Large-scale ASR Domain Adaptation using Self- and Semi-supervised Learning

SPE-22.1 PaperNum=1909

Dongseong Hwang, Ananya Misra, Zhouyuan Huo, Nikhil Siddhartha, Shefali Garg, David Qiu, Khe Chai Sim, Trevor Strohman, Françoise Beaufays, Yanzhang He

Google Research

Motivation

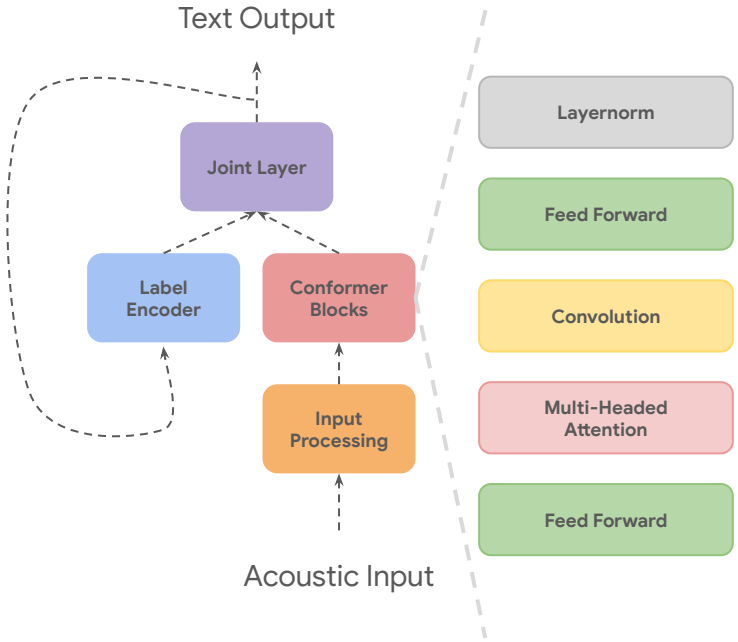
- How much effective are Self and Semi-supervised learning on Large scale production dataset (400k hours)?
  - Literatures mostly focused on LibriSpeech (labeled, 1k hours) and LibriLight (unlabeled, 60k hours) data.

- How much effective are Self and Semi-supervised learning for domain adaptation on large scale data?

Paper Summary

- Semi-supervised learning:
  - Main contributor to close domain mismatching gap.
  - Still effective on large scale data.
- Self-supervised learning:
  - Diminishing return as labeled data size is growing.
  - Complementary to Semi-supervised learning.
  - Improve WER last mile.

Models and Datasets

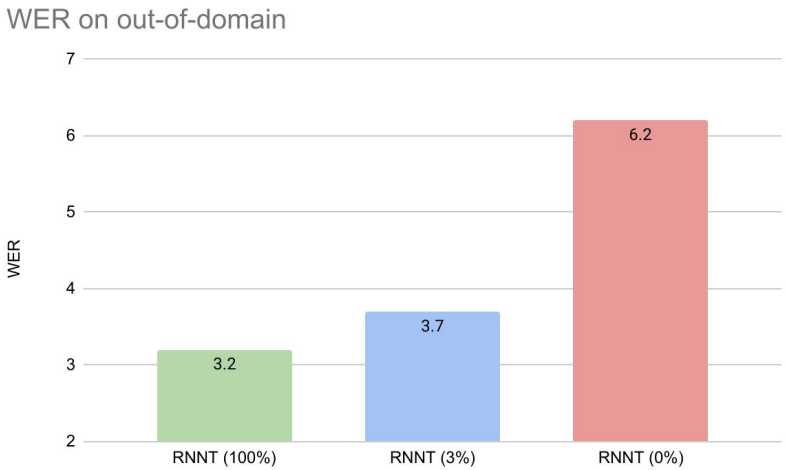


Streaming Conformer

- RNN-T architecture
- 137M Parameters
- Audio Encoder uses 17 Conformers
- Label Encoder uses 2 LSTM
- Trained on Google scale multi-domain (MD:400k hours) data.

