SpeechCLIP: Integrating Speech with Pre-Trained Vision and Language Model

Yi-Jen Shih¹, Hsuan-Fu Wang¹, Heng-Jui Chang¹,², Layne Berry³, Hung-yi Lee¹, David Harwash³
¹National Taiwan University ²MIT CSAIL ³The University of Texas at Austin

Motivation

- Paired speech-image data is cheaper and easier to collect compared to transcription data
- Compared to previous works, SpeechCLIP utilizes visually grounded speech and indirect text signals for training

Methods

- CLIP [1]
- Align image and speech into same embedding space and able to transfer to Computer Vision Tasks

SpeechCLIP

- Parallel: the spoken utterance embedding is aligned with CLIP image encoder’s output
- Cascaded: K keywords are selected and inserted into CLIP text encoder to align with CLIP image encoder’s output

Results

- Keyword Discovery
  - Keywords are retrieved by finding subwords with the highest cosine similarities between $z_k$ and the corresponding subword embeddings

<table>
<thead>
<tr>
<th>Model</th>
<th>Audio Encoder</th>
<th>CLIP Image Encoder</th>
<th>Trainable Params (M)</th>
<th>Total Params (M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>HuBERT Base (95M)</td>
<td>ViT-B/32 (250M)</td>
<td>2.8 - 7.5</td>
<td>252 - 257</td>
</tr>
<tr>
<td>Large</td>
<td>HuBERT Large (216M)</td>
<td>ViT-L/14 (422M)</td>
<td>6.1 - 13.4</td>
<td>765 - 772</td>
</tr>
</tbody>
</table>

Model Details

- Keyword hit rates for cascaded SpeechCLIP. Avg denotes averaged hit rate

Reference