Layer-wise analysis of a self-supervised speech representation

Ankita Pasad, Ju-Chieh Chou, Karen Livescu

Toyota Technological Institute at Chicago

{ankitap, jcchou, klivescu}@ttic.edu

Abstract

Recently proposed self-supervised learning approaches have been successful for pre-training speech representation models. The utility of these learned representations has been observed empirically, either as a boost in performance or as a relaxation in labeling requirements for supervised downstream tasks. However, not much has been studied about the type or extent of information encoded in the pre-trained representations themselves. Developing such insights can help understand model capabilities and limits, and thus enable the research community to more efficiently develop their usage for downstream applications. In this work, we begin to fill this gap in understanding by examining one recent and successful pre-trained model (wav2vec 2.0), via its intermediate representation vectors, using a suite of analysis tools. We use the metrics of canonical correlation, mutual information, and performance on simple downstream tasks with non-parametric probes, in order to (i) query for acoustic and linguistic information content, (ii) characterize the evolution of information across model layers, and (iii) understand how fine-tuning the model for automatic speech recognition (ASR) affects these observations. We use these findings to modify the fine-tuning protocol for ASR, and obtain an improved WER with 10 minutes of training data.

Index Terms: Self-supervised pre-training, representation analysis, speech representation learning

1. Introduction

Various self-supervised learning techniques have recently been proposed to learn speech representations (e.g., [1, 2, 3, 4]). Although new and improved approaches are being proposed at a rapid rate, very little is understood about the pre-trained models themselves apart from their empirical successes on downstream tasks, leaving their development and application as a time- and resource-consuming process of trial and error. We seek to fill that gap by analyzing pre-trained models to understand how the representations evolve across layers and how they change when fine-tuned for a downstream task.

We investigate the layer-wise evolution of representations in a self-supervised model. We are especially interested in studying representations directly, rather than training additional classifiers as probes, to avoid the overhead of training many classifiers. We study the relationship between representation layers and a range of linguistic properties including phonetic content, word identity, and word meaning. We perform all our analysis on wav2vec 2.0 (W2V2) [4], which has been successful and broadly studied for ASR. Several variants of W2V2 are publicly available; we study all these variants and their fine-tuned counterparts.[1]

[1] https://github.com/pytorch/fairseq/blob/master/examples/wav2vec

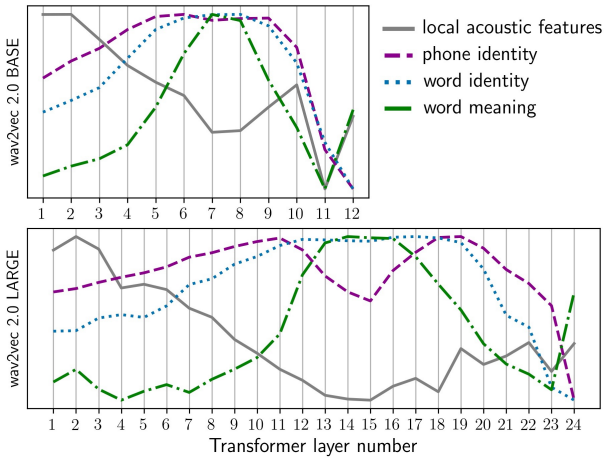

Figure 1: *Visualization of the degree to which several properties are encoded in different transformer layers of pre-trained models. The curves measure different quantities on different scales; they are shown together only to compare where major peaks and valleys occur in each. Section 4 provides more details.*

1.1. Summary of findings

We make the following observations from our analyses (i) the W2V2 transformer layers follow an autoencoder-style behaviour, where as we go deeper into the model, the representation starts deviating from the input speech features followed by a reverse trend where even deeper layers become more similar to the input, as if reconstructing the input; (ii) the layer-wise evolution of the representations follows the linguistic hierarchy of speech understanding, with the shallowest layers encoding acoustic features, followed by phonetic, word identity, and word meaning information, in that order (and then followed by a reverse trend as described above) as illustrated in Figure 1; (iii) fine-tuning the model for ASR breaks the autoencoder-style behaviour, especially in the final few layers, which accordingly also get better at encoding word identity, (iv) the final convolutional layers and initial transformer layers are highly correlated with mel spectrogram features, suggesting that the model has learned to extract features similar to human-engineered ones; (v) the model seems to encode some word meaning information, though the extent of the encoded semantics is unclear; (vi) the last two layers often defy the trends observed for other layers; and (vii) a modified fine-tuning protocol for ASR, designed based on these findings, improves the WER, and the layer-wise trends in WER also correlate with the extent of word identity information according to our analyses.

2. Related Work

Some very recent work has begun to explore the phonetic content in pre-trained models using a classifier probe [5] and relationships between models trained with different training objec-

tives and model architectures [6]. We study a broader range of linguistic content and focus on lightweight methods that don't require training classifiers. Our work also shares much of the motivation of the 2021 Zero Resource Speech Benchmark [7], but our approach is dataset-agnostic, includes less implementation overhead, and performs layer-wise analysis. Our methods are closest to Voita et al.'s work on analyzing text representations [8]. We use a similar set of analysis tools based on canonical correlation analysis (CCA) and discrete mutual information (MI) estimates and perform layer-wise analysis, but apply them to the continuous domain of speech (as opposed to discrete text tokens), analyze the relationship between representations and both discrete and continuous labels, and analyze the relationship between pre-trained and fine-tuned models. To our knowledge, this is the first work to perform layer-wise analyses of a pre-trained speech representation model to assess a range of linguistic properties and to also report the effect of ASR fine-tuning on these trends.

