Towards Better Egocentric Action Understanding in a Multi-Input Multi-Output View

Wenxuan Hou[1], Ruoxuan Feng[1], Yixin Xu[1], Yapeng Tian[2], Di Hu[1*]
[1] Renmin University of China, Beijing, China [2] The University of Texas at Dallas, Dallas, USA
wxhou@ruc.edu.cn, fengruoxuan@ruc.edu.cn, xu_yixin@ruc.edu.cn
yapeng.tian@utdallas.edu, dihu@ruc.edu.cn

Abstract

Egocentric videos are close records of the human's field of vision, making action recognition within them crucial for understanding human behavior and intentions. When recognizing actions in egocentric videos, we usually extract and use the data of several modalities (such as RGB, optical flow, and audio), and simultaneously predict the verb and noun of actions. Thus, egocentric action recognition can be viewed as a Multi-Input Multi-Output (MIMO) problem as it contains characteristics of both multimodal and multitask. These characteristics bring some challenges that have not been adequately explored so far. First, the optimization objectives between different tasks could conflict, impacting the quality of feature representations. Second, the phenomenon of modality imbalance becomes more complex, different tasks may bias on different modalities. To address the first challenge, we propose a query-based fusion architecture with specific task tokens to query task-specific features. For the second challenge, we propose an attention modulation strategy. An attention allocating loss and reallocating strategy are performed to improve the training and inference phase respectively. By identifying and addressing the challenges in the MIMO problem, our proposed methods boost the performance of egocentric action recognition.

1. Introduction

Learning and understanding egocentric videos has become a research hotspot in the computer vision community in recent years. An egocentric video is generally recorded by head-mounted cameras, showing the wearer's field of vision. Thus, egocentric videos can indicate the attention and intentions of the wearer, playing an important role in human gaze prediction [8], human object interaction [14], and human action anticipation [6]. The research on egocentric video can further be applied to many fields, such as AR/VR [13], and embodied AI [10].

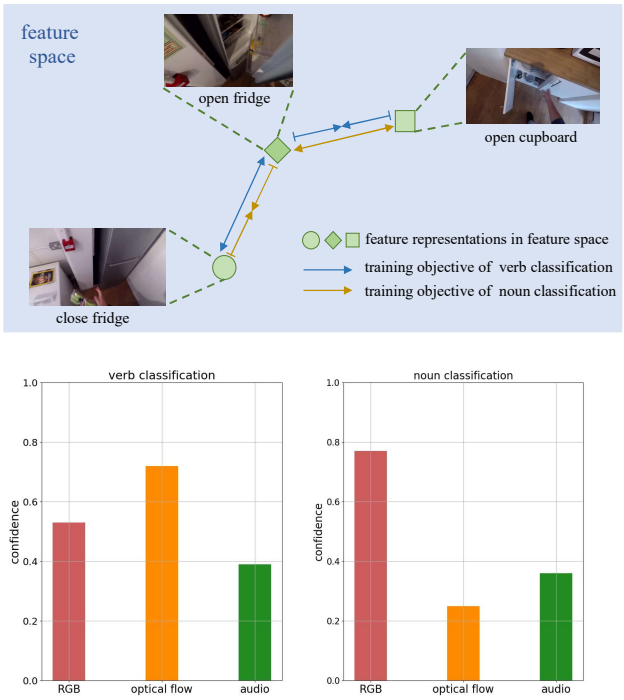

Figure 1. An illustration of challenges in a MIMO problem. **Top:** An example of task conflict, the training objectives of two tasks are conflicting. **Bottom:** An example of task-aware modality bias, for different tasks, the modalities with highest unimodal confidence are not the same.

Action recognition is a basic and important task in video understanding. The action recognition task in egocentric videos has two characteristics: First, egocentric videos usually contain multiple different modalities (*e.g.,* RGB, optical flow, and audio), which describe egocentric scenes from various aspects and all provide valuable information for action recognition. Second, in egocentric videos, an action can be usually described as a verb-noun phase [4], thus the action recognition task in egocentric videos requires the accurate classification of both verb and noun. In conclu-

*Corresponding author.

