ENHANCING CONTRASTIVE LEARNING WITH TEMPORAL COGNIZANCE FOR AUDIO-VISUAL REPRESENTATION GENERATION

Chandrashekhar Lavania, Shiva Sundaram, Sundararajan Srinivasan, Katrin Kirchhoff

Amazon
{clavania,sssundar,sundarsr,katrinki}@amazon.com

ABSTRACT

Audio-visual data allows us to leverage different modalities for downstream tasks. The idea being individual streams can complement each other in the given task, thereby resulting in a model with improved performance. In this work, we present our experimental results on action recognition and video summarization tasks. The proposed modeling approach builds upon the recent advances in contrastive loss based audio-visual representation learning. Temporally cognizant audio-visual discrimination is achieved in a Transformer model by learning with a masked feature reconstruction loss over a fixed time window in addition to learning via contrastive loss. Overall, our results indicate that the addition of temporal information significantly improved the performance of the contrastive loss based framework. We achieve an action classification accuracy of 66.2% versus the next best baseline at 64.7% on the HMDB dataset. For video summarization, we attain an F1 score of 43.5 verses 42.2 on the SumMe dataset.

Index Terms— representation learning, action recognition, video summarization, contrastive loss, transformers

1. INTRODUCTION

Large volume of multimodal data is generated everyday in several forms such as videos, texts, audios, sensor streams and many more. Multiple modalities in the data can be processed individually, or combined for downstream tasks. Any improvement in performance based on multimodal input is a function of the available signals and downstream tasks. For example, utilizing both the speech and facial expressions [1] in a video can give information about the emotions being conveyed by a speaker. The success of multimodality for emotion tracking is due to integration of audio-visual cues that are beneficial in tracking valence and arousal states [2]. The aim of this work is to produce such useful multimodal representations. While the focus of this work is in combining audio-visual information, it can also be used for combining other modalities such as speech, text and video.

The proposed temporally cognizant audio-visual representation learning approach uses contrastive learning along with Transformer based masked reconstructions. Temporal cognizance is important for scenarios where the knowledge of context is valuable (such as speech, action events and many more). Therefore, these representations of a video snippet need to not only encode information from the visual and audio modalities of the snippet but also incorporate the influence from any past and future snippets from the same video. The contrastive loss focuses on the synchronization and similarity between the audio and visual representations. On the other hand, the Transformer based masked reconstruction helps to incorporate the temporal relationships between snippets from the same video. The framework is an offline model with access to the whole video during training. The performance of the learned representations is evaluated on publicly available benchmarks for action recognition and video summarization.

2. RELATED WORK

A large number of works have explored audio-visual representations through the contrastive training [3] of their models. For example, Arandjelovic and Zisserman [4] and Arandjelovic et al. [5] learn models to determine if the audio and visual feeds correspond to each other. Owens and Efros [6] use a self-supervised approach to learn a model that can determine whether a given pair of audio and visuals are synchronized. Alayrac et al. [7] and Akbari et al. [8] use not just audio and visuals, but also text while incorporating contrastive loss to learn representations for multiple modalities.

In addition to this, Cheng et al. [9] utilize cross-modal attention to determine the synchronization between audio and visuals. Morgado et al. [10] utilize cross-modal instance discrimination along with cross modal agreement to train their audio-visual model in a contrastive manner. Piergiovanni et al. [11] use the concept of evolving loss to train their model. They employ multiple self-supervised tasks such as reconstruction, prediction, temporal ordering and multimodal alignment. An evolutionary algorithm is used to find the optimum combination of these tasks. They use audio, RGB based visuals and optical flow as the three modalities to train their framework. These approaches demonstrate the benefits of contrastive training for audio-visual representation learning.

There are, however, other approaches that do not directly use contrastive loss based training. Alwassel et al. [12] learn

an audio-visual representations through self-supervised cross modal audio-visual clustering. They use visual and audio encoders to produce representations. These representations are clustered to produce pseudo labels. Their procedure alternates between clustering and training using these pseudo labels. Aytar et al. [13] on the other hand use a teacher-student framework. The teacher is a pretrained network that is capable of extracting probabilities from visual images. The student network operates on audio and produces another probability distribution. The aim is to reduce the KL divergence between the distributions produced by the teacher and the student networks.

