ARIA
APPLIED RESEARCH IN ACTION

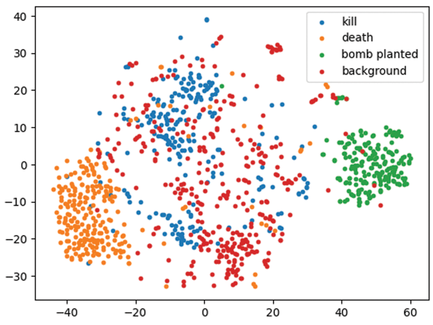
Automatic Gameplay Highlight Reel Generation Using Multimodal Learning

Detect and merge interesting events in first person shooting games using multimodal video transformer model to generate highlight clips

Vignesh Edithal

Chris McIntosh
ACADEMIC SUPERVISOR

Le Zhang
INDUSTRY SUPERVISOR



	X-CLIP	VideoMAE
CS:GO	95.2 %	85.2 %
Valorant	95.7 %	84.9 %

Event detection accuracy using X-CLP and VideoMAE models in popular first-person shooting games such as CS:GO and Valorant

PROJECT SUMMARY

In this work, we enable gamers to share their gaming experience on social media by automatically generating eye-catching highlight reels from their gameplay session. Gaming is a fast-growing segment of entertainment, particularly around E-sports and Twitch. Our automation will save time for gamers while increasing audience engagement. We approach the highlight generation problem by first identifying intervals in the video where interesting events occur and then concatenate them. We develop an in-house gameplay event detection dataset containing interesting events annotated by humans using VIA video annotator [1] [2]. We finetune a general-purpose video understanding model such as X-CLIP [3] using our dataset. Prompt engineering was performed to improve the classification performance of this multimodal model. Our evaluation shows that such a fine-tuned model can detect interesting events in first person shooting games from unseen gameplay footage with more than 90% accuracy. We also fine-tuned a video encoder model such as VideoMAE [4]. However, only the final classification layer could be finetuned while keeping the encoder weights frozen. This was done to prevent overfitting due to the small size of our dataset. To conclude, we show that natural language supervision leads to data efficient video recognition models.

REFERENCES

[1] A. Dutta, A. Gupta, and A. Zissermann. VGG image annotator (VIA). http://www.robots.ox.ac.uk/ vgg/software/via/, 2016. Version: 3.0.12, Accessed: 24 August 2023.

[2] Abhishek Dutta and Andrew Zisserman. The VIA annotation software for images, audio and video. In Proceedings of the 27th ACM International Conference on Multimedia, MM '19, New York, NY, USA, 2019. ACM.

[3] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition, 2022.

[4] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training, 2022.

AMD

Computer Science
UNIVERSITY OF TORONTO
Master of Science in Applied Computing