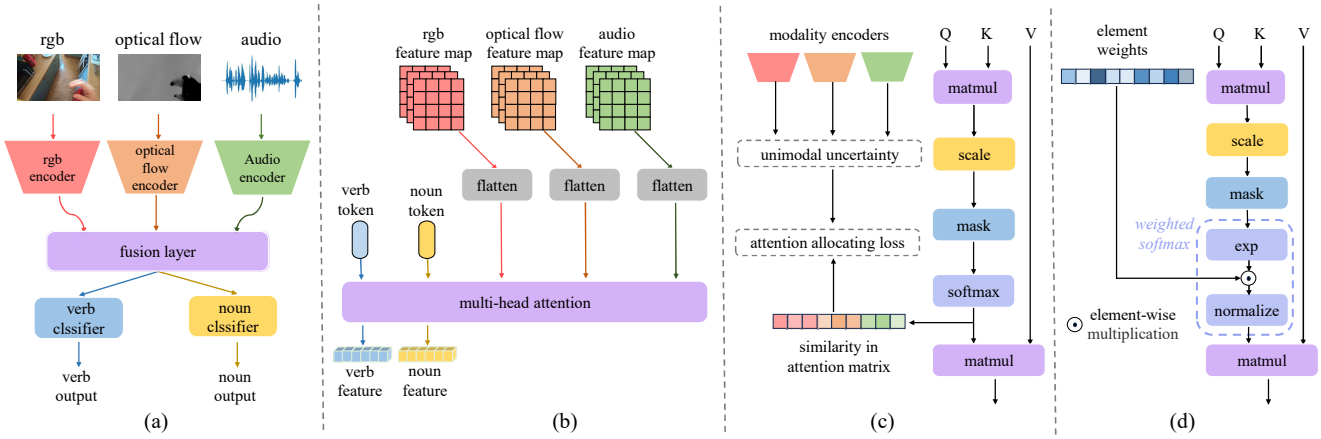

Figure 2. **(a)**: A classic model architecture for multimodal learning, we use it as our overall model architecture. **(b)**: Our proposed query-based fusion architecture, which contains several specific task tokens. **(c)**: Scaled dot-product attention module during the training phase, our proposed attention allocating loss is performed to improve training. **(d)**: Scaled dot-product attention module during the inference phase, *exp* represents exponential operation. The weighted softmax operation is performed to reallocate the contribution of each modality.

sion, the egocentric action recognition task is a *Multi-Input Multi-Output* (MIMO) problem.

The characteristic of MIMO brings two challenges on recognizing egocentric actions. **(1) Task conflict.** The characteristic of multi-output could lead to conflict between two tasks when these two tasks share the same feature space. An example is shown in the top of Fig. 1. For the task of verb classification, samples *open fridge* and *open cupboard* should be close in the feature space, while samples *open fridge* and *close fridge* should be far apart. However, for the task of noun classification, an opposite optimization objective is expected. Previous methods [12, 11] overlooked this point. They simply obtain a shared feature representation for all tasks. **(2) Task-aware modality bias.** Due to the heterogeneity of multimodal data, the modality imbalance phenomenon [17] is widely existing in multimodal learning. In the MIMO problem, a more complex phenomenon exists as different tasks may bias on different modalities. The unimodal confidence of an example is shown in the bottom of Fig 1. For the verb classification task, the optical flow modality has the highest confidence, while for the noun classification task, the RGB modality has the highest confidence, the optical flow modality even has the lowest confidence. Although some balanced multimodal learning methods have been proposed to address the modality imbalance phenomenon [20, 17, 21], they are not applicable for addressing the imbalance issue in the MIMO problem. These methods are mainly designed for modulating a single task and could not tackle the difference of modality bias between different tasks.

