DISTRIBUTED AUDIO-VISUAL PARSING BASED ON MULTIMODAL TRANSFORMER AND DEEP JOINT SOURCE CHANNEL CODING

Penghong Wang[⋆] *Jiahui Li*[††] *Mengyao Ma*[††] *Xiaopeng Fan*[⋆,†]

[⋆] School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China
[††] Wireless Technology Lab, Huawei, Shenzhen 518129, China
[†] PengCheng Lab, Shenzhen 518055, China

ABSTRACT

Audio-visual parsing (AVP) is a newly emerged multimodal perception task, which detects and classifies audio-visual events in video. However, most existing AVP networks only use a simple attention mechanism to guide audio-visual multimodal events, and are implemented in a single end. This makes it unable to effectively capture the relationship between audio-visual events, and is not suitable for implementation in the network transmission scenario. In this paper, we focus on these problems and propose a distributed audio-visual parsing network (DAVPNet) based on multimodal transformer and deep joint source channel coding (DJSCC). Multimodal transformers are used to enhance the attention calculation between audio-visual events, and DJSCC is used to apply DAVP tasks to network transmission scenarios. Finally, the Look, Listen, and Parse (LLP) dataset is used to test the algorithm performance, and the experimental results show that the DAVPNet has superior parsing performance.

***Index Terms*—** distributed audio-visual parsing network, multimodal transformer, deep joint source channel coding

1. INTRODUCTION

Recently, multimodal tasks have been a hot issue in computer vision, including cross-modal generation [1, 2], sound source localization [3, 4] , audio-visual representation [5, 6] and action recognition [7, 8]. As a common audio-visual representation task, audio-visual parsing (AVP) [5] task is to detect and analyze audible, visible or audio-visual segments in the video, and record their start and end times, as shown in Fig. 1. Most existing AVP networks use a simple attention mechanism to guide audio-visual multimodal events, which leads to the inaccuracy of parsing performance. Moreover, the acquisition of audio and video may come from different terminals. In the existing work, the parsing task is implemented in a single terminal, which makes it not suitable for distributed transmission scenarios (DTS).

DTS refers to that different terminals collecting audio data and visual data preprocess the data in a distributed manner to obtain intermediate features and then transmit these features to the far-end data fusion center to perform further processing. These terminals should encode or compress the source data to reduce the consumed bandwidth. To improve the parsing performance in DTS, we draw lessons from the existing excellent work. Such as transformer [9] provides a powerful attention mechanism and has been successfully applied to many tasks of computer vision [10, 11], and deep joint source channel coding (DJSCC) [12] has been applied to wireless image transmission [13], multi task learning [14, 15], etc. In addition, Slepian-Wolf theorem [16] shows that correlated sources can be efficiently compressed separately, and it has been applied to distributed video coding [17] and joint source channel coding [18]. It shows that the distributed coding transmission of audio and video sources is feasible. Therefore, we design a multimodal transformer model to improve the ability to capture the relationship between different modal data, and propose a distributed deep joint source channel coding (DJSCC) method to support distributed AVP task in distributed transmission scenarios. The main contributions of this paper include:

1) The proposed multimodal transformer model provides a powerful ability of self attention and cross attention for DAVP tasks.

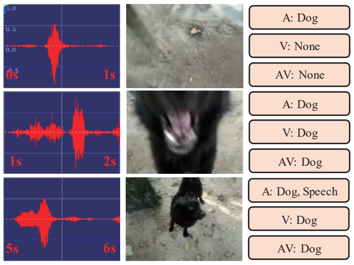

Fig. 1. The task of AVP is to identify audio, video and audio-visual events, and determine their categories and boundaries. For instance, barking occurs from 0 to 2^{nd} second and 5^{th} to 6^{th} second, and the dog is visible from 1^{st} to 2^{nd} second and 5^{th} to 6^{th} second. Therefore, the audio-visual parsing of Dog event category is: audio event (0s-2s, 5s-6s), visual event (1s-2s, 5s-6s), and audio-visual event (1s-2s, 5s-6s).

