World Academy of Science, Engineering and Technology
International Journal of Electrical and Computer Engineering
Vol:17, No:11, 2023

Audio-Visual Co-Data Processing Pipeline

Authors : Rita Chattopadhyay, Vivek Anand Thoutam

Abstract : Speech is the most acceptable means of communication where we can quickly exchange our feelings and thoughts. Quite often, people can communicate orally but cannot interact or work with computers or devices. It's easy and quick to give speech commands than typing commands to computers. In the same way, it's easy listening to audio played from a device than extract output from computers or devices. Especially with Robotics being an emerging market with applications in warehouses, the hospitality industry, consumer electronics, assistive technology, etc., speech-based human-machine interaction is emerging as a lucrative feature for robot manufacturers. Considering this factor, the objective of this paper is to design the "Audio-Visual Co-Data Processing Pipeline." This pipeline is an integrated version of Automatic speech recognition, a Natural language model for text understanding, object detection, and text-to-speech modules. There are many Deep Learning models for each type of the modules mentioned above, but OpenVINO Model Zoo models are used because the OpenVINO toolkit covers both computer vision and non-computer vision workloads across Intel hardware and maximizes performance, and accelerates application development. A speech command is given as input that has information about target objects to be detected and start and end times to extract the required interval from the video. Speech is converted to text using the Automatic speech recognition QuartzNet model. The summary is extracted from text using a natural language model Generative Pre-Trained Transformer-3 (GPT-3). Based on the summary, essential frames from the video are extracted, and the You Only Look Once (YOLO) object detection model detects You Only Look Once (YOLO) objects on these extracted frames. Frame numbers that have target objects (specified objects in the speech command) are saved as text. Finally, this text (frame numbers) is converted to speech using text to speech model and will be played from the device. This project is developed for 80 You Only Look Once (YOLO) labels, and the user can extract frames based on only one or two target labels. This pipeline can be extended for more than two target labels easily by making appropriate changes in the object detection module. This project is developed for four different speech command formats by including sample examples in the prompt used by Generative Pre-Trained Transformer-3 (GPT-3) model. Based on user preference, one can come up with a new speech command format by including some examples of the respective format in the prompt used by the Generative Pre-Trained Transformer-3 (GPT-3) model. This pipeline can be used in many projects like human-machine interface, human-robot interaction, and surveillance through speech commands. All object detection projects can be upgraded using this pipeline so that one can give speech commands and output is played from the device.

Keywords : OpenVINO, automatic speech recognition, natural language processing, object detection, text to speech
Conference Title : ICHCITA 2023 : International Conference on Human Computer Interaction Technologies and Applications
Conference Location : New York, United States
Conference Dates : November 06-07, 2023