Interactive Multimodal Learning via Flat Gradient Modification

Qing-Yuan Jiang, Zhouyang Chi, and Yang Yang*

IJCAI, 2025, Guangzhou,

Nanjing University of Science and Technology

August 30, 2025

1. Introduction
2. Methodology
3. Experiments
4. Conclusion

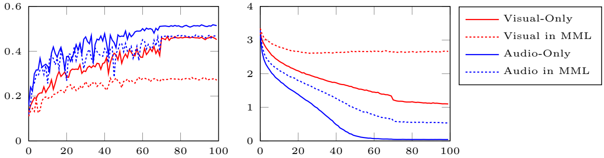
Imbalance in Multimodal Learning

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.

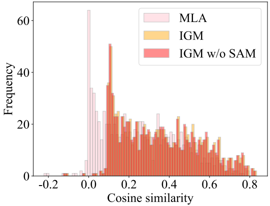
Issue in Orthogonal Projection

Alternating Learning: Learning each modality one-by-one

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**.

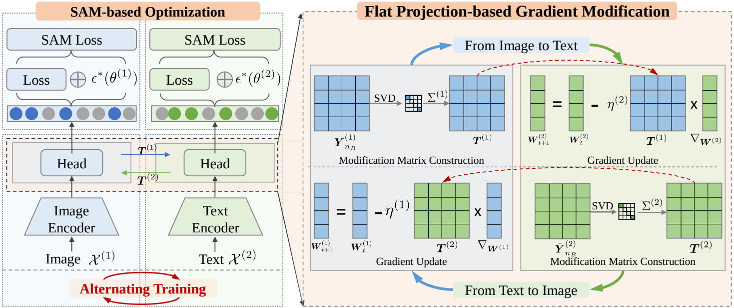
Gradient Change:



1. Introduction
2. Methodology
3. Experiments
4. Conclusion

Overall Framework

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



Flat Projection Gradient Modification

- **Find Flat Direction**: $\mathbf{U}^v\Lambda^v[\mathbf{V}^v]^\top = svd(\mathbf{Y}^v)$
- **Relationship of Flatness and Singular Value**: Small $\lambda^v \mapsto$ Flat Area
- **Conduct Projection Matrix**: $\Sigma^v = exp(-\frac{\tau}{\lambda^v_{\text{max}}-\lambda^v_{\text{min}}}(\Lambda^v - \lambda^v_{\text{min}}\mathbf{I}))$
- **Project during Updating**:

$$\mathbf{T}^v = \mathbf{U}^v\Sigma^v[\mathbf{V}^v]^\top$$
$$\omega^a_{t+1} = \omega^a_t - \eta \cdot \boldsymbol{T}^v \cdot \nabla_\omega \mathcal{L}$$

SAM-based Optimization

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

Overall Algorithm

Updating rule: $\omega_{t+1}^{a} = \omega_{t}^{a} - \eta \cdot \boldsymbol{T}^{v} \cdot \nabla_{\omega^{a}} \mathcal{L}^{\mathrm{SAM}}(\omega^{a})$

Algorithm 1 Algorithm for IGM

Input: Training set $\mathcal{D}$ and labels $\boldsymbol{Y}$;

Output: The learned parameters $\{\theta^{(j)}\}_{j=1}^{(m)}$;

INIT: Initialize gradient modification matrix. Initialize $\{\boldsymbol{T}^{(k)}\}_{j=1}^{(m)}$: $\forall k \in \{1, \cdots, m\}$, $\boldsymbol{T}^{(k)} = \boldsymbol{I}$;

for $i = 1 \rightarrow$ *Out_Iters* **do**
  for $j = 1 \rightarrow m$ **do** ▷ Main iteration.
    for $t = 1 \rightarrow$ *Inner_Iters* **do**
      Randomly construct a mini-batch $\mathcal{X}_t^{(j)}$.
      Calculate loss $L(\theta^{(j)})$ for data in $\mathcal{X}_t^{(j)}$.
      Calculate $\epsilon^*(\theta^{(j)})$ according to Eq. (7).
      Calculate $\nabla_{\theta^{(j)}} L^{\mathrm{SAM}}$ according to Eq. (8).
      Calculate modality index:
        $k = \mathtt{mod}(j + m - 2, m) + 1$.
      Update $\theta^{(j)}$: $\theta_{t+1}^{(j)} = \theta_t^{(j)} - \eta^{(j)} \boxed{\boldsymbol{T}^{(k)} \nabla_{\theta^{(j)}} L^{\mathrm{SAM}}}$.
    for $j = 1 \rightarrow n_B$ **do** ▷ Update $\{\bar{Y}_{n_B}^{(k)}\}$.
      Update cumulative variance according to Eq. (2).
    Update $\boldsymbol{T}^{(j)}$ according to Eq. (4). ▷ Update $\boldsymbol{T}^{(j)}$.

Settings

Datasets:

- **CREMA-D:** 7,442 audio-video pairs with 6 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.
- **NVGesture:** 1,532 dynamic hand gestures with 3 modalities.

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].

Comparison with SOTAs

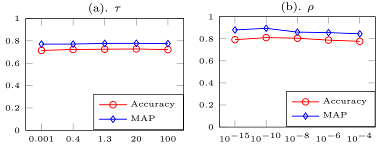
Method	CREMA-D		KSounds		Twitter2015		Sarcasm		NVGesture	
	Acc.	MAP	Acc.	MAP	Acc.	Mac-F1	Acc.	Mac-F1	Acc.	Mac-F1
Unimodal-1	.6317	.6861	.5312	.5669	.7367	.6849	.8136	.8065	.7822	.7833
Unimodal-2	.4583	.5879	.5462	.5837	.5863	.4333	.7181	.7073	.7863	.7865
Unimodal-3	-	-	-	-	-	-	-	-	.8154	.8183
OGR-GB	.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**

Further Analysis

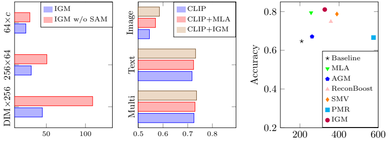
Ablation Study:

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%**

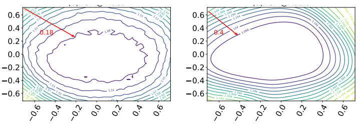
Params. Sensitivity:



Singular Values, Pretrained Model, and Training Time:



Loss Landscape Visualization:



1. Introduction
2. Methodology
3. Experiments
4. Conclusion

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:

Thanks