Balanced Multimodal Learning via On-the-fly Gradient Modulation (Supplementary Materials)

Xiaokang Peng[1,†], **Yake Wei**[1,†], **Andong Deng**[2], **Dong Wang**[3], **Di Hu**[1,*]

[1]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing
[1]Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing
[2]Shanghai Jiao Tong University, Shanghai
[3]Shanghai Artificial Intelligence Laboratory, Shanghai
[1]{xiaokangpeng, yakewei, dihu}@ruc.edu.cn, [2]{andongdeng69, dongwang.dw93}@gmail.com

1. Imbalance analysis of summation structure

In this section, we introduce the analysis of the optimization imbalance phenomenon for the model with summation as fusion method. For convenience, here we continue to use most of the notations in Section 3.1. Different from the case of concatenation, model with summation has two independent linear classifiers for corresponding modalities, whose parameters consist of W^a, b^a, W^v and b^v. Then the logits output of the multimodal model is given by:

$$f(x_i) = W^a \cdot \varphi^a(\theta^a, x_i^a) + b^a + W^v \cdot \varphi^v(\theta^v, x_i^v) + b^v. \quad (1)$$

Similarly, we denote the logits output for class c as $f(x_i)_c$, thus the cross-entropy loss of the discriminative model becomes $L = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{f(x_i)_{y_i}}}{\sum_{k=1}^{M} e^{f(x_i)_k}}$. With the *Gradient Descent* (GD) optimization method, W^a, b^a, and the parameters of encoder $\varphi^a(\theta^a, \cdot)$ are updated as (similarly for W^v and $\varphi^v(\theta^v, \cdot)$):

$$\begin{aligned} W_{t+1}^a &= W_t^a - \eta \nabla_{W^a} L(W_t^a) \\ &= W_t^a - \eta \frac{1}{N} \sum_{i=1}^{N} \frac{\partial L}{\partial f(x_i)} \varphi^a(\theta^a, x_i^a), \end{aligned} \quad (2)$$

$$\begin{aligned} b_{t+1}^a &= b_t^a - \eta \nabla_{b^a} L(b_t^a) \\ &= b_t^a - \eta \frac{1}{N} \sum_{i=1}^{N} \frac{\partial L}{\partial f(x_i)}, \end{aligned} \quad (3)$$

$$\begin{aligned} \theta_{t+1}^a &= \theta_t^a - \eta \nabla_{\theta^a} L(\theta_t^a) \\ &= \theta_t^a - \eta \frac{1}{N} \sum_{i=1}^{N} \frac{\partial L}{\partial f(x_i)} \frac{\partial (W_t^a \cdot \varphi_t^a(\theta^a, x_i^a))}{\partial \theta_t^a}, \end{aligned} \quad (4)$$

where η is the learning rate. According to Equation 2 to 4, we can tell that the optimization of W^a, b^a and φ^a has nearly no correlation with that of the visual modality (vice versa), except the term related to the training loss ($\frac{\partial L}{\partial f(x_i)}$). The modal-specific encoders thus can hardly make adjustment according to the feedback from each other. Therefore, combined with Equation 1, the gradient term $\frac{\partial L}{\partial f(x_i)}$ can be rewritten as follows:

$$\frac{\partial L}{\partial f(x_i)_c} = \frac{e^{(W^a \cdot \varphi_i^a + b^a + W^v \cdot \varphi_i^v + b^v)_c}}{\sum_{k=1}^{M} e^{(W^a \cdot \varphi_i^a + b^a + W^v \cdot \varphi_i^v + b^v)_k}} - \mathbb{1}_{c=y_i}. \quad (5)$$

For convenience, we simplify $\varphi^a(\theta^a, x_i^a)$ and $\varphi^v(\theta^v, x_i^v)$ as φ_i^a and φ_i^v, respectively. Then, we can infer that for sample x_i belonging to class y_i, when one modality, say, visual modality, shows better performance, it contributes more to shared $\frac{\partial L}{\partial f(x_i)_{y_i}}$ via larger $(W^v \cdot \varphi_i^v + b^v)$, leading to lower loss globally. Consequently, the audio modality, which is less confident for the correct category, could obtain only limited optimization efforts w.r.t. its modal-specific parameters during the back propagation.

