Towards Better Domain Adaptation for Self-supervised Models: A Case Study of Child ASR

Ruchao Fan, Yunzheng Zhu, Jinhan Wang, Abeer Alwan

fanruchao@g.ucla.edu

Published in IEEE Journal of Selected Topics in Signal Processing, 2022

ICASSP 2023, 4 - 10 JUNE, RHODES ISLAND, GREECE

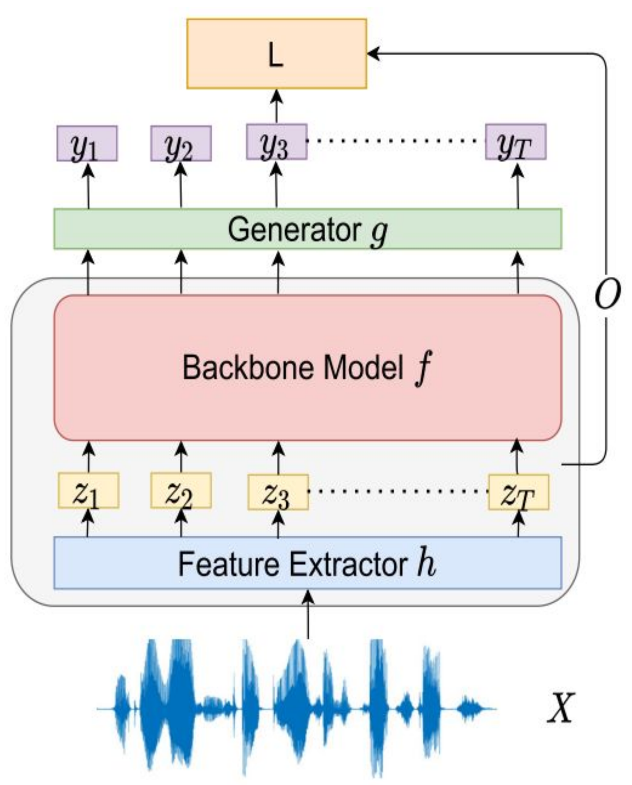
Introduction

- Recently, self-supervised learning (SSL) has achieved a considerable success for low-resource ASR tasks.
- However, SSL models are biased to the pretraining data.
- Current solutions:
  - Including target domain data pretraining. However, the target domain is unknown during the pretraining, and retraining the SSL model is time-consuming,
  - Additional pretraining of the SSL model with the target domain data. But this requires large amounts of the target domain data, catastrophic forgetting problem
- Can we use the finetuning data (typically in a low-resource setting) to reduce the effect of domain shifting in the pretrained models? -> **Yes!**
- **Novel Contributions in this paper:**
  - We propose a **d**omain **r**esponsible **a**daptation and **f**ine**t**uning (DRAFT) framework to reduce the domain shifting in both the causal and non-causal pretrained models with the finetuning data.
  - For causal SSL (autoregressive predictive coding, APC), we propose to use multiple temporally-shifted sequences as a multi-task training objective APC, denoted as E-APC.
  - For non-causal SSL (Bidirectional APC, Bi-APC), we extend it to transformer architectures and explore various parameter-sharing solutions to achieve Bi-APC mechanism for a transformer.

A General SSL Framework

- X: raw waveform of an utterance
- Z: speech representations for each frame
- h: feature extractor
- f: backbone model
- g: generator
- Y: model output
- O: an operation
- A general SSL loss can be written as:

