Explainable Audio-Visual Representation Learning via Prototypical Contrastive Masked Autoencoder

Anonymous Author(s)
Affiliation
email

Abstract

In this paper, we propose a self-supervised prototypical contrastive audio-visual masked autoencoder (PCAV-MAE) to learn a joint and coordinated audio-visual representation. Different from conventional techniques, we calculate prototypes as latent variables and reconstruct the masked tokens by encouraging them to be closer to their assigned prototypes with contrastive learning. This design not only allows us to learn a joint representation but also helps to learn the intrinsic semantic information of videos. We demonstrate the transferability of our representations, achieving state-of-the-art audio-visual results in downstream tasks. As a result, our fully self-supervised pre-trained CAV-MAE achieves a new SOTA accuracy of 69.9% on AudioSet and is comparable with the previous best supervised pre-trained model on VGGSound over audio-visual event classification.

1 Introduction

Acoustic and visual modalities have different properties, yet humans can seamlessly connect and integrate them to perceive the world. Developing deep learning algorithms to replicate these abilities, especially for multi-modal audio-visual fusion and retrieval, is of great interest (1; 2). Since manually annotating audio and video is expensive and difficult to scale, utilizing web-scale unlabeled video data in a self-supervised manner has become a core research question. Recent advances, such as the development of contrastive learning techniques (3; 4), have significantly enhanced the capability of models to learn from multi-modal data in a self-supervised manner. Audio-visual representation learning leverages the complementary nature of audio and visual information to improve the performance of various downstream tasks, including speech recognition (5), video understanding (6), and emotion recognition (7). By integrating data from both modalities, models can achieve a more comprehensive understanding of the environment or context (2), leading to more robust and accurate results.

Despite their advancements, audio-visual models (1; 8) share a common weakness: the representation is not encouraged to encode the semantic structure of data. For example, Gong et al. combine masked data modeling and contrastive learning, two major self-supervised learning frameworks, to learn a fused audio-visual representation (3). However, two separate amples are treated as a negative pair as long as they are from different instances, regardless of their semantic similarity. This issue is magnified by the fact that thousands of negative samples are generated to form the contrastive loss, leading to many negative pairs that share similar semantics but are undesirably pushed apart in the embedding space.

To overcome this drawback, in this paper, we propose a prototypical contrastive audio-visual masked autoencoder (PCAV-MAE) to learn a joint audio-visual representation that encodes the semantic structure of data into the embedding space. First, we tokenize input video frames and audio spectra and mask the majority of them. Only the remaining visible subsets are fed into the visual encoder and audio encoder. Moreover, different from conventional techniques, we calculate prototypes as "a representative embedding for a group of semantically similar instances" and assign several prototypes of different granularity to each instance. We reconstruct the masked tokens by encouraging them to be closer to their assigned prototypes with contrastive learning. In practice, we find prototypes by performing clustering on the embeddings. The goal of prototypical contrastive learning is to find the network parameters that best describe the data.

2 Proposed Method

Our proposed framework is presented in Figure 1.

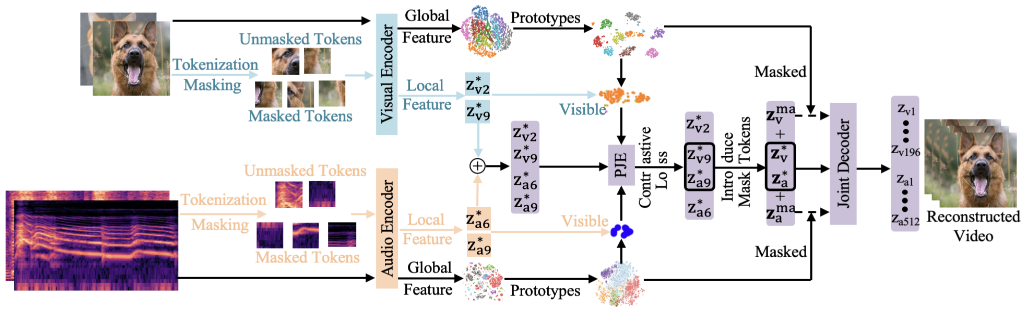


Figure 1: Proposed prototypical contrastive audio-visual masked autoencoder (PCAV-MAE).

