AVSpeech Data Filtering

Ami Baid

May 2023

The AVSpeech dataset [1] consists of 3-10 second clips from 290k YouTube videos, most of which feature a single speaker and little background noise. We utilize a subset of the AVSpeech dataset, filtering out videos that are not well-suited for visual acoustic matching—i.e., videos that have little audio-visual correspondence. To filter the dataset, we use the Vision-and-Language Transformer model (ViLT) [2]. Given an image-question pair, the ViLT model performs a classification task with 3,129 different answer classes, where each class corresponds to a natural language word.
We observe that the model accurately answered a variety of yes-or-no questions about the visible setting in each clip. Since the clips are primarily recorded using a static camera, we ran the model for inference using a representative frame from each video. For each question, we selected a confidence threshold that filtered out clips likely to have the corresponding failure modes based on prior manual inspection.
Videos deemed to have low audio-visual correspondence (e.g., images where the speaker is using a microphone, images with a virtual background) were removed from the training set. The initial set of 123,841 video clips is reduced by 38%, resulting in 76,473 video clips with high audio-visual correspondence that were used for training. See Supp. for the specific questions and confidence thresholds used.

Table 1: **Questions and thresholds used for filtering AVSpeech data**

Question	Answer	Threshold
Is a microphone or headset visible in the image	yes	0.35
Is there a whiteboard/blackboard in the background	yes	0.993
Is the entire background one solid color and material?	yes	0.995
Is there a large projector screen covering most of the background?	yes	0.999
Is part or all of the background virtual?	yes	0.6
Are there multiple screens in the image?	yes	0.995
Is the wider room clearly visible?	no	0.985

References

[1] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *CoRR*, abs/1804.03619, 2018.

[2] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision, 2021.