Wav2Vec2 2.0

•••

https://arxiv.org/pdf/2006.11477.pdf

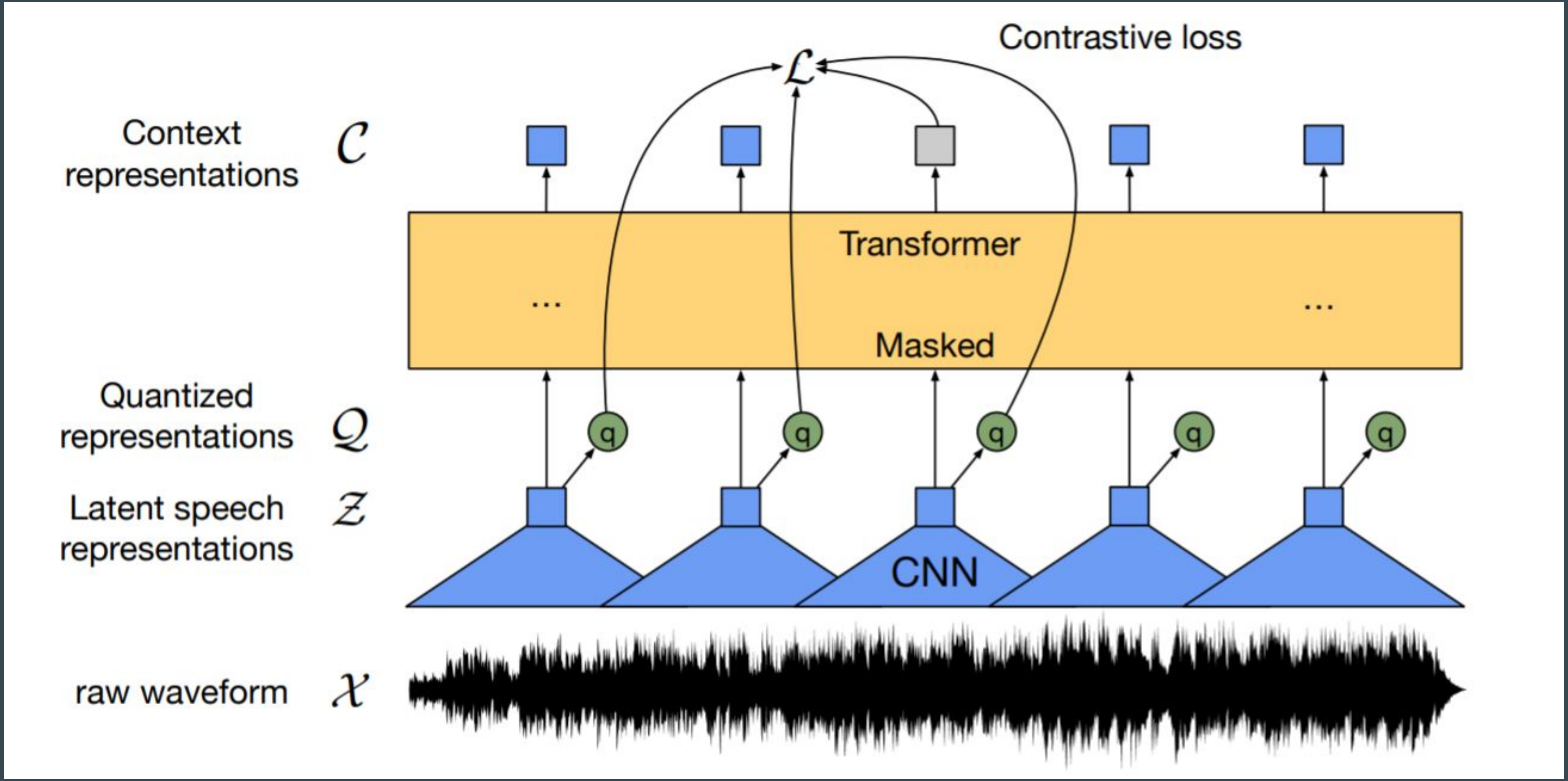
Compiled by - Vaibhav

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.

