Self-Supervised Learning for speech recognition with Intermediate layer supervision

李思瑞

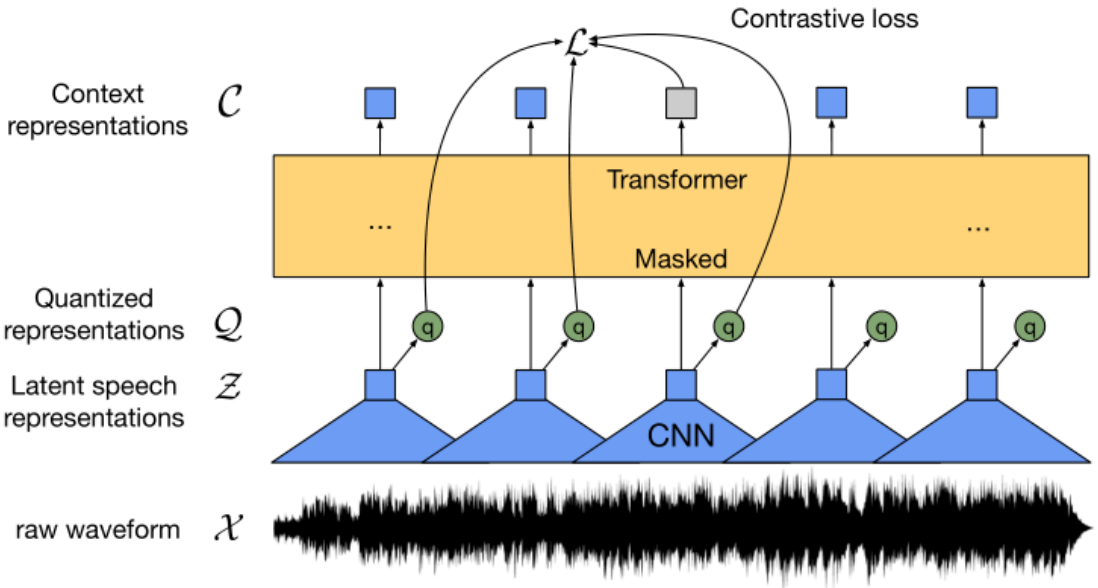
2022/1/5

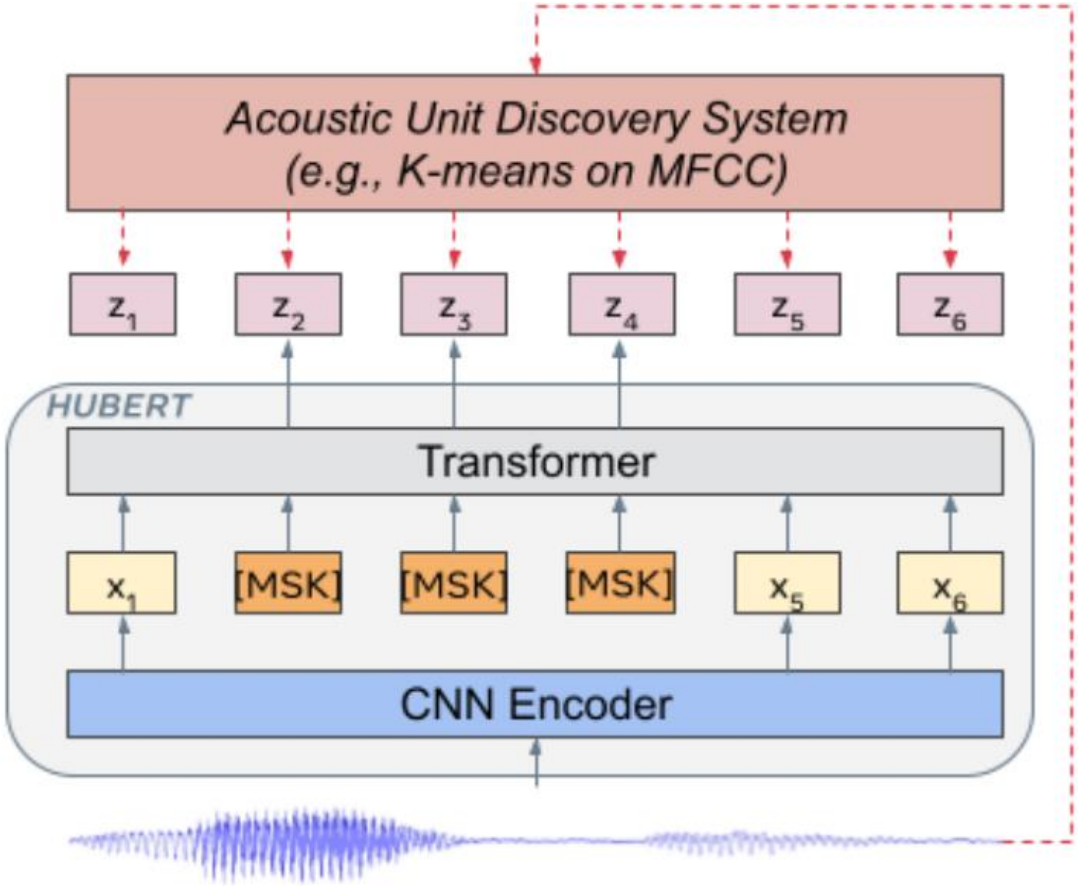
https://arxiv.org/abs/2112.08778

研究背景

- 无监督学习在语音领域逐渐成为热点
- wav2vec 2.0 HuBERT
- 网络各层次学习的特征都是不同
- 模型可以容纳的学习内容容量受限

从wav2vec2.0到HuBERT





从wav2vec2.0到HuBERT

- 在线聚类 到 离线聚类
- 对比损失 到 交叉熵损失
- HuBERT增加了Xlarge版本模型，参数10亿

$$L_m(f; X, M, Z) = \sum_{t \in M} \log p_f\left(z_t \mid \tilde{X}, t\right)$$

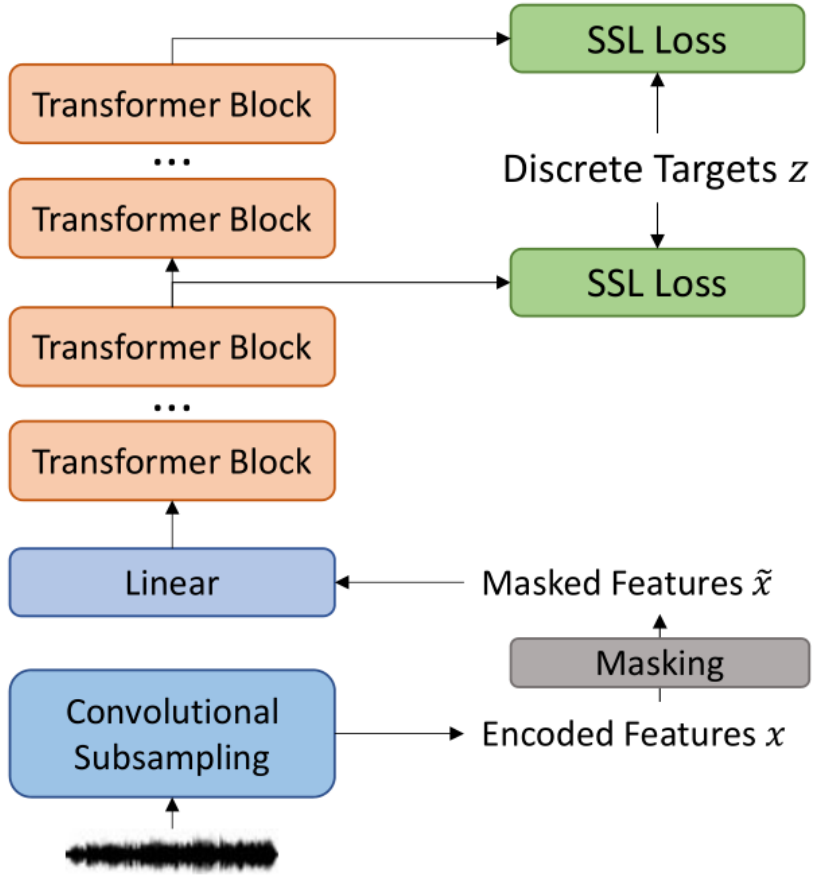
$$L = \alpha L_m + (1 - \alpha) L_u$$

$$p(c \mid \mathbf{h}_t) = \frac{\exp(\text{sim}(\mathbf{W}\mathbf{h}_t, \mathbf{e}_c)/\tau)}{\sum_{c'=1}^{C} \exp(\text{sim}(\mathbf{W}\mathbf{h}_t, \mathbf{e}_{c'})/\tau)}$$

intermediate Layer Supervision for Self-Supervised Learning (ILS-SSL)

$$L = -\sum_{l \in K} \sum_{t \in M} \log p^l\left(z_t \mid \mathbf{h}_t^l\right)$$

$$p^l\left(c \mid \mathbf{h}_t^l\right) = \frac{\exp\left(\mathrm{sim}\left(\mathbf{W}^l \mathbf{h}_t^l, \mathbf{e}_c^l\right)/\tau\right)}{\sum_{c'=1}^{C} \exp\left(\mathrm{sim}\left(\mathbf{W}^l \mathbf{h}_t^l, \mathbf{e}_{c'}^l\right)/\tau\right)}$$