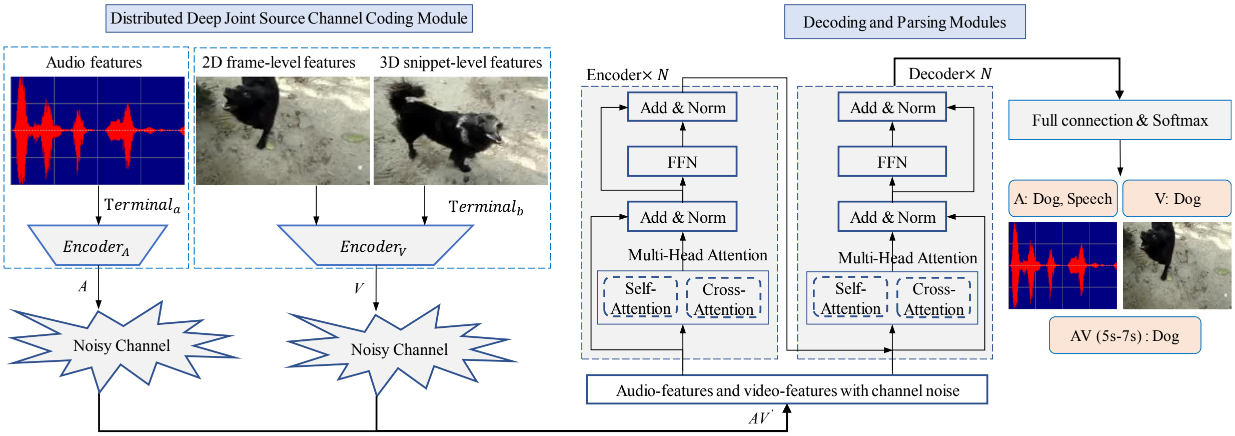

Fig. 2. Overview of distributed DAVPNet based on multimodal transformer and distributed DJSCC. A: compressed audio features, V: compressed visual features, $AV^{'}$: noisy compressed audio-visual features received by decoder.

2) The design of distributed DJSCC provides powerful compression capability, makes the AVP task suitable for DTS, and reduces network parameters.

3) The experimental results show that our model has better parsing performance in both noisy and noiseless channels.

2. PROPOSED METHOD

A distributed audio-visual parsing network (DAVPNet) is proposed, and the network framework is shown in Fig. 2. Specifically, DAVPNet contains a distributed DJSCC module and a decoding parsing module. The audio-visual source is encoded into compact audio A and video V features respectively. After transmission through the noise channel, these noisy compressed features $AV^{'}$ are jointly decoded and parsed at the receiving terminal.

2.1. Distributed Deep Joint Source Channel Coding Module

As shown in Fig. 3, distributed DJSCC contains two encoders: $Encoder_a$ and $Encoder_v$. Each encoder needs to carry out feature extraction and feature compression on the source data independently. In $Encoder_a$, VGGish model is used to extract audio features, and then we use a one-layer full connection to compress the features. In $Encoder_v$, ResNet152 and ResNet (2+1)D are used to extract visual features (including 2D frame-level and 3D snippet-level features). In order to improve the transmission efficiency of the network, we further compress the extracted features. Firstly, 2D and 3D features are compressed by one-layer full connection. Then, pool the temporal features in 2D with the average pooling layer to obtain feature vectors of the same size (T × $\frac{X_2}{2}$), and connect the 2D features with the 3D features. Finally, the features are compressed to the same dimension of audio features with one-layer full connection.

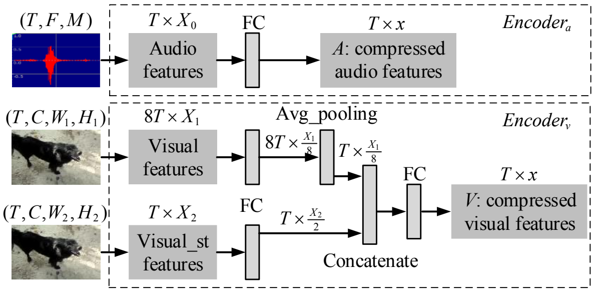

Fig. 3. Design of encoder. Where, *T, C, W* and *H* refer to time, channel, width and height respectively. *X* represents the feature size, and *x* denotes the compressed feature size. *F* and *M* represent the data dimension of audio when using VGGish model.