Wav2Vec2 Framework


Contrastive loss
$\mathcal{L}$
Context representations $\mathcal{C}$
Transformer
...
Masked
...
Quantized representations $\mathcal{Q}$
q
Latent speech representations $\mathcal{Z}$
CNN
raw waveform $\mathcal{X}$

Wav2Vec2 Framework

The approach encodes speech audio via a multi-layer convolutional neural network and then masks spans of the resulting latent speech representations, similar to masked language modeling.

The latent representations are fed to a Transformer network to build contextualized representations and the model is trained via a contrastive task where the true latent is to be distinguished from distractors

Model

Feature Encoder - The encoder consists of several blocks containing a temporal convolution followed by layer normalization and a GELU activation function.

Contextualized representations with Transformers

Quantization module

Model

Feature Encoder

Contextualized representations with Transformers - The output of the feature encoder is fed to a context network which follows the Transformer architecture. we use a convolutional layer to act as relative positional embedding.

Quantization module

Model

Feature Encoder

Contextualized representations with Transformers

Quantization module - For self-supervised training we discretize the output of the feature encoder to a finite set of speech representations via product quantization. Product quantization amounts to choosing quantized representations from multiple codebooks and concatenating them

Training

Masking - We mask a proportion of the feature encoder outputs, or time steps before feeding them to the context network and replace them with a trained feature vector shared between all masked time steps; we do not mask inputs to the quantization module.

Objective

- **Contrastive Loss**
- **Diversity Loss**

Training

Masking

Objective

Contrastive Loss - Given context network output ct centered over masked time step t, the model needs to identify the true quantized latent speech representation qt in a set of quantized candidate representations

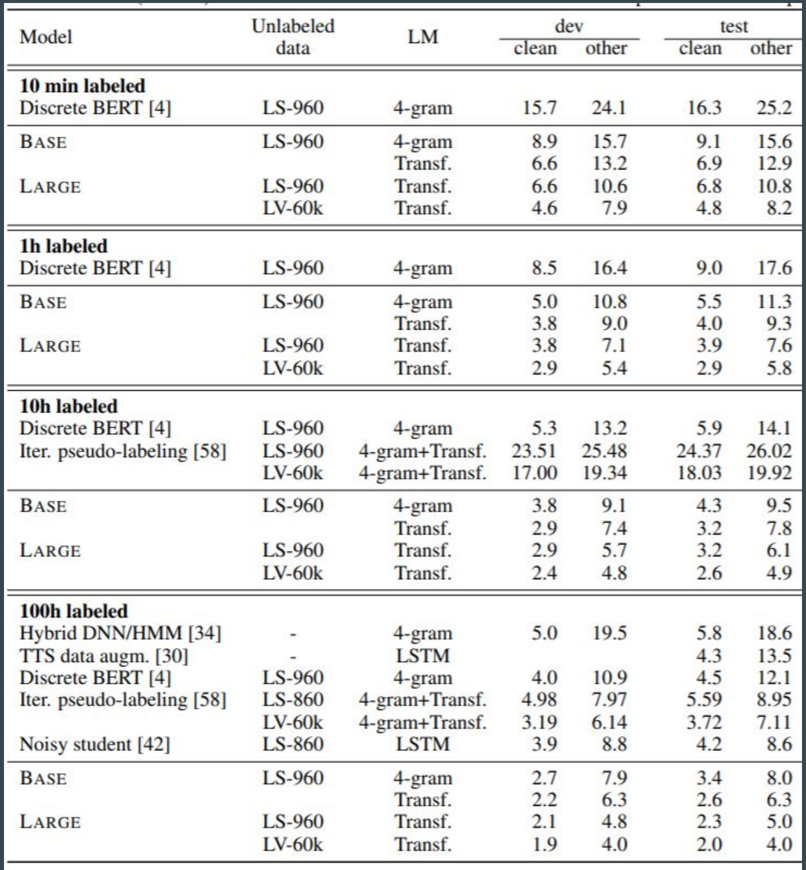
Diversity Loss - The contrastive task depends on the codebook to represent both positive and negative examples and the diversity loss Ld is designed to increase the use of the quantized codebook representations

Fine Tuning

Pre-trained models are fine-tuned for speech recognition by adding a randomly initialized linear projection on top of the context network into C classes representing the vocabulary of the task.

