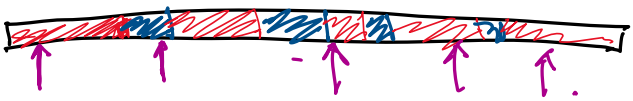
Video Description Generation.

Saturday, July 8, 2023 6:21 PM



All the timeline of the video is sorted into 2 portions

dialogue and pauses

In a video, we define the "most important" parts as keyframes.

Implementation

At each keyframe, we extract the image portion of the video.

Then we grab the caption of the nearest previous dialogue.

Using the caption of dialogue + visual data of keyframe

We can generate an "audio description" of what is happening

How to make sure the new "audio description" doesn't interfere with the original audio?

→ Intersperse. By replacing the pauses with "audio description"

Process

① Extract keyframe timestamps; extract screenshot data

② Extract captions at each keyframe.

③ Gen. description from keyframe + caption.

④ Find nearest "pause" and the duration.
Based on pause duration, truncate the description

⑤ Convert description to speech mp3
Speed up mp3 to fit in "pause" duration

⑥ Replace the pause in the original

(6) Replace the pause in the original audio with the audio description

(7) After processing entire video, in-browser replace video audio with processed .mp3.

Tech

- VideoIndexer.AI (Azure)
- Text → Speech (Azure)
- Azure OpenAI
- Azure Computer Vision,

Parts

(1) Extract images

(2)