3M-Game: Multi-Modal Multi-Task Multi-Teacher Learning for Game Event Detection (Student Abstract)

Thye Shan Ng[1], **Feiqi Cao**[2], **Soyeon Caren Han**[1,2]
[1]University of Melbourne
[2]University of Sydney
shan.ng@student.unimelb.edu.au, fcao0492@uni.sydney.edu.au, caren.han@unimelb.edu.au

Abstract

Esports has rapidly emerged as a global phenomenon with an ever-expanding audience on livestream platforms. However, due to the complex nature of the game, it becomes challenging for newcomers to comprehend the gaming situation. This research introduces a 3M-Game that integrates multi-modal (MM) information from the livestream platform, including chat and livestream, to uncover the event. While conventional MM models typically prioritise aligning MM data through concurrent training towards a unified objective, our framework leverages multiple independent teachers trained on different tasks to accomplish game event detection. The results show the effectiveness of the proposed framework. The code and appendix are in https://github.com/adlnlp/3m_game.

Introduction

The advent of online streaming platforms[1] gave rise to a new era for gaming channels, providing audiences with the opportunity to spectate gaming events in real time. However, it is challenging for newcomers to understand the overall game situation as the commentators' speech is often delivered rapidly, and the chat function accompanying the live streaming is typically filled with numerous misspellings and abbreviations, further complicating the understanding of the event. To help the audience better understand the situation, previous research has explored game situation understanding using various deep learning models, ranging from seq2seq (Ishigaki et al. 2021) to Transformers (Wang and Yoshinaga 2022). However these approaches typically rely on data sources extracted solely from the game itself, utilising a single modality and ignoring the rich information that could be provided by other modalities, such as audio tone and chat emotions. Some efforts have been made for the comprehension of chat contents (Weld et al. 2021; Han et al. 2021; Weld et al. 2023), but there remains a gap in the literature concerning the integration of multiple aspects from the livestream content and chat platforms together to establish a comprehensive understanding of the overall gaming situation. To address this, we propose a 3M framework that combines multiple modalities, drawing from both the game and audience reactions on the livestream platform to create a more robust game event detection system. More specifically, we focus on three modalities: *audio* from the gameplay, *transcript* from the commentators, and *chat* information, inspired by several multi-modal and multi-aspect knowledge distillation frameworks(Cabral et al. 2024; Li et al. 2024; Ding et al. 2024). Each modality is then fine-tuned using specialised teacher models to their own respective tasks, with their collective knowledge distilled into a unified student model designed for effective game event detection.

[1]For example, Twitch. https://twitch.tv/

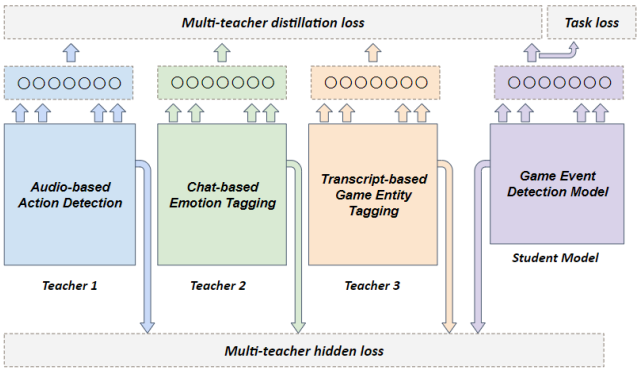

Figure 1: Architecture of proposed 3M-GAME Framework

Methodology

This paper utilises the Game-MUG dataset (Zhang et al. 2024), derived from LoL matches broadcasted on Twitch and YouTube. The dataset includes game match logs, audio, and textual discussions. Using OpenAI's Whisper model (Radford et al. 2023), we segment the continuous stream into distinct chunks by transcribing the raw audio into transcripts. This process yields 28,145 data instances, each containing all three modalities. These instances are split into a 95/5 train/test set. Our framework employs three teachers: audio (AST), chat (XLM-RoBERTa), and transcript (RoBERTa), each fine-tuned for their respective tasks with cross-entropy loss. Our proposed 3M-GAME framework (shown in Figure 1) is for multimodal multitask multiteacher learning for game event detection. We use three loss functions for knowledge distillation: multi-teacher hidden loss,

multi-teacher distillation loss and task-specific loss.

