UniLip: Learning Visual-Textual Mapping with Uni-Modal Data for Lip Reading

Bingquan Xia, Shuang Yang, Shiguang Shan, Xilin Chen

Background
- Existing lip reading methods rely on large-scale labelled video-text pairs to perform supervised training.
- Collecting labeled video-text pairs are time-consuming, while collecting uni-modal videos and uni-modal texts are much easier.
- Uni-modal texts contain rich linguistic prior information of the target language which could facilitate lip reading.

Motivation
- Utilize uni-modal videos and uni-modal texts to perform lip reading.

Video&Text Data Examples
- LRS3: TED talks, 433h.
- LRS2: BBC shows, 224h.
- Vox2-433h: English sub-set of VoxCeleb2, 433h.

Supervised Model

An example of linguistic priors

Supervised Approach

The proposed UniLip
- Decompose lip reading into two sub-tasks: (S1) learn linguistic priors from uni-modal texts (language modelling); (S2) generate text distributions conditioned on uni-modal videos (conditional generation).
- Propose a unified adversarial training framework to finish both (S1) and (S2).
- (S1): \( \mathcal{D} \) maximizes the log likelihood of real samples; (S2): \( \mathcal{G} \) generates text distributions that could deceive \( \mathcal{D} \) conditioned on visual inputs.
- Multi-grained Learning of Linguistic Priors: alleviate the biases of text sources and domains by ngram sampling.
- Multi-grained Visual-Textual Mapping: adapt features of pre-trained models by integrating both local information and the global context.

Unsupervised Results
- UniLip's performance scales with the size of texts.
- UniLip can effectively accommodate videos and texts from different sources.

<table>
<thead>
<tr>
<th>Training Video</th>
<th>Training Text</th>
<th>Test WER/% (Base)</th>
<th>Test WER/% (Large)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LRS3</td>
<td>LRS3</td>
<td>51.9 (1.9)</td>
<td></td>
</tr>
<tr>
<td>LRS2</td>
<td>TEDLIUM</td>
<td>51.2 (1.9)</td>
<td>53.1 (1.9)</td>
</tr>
<tr>
<td></td>
<td>Cantab</td>
<td>61.8 (1.0)</td>
<td>60.8 (1.0)</td>
</tr>
<tr>
<td></td>
<td>LibriSpeech</td>
<td>N/A</td>
<td>64.9 (1.0)</td>
</tr>
<tr>
<td>LRS3</td>
<td>LRS3</td>
<td>57.2 (1.9)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TEDLIUM</td>
<td>57.8 (1.9)</td>
<td>57.8 (1.0)</td>
</tr>
<tr>
<td></td>
<td>Cantab</td>
<td>60.7 (1.8)</td>
<td>58.9 (1.8)</td>
</tr>
<tr>
<td></td>
<td>LibriSpeech</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Semi-supervised Results
- \( L = L_{\text{seq2seq}} + aL_{\text{GAN}} \).
- UniLip could effectively incorporate extra uni-modal data into the popular supervised seq2seq framework.

Visualization
- perform phoneme-level decoding and retrieve corresponding input lip images.
- UniLip successfully maps different phonemes to different lip shapes, such as "CH" and "M".