Speech Representation Learning with Contrastive Predictive Coding

Alena Pestova

Consultant: Alexander Samarin, HUAWEI

St. Petersburg School of Physics, Mathematics and Computer Science
Higher School of Economics

Master thesis
22.06.2023

Introduction

- Self-supervised models have become increasingly popular in recent years.
- Their advantage is no need in labeled data (can be trained on large amount of unlabeled data).
- After pre-training, such models are used as feature extractors for downstream tasks.

Self-supervised Models in Speech Processing

- **Contrastive Predictive Coding (CPC)**
  Problems: it is not clear how generalizable are representations from CPC, whether this method can be used for very different speech processing tasks.
- Large models based on Transformer: **Wav2vec2**[Baevski et al. 2020] and **HUBERT**[Hsu et al. 2021].
  Problems: big size, which can be a problem in their use in real tasks, for example, in different speech processing tasks in phones (ex: speaker verification, speech recognition)

Goal and Tasks

Goal: Explore the applicability of Contrastive Predictive Coding models for learning generalizable speech representations and evaluate its competitiveness with large language models.

Tasks:

- Reproduce the CPC model from the original article and its results.
- Experiment with the architecture of the original model, with increasing the number of parameters and training dataset.
- Compare the CPC models with each other and with baselines on additional downstream tasks.

Contrastive Predictive Coding

paper: *Representation Learning with Contrastive Predictive Coding* [Oord, Li, and Vinyals 2018]

CPC model architecture

- encoder (CNN) g_{enc}
- context network (GRU) g_{ar}
- networks for predicting K(K=12) timestamps W_k

x - audio of length 20480.

Training step:

- $z = g_{enc}(x)$
- $c_t = g_{ar}(z_t), z < t$
- $z_{pred\_t+k}(c_t) = W_k c_t$ for each k.

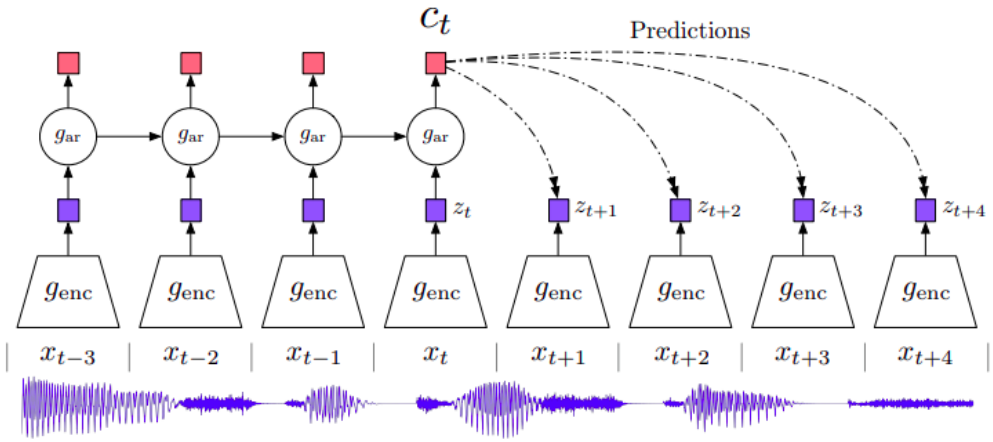

Figure: Visualization of predictive coding [Oord, Li, and Vinyals 2018].

Either z or c embeddings can be used as speech representations later.

Training CPC model: InfoNCE loss

Training objective:

- minimizing the distance $d(z_{t+k}, z_{pred\_t+k}(c_t))$
- maximizing $d(z_n, z_{pred\_t+k}(c_t))$ for $x_n \in X$ - a set of negative examples, sampled from other sequences.

