Ruchira Ray EID:rr52486

Paper Response: Unsupervised Speech Recognition

Multiple natural language processing, speech and vision tasks have been successful using self-supervised learning; nevertheless, downstream processes still need labels. These authors provide a novel approach to automated speech recognition that does not rely on any labelled data. This research advances the use of unlabeled data for speech-related problems. This approach only uses unlabeled speech audio and unlabelled text which may or may not be related to the speech audio. It starts by learning the structure of speech from unlabeled audio. We initially learn good speech representations using a previously published model self-supervised model, wav2vec 2.0. We use a k-means clustering algorithm to find cluster ids, and we segment the speech using this into phonemic units. This is done by placing a segment boundary whenever the cluster-ID changes. Following that, segment representations are generated by employing mean-pooling the wav2vec 2.0 representations, PCA, and a second mean-pooling step between adjacent segments. Then we feed these representations into the generator network which generates phenome sequences. The authors train a generative adversarial network (GAN) composed of a generator and a discriminator network. The generator takes each audio segment contained in self-supervised representations and predicts a phoneme matching to a sound in language. It is trained by attempting to trick the discriminator, which determines if the anticipated phonemes sequences are realistic. The discriminator is a neural network which is trained by giving it the generator's output as well as presenting it with phonemized text from multiple sources. The authors use off-the-shelf toolkits for phenomizing text. It learns to discriminate between the generator's voice recognition output and genuine text in this manner. The transcriptions are initially quite bad, but with time and discriminator feedback, they improve.

Wav2vec-U proposes an unsupervised technique for training speech recognition models without using any labelled data. It utilizes self-supervised speech representations to segment unlabeled language and learns an adversarial mapping from these representations to phonemes. The article is well-motivated: unsupervised speech recognition is particularly fascinating given the abundance of unlabeled data that can be accessed. Wav2vec-U Training requires no labels and lives up to its promises. It's also worth noting that when typical unsupervised machine translation approaches are used to match discretized speech audio units to phonemic units, they perform poorly. This may be reinforced by the fact that speech carries significantly more information, as well as variances in pronunciation, silences, and other nuances. Wav2vec-U significantly decreases the TIMIT phone error rate. when compared to the previous best-performing unsupervised speech recognition technique. As a result, it proves to be SOTA in unsupervised scenarios, however, it is not nearly comparable with current supervised models, but it may be competitive with models from 2018. People acquire speech-related abilities simply by listening to others. This implies that it should be possible to train speech recognition algorithms that do not need enormous volumes of labelled data. Reducing our dependence on labelled data is an important step toward widening and speeding up the development of technology. This research has the potential to make way for the development of very effective voice recognition technologies for many more languages and dialects throughout the world. It would be interesting to see how much better this would be at picking up nuanced meanings that get lost in literal translations of speech. For future studies, it would be interesting to see how such models can take intonation and emotion into account when embedding projecting embedding space. It would also be fascinating to investigate what linkages exist inside the embedding space between identical terms in other languages.