2.1 Pre-processing, Tokenization and Masking

In this work, we utilize 10-second videos (with parallel audios) from VGGSound (9) and AudioSet (10) for pre-training and fine-tuning the model. Each 10-second video is sampled at 1 frame per second (FPS). In the training stage, one RGB frame is randomly selected as the training data. We resize and center crop each RGB frame to 224×224, and then split it into 196 16×16 square patches $\mathbf{v} = [v_1 ... v_{196}]$. For audio, we convert each 10-second audio waveform into a sequence of 128-dimensional log Mel filterbank features, computed with a 25ms Hanning window every 10ms (11). Then, we split the obtained 1024 (time) $\times$ 128 (frequency) spectrogram into 512 16×16 square patches $\mathbf{a} = [a_1 ... a_{512}]$. In the inference stage, we average the model's prediction for each RGB frame to produce the video prediction. Inspired by (12; 4), we randomly mask 75% of video $\mathbf{v}^{ma}$ and 50% of audio $\mathbf{a}^{ma}$ tokens.

2.2 Prototypical Joint Enocder

Given a full-set of video data v and audio data a, in the global representation learning routine (i.e., black arrows), we input them into the visual encoder and audio encoder to obtain the representation $\mathbf{z}_v$ and $\mathbf{z}_a$, respectively. We then calculate global prototypes of the full-set data which are latent variables. To achieve that, we use the local peaks of the density (13) as the prototype, in other words, the most representative data samples of v and a. The goal of the proposed prototypical joint encoder (PJE) is to find a network parameter that maximizes the log-likelihood function between representation of visible video and audio patches by a prototype-wise contrastive audio-visual learning (PCAV). The loss, namely $\mathcal{L}_{\text{PJE}}$, is defined as:

$$\mathcal{L}_{\text{PJE}} = \frac{1}{\tau\,|\mathcal{M}|} \sum_{p_{vi}^{+} \in \mathcal{M}} - \log \frac{\exp\left(z_{aj}^{*} \cdot p_{vi}^{+}/\gamma\right)}{\sum_{p_{vi}^{-} \in \mathcal{N}} \exp\left(z_{aj}^{*} \cdot p_{vi}^{-}/\gamma\right)} \tag{1}$$

$\mathcal{M}$ and $\mathcal{N}$ are prototype collections of the positive and negative samples, respectively. The prototype of the i-th visual patch and the visible representation of the j-th audio patch are denoted as p_{vi} and z_{aj}, respectively. We set the temperature τ to 0.1 as shown in Sec. **??**. Inspired from previous supervised learning work (14)(15), we find different levels of concentration distributes around each prototype embeddings. Therefore, we exploit γ as the concentration level around the joint prototype p^m within $I \times J$ potential combinations of audio and video patches as:

$$\gamma = \frac{\sum_{i=1}^{I} \sum_{j=1}^{J} (\|p^m - z_{vi}^{*}\|_2 + \|p^m - z_{aj}^{*}\|_2)}{IJ \log(I + J + \beta)} \tag{2}$$

where the momentum features are denoted as $\{v_i^m\}_{i=1}^n$ within the same cluster as a prototype p. We set a smooth parameter β to ensure that small clusters do not have an overly-large γ. In the proposed prototype clustering, γ acts as a scaling factor on the similarity between an embedding v and its prototype p.

2.3 Prototypical Joint Decoder

In conventional masked autoencoder frameworks (12; 8; 16), decoders utilize Transformers that reconstruct the masked tokens given the encoded tokens as context, audio, and images. These Transformer-based decoders have less capacity than encoders to force encoders learn discriminative features which can be utilized for reconstruction. Moreover, this also improves training efficiency, as masked tokens are also processed by decoders. Therefore, we follow a vanilla Transformer (17) architecture, whilst also being shallower, to build up the joint decoder.

