wav2vec

What is wav2vec?

- Title of the paper: wav2vec 2.0: A Framework for **Self-Supervised** Learning of **Speech Representations**

- Learning **powerful representations** from speech audio using **self-supervision**

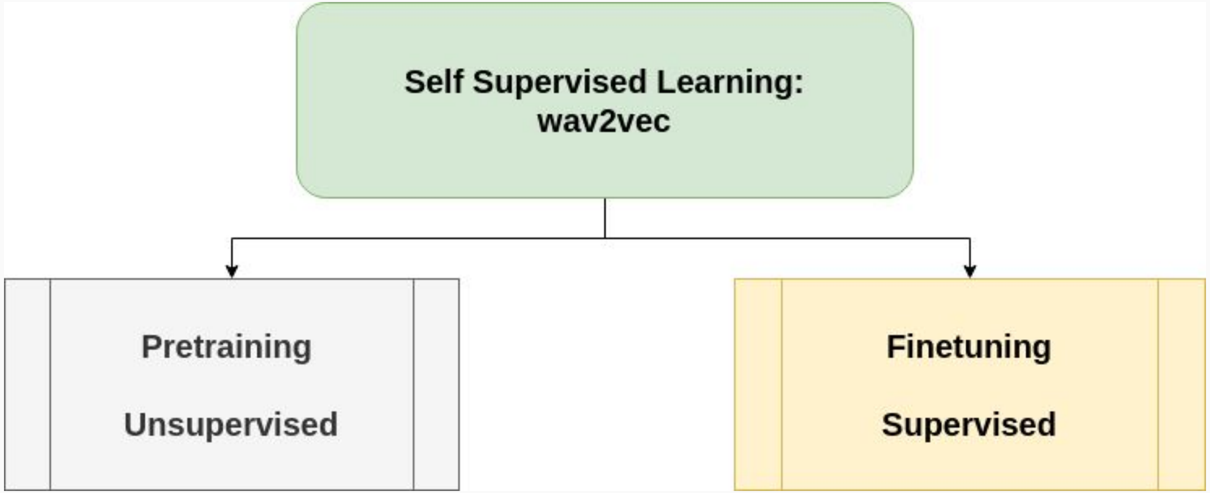
Self Supervision

Self Supervised = Unsupervised + Supervised

- This means there are two components in self supervised learning. One is unsupervised component and another is supervised component.
- Unsupervised learning is called pretraining.
- Supervised learning is called finetuning.

Self Supervision

Self Supervised = Unsupervised + Supervised



What are the building blocks of any Deep learning?

What are the building blocks of any Deep learning?

1. What is the data?

What are the building blocks of any Deep learning?

1. What is the data?
2. What is the model?

What are the building blocks of any Deep learning?

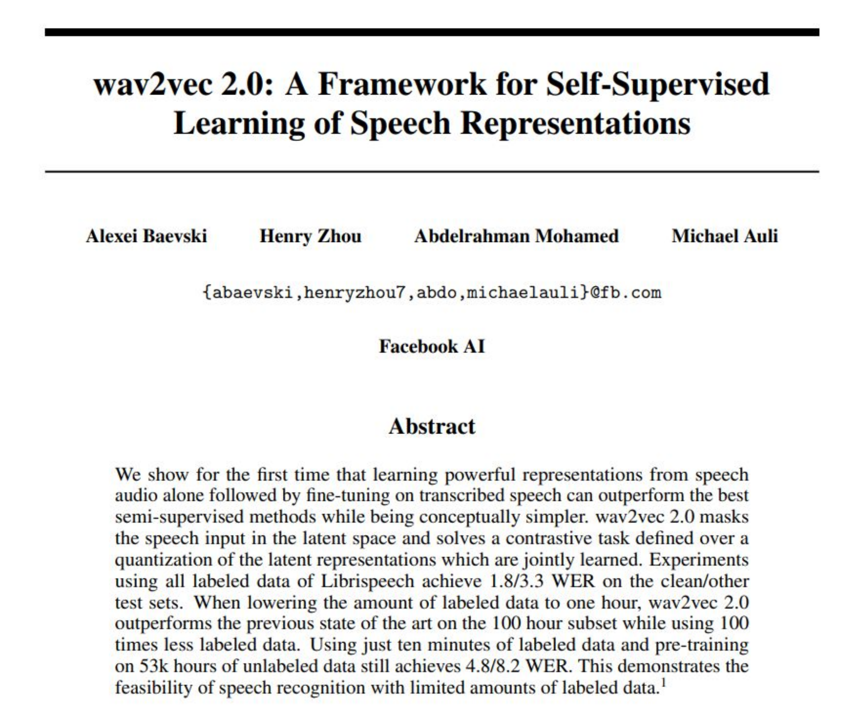
1. What is the data?
2. What is the model?
3. What is the loss function?

