Self-supervised Learning with Random-projection Quantizer for Speech Recognition

Chung-Cheng Chiu*, James Qin*, Yu Zhang, Jiahui Yu, Yonghui Wu

Google Research

Self-supervised learning for ASR

Motivation: designing a BERT-style pre-training for ASR

- Challenge: BERT use discrete tokens but speech signals are continuous
- How can we bridge such a gap?

Previous belief: "One must learns the content representation of the speech"

“We need representation learning for self-supervised learning”

But we now need to develop both self-supervised learning AND representation learning

The two objectives are not necessarily compatible and limit the design of the model architecture

Can we challenge the status quo and avoid representation learning?

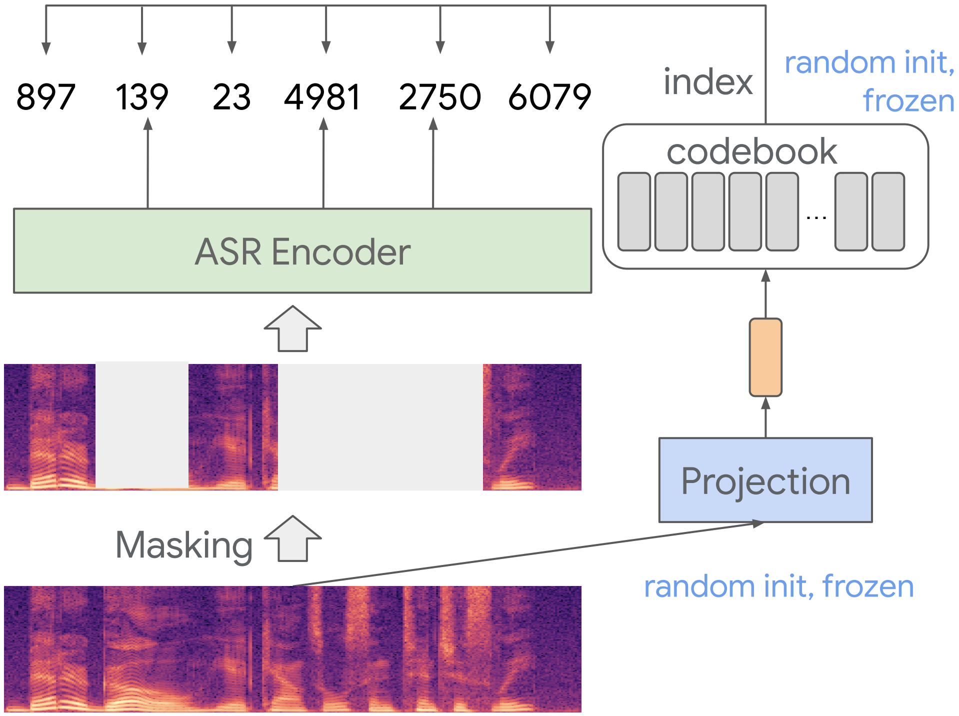
BEST-RQ

BERT-based Speech pre-Training with Random-projection Quantizer

Masked language modeling

Generate quantized prediction targets with randomly-initialized codebook and projection matrix

Freeze the codebook and the projection matrix



LibriSpeech

Non-streaming

Method	Size (B)	No LM				With LM			
		dev	dev-other	test	test-other	dev	dev-other	test	test-other
wav2vec 2.0 (Baevski et al., 2020b)	0.3	2.1	4.5	2.2	4.5	1.6	3.0	1.8	3.3
HuBERT Large (Hsu et al., 2021)	0.3	–	–	–	–	1.5	3.0	1.9	3.3
HuBERT X-Large (Hsu et al., 2021)	1.0	–	–	–	–	1.5	**2.5**	1.8	2.9
w2v-Conformer XL (Zhang et al., 2020)	0.6	1.7	3.5	1.7	3.5	1.6	3.2	**1.5**	3.2
w2v-BERT XL (Chung et al., 2021)	0.6	**1.5**	2.9	**1.5**	**2.9**	**1.4**	2.8	**1.5**	2.8
BEST-RQ (Ours)	0.6	**1.5**	**2.8**	1.6	**2.9**	**1.4**	2.6	**1.5**	**2.7**

Streaming

Method	Size (B)	dev	dev-other	test	test-other	Relative latency (ms)
Conformer 0.1B	0.1	4.1	10.3	4.5	9.8	0
Conformer 0.6B	0.6	3.9	9.8	4.4	9.4	15.3
Non-Streaming pre-train						
wav2vec 2.0	0.6	2.6	7.3	3.0	7.2	-10.1
w2v-BERT	0.6	2.8	7.2	3.3	6.9	-0.7
BEST-RQ (Ours)	0.6	**2.5**	**6.9**	**2.8**	**6.6**	-16.3
Streaming pre-train						
wav2vec 2.0	0.6	2.7	8.0	2.9	7.9	-130.6
w2v-BERT	0.6	2.7	8.4	3.0	8.1	-117.1
BEST-RQ (Ours)	0.6	**2.5**	**6.9**	**2.8**	**6.6**	**-130.9**