Different from decoders in previously mentioned masked autoencoders, we propose to use prototypes of masked tokens to assist the reconstruction. As described previously, the representation is not encouraged to encode the semantic structure of data. However, two samples are treated as a negative pair as long as they are from different instances, regardless of their semantic similarity. Therefore, to address the limitation and achieve a high accuracy of reconstruction, we learn the semantic structure of data.

During reconstruction, the contrastive learning objective aligns the features of the masked patches with their closest prototypes, ensuring that the reconstructed patches are accurate and semantically consistent. The loss, namely $\mathcal{L}_{\text{PCPE}}$, is defined as:

$$\frac{1}{\tau\,|\mathcal{M}|}\sum_{p_{vi}^{+},p_{aj}^{+}\in\mathcal{M}} -\log\frac{\exp((z_{p_{aj}}^{*}\cdot z_{p_{aj}}^{+}+z_{p_{vi}}^{*}\cdot z_{p_{vi}}^{+})/\gamma)}{\sum\limits_{p_{vi}^{-},p_{aj}^{-}\in\mathcal{N}}\exp((z_{p_{aj}}^{*}\cdot z_{p_{aj}}^{-}+z_{p_{vi}}^{*}\cdot z_{p_{vi}}^{-})/\gamma)} \tag{3}$$

This method enhances the PCAV-MAE's ability to handle complex scenes, ultimately leading to better video reconstruction by effectively linking masked patches to their corresponding positions through the use of prototypes. Our reconstruction loss function computes the mean squared error (MSE) between the masked patches of the reconstructed and original images as:

$$\mathcal{L}_{\text{r}} = \sum_{j=1}^{J}(\hat{a}_j^* - a_j^*)^2 + \sum_{i=1}^{I}(\hat{v}_i^* - v_i^*)^2 \tag{4}$$

where $\hat{a}_j^*$ and $\hat{v}_i^*$ are reconstructed unmasked tokens. Our overall objective in the pre-training is the sum of equations (1), (3) and (4) as $\mathcal{L}_{\text{PCAV-MAE}} = \mathcal{L}_{\text{PJE}} + \mathcal{L}_{\text{PCPE}} + \mathcal{L}_{\text{r}}$. After pre-training, we abandon the decoder and only keep the encoders of the model for downstream tasks. We can use the sum of the single-modality stream output and the multi-modal modality stream output, or just the multi-modal stream output for fine-tuning. They perform similarly in our experiments.

3 Experiments

3.1 Datasets and Attacks

We use the full training set (unbalanced + balanced) of AudioSet (10) pre-training. In the AS-2M task, we fine-tune on the full training set. In the AS-20K task, we fine-tune only on the 20K balanced training set. We randomly select 170,000 clips from VGGSound (9) for fine-tuning and 14,448 clips for inference.

3.2 Implementation Details

In the pre-training, we use two 11-layer Transformers (each is 768-dimensional) as the audio and visual encoders, respectively. The decoder is a shallower vanilla Transformer with a hidden dimension of 384, 4 layers, 6 attention heads, and an MLP dimension of 1536. The joint decoder is discarded after pre-training. We set $\beta = 10$. We pre-train the model using the AdamW optimizer with a momentum of 0.9, an accumulated batch size of 512, and a learning rate of 0.0002. We pre-train for 400 epochs for the PJE and 200 epochs for the joint decoder.

3.3 Results

The learned representation is evaluated on fine-tuning performance over AS-20K and VGGSound, alongside recent competitor models. Tables 1&2 show the results.

