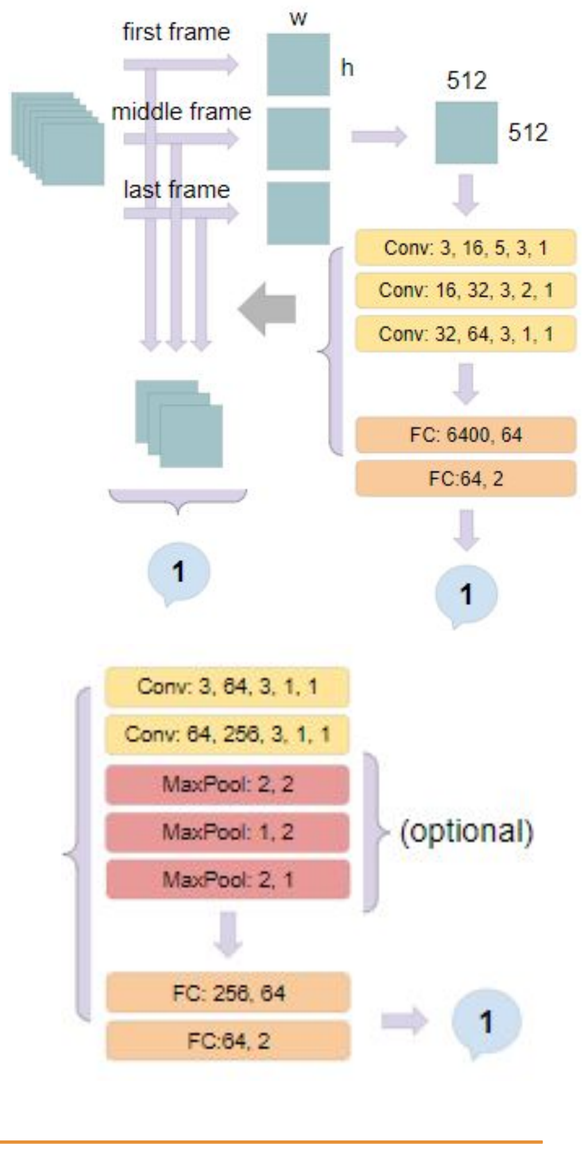
DLCV Final Project 1 - Talking To Me

Team BCNM
B08901169 梁正
B08901019 吳冠緯
B08901027 吳育丞
B08901048 陳宥辰

A. Vision Model

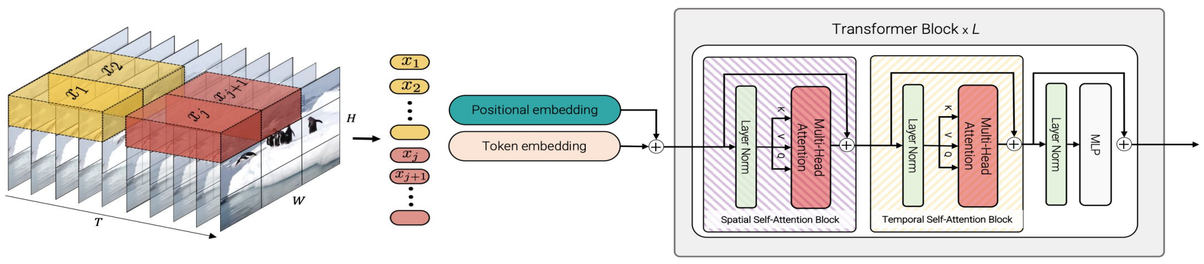
➢ Few-Frame Classification

- Image Preprocessing
  Crop bbox regions.
- Method 1 - Single Frame
  Only use a single frame in the middle.
- Method 2 - Three Frames
  Concate the first, middle and the last frame.
- **Result**
  Single-frame method yields better performance. (Accuracy 0.6)
- Drawbacks
  - Unable to capture continuous information.
  - Might use a bad image.
  - Sizes of bbox regions are quite different.
- FCN
  - Try to handle cropped bboxs with different sizes without directly reshaping.
  - This method doesn't improve performance.



➢ ViViT (Video Vision Transformer) - Model 3

- Tubelet Embedding
  Extract non-overlapping, spatio-temporal "tubes" from the input volume, which is similar to "patches" in ViT.
- Factorized Self-Attention
  Do temporal attetion after spatial attention in each transformer block.
- Image Preprocessing
  - Method 1 - Focus on the person's face
    Crop a square box which just contains all bbox regions of a certain person in a video clip.
  - Method 2 - Include more backgrounds
    Crop a square box as big as possible and add an extra channel for each frame, which pixels in the bbox region are assigned 1, otherwise, 0.
- Padding
  - Method 1 - Zero Padding
  - Method 2 - Repeating Frames
- **Result**
  The performance is not even better than single-frame method.
  (Accuracy 0.58~0.6)
- **Conclusion**
  - **Continuous visual information might not really help.**
  - **ViViT is assumed to require a much bigger dataset.**

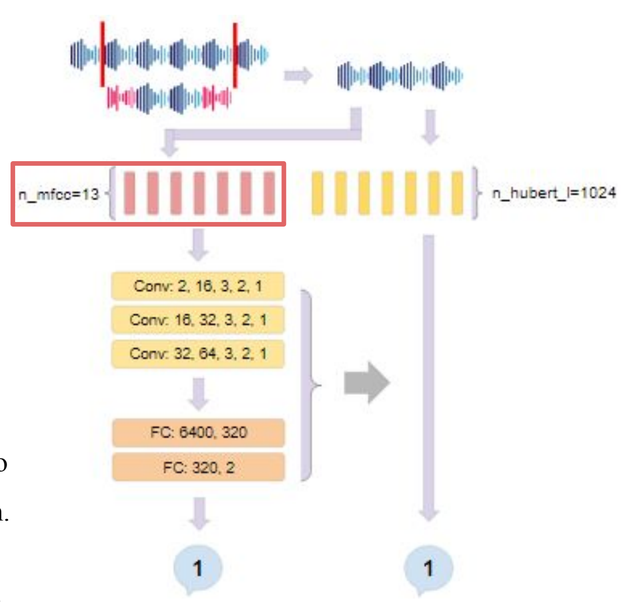


ViViT: A Video Vision Transformer ICCV 2021 · Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, Cordelia Schmid

