DLCV 2022 Final Challenge 1 - Talking to Me

Team name : WI11yT34m

Tun-Hsiang, Chou
R11922163

Ding-Jiun, Huang
B08902028

Szu-Yen, Kuo
B08508002

Po-Heng, Chen
R11922044

Introduction

The task is to identify whether the visible face in a video clip is talking to the camera wearer. There are three difficult parts in this task.

1. Audio/Video modal
   - Different model design for two modalities' data
     - Video : RNN, 3D-CNN, Transformer
     - Audio : MFCC, Spectrogram, RNN, Transformer
2. Align different modalities' length
   - Naive : Do not align
   - Change MFCC window size to make output length same as frame number
3. Align video content and audio
   - Fuse two modalities' representation and do self-attention
   - Use cross-attention to align two modalities' representation

Related Works

- Video classification :
  - Classify the full video or each frames in it to certain labels.
  - Dataset : Kinetics, Action Recognition
  - Model : 3D-CNN, TimeSformer, ViViT
- Audio classification :
  - Classify the audio to certain labels.
  - Dataset : Command Recognition, Emotion recognition
  - Model : Hubert, Wav2Vec2.0
- Active Speaker Detection (ASD) :
  - Detects whether a person is speaking in the video clip. Each frame is labeled as True or False.
  - Dataset : AVA-ActiveSpeaker
  - Model : TalkNet-ASD, SPELL
- Audio-visual speech recognition:
  - Use both audio and lip video to do speech recognition.
  - Dataset : VoxCeleb2
  - Model : Av-Hubert

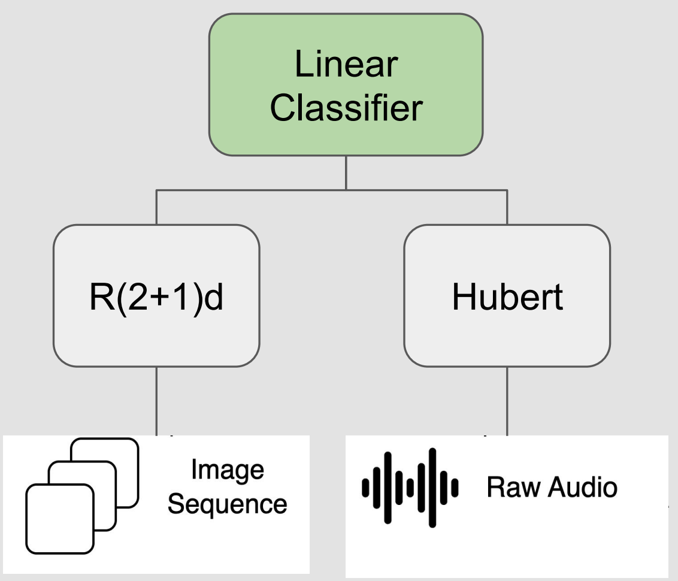
Data Preprocess

- Video:
  - Crop each frames to only ROI
  - Remove empty frames or replace empty frames with mean value
- Audio:
  - Train with raw waveform
  - Waveform to mfcc: adjust and select the window size that can match output length to frame number
- Data augmentation
  - Video: video-wise random horizontal flip, random rotation, add random noise...
  - Audio: add gaussian noise, random shifting and stretching