$$L_{\mathrm{SSL}} = L(f(h(X)), O(h(X)))$$



Proposed Methods

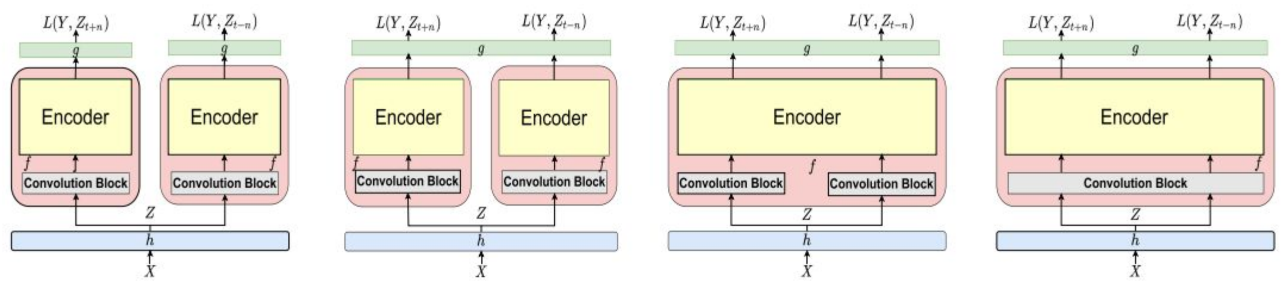
1. An extension of Autoregressive Predictive Coding (E-APC)

- APC uses one temporally-shifted sequence during pretraining. A model may learn differently with different temporal lags n:
  - Local smoothness of the signal with a small value of n
  - Global structure with a large value of n
- E-APC: reformulate APC as a multi-task training with multiple temporally-shifted sequences as the targets.

$$L_{\text{E-APC}} = \sum_{n=s}^{s+k} L^n_{\text{APC}} = \sum_{n=s}^{s+k} \sum_{t=1}^{T-n} (|y_t - z_{t+n}|_p)$$

2. Bi-APC for Non-causal Transformer

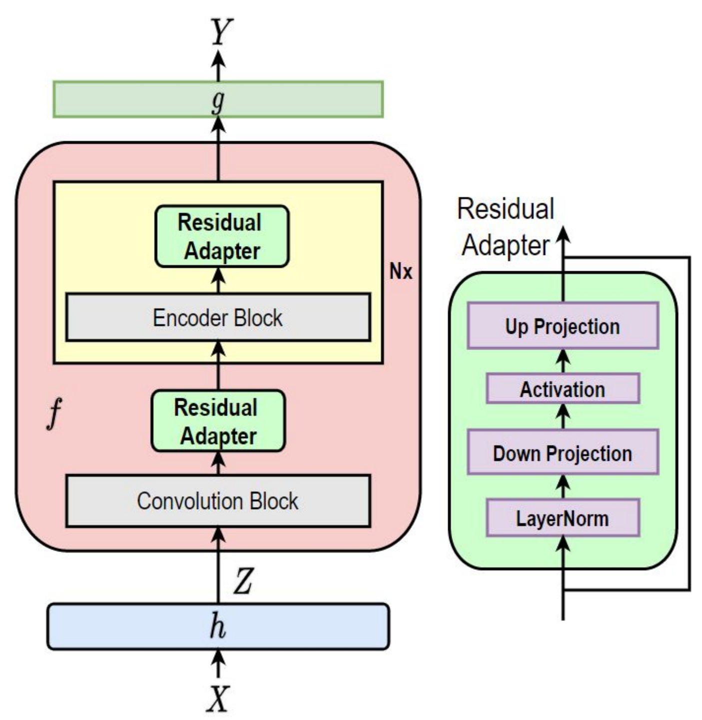
- The parameters of BLSTM are designed to be separated into a left-to-right and right-to-left context modelling LSTMs.
- The Bi-APC framework takes each LSTM as an individual APC and ignores the parameters that induce information exchange between the two LSTMs.
- The parameters in non-causal transformer, however, are not separate for contextual modelling from both directions.
- Four parameter-sharing solutions:
  - No modules are shared
  - Only the generator is shared
  - Only the convolution blocks is not shared
  - All modules are shared



(a) No Sharing (b) Share Generator g (c) Share g and Encoder (d) Share All

3. DRAFT: Adaptation of Self-supervised Models

- Residual adapters (RA):
  - Linear (down proj) + activation + Linear (up proj)
- SAFT (simple adaptation and finetuning): an additional adaptation stage to continually pretrain the SSL model using the finetuning data.
- DRAFT (domain responsible adaptation and finetuning): RAs are inserted after the convolution and each encoder block.
- DRAFT has 3 stages:
  - Pretraining stage: update the backbone model
  - **Adaptation stage: Freeze backbone and update RAs**
  - Finetuning stage: update all modules
- RAs learn domain related information during the adaptation stage and prevent the catastrophic forgetting Problem in finetuning.



