Towards Unified, Explainable, and Robust Multisensory Perception

Yapeng Tian

Department of Computer Science, University of Texas at Dallas, USA
yapeng.tian@utdallas.edu

Abstract

Humans perceive surrounding scenes through multiple senses with multisensory integration. For example, hearing helps capture the spatial location of a racing car behind us; seeing peoples' talking faces can strengthen our perception of their speech. However, today's state-of-the-art scene understanding systems are usually designed to rely on a single audio or visual modality. Ignoring multisensory cooperation has become one of the key bottlenecks in creating intelligent systems with human-level perception capability, which impedes the real-world applications of existing scene understanding models. To address this limitation, my research has pioneered marrying computer vision with computer audition to create multimodal systems that can learn to understand audio and visual data. In particular, my current research focuses on asking and solving fundamental problems in a fresh research area: audio-visual scene understanding and strives to develop *unified*, *explainable*, and *robust* multisensory perception machines. The three themes are distinct yet interconnected, and all of them are essential for designing powerful and trustworthy perception systems. In my talk, I will give a brief overview about this new research area and then introduce my works in the three research thrusts.

Unified Multisensory Perception My research has provided some critical first steps in asking core problems in audio-visual scene understanding and proposing innovative solutions that can unify senses from sound and sight (Tian et al. 2018; Tian, Li, and Xu 2020; Mo and Tian 2022). In particular, I posed and tried to tackle a fundamental problem: *audio-visual video parsing* in (Tian, Li, and Xu 2020) that aims to group video segments and parse a video into different temporal audio, visual, and audio-visual events associated with semantic labels. This work pushed state-of-the-art video understanding to a whole new level by considering modality-aware video parsing.

Explainable Multisensory Perception Although the audio-visual models have achieved promising scene understanding performance by multisensory integration, a basic question: *how do the different modalities cooperate in perceiving scenes from sound and sight inside black-box models* remains unexplored. It is seemingly impossible to answer the question since there is no annotation denoting the individual contributions of the auditory or visual modality made to perceived scenes in any of the existing datasets– such a process is difficult to quantify without breakthroughs in Neurophysiology. Instead, I investigated the multisensory interpretability problem from a computational perspective and used audio-visual video captioning as a proxy (Tian et al. 2019), where I mined signals from audio and video and compete their associations to language descriptions. In the end, the trained model can disentangle the interplay of the two modalities and explain to what extent different modalities contribute to a particular predicted natural language sentence, and furthermore, to individual words in a sentence.

Robust Multisensory Perception My research has shown that robust auditory or visual perception can be achieved by integrating multisensory information. However, *whether these computational perception models still exhibit robustness under attacks*. Inspired by the auditory-visual illusion in human perception, I have presented the first systematic study on machines' multisensory integration under attacks (Tian and Xu 2021). The study reveals that an audio-visual-fused model would underperform a unimodal model when the likelihood of receiving compromised audio/visual/mixed signals changes. Thus, instead of only focusing on clean signals (the tip of the iceberg) as our research community currently does, we should reconsider audio-visual integration in the context of receiving possibly attacked signals.

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

References

Mo, S.; and Tian, Y. 2022. Multi-modal Grouping Network for Weakly-Supervised Audio-Visual Video Parsing. In *NeurIPS*.

Tian, Y.; Guan, C.; Justin, G.; Moore, M.; and Xu, C. 2019. Audio-Visual Interpretable and Controllable Video Captioning. In *CVPR Workshops*.

Tian, Y.; Li, D.; and Xu, C. 2020. Unified Multisensory Perception: Weakly-Supervised Audio-Visual Video Parsing. In *ECCV*.

Tian, Y.; Shi, J.; Li, B.; Duan, Z.; and Xu, C. 2018. Audio-Visual Event Localization in Unconstrained Videos. In *ECCV*.

Tian, Y.; and Xu, C. 2021. Can Audio-Visual Integration Strengthen Robustness Under Multimodal Attacks? In *CVPR*.