The multi-teacher hidden loss $\mathcal{L}_{MT-Hid}$ transfers knowledge from the hidden states of N fine-tuned teacher models, each with K Transformer layers, to a smaller student model with M layers. We define $\mathcal{L}_{MT-Hid}$ as the sum of two components: $\mathcal{L}_{i2i}$, representing the alignment loss between each student layer and its corresponding teacher layers, and $\mathcal{L}_{M2j}$, representing the alignment loss between the student's final student layer and the remaining teacher layers:

$$\mathcal{L}_{i2i} = \sum_{n=1}^{N} \sum_{i=1}^{M-1} \text{MSE}(H_i^S, W_{ni} H_{ni}^T) \quad (1)$$

$$\mathcal{L}_{M2j} = \sum_{n=1}^{N} \sum_{j=M}^{K} \text{MSE}(H_M^S, W_{nj} H_{nj}^T) \quad (2)$$

In $\mathcal{L}_{i2i}$, each ith student layer H_i^S is aligned with the corresponding ith teacher layer H_{ni}^T for each teacher n. In $\mathcal{L}_{M2j}$, the final student layer H_M^S is aligned with the remaining teacher layers H_{nj}^T for $j = M, ..., K$ using a trainable transformation matrix $W_{n.}$. Mean squared error (MSE) is applied to minimise the discrepancy between the teacher and student hidden representations, while the teacher layers are frozen during distillation to retain their task-specific knowledge.

Next, the goal of the multi-teacher distillation loss $\mathcal{L}_{MT-Dis}$ is to transfer knowledge from the soft labels of multiple teacher models to the student model. We replace the teacher models' prediction layers with trainable output layers matching the student's output layer. Since each teacher is fine-tuned for specific game expertise, exact game classification is not expected. Instead, the focus is on distilling modality-specific information from the teacher models to the student. Thus, the loss defined as:

$$\mathcal{L}_{MT-Dis} = \sum_{i=1}^{N} \frac{\text{CE}(y_i, y_s)}{1 + \text{CE}(y, y_i)} \quad (3)$$

where the output logits of the teacher model are compared with the student model, as well as the ground truth label via cross entropy.

Finally, we incorporate ground truth labels to compute the task-specific loss $L_{Task} = \text{CE}(y, y_s)$, based on the predictions of the student model for game event detection using transcripts as the input for the student model. The final loss function $\mathcal{L}$ is the summation of multi-teacher hidden loss, multi-teacher distillation loss and task-specific loss, which is formulated as $\mathcal{L} = \mathcal{L}_{MT-Hid} + \mathcal{L}_{MT-Dis} + \mathcal{L}_{Task}$. After knowledge distillation, only the student model is used for game event detection. Although the input is based on transcripts, the hidden layers are enriched with multimodal information.

Experiment and Results

In our empirical study, we use K=12 teacher layers and M=8 RoBERTa layers for the student model. The learning rate for fine-tuning the teacher is (1e-5, 1e-7), while for 3M knowledge distillation, it is (1e-4, 1e-7) using a cyclical approach. We use the AdamW optimizer with a dropout rate of 0.1 for all models. The implementation is done in PyTorch and HuggingFace, and we evaluate performance using Precision-*score*, comparing against the baseline Game-MUG BERT$_{\text{BASE}}$ model (Zhang et al. 2024).

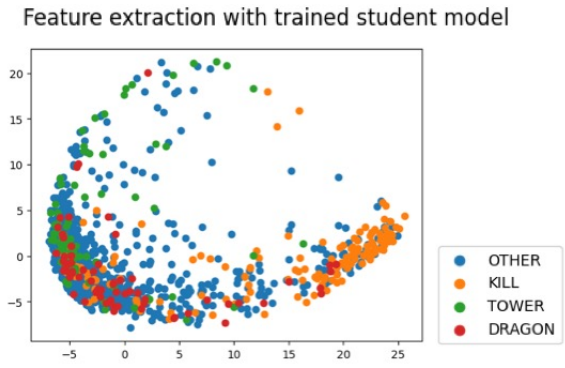

Figure 2: PCA of student model after knowledge distillation

Models	All	Mains Only
(Zhang et al. 2024)	0.340	0. 185
3M-Game (All)	**0.495**	**0. 638**
3M-Game ($\times$ Audio)	0.483	-
3M-Game ($\times$ Chat)	0.325	-
3M-Game ($\times$ Transcript)	0.447	-