2.1. Our Contributions

Contrastive loss based audio-visual feature learning and sequence to sequence networks such as RNNs and Transformers have seen application in multimodal setups, especially those involving images and text [10, 7, 5, 14, 6, 4]. However, an end-to-end system that is a combination of contrastive loss based audio-visual synchronization framework that feeds into a sequence-to-sequence model has not been extensively explored. Specifically, the proposed architecture is capable of producing joint representations that are learned directly as part of the training instead of concatenating or averaging of the features from the audio and visual arms [5]. Such training encourages the learned representations to not only extract information specific to the modalities but also any knowledge that arises due to the presence of both the modalities in the data source. In addition to this, the temporal knowledge that arises due to the visual and auditory continuity in videos is also absorbed into the network. The added benefit of this temporal cognizance via masked reconstruction is demonstrated through the performance of the learned model on downstream tasks of action recognition and video summarization as shown in Section 6.

3. DETAILED APPROACH

Morgado et al. [10] (AVID-CMA) propose an approach for learning representations using a contrastive loss. In their approach, the similarity is calculated based on cross-modal comparisons. In addition to this, the sampling of positive samples also takes into account video segments that are semantically related. They claim that two segments have high semantic similarity when both the visuals and audios of the two segments have high similarities. Their work is state of the art for producing visual and audio representations through contrastive learning.

Our approach is a Transformer based encoder on top of the audio and visual branches learned using the AVID-CMA procedure. The architecture can be described as shown in Figure 1. Since the self attention layer ranges over the entire input signal, the resulting embedding of a given video snippet incorporates the influence from both past and future snippets in the video. The outputs from these branches are fed into a projection block that combines these outputs from different modalities into a single joint representation. These joint representations are further combined with a clip index based positional encoding. The resultant representations are the input to the Transformer based encoder that is trained using the masked language modeling (MLM) procedure [15]. Therefore, the learned representations of random video snippets are masked at the input to the Transformer. The corresponding output is then passed through splitting layers to reconstruct the visual and audio features.

The framework is trained using two losses: 1) The noise contrastive estimation (NCE) [16] based loss used in AVID-CMA [10] (described in equation 3) and 2) a reconstruction error (equation 4) between the outputs from the audio and visual branches for the masked time step and their corresponding split reconstructions (produced from the Transformer encoder output). Let v_i and a_i be the representations produced by the network for the visual and audio modalities of video snippet $i \in \mathcal{S}$, where $\mathcal{S}$ is the set of all snippets in all videos. These losses are therefore defined as follows:

$$L_{\text{AVID}}(v_i, a_i) = L_{NCE}(v_i; \bar{a}_i, \mathcal{N}_i) + L_{NCE}(a_i; \bar{v}_i, \mathcal{N}_i) \quad (1)$$

$$L_{\text{M}}(v_i, a_i) = \sum_{p \in \mathcal{P}_i} L_{NCE}(v_i; \bar{v}_p, \mathcal{N}_i) + L_{NCE}(a_i; \bar{a}_p, \mathcal{N}_i) \quad (2)$$

$$L_{\text{AVID-CMA}} = L_{\text{AVID}} + \lambda L_{\text{M}} \quad (3)$$

$$L_{\text{rec}} = \frac{1}{|\Gamma|} \sum_{t \in \Gamma} ||\hat{v}_t - v_t||_2 + ||\hat{a}_t - a_t||_2 \quad (4)$$

Here L_{NCE} is the noise contrastive estimation based loss [10]. $\bar{a}_i$ is a moving average of the features a_i of the audio modality in snippet i. $\bar{v}_i$ is defined in a similar manner for visual features. $\mathcal{N}_i$ is the set of negative samples corresponding to snippet i and $\mathcal{P}_i$ is the set of positive samples (including semantically similar samples) for snippet i. F_t is the joint multimodal representation of the snippet at time step t. Therefore, F_1 to F_T are the audio-visual representations of snippets from a video that has T time steps. The reconstructed outputs $\hat{v}_t$ and $\hat{a}_t$ are compared with the corresponding outputs from the visual and audio branches to produce L_{rec}. Γ is the set of time steps in the Transformer encoder input that have been masked and $\lambda > 0$ is the weight of L_M.

