The 34th International Joint Conference on Artificial Intelligence (IJCAI)

IJCAI 2025 Guangzhou August 29-31

Paper ID: 4373

Balance-aware Sequence Sampling Makes Multimodal Learning Better

Zhi-Hao Guan, Qing-Yuan Jiang*, Yang Yang*

Nanjing University of Science and Technology, Nanjing, China

Background

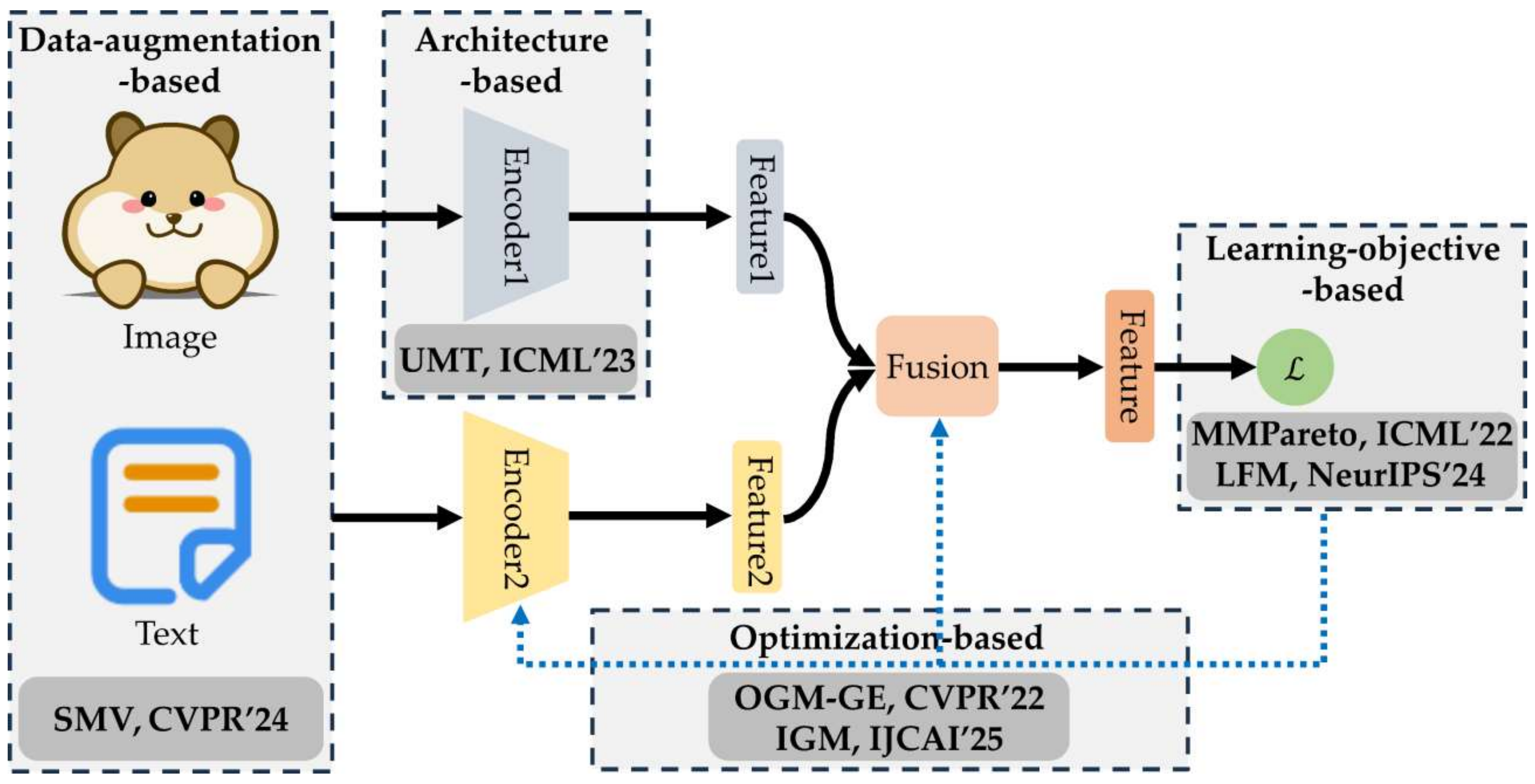
◆ Modality Imbalance

Due to **modality heterogeneity**, multimodal learning (MML) is often dominated by stronger modalities, resulting in insufficient learning of weaker ones and suboptimal overall performance.