Google Research

Problem: domain mismatch



There are 2 datasets:

- in-domain: short-form and YouTube (374k hrs)
- out of domain: long-form (26k hrs)

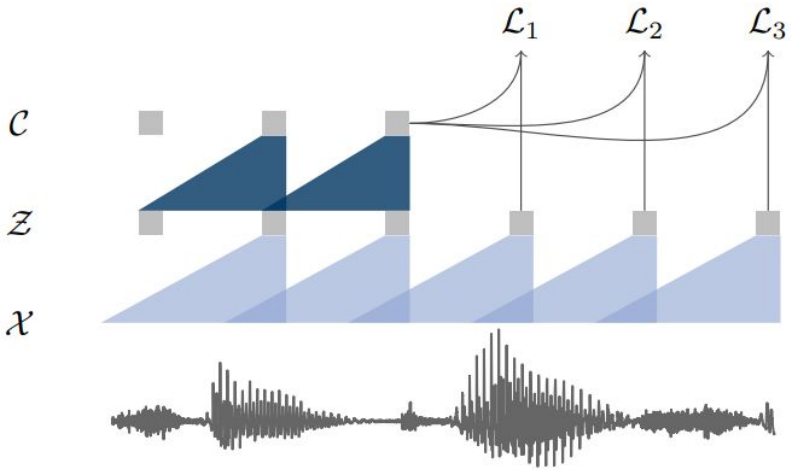
The RNNT model trained with in-domain has poor WER on OOD.

Goal: minimize domain mismatch gap.

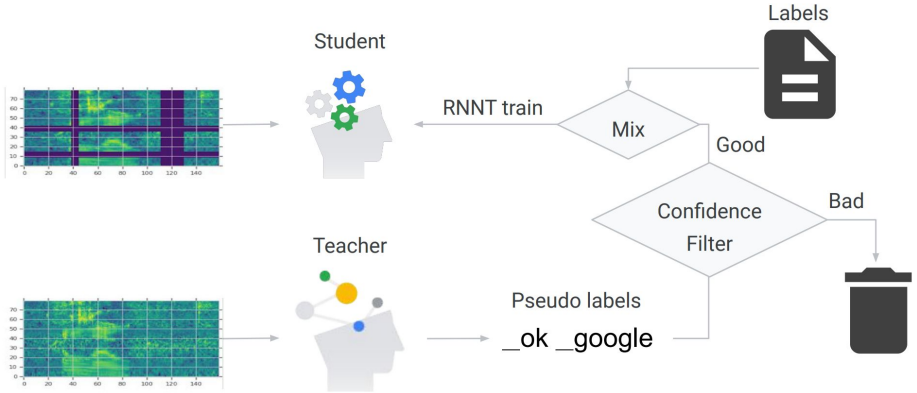
Google Research

Solution: self-sup and semi-sup

Self-sup (Wav2vec)
Wav2Vec, Wav2Vec 2, APC



Semi-sup
(Noisy student training)