2.2. Decoding and Parsing Modules

The multimodal transformer is regarded as a decoder that processes the received audio-visual noisy features. The highlight is the design of audio-visual cross attention mechanism in the process of encoding and decoding. Fig. 4 shows the attention mechanism of multimodal fusion in the process of encoding and decoding. In the encoding stage, self-attention and cross-attention are applied to single-mode and audio-visual mode data respectively. In the decoding stage, the input data includes audio feature vector, visual feature vector and fusion attention weight output by the encoder. Therefore, we use the audio and video feature vectors as the query respectively for cross attention operation.

In addition, the remaining network structures in the encoding and decoding blocks are the same as the classical transformer structure [9], except that the number of blocks is different (N=2 in this paper). This avoids the loss of attention weight caused by too many decoder blockers. Finally, we use one-layer full connection and Softmax to detect and classify

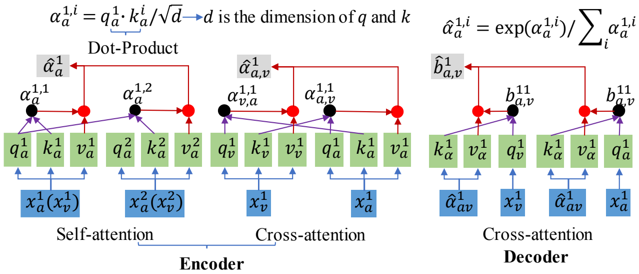

Fig. 4. Multimodal attentional mechanism. Where, q^i, k^i and v^i denote the *query, key* and *value* vector of the i^{th} time sequence, x_a^i and x_v^i denote the audio and visual features of the i^{th} time sequence, $\alpha_{av}^{i,j}$ and $b_{av}^{i,i}$ represent the attention weight obtained from the dot product of k and q, $\hat{\alpha}_{av}^1$ and $\hat{b}_{av}^1$ denote the attention matrix output by the encoder and decoder.Black and red spots represent Dot-Product operations.

(25 categories) the weight matrix of transformer output.

2.3. Noisy Channel Model

The intermediate features obtained by the encoder need to be transmitted to the far-end data fusion center to perform further processing. In this paper, we employ an additive white Gaussian noise (AWGN) model to simulate the noisy channel. The channel signal-to-noise ratio (SNR) is defined as:

$$SNR = 10 \log_{10} \frac{P}{\delta^2} \ (dB) \tag{1}$$

where, P represents the average power of the coded signal (channel input signal).

2.4. Training Strategy

When training DAVPNet, the loss function $\mathcal{L}$ is the cross-entropy between the distribution of prediction class and target class. And it is employed to train the DAVPNet. To fully train the network, we use a two-step training strategy. Firstly, the decoding and parsing module is trained end-to-end to test the performance gain of multimode transformer (Where, the data input dimension is the same as MMIL [5]). Then, the whole DAVPNet is trained under different noise powers to make it robust to channel noise. The loss function is defined as follows:

$$\mathcal{L} = CE(\overline{P}_a,\ \overline{y}_a) + CE(\overline{P}_v,\ \overline{y}_v) + CE(\overline{P}_{a,v},\ \overline{y}_{a,v}) \tag{2}$$

where $\overline{P}_a$, $\overline{P}_v$ and $\overline{P}_{a,v}$ refer to the estimated probabilities of audio, visual and audio-visual events respectively, $\overline{y}_a$, $\overline{y}_v$ and $\overline{y}_{a,v}$ are ground truth labels.

3. EXPERIMENTS AND ANALYSIS

3.1. Dataset

LLP dataset [5] is used for performance testing, it contains 11,849 video clips (each video is 10s long) with video level event annotations. The dataset contains 25 categories such as human speaking, dog barking, human and animal activities, transportation, performance. In our experiment, the whole data set is divided into three parts: training set, verification set and testing set. Specifically, training set contains 10,000 videos with weak labels, verification set and testing set have 649 videos and 1200 videos with fully annotated labels, respectively. The verification and test videos[1] include 4131 audio events, 2495 video events and 2488 audio-visual events.

3.2. Evaluation Metrics

To test and compare the performance of the algorithms fairly, we adopt the same evaluation criteria as in [5]. We evaluate the accuracy of all event types under segment-level and event-level criteria separately, where the segment-level criteria indicators are used to evaluate the performance of video-snippet event labeling in the sample, and the event-level criteria is used to evaluate the performance of event labeling in the sample. Meanwhile, we keep Type@AV and Event@AV[2] metrics to evaluate the overall audio-visual scene analysis performance of the algorithm, where Type@AV is the mean of the results of audio, visual, and audio-visual event, and Event@AV computes the F-score considering each the audio and visual events in sample.