During training, every alternate epoch uses the same type of loss. Therefore, in the first epoch L_{rec} is used for training followed by $L_{\text{AVID-CMA}}$ in the next epoch and so on. This procedure allows for course correction during training without the need for balancing both losses at the same time. L_{rec}

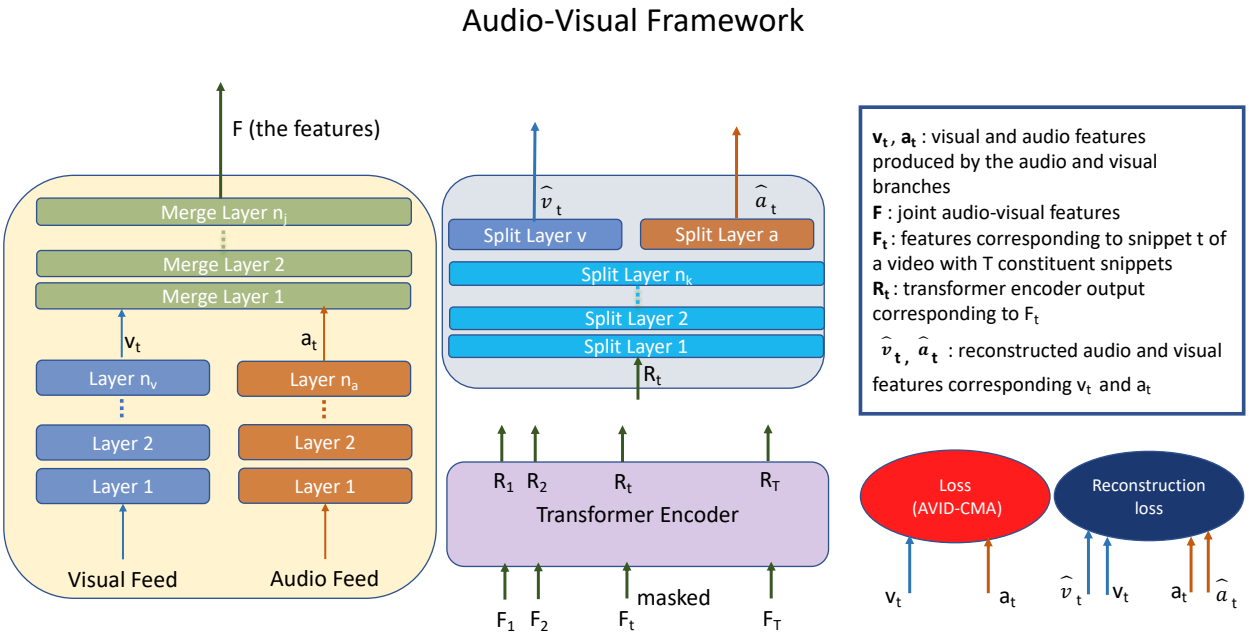

Fig. 1. The overall architecture of the end-to-end training framework. The visual and audio feeds are passed through their respective branches before being merged into a joint representation. This combined representation is used as input to a Transformer based encoder. Some of the representations (randomly selected) from the video are masked before being fed into the Transformer. The outputs corresponding to these masked inputs are then passed through a splitting network to produce representations corresponding to the visual and audio branches. Outputs at different stages utilize $L_{\text{AVID-CMA}}$ or L_{rec} during training.

attempts to ensure that any temporal knowledge gets incorporated into the learned representation. However, using only L_{rec} does not encourage the model to encode any knowledge due to the alignment of audio and visual streams. Hence, $L_{\text{AVID-CMA}}$ is used to incorporate such knowledge. As a result the overall trained model is temporally cognizant along with assimilating information due to the alignment between modalities.

4. DATA

HMDB [17] dataset is used for the task of action recognition. This data consists of video clips extracted from public sources such as Prelinger archive, YouTube and Google videos. The data is divided into 51 action categories. The categories include several commonly performed actions such as clapping, laughing, eating, and several type of body movements. In addition to this, the dataset also includes metadata such as visible body parts, occlusions, camera motion and many more. Three train-test splits are provided as part of the dataset.