InfoNCE loss

$$L_N = - \mathop{\mathbb{E}}_{x \in X} \log \frac{\exp(z_{t+k}^T z_{pred_{t+k}})}{\sum_{x \in X, z = g_{enc}(x)} \exp(z_{t+k}^T z_{pred_{t+k}})} \tag{1}$$

where X - a set of all samples, including one positive example and $N - 1$ negative ones.

The authors show that minimization of this loss also minimizes mutual information between c_t and x_{t+k}.

Reproducing CPC model and my modifications

Original architecture:

- **Encoder** g_{enc}: 5 residual blocks with convolution layers (kernel sizes [10, 8, 4, 4, 4] and strides [5, 4, 2, 2, 2], number of channels - 512). Batch normalization and ReLU activations are used between the blocks.
- **Context network** g_{ar}: GRU, hidden state 256.
- **Prediction networks**: 12 Linear layers (Linear(256, 512)).

Training dataset: English corpus of clean audios from audiobooks **LibriSpeech** [Panayotov et al. 2015], *train-clean-100* subset.

Modifications:

- **augmentation**: random noise from MUSAN [Snyder, Chen, and Povey 2015] and reverbation.
- train on **bigger datasets**:
  - 960h of LibriSpeech
  - Voxceleb2 [Chung, Nagrani, and Zisserman 2018] (1.2 million utterances from 6112 speakers).
- adding **normalization** of raw audio and layer normalization to encoder (instead of batch normalization) - like in wav2vec2 encoder.
- (scaling the number of parameters in encoder did not influence the results)

Downstream tasks for evaluation

From original paper:

- *LBS*: Speaker Identification on LibriSpeech train-clean-100 (251 classes)
- *Phone*: Phoneme Classification on LibriSpeech train-clean-100 (41 classes)

Additional:

- **Speaker identification**
  LBS_test: LibriSpeech test-clean subset (40 speakers);
  Vox1: Voxceleb1 [Nagrani, Chung, and Zisserman 2017] (1251 speakers);
  Vox1_40: subset of Voxceleb1 (40 speakers);
  SHAL: SHALCAS22A [1] (60 speakers).
- **Age—gender binary classification**
  Gender: Samrómur Icelandic Speech corpus [Mollberg et al. 2020] (males—females);
  Age: Samrómur Children [Mena et al. 2021] (younger than 12 years — older).
- **Keyword spotting**
  Cmd1 and *Cmd2*: Speechcommands1 and Speechcommands2 [Warden 2018] dataset (30 and 35 words)

[1] SHALCAS22A, a free Chinese Mandarin corpus by Shanghai Acoustics Laboratory, CAS and Wuxi Sandu Intelligent Technology Co., Ltd., 2022; `https://www.openslr.org/138/`

Results: Experiments with Architecture and Training Dataset

Classification: frozen CPC features + one linear layer.

model	Phone	Age	Gender	SHAL	Cmd1	Cmd2	Vox1	Vox1_40
Base arhitecture, LibriSpeech 100h								
c embeddings								
CPC_noaug	0.51	0.81	0.95	0.62	0.57	0.58	0.09	0.41
z embeddings								
CPC_noaug	0.51	0.84	0.95	0.80	0.65	0.66	0.12	0.40
+augmentation								
CPC	0.53	0.86	0.96	0.89	0.74	0.73	0.14	0.48
+more training data								
CPC_960h	0.54	0.85	**0.97**	0.86	0.75	0.74	0.14	0.53
CPC_vox2	0.54	0.86	**0.97**	0.86	0.75	0.74	0.16	0.54
Wav2vec2-like encoder								
CPC_norm_100h	0.53	0.84	**0.97**	0.89	0.77	0.75	0.23	0.63
CPC_norm_960h	0.54	**0.87**	**0.97**	0.93	0.78	0.76	0.25	**0.74**
concatting z and c embeddings								
CPC_norm_960h	**0.61**	0.85	**0.97**	**0.95**	**0.86**	**0.84**	**0.32**	0.72
CPC_norm_vox2	0.59	0.84	**0.97**	0.90	0.85	**0.84**	0.30	0.72

Table: Ablation studies results. Classification accuracy on downstream tasks. *z* features from CPC are used if the opposite is not stated.