Experimental Settings

1. Data
- Pretraining data: Librispeech 960-hour adult speech corpus
- Finetuning data:
  - Librispeech 10-hour subset
  - OGI 50-hour child corpus (read speech)
  - MyST 240-hour child corpus (spontaneous speech)

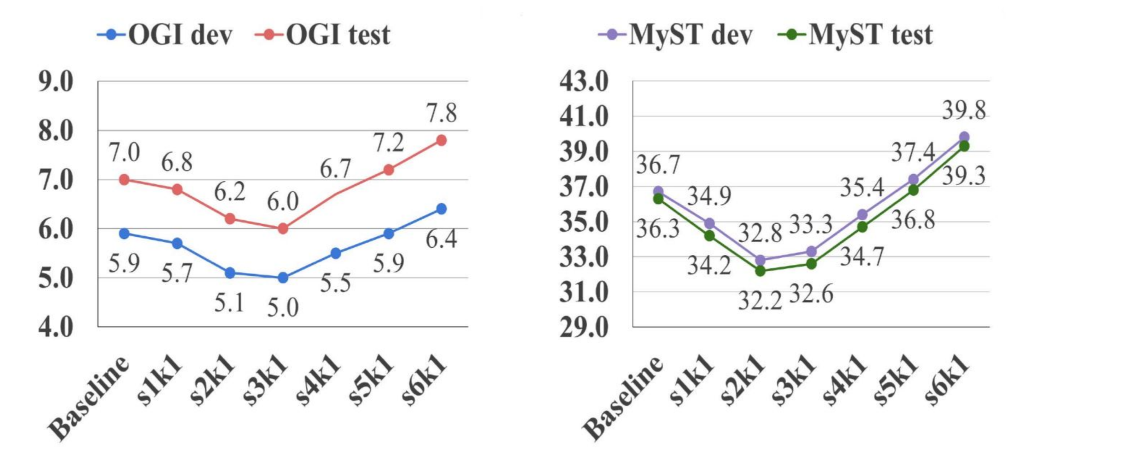
2. Pretraining Stage Setup
- E-APC and Bi-APC:
  - Z: 80-dimensional filter-bank features
  - f: two-layer cov block (4x subsample) + 12 transformer encoder blocks
  - Y: 320-dimensional output because of 4x subsampling
  - Adam optimizer for 130k steps
  - Model size: 39M
- Wav2vec2.0 and HuBERT
  - Pretrained models in the Fairseq toolkit
  - Model size: 95M

3. Adaptation Stage Setup
- E-APC and Bi-APC:
  - OGI: RAs are updated in 55k steps, a noam factor of 8
  - MyST: RAs are updated in 74k steps, a noam factor of 4
  - Batch size: 64
- Wav2vec2.0 and HuBERT:
  - RAs are updated 200k/100k steps with parking learning rate 5e-4 and the batch size is 16

4. Finetuning Stage Setup
- Loss function: CTC
- Data augmentation: speed perturbation + SpecAug
- Greedy search decoding is used during evaluation

Results

1. APC for casual transformer and its extension



APC	L_p	OGI		MyST		Libri-10h	
		dev	test	dev	test	clean	other
Baseline	-	5.9	7.0	36.7	36.3	56.0	74.9
APC-s2k1	L_1	5.1	6.2	32.8	32.2	47.6	67.2
EAPC-s1k4	L_1	5.1	6.0	35.5	34.8	45.7	65.1
	L_1	5.0	6.1	32.2	31.6	45.6	65.1
EAPC-s2k2	L_2	5.4	6.3	32.9	32.2	48.5	67.1
	$L_1 + L_2$	5.2	6.3	32.4	31.7	45.6	65.1

- APC achieves the best performance at s={2,3}, which are approximately the duration of an acoustic unit (a vowel or a short syllable).
- S1k4: four temporally-shifted sequences with the lag prediction of {1,2,3,4}.
- Combining multiple temporally-shifted sequences can achieve better performance for finetuning.