B. Audio Model

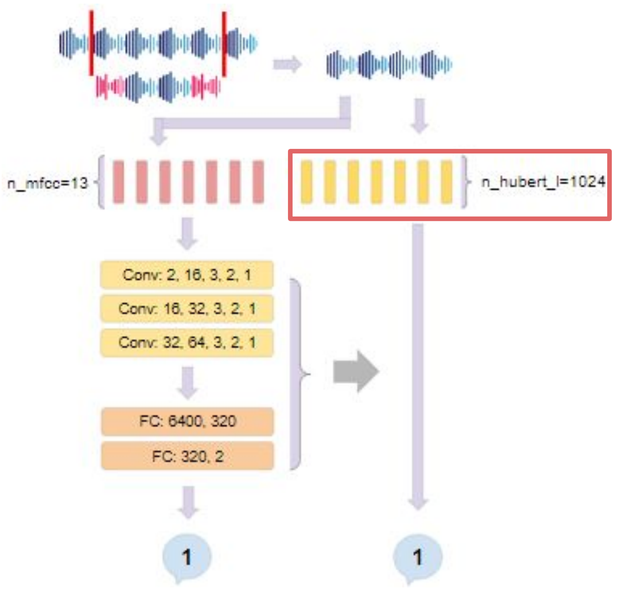
➢ MFCC

- Data Preprocessing - Padding / Truncation
  For each audio segment (start frame ~ end frame), we either zero-pad/truncate the audio to 4 seconds based on its length.
- Audio Feature Extraction
  We extract audio features from MFCC in torchaudio with default setting.
- Drawbacks
  - Some segments are too short (<1sec) and MFCC may have limited ability to extract enough information from them.
- Improved Preprocessing Method
  - Use complete 4-sec data from original video instead of zero-padding.



➢ HuBERT

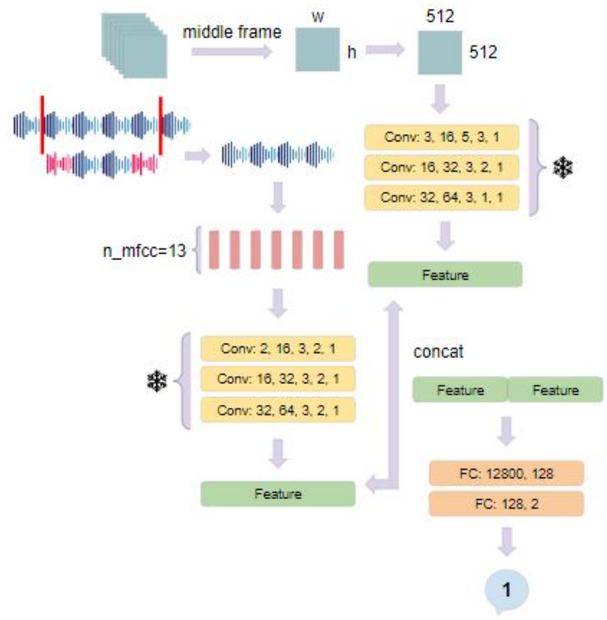
- Data Preprocessing - Padding / Truncation
  For each audio segment (start frame ~ end frame), we either zero-pad/truncate the audio to 5 seconds based on its length.
- Audio Feature Extraction
  We extract audio features from pretrained HuBERT's (hubert_large_ll60k, using S3PRL toolkit) 22nd layer.
- Drawbacks
  - Though we could obtain robust acoustic features from the SSL models, the large size of features results in longer training time.



C. Hybrid Model

➢ Single-Frame + MFCC / HuBERT

- Flow
  - Pretrain single-frame model and MFCC model as backbone.
  - Use both models (without FC layers) to generate features, respectively.
  - Concate two feature vectors and feed it to a FC classifier .
  - Train the whole network, including models with pretrained parameters that can either be freezed or not..
- Drawbacks
  - Some video clips lack people's visiual information, which leads to worse performance than pure audio model on this kind of data.
- Improved Embedded Model
  - Use pure audio model for data without people's visiual information, otherwise, use the hybrid model .



D. Results & Analysis

Vision Model	Validation Acc.
Single-Frame	**0.6** (Test: 0.59)
All-Frame	0.56
Concat 3 Frames	0.58
Averaging 3 Frames	0.57
Vivit Model - Method 1	0.58
Vivit Model - Method 2	0.56

Audio Model	Validation Acc.
MFCC 4sec	0.63 (Test: 0.62)
HuBERT 5sec	0.65

Hybrid Model	Validation Acc.
Single-Frame + MFCC 2sec	0.6
Single-Frame + MFCC 4sec	0.66 (Test: 0.63)
Single-Frame + HuBERT 5sec	0.61
Embedded Model	(Test: 0.63)

➢ Ablation Study

- CNN Model: **single-frame** > **three-frame** (**concat** > **avg**) > **all-frame**
- Vivit+**Method1** > Vivit+**Method2**
- CNN Model: **HuBERT**-based > **MFCC**-based (**4sec** > **2sec**)
- **MFCC** hybrid model > **HuBERT** hybrid model
- MFCC **embedded model** > MFCC **hybrid model**