Language-guided Multi-Modal Fusion for Video Action Recognition

Jenhao Hsiao, Yikang Li, Chiuman Ho
OPPO US Research Center
{mark, yikang.li1, chiuman}@oppo.com

Abstract

A recent study [30] has found that training a multi-modal network often produces a network that has not learned the proper parameters for video action recognition. These multi-modal network models perform normally during training but fall short to its single modality counterpart when testing. The main cause for this performance drop could be two-fold. First, conventional methods use a poor fusion mechanism, where each modality is trained separately and then simply combine together (e.g., late feature fusion). Second, collecting videos is much more expensive than images. The insufficient video data can hardly provide support for training a multi-modal network that has a larger and more complex weight space.

In this paper, we proposed the Language-guided Multi-Modal Fusion to address the above poor fusion problem. A sophisticatedly designed bi-modal video encoder is used to fuse audio and visual signal to generate a finer video representation. To ensure the over-fitting can be avoid, we use a language-guided contrastive learning to largely augment the video data to support the learning of multi-modal network. On a large-scale benchmark video dataset, the proposed method successfully elevates the accuracy of video action recognition.

1. Introduction

The explosive growth in video and its applications has drawn considerable interest in the computer vision community and boost the need of high-level video understanding and effectively recognize human actions. However, it is a particularly challenging problem due to the complicated nature of videos, including large intra-class variations and complex temporal structures. In recent years, the accuracy gains for video action recognition have come from the newly designed CNN architectures (e.g., 3D-CNNs), and most contemporary models for video analysis exploit only the visual signal and ignore the audio signal. Traditional visual-only 3D-CNNs are thus prone to have limited recognition accuracy.

The fact that videos are intrinsically multimodal requires solutions that can explore not only static visual information but also audio clues. Given its high potential in facilitating video action recognition, researchers have attempted to utilize both audio and visual information in videos. However, end-to-end training of multi-modal video action recognition is non-trivial. In theory, a well-optimized multi-modal classifier should always match or outperform the best uni-modal classifier since the multi-modal network receives more information. However, we usually observe the opposite that the best uni-modal network often outperforms the multi-modal network. The main cause for this performance drop could be two-fold. First, conventional independent and separate training of multiple modalities (e.g., late feature fusion) may pose a poor modalities fusion mechanism. Since an action is usually complex and could span several video segments, simply concatenating features from different modalities could reversely increase the difficulty of network learning. Second, jointly trained multi-modal network have weight space that is too large for effective training. Since multi-modal network, especially for video action recognition, is much more complex than single modality counterpart or conventional 2D image network, it requires a great amount of video training data. However, acquiring a large amount of labeled video data will be prohibitively difficult and expensive comparing to image data collection.

In this paper, we propose the Language-guided Multi-Modal Fusion to address the above poor fusion and data scarcity problem. Our contributions are two-fold. First, A sophisticatedly designed bi-modal video encoder is used to jointly fuse both audio and visual signal to optimize the network prediction, which is able to exploit multi-modal features that are more comprehensive than those previously attempted. Second, to ensure the over-fitting can be avoid, we use a novel language-based contrastive learning to largely augment the video data. On large-scale video dataset, the proposed method successfully elevates the accuracy of video action recognition.

The rest of this paper is organized as follows. We discuss the related work in Section 2 and describe our contrastive audio-visual fusion network in Section 3. Section 4 presents

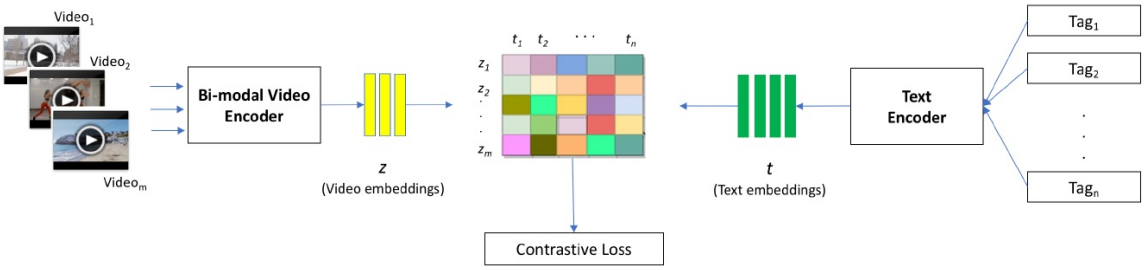

Figure 1. The proposed Contrastive Audio-Visual Fusion (CAV-Net) Architecture

the details of our experiment and Section 5 concludes this paper.