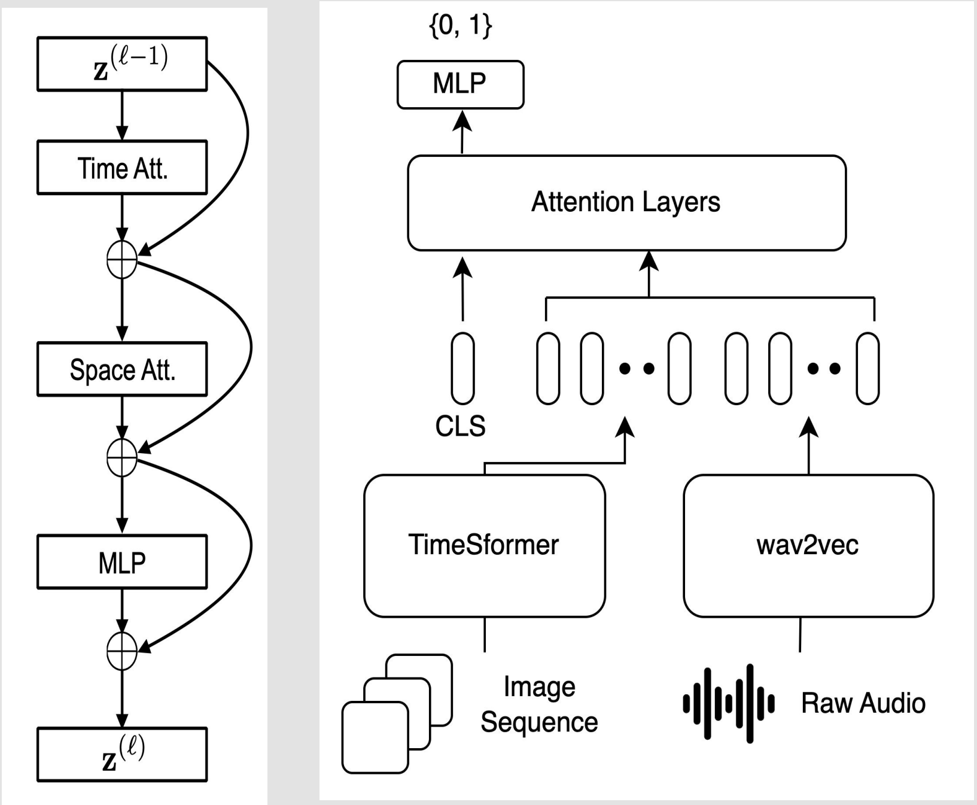
Model Design

Late fusion

- **Serparete encoder (R(2+1)d & Hubert)**
  - Use separate encoder models for video and audio. Use only one vector as representation for each modality.
  - Concat two vector and do linear classification



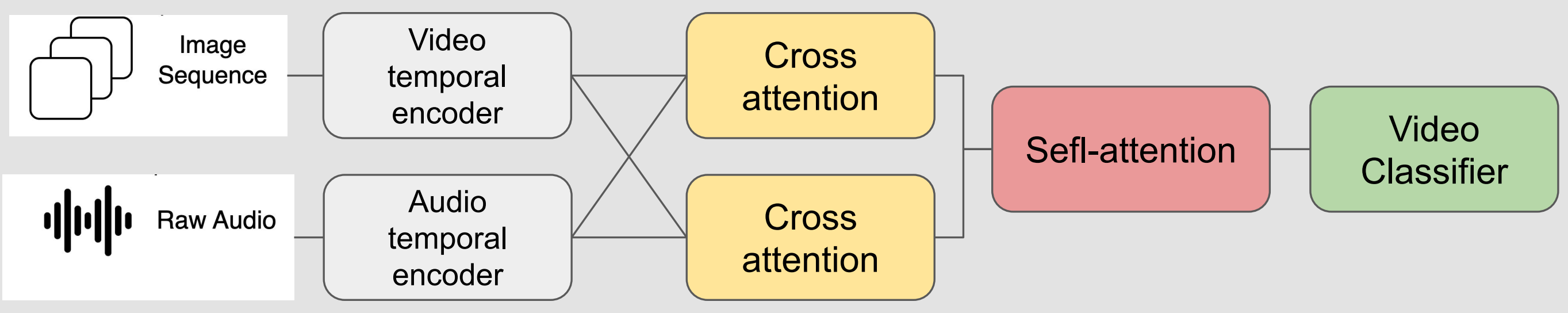
Serparete encoder
(R(2+1)d & Hubert)



Serparete encoder and self-attention
(TimeSformer & Wav2vec)