This analysis, which is analogous with that in Section 3.1, consolidates that the imbalanced phenomenon is common in the training process of multimodal model with different fusion method.

2. Supplementary experiment and analysis

2.1. Applications beyond classification.

We further employ OGM-GE in multimodal representation learning (MMRL) task to balance the different learning pace between positive and negative sample pairs, which may bring negative effects for the quality of the learned representation. Performance of positive pairs and negative pairs are estimated separately and adjusted according to their discrepancy. We take L^3-Net framework [1], Audio-visual Scene Analysis [3], and Audio-visual Co-attention [2], respectively, to perform multimodal self-supervised pretraining on Kinetics-Sounds and then conduct audio classification with the pretrained audio encoder on ESC50 [4] to evaluate the representation quality. Results

shown in Tabel 1 demonstrate the effectiveness of OGM-GE when applied in imbalanced learning problem brought by positive and negative sample pairs in representation learning.

For the audio-visual event localization task, we simply insert OGM-GE without other operations in AGVA [5]. In PSP [6], considering its sophisticated cross-modal interaction, which somehow alleviate the modality imbalance thanks to its well-designed attention mechanism, we only modulate gradients of part of the parameters, specifically, only gradients before positive sample propagation module are modulated.

Multimodal Representation Learning	
L^3-Net [1]	53.9
AVSA [3]	53.1
AVCA [2]	52.8
L^3-Net†	**57.3**
AVSA†	**57.2**
AVCA†	**54.6**

Table 1. Experimental results on multimodal representation learning. Classification accuracy in downstream task is utilized as the evaluation metric. † indicates OGM-GE is applied.

2.2. Fine-grained effectiveness analysis

In this part, we analyze our OGM-GE training strategy from the perspective of both category and sample. As shown in Figure 1, our method improves the performance of most categories to a certain degree. Further, we notice that the modality with less confidence tends to gain more performance improvement after being equipped with our method. It is also validated that no matter which modality dominates the training for a category, the OGM-GE method is capable of alleviating the imbalanced situation.

To further explore whether the model gains improvement in the sample-level, we show the training process of some samples in Figure 2. We compare the audio performance, visual performance, and multimodal performance in multimodal learning under three settings: concatenation, concatenation with OGM, and concatenation with OGM-GE. Here we introduce the confidence in classification (probability of the correct category) to measure the learning quality of a single sample. The results demonstrate the effectiveness of our OGM-GE method.

References

[1] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617, 2017. 1, 2

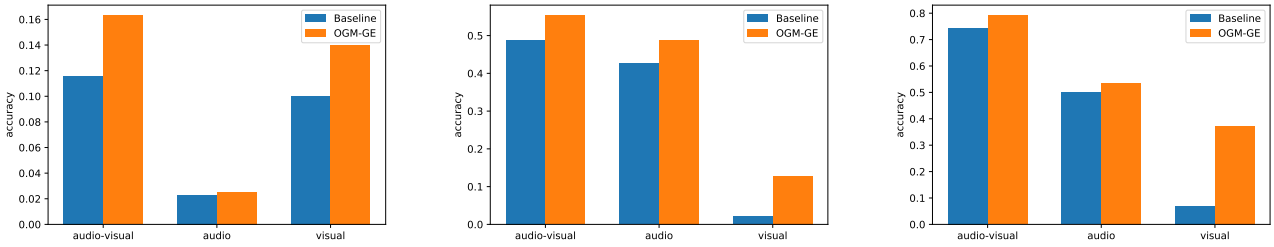
[2] Ying Cheng, Ruize Wang, Zhihao Pan, Rui Feng, and Yuejie Zhang. Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3884–3892, 2020. 1, 2

