Enhancing Multimodal Cooperation via Sample-level Modality Valuation

Supplementary Material

1. Proof of Remark 2

Remark 2. *Suppose the marginal contribution of modality is non-negative and the numerical benefits of one modality's marginal contribution follow the discrete uniform distribution. Enhancing the discriminative ability of low-contributing modality i can increase its contribution ϕ^i.*

Proof. Denote after discriminative ability enhancement, the benefits that only having x^i as model input is $v'(\{x^i\})$. And the contribution after enhancement is ϕ'^i. The marginal contribution of x^i after enhancement with respect to a permutation π is denoted by $\Delta'_\pi(x^i)$.

When x^i is the first one in the permutation, based on the definition, $v(\{x^i\})$ is the marginal contribution of x^i. Based on the definition of function v, benefits reflects the discriminative ability, then the benefits that only having x^i as model input would accordingly increase after enhancing its discriminative ability:

$$v'(\{x^i\}) - v(\{x^i\}) > 0. \tag{1}$$

When x^i is not the first one in the permutation, for a specific permutation π, suppose the marginal contribution of modality is non-negative, since the introduction of additional modality tends to not bring negative effects in practice. Suppose $S_\pi(x^i)$ has $c-1$ modalities. Then, based on the definition of function v, the marginal contribution of x^i for permutation π, $\Delta_\pi(x^i)$, can be 0 ($v(S_\pi(x^i) \cup x^i) = v(S_\pi(x^i))$), 1 ($v(S_\pi(x^i) \cup x^i) = c, v(S_\pi(x^i)) = c-1$), and c ($v(S_\pi(x^i) \cup x^i) = c, v(S_\pi(x^i)) = 0$). After modality enhancement, $\Delta'_\pi(x^i)$ also have these possible value.

Suppose the numerical value of one modality's marginal contribution follows the discrete uniform distribution, then $\Delta'_\pi(x^i) - \Delta_\pi(x^i)$ have following cases with equal probability:

(1) $\Delta'_\pi(x^i) = \Delta_\pi(x^i) = 0$. $\Delta'_\pi(x^i) - \Delta_\pi(x^i) = 0$.

(2) $\Delta'_\pi(x^i) = 0, \Delta_\pi(x^i) = 1$. $\Delta'_\pi(x^i) - \Delta_\pi(x^i) = -1$.

(3) $\Delta'_\pi(x^i) = 0, \Delta_\pi(x^i) = c$. $\Delta'_\pi(x^i) - \Delta_\pi(x^i) = -c$.

(4) $\Delta'_\pi(x^i) = \Delta_\pi(x^i) = 1$. $\Delta'_\pi(x^i) - \Delta_\pi(x^i) = 0$.

(5) $\Delta'_\pi(x^i) = 1, \Delta_\pi(x^i) = 0$. $\Delta'_\pi(x^i) - \Delta_\pi(x^i) = 1$.

(6) $\Delta'_\pi(x^i) = 1, \Delta_\pi(x^i) = c$. $\Delta'_\pi(x^i) - \Delta_\pi(x^i) = 1 - c$.

(7) $\Delta'_\pi(x^i) = \Delta_\pi(x^i) = c$. $\Delta'_\pi(x^i) - \Delta_\pi(x^i) = 0$.

(8) $\Delta'_\pi(x^i) = c, \Delta_\pi(x^i) = 0$. $\Delta'_\pi(x^i) - \Delta_\pi(x^i) = c$.

(9) $\Delta'_\pi(x^i) = c, \Delta_\pi(x^i) = 1$. $\Delta'_\pi(x^i) - \Delta_\pi(x^i) = c-1$.

