Classifier-guided Gradient Modulation for Enhanced Multimodal Learning

Zirun Guo, Tao Jin, Jingyuan Chen, Zhou Zhao

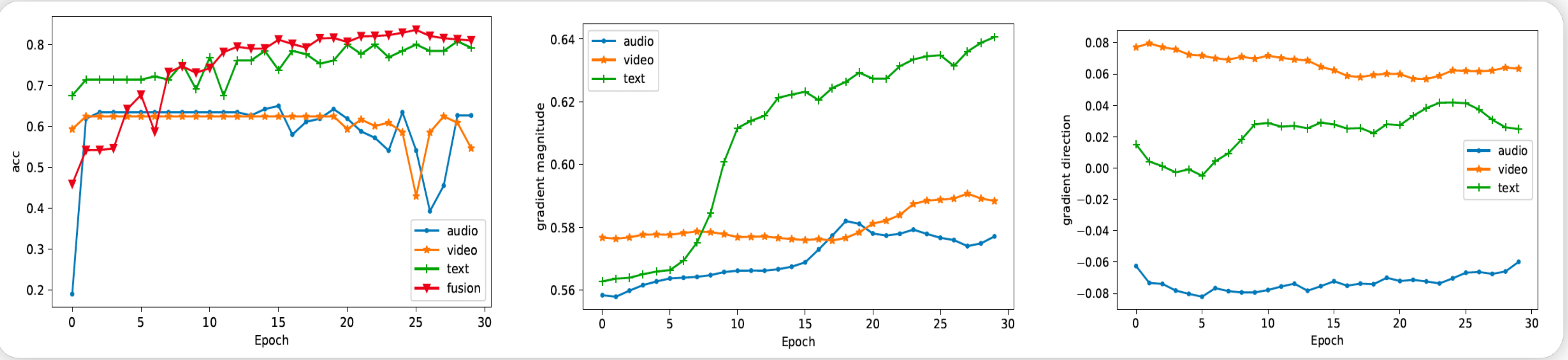
Zhejiang University

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

Challenge in Multimodal Learning: the model tends to rely on only one modality based on which it could learn faster, thus leading to inadequate use of other modalities.

Sometimes, the performance of joint training even worse than that of the unimodal training.