2. Related work

The goal of video action recognition aims to identify a single or multiple actions per video. In recent years, most of the accuracy gains for video action recognition have come from the introduction of new powerful architectures. 3D-CNNs [3, 9, 16, 21, 26, 27, 28, 29, 33] have been widely used to learn video features and classify video in a end-to-end manner. However, the 3D-CNNs proposed by the above methods are mainly focused on the design of convolution network architecture, and trained by single clip (e.g., gradients are updated based on one clip point of view), where irrelevant video segments could lead the gradient to the wrong direction and thus disrupt training performance.

Multi-modal networks [2, 10, 11, 19, 24] is another research track that aims to boost the video action recognition accuracy. However, simply concatenating output from individual network often reversely decrease the video-level prediction result since multi-modalities are trained independently and didn't be fused properly during the training process. Some recent works [35, 31, 5] has explored the idea of allowing different modalities to attend to each other. Unfortunately, previous works focus mostly on the text-visual retrieval, and has less attention on the video action recognition task. In [5], the author tried to use co-attention to better fuse audio-visual signal. However, it still faces the difficulty of insufficient video data for training a complex multi-modal network. Contrastive learning [20, 22] has recently been used for large-scale multi-modal network training, but mostly focus on the text-visual domain as well. In [20], MIL-NCE objective was proposed, but an averaged pooled frame features were used, which is a less effective video representation encoder.

In contrast to all the previous works, we sophisticatedly integrate the idea of bi-modal attention with language-guided contrastive learning to learn a better video representation and boost the accuracy of video action recognition application.

3. Method

Figure 1 shows an overview of the proposed Contrastive Audio-Visual Fusion framework (CAV-Net). The CAV-Net is composed by two neural networks: A Bi-modal Video Encoder and a Text Encoder. The two encoders produce similar embeddings if the video and the text contain similar visual and textual concepts. Below are the details of each module and training process.

3.1. Bi-modal Video Encoder

Figure 2 shows the architecture of the proposed bi-modal video encoder that can help generate a audio-visual-fused video representation. To represent a visual stream, we use a 3D-CNN network while for the audio stream we employ an audio CNN network (e.g., VGGish [7]), where both 3D-CNN or VGGish contain a set of convolutional layers (either 3D or 2D) to represent each video segment. The output of visual and audio CNN feature can be represented as $V = \{v_1, v_2, \ldots, v_n\} \in R^{d_v * n}$ and $A = \{a_1, a_2, \ldots, a_m\} \in R^{d_a * m}$ where n and m is the number of clips (or segments) of visual and audio feature respectively, and d_v and d_a is the feature dimension of visual and audio feature respectively.

Since each clip descriptor is produced by the visual- or audio-CNN module separately, the inter-clip relationships modeled by convolution are inherently implicit and local (e.g., each clip descriptor can only observe an extremely limited local event). This will become a performance bottleneck since the duration of different actions are variant and complex actions could span across multiple video segments. Hence, to capture the inter-clip dependencies for both short- and long-range dependencies, we first apply the bi-directional fusion module to strengthen the local clip descriptor of the target position via aggregating information from other positions (e.g., other video segments). The inter-clip relationships can be fused by a bi-directional attention layer to link different clips and can be expressed as:

$$B(S,T) = softmax(\frac{(W_q S)(W_k T)^T}{N_d})(W_v T), \quad (1)$$

where S and T are source and target vector, and W_q, W_k, and W_v denote linear transform matrices for query, key, value vector transformation. $(W_q S)(W_k T)^T$ model the bi-directional relationship between source and target (e.g., raw video or audio descriptor at different time segments), and N_d is the normalization factor.

$$V^{self} = B(V, V) \tag{2}$$

$$A^{self} = B(A, A) \tag{3}$$

Here we call V^{self} (resp. A^{self}) the visual (resp. audio) self-attended vector.

