Team AZBYCX: Submission Details to EPIC-SOUNDS Audio-Based Interaction Detection

Siqi Yu   Yichen Lu   Shuo Li   Fang Liu   Xu Liu
School of Artificial Intelligence, Xidian University
24171213877@stu.xidian.edu.cn

Abstract

This report presents the solution of Team AZBYCX in the EPIC-SOUNDS Audio-Based Interaction Detection Challenge. The challenge aims to identify interaction behaviors between humans and the environment by processing audio signals from kitchen scenarios. This work draws inspiration from two open-source solutions: the CausalTAD configuration in the OpenTAD project (for temporal action detection) and the ActionFormer network architecture in the official EPIC-SOUNDS baseline system, followed by fusion and adaptation. On the EPIC-SOUNDS dataset, our method significantly improves the mean average precision (mAP), increasing the average mAP from 12.39% (baseline ActionFormer) to 14.87%. Experimental results show that the constructed model achieves both high precision and good temporal localization performance, providing a feasible path for audio-based action recognition research. The model ultimately ranked first on the leaderboard.

1. Introduction

1.1. Background and Task Definition

Audio-visual interaction detection is a crucial task in computer vision, aiming to identify and localize audio-related interaction behaviors from videos. The EPIC-Kitchens-100 dataset[1], a large-scale, untrimmed first-person vision dataset, captures activities in 45 kitchens across 4 global cities, totaling 100 hours of footage. Building upon this, as illustrated in Figure 1, the EPIC-SOUNDS dataset[2] further annotates audio-visual interactions, featuring 79k audio annotations with 223 unique free-form descriptions spanning 44 interaction categories. The objective of audio-visual interaction detection is to predict a set of audio-related interaction instances from a given video, including the start time, end time, and predicted action category for each instance.

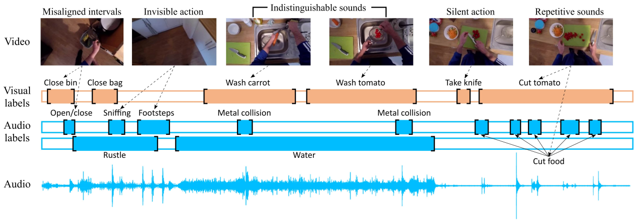

Figure 1. The EPIC-SOUNDS dataset.

1.2. Contributions

This report describes work centered around the tasks of the EPIC-SOUNDS Audio-Based Interaction Detection Challenge, which involves conducting in-depth exploration and innovation while fully drawing on existing research achievements. Specifically, we first adopted the publicly available EPIC-SOUNDS Baseline configuration[5], which provides a fundamental implementation framework and performance reference standard for audio interaction detection tasks. Meanwhile, we referenced the temporal modeling structure of CausalTAD in OpenTAD[4] and introduced it into our model design. Through its unique temporal causal modeling approach, CausalTAD can effectively capture the causal relationships of actions in the temporal dimension, providing strong support for the model to accurately identify and localize interaction behaviors in audio sequences.

2. Related Work

Audio interaction detection can be regarded as a special case of Temporal Action Detection (TAD), whose core is to localize and classify action instances from long videos. TAD methods can be divided into feature-based methods and end-to-end methods. Feature-based methods extract video features through pre-trained models and then use detectors for action localization and classification. End-to-end methods jointly optimize video encoders and action detectors, garnering increasing attention in recent years. In temporal modeling, convolutional networks, graph networks, and recurrent networks are widely used to capture tempo-