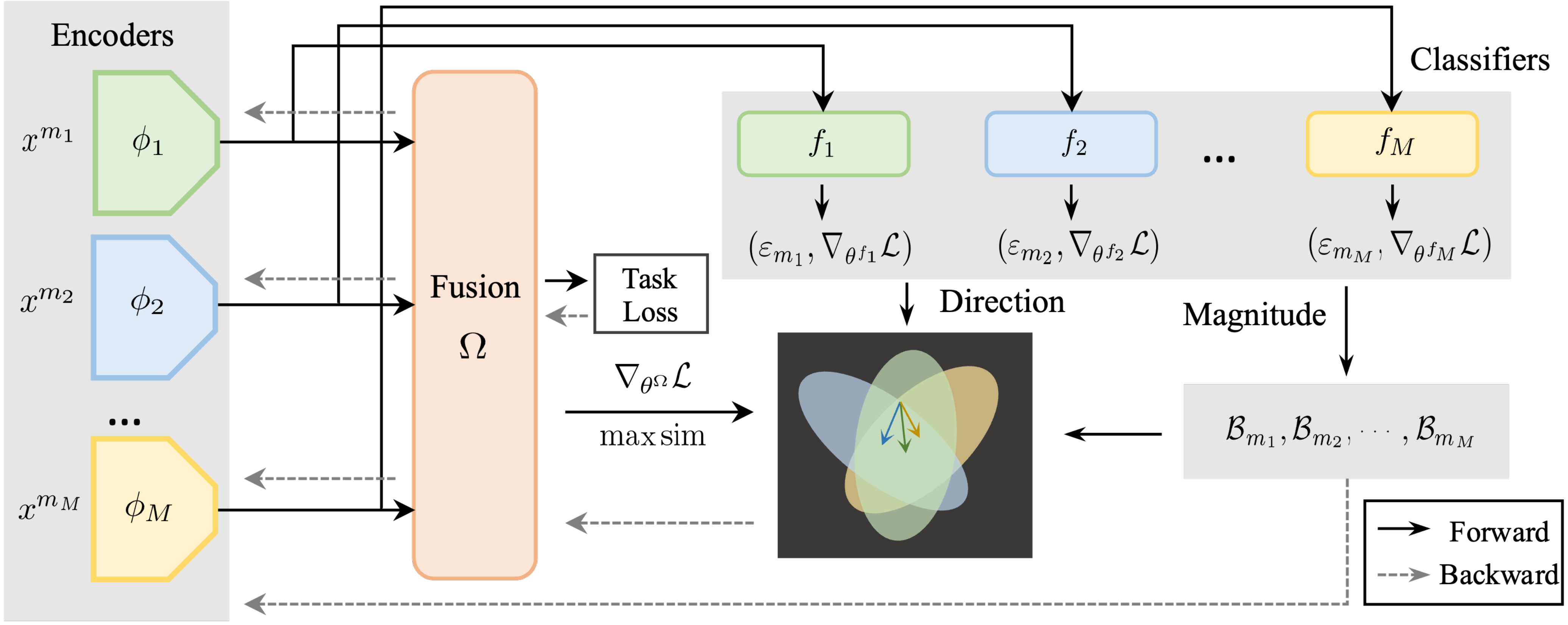
Limitations of Existing Methods

NEURAL INFORMATION PROCESSING SYSTEMS

- The number of modalities
- Task type (loss function)
- Optimizers
- …………
- Insufficient exploration of gradient direction

More general situations

Methodology



Methodology

Gradient Magnitude

(1) Add light classifiers for each modality to make unimodal predictions

(2) Calculate the difference of performance between two consecutive iteration:

$$\begin{aligned}\Delta\boldsymbol{\varepsilon}^{\boldsymbol{t+1}} &= \boldsymbol{\varepsilon}^{\boldsymbol{t+1}} - \boldsymbol{\varepsilon}^{\boldsymbol{t}} = (\Delta\varepsilon_{m_1}^{t+1}, \Delta\varepsilon_{m_2}^{t+1}, \cdots, \Delta\varepsilon_{m_M}^{t+1}) \\ &= (\varepsilon_{m_1}^{t+1} - \varepsilon_{m_1}^{t}, \varepsilon_{m_2}^{t+1} - \varepsilon_{m_2}^{t}, \cdots, \varepsilon_{m_M}^{t+1} - \varepsilon_{m_M}^{t})\end{aligned}$$

(3) Calculate the balancing term for each modality:

$$\mathcal{B}_{m_i}^t = \rho \frac{\sum_{k=1, k\neq i}^{M} \Delta\varepsilon_{m_k}^t}{\sum_{k=1}^{M} \Delta\varepsilon_{m_k}^t}$$

(4) Update the gradient of encoders of each modality:

$$\theta_{t+1}^{\phi_i} = \theta_t^{\phi_i} - \alpha \mathcal{B}_{m_i}^{t+1} \nabla_{\theta^{\phi_i}} \mathcal{L}(\theta_t^{\phi_i})$$

Methodology

Gradient Direction

(1) Calculate the gradient of each modality encoder:

$$\nabla_{\theta^{f_i}} \mathcal{L}(\theta^{f_i}) = \frac{\partial \mathcal{L}(\theta^{f_i})}{\partial f_i} = \left[\frac{\partial \mathcal{L}(\theta^{f_i})}{\partial \theta_1^{f_i}}, \frac{\partial \mathcal{L}(\theta^{f_i})}{\partial \theta_2^{f_i}}, \cdots, \frac{\partial \mathcal{L}(\theta^{f_i})}{\partial \theta_n^{f_i}} \right]^\top$$

(2) Calculate the gradient of the fusion module:

$$\nabla_{\theta^{\mathcal{F}}} \mathcal{L}(\theta^{\mathcal{F}}) = \frac{\partial \mathcal{L}(\theta^{\mathcal{F}})}{\partial \mathcal{F}} = \left[\frac{\partial \mathcal{L}(\theta^{\mathcal{F}})}{\partial \theta_1^{\mathcal{F}}}, \frac{\partial \mathcal{L}(\theta^{\mathcal{F}})}{\partial \theta_2^{\mathcal{F}}}, \cdots, \frac{\partial \mathcal{L}(\theta^{\mathcal{F}})}{\partial \theta_n^{\mathcal{F}}} \right]^\top$$

(3) Enforce the gradient direction of the fusion module as close as possible to the weighted average of the unimodal gradient directions:

$$\max \sum_{i=1}^{M} \mathcal{B}_{m_i}^t \mathrm{sim}\left(\nabla_{\theta^{\mathcal{F}}} \mathcal{L}, \nabla_{\theta^{f_i}} \mathcal{L}\right)$$

