Lip Reading based on Audio-Visual Hidden Unit BERT

指導教授:陳冠文
組員:何存益 110550165 房天越 110550085

I. Abstract

Human perception of speech relies on multiple sensory modalities, primarily auditory and visual. Traditional lip-reading models often struggle to capture subtle variations in lip shapes and generally rely only on visual datasets for training. Audio-Visual Hidden Unit BERT (AV-HuBERT) uses self-supervised learning to cluster audio and visual features into hidden units, thus improving lip-reading accuracy. In this project, we aim to apply AV-HuBERT to Chinese language, addressing the challenges that arise and exploring the solutions.

II. AV-HuBERT

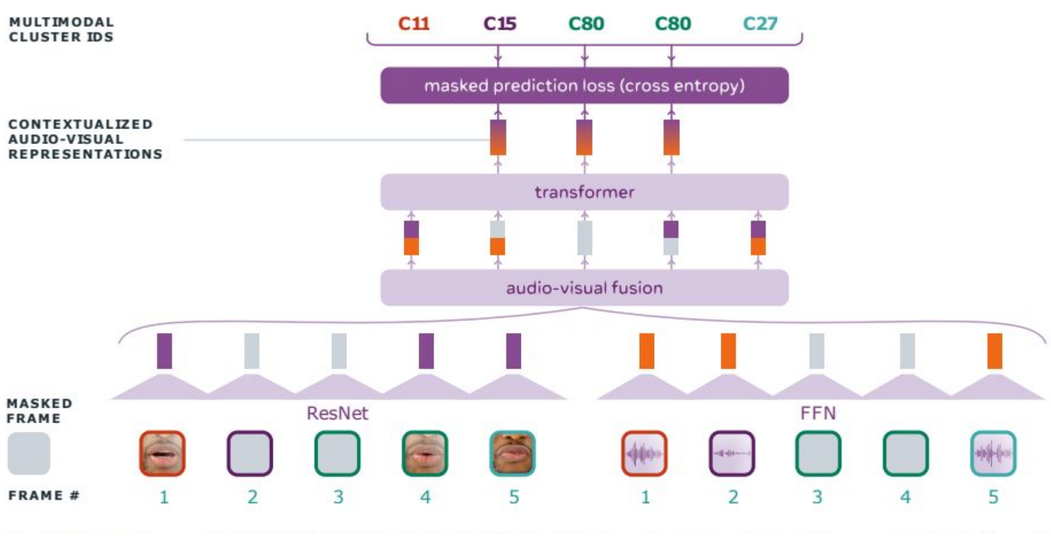


Figure 1 AV-HuBERT model

Figure 1 illustrates the AV-HuBERT model, which combines audio and visual inputs for lip reading. Visual frames are processed through ResNet, and audio frames through an audio extractor, with certain frames masked to predict missing parts. Fused audio-visual features are passed through a transformer, and a masked prediction loss is computed based on cluster IDs.

III. Method

1. Audio-Visual Input

   Visual and Audio features go through encoders to produce intermediate features, then be concatenated and be fed to a shared transformer for masked prediction.

2. Modality Dropout

   To avoid the problem that the model is dominated by audio stream, use only one single linear layer to encode the acoustic input. Also, before feeding the features to the transformer, use modality dropout:

$$f_t^{av} = \begin{cases} \text{concat}(f_t^a, f_t^v), \text{ with } p_m \\ \text{concat}(f_t^a, 0), \text{ with } (1\text{-}p_m)p_a \\ \text{concat}(0, f_t^v), \text{ with } (1\text{-}p_m)(1\text{-}p_a) \end{cases}$$

3. Audio-Visual Clustering

   Pretraining on both modalities enables the generation of multi-modal clusters, serving as the intermediate targets for future generational prediction tasks.

4. Masking by Substitution

   Random segments from the same visual sequence are used to replace masked visual input segments.

5. Training

   The training loss function:

$$L = -\sum_{t \in M^a \cup M^v} \log p_t(z_t) - \alpha \sum_{t \notin M^a \cup M^v} \log p_t(z_t)$$

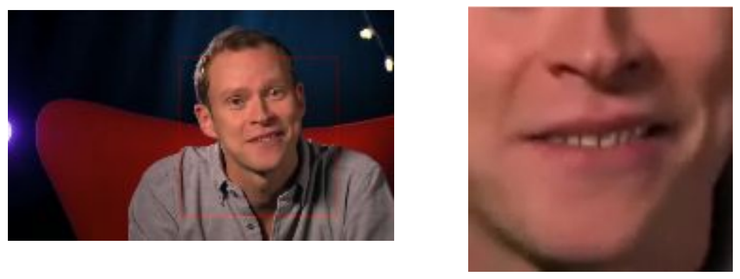
   M^a and M^v are masked frames in the audio and visual stream.

6. Fine-tuning

   Finetune on the lip-reading tasks with CTC or Seq2Seq Loss.

IV. Result

Prediction - so rather than just relying on this observation
Ground Truth - so rather than just relying on this information



Prediction - four little titles kept off the attacky related targets
Ground Truth - four little turtles named after Italian Renaissance artists

From the results above, the shorter words with simpler syllables tend to have more accurate predictions. However, for longer words with more syllables, there is a tendency for them to be split or partially merged with the following word, which can increase the total word count in the sentence.

V. Future Work

The AV-HuBERT model is originally designed for English lip reading, so we plan to fine-tune it using the CMLR dataset to enhance its ability to predict spoken Chinese based on lip movements. Since Chinese has distinct phonetic features compared to English, this fine-tuning process will allow the model to adapt to Chinese characteristics and generate more accurate predictions in Chinese.