What are the building blocks of any Deep learning?

1. What is the data?
2. What is the model?
3. What is the loss function?
4. What is the metric?

wav2vec2 Architecture

wav2vec2 Architecture



wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations

Alexei Baevski   Henry Zhou   Abdelrahman Mohamed   Michael Auli

{abaevski,henryzhou7,abdo,michaelauli}@fb.com

Facebook AI

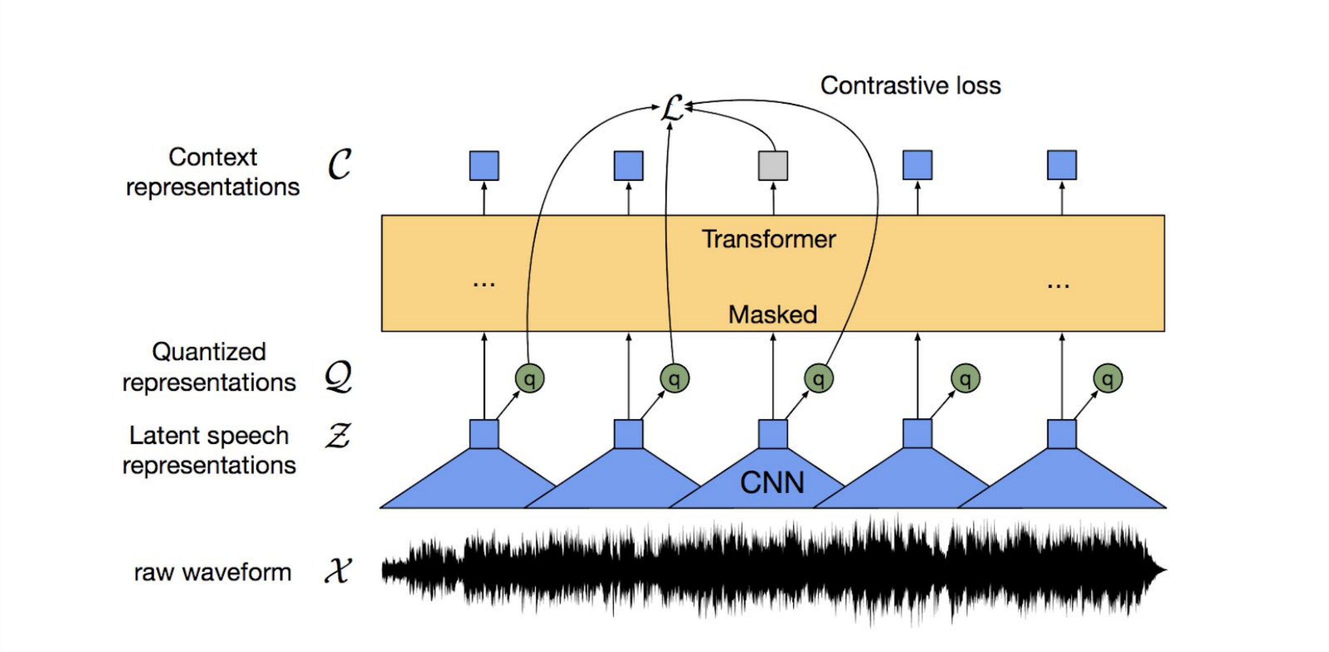
Abstract

We show for the first time that learning powerful representations from speech audio alone followed by fine-tuning on transcribed speech can outperform the best semi-supervised methods while being conceptually simpler. wav2vec 2.0 masks the speech input in the latent space and solves a contrastive task defined over a quantization of the latent representations which are jointly learned. Experiments using all labeled data of Librispeech achieve 1.8/3.3 WER on the clean/other test sets. When lowering the amount of labeled data to one hour, wav2vec 2.0 outperforms the previous state of the art on the 100 hour subset while using 100 times less labeled data. Using just ten minutes of labeled data and pre-training on 53k hours of unlabeled data still achieves 4.8/8.2 WER. This demonstrates the feasibility of speech recognition with limited amounts of labeled data.[1]