Table 1: Precision comparison with Baseline from (Zhang et al. 2024) on Game-MUG (All and Main events only)

As shown in Table 1, the results indicate that the 3M-Game framework outperforms the baseline model in terms of precision (average macro). There are four types of game events, three main events and other: 'KILL', 'TOWER', 'DRAGON', and 'OTHER'. When considering all modalities, 3M-Game achieved the highest performance with scores of 0.495 and 0.638 for the datasets including and excluding the 'OTHER' game event. When certain modalities are excluded, the performance of 3M-Game slightly drops, indicating the importance of contribution from each teacher models. Overall, the results demonstrate the superior performance of baseline 3M-Game framework, particularly when all modalities are utilised. We then analyse game event predictions by the trained student model via its embedding outputs, as shown in Figure 2. The PCA plot shows distinct clustering of categories, with 'KILL', 'TOWER' and 'DRAGON' forming tighter groups, indicating the model's ability to differentiate these categories post-distillation.

Conclusion

This paper introduces the Multi-teacher Multi-task Multi-teacher framework named 3M-Game for game event detection, which imbues the student model with game knowledge from multiple teacher game expertise models via knowledge distillation. Our study showcases the capabilities of 3M-Game, leveraging diverse game expertise for improved performance.

Acknowledgements

This study was supported by funding from the Google Award for Inclusion Research Program (G222897). We thank David Chung and Zhihao Zhang for fruitful suggestions and discussions.

References

Cabral, R. C.; Luo, S.; Poon, J.; and Han, S. C. 2024. 3M-Health: Multimodal Multi-Teacher Knowledge Distillation for Mental Health Detection. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 152–162.

Ding, Y.; Vaiani, L.; Han, C.; Lee, J.; Garza, P.; Poon, J.; and Cagliero, L. 2024. M3-VRD: Multimodal Multi-task Multi-teacher Visually-Rich Form Document Understanding. *arXiv preprint arXiv:2402.17983*.

Han, S. C.; Long, S.; Li, H.; Weld, H.; and Poon, J. 2021. Bi-directional joint neural networks for intent classification and slot filling. In *22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021*, 3931–3935. International Speech Communication Association.

Ishigaki, T.; Topic, G.; Hamazono, Y.; Noji, H.; Kobayashi, I.; Miyao, Y.; and Takamura, H. 2021. Generating Racing Game Commentary from Vision, Language, and Structured Data. In Belz, A.; Fan, A.; Reiter, E.; and Sripada, Y., eds., *Proceedings of the 14th International Conference on Natural Language Generation*, 103–113. Aberdeen, Scotland, UK: Association for Computational Linguistics.

Li, Y.; Kim, S.-E.; Park, S.-B.; and Han, S. C. 2024. MIDAS: Multi-level Intent, Domain, And Slot Knowledge Distillation for Multi-turn NLU. *arXiv preprint arXiv:2408.08144*.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, 28492–28518. PMLR.

Wang, Z.; and Yoshinaga, N. 2022. Esports Data-to-commentary Generation on Large-scale Data-to-text Dataset.

Weld, H.; Hu, S.; Long, S.; Poon, J.; and Han, S. C. 2023. Tri-level Joint Natural Language Understanding for Multi-turn Conversational Datasets. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2023, 700–704.

Weld, H.; Huang, G.; Lee, J.; Zhang, T.; Wang, K.; Guo, X.; Long, S.; Poon, J.; and Han, C. 2021. CONDA: a CONtextual Dual-Annotated dataset for in-game toxicity understanding and detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2406–2416.

Zhang, Z.; Cao, F.; Mo, Y.; Zhang, Y.; Poon, J.; and Han, C. 2024. Game-MUG: Multimodal Oriented Game Situation Understanding and Commentary Generation Dataset. arXiv:2404.19175.