For Librispeech, we have 29 tokens for character targets plus a word boundary token. Models are optimized by minimizing a CTC loss

Results - Low Resource



Model	Unlabeled data	LM	dev clean	dev other	test clean	test other
10 min labeled						
Discrete BERT [4]	LS-960	4-gram	15.7	24.1	16.3	25.2
BASE	LS-960	4-gram	8.9	15.7	9.1	15.6
		Transf.	6.6	13.2	6.9	12.9
LARGE	LS-960	Transf.	6.6	10.6	6.8	10.8
	LV-60k	Transf.	4.6	7.9	4.8	8.2
1h labeled						
Discrete BERT [4]	LS-960	4-gram	8.5	16.4	9.0	17.6
BASE	LS-960	4-gram	5.0	10.8	5.5	11.3
		Transf.	3.8	9.0	4.0	9.3
LARGE	LS-960	Transf.	3.8	7.1	3.9	7.6
	LV-60k	Transf.	2.9	5.4	2.9	5.8
10h labeled						
Discrete BERT [4]	LS-960	4-gram	5.3	13.2	5.9	14.1
Iter. pseudo-labeling [58]	LS-960	4-gram+Transf.	23.51	25.48	24.37	26.02
	LV-60k	4-gram+Transf.	17.00	19.34	18.03	19.92
BASE	LS-960	4-gram	3.8	9.1	4.3	9.5
		Transf.	2.9	7.4	3.2	7.8
LARGE	LS-960	Transf.	2.9	5.7	3.2	6.1
	LV-60k	Transf.	2.4	4.8	2.6	4.9
100h labeled						
Hybrid DNN/HMM [34]	-	4-gram	5.0	19.5	5.8	18.6
TTS data augm. [30]	-	LSTM			4.3	13.5
Discrete BERT [4]	LS-960	4-gram	4.0	10.9	4.5	12.1
Iter. pseudo-labeling [58]	LS-860	4-gram+Transf.	4.98	7.97	5.59	8.95
	LV-60k	4-gram+Transf.	3.19	6.14	3.72	7.11
Noisy student [42]	LS-860	LSTM	3.9	8.8	4.2	8.6
BASE	LS-960	4-gram	2.7	7.9	3.4	8.0
		Transf.	2.2	6.3	2.6	6.3
LARGE	LS-960	Transf.	2.1	4.8	2.3	5.0
	LV-60k	Transf.	1.9	4.0	2.0	4.0

