Great Minds Think Alike, Great Models Think All at Once: MAESTRO, A Unified Framework for Video-Audio Understanding & Reasoning

Tan, Felicia (School: Raffles Institution)
Low, Amy (School: Raffles Institution)

Multimodal contents' increasing complexity poses significant challenges for Vision-Language Models (VLMs), particularly in hateful video detection. Hateful videos are inherently multimodal as hatefulness often only emerges when visual and auditory cues are analysed together, making isolated modalities insufficient. Additionally, hatefulness often appears in only a few frames, rendering frame-by-frame processing inefficient. Existing VLMs struggle with weak auditory integration, static reasoning, and high computational costs. To overcome these, this study presents MAESTRO, a plug-and-play framework enhancing VLMs through three innovations: (1) Semantic Segmentation Mechanism: a BERT-based module that segments videos into semantic chunks, ensuring temporal alignment between audio and visual modalities; (2) Unified Modality Alignment, which maps video, non-speech, and speech-audio into a shared space, facilitating deeper multimodal interactions; and (3) Global-Local Reasoning Loop, dynamically refining analysis by integrating local details with global context, eliminating inefficient frame-by-frame processing. MAESTRO achieves state-of-the-art, attaining 93% F1-score on MultiHateClip for hateful video detection. Beyond this use case, it establishes new benchmarks in Video Question-Answering (VQA), achieving F1-scores of 82.0% on MSRVTT-QA, 86.9% on MSVD-QA, and 87.2% on ActivityNet-QA. These highlight its ability to enhance VLMs for a wide range of multimodal tasks while maintaining relative computational efficiency (2 T4 GPUs) to traditional VLMs. Being scalable and efficient, MAESTRO has broad applications like improving general VQA, enabling greater automation in multimodal content analysis, and enhancing AI-driven reasoning for complex video-language tasks.

Awards Won: