R(2+1)D network of [22] as the video encoder and a 9-layer (2D) CNN with batch normalization as the audio encoder. Models are trained either on Kinetics-400 [23] or the full Audioset [6] datasets.

Following prior work [9, 10, 21], we evaluate visual representations on the UCF-101 [18] and HMDB-51 [11] datasets by full network fine-tuning using clips with both 8 and 32 frames. At inference time, predictions are computed by averaging 10 sub-clips [10]. Tab 2a shows that Cross-AVID achieves state-of-the-art performance for equivalent data settings in most cases. When pre-trained on

Audioset, Cross-AVID outperformed other audio-visual SSL methods such as L3 and AVTS by at least 2.0% on UCF and 2.5% on HMDB. Similar to Cross-AVID, L3 and AVTS seek to predict whether audio/video pairs are in-sync. However, these methods optimize for the binary audio visual correspondence task, which fails to reason about the data distribution at large. The concurrently proposed XDC relies on clusters in the visual and audio spaces to provide cross-modal supervision. The instance discrimination approach of Cross-AVID outperforms XDC in most settings.

Audio representations are evaluated on the ESC-50 [16] and DCASE [19] datasets by linear probing using a linear one-vs-all SVM classifier (as in [10]). At test time, sample level predictions are obtained by averaging 10 clip level predictions. Tab 2b shows that Cross-AVID also outperforms prior work by significant margins (2.7% on ESC and 3% on DCASE).

Bibliography

[1] Alwassel, H., Mahajan, D., Torresani, L., Ghanem, B., Tran, D.: Self-supervised learning by cross-modal audio-video clustering. arXiv:1911.12667 (2019)
[2] Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: ICCV (2017)
[3] Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: Learning sound representations from unlabeled video. In: Advances in Neural Information Processing Systems (NeurIPS) (2016)
[4] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. arXiv:2002.05709 (2020)
[5] Dosovitskiy, A., Fischer, P., Springenberg, J.T., Riedmiller, M., Brox, T.: Discriminative unsupervised feature learning with exemplar convolutional neural networks. TPAMI **38**(9), 1734–1747 (2016)
[6] Gemmeke, J., Ellis, D., Freedman, D., Jansen, A., Lawrence, W., Moore, R., Plakal, M., Ritter, M.: Audioset: An ontology and human-labeled dataset for audio events. In: ICASSP (2017)
[7] Gutmann, M., Hyvärinen, A.: Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In: ICAIS (2010)
[8] Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: CVPR (2006)
[9] Han, T., Xie, W., Zisserman, A.: Video representation learning by dense predictive coding. In: Workshop on Large Scale Holistic Video Understanding, ICCV (2019)
[10] Korbar, B., Tran, D., Torresani, L.: Cooperative learning of audio and video models from self-supervised synchronization. In: NeurIPS (2018)
[11] Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: a large video database for human motion recognition. In: ICCV. IEEE (2011)
[12] Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv:1807.03748 (2018)
[13] Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: ECCV (2018)
[14] Owens, A., Wu, J., McDermott, J.H., Freeman, W.T., Torralba, A.: Ambient sound provides supervision for visual learning. In: ECCV (2016)
[15] Piczak, K.J.: Environmental sound classification with convolutional neural networks. In: IEEE International Workshop on Machine Learning for Signal Processing (MLSP) (2015)
[16] Piczak, K.J.: ESC: Dataset for environmental sound classification. In: ACM Multimedia (2015)
[17] Sailor, H.B., Agrawal, D.M., Patil, H.A.: Unsupervised filterbank learning using convolutional restricted boltzmann machine for environmental sound classification. In: InterSpeech (2017)
[18] Soomro, K., Zamir, A.R., Shah, M.: UCF101: A dataset of 101 human actions classes from videos in the wild. Tech. Rep. CRCV-TR-12-01 (2012)
[19] Stowell, D., Giannoulis, D., Benetos, E., Lagrange, M., Plumbley, M.D.: Detection and classification of acoustic scenes and events. IEEE Trans on Multimedia **17**(10), 1733–1746 (2015)
[20] Sun, C., Baradel, F., Murphy, K., Schmid, C.: Contrastive bidirectional transformer for temporal representation learning. arXiv preprint arXiv:1906.05743 (2019)
[21] Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. In: Workshop on Self-Supervised Learning, ICML (2019)
[22] Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: CVPR (2018)
[23] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman: The kinetics human action video dataset. arXiv:1705.06950 (2017)
[24] Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: CVPR (2018)