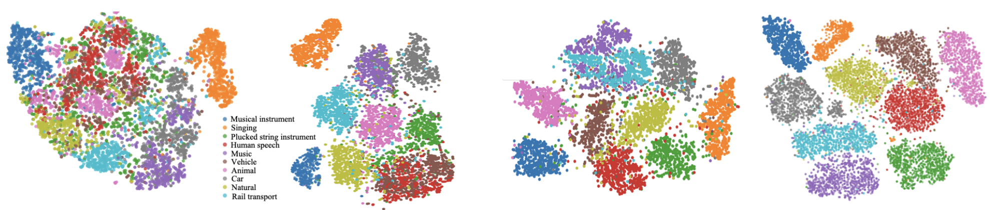
Table 1: Top-1 testing classification accuracy (Acc) on VGGSound (VS).

Method	A	V	AV
AV-MAE (8)	57.2	50.3	65.0
TSS (18)	39.1	39.7	53.9
AVS (19)	38.5	39.0	53.4
AV-LLM (6)	42.3	40.3	53.7
CAV-MAE (3)	59.5	47.0	65.5
MAViL (4)	**60.8**	<u>50.9</u>	67.1
Fusion (1)	47.0	40.9	59.1
MMT (20)	57.6	44.8	66.2
Mirasol3B (21)	59.9	50.1	<u>69.8</u>
Ours	<u>60.5</u>	**51.8**	**69.9**

Table 2: Mean Average Precision (mAP) comparison of AV classification on AudioSet-20K and AudioSet-2M.

Method	AudioSet-20K			AudioSet-2M		
	A	V	A-V	A	V	A-V
AV-MAE (8)	35.8	23.9	45.9	46.6	<u>31.1</u>	51.8
TSS (18)	20.4	14.8	37.3	36.2	21.1	42.5
AVS (19)	22.0	40.3	51.7	34.2	24.7	43.6
AV-LLM (6)	27.5	40.4	52.2	38.3	23.9	46.8
CAV-MAE (3)	37.7	19.8	42.0	46.6	26.2	51.2
MAViL (4)	<u>41.8</u>	24.8	44.9	<u>48.7</u>	30.3	<u>53.3</u>
Fusion (1)	31.5	39.3	<u>54.6</u>	35.6	26.7	50.2
MMT (20)	39.3	<u>40.7</u>	51.1	41.5	26.7	49.0
Mirasol3B (21)	38.6	29.5	52.0	46.7	28.7	53.0
Ours	**42.0**	**42.3**	**57.8**	**49.1**	**33.8**	**55.4**

It can be observed that: (1) The proposed PCAV-MAE offers the best effectiveness. In particular, PCAV-MAE surpasses CAV-MAE (3) in A, V, and A+V tasks by a large margin. (2) PCAV-MAE ranks as the second-best in the audio classification task, showing slightly lower accuracy compared to the state-of-the-art model, MAViL (4). This is likely because MAViL incorporates context along with audio and video inputs, enabling a better understanding and interpretation of the audio signals.



(a) AV-MAE. (b) Contrastive fusion encoder (c) (a) + PJE (d) PCAE-MAE

Figure 2: T-SNE feature visualization of the unsupervised learned representation for AudioSet training samples from the first 10 classes.

As qualitative analysis, Figure 2 presents the t-distributed stochastic neighbour embedding (t-SNE) visualisation of the baseline and proposed models on AudioSet. Compared to the representation learned by AV-MAE and contrastive fusion encoder, the representation learned by PCAV-MAE forms more separated clusters, which also suggests representation of lower entropy. In Figure 5(b), it can be observed that the feature embeddings within a single prototype are not separable. However, when the PCPE is added in Figure 5(c), individual instances become separated. This demonstrates that the proposed methods can learn better semantic structure of data that enhances discriminative feature representation learning.

4 Conclusion

In this paper, we have proposed a self-supervised audio-visual representation learning approach, offering an effective alternative to traditional supervised pipelines. We reconstructed masked tokens in multi-modal MAE by encouraging them to be closer to their assigned prototypes through contrastive learning. The model learned not only joint representation learning but also intrinsic semantic information of multi-modal data. Our extensive experiments on multiple benchmarks demonstrated the advantage of PCAV-MAE for unsupervised representation learning. Additionally, prototypes offered interpretations compared to baselines, enabling PCAV-MAE to provide more insights into performance improvement on downstream tasks.