In this paper, we betake to address the above ignored or unsolved challenges. First, for the challenge of task conflict, we propose a query-based fusion architecture to integrate different modalities, as shown in Fig. 2 (b). We set a specific token for each task (task token), and concatenate the task tokens with feature tokens. Then tokens are fed into a multi-head attention layer. Task tokens in the output are used to perform the corresponding task. In this manner, task tokens can query task-specific features, thereby eliminating the commonality of the feature space among distinct tasks. Second, for the challenge of task-aware modality bias, we further propose an attention modulation strategy, which draws inspiration from the learning and testing process of middle school students. During regular practice, more attention should be devoted to hard and mistaken questions, while during the exam, correctness is the most important, and easy questions should be answered at first. We apply a similar modulation strategy on the query-based fusion module. In the training phase, we perform an attention allocating loss to encourage each task token to pay more attention to corresponding non-dominated modalities [16] via a task-aware metric, guiding the model to be trained more adequately. In the inference phase, we reallocate the contribution of each modality, encouraging each task token to pay more attention to the corresponding dominated modality, as the dominated modality is usually more reliable. Extensive experiment results indicate that our proposed methods improve the recognition performance. Our contributions can be summarized as follows:

- We view egocentric action recognition as a MIMO problem and identify two ignored but important challenges: task conflict and task-aware modality bias.
- We propose a query-based fusion architecture with specific task tokens to tackle the challenge of task conflict and an attention modulation strategy for the challenge of task-aware modality bias.
- Experiment results show that our proposed methods effectively improved recognition performance.

2. Method

2.1. Query-based Fusion Architecture

In multimodal learning, a classic model architecture usually consists of modality-specific encoders, a fusion layer, and several task heads, as shown in Fig. 2 (a). Most existing fusion layers simply produce a shared feature representation for all tasks, thus suffering from the task conflict because all task heads share the same feature space.

Inspired by DETR [2], we propose a query-based fusion architecture. For each task, we set a specific token to query task-specific features individually, thus different tasks no longer need to share the same feature space. Specifically, we use Video Swin Transformer [15] as the encoder for all modalities. Data of the audio modality is transformed into spectrograms at first. We represent the output feature maps of the modality *m* as $f_m \in \mathbb{R}^{t\times h\times w\times c}$, where t, h, w, c are temporal length, height, width and the number of channels of feature maps. Following Pixel-BERT [9], we tokenize feature maps by flatting them on temporal and spatial dimensions. All feature tokens and task tokens are concatenated to a feature sequence F. Then, F is fed into a standard multi-head attention layer, guiding different task tokens to query task-specific features. Eventually, we obtain distinct representations for different tasks.

In practice, we set two task tokens (one for verb and another for noun) and three input modalities (RGB, optical flow, and audio), hence the length of F is $l = 3 \times t \times h \times w + 2$. Fig. 2 (b) shows the details of our proposed query-based fusion architecture.

2.2. Attention Allocating in the Training Phase

By performing the query-based fusion architecture, we have solved the challenge of task conflict. To further overcome the challenge of task-aware modality bias, we propose the attention modulation strategy, which contains two parts: attention allocating in the training phase, and attention reallocating in the inference phase. We will introduce them in this section, and the next section, respectively.

Due to the widely existing modality imbalance phenomenon, in the multi-head attention layer, task tokens tend to query features from the dominated modality [16], as the dominated modality tends to converge more quickly. This could lead to optimization suppression of non-dominated modalities and inadequate training of the model.

To tackle the problem, we propose to allocate the similarity in attention matrix among different modalities, guiding task tokens to pay more attention to non-dominated modalities, to alleviate the suppression of non-dominant modalities, as shown in Fig. 2 (c). We select uncertainty as the metric to discern the dominated modality. We set an extra unimodal classification layer for each modality to obtain the unimodal uncertainty. The unimodal uncertainty is calculated as the entropy of the prediction distribution [18].

FA	AA	AR	verb	noun	action
concatenation (baseline)	✗	✗	52.12%	36.37%	24.47%
query-based	✗	✗	54.38%	41.04%	27.81%
query-based	✓	✗	55.33%	42.31%	28.52%
query-based	✓	✓	55.46%	42.66%	28.86%

Table 1. Experiment results of our proposed method. *FA* represents fusion architecture; *AA* represents attention allocating; *AR* represents attention reallocating.