Results: Comparison with Baselines on Downstream Tasks

model	LBS speaker test	LBS speaker train	LBS phone
CPC_noaug(c, paper)	-	0.97	**0.65**
CPC_noaug(c)	**1.00**	**1.0**	0.51
CPC_noaug(z)	0.99	**1.0**	0.51

Table: Clasification accuracy on the downstream tasks using encoder and context embeddings.

model	Phone	Age	Gender	SHAL	Cmd1	Cmd2	Vox1	Vox1_40
CPC-supervised	0.80	0.89	0.98	0.88	0.94	0.93	0.29	0.54
MFCC	0.42	**0.85**	0.93	0.48	0.26	0.28	0.12	0.61
HUBERT_base	-	0.84	**0.98**	0.82	**0.97**	**0.96**	**0.67**	**0.93**
Wav2vec2_base	-	**0.85**	**0.98**	0.88	0.96	**0.96**	0.63	0.86
CPC_norm_960h	**0.61**	**0.85**	0.97	**0.95**	0.86	0.84	0.32	0.72

Table: Comparison of the CPC model with baselines on downstream tasks. Classification accuracy on downstream tasks. Concatenated *z* and *c* embeddings from CPC are used.

Number of Parameters

model	Total	Encoder	Context	Prediction
wav2vec2—HUBERT (base)	94.4			
CPC	4.6	2.5	0.6	1.7
CPC_norm	7.4	5.3	0.6	1.7

Table: The number of model parameters.

Interactive Speaker Identification

model	Random	RL agent
x-vectors	0.75	0.91
CPC_norm_960h	**0.945**	**0.99**

Table: Accuracy of speaker identification with requesting 2 words chosen randomly or with the use of trained RL agent. The total number of speakers is 20.

Conclusion

- The model performance was improved on all downstream tasks.
- CPC features perform quite well on cleaner datasets, but there are problems with noisy data such as Voxceleb.
- Larger models perform better on harder (noisier) datasets, although for other tasks, CPC works at the same level. So, in conditions of limited resources, the use of the CPC can be justified in these cases.
- Probably, changing the context network architecture to more complex model (like Transformer) could give better results, however, the advantage in the small size of the model will be lost in this case.

References I

Baevski, Alexei et al. (2020). *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations*. arXiv: 2006.11477 [cs.CL].

Chung, J. S., A. Nagrani, and A. Zisserman (2018). "VoxCeleb2: Deep Speaker Recognition". In: *INTERSPEECH*.

Hsu, Wei-Ning et al. (2021). *HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units*. arXiv: 2106.07447 [cs.CL].

Mena, Carlos et al. (2021). "Samrómur Children Icelandic Speech 21.09". In: Reykjavik University: Language and Voice Lab.

Mollberg, David Erik et al. (May 2020). "Samrómur: Crowd-sourcing Data Collection for Icelandic Speech Recognition". English. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 3463–3467. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.425.

Nagrani, A., J. S. Chung, and A. Zisserman (2017). "VoxCeleb: a large-scale speaker identification dataset". In: *INTERSPEECH*.

Oord, Aaron van den, Yazhe Li, and Oriol Vinyals (2018). *Representation Learning with Contrastive Predictive Coding*. DOI: 10.48550/ARXIV.1807.03748. URL: https://arxiv.org/abs/1807.03748.

References II

Panayotov, Vassil et al. (2015). “Librispeech: An ASR corpus based on public domain audio books”. In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964.

Snyder, David, Guoguo Chen, and Daniel Povey (2015). *MUSAN: A Music, Speech, and Noise Corpus*. arXiv: 1510.08484 [cs.SD].

Warden, Pete (2018). *Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition*. arXiv: 1804.03209 [cs.CL].