◆ Modality Rebalance Method

- Learning-objective-based: **MMPareto, LFM**
- Optimization-based: **OGM-GE, IGM**
- Architecture-based: **UMT**
- Data-augmentation-based: **SMV**

Outstanding Performance!



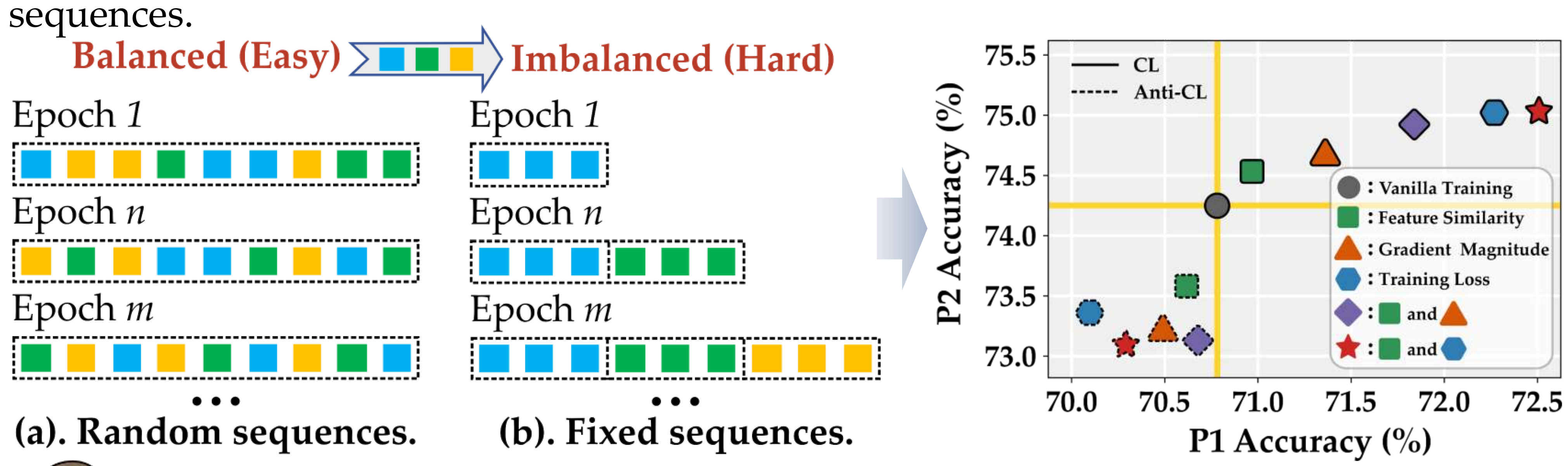
Motivation

◆ Viewpoint

Although existing methods have shown promising results, they generally overlook a key aspect: **MML can be highly sensitive to the training sequence**. Since the standard training paradigm is characterized by random data shuffling, this process inevitably introduces imbalanced samples into early training stages, which may further exacerbate modality imbalance and ultimately degrade MML performance.

◆ Toy Experiment

We investigate the relationship between different training sequences and MML performance. Inspired by **curriculum learning (CL)**, we first evaluate the balance degree of sample pairs based on various criteria, and then rank them to construct new training sequences.



CL effectively boosts MML performance, whereas anti-CL degrades it!

Method

◆ Multi-perspective Measurer

The balance score of a sample $\boldsymbol{x}_i$ can be formulated as the combination of **correlation criterion** (prediction similarity) and **information criterion** (training loss):

$$s(\boldsymbol{x}_i)=\frac{\mathtt{sim}(\boldsymbol{x}_i^{(u)},\boldsymbol{x}_i^{(v)})-\min(\mathcal{S})}{\max(\mathcal{S})-\min(\mathcal{S})}-\frac{\ell_{total}(\boldsymbol{x}_i^{(u)},\boldsymbol{x}_i^{(v)},\boldsymbol{y}_i)-\min(\mathcal{L})}{\max(\mathcal{L})-\min(\mathcal{L})}.$$

