Audio Description Generation in the Era of LLMs and VLMs: A Review of Transferable Generative AI Technologies

Yingqiang Gao, Lukas Fischer, Alexa Lintner, Sarah Ebling

Audio descriptions (ADs) function as acoustic commentaries designed to assist blind persons and persons with visual impairments in accessing digital media content on television and in movies, among other settings. As an accessibility service typically provided by trained AD professionals, the generation of ADs demands significant human effort, making the process both time-consuming and costly. Recent advancements in natural language processing (NLP) and computer vision (CV), particularly in large language models (LLMs) and vision-language models (VLMs), have allowed for getting a step closer to automatic AD generation. This paper reviews the technologies pertinent to AD generation in the era of LLMs and VLMs: we discuss how state-of-the-art NLP and CV technologies can be applied to generate ADs and identify essential research.