相对位置编码

$$\mathbf{q}_i, \mathbf{k}_i, \mathbf{v}_i = \mathbf{h}_i^{l-1}\mathbf{W}^Q, \mathbf{h}_i^{l-1}\mathbf{W}^K, \mathbf{h}_i^{l-1}\mathbf{W}^V$$

$$a_{ij} \propto \exp\left(\frac{\mathbf{q}_i \cdot \mathbf{k}_j}{\sqrt{d}} + r_{i-j}\right)$$

$$\tilde{\mathbf{h}}_i^l = \sum_{j=1}^{T} a_{ij}\mathbf{v}_j$$

训练策略

- 对于第一次迭代，通过对训练数据的MFCC特征进行聚类来确定离散目标Z

- 对于第二次迭代，通过对第一次迭代模型中学习到的潜在表示进行聚类，生成新一代训练目标

训练配置

- BASE

BASE模型包含12层，第一次迭代在39维MFCC上聚类到100类，第二次迭代在第6层的输出上聚类到500类

- LARGE

LARGE模型包含24层，第一次迭代在39维MFCC上聚类到100类，第二次迭代在第9层的输出上聚类到500类

Model	LM	test-clean	test-other
1-hour labeled			
wav2vec 2.0 [11]	None	24.5	29.7
HuBERT	None	20.9	27.5
WavLM [27]	None	24.5	29.2
DeCoAR 2.0 [4]	4-gram	13.8	29.1
DiscreteBERT[28]	4-gram	9.0	17.6
wav2vec 2.0 [11]	4-gram	5.5	11.3
HuBERT [13]	4-gram	6.1	11.3
WavLM [27]	4-gram	5.7	10.8
ILS-SSL	None	17.9	23.1
ILS-SSL	4-gram	5.4	10.2
10-hour labeled			
wav2vec 2.0 [11]	None	11.1	17.6
HuBERT	None	10.1	16.8
WavLM [27]	None	9.8	16.0
DeCoAR 2.0 [4]	4-gram	5.4	13.3
DiscreteBERT[28]	4-gram	5.9	14.1
wav2vec 2.0 [11]	4-gram	4.3	9.5
HuBERT [13]	4-gram	4.3	9.4
WavLM [27]	4-gram	4.3	9.2
ILS-SSL	None	8.3	13.6
ILS-SSL	4-gram	3.8	8.1
100-hour labeled			
wav2vec 2.0 [11]	None	6.1	13.3
HuBERT	None	6.3	13.2
WavLM [27]	None	5.7	12.0
DeCoAR 2.0 [4]	4-gram	5.0	12.1
DiscreteBERT[28]	4-gram	4.5	12.1
wav2vec 2.0 [11]	4-gram	3.4	8.0
HuBERT [13]	4-gram	3.4	8.1
WavLM [27]	4-gram	3.4	7.7
ILS-SSL	None	4.7	10.1
ILS-SSL	4-gram	3.0	6.9

Table 1. Model comparisons in the BASE setting.

Table 2. Model comparisons in the LARGE setting.

Model	LM	test-clean	test-other
1-hour labeled			
wav2vec2.0 [11]	None	17.2	20.3
HuBERT	None	17.4	20.3
wav2vec 2.0 [11]	4-gram	3.8	7.1
HuBERT	4-gram	3.9	6.6
wav2vec 2.0 [11]	Transf	2.9	5.8
HuBERT [13]	Transf	2.9	5.4
ILS-SSL	None	14.3	16.9
ILS-SSL	4-gram	3.6	6.5
ILS-SSL	Transf	2.8	5.3
10-hour labeled			
wav2vec2.0 [11]	None	6.3	10.0
HuBERT	None	6.2	9.6
wav2vec 2.0 [11]	4-gram	3.0	5.8
HuBERT	4-gram	2.9	5.4
wav2vec 2.0 [11]	Transf	2.6	4.9
HuBERT [13]	Transf	2.4	4.6
ILS-SSL	None	6.1	9.1
ILS-SSL	4-gram	2.8	5.2
ILS-SSL	Transf	2.5	4.5
100-hour labeled			
wav2vec2.0 [11]	None	3.1	6.3
HuBERT	None	2.9	6.0
wav2vec 2.0 [11]	4-gram	2.3	4.6
HuBERT	4-gram	2.3	4.5
wav2vec 2.0 [11]	Transf	2.0	4.0
HuBERT [13]	Transf	2.1	3.9
ILS-SSL	None	2.9	5.8
ILS-SSL	4-gram	2.2	4.5
ILS-SSL	Transf	2.0	4.0
960-hour labeled			
Transformer-CTC [29]	Transf	2.5	5.5
Transformer-S2S [29]	Transf	2.3	5.2
Transformer-T [30]	Transf	2.0	4.6
Conformer-T[31]	LSTM	1.9	3.9
wav2vec 2.0 [11]	None	2.2	4.5
HuBERT	None	2.1	4.3
wav2vec 2.0 [11]	4-gram	2.0	3.6
HuBERT	4-gram	2.0	3.7
wav2vec 2.0 [11]	Transf	1.8	3.3
HuBERT [13]	Transf	1.9	3.3
ILS-SSL	None	1.9	3.8
ILS-SSL	4-gram	1.9	3.4
ILS-SSL	Transf	1.8	3.2