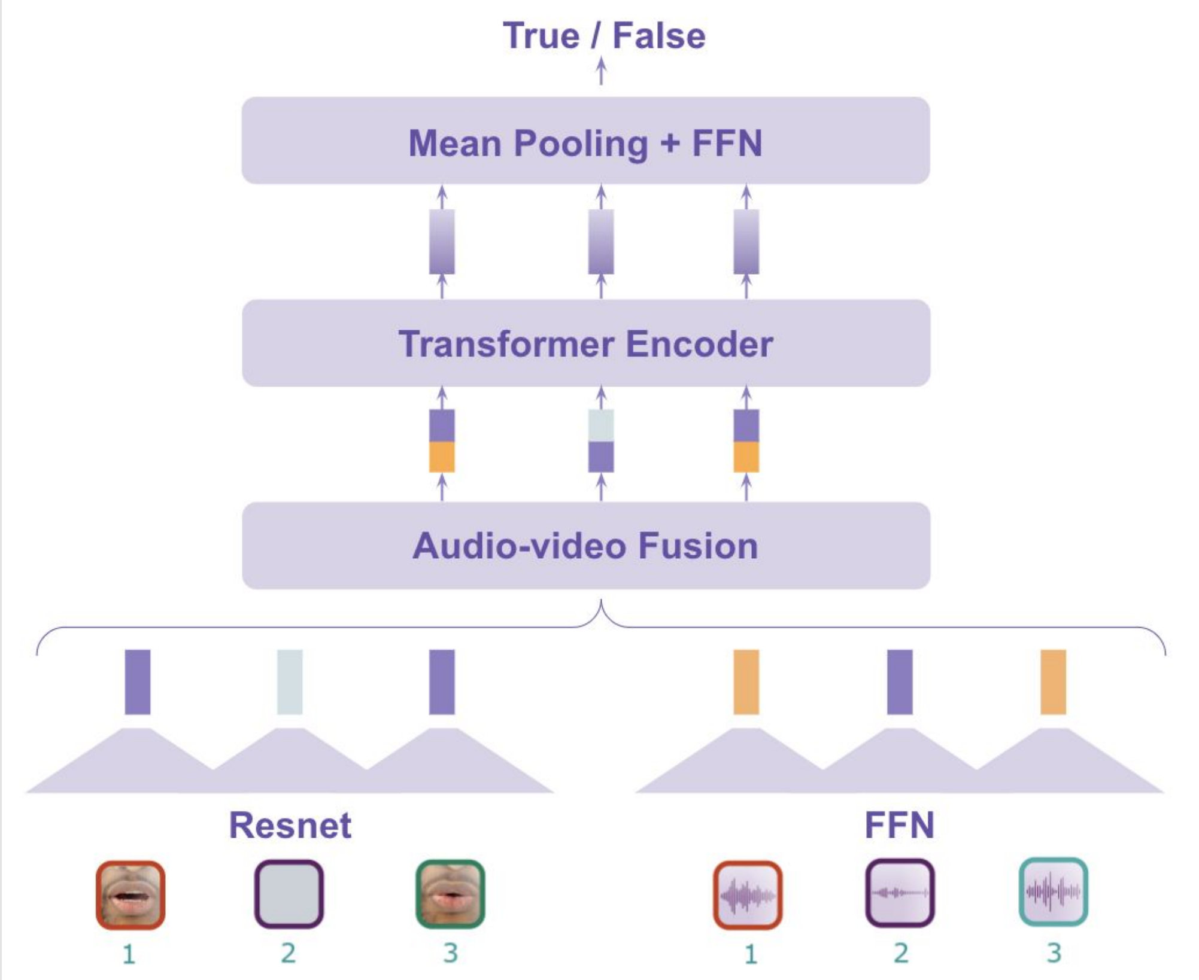
Middle fusion

- **Serparete encoder and self-attention (TimeSformer & Wav2vec)**
  - Audio and video features are first extracted by pretrained TimeSformer and wav2vec respectively.
  - The tokenized output features, along with a CLS token, are fed into multihead attention blocks.
  - Input from two modals aggregate after a respective extraction process.
  - We use the CLS token for the decision.
- **Serparete encoder and cross-attention**
  - Audio and video features are first extracted by temporal encoder trained from scratch
  - Align temporal dimension of two modality and do cross-attention
  - Concat cross-attention output and fed to self-attention layer
  - Use mean pool of self-attention output for classification