To fuse visual and audio feature, we use a cross-attention layer to integrate different modalities. Specifically, we use the visual self-attended vector to fuse with audio self-attended vector (and vice versa) to build the fused visual (audio) signal:

$$V^{fuse} = B(V^{self}, A^{self}) \tag{4}$$

$$A^{fuse} = B(A^{self}, V^{self}) \tag{5}$$

Here we call V^{fuse} (resp. A^{fuse}) the fused visual (resp. audio) vector.

Since an action typically represents only a subset of objects and events which are most relevant to the context of the video, directly taking average of fused vector for action class prediction would decrease the accuracy due to those irrelevant clips. Here we further introduce the adaptive pooling to adaptively pool fused-descriptors based on their significance so that the final video-level action recognition decision can be further improved. Here a gating module r is used to achieve the adaptive pooling goal:

$$X = [V^{fuse}, A^{fuse}] \tag{6}$$

$$r(X) = \sigma_{sigmoid}(W_2 \sigma_{ReLU}(W_1 X)) \tag{7}$$

$$z = \sum_{i}^{m+n} X_i * r_i(X) \tag{8}$$

Where z is the final bi-modal video embedding.

3.2. Language-guided Contrastive Learning

End-to-end training of multi-modal video action recognition is non-trivial. Due to the high dimensionality of the parameterization in multi-modal network and the lack of large-scale labeled video data, previous multi-modal method cannot be trained well and tend to generate inferior recognition accuracy. In practice, we often observe that the best uni-modal network can even outperforms the multi-modal network.

Inspired by [22], we propose to use a language-guided contrastive learning to relax the limitation of data scarcity and learn directly from the vast amount of video with noisy

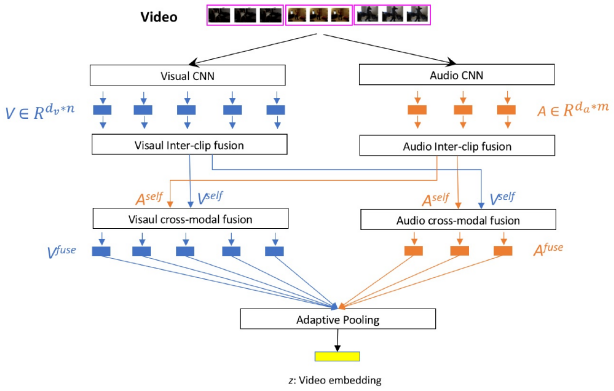

Figure 2. The proposed Bi-modal video encoder

text on the internet. Here, we use Sentence-BERT [23] , a modification of the pretrained BERT [6] network that use siamese and triplet network structures to derive semantically meaningful sentence embeddings, as the text encoder to generate text embedding. For videos with various text information (e.g., labels, titles, description, caption, and etc.), we encode the corresponding text into a sentence vector $t \in R^{d_t}$, where d_t is the dimension of sentence vector. Given a randomly sample a mini-batch of N (video, text) pairs, the proposed CAV-Net is trained to predict which of the $N \times N$ possible (image, text) pairings across a batch actually occurred. To do this, CAV-Net learns a visual-semantic embedding space by jointly training an video encoder and text encoder to maximize the cosine similarity of the video and text embeddings of the N real pairs in the batch. We do not sample negative examples explicitly. Instead, we treat the other $2(N-1)$ examples within a mini-batch as negative examples.