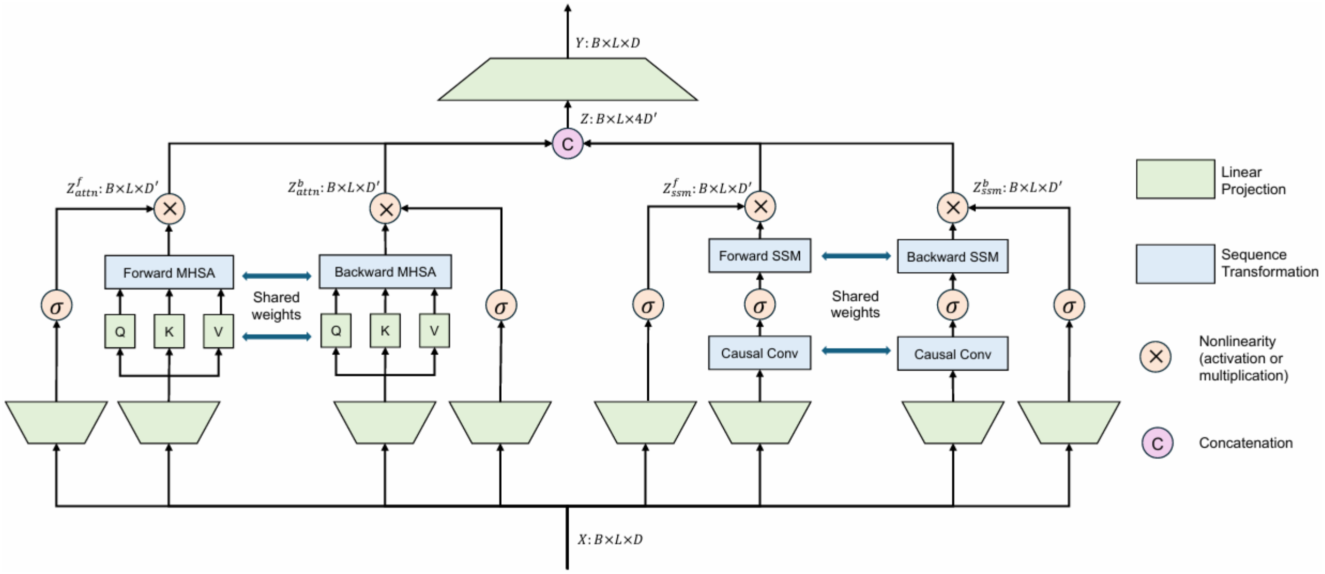

Figure 2. Hybrid Causal Block.

ral dependencies. However, these methods often struggle to handle long-range dependencies effectively. The emergence of Transformer has brought new hope for long-range temporal modeling, but traditional Transformer models symmetrically process past and future temporal contexts, ignoring the causality in action transitions. Recent studies have begun to explore causal modeling—for example, the causal Mamba module proposed in VideoMambaSuite enhances temporal causality by restricting the model to access only past or future contexts.

In audio interaction detection, the emergence of the EPIC-Sounds dataset has posed new challenges and opportunities for the field. Existing methods mainly rely on audio features or audio-visual fusion features for detection. However, how to effectively leverage temporal causality to improve detection performance remains an open question. Therefore, we adopted the CausalTAD model, which utilizes the temporal causality of actions by restricting the model to access only past or future contexts, thereby enhancing audio interaction detection performance.

3. Methodology

3.1. One-Stage Detection Framework

Our audio interaction detection framework is based on a one-stage detection method, using a pre-trained action recognition model as a video encoder to extract semantically rich video features. Untrimmed videos are divided into multiple short clips, and a sliding window approach is used to extract features from each clip independently. Video clips may overlap with each other according to the sliding window stride. Spatiotemporal average pooling is applied after the video backbone network to obtain dense video features for each clip.

3.2. Temporal Causal Modeling

To leverage temporal causality, we employ a hybrid causal block that combines causal attention and causal Mamba modules, as shown in Figure 2. The causal Mamba module is a structured state space model (SSM) that effectively captures long-range dependencies through parallel scanning and data-dependent SSM layers. However, the original Mamba module is only suitable for causal language tasks and lacks the ability to process future information common in visual tasks. Therefore, we adopted the bidirectional Mamba module proposed in VideoMambaSuite, which shares input projectors and SSM parameters in forward and backward scanning directions, outperforming traditional ViM models in performance.