Results

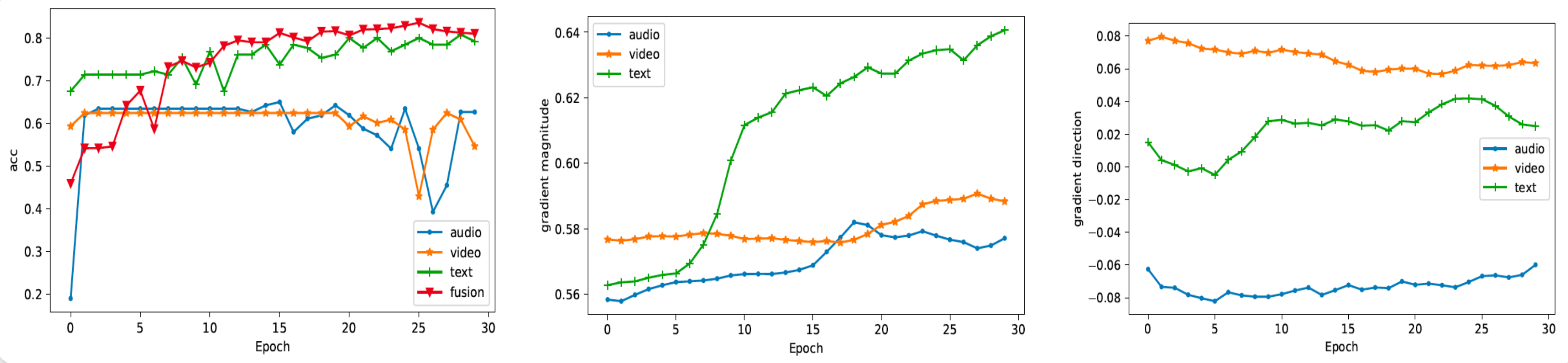
Comparison with SOTA

Dataset	Task type	No. of modality
UPMC-Food 101	Classification	2
CMU-MOSI	Regression	3
IEMOCAP	Classification	3
BraTS 2021	Segmentation	4

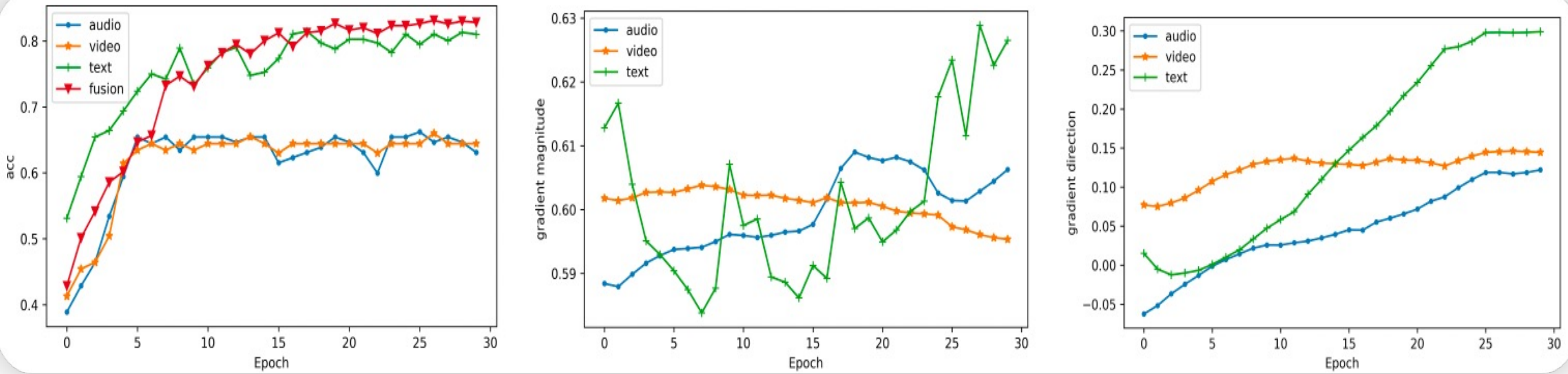
- Consistent improvement on four different multimodal datasets, covering classification, regression and segmentation
- Outperforms other SOTA methods

Results

w/o. CGGM



CGGM



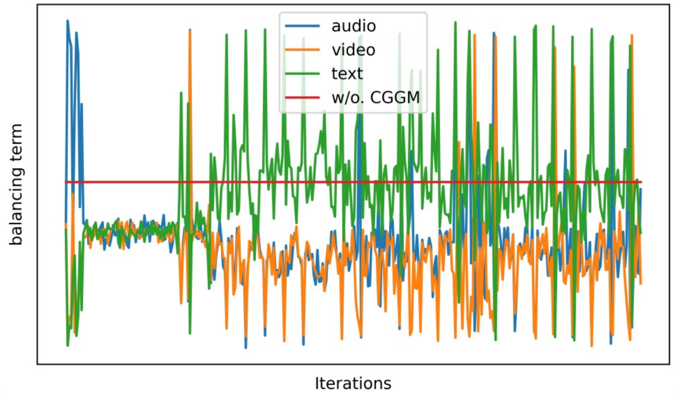
Results

Ablation

Model	Acc	F1
Baseline	70.74	69.53
CGGM ($\rho = 1.0, \lambda = 0$)	72.35	71.56
CGGM ($\rho =$ None, $\lambda = 0.1$)	72.41	72.07
CGGM ($\rho = 1.0, \lambda = 0.1$)	73.74	73.18

Effectiveness of CGGM

Balancing term



Dynamic adjustments during the training process

Additional computational resources

Setting	Food101	MOSI	IEMOCAP	BraTS
With classifiers	+8MB	+8MB	+8MB	+24MB

Low additional gpu memory cost