References

[1] A. Senocak1, J. Kim, T.-H. Oh, D. Li, and I. S. Kweon, "Event-specific audio-visual fusion layers: a simple and new perspective on video understanding," *Proceedings of IEEE/CVF Winter Conference on Computer Vision (WACV)*, 2023.

[2] R. G. Praveen and J. Alam, "Audio-visual person verification based on recursive fusion of joint cross-attention," *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2024.

[3] Y. Gong, A. Rouditchenko, A. H. Liu, D. Harwath, L. Karlinsky, H. Kuehne, and J. Glass, "Contrastive audio-visual masked autoencoder," *Proceedings of International Conference on Learning Representations (ICLR)*, 2023.

[4] P.-Y., Huang, V. Sharma, H. Xu, C. Ryali, H. Fan, Y. Li, S.-W. Li, G. Ghosh, J. Malik, and C. Feichtenhofer, "MAViL: masked audio-video learners," *arXiv preprint arXiv:2212.08071*, 2022.

[5] C. Chen, Y. Hu, Q. Zhang, H. Zou, B. Zhu, and E. S. Chng, "Leveraging modality-specific representations for audio-visual speech recognition via reinforcement learning," *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*, 2023.

[6] F. Shu, L. Zhang, H. Jiang, and C. Xie, "Audio-visual LLM for video understanding," *arXiv preprint arXiv:2312.06720*, 2023.

[7] L. Sun, Z. Lian, B. Liu, and J. Tao, "HiCMAE: hierarchical contrastive masked autoencoder for self-supervised audio-visual emotion recognition," *Information Fusion*, vol. 108, p. 102382, 2024.

[8] M.-I. Georgescu, E. Fonseca, R. T. Ionescu, M. Lucic, C. Schmid, and A. Arnab, "Audiovisual masked autoencoders," *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[9] H. Chen, W. Xie, and a. A. Z. A. Vedaldi, "Vggsound: a large-scale audio-visual dataset," *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[10] J. Gemmeke, D. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: an ontology and human-labeled dataset for audio events," *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[11] Y. Li, Y. Sun, W. Wang, and S. M. Naqvi, "U-shaped Transformer with frequency-band aware attention for speech enhancement," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 31, p. 1511–1521, 2023.

[12] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[13] P. Angelov and E. Soares, "Towards explainable deep neural networks (xDNN)," *Neural Networks*, vol. 130, p. 185–194, 2020.

[14] Y. Li, P. Angelov, and N. Suri, "Self-supervised representation learning for adversarial attack detection," *Proceedings of European Conference on Computer Vision (ECCV)*, 2024.

[15] E. Soares, P. Angelov, and N. Suri, "Similarity-based deep neural network to detect imperceptible adversarial attacks," *Proceedings of IEEE Symposium Series on Computational Intelligence (SSCI)*, 2022.

[16] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image Transformers," *Proceedings of International Conference on Learning Representations (ICLR)*, 2021.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Proceedings of Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

[18] S. Jenni, A. Black, and J. Collomosse, “Audio-visual contrastive learning with temporal self-supervision,” *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*, 2023.

[19] J. Zhou, J. Wang, J. Zhang, W. Sun, J. Zhang, S. Birchfield, D. Guo, L. Kong, M. Wang, and Y. Zhong, “Audiovisual segmentation,” *Proceedings of European Conference on Computer Vision (ECCV)*, 2022.

[20] W. Zhu, J. Yi, X. Sun, X. Hao, L. Liu, and M. Omar, “Multiscale multimodal Transformer for multimodal action recognition,” *Proceedings of International Conference on Learning Representations (ICLR)*, 2022.

[21] A. Piergiovanni, I. Noble, D. Kim, M. S. Ryoo, V. Gomes, and A. Angelova, “Mirasol3B: a multimodal autoregressive model for time-aligned and contextual modalities,” *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.