Interactive Multimodal Learning via Flat Gradient Modification (#1049)

Qing-Yuan Jiang, Zhouyang Chi, & Yang Yang
Nanjing University of Science and Technology, Nanjing, China.
{jiangqy, yyang}@njust.edu.cn, zhouyangchi0@gmail.com

IJCAI 2025 MONTREAL

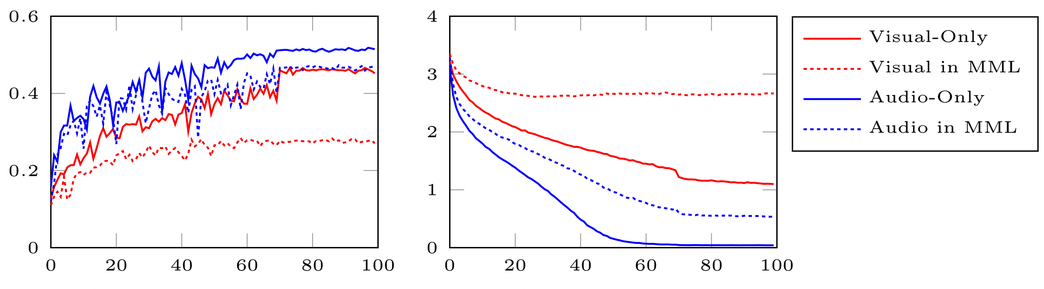
Introduction

Multimodal Learning

- **Goal:** fuse multimodal data to boost model performance.
- **Optimal Scenario:** maximize information extraction from multiple modalities for better performance.

Modality Imbalance

- **Phenomenon:** MML underperforms single-modality models.
- Strong-Weak Modality $\mapsto$ Modality Imbalance



(a). Training Accuracy. (b). Training Loss.

- **Alternating Learning:** a novel MML learning paradigm to enhance the cross-modal interaction.
- **MLA[CVPR'24]:** enhance interaction through orthogonal projection:

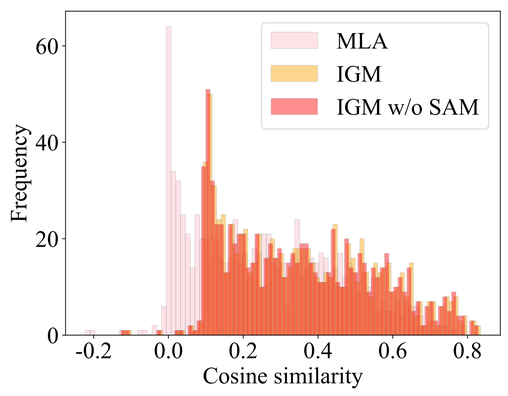
$$\boldsymbol{\omega}_{t+1}^{a} = \boldsymbol{\omega}_{t}^{a} - \eta \cdot \boldsymbol{P}^{v} \cdot \nabla \mathcal{L}$$

Issue in Orthogonal Projection

Poor Plasticity:

- The orthogonal projection suffers from **poor plasticity** problem, i.e., leading to **feasible gradient direction becomes narrow**.
- The poor plasticity problem results in **suboptimal solution**.

Gradient Change:

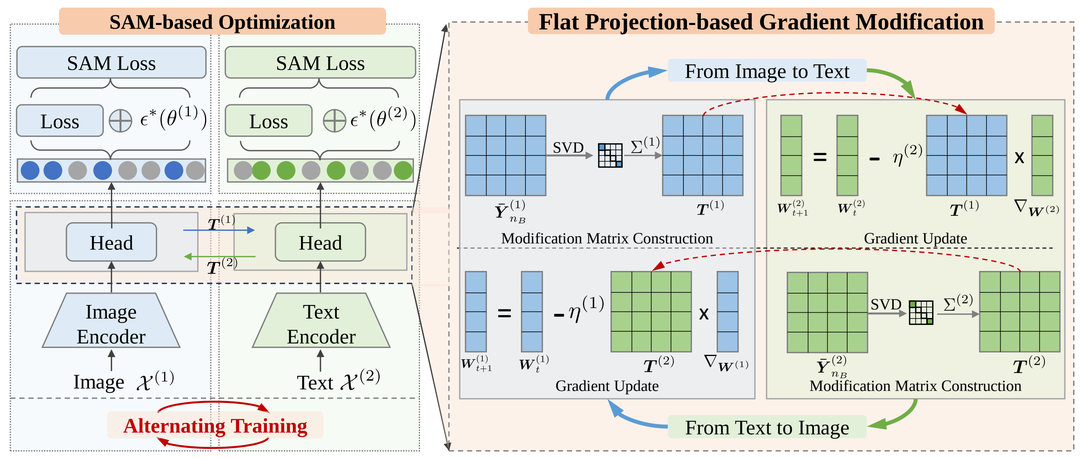


Contributions

- Propose flat projection-based gradient modification strategy.
- Introduce SAM based optimization to smooth objectives.
- IGM can achieve the best performance compared to STOAs.