The causal attention module aims to explicitly capture global long-range temporal dependencies and temporal causality. It shares a similar architectural design with the causal Mamba module, including gated projection and bidirectional multi-head self-attention (MHSA). Bidirectional MHSA follows the causal modeling in VideoMambaSuite, restricting attention contexts to only past or future information. This allows the model to more accurately simulate temporal causality, thereby improving audio interaction detection performance.

The hybrid causal block combines the causal Mamba module and causal attention module, leveraging their complementary strengths to achieve more accurate and context-aware audio interaction detection. We separately compute the outputs of the causal Mamba and causal attention mod-

Method	mAP@0.1(%)	mAP@0.2(%)	mAP@0.3(%)	mAP@0.4(%)	mAP@0.5(%)	Avg. mAP(%)
ActionFormer	16.17	14.36	12.59	10.41	8.44	12.39
CausalTAD	19.40	17.35	15.00	12.58	10.02	14.87

Table 1. Comparison of results between ActionFormer and CausalTAD on the test set.

ules, concatenate them, and then use a linear layer to reduce the channel dimension to match the input.

4. Experiments

4.1. Dataset and Metrics

We evaluated our method on the EPIC-SOUNDS dataset, which contains untrimmed first-person videos. We followed the standard training/validation/testing splits and reported performance accordingly. We used the mean average precision (mAP) at different intersection-over-union (IoU) thresholds as the evaluation metric.

4.2. Implementation Details

The method was implemented based on the OpenTAD codebase and trained on three RTX2080 GPUs. We used mixed-precision training and flash attention to accelerate the training process. To achieve stronger performance, we optimized hyperparameters of the detection head, including the number of feature pyramid layers, regression loss weights, input channel dropout probability, and training epochs. For audio interaction detection, we used the Auditory-SlowFast model[3] as the feature extractor—a two-stream convolutional network specialized for audio recognition that processes time-frequency spectra and was pre-trained on the EPIC-Kitchens action recognition task.

4.3. Comparison with Existing Methods

On the EPIC-SOUNDS dataset, our method significantly outperformed the baseline ActionFormer, with specific results shown in Table Tab. 1. The CausalTAD method achieved an average mAP of 14.87% on the test set, compared to 12.39% for the baseline ActionFormer. This result demonstrates the significant advantages of causal temporal modeling in audio interaction detection tasks.

5. Conclusion

Team AZBYCX proposed an audio interaction detection method based on causal attention mechanisms, effectively leveraging temporal causality to enhance representation capabilities through the combination of causal attention and causal Mamba modules. Experimental results on the EPIC-SOUNDS dataset show that the method significantly improves audio interaction detection performance, achieving a 2.48% increase in average mAP on the test set compared to the baseline, and ultimately ranking first on the leaderboard.

References

[1] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, , Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022. 1

[2] Jaesung Huh, Jacob Chalk, Evangelos Kazakos, Dima Damen, and Andrew Zisserman. EPIC-SOUNDS: A Large-Scale Dataset of Actions that Sound. In *IEEE International Conference on Acoustics, Speech, Signal Processing (ICASSP)*, 2023. 1

[3] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Slow-fast auditory streams for audio recognition. *CoRR*, abs/2103.03516, 2021. 3

[4] Shuming Liu, Chen Zhao, Fatimah Zohra, Mattia Soldan, Alejandro Pardo, Mengmeng Xu, Lama Alssum, Merey Ramazanova, Juan León Alcázar, Anthony Cioppa, Silvio Giancola, Carlos Hinojosa, and Bernard Ghanem. Opentad: A unified framework and comprehensive study of temporal action detection. *arXiv preprint arXiv:2502.20361*, 2025. 1

[5] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, pages 492–510, 2022. 1