The task of video summarization is performed on the SumMe [18] dataset. There are 25 videos in this dataset. Each video has a set of user produced summaries (the ground truth). The dataset is a combination of egocentric videos, static videos and those with camera motion.

For both these datasets, temporal cognizance plays an important role. In the case of action recognition, individual actions are performed over several video frames. Therefore recognizing that a given sequence of frames jointly represent an action is important. Similarly, in video summarization, salient events transpire over a period of time. Hence, it is important that the model used for these tasks is temporally cognizant.

5. EXPERIMENT SETUP

The individual visual and audio branches of the frame work follow the AVID-CMA setup. Therefore, the visual branch uses a 18 layer R(2+1)D network [19] and the audio branch utilizes a 9 layer (2D) CNN. There is a projection network that combines the two branches. The projection is done through two fully connected layers. The projection network feeds into the Transformer encoder that consists of 8 attention heads, and 2 Transformer layers. $|\Gamma|$ is set to 30% of the sequence length. λ here is used for scaling and can be set to $1/|\mathcal{P}_i|$ with $|\mathcal{P}_i| = 32$. In addition to this, the splitting network uses two fully connected layers to split the Transformer encoder output into audio and visual representations. The input of the audio branch is a 200x257 sized spectrogram for each video snippet. The visual branch takes in 'k' 224x224 sized visual frames where k is the number of frames in the video snippet. During training, the visual and audio branches of the proposed framework are initialized using a pretrained AVID-CMA model that

Fig. 2. The top row shows a summary when the temporal cognizance based proposed model is used in the summarization procedure. The bottom row does not use temporal cognizance model. It can be seen that unlike the bottom row (which contains only chopping of food) the whole procedure of building the food tower (columns 1-11) is considered as part of summary in the top row thereby showing the importance of temporal cognizance.

was trained on Audioset [20]. Once initialized, the balanced set from Audioset is used to train the whole framework.

For the task of action recognition, thirty two 224x224 frames are used as input and the trained network is finetuned on the HMDB dataset. We follow the procedure used in AVID-CMA [10] for evaluation. Therefore, snippet level predictions from 10 uniformly sampled snippets are averaged to produce video level predictions. The overall reported performance is the average score from the 3 splits made available in the dataset.

In the case of video summarization, for a fair comparison we use the same framework as Zhou et al. [21] (unsupervised approach) and change the input based on the features that are being compared. The same F1 score based mechanism as used in [21, 22] is utilized for comparing performance.

6. RESULTS

We evaluate the effectiveness of the learned model for the downstream tasks of action recognition and video summarization. For each of these tasks, a comparison is made with several baselines such as XDC [12] and AVID-CMA [10].

The proposed procedure can significantly outperform the baselines (p-value < 0.05 using a t-test). Table 1 shows the performance for the task of action recognition using the HMDB dataset. Similarly, Table 2 displays the capability of our framework for video summarization using the SumMe dataset. Figure 2 shows an example that demonstrates the benefit of out temporally cognizant procedure. These results give support to our usage of Transformer based masked reconstruction loss on top of contrastive loss based training to introduce temporal cognizance.

7. CONCLUSIONS AND FUTURE WORK

We propose an audio-visual framework that introduces temporal cognizance to contrastive loss based audio-visual representation learning. It uses a combination of contrastive loss ($L_{\text{AVID-CMA}}$) and a Transformer based masked reconstruction of visual and audio features from joint audio-visual representations. The proposed framework was utilized for the tasks of action recognition (using HMDB [17]) and video summarization (using SumMe [18]). The addition of temporal cognizance via masked reconstruction on top of contrastive learning demonstrates significant improvement over several baselines including contrastive loss based AVID-CMA [10].

Approach	Accuracy
L3 [4, 10]	51.6
AVTS [14]	61.6
XDC [12]	63.7
AVID-CMA [10]	64.7
Ours	**66.2**

Table 1. Performance comparison for the action recognition task using the HMDB dataset. The reported scores are classification accuracy.

Approach	Score
Reinforce [21]	41.4
XDC [12]	41.4
AVID-CMA [10]	42.2
Ours (visual only branch)	*42.7*
Ours (audio-visual)	**43.5**

Table 2. Performance comparison for the video summarization task using the SumMe dataset. The F1 scores are reported as used in [21, 22].

As mentioned in Morgado et al. [10], the size of input during training of the multimodal model influences the performance on downstream tasks. Therefore, as part of future work, the influence of various input sizes during training can be explored on the action recognition and video summarization tasks. In addition to this, another potential avenue for exploration is into making the model more robust to missing modalities [23]. Furthermore, integration of other sources of information such as text in video captioning can also be explored. In addition to this an online setting can also be explored. For example, the block processing mechanism of [1] or the streaming Transformer procedure of [24] can be explored further in our context.

8. REFERENCES

[1] Srinivas Parthasarathy and Shiva Sundaram, "Detecting expressions with multimodal transformers," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 636–643.

[2] James A Russell, "A circumplex model of affect.," *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161, 1980.

[3] Joon Son Chung and Andrew Zisserman, "Out of time: automated lip sync in the wild," in *Asian conference on computer vision*. Springer, 2016, pp. 251–263.

[4] Relja Arandjelovic and Andrew Zisserman, "Look, listen and learn," in *ICCV*, 2017, pp. 609–617.

[5] Relja Arandjelovic and Andrew Zisserman, "Objects that sound," in *ECCV*, 2018, pp. 435–451.

[6] Andrew Owens and Alexei A Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *ECCV*, 2018, pp. 631–648.

[7] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelovic, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman, "Self-supervised multimodal versatile networks.," *NeurIPS*, vol. 2, no. 6, pp. 7, 2020.

[8] Hassan Akbari, Linagzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong, "Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text," *arXiv preprint arXiv:2104.11178*, 2021.

[9] Ying Cheng, Ruize Wang, Zhihao Pan, Rui Feng, and Yuejie Zhang, "Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning," in *ACM Multimedia*, 2020, pp. 3884–3892.

[10] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra, "Audio-visual instance discrimination with cross-modal agreement," in *CVPR*, 2021, pp. 12475–12486.

[11] AJ Piergiovanni, Anelia Angelova, and Michael S Ryoo, "Evolving losses for unsupervised video representation learning," in *CVPR*, 2020, pp. 133–142.

[12] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran, "Self-supervised learning by cross-modal audio-video clustering," *NeurIPS*, vol. 33, 2020.

[13] Yusuf Aytar, Carl Vondrick, and Antonio Torralba, "See, hear, and read: Deep aligned representations," *arXiv preprint arXiv:1706.00932*, 2017.

[14] Bruno Korbar, Du Tran, and Lorenzo Torresani, "Cooperative learning of audio and video models from self-supervised synchronization," *arXiv preprint arXiv:1807.00230*, 2018.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[16] Michael Gutmann and Aapo Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 297–304.

[17] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre, "Hmdb: a large video database for human motion recognition," in *ICCV*, 2011, pp. 2556–2563.

[18] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool, "Creating summaries from user videos," in *ECCV*, 2014.

[19] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *CVPR*, 2018, pp. 6450–6459.

[20] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *ICASSP*, 2017.

[21] Kaiyang Zhou, Yu Qiao, and Tao Xiang, "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward," in *AAAI*, 2018, vol. 32.

[22] Li Yuan, Francis EH Tay, Ping Li, Li Zhou, and Jiashi Feng, "Cycle-sum: cycle-consistent adversarial lstm networks for unsupervised video summarization," in *AAAI*, 2019, vol. 33, pp. 9143–9150.

[23] Srinivas Parthasarathy and Shiva Sundaram, "Training strategies to handle missing modalities for audio-visual expression recognition," in *Companion Publication of the 2020 International Conference on Multimodal Interaction*, 2020, pp. 400–404.

[24] Niko Moritz, Takaaki Hori, and Jonathan Le, "Streaming automatic speech recognition with the transformer model," in *ICASSP*. IEEE, 2020, pp. 6074–6078.