Self Training for Speech Recognition

P. R. Sullivan[1]

[1]School of Information
UBC

UBC DL-NLP, November 2020

Table of Contents

Overview

Self-training represents a clever way around data sparsity (a pressing issue in non-English speech tasks) by allowing models to pseudo-label raw data. FAIR demonstrates how this techniques can be used both in Spoken Language Translation (SLT) as well as Automatic Speech Recognition (ASR).

Quick Speech Review

- Cascade Models have ASR unit and Machine Translation (MT) unit trained separately (vs. End-to-End)
- To date SLT End-to-End lags Cascade performance
- ASR models are all End-to-End.

Diagram of End-to-End from [1]

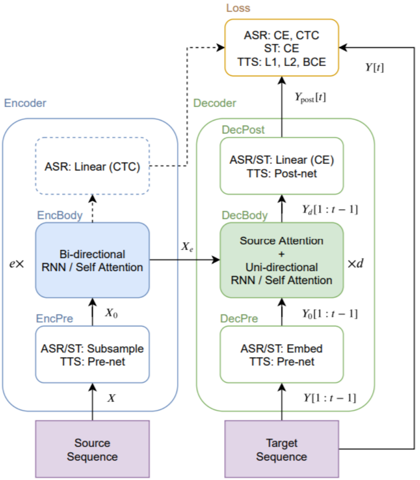


Table of Contents

Self-Training for End-to-End Speech Translation [2]

High Level:

- Compare Performance of self-training (ST) for SLT among a number of different models, language pairs, and data conditions.
- Pseudo-labels from large corpuses of unlabeled data (either LibrisLight or Facebook videos)
- Results: ST allows increase in model size dramatically and performance.

Data Used

- MuST-C for test sets
- Librispeech for additional data to simulate High resource setting
- LibrisLight unlabeled data to be pseudo-labeled for ST
- Facebook proprietary data for scale comparison (not particularly interesting)

Domain	Language	Dataset	# utterances	# hours
Open	En-Fr	MuST-C	275k	479
		dev	1412	2.6
		tst-COMMON	2632	4.2
	En-De	MuST-C	230k	395
		dev	1423	2.5
		tst-COMMON	2641	4.1
	En	LIBRISPEECH	281k	960
		LIBRILIGHT	15.8M	56k
FBVideos	En-Fr	train	20.7	30k
		dev	925	6.3
		test	3909	24.3
	En-Es	train	20.6M	30k
		dev	935	6.4
		test	3915	24.3
	En	unlabeled	32.2M	255k

Models Overview

- End-to-End models compare include a tiny LSTM model, and VGG-based Transformers (transformer with beefed up deep ConvNet layers see [3] and [4])
- ASR models (for cascade) are all Transformer based trained using wav2letter++, except for facebook video specific data which uses a large model built on Time-depth separable convolution blocks (TDS).
- MT models (for cascade) are all Transformer based.

Task	Model	# Parameters
ST	LSTM	13.5M
	VGGT	260.0M
	VGGTLARGE	435.0M
ASR	Transformer 1024	339.9M
	Transformer 768	204.7M
	TDS	292.0M
MT	En-Es FB Video	320.1M
	En-Fr FB Video	300.6M
	En-De [29]	209.9M
	En-Fr [29]	221.9M

Methods

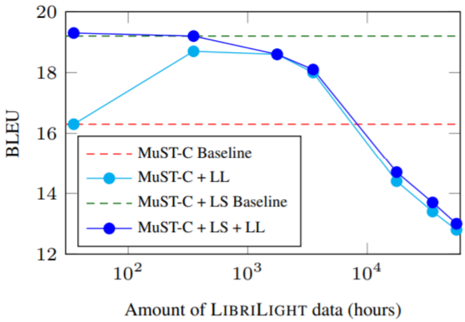
Using pseudolabels generated from the cascade model on LibrisLight

- compare performance of training with addition pseudo-labeled data in low and high-resource setting (supplement MuST-C with Librispeech)
- compare model size performance with fine-tuning on target data

Finally, compare performance of end-to-end models with addition of pseudo-labeled data (either from cascade or from end-to-end self-training).