Serparete encoder
and cross-attention

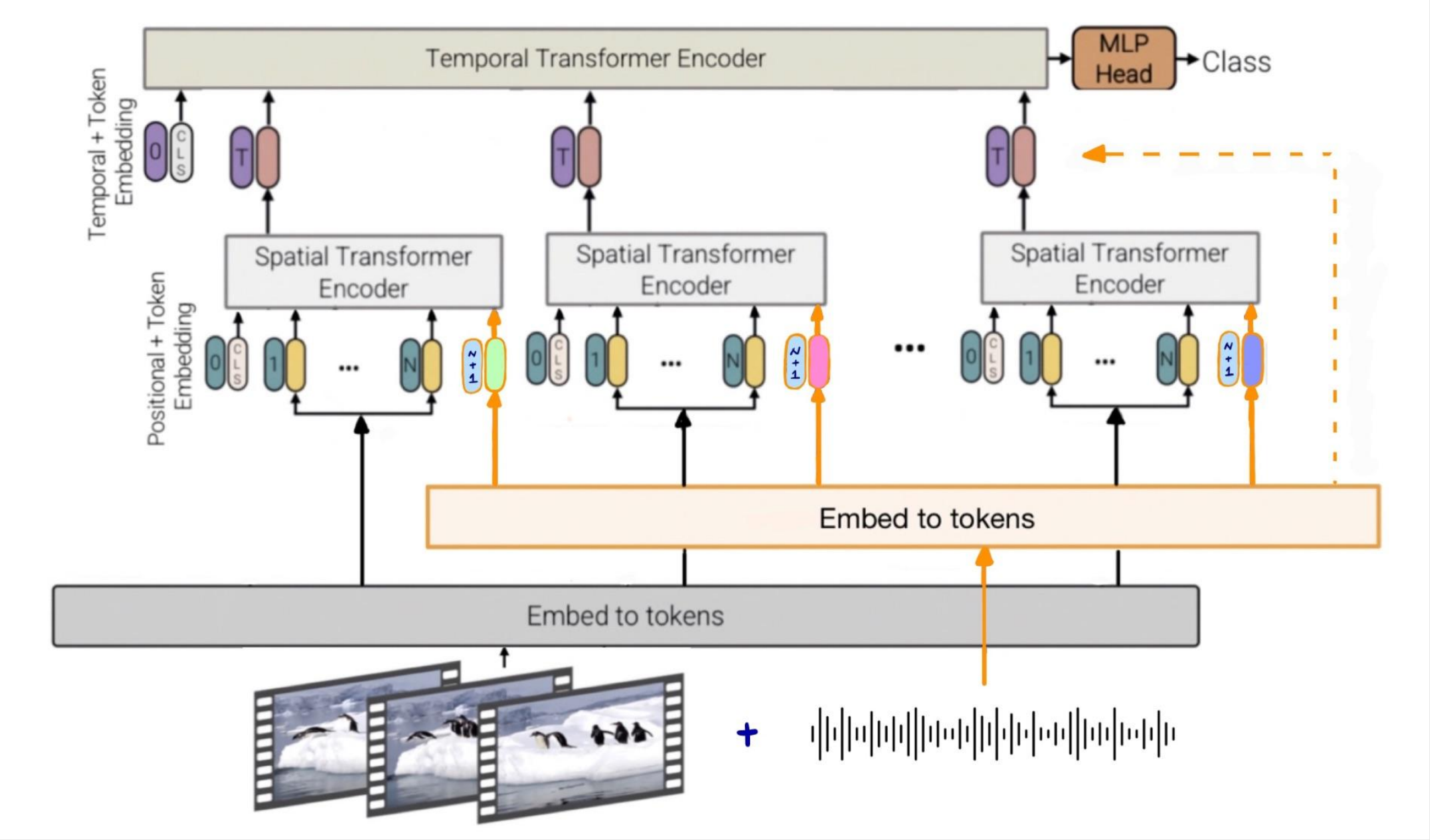
- Av-Hubert
  - AV-HuBERT is a self-supervised representation learning framework for audio-visual speech, which masks multi-stream video input and predicts automatically discovered and iteratively refined multimodal hidden units.
  - Av-Hubert take video and logfBank of audio as input, which are align in time dimension
  - Use mean pooling of Av-Hubert last layer output for classification



Av-Hubert

Early fusion

- **ViViT-based Video and Audio Transformer Model**
  - ViViT is a model for video based on ViT, which cut each frames into patches.
  - To add audio information, transform audio waveform to mfcc and align the length to frame number
  - After projecting to the same hidden dimension, concatenate audio information as an audio token to each frame of the video.


Temporal Transformer Encoder
MLP Head
Class
Temporal + Token Embedding
Positional + Token Embedding
Spatial Transformer Encoder
Spatial Transformer Encoder
Spatial Transformer Encoder
Embed to tokens
Embed to tokens
+

ViViT-based Video and Audio Transformer Model

Result

	Single Modal	Late Fusion	Middle Fusion			Early Fusion
Model	R(2+1)d	R(2+1)d+ Hubert	Wav2Vec + TimeSFormer + Attention	Temporal encoder + cross-attention	Av-Hubert	Vivit-base video&audio transformer
Test Acc						

Conclusion

- Pretrained weight can help us capture information from data easily
  - Train on only video with pretrained weight has quite great performance
- When we align audio and video data, model use two modality information on each temporal dimension simultaneously, which can improve model performance
  - Model with aligned data has better performance
- If model can learn the relationship between video and audio (ex. Relationship between lip movement and speech), it can make prediction much more accurately
  - Av-Hubert has best performance among all the model we tried

Reference

- R(2+1)d : https://arxiv.org/abs/1711.11248v3
- VIvit : https://arxiv.org/abs/2103.15691
- TImeSformer : https://arxiv.org/abs/2102.05095
- TalkNet-ASD : https://arxiv.org/abs/2107.06592
- Av-Hubert : https://arxiv.org/abs/2201.02184
- Hubert : https://arxiv.org/abs/2106.07447
- Wav2vec2.0 : https://arxiv.org/abs/2006.11477