Based on the assumption, all cases are with equal probability, then, for a specific permutation π (except x^i is the first one):

$$\mathbb{E}(\Delta'_\pi(x^i) - \Delta_\pi(x^i)) = 0. \tag{2}$$

Combined Equation 1 and Equation 2, we have,

$$\mathbb{E}(\phi'^i - \phi^i) = \mathbb{E}(\frac{1}{n!}\sum_{\pi\in\Pi_N}\Delta'_\pi(x^i) - \frac{1}{n!}\sum_{\pi\in\Pi_N}\Delta_\pi(x^i)), \tag{3}$$

$$\mathbb{E}(\phi'^i - \phi^i) = \mathbb{E}(\frac{1}{n!}\sum_{\pi\in\Pi_N}(\Delta'_\pi(x^i) - \Delta_\pi(x^i))), \tag{4}$$

$$\mathbb{E}(\phi'^i - \phi^i) = \mathbb{E}(\underbrace{\frac{(n-1)!}{n!}(v'(\{x^i\}) - v(\{x^i\}))}_{\text{only cases } x^i \text{ is the first one, which have } (n-1)! \text{ permutations}}), \tag{5}$$

$$\mathbb{E}(\phi'^i - \phi^i) > 0. \tag{6}$$

Then, we can have enhancing the discriminative ability of low-contributing modality i can increase its contribution in the multimodal learning.

□

2. Experimental settings

When not specified, ResNet-18 [5] is used as the backbone in experiments. Concretely, for the visual encoder, we take multiple frames as the input, and feed them into the 2D network like [15] does; for the audio encoder, we modified the input channel of ResNet-18 from three to one like [1] does and the rest parts remain unchanged; for the optical flow encoder, we stack the horizontal vector u and vertical vector v in the way of $[u, v]$ to form as one frame, then multiple frames are also put into the ResNet-18 as [15] does; for the text data, the pre-trained BERT [2] is used to extract embeddings. Encoders used for UCF-101 are ImageNet pre-trained. Encoders of other datasets are trained from scratch.

Videos are extracted frames with 1fps and three frames are uniformly sampled as the visual input. Three optical flow frames are also uniformly sampled in each video. Audio of Kinetics Sounds is converted to a 128×1024 spectrogram with 128-dim log-mel filterbank and a 25ms Hamming window. The audio data of MM-Debiased dataset is transformed into a spectrogram with size $257 \times 1,003$ using a window with length of 512 and overlap of 353.

Method	MELD	UCF-101-Three	CMU-MOSI
Concatenation	63.56	82.29	75.07
Decision fusion	60.84	82.18	74.78
OGM-GE (our extension) [9]	62.84	82.43	75.80
G-Blending [12]	63.64	82.67	76.16
Greedy [13]	Inapplicable	Inapplicable	Inapplicable
PMR (our extension) [3]	63.58	82.32	76.28
AGM [7]	63.45	Not converge	76.08
Our-Sample-level	**63.95**	**82.90**	**76.82**
Our-Modality-level	63.91	82.76	76.53

Table 1. Accuracy of multimodal models on the three-modality dataset.

During training, we use SGD with momentum (0.9) and set the learning rate at $1e-3$. All models are trained on 2 NVIDIA RTX 3090 (Ti). In experiments, we randomly split a subset with 20% training samples for the average uni-modal contribution estimation in the modality-level method. During modality valuation, for input modality set C, input of modalities not in C are zeroed out, similar to related work [4]. During testing, all modalities are taken as the model input.

3. Construction of MM-Debiased dataset

To evaluate our proposed methods on the dataset with less modality preference of low-contributing phenomenon, we construct the MM-Debiased dataset. We first train uni-modal ResNet-18 model on the audio and visual modality of VGG-Sound [1] and Kinetics-400 dataset [6]. During training, we record the mean uni-modal softmax scores of each training and testing sample, which reflects the confidence for the sample [8]. Then, we select the training samples of 10 classes that have a close summation of uni-modal softmax scores on both modalities from the two datasets. The testing samples are then selected from the 10 classes with the same strategy. The selected 10 classes are *playing piano*, *playing cello*, *lawn mowing*, *singing*, *cleaning floor*, *bowling*, *whistling*, *motorcycling*, *playing flute* and *writing on blackboard*. Based on the result in Figure 4a of the manuscript, the average contribution of each modality over all training samples during training on MM-Debiased is apparently closer than Kinetics Sounds and UCF-101.