Pre-train on LibriLight, fine-tune on LibriSpeech

Multilingual LibriSpeech

Exp.	Languages								Avg.
	en	de	nl	fr	es	it	pt	pl	
MLS-full									
wav2vec 2.0 from XLSR-53 (Conneau et al., 2020)	-	7.0	10.8	7.6	6.3	10.4	14.7	17.2	10.6
w2v-BERT from JUST (Bai et al., 2021)	6.6	4.3	9.9	5.0	3.8	9.1	14.6	8.1	7.8
JUST (Bai et al., 2021) (co-train)	6.5	4.1	9.5	5.2	3.7	8.8	8.0	6.6	6.5
w2v-BERT (0.6B)	5.5	4.3	10.9	5.6	4.5	10.1	13.4	11.2	8.2
BEST-RQ (Ours, 0.6B)	6.8	4.1	9.7	5.0	4.9	7.4	9.4	5.2	6.6
MLS-10hrs									
XLSR-53 (Conneau et al., 2020)	14.6	8.4	12.8	12.5	8.9	13.4	18.2	21.2	13.8
XLS-R(0.3B) (Babu et al., 2021)	15.9	9.0	13.5	12.4	8.1	13.1	17.0	13.9	12.8
XLS-R(1B) (Babu et al., 2021)	12.9	7.4	11.6	10.2	7.1	12.0	15.8	10.5	10.9
XLS-R(2B) (Babu et al., 2021)	14.0	7.6	11.8	10.0	6.9	12.1	15.6	9.8	11.0
w2v-BERT (0.6B)	12.7	7.0	12.6	8.9	5.9	10.3	14.6	6.9	9.9
BEST-RQ (Ours, 0.6B)	12.8	7.4	12.7	9.6	5.4	9.9	12.1	7.1	9.6

Pre-train on XLS-R unsupervised data without VoxLingua-107.

Large-scale Multilingual Set

Exp.	Avg. on 15 langs (VS)
Baseline (0.6B)	12.6
wav2vec 2.0 (0.6B)	12.0
w2v-bert (0.6B)	11.5
BEST-RQ (Ours) (0.6B)	**10.9**

Pre-train on Multilingual YouTube (250k~800k hrs per language).
Fine-tune on Multilingual Voice Search (1k hrs per language).
Same recipe as (Zhang et al., 2021)

Better understand random-projection quantizers

Do random-projection quantizers provide good speech representations?

Study: compare two types of quantizers and two types of experiments

Two quantizers

- Random-projection quantizer: No representation learning
- VQ-VAE: Has representation learning

Two experiments

- Use quantized code as **input to train ASR:** Tells us the representation quality
- Use quantized code as self-supervised learning prediction **targets:** Tells us the effectiveness for self-supervised learning

Quantization quality

Configuration	Quantizer size (M)	Direct ASR WER				Pretrain-finetune WER			
		dev	dev-other	test	test-other	dev	dev-other	test	test-other
Random quantizer	1	58.8	78.8	57.9	72.8	1.5	**2.8**	**1.6**	**2.9**
Projection VQ-VAE	1	61.4	74.8	60.9	75.2	1.5	**2.8**	**1.6**	**2.9**
Transformer VQ-VAE	10	**17.8**	**35.8**	**17.6**	**36.1**	**1.4**	2.9	**1.6**	3.1

- As input: VQ-VAE provides much better quality
- As targets: no difference in self-supervised learning

Representation quality does not directly translate to self-supervised learning quality

Hypothesis: self-supervised learning learn to mitigate the quality gap from sufficient amount of unsupervised data

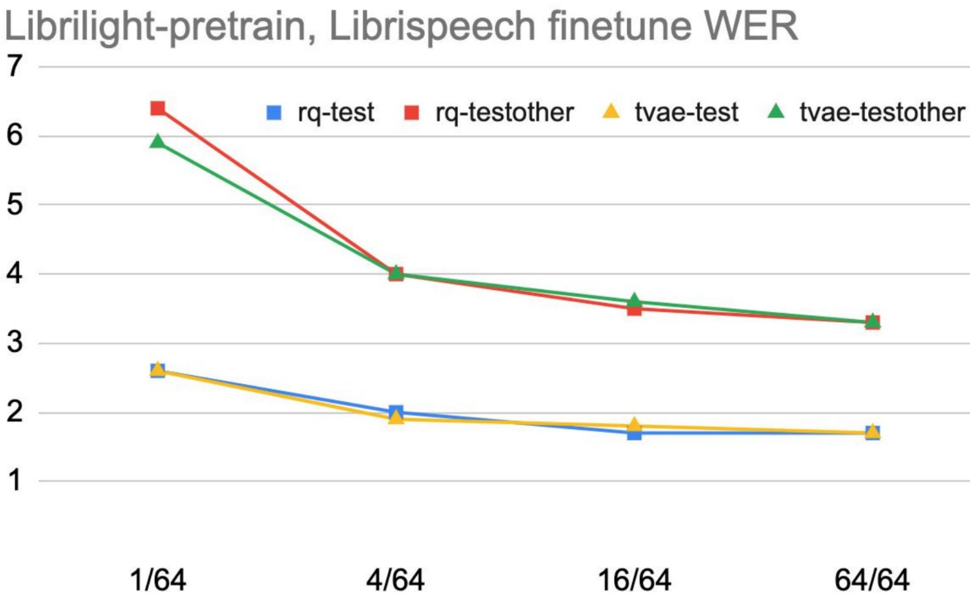
Study: compare different unsupervise data size

Quantization quality matters more when unsupervised data size is limited

The gap disappear as the unsupervised data size increase

rq: random-projection quantizer

tvae: transformer VQ-VAE



Conclusions

- Random quantizer is simple and effective for self-supervised learning
  - Does not require representation learning
- Random quantizer do not capture content information as efficient as other learned representations
  - But it capture essential information for self-supervised learning
- Codebook utilization is the most critical metric for pre-training