Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction

(ICLR 2022) – Review

Summary

This paper presents **AV-HuBERT**, a self-supervised audio-visual speech representation learning framework. The authors extend HuBERT to *multimodal* input (speech audio + corresponding video of lip movements) by masking parts of both modalities and predicting discrete cluster IDs for the masked portions. A separate encoder is used for each modality (a linear feed-forward layer for audio features and a modified ResNet-18 for video frames), after which the audio and visual features are fused and fed into a shared Transformer encoder. The model is trained to predict target cluster assignments for each time step using a cross-entropy loss. These targets are obtained via k-means clustering: starting with clusters of acoustic MFCC features for the first iteration, and then using the learned embeddings from the previous iteration to refine the clusters in subsequent iterations. This iterative *masked cluster prediction* setup allows the model to discover improved audio-visual units over time. The training objective is a masked prediction loss that operates only on positions where at least one modality is masked, while optionally also giving a smaller weight to unmasked positions:

$$\mathcal{L} = -\sum_{t \in M_a \cup M_v} \log p_t(z_t) - \alpha \sum_{t \notin M_a \cup M_v} \log p_t(z_t) ,$$

where M_a, M_v are the sets of masked timesteps in the audio and video streams respectively, z_t is the cluster label for time t, and α is a down-weighting factor for the loss on unmasked frames (usually $\alpha < 1$). In practice, α encourages the model to focus learning on masked regions while still using the context of unmasked regions.

To create a challenging prediction task for the visual modality, the paper introduces a novel *masking by substitution* strategy: instead of simply dropping/blanking video frames, AV-HuBERT replaces some segments of the input video with random segments (an *impostor*) from a different time in the same video. This way, the model must detect which video frames have been corrupted and still predict the correct cluster IDs for the original content. The audio stream is masked in a more standard way (random audio feature frames replaced by a mask embedding, as in HuBERT). The mask probabilities for audio vs. video are set differently (m_a vs m_v) because predicting masked audio is easier than masked lip images; a higher masking rate on audio (e.g. mask a larger fraction of audio frames) is used to force the model to learn linguistic content from audio, while the video masking rate is kept lower to avoid overwhelming the visual encoder with too many fake frames.

An important technique in AV-HuBERT is **modality dropout**, used to prevent the model from simply relying on the easier audio modality. In each training sequence, with probability p_m the model uses both audio and visual inputs together; otherwise, it drops one modality entirely (feeding all-zeros for that modality). When only one modality is used, the audio-alone case is chosen with probability p_a (and video-alone with probability $1 - p_a$). Formally, if $f_t^{(a)}$ and $f_t^{(v)}$ are the encoded features at time t for audio and video respectively, the fused feature $f_t^{(av)}$ is constructed as:

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

In other words, the model sometimes sees both modalities, and other times sees only audio or only video. This sequence-level dropout forces the Transformer to learn to handle audio-only and visual-only cases as well, which is handy for fine-tuning: e.g. for a lip-reading task, the pre-trained AV-HuBERT can be fine-tuned with just the visual stream (audio dropped) without a big modality mismatch, and similarly for audio-only ASR.

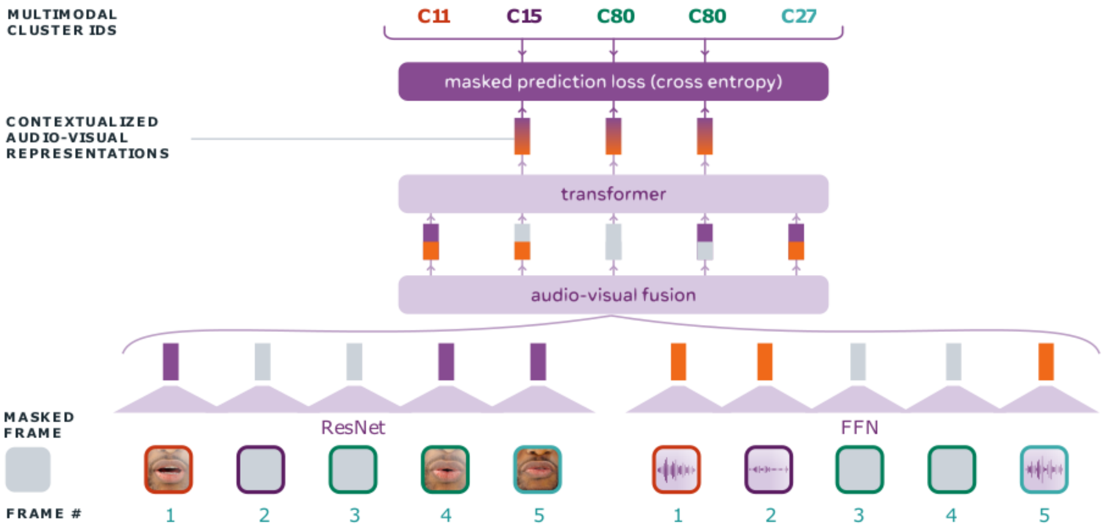

Figure 1: *AV-HuBERT architecture overview. Masked prediction losses are computed on frames where at least one modality is masked. ResNet and FFN process visual and audio inputs respectively.*

Strengths

- **Novel multimodal approach:** good extension of HuBERT for AV data.
- **Masking strategy:** visual masking via substitution is clever.

- **Modality dropout:** prevents model from relying too much on audio.
- **Strong results:** SOTA on LRS3 with only 30h labels.
- **Thorough experiments:** good ablations and baselines.

Weaknesses

- **Training complexity:** 5-stage clustering + masked prediction is costly.
- **Audio bias in early stage:** first cluster iteration uses MFCC audio only.
- **Somewhat incremental:** builds closely on HuBERT ideas. So, there is a question to actual novelty.
- **Lack of interpretability:** no deep dive into what clusters mean.
- **Limited robustness:** only tested on LRS3 domain.

Minor Questions

- Sensitivity of m_a, m_v, and α?
- Any experiments on stronger visual encoders than ResNet-18?
- Do substituted video segments get detected well during training?

Reviewer Suggestions

To improve the work:

- Try AV-HuBERT on noisy speech/video domains.
- Analyze clusters for phoneme/viseme alignment.
- Explore end-to-end alternatives to k-means.
- Try multilingual or cross-lingual pretraining.
- Fine-tune end-to-end AV-ASR with external LMs.

Rating

7/10 (Accept). A good paper with clear wins over previous works. Not radically new but has clever ideas and achieves impressive results. Training cost is high but justified. Some more novelty would have been better.

Bonus Question

To evaluate the robustness and transferability of AV-HuBERT beyond the LRS3 domain, we examine its performance on three datasets: CREMA-D, VoxCeleb2, and LRS3. Notably, for CREMA-D, we apply **DoRA (Downstream Re-Alignment)** fine-tuning on top of the pretrained AV-HuBERT base model to better adapt it to the dataset.

- **LRS3**: Serving as the original benchmark domain, AV-HuBERT achieves a WER of approximately **39%**; where we used a subset of data. This dataset features TED and TEDx talks, offering relatively clean and structured speech but with high vocabulary variability.

- **VoxCeleb1**: Featuring unstructured, real-world interviews with substantial background noise and speaker variation, VoxCeleb2 remains challenging even for pretrained models.AV-HuBERT has got a WER of **47%**, owing to the absence of fixed scripts and inconsistent audio quality.

- **CREMA-D**: This dataset includes short, emotionally acted clips of 12 fixed sentences spoken by 91 actors. After applying DoRA fine-tuning to the pretrained model, the expected WER reduces significantly due to the constrained vocabulary and consistent structure, approximately **25%**.