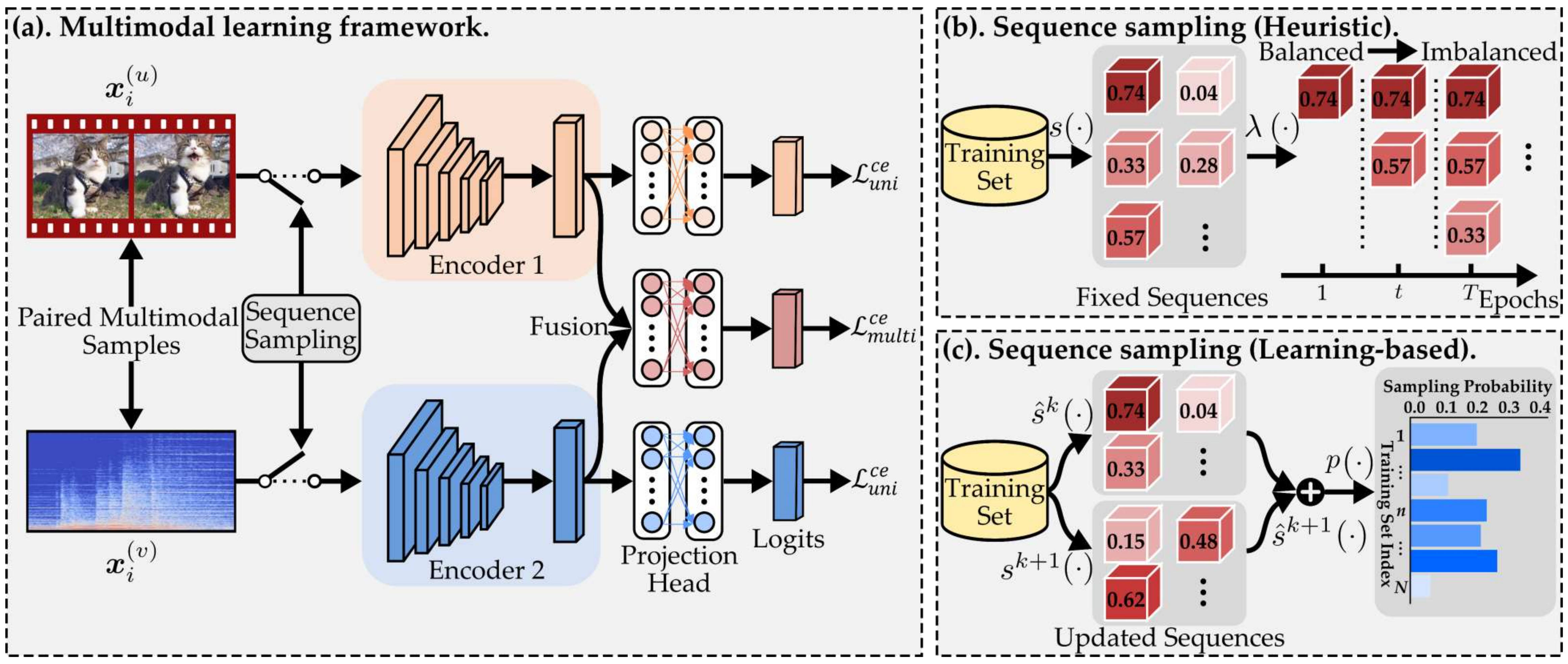
◆ Training Scheduler

Heuristic Scheduler: Following curriculum learning, we adopt a widely-used pacing function $\lambda(t)$ to achieve this:

$$\lambda(t)=\min\left(1,\sqrt{\frac{1-\lambda_0^2}{T_{grow}}\cdot t+\lambda_0^2}\right).$$

At epoch t, current batch data $\boldsymbol{X}_{batch}$ is randomly sampled from top λ proportion of training data in the entire ranked sequence $\boldsymbol{X}_{rank}$:

$$\boldsymbol{X}_{batch}(t)=\mathtt{Sampling}(\{\boldsymbol{x}_i|\boldsymbol{x}_i\in\boldsymbol{X}_{rank},i<\lfloor n\cdot\lambda(t)\rfloor\}).$$



Leaning-based Scheduler: Since heuristic scheduler may neglect model feedback. We further propose a learning-based scheduler that reconstructs the dynamic sequence by learning a sampling probability for each sample, considering both the balance of past and current samples in a more fine-grained manner.