Compiled by - Vaibhav

Original Paper

Results - 960 hours

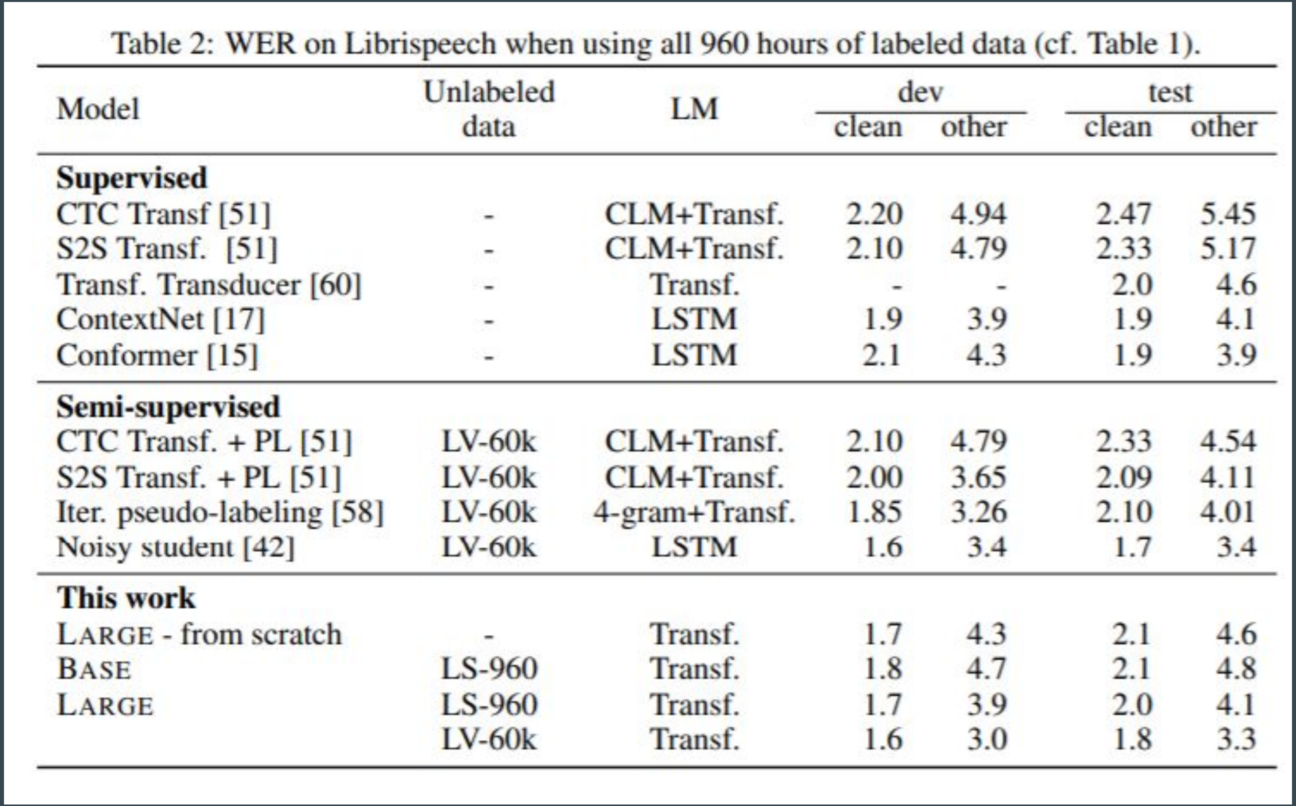


Table 2: WER on Librispeech when using all 960 hours of labeled data (cf. Table 1).

Model	Unlabeled data	LM	dev clean	dev other	test clean	test other
Supervised						
CTC Transf [51]	-	CLM+Transf.	2.20	4.94	2.47	5.45
S2S Transf. [51]	-	CLM+Transf.	2.10	4.79	2.33	5.17
Transf. Transducer [60]	-	Transf.	-	-	2.0	4.6
ContextNet [17]	-	LSTM	1.9	3.9	1.9	4.1
Conformer [15]	-	LSTM	2.1	4.3	1.9	3.9
Semi-supervised						
CTC Transf. + PL [51]	LV-60k	CLM+Transf.	2.10	4.79	2.33	4.54
S2S Transf. + PL [51]	LV-60k	CLM+Transf.	2.00	3.65	2.09	4.11
Iter. pseudo-labeling [58]	LV-60k	4-gram+Transf.	1.85	3.26	2.10	4.01
Noisy student [42]	LV-60k	LSTM	1.6	3.4	1.7	3.4
This work						
LARGE - from scratch	-	Transf.	1.7	4.3	2.1	4.6
BASE	LS-960	Transf.	1.8	4.7	2.1	4.8
LARGE	LS-960	Transf.	1.7	3.9	2.0	4.1
	LV-60k	Transf.	1.6	3.0	1.8	3.3

Conclusion

Our experiments show the large potential of pre-training on unlabeled data for speech processing: when using only 10 minutes of labeled training data, or 48 recordings of 12.5 seconds on average, we achieve a WER of 4.8/8.2 on test-clean/other of Librispeech

On the clean 100 hour Librispeech setup, wav2vec 2.0 outperforms the previous best result while using 100 times less labeled data

We expect performance gains by switching to a seq2seq architecture and a word piece vocabulary.

Try Wav2Vec2 models on HuggingFace

-> All the models - https://huggingface.co/models?search=wav2vec2

-> Test out the ASR performance on the base model via hosted inference or code - https://huggingface.co/facebook/wav2vec2-base-960h

Discussion

- How do you think this architecture can be improved?
- Can the quantization step be removed/ replaced?
- Effect of increasing masking% wrt current 25%?
- Extensibility of Wav2Vec2 to other languages - XLSR model
- Thoughts on the error analysis table in the appendix