3.3. Experimental Results

To evaluate the effectiveness of the DAVPNet, we compare the current popular algorithms: AVE [3], ABSDN [19] and MMIL [5]. Table 1 presents the experimental results with noiseless condition. That is, the performance gain of multimode transformer. Among these algorithms, DAVPNet achieves the best results except that it is inferior to MMIL in audio metrics. And the MMIL algorithm has obtained the suboptimal performance. Moreover, the total number of parameters in the MMIL network model is 9.31×10^6, DAVPNet model is 8.39×10^6.

In addition, we test the parsing performance of DAVPNet under different noise powers, and compare it with MMIL algorithm. The parameters of the compressed model are only 0.87×10^6, which is nearly 10.7× smaller than the MMIL model. Table 2 shows the experimental results of DAVPNet and MMIL algorithm under different noise powers. When

[1] Notice: During the experiment, we did not obtain the original dataset, only the audio and visual features of videos in the LLP dataset [5]. And these audio, 2D frame-level, and 3D snippet-level features are extracted from the existing general models, VGGish, ResNet152, and ResNet (2+1)D, respectively.

[2] Both Type@AV and Event@AV are named from literature [5]

Table 1. Video parsing accuracy (%) of different algorithms with noiseless condition.

Criteria	Methods	Audio	Visual	Audio-Visual	Type@AV	Event@AV
Segment-level	AVE [3]	47.2	37.1	35.4	39.9	41.6
	ABSDN [19]	47.8	52.0	37.1	45.7	50.8
	MMIL [5]	**60.1**	52.9	48.9	54.0	55.4
	DAVPNet	59.8	**55.7**	**50.0**	**55.2**	**56.8**
Event-level	AVE [3]	40.4	34.7	31.6	35.5	36.5
	ABSDN [19]	34.1	46.3	26.5	35.6	37.7
	MMIL [5]	**51.3**	48.9	43.0	47.7	48.0
	DAVPNet	50.0	**52.2**	**43.4**	**48.6**	**48.9**

Table 2. Video parsing accuracy (%) of the algorithm under different noise powers. The $+\infty$ indicates transmission under noiseless channel.

Criteria	SNR (dB)	Methods	Audio	Visual	Audio-Visual	Type@AV	Event@AV
Segment-level	$+\infty$	MMIL [5]	60.1	52.9	48.9	54.0	55.4
		DAVPNet	**60.2**	**54.8**	**49.8**	**54.9**	**56.6**
	15	MMIL [5]	**59.3**	52.8	48.1	53.4	54.8
		DAVPNet	58.9	**55.0**	**49.8**	**54.5**	**55.8**
	5	MMIL [5]	59.2	52.2	48.0	53.1	54.9
		DAVPNet	**59.5**	**53.2**	**48.3**	**53.7**	**55.6**
Event-level	$+\infty$	MMIL [5]	51.3	48.9	43.0	47.7	48.0
		DAVPNet	**51.5**	**51.4**	**43.5**	**48.8**	**49.6**
	15	MMIL [5]	50.8	48.5	41.5	46.7	48.0
		DAVPNet	**51.0**	**51.0**	**43.5**	**48.5**	**48.7**
	5	MMIL [5]	50.0	48.0	41.1	46.4	47.2
		DAVPNet	**50.6**	**49.5**	**41.5**	**47.2**	**48.3**

Table 3. Compression ratio. Different lines represent audio, 2D frame level, and 3D snippet level features respectively.

Source Data Size	Feature Size	Encoder Output	Compression Ratio	
(T,C,W,H)	(T,X)	(T,x)	Source	Feature
(10,96,64)	(10,512)	(10,64)	96×	8×
(10,3,224,224) (10,3,112,112)	(80,2048) (10,512)	(10,64)	**2940×**	**264×**

SNR is equal to $+\infty$ or 5 dB, DAVPNet achieves the best results in all evaluation metrics. When SNR is equal to 15 dB, the audio parsing performance (Segment-level) of our method is slightly inferior to that of MMIL algorithm, and other indexes are significantly better than it. The experimental results show that DAVPNet achieves the optimal performance under different noise powers and shows strong robustness. In addition, compared to the noiseless condition, the accuracy of DAVPNet is reduced by 1-2%.