Update Formula: $\hat{s}^{k+1}(\boldsymbol{x}_i)=\begin{cases}s^{k+1}(\boldsymbol{x}_i), & \text{if } k=0,\\(1-\beta)\hat{s}^k(\boldsymbol{x}_i)+\beta s^{k+1}(\boldsymbol{x}_i), & \text{otherwise.}\end{cases}$

Sampling Probability: $p(\boldsymbol{x}_i)=\frac{e^{\hat{s}^{k+1}(\boldsymbol{x}_i)}}{\sum_{j=1}^n e^{\hat{s}^{k+1}(\boldsymbol{x}_j)}}.$

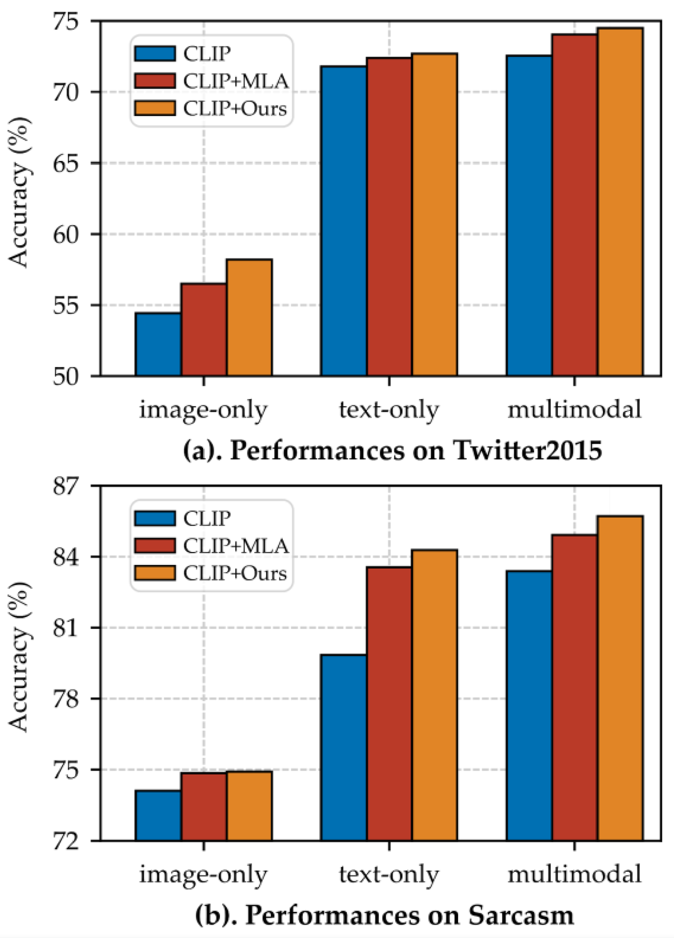
Current Batch Data: $\boldsymbol{X}_{batch}(t)=\mathtt{Sampling}(\{p(\boldsymbol{x}_1),p(\boldsymbol{x}_2),\ldots,p(\boldsymbol{x}_n)\}).$