To be more specific, let $sim(z, t) = \frac{z^T t}{\|z\|\|t\|}$ denote the dot product between L_2 normalized video embedding z and sentence embedding t. Then the loss function for a positive pair of examples (i, j) is defined as

$$L_{i,j} = -log(\frac{exp(\frac{sim(z_i, t_j)}{\tau})}{\sum_{k=1}^{2N} 1_{[i \neq k]} exp(\frac{sim(z_i, t_k)}{\tau})}) \tag{9}$$

where $1_{[i \neq k]} \in \{0, 1\}$ is an indicator function evaluating to 1 iff $i \neq k$, and τ denotes a temperature parameter. The final loss is computed across all positive pairs, both (i, j) and (j, i), in a mini-batch.

In testing stage, $sim(.)$ function can be used to link the test video with related action category (e.g., comparing the similarity between video embedding and action label) to get the video action prediction result.

Method	Accuracy(%) Top-1	Top-5
Audio Only		
VGGish [14]	10.5	23.1
Visual Only		
I3D [3]	71.1	90.3
MF-Net [4]	72.8	90.4
R(2+1)D [28]	72.0	90.0
TSM [18]	74.7	N/A
SlowFast [9] 4×16	75.6	92.1
Nonlocal [33]	76.5	92.6
X3D [8]	77.5	92.9
CAV-Net (visual only)	**78.1**	**93.3**
Multi-modal		
Two-Stream (Audio-visual) I3D [25]	68.5	87.3
AVSlowFast 4×16 [32]	77.0	92.7
CAV-Net	79.2	94.4
CAV-Net + pretrain	**80.7**	**95.2**

Table 1. Accuracy comparison of different methods

4. Experiments

4.1. Datasets

To compare to previous research, we use Kinetics-400 as the benchmark dataset. Kinetics-400 dataset contains 400 human action classes, with at least 400 video clips for each action. Each clip is approximately 10 seconds long and is taken from a different YouTube video.

4.2. Multi-Modal Action Recognition Accuracy

In this subsection, we study the effectiveness of the proposed model on learning multi-modal classifier on different datasets. Our CAV-Net can be used with any existing clip-based action classifier and immediately boost the recognition accuracy. Here we use the classical 3D-ResNext [13] [34] as the backbone of visual network, and ResNet [14] with 50 layers as the audio network.

Table 1 shows the action recognition results of uni-modal and multi-modal methods. For audio only method, as expected that since audio alone cannot handle the video recognition task, it thus has the lowest top-1 accuracy. On the other hand, the best visual only model (e.g., X3D [8]) can achieve 77.5% accuracy in Kinetics 400.

For multi-modal fusion methods, the popular two-stream network [25] that adopted a late-fusion strategy based on I3D [3] (noted as Two-Stream Audio-visual I3D in Table 1) surprisingly achieves merely 68.5% top-1 accuracy, which has worse performance than its counterpart uni-modal version I3D [3], where it has 71.1% top-1 accuracy. It shows the difficulty of training a multi-modal network for video action recognition, and the importance of a proper fusion process for multi-modalities. AVSlowFast [32] achieves a slightly better accuracy than its counterpart visual-only version SlowFast [9], where they achieve 77.0% and 75.6% top-1 accuracy respectively. The accuracy improvement could come from the adoption of an early fusion strategy in AVSlowFast.

In contrast, our proposed CAV-Net achieves the best accuracy among all different type of methods. As shown in table 1, our CAV-Net (train mainly based on Kinetics dataset) has 79.2% top-1 accuracy, which significantly outperforms all uni-modal and conventional audio-visual fusion methods. It shows the effectiveness of the propose bi-modal fusion strategy in integrating the multi-modal signal into deep network.

To verify the effectiveness of language-guided learning, we also collect videos with noisy text information (such as caption, description, and etc) as the augmented video datasets to pretrain the CAV-Net, including public available video datasets (such as Youtube-8M [1], Charades-ST [12], DiDeMo [15], ActivityNet Captions [17], and etc) and videos crawled from the Internet, where the title will be used as the label. As can be seen that 'CAV-Net + pretrain' has the best accuracy among all the methods, which achieves 80.7% accuracy. It shows that the proposed language-based contrastive learning strategy can bring more generalization to the model, and employ the video data that is traditionally hard to be used for training (e.g., noisy text and sentences as labels).