4. Experiments of more-than-two modalities

To further validate the effectiveness of our methods under scenarios with more modalities, we further conduct experiments on MLED [10], UCF-101-Three and CMU-MOSI [14] dataset. The UCF-101-Three dataset introduces the additional RGB-Difference modality based on the UCF-101 dataset. Results are shown in Table 1. Many existing imbalanced multimodal learning methods do not consider the 3-modality case. Greedy [13] is not applicable for more-than-two cases. We modify OGM-GE [9] and PMR [3] while keeping the core strategy to extend them. As the results, they only obtain a marginal improvement in more complex 3-modality case. OGM-GE even loses its efficacy on MELD. In comparison, our method obtained a superior improvement. Moreover, we also consider more modalities case on the Caltech101-20 dataset. As Table 2, our methods remain effective even with 5 modalities (views).

Num of modalities	Concat	Our-Sample	Our-Modality
2	82.91	**83.62**	83.47
3	87.71	**88.42**	87.99
4	93.64	**94.07**	93.79
5	94.63	**94.77**	94.73

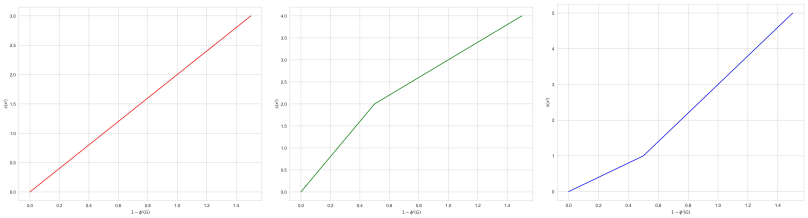
Table 2. Accuracy of our methods on Caltech101-20 dataset.

5. Comparison for audio-visual event localization task

Method	Accuracy
Baseline [11]	71.64
OGM-GE [9]	72.04
Sample-level	**72.39**
Modality-level	72.14

Table 3. Accuracy of comparison for audio-visual event localization task on AVE dataset.

To further evaluate our proposed methods in more general cases, we employ our methods on a representative scene understanding task, audio-visual event localization, which aims to temporally demarcate both audible and visible events from a video. The experiments are conducted on the widely-used AVE dataset [11]. We use the official codes



(a) $f_s^1(\cdot)$ (b) $f_s^2(\cdot)$ (c) $f_s^3(\cdot)$

Figure 1. Different function $f_s(\cdot)$ in sample-level method.

and run them in our environment for fairness. The experiment results are shown in Table 3. Our sample-level and modality-level methods are based on Baseline [11] in this table. According to the experiment results, OGM-GE [9] outperforms the Baseline [11]. Our methods further have improvement and maintain effective on the more challenging audio-visual event localization task on the AVE dataset.

6. Experiments of the scale of split subset in modality-level method

Method	Accuracy	
	Kinetics Sounds	UCF-101
Concatenation	62.30	81.15
Decision fusion	62.65	79.81
20%	66.65	83.46
50%	66.69	83.94
100%	**66.77**	**84.18**

Table 4. Comparison with different scale subset in the modality-level method.

Considering that conducting modality valuation for each sample in the sample-level method would be high additional computational cost when the scale of dataset is quite large, the more efficient modality-level method is proposed, which estimates the average uni-modal contribution via only conducting modality valuation on the subset of training samples. Here we conduct experiments about the scale of split subset. Based on the results in Table 4, we can find that although the accuracy is improved with larger subset, only 20% samples are adequate to achieve considerable performance. These experiments indicate that our modality-level method is efficient and effective.

7. Experiments of different $f_s(\cdot)$ and $f_m(\cdot)$

In our sample-level strategy, the re-sample frequency of modality i for specific sample x is:

$$s(x^i) = \begin{cases} f_s(1-\phi^i) & \phi^i < 1, \\ 0 & \text{others}, \end{cases} \tag{7}$$

Method	Accuracy	
	Kinetics Sounds	UCF-101
Concatenation	62.30	81.15
Decision fusion	62.65	79.81
$f_s^1(\cdot)$	66.92	83.52
$f_s^2(\cdot)$	68.58	83.12
$f_s^2(\cdot)$	67.12	83.68

(a) Sample-level method.

Method	Accuracy	
	Kinetics Sounds	UCF-101
Concatenation	62.30	81.15
Decision fusion	62.65	79.81
$f_m(x) = x$	66.65	83.46
$f_m(x) = \tanh(x)$	66.50	83.86
$f_m(x) = \text{power}(1.5, x)$	65.69	83.25

(b) Modality-level method.

Table 5. Comparison with different function $f(\cdot)$.

where $f_s(\cdot)$ is a monotonically increasing function.