We represent the attention matrix in the multi-head attention layer as $A \in \mathbb{R}^{n\times l\times l}$, where n is the number of heads. In the k-th head, suppose the sum of attention similarity between the verb task token and all tokens from the i-th modality is s_i^k, the unimodal uncertainty of the i-th modality is u_i. Then, the allocating loss of the verb classification task is:

$$L^v_{alloc} = \frac{1}{n}\Sigma^3_{i=1}\Sigma^3_{j=1}\Sigma^n_{k=1} max(0, u_j - u_i) * max(0, s^k_j - s^k_i), \tag{1}$$

where u_i, u_j are unimodal uncertainty of i-th and j-th modality. The loss of the noun classification task L^n_{alloc} is calculated in the same way. We also set two cross-entropy losses for verb and noun classification tasks, respectively, represent them as L^v_{ce} and L^n_{ce}. The total loss is:

$$L = L^v_{ce} + L^n_{ce} + \lambda(L^v_{alloc} + L^n_{alloc}), \tag{2}$$

where λ is the hyper-parameter that balance the cross-entropy loss and the allocating loss.

2.3. Attention Reallocating in the Inference Phase

In the training phase, we add an allocating loss to improve the optimization of the model. In the inference phase, the parameters of the model will not be updated, we propose to guide the model to pay more attention to the dominated modality as the feature representation of the dominated modality is usually more powerful. To achieve this, we propose a weighted softmax operation to reallocate the contribution of each modality, as shown in Fig. 2 (d).

Suppose the element in the x-th row, y-th column and k-th head in the pre-normalized attention matrix in the multi-head attention module is $m^k_{x,y}$, the standard softmax operation in the scaled dot-product attention is calculated as:

$$a^k_{x,y} = \frac{exp(m^k_{x,y})}{\Sigma^l_{y'=1} exp(m^k_{x,y'})}, \tag{3}$$

where $a^k_{x,y}$ is the element in normalized attention matrix. Compared to this, our proposed weighted softmax operation apply weighting to each element of the matrix, it is calculated as:

$$a^k_{x,y} = \frac{w_y * exp(m^k_{x,y})}{\Sigma^l_{y'=1} w_{y'} * exp(m^k_{x,y'})}, \tag{4}$$

fusion method	verb	noun	action
concatenation	52.12%	36.37%	24.47%
summation	51.62%	37.90%	24.90%
gating fusion [1]	53.76%	36.34%	25.25%
mid-level fusion [12]	52.52%	34.71%	23.92%
query-based (ours)	54.38%	41.04%	27.81%

Table 2. Comparison of existing fusion methods and our proposed query-based fusion architecture.

modulation method	verb	noun	action
None	52.12%	36.37%	24.47%
OGM (v) [17]	52.53%	36.91%	25.05%
OGM (n) [17]	52.01%	36.77%	24.56%

Table 3. Results of applying the balanced multimodal learning method OGM, OGM (v) represents obtaining the discrepancy ratio from verb classification layer, and OGM (n) represents obtaining the discrepancy ratio from noun classification layer.

where w_y is the weight of the $m^k_{x,y}$. For each modality, we set a unified weight for all tokens from it. By adjusting the weight, we can achieve a simple but effective modulation to reallocate the contribution of each modality.

3. Experiment

3.1. Experiment Settings

We perform all experiments on the EPIC-KITCHENS-100 [4] dataset, which contains egocentric videos of 100 hours that are recorded in kitchen scenes with annotations of 89,777 egocentric actions. To quickly verify the effectiveness of the proposed method, We sampled a quarter of the data from the training set for model training. For all modalities, the backbone is Video Swin-T [15] from Omnivore [7] pretrained on ImageNet [5], Kinetics-400 [3], and SUN RGB-D [19] datasets. The hyper-parameter λ is set to 0.1. For each video clip, we sample 8 frames for training. We train the model for 60 epochs. Except for the aforementioned settings, the rest of the configurations are the same as those in Omnivore [7]. All experiments are performed on two GeForce RTX 3090 GPUs.