[3] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018. 1, 2

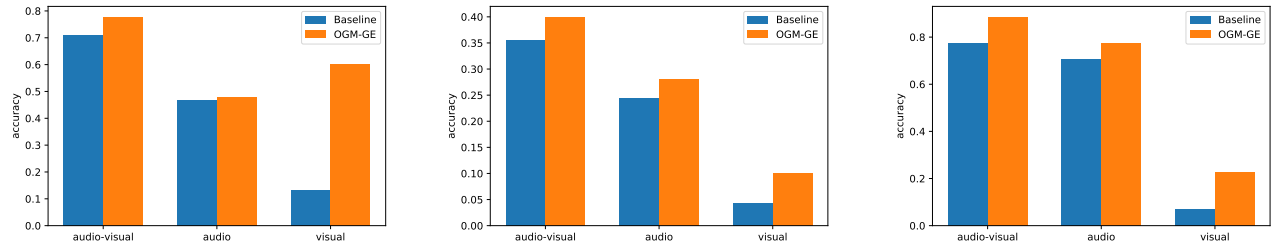
[4] Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018, 2015. 1

[5] 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

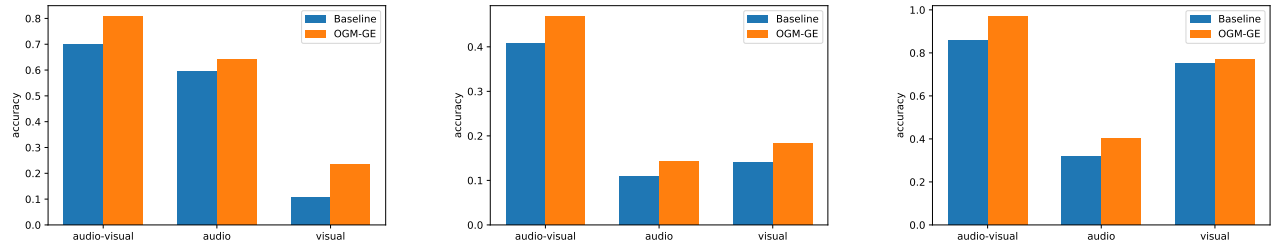
[6] Jinxing Zhou, Liang Zheng, Yiran Zhong, Shijie Hao, and Meng Wang. Positive sample propagation along the audio-visual event line. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8436–8444, 2021. 2



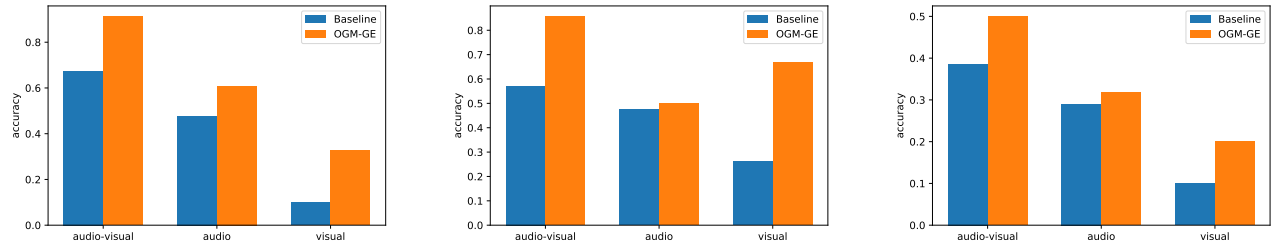
(a) people running (b) basketball bounce (c) airplane flyby



