TALKING TO ME (TTM)

Team-02 CudaOutOfMemory

National Taiwan University

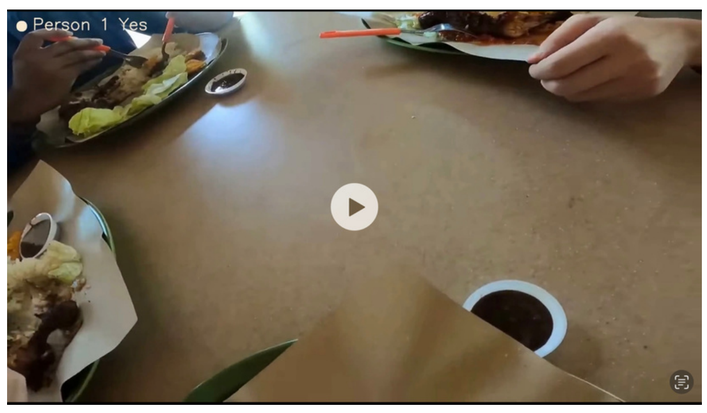
R11922A10 資工碩一 林昱辰
R11922A04 資工碩一 陳奕嘉
R10922132 資工碩二 吳峻銘
R11922069 資工碩一 謝欣玉

Introduction

We propose a novel model, **WARMEST**(**wa**vefo**rm**-R**es**Ne**t**), which ensembles WavLM and ResNet152 model for video classification. Given videos of wearer's view, our model identifies whether each person showing up in the video is talking to the wearer in the specific time interval. In this task, we need to consider both visual and audio information.

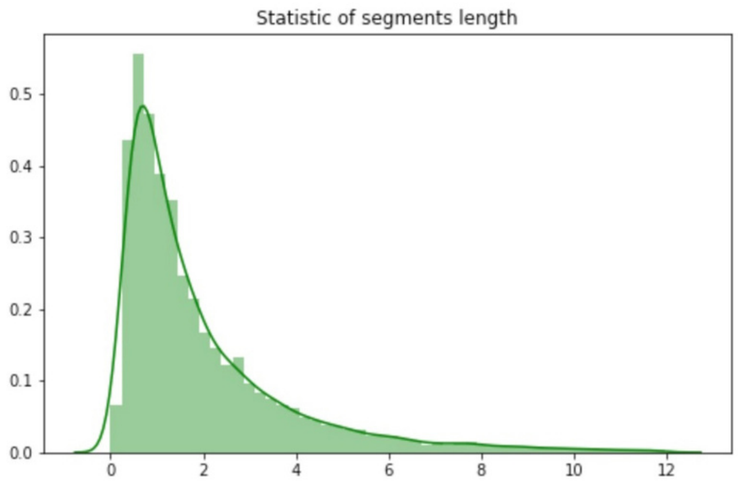
Data Analysis and Preprocessing

- Visual data



We found that sometimes the speaker is **NOT** in the frame, but he/she **is** talking to the wearer.

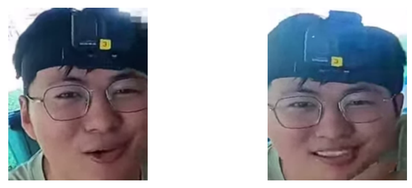
- Audio data



About 96% of the segments are less than 8 seconds.

Preprocessing

- Visual

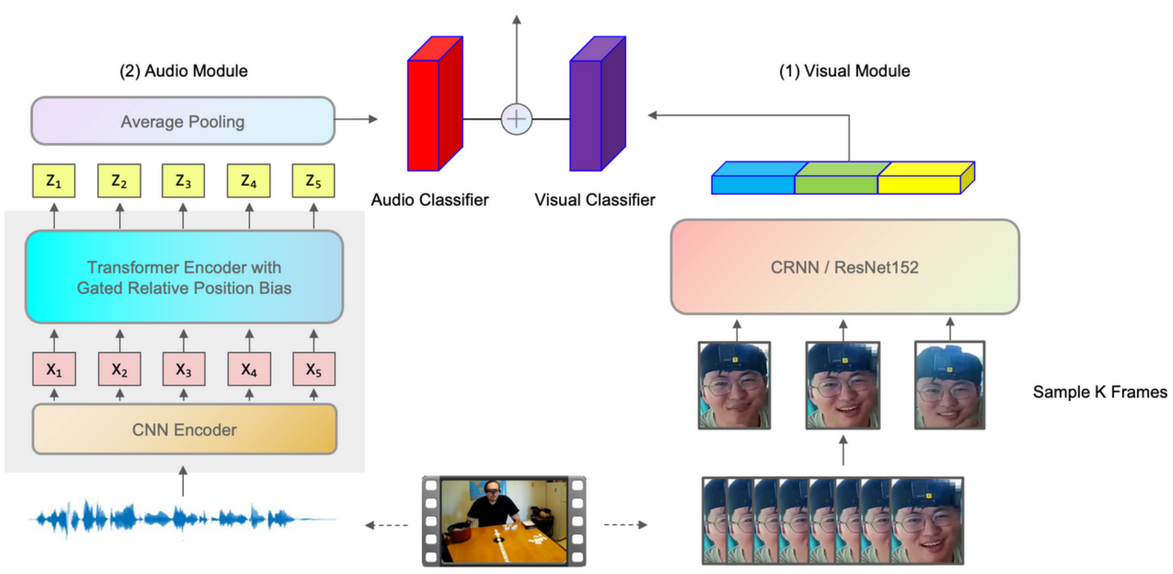


  - Sample 5 **frames** in the segment
  - Crop the face according to bbox
  - Resize to 64x64
- Audio
  - Pad or truncate to 8 seconds

Others

- When no bbox is available, we will use blank images for inference

Methodology



- **Model Overview**

We use different modules to encode audio and image data. Then, we fusion two classifier outputs as our final output.

- **Visual Module**

We sample k frames as CRNN/ResNet152 input and concatenate each frame hidden state together to get speaker facial feature.

- **Audio Module**

We use WavLM pre-trained model to get speakers audio interaction feature .

- **Classifier Fusion Module**

$$\text{final classifier head} = \alpha * \text{audio classifier head} + (1 - \alpha) * \text{visual classifier head}$$

Experiment

- Compare **different number of frames** used for training CRNN

Frames	5	100
Accuracy	**0.60524**	0.56052

According to the table, we can see less frames achieve better performance. It is probably caused by **noisy frames**.

- Compare **different visual encoder**

Model	CRNN	ResNet152
Accuracy	0.60524	**0.6208**

- Compare **different length of audio** for WavLM audio encoder

Second	5	8	10
Accuracy	0.68	**0.70**	0.69

- **Ensembling** different visual encoder(CRNN, ResNet152) with audio encoder(WavLM)

Model	WavLM-CRNN	WavLM-ResNet152
Accuracy	0.69956	**0.71803**

Finally we found that **ensembling WavLM & ResNet152 embedding with weight 7:3 can achieve the best performance.**

Conclusion

1. The quality of image frames greatly influence the performance of visual encoder. In the future, researchers can further study on how to select **key-frames** to collect more decisive visual features.
2. Sometimes the talking-to-me speaker is out of wearer's view. In this case, there won't be any visual features, and model will make prediction totally according to audio features. Here's the **powerful audio module** that comes into play.