Table 3 shows the compression ratio of the data in this experiment. Feature size is compressed to (10, 64), in which the audio is independent coding, and the 2D frame level, and 3D snippet level features are joint coding. Compared with source data and feature data, our encoder achieves 96× and 8× the compression rate for audio features, 2940× and 264× for video features.

4. CONCLUSION

In this work, we propose a distributed audio-visual parsing network (DAVPNet). Specifically, DAVPNet contains a distributed DJSCC module and a decoding parsing module. The purpose of distributed DJSCC is to compress visual and audio features and improve transmission efficiency. The decoding and parsing module designs a multimodal transformer model to improve the performance of audio-visual parsing. Then, the LLP dataset is employed for model training and performance testing. The performance of DAVPNet under compressed transmission and uncompressed transmission is tested respectively. Moreover, the compression ratio of the features and the parameters of the model are calculated. The experimental results show that DAVPNet is significantly superior to the state of the art in terms of video parsing performance and model robustness.

5. ACKNOWLEDGMENT

This work was supported in part by the National Key Research and Development Program of China (2021YFF0900500), and the National Science Foundation of China (NSFC) under grants 61972115 and 61872116.

6. REFERENCES

[1] Peihao Chen, Yang Zhang, Mingkui Tan, Hongdong Xiao, Deng Huang, and Chuang Gan, "Generating visually aligned sound from videos," *IEEE Transactions on Image Processing*, vol. 29, pp. 8292–8302, 2020.

[2] Moitreya Chatterjee and Anoop Cherian, "Sound2sight: Generating visual dynamics from sound and context," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 701–719.

[3] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu, "Audio-visual event localization in unconstrained videos," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 247–263.

[4] Janani Ramaswamy, "What makes the sound?: A dual-modality interacting network for audio-visual event localization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 4372–4376.

[5] Yapeng Tian, Dingzeyu Li, and Chenliang Xu, "Unified multisensory perception: Weakly-supervised audio-visual video parsing," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 436–454.

[6] Mathew Monfort, SouYoung Jin, Alexander Liu, David Harwath, Rogerio Feris, James Glass, and Aude Oliva, "Spoken moments: Learning joint audio-visual representations from video descriptions," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 14871–14881.

[7] Arsha Nagrani, Chen Sun, David Ross, Rahul Sukthankar, Cordelia Schmid, and Andrew Zisserman, "Speech2action: Cross-modal supervision for action recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10317–10326.

[8] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen, "Epic-fusion: Audio-visual temporal binding for egocentric action recognition," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 5492–5501.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (ANIPS)*, 2017, pp. 5998–6008.

[10] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 213–229.

[11] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia, "End-to-end video instance segmentation with transformers," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 8741–8750.

[12] Fan Zhai, Yiftach Eisenberg, and Aggelos K Katsaggelos, "Joint source-channel coding for video communications," *Handbook of Image and Video Processing*, pp. 1065–1082, 2005.

[13] Eirina Bourtsoulatze, David Burth Kurka, and Deniz Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 3, pp. 567–579, 2019.

[14] Zhicong Zhang, Mengyang Wang, Mengyao Ma, Jiahui Li, and Xiaopeng Fan, "Msfc: Deep feature compression in multi-task network," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2021, pp. 1–6.

[15] Mengyang Wang, Zhicong Zhang, Jiahui Li, Mengyao Ma, and Xiaopeng Fan, "Deep joint source-channel coding for multi-task network," *IEEE Signal Processing Letters*, pp. 1–1, 2021, doi:10.1109/LSP.2021.3113827.

[16] David Slepian and Jack Wolf, "Noiseless coding of correlated information sources," *IEEE Transactions on Information Theory*, vol. 19, no. 4, pp. 471–480, 1973.

[17] B. Girod, A.M. Aaron, S. Rane, and D. Rebollo-Monedero, "Distributed video coding," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 71–83, 2005.

[18] Qian Xu and Zixiang Xiong, "Layered wyner–ziv video coding," *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3791–3803, 2006.

[19] Yan-Bo Lin, Yu-Jhe Li, and Yu-Chiang Frank Wang, "Dual-modality seq2seq network for audio-visual event localization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2002–2006.