2. Bi-APC with a non-causal transformer

	Share?			OGI		MyST		Libri-10h	
	Conv.	Enc.	G.	dev	test	dev	test	clean	other
Baseline	-	-	-	2.9	3.3	28.0	27.8	51.9	70.2
				3.2	3.9	27.8	27.3	60.7	77.0
Bi-APC			✓	3.3	4.1	32.1	31.8	58.6	75.7
		✓	✓	2.8	3.4	26.2	25.7	40.1	58.9
	✓	✓	✓	2.8	3.3	25.5	25.0	40.3	58.9

- With more modules shared, the WER performance tends to be better. But the improvements are not as large as other non-causal pretraining methods like Wav2vec and HuBERT.

3. DRAFT for causal and non-causal transformers

	E-APC				Bi-APC				Wav2vec2.0				HuBERT			
	OGI		MyST		OGI		MyST		OGI		MyST		OGI		MyST	
	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test
Baseline	5.9	7.0	36.7	36.3	2.9	3.3	28.0	27.8	-	-	-	-	-	-	-	-
+ Finetune	5.0	6.1	32.2	31.6	2.8	3.3	25.5	25.0	2.3	2.7	17.84	17.16	2.1	2.5	17.40	16.71
+ Adapter Finetune [67]	8.6	10.1	47.4	47.3	-	-	-	-	100	100	100	100	100	100	100	100
+ SAFT	5.0	5.9	33.4	32.9	-	-	-	-	2.2	2.7	17.85	17.28	2.0	2.4	17.52	16.89
+ DRAFT	4.4	4.9	30.1	29.4	2.7	3.2	24.8	24.3	2.1	2.5	17.21	16.70	1.9	2.1	16.79	16.53

- SAFT has a performance degradation compared to the finetuning baseline, caused by catastrophic forgetting.
- DRAFT improves the finetuning performance consistently for both the causal and non-causal pretrained models.
- Adapter finetuning fails to converge for the child ASR tasks.

4. What do Residual Adapters Learn?

E-APC	RA Initialization	Update RA?	OGI	
			dev	test
Baseline	None	No	5.9	7.0
+ RA	θ^0_{ada}	Yes	5.5	6.4
	θ^0_{ada}	Yes	4.8	5.6
DRAFT	θ^1_{ada}	Yes	4.4	4.9
	θ^1_{ada}	No	4.7	5.4

"RA initialization" and "update ra?" are describing the finetuning stage. "+ RA" indicates adding randomly initialized RA at the finetuning stage for a fair comparison to DRAFT. θ^1_{ada} indicates the RA learned in the adaptation stage and θ^0_{ada} is the RA with random initialization.

- RAs learn target domain information WER 5.6% - > 4.9%
- Even without updating RAs in finetuning, the learned RAs can also improve the performance (5.4% v.s. 5.6%)

Conclusion

- The DRAFT framework performed well on E-APC, Bi-APC, Wav2vec2.0 and HuBERT methods, showing that it can improve the finetuning performance by reducing the domain mismatch between the pretraining and finetuning data.
- E-APC had a 1.8% relative WER improvement on the OGI and MyST data compared to APC.
- Bi-APC for a transformer can have an improvement over the baseline without pretraining, but the results are worse than bidirectional pretraining methods like Wav2vec2.0 and HuBERT.

References and Acknowledgement

- The reference numbers are the same as appeared in the paper.

- This work was supported in part by the NSF and UCLA-Amazon Science-Hub.