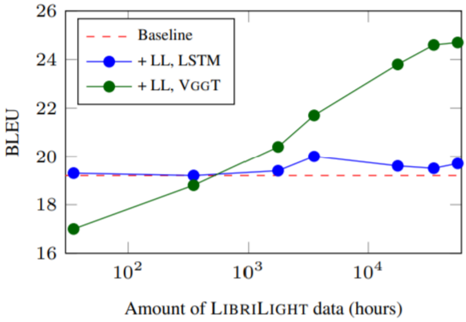
Results Low-High resource settings

How does a tiny cap. model do with additional pseudo-labels (+LL) in a low (light blue) or high (dark blue) setting?



Results Model Size

How does fine-tuning on the target data improve performance?
Low cap. LSTM (Bleu) vs. High cap. Transformer (Green)



Results Overall Cascade

Language	Data	Model	BLEU
En-Fr	MuST-C	LSTM	24.8
	MuST-C + LS		26.2
	MuST-C + LS	VGGT	23.9
	+ 35,217h LL + fine-tuning		**34.5**
	State-of-the-art baseline [16]		34.05
En-De	MuST-C	LSTM	15.6
	MuST-C + LS		19.5
	MuST-C + LS	VGGT	3.5
	+ 35,217h LL + fine-tuning		24.8
	+ 35,217h LL + fine-tuning	VGGTLARGE	**25.2**
	State-of-the-art baseline [16]		22.11
En-Fr (FB Videos)	baseline	VGGT	20.3
	+ 96k h unlabeled + fine-tuning		**21.6**
En-Es (FB Videos)	baseline	VGGT	18.5
	+ 96k h unlabeled + fine-tuning		**19.9**

Results End to End

How do End-to-End models do with labels either generated from other end-to-end models (LSTM Pseudo-Labels), bootstrapped from Cascade (VGGT Pseudo-Labels), or from Cascade directly.

Data	Pseudo-Labeling Model	Model	BLEU
MuST-C	N/A	LSTM	16.3
+ 3523h LL	Cascade	LSTM	20.8
	LSTM		18.5
	VGGT		20.6
MuST-C + LS	N/A	LSTM	19.2
+ 17,607h LL	Cascade	VGGT	23.8
	LSTM		20.7
	VGGT		**24.5**

Discussion

- Cool idea, SOTA results if you can get a ton of unlabeled data. Is this practical and reproducible for non-English languages? Maybe 1000hrs but 17k?
- ST + bootstrapping End-to-End models from cascade labels appears to close the gap between End-to-end and Cascade, at the cost of training both!
- Paper writing is somewhat of a mess, both in what it tries to do as well as with some gaps you wouldn't expect from an Interspeech paper.

Table of Contents

Self-training and Pre-training are Complementary for Speech Recognition [5]

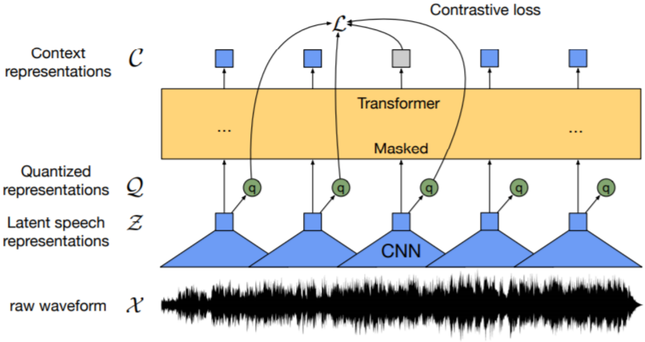
High Level:

- Compare unsupervised pretraining (wav2vec 2.0) with iterative pseudo-label self-training and the combination of both on ASR task.
- Librispeech and LibrisLight used for data
- Results: 3.1% WER on Librispeech-other (using 960hr unlabeled and 1h labeled), 5.2% WER with just 10 minutes (with 50h unlabeled)!

