wav2vec 2.0 : Self-Supervised Learning for Speech Recognition

https://arxiv.org/pdf/2006.11477

0. Introduction

- 음성 인식에는 대규모 레이블 데이터가 필요하지만, 대부분 언어에서 충분한 레이블 데이터가 없음.
- 인간이 언어를 습득하듯, 음성 데이터의 레이블 없이도 표현 학습 가능 필요.
- 지도학습 기반 음성 인식은 레이블 데이터에 의존하며, 이전 자가 지도 학습 방법은 두 단계 학습 또는 입력 특징 복원에 의존.
- 심 기여: 음성 데이터로부터 self-supervised 학습을 통해 강력한 표현 학습, 소량 레이블 데이터로도 높은 성능 달성.

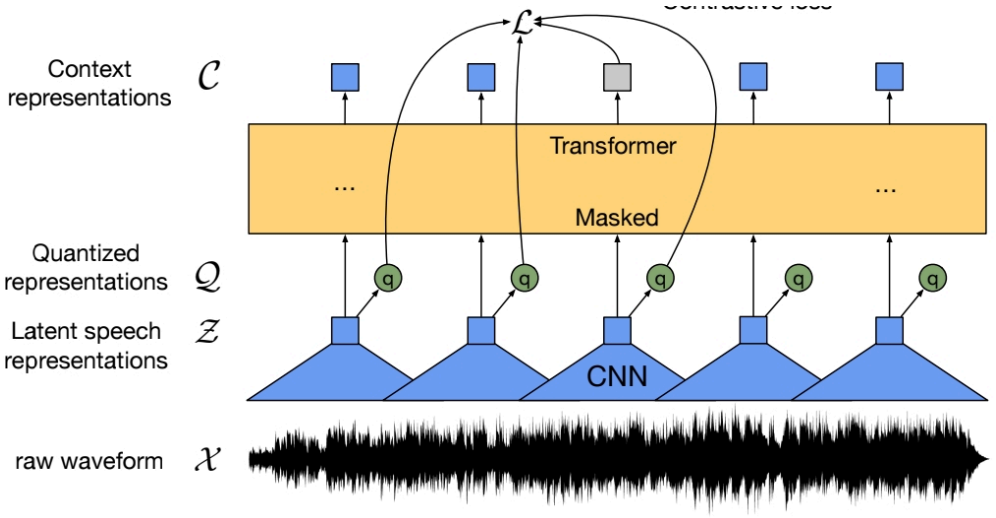
1. Overview

- 음성 잠재 표현 마스킹 후, 양자화된 벡터 간 대조 학습 수행.
- 모델 구조 : Feature Encoder (컨볼루션) → Context Network (Transformer) → Quantization Module.
- 소량 레이블 환경에서도 고성능 음성 인식 가능성 검증.

2. Challenges

- 음성 데이터 레이블 부족 문제.
- 연속 음성 신호의 잠재 표현 학습 및 양자화 어려움.
- Transformer 기반 모델의 장기 의존성 학습 문제.
- 단일 모델에서 양자화와 컨텍스트 표현 학습 동시 수행 난이도.

3. Method



- Feature Encoder : 오디오 입력 → 다층 시간 컨볼루션 → Layer Norm → GELU 활성화.
- Context Network : Transformer 구조, 상대 위치 임베딩 위해 1D 컨볼루션 사용.
- Quantization Module : Product Quantization + Gumbel Softmax, 출력 벡터 이산화.
- Pre-training Objective : 마스킹된 latent timestep 예측, 대조 학습 수행.
- Fine-tuning : 소량 레이블 데이터에서 CTC 손실 기반 학습.

4. Experiments

- 데이터셋 : Librispeech 960h, 10분~100시간 레이블 서브셋.
- 실험 설계 : 마스크 비율, 코드북 그룹 수 등 하이퍼파라미터 탐색.
- 비교 모델 : vq-wav2vec, self-training 기반 모델.
- 평가 지표 : Word Error Rate (WER).
- 분석 : 레이블 데이터량 변화에 따른 성능, Ablation study로 마스킹/양자화 영향 분석.

5. Results

Model	Unlabeled data	LM	dev		test	
			clean	other	clean	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

Model	Unlabeled data	LM	dev		test	
			clean	other	clean	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

	dev PER	test PER
CNN + TD-filterbanks [59]	15.6	18.0
PASE+ [47]	-	17.2
Li-GRU + fMLLR [46]	–	14.9
wav2vec [49]	12.9	14.7
vq-wav2vec [5]	9.6	11.6
This work (no LM)		
LARGE (LS-960)	7.4	8.3

	avg. WER	std.
Continuous inputs, quantized targets (Baseline)	7.97	0.02
Quantized inputs, quantized targets	12.18	0.41
Quantized inputs, continuous targets	11.18	0.16
Continuous inputs, continuous targets	8.58	0.08

- 960h 레이블 : 1.8/3.3 WER (clean/other).
- 100h 레이블 : 이전 SOTA 대비 성능 향상, 레이블 100배 절약.
- 10분 레이블 : 4.8/8.2 WER 달성, 극저자원 환경에서도 실용적 성능.

- Ablation study : 양자화 + 컨텍스트 학습 결합이 단일 단계 학습 대비 우수.
- 실무 적용 : 소규모 레이블 환경에서도 실시간 음성 인식 가능.

6. Insight

- 양자화된 잠재 표현과 컨텍스트 표현의 공동 학습이 핵심.
- 소량 레이블 환경에서도 self-supervised 학습으로 높은 성능 달성 가능.
- 기존 두 단계 학습 대비 학습 단순화 및 성능 향상.
- 후속 연구: 다른 언어, 방언, 잡음 환경에서도 초저자원 음성 인식 확장 가능.