Experiments

◆ Classification Results

Method	CREMA-D		Kinetics-Sounds		Twitter2015		Sarcasm		NVGesture	
	ACC (%)	MAP (%)	ACC (%)	MAP (%)	ACC (%)	F1 (%)	ACC (%)	F1 (%)	ACC (%)	F1 (%)
Audio/Text/RGB	63.17	68.61	54.12	56.69	73.67	68.49	81.36	80.65	78.22	78.33
Video/Image/OF	45.83	58.79	55.62	58.37	58.63	43.33	71.81	70.73	78.63	78.65
Depth	-	-	-	-	-	-	-	-	81.54	81.83
MSES	61.56	68.83	64.71	70.63	71.84	66.55	84.18	83.60	81.12	81.47
OGR-GB	64.65	84.54	67.10	71.39	74.35	68.69	83.35	82.71	82.99	83.05
DOMFN	67.34	85.72	66.25	72.44	74.45	68.57	83.56	82.62	-	-
OGM	66.94	71.73	66.06	71.44	74.92	68.74	83.23	82.66	-	-
MSLR	65.46	71.38	65.91	71.96	72.52	64.39	84.23	83.69	82.86	82.92
AGM	67.07	73.58	66.02	72.52	74.83	69.11	84.02	83.44	82.78	82.82
PMR	66.59	70.30	66.56	71.93	74.25	68.60	83.60	82.49	-	-
ReconBoost	74.84	81.24	70.85	74.24	74.42	68.34	84.37	83.17	84.13	86.32
MMPareto	74.87	85.35	70.00	78.50	73.58	67.29	83.48	82.48	83.82	84.24
SMV	78.72	84.17	69.00	74.26	74.28	68.17	84.18	83.68	83.52	83.41
MLA	79.43	85.72	70.04	74.13	73.52	67.13	84.26	83.48	83.40	83.72
AMSS	70.30	76.14	72.25	79.13	75.12	69.23	84.35	83.77	84.64	84.94
BSS-H	80.78	87.86	72.67	78.61	74.73	68.67	84.41	83.86	85.06	85.15
BSS-L	**82.80**	**88.61**	**73.95**	**79.43**	**75.22**	**69.51**	**85.01**	**84.62**	**86.72**	**87.04**

Our BSS achieves SOTA performance across various datasets!

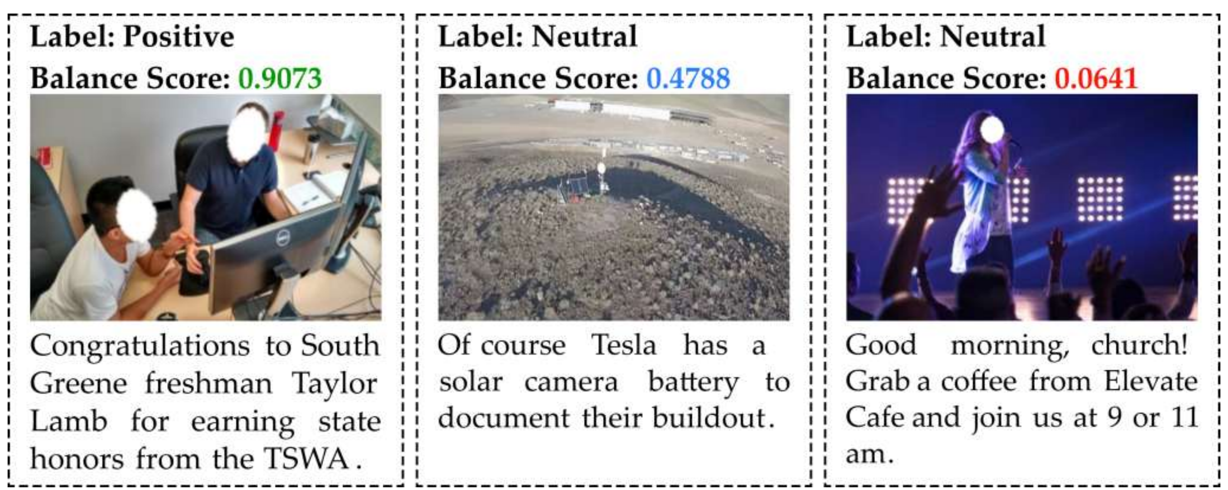


(a). Performances on Twitter2015

(b). Performances on Sarcasm

◆ Ablation Study

Criterion		ACC (%) / MAP (%)		
PreSim	Loss	Audio	Video	Multi
✗	✗	49.37/51.07	54.03/57.48	70.44/76.62
✗	✓	52.11/54.40	54.23/57.91	72.44/79.41
✓	✗	52.38/54.32	**54.93/58.52**	73.25/78.98
✓	✓	**52.73/54.43**	54.74/58.46	**73.95/79.43**



Conclusion

BSS mitigates modality imbalance problem by evaluating sample balance with a multi-perspective measurer and constructing balanced-to-imbalanced training sequences using both heuristic and learning-based schedulers.

Contact Info

zhguan@njust.edu.cn
jiangqy@njust.edu.cn
yyang@njust.edu.cn

KMG Group

WeChat