3. Methods

We extract layer-wise representations of LibriSpeech utterances [9] from W2V2 models. We encode each utterance in its entirety, and then extract local representations at the frame level, phone level, or word level.

We use *projection-weighted CCA* [10] to measure similarity between the W2V2 layer representations and various continuous-valued quantities of interest, either (i) from a different layer of the same model, (ii) from a fine-tuned version of the model, or (iii) from an external representation. For the third type of analysis we use mel spectrogram features, acoustically grounded word embeddings [11] and GloVe word embeddings [12] as ways to assess the local acoustic, word-level acoustic and word meaning information encoded in the W2V2 representations respectively.

While CCA is a natural choice for relating continuous-valued vectors, we use *mutual information (MI)* to measure how learned representations relate to discrete-valued properties, specifically phone and word identities. Similarly to previous work using MI to analyze text representations [8], we cluster the continuous-valued representation vectors to obtain discrete cluster IDs. We then estimate MI using the co-occurrence counts of the cluster IDs and the phone/word labels.

We also measure performance on two downstream tasks, *word-discrimination* [13] and *semantic word similarity* [14], using non-parametric predictors based on the cosine similarities between representations. These tasks provide a concrete evaluation measure, and also help corroborate our findings from the CCA and MI experiments.

Finally, we measure performance on a downstream ASR task, specifically we train a character-based model on the 10-minute split of the LibriSpeech corpus, using different layers of the pre-trained model and different fine-tuning protocols.

4. Results

We report our main findings in Section 1.1. Figure 1 illustrates some of the layer-wise analyses, showing the ability of the W2V2 models to encode certain linguistic properties. For the purpose of this illustration, we include a subset of our analyses performed on the wav2vec 2.0 Base and Large models, pre-trained on 960 hrs LibriSpeech and 60k hrs LibriVox respectively. *Local acoustic feature* content is represented as the CCA similarity between W2V2 frame representations and mel-spectrogram features. As a measure of the *word meaning* content in each layer, we use the CCA similarity between W2V2 representations and GloVe word embeddings. *Phone and word identity* content is measured using the estimated MI between the representations and the discrete phone and word labels respectively. Note that these measures are not comparable to each other, so for the purpose of illustration within a single figure, they have been linearly rescaled and no units are given.

Based on our observation that the top two layers are often poorer at encoding linguistic information, we re-initialize the final two layers before fine-tuning the model for ASR. This protocol improves the WER from 41.5% to 39.2%.

5. Summary

We have presented a suite of analytical tools to assess the layer-specific information in pre-trained speech representations, applied to wav2vec 2.0 models. We have found that information about various linguistic levels tends to be encoded in different layers of the model, and that our analytical measures correlate well with performance on certain downstream tasks. Some of these findings have motivated a modification to the fine-tuning protocol for wav2vec 2.0, which leads to improved downstream ASR performance.

The analytical tools presented here can be easily applied to other pre-trained speech models. The insights from these analyses can help direct the research community toward additional useful modifications and also help understand the limitations of these models trained without any external supervision.

6. References

[1] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," arXiv:1807.03748, 2018.

[2] S. Pascual et al., "Learning problem-agnostic speech representations from multiple self-supervised tasks," in Interspeech, 2019.

[3] Y.-A. Chung and J. Glass, "Generative pre-training for speech with autoregressive predictive coding," in ICASSP, 2020.

[4] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in NeurIPS, 2020.

[5] A. Baevski, W.-N. Hsu, A. Conneau, and M. Auli, "Unsupervised speech recognition," arXiv:2105.11084, 2021.

[6] Y.-A. Chung, Y. Belinkov, and J. Glass, "Similarity analysis of self-supervised speech representations," in ICASSP, 2021.

[7] T. A. Nguyen et al., "The Zero Resource Speech Benchmark 2021: Metrics and baselines for unsupervised spoken language modeling," arXiv:2011.11588, 2020.

[8] E. Voita, R. Sennrich, and I. Titov, "The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives," in NAACL, 2019.

[9] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in ICASSP, 2015.

[10] A. S. Morcos, M. Raghu, and S. Bengio, "Insights on representational similarity in neural networks with canonical correlation," in NeurIPS, 2018.

[11] S. Settle, K. Audhkhasi, K. Livescu, and M. Picheny, "Acoustically grounded word embeddings for improved acoustics-to-word speech recognition," in ICASSP, 2019.

[12] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in EMNLP, 2014.

[13] M. A. Carlin, S. Thomas, A. Jansen, and H. Hermansky, "Rapid evaluation of speech representations for spoken term discovery," in Interspeech, 2011.

[14] M. Faruqui and C. Dyer, "Community evaluation and exchange of word vectors at wordvectors. org," in ACL, 2014.