We are also interested in the effectiveness of the proposed fusion strategy in uni-modality. We thus remove the audio branch in video encoder (i.e., only with inter-clip fusion and no cross-modal fusion) and train the CAV-Net. The result is shown in 'CAV-Net(visual only)'. As can be seen in 1, despite that the uni-modal video encoder is less accurate than the bi-modal counterpart, it still achieves 78.1% accuracy, which outperforms the best uni-modal visual modal (i.e., X3D). It shows the benefit brought by the inter-clip fusion strategy.

5. Conclusion

In theory, a well-optimized multi-modal video action classifier should always match or outperform the best uni-modal classifier. However, as we've shown in the experiment that the best uni-modal network often outperforms the multi-modal network counter-intuitively due to a poor design of multi-modal fusion and training process. In this work, the proposed language-guided multi-modal fusion successfully addresses the above problem. The language-based contrastive learning strategy also largely augment the available video data (though with noisy labels/texts/sentences) to further optimize the classifier. On a large-scale video dataset, the proposed method successfully elevates the accuracy of video action recognition

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Apostol (Paul) Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. In *arXiv:1609.08675*, 2016. 4

[2] J. Arevalo, T. Solorio, M. M. Gmez, and F. A. Gonzlez. Gated multimodal units for information fusion. In *ICLR Workshop*, 2017. 2

[3] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. *CVPR*, pages 4724–4733, 2017. 2, 4

[4] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. Multi-fiber networks for video recognition. In *ECCV*, 2018. 4

[5] 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 *Proceedings of the 28th ACM International Conference on Multimedia*, page 3884–3892, 2020. 2

[6] J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova less. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019. 3

[7] Shawn Hershey et al. Cnn architectures for large-scale audio classification. In *ICASSP*, 2017. 2

[8] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *CVPR*, 2020. 4

[9] C. Feichtenhofer, H. Fan, J. Malik, and K. He. Slowfast networks for video recognition. In *ICCV*, pages 6201–6210, 2019. 2, 4

[10] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, 2016. 2

[11] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*, 2016. 2

[12] J. Gao, C. Sun, Z. Yang, and R. Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, 2017. 4

[13] K. Hara, H. Kataoka, and Y. Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*, pages 6546–6555, 2018. 4

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[15] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, 2017. 4

[16] Jenhao Hsiao, Jiawei Chen, and Chiuman Ho. Gcf-net: Gated clip fusion network for video action recognition. In *ECCV Workshops*, pages 699–713, 2020. 2

[17] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017. 4

[18] Ji Lin, Chuang Gan, and Song Han. Temporal shift module for efficient video understanding. *CoRR*, abs/1811.08383, 2018. 4

[19] L.Wang, Y. Xiong, Z.Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016. 2

[20] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, I. Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, 2020. 2

[21] Z. Qiu, T. Yao, and T. Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *ICCV*, 2017. 2

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 3

[23] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP*, 2019. 3

[24] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014. 2

[25] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *NIPS*, pages 568–576. Curran Associates, Inc., 2014. 4

[26] Graham W. Taylor, Rob Fergus, Yann Lecun, and Christoph Bregler. Convolutional learning of spatio-temporal features. In *ECCV*, 2010. 2

[27] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 2

[28] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018. 2, 4

[29] G. Varol, I. Laptev, and C. Schmid. Long-term temporal convolutions for action recognition. In *IEEE PAMI*, 2018. 2

[30] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *CVPR*, 2020. 1

[31] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. Attention network for image and sentence matching. In *CVPR*, pages 10941–10950, 2020. 2

[32] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. In *arXiv preprint arXiv:2001.08740*, 2020. 4

[33] Wang Xiaolong, Girshick Ross, Gupta Abhinav, and He Kaiming. Non-local neural networks. In *CVPR*, June 2018. 2, 4

[34] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *CVPR*, pages 5987–5995, 2017. 4

[35] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *CVPR*, pages 8746–8755, 2020. 2