Wav2Vec(s) Recap

- wav2vec [6]: Contrastive loss used unsupervised training to create a representation of audio
- vq-wav2vec [7]: Improve on wav2vec through quantized outputs
- wav2vec 2.0 [8]: Incorporate tranformer context network. Basically learns contextualized representation alongside speech units.

Wav2Vec 2.0 Model



Methods

- Pre-train with wav2vec 2.0
- Fine tune on labeled data
- Create pseudo-labels using fine-tuned model (more details next slide).
- Train final model with additional pseudo-label data.

Methods cont.

A major piece of the pseudo-label generation is using large language models to improve the pseudo-label quality. In particular they use use beam size 800 with a 4-gram LM, which is then pruned to top 50 and then rescored with a large tranformer LM.

Results

Model	Unlbld data	dev clean	dev other	test clean	test other
10 min labeled					
Discr. BERT [27]	LS-960	15.7	24.1	16.3	25.2
wav2vec 2.0 [24]	LS-960	6.6	10.6	6.8	10.8
+ ST (s2s scratch)	LS-960	4.1	7.0	5.0	8.1
+ ST (ctc ft)	LS-960	3.6	6.6	4.0	7.2
wav2vec 2.0 [24]	LV-60k	5.0	8.4	5.2	8.6
+ ST (s2s scratch)	LV-60k	2.6	4.7	3.1	5.4
+ ST (ctc ft)	LV-60k	2.8	4.6	3.0	5.2
1h labeled					
Discr. BERT [27]	LS-960	8.5	16.4	9.0	17.6
wav2vec 2.0 [24]	LS-960	3.8	7.1	3.9	7.6
+ ST (s2s scratch)	LS-960	2.9	5.6	3.4	6.6
+ ST (ctc ft)	LS-960	2.8	5.5	3.1	6.3
10h labeled					
Discr. BERT [27]	LS-960	5.3	13.2	5.9	14.1
IPL [14]	LS-960	23.5	25.5	24.4	26.0
wav2vec 2.0 [24]	LS-960	2.9	5.7	3.2	6.1
+ ST (s2s scratch)	LS-960	2.5	5.1	3.5	5.9
+ ST (ctc ft)	LS-960	2.6	5.2	2.9	5.7

Table of Contents

Discussion

- Self-Training appears to be a solid boost to performance especially in low-resource settings.
- How realistic is using this, when very few public raw corpora exist for non-English at the size used?
- Could multilingual with phone targets be a compromise (especially for ST pre-training?)

References I

[1] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang *et al.*, "A comparative study on transformer vs rnn in speech applications," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 449–456.

[2] J. Pino, Q. Xu, X. Ma, M. J. Dousti, and Y. Tang, "Self-training for end-to-end speech translation," *arXiv preprint arXiv:2006.02490*, 2020.

[3] C.-F. Yeh, J. Mahadeokar, K. Kalgaonkar, Y. Wang, D. Le, M. Jain, K. Schubert, C. Fuegen, and M. L. Seltzer, "Transformer-transducer: End-to-end speech recognition with self-attention," *arXiv preprint arXiv:1910.12977*, 2019.

References II

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[5] Q. Xu, A. Baevski, T. Likhomanenko, P. Tomasello, A. Conneau, R. Collobert, G. Synnaeve, and M. Auli, "Self-training and pre-training are complementary for speech recognition," *arXiv preprint arXiv:2010.11430*, 2020.

[6] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

References III

[7] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," *arXiv preprint arXiv:1910.05453*, 2019.

[8] A. Baevski, Y. Zhou, A.-r. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, 2020.