3.2. Results and Analysis

Effectiveness of the proposed method. Experiment results are shown in Tab. 1. We should focus our attention on three points. First, all of our proposed query-based fusion architecture, attention allocating loss, and attention reallocating strategy can bring progress in recognition performance. The overall performance boosting is 4.39%. Second, compared to the verb classification task, our proposed query-based fusion architecture shows a more noticeable improvement in the noun classification task. This could be due to that noun classification is more challenging than verb classification. The noun classification task is influenced to a greater extent when two tasks share the same feature space. Third, though the performance improvement of our proposed attention reallocating strategy is not very significant, it should be noted that it is a train-free method, it can be directly applied to more pretrained models which contain our proposed query-based fusion architecture with almost no additional computational cost.

Comparison of existing fusion methods. To further illustrate the superiority of the proposed query-based fusion architecture, we perform experiments with more fusion methods, including summation, gating fusion [1], and mid-level fusion [12], they are all widely-used fusion methods in multimodal learning. Results are shown in Tab. 2, our method surpasses all of them by at least 2.5%, this could because these previous methods still just output a shared feature representation for all tasks. These results further demonstrate the importance of special task tokens.

Applying balanced multimodal learning method. We also show the results of applying the balanced multimodal learning method OGM [17]. OGM is a representative balanced multimodal learning method, it obtains a discrepancy ratio from the final classification layer and alleviates modality imbalance via modulating backward gradients of different modalities. The fusion method is set to concatenation, as concatenation is the default fusion method in OGM. We both show the results of obtaining the discrepancy ratio from verb and noun classification layer, as in Tab. 3. OGM can only achieve very limited performance improvements, indicating that OGM is not suitable for addressing the issue of modality imbalance in the MIMO problem.

4. Conclusion

In this paper, we view the egocentric action recognition task as a MIMO problem. We point out the challenges of task conflict and task-aware modality bias. A query-based fusion architecture and an attention modulation strategy are proposed to tackle these challenges, respectively. Experiment results show that the recognition performance can be significantly enhanced by solving these challenges. We hope our work can bring further inspiration in the area of multimodal learning and egocentric video understanding. In the future, we will perform large-scale experiments and perform our proposed method on more egocentric datasets, to further evaluate the effectiveness of our proposed method.

5. Acknowledgments

This research was supported by National Natural Science Foundation of China (NO.62106272), the Young Elite Scientists Sponsorship Program by CAST (2021QNRC001), in part by the Research Funds of Renmin University of China (NO. 21XNLG17) and Public Computing Cloud, Renmin University of China.

References

[1] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[4] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23, 2022.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[6] Antonino Furnari and Giovanni Maria Farinella. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6252–6261, 2019.

[7] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16102–16112, 2022.

[8] Yifei Huang, Minjie Cai, Zhenqiang Li, Feng Lu, and Yoichi Sato. Mutual context network for jointly estimating egocentric gaze and action. *IEEE Transactions on Image Processing*, 29:7795–7806, 2020.

[9] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020.

[10] Ya Jing, Xuelin Zhu, Xingbin Liu, Qie Sima, Taozheng Yang, Yunhai Feng, and Tao Kong. Exploring visual pre-training for robot manipulation: Datasets, models and methods. *arXiv preprint arXiv:2308.03620*, 2023.

[11] Evangelos Kazakos, Jaesung Huh, Arsha Nagrani, Andrew Zisserman, and Dima Damen. With a little help from my temporal context: Multimodal egocentric action recognition. *arXiv preprint arXiv:2111.01024*, 2021.

[12] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5492–5501, 2019.

[13] Xuhua Huang Ye Yuan Xingyu Liu and Qichen Fu Kris M Kitani. Egoaugment: Cmu-klab submission to the epic-kitchens action recognition 2021 challenge.

[14] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21013–21022, 2022.

[15] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022.

[16] Kranti Parida, Neeraj Matiyali, Tanaya Guha, and Gaurav Sharma. Coordinated joint multimodal embeddings for generalized audio-visual zero-shot classification and retrieval of videos. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3251–3260, 2020.

[17] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8238–8247, 2022.

[18] Burr Settles and Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In *proceedings of the 2008 conference on empirical methods in natural language processing*, pages 1070–1079, 2008.

[19] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015.

[20] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12695–12705, 2020.

[21] Ruize Xu, Ruoxuan Feng, Shi-Xiong Zhang, and Di Hu. Mmcosine: Multi-modal cosine loss towards balanced audio-visual fine-grained learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.