VI. Reference

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, Abdelrahman Mohamed, “Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction”, ICLR 2022

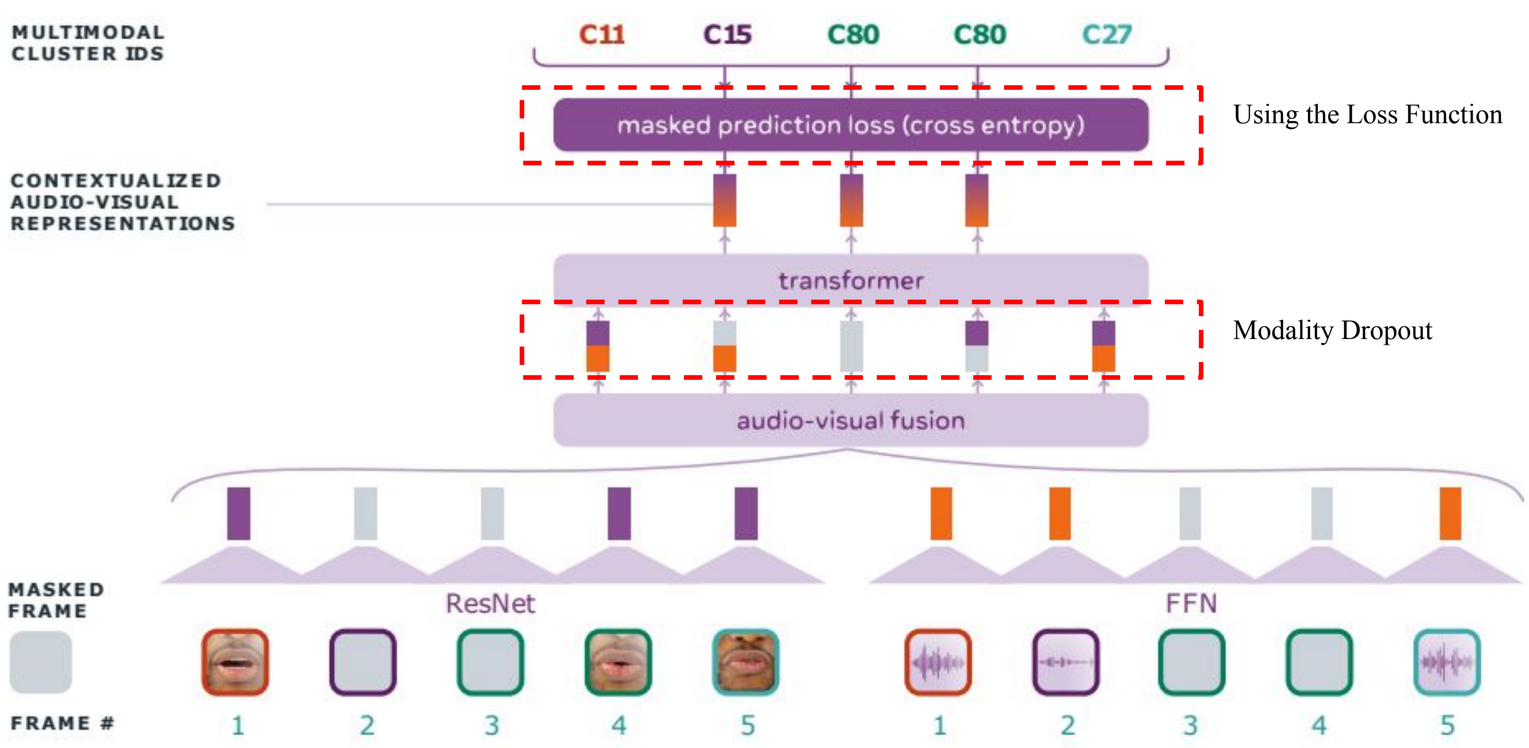
Lip Reading based on Audio-Visual Hidden Unit BERT

指導教授:陳冠文
組員:何存益 110550165 房天越 110550085

Abstract

Human perception of speech relies on multiple sensory modalities, primarily auditory and visual. Traditional lip-reading models often struggle to capture subtle variations in lip shapes and generally rely only on visual datasets for training. Audio-Visual Hidden Unit BERT (AV-HuBERT) uses self-supervised learning to cluster audio and visual features into hidden units, thus improving lip-reading accuracy. In our project, we aim to apply AV-HuBERT to Chinese language, addressing the challenges that arise and exploring the solutions.

Method



Modality Dropout

$$f_t^{av} = \begin{cases} \text{concat}(f_t^a, f_t^v), \text{ with } p_m \\ \text{concat}(f_t^a, 0), \text{ with } (1-p_m)p_a \\ \text{concat}(0, f_t^v), \text{ with } (1-p_m)(1-p_a) \end{cases}$$

Training Loss Function

$$L = -\sum_{t \in M^a \cup M^v} \log p_t(z_t) - \alpha \sum_{t \notin M^a \cup M^v} \log p_t(z_t)$$

Result

Prediction - so rather than just relying on this observation
Ground Truth - so rather than just relying on this information

Prediction - four little titles kept off the attacky related targets
Ground Truth - four little turtles named after Italian Renaissance artists

From the results above, the shorter words with simpler syllables would be more accurate.
For longer words with more syllables, it will be split or partially merged with the following word, decreasing the accuracy.

Future Work

The AV-HuBERT model is originally designed for English lip reading, so we plan to use the CMLR dataset to enhance its ability to predict spoken Chinese contexts.
However, using AV-HuBERT for Chinese lip reading is challenging due to tonal distinctions, visually similar lip shapes, and flexible syntax, which all require precise context understanding beyond English requirements. To solve these problems, we would apply language models to help us distinguish between similar words and intonation.

Reference

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, Abdelrahman Mohamed, “Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction”, ICLR 2022