wav2vec2 architecture

- Speech representations are masked in the **latent space** and not the input space.
- Solves a **contrastive task** defined over a **quantization** of the **latent representations** which are jointly learned

wav2vec2 architecture

- 

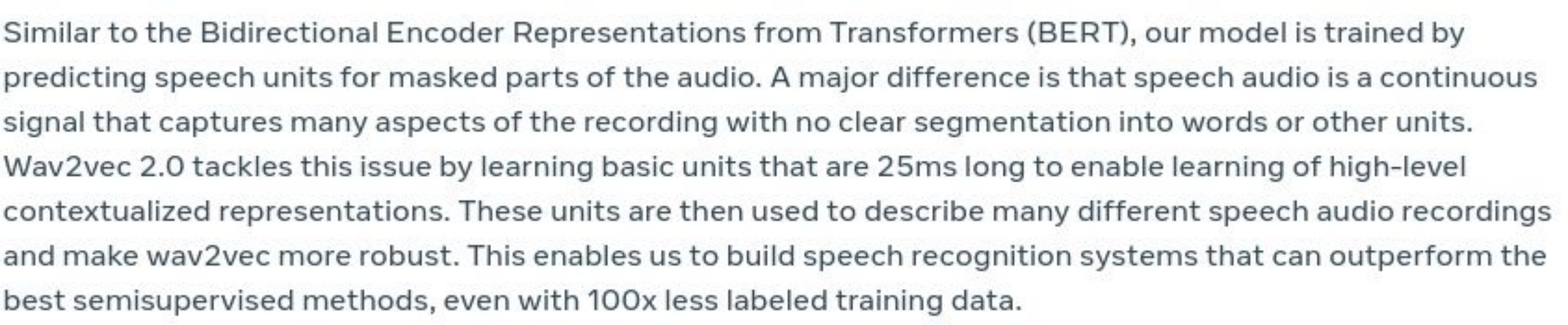
wav2vec2 architecture

- From raw waveform , latent speech representations are created by using set of CNN layers.
- Latent speech representations are then fed to transformer layers to capture the sequence information. This is a continuous representation.
- The latent speech representations is also sent to a quantization module which converts the continuous representation to discrete representations which represents the target in this task.

wav2vec2 training

- Latent speech representation are masked before passing to the transformer module by randomly sampling portions from the sentence.
- Masking is not done to the inputs that are passed to quantization module.
- During pre-training, representations of speech audio are learnt by solving a contrastive task Lm which requires to identify the true quantized latent speech representation for a masked time step within a set of distractors.

wav2vec2 architecture



Similar to the Bidirectional Encoder Representations from Transformers (BERT), our model is trained by predicting speech units for masked parts of the audio. A major difference is that speech audio is a continuous signal that captures many aspects of the recording with no clear segmentation into words or other units. Wav2vec 2.0 tackles this issue by learning basic units that are 25ms long to enable learning of high-level contextualized representations. These units are then used to describe many different speech audio recordings and make wav2vec more robust. This enables us to build speech recognition systems that can outperform the best semisupervised methods, even with 100x less labeled training data.

Pretraining

Pretraining: Unsupervised

1. What is the data?

Unlabelled data.

Unlabelled speech data. Only Audio data. Audio data can be from multiple languages.

Pretraining: Unsupervised

2. What is the model?

Wav2vec 2.0

Pretraining: Unsupervised

3. What is the loss function?

Contrastive Loss

Diversity Loss

Pretraining: Unsupervised

4. What is the metric?

Accuracy

Finetuning

Finetuning: Supervised

1. What is the data?

Labelled data.

Labelled speech data. Audio data + corresponding text. Audio data has to be from the language for which you want to create a ASR.

Pretraining: Unsupervised

2. What is the model?

Wav2vec 2.0

Just a final softmax layer on top of the wav2vec network

Pretraining: Unsupervised

3. What is the loss function?

CTC Loss

Pretraining: Unsupervised

4. What is the metric?

WER

Number of substitutions, insertions and deletions required to make two strings equal

(Minimum Edit Distance)