(d) volcano explosion (e) ambulance siren (f) hail

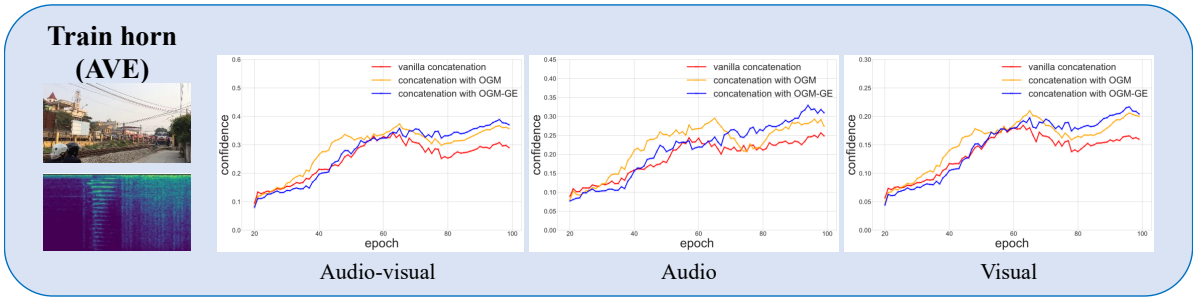


(g) playing double bass (h) hedge trimmer running (i) slot machine

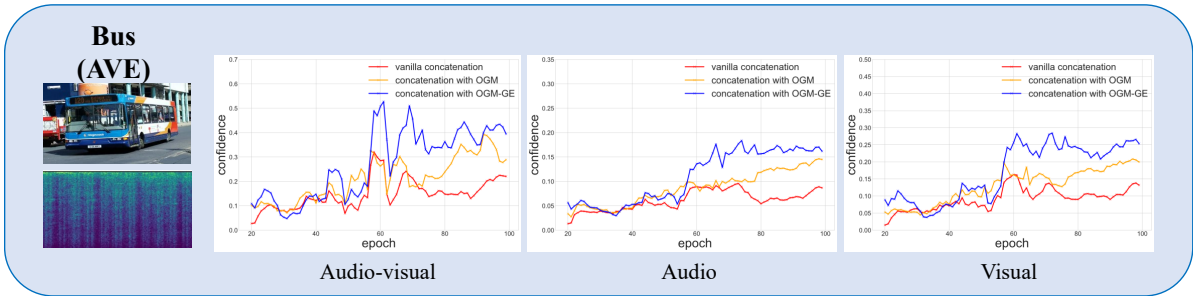


(j) tap dancing (k) bowling impact (l) wood thrush calling

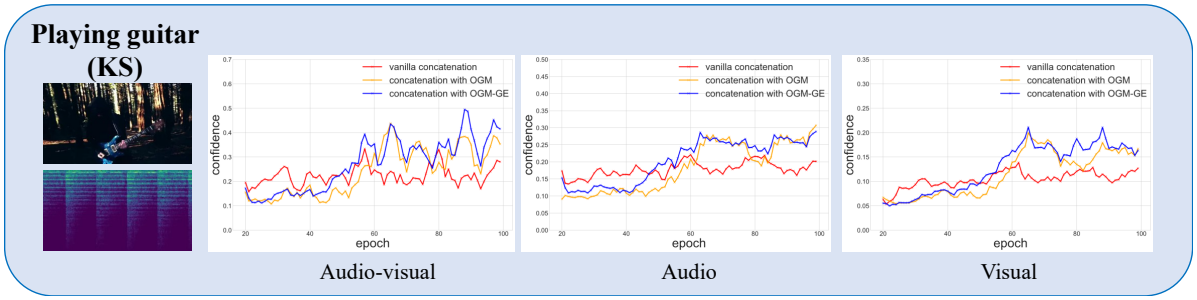
Figure 1. Performance of different categories on VGGSound, with vanilla concatenation framework and that applied with our proposed OGM-GE. Classes dominated by audio and visual examples are both provided.



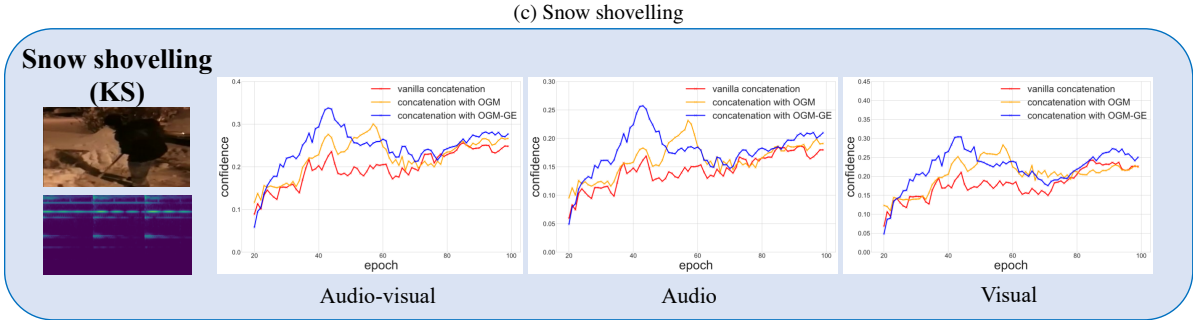
(a) Train horn



(b) Bus



(c) Snow shovelling



(d) Playing guitar

Figure 2. Performance of different samples on AVE and Kinetics-Sounds. For each sample we show results of three training settings: vanilla concatenation, OGM, and OGM-GE.