Methodology

Overall Framework



Two Key Components of IGM:

- **Flat Projection Gradient Modification**: Project updating direction, enhancing the interaction.
- **SAM-based Optimization**: Smooth the objective function, enhancing the flatness.

Flat Projection Gradient Modification:

- ①. **Find Flat Direction**: Decompose variance of activation by SVD: $\mathbf{U}^{v}\Lambda^{v}[\mathbf{V}^{v}]^{\top} = svd(\mathbf{Y}^{v})$.
- ②. **Relationship of Flatness and Singular Value**: The flatness of $\mathbf{v}_i^v$ is determined by singular value λ_i^v.
- ③. **Conduct Projection Matrix**: Large $\lambda^v \mapsto$ More Sharp Loss: $\Sigma^{v} = exp(-\frac{\tau}{\lambda_{\max}^{v} - \lambda_{\min}^{v}}(\Lambda^{v} - \lambda_{\min}^{v}\mathbf{I}))$
- ④. **Project during Updating**: Learning Another Modality: $\mathbf{T}^{v} = \mathbf{U}^{v}\Sigma^{v}[\mathbf{V}^{v}]^{\top}$; $\boldsymbol{\omega}_{t+1}^{a} = \boldsymbol{\omega}_{t}^{a} - \eta \cdot \boldsymbol{T}^{v} \cdot \nabla_{\omega}\mathcal{L}$.

SAM-based Optimization:

- ①. **Perturb the Loss**: $\mathcal{L}^{\text{SAM}}(\omega) = \max_{\epsilon:||\epsilon||\leq\rho} \mathcal{L}(\omega + \epsilon)$.
- ②. **Optimal perturbation**: $\epsilon^{*}(\omega) = \text{argmax}_{||\epsilon||\leq\rho} \mathcal{L}(\omega + \epsilon)$.
- ③. **Gradient of SAM Loss**: $\nabla_{\omega}\mathcal{L}^{\text{SAM}}(\omega) = \nabla_{\omega}\mathcal{L}(\omega)|_{\omega+\epsilon^{*}(\omega)}$.

Integrating Two Key Components:

- Updating rule during learning:

$$\boldsymbol{\omega}_{t+1}^{a} = \boldsymbol{\omega}_{t}^{a} - \eta \cdot \boldsymbol{T}^{v} \cdot \nabla_{\omega}\mathcal{L}^{\text{SAM}}(\omega)$$

Experiments

Experimental Settings

Datasets:

- **CREMA-D:** 7,442 audio-video pairs with 6 emotional categories.
- **KSounds:** 19,000 audio-video pairs with 31 action categories.
- **Twitter2015:** 5,338 image-text pairs with 3 categories.
- **Sarcasm:** 24,635 image-text pairs with 2 categories.

Baselines:

- **Unimodal:** audio, video, image, text, RGB, OF, Depth.
- **Joint-Training:** MSES [ACPR'19], OGR-GB [CVPR'20], OGM [CVPR'22], DOMFN [MM'22], MLSR [ACL'22], PMR [CVPR'23], AGM [ICCV'23], SMV [CVPR'24], MMPareto [ICML'24].
- **Alternating Learning:** ReconBoost [ICML'24] and MLA [CVPR'24].

Evaluation Protocols:

- Accuracy + MAP: audio-video.
- Accuracy + Mac-F1: image-text, RGB-OF-Depth.