聚类效果分析

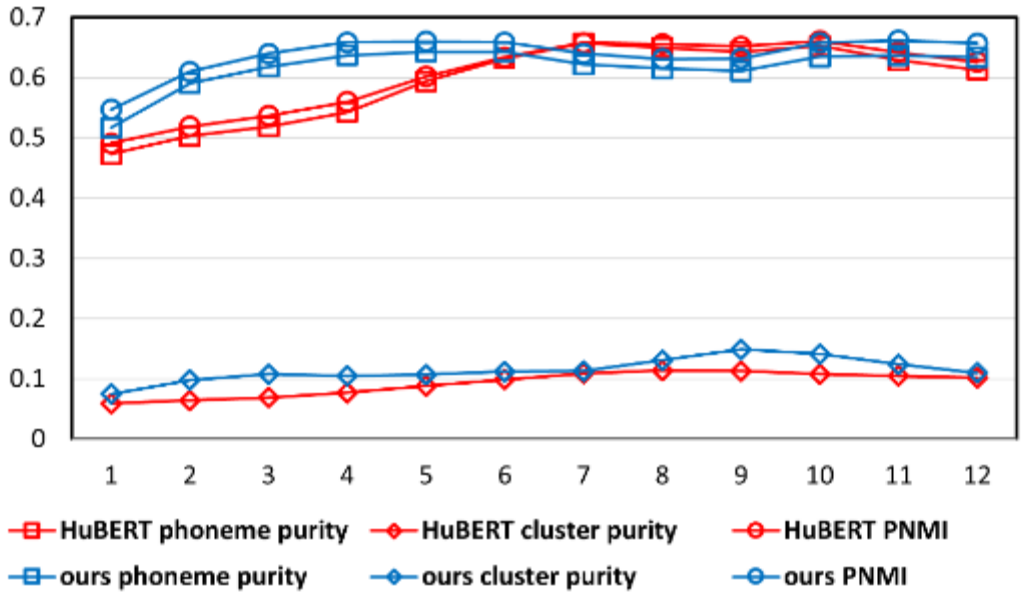


Fig. 2. Quality of the cluster assignments obtained by running k-means clustering on features extracted from each Transformer layer of the BASE model (after the 2nd iteration).

非ASR任务

Method	#Params	Corpus	Speaker			Content				Semantics			ParaL
			SID	ASV	SD	PR	ASR	KS	QbE	IC	SF		ER
			Acc ↑	EER ↓	DER ↓	PER ↓	WER ↓	Acc ↑	MTWV ↑	Acc ↑	F1 ↑	CER ↓	Acc ↑
FBANK	0	-	8.5E-4	9.56	10.05	82.01	23.18	8.63	0.0058	9.10	69.64	52.94	35.39
HuBERT BASE [13]	94.68M	LS 960 hr	**81.42**	**5.11**	**5.88**	5.41	6.42	96.30	0.0736	98.34	88.53	25.20	64.92
ILS-SSL BASE	94.68M	LS 960 hr	79.29	5.24	6.31	**5**	**5.45**	**4.38**	**0.0789**	**98.47**	**89.16**	**24.29**	**65.79**

Table 3. Universal speech representation evaluation on SUPERB benchmark. ParaL denote Paralinguistics aspect of speech.

Speaker Identification (SID), Auto-matic Speaker Verification (ASV), Speaker Diarization (SD),Phoneme Recognition (PR), Automatic Speech Recognition(ASR), Keyword Spotting (KS), Query by Example SpokenTerm Detection (QbE), Intent Classification (IC), Slot Filling(SF), Emotion Recognition (ER)

谢谢大家