In our modality-level strategy, we randomly split a subset with Z samples in the training set to approximately estimate the average uni-modal contribution. The overall low-contributing modality i can be approximately identified. Then, other modalities remain unchanged, and modality i in sample x is dynamically re-sampled with specific probability during training via:

$$p(i) = f_m(\text{Norm}(d)), \tag{8}$$

where $d = \frac{1}{n-1}(\sum_{j=1,j\neq i}^{n}(\frac{\sum_{k=1}^{Z}\phi_k^j}{Z} - \frac{\sum_{k=1}^{Z}\phi_k^i}{Z}))$. The discrepancy in average contribution between overall low-contributing modality x^i compared to others (*i.e.,* d) is first $0-1$ normalized, then fed into $f_m(\cdot)$, a monotonically increasing function with a value between 0 and 1. n is the number of modalities.

According to Equation 7 and Equation 8, $f_s(\cdot)$ and $f_m(\cdot)$ are not limited to a specific function. In this section, we perform experiments on different $f_s(\cdot)$ and $f_m(\cdot)$ to validate

the effectiveness of our method. The results are shown in Table 5. Based on the results, it does not require much effort to specifically choose $f_s(\cdot)$ and $f_m(\cdot)$. Different $f_s(\cdot)$ and $f_m(\cdot)$ can achieve consistent improvements over the compared baseline across different datasets. Results show that our sample- and modality-level methods are flexible and their effectiveness does not depend on the specific design.

8. Experiments of fixed re-sample rate

Method	UCF-101	
	Acc	Num of re-sampled samples
Concatenation	81.15	-
Decision fusion	79.81	-
Low re-sample rate	83.41	0.96×
High resample rate	82.24	2.68×
Ours-Sample-level	**83.52**	1.00×

Table 6. Comparison with fixed re-sample rate methods on the UCF-101 dataset.

In our methods, the specific re-sample rate is dynamically determined by the exact contribution value during training. Concretely, the low-contributing modality in sample x is re-trained with a re-sample rate inversely proportional to its contribution. To validate the effectiveness of our sample-level method, we perform experiments of fixed re-sample rate on the UCF-101 dataset and the results are shown in Table 6. Two types of fixed re-sample rate methods are compared: low re-sample rate one and high re-sample rate one. Both the performance of low re-sample rate method and high re-sample rate method are inferior to our sample-level method. In addition, compared to our sample-level method, the number of re-sampled samples of low re-sample rate method is close, since the dataset has a long-tail phenomenon that the contribution of low-contributing modality in the majority of samples is not severely low. But the number of re-sampled samples of high re-sample rate method is obviously greater but with worse performance. The reason could be that the high re-sample rate leads to more server over-fitting. These experiments indicate the effectiveness of our dynamic re-sample rate.

9. Experiments of other enhancement strategy

Dataset	Concat	Modality drop	Data mask
KS	62.30	65.85	65.54
UCF-101	81.15	84.15	83.38

Table 7. Accuracy of our valuation with other strategies.

In addition, the resample is one of the simple but effective tools to enhance uni-modal discriminative ability. In fact, other enhancement strategies can also be attempted. In Table 7, we attempt the modality drop and data mask strategies to replace resample. Their effectiveness indicates that our fine-grained modality valuation is universal.

References

[1] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 1, 2

[2] 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. 1

[3] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. Pmr: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20029–20038, 2023. 2

[4] Amirata Ghorbani and James Y Zou. Neuron shapley: Discovering the responsible neurons. *Advances in Neural Information Processing Systems*, 33:5922–5932, 2020. 2

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[6] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 2

[7] Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, and Yi Zhou. Boosting multi-modal model performance with adaptive gradient modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22214–22224, 2023. 2

[8] Lukas Neumann, Andrew Zisserman, and Andrea Vedaldi. Relaxed softmax: Efficient confidence auto-calibration for safe pedestrian detection. 2018. 2

[9] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8238–8247, 2022. 2, 3

[10] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*, 2018. 2

[11] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018. 2, 3

[12] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12695–12705, 2020. 2

[13] Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International Conference on Machine Learning*, pages 24043–24055. PMLR, 2022. 2

[14] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016. 2

[15] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 570–586, 2018. 1