Comparison with SOTAs

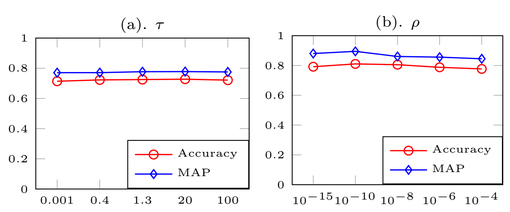
Performance (Accuracy/MAP/Mac-F1):

Method	CREMA-D		KSounds		Twitter2015		Sarcasm		NVGesture	
	Acc.	MAP	Acc.	MAP	Acc.	Mac-F1	Acc.	Mac-F1	Acc.	Mac-F1
Unimodal-1	.6317	.6861	.5312	.5669	.7367	.6849	.8136	.8065	.7822	.7833
Unimodal-2	.4583	.5879	.5462	.5837	.5863	.4333	.7181	.7073	.7863	.7865
Unimodal-3	-	-	-	-	-	-	-	-	.8154	.8183
OGR-GB	.6465	.6854†	.6710	.7139	.7435	.6869	.8335	.8271	.8299	.8305
OGM	.6694	.7173	.6606	.7144	**.7492**	.6874	.8323	.8266	-	-
DOMFN	.6734	.7372	.6625	.7244	.7445	.6857	.8356	.8262	-	-
MSES	.6156†	.6683†	.6471	.7063	.7184†	.6655†	.8418	.8360	.8112†	.8147†
PMR	.6659	.7030	.6656	.7193	.7425	.6860	.8360	.8249	-	-
AGM	.6707	.7358	.6602	.7252	.7483	.6911	.8402	.8344	.8278	.8282
MSLR	.6546	.7138	.6591	.7196	.7252†	.6439†	.8423	.8369	.8286	.8292
ReconBoost	.7484	.8124	.7085	.7424	.7442	.6834	.8437	.8317	.8413	.8632
SMV	.7872	.8417	.6900	.7426	.7428	.6817	.8418	.8368	.8352	.8341
MMPareto	.7487	.8535	.7000	.7850	.7358	.6729	.8348	.8284	.8382	.8424
MLA	.7943	.8572	.7004	.7413	.7352†	.6713†	.8426	.8348	.8373	.8387
IGM w/o SAM	.8026	.8830	.7159	.7623	.7395	.6912	.8455	.8390	.8487	.8634
IGM	**.8105**	**.8948**	**.7403**	**.7855**	.7489	**.6917**	**.8468**	**.8392**	**.8693**	**.8703**

Ablation Study:

SAM	GM	Audio	Video	Multi
✘	✘	45.83%	63.17%	64.52%
✔	✘	58.60%	64.79%	73.42%
✘	✔	60.13%	65.06%	80.26%
✔	✔	**61.16%**	**67.82%**	**81.05%**

Params. Sensitivity:

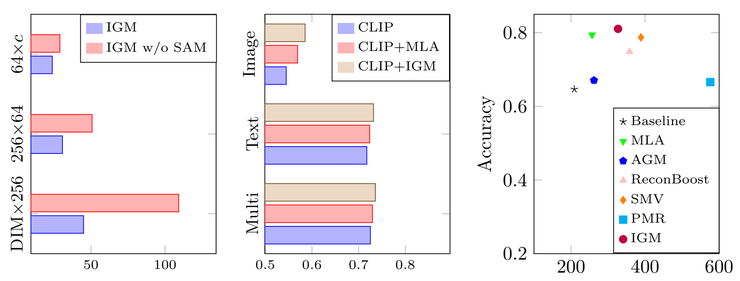


Further Analysis

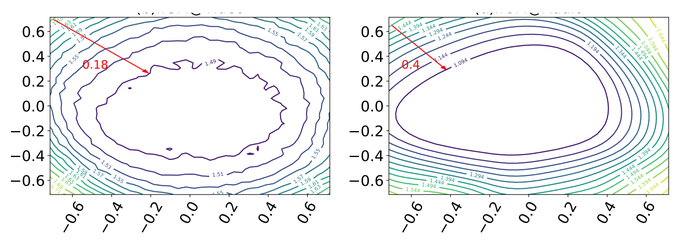
Interactive Enhancement:

Method	*Initial*	*Out_Iters*=1		*Out_Iters*=2	
		Audio	Video	Audio	Video
w/o a-GM	.0325	.5312	.6803	.7231	.7482
w/o v-GM	.0325	.5312	.7023	.7472	.7646
IGM	.0325	.5312	.7023	**.7557**	**.8105**

Singular Values[L], Pretrained Model[M], and Training Time[R]:



Loss Landscape Visualization:



Conclusion

- A **flat-projection gradient modification** based MML method is proposed to address the **poor plasticity** issue.
- SAM optimization algorithm is integrated in the loss function to **smooth the objective function**.
- Comprehensive experiments are conducted to demonstrate the superiority and effectiveness of IGM.

Contact Us