wav2vec: Unsupervised Pre-Training for Speech Recognition

Self-training with Noisy Student improves ImageNet classification

Google Research

Solution: self-sup and semi-sup

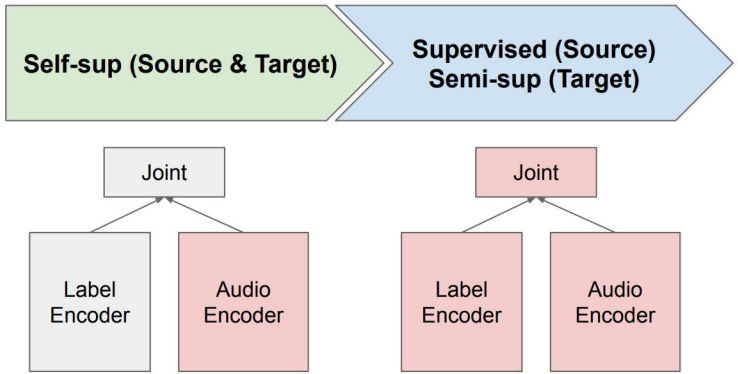


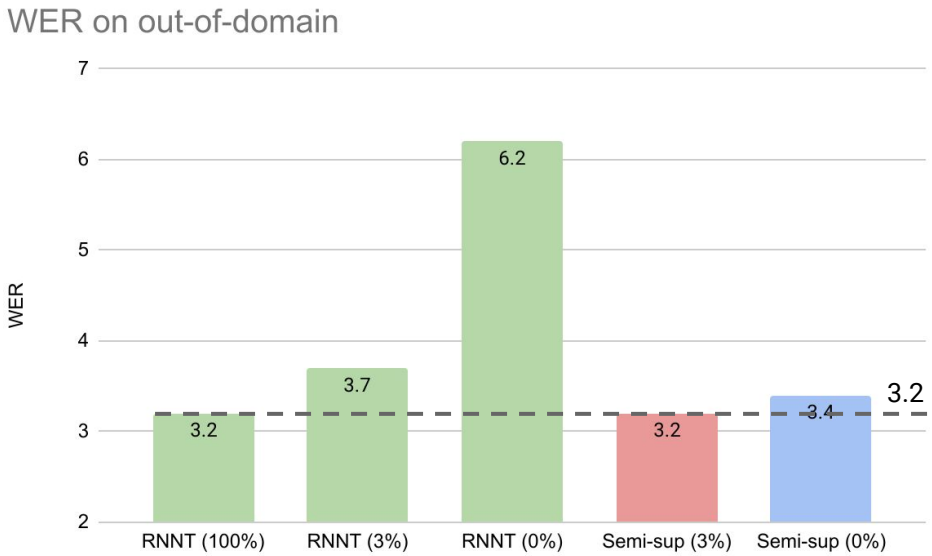
Fig. 1: *Domain adaptation: Self-sup pre-trains the audio encoder. Supervised and semi-sup train all the modules.*

2 stage training

- Self-sup: all the data (400k hrs)
- Semi-sup
  - in-domain: human label
  - OOD: pseudo label

Google Research

Ablation: Semi-supervised Learning



2 stage training

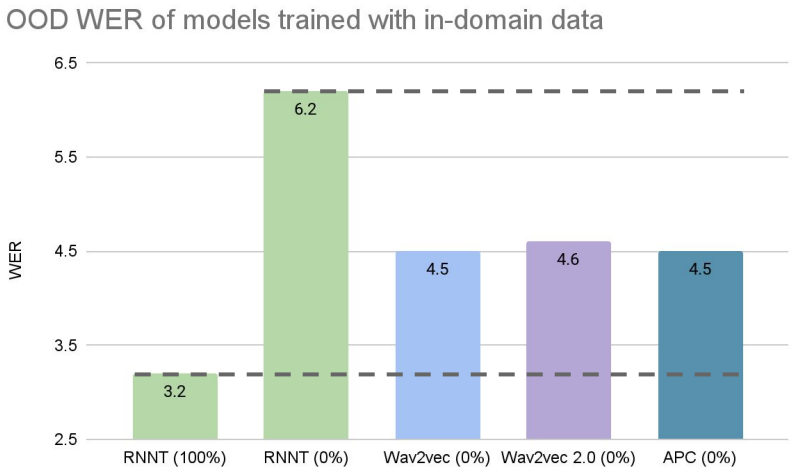
- Pretrain: RNNT with in-domain data
- Finetune: Semi-sup
  - Human label: x% OOD, in-domain
  - Pseudo label: rest of OOD

Semi-sup closes all the gap

- 3% of human label is enough
- 100% pseudo label works well

Google Research

Ablation: Self-supervised Learning



2 stage training

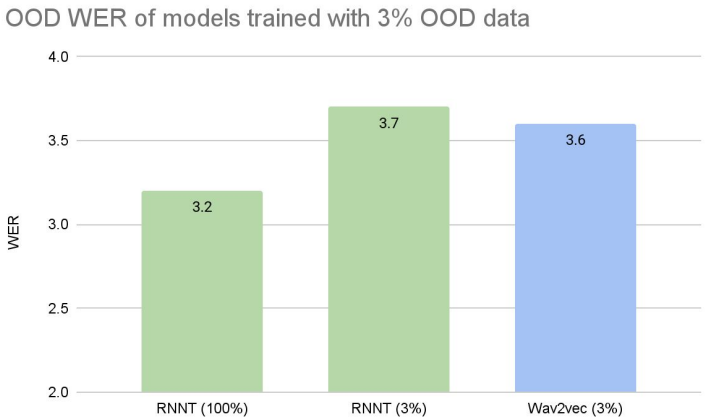
- Self-sup: all the data
- Finetune: RNNT with in-domain data

Wav2vec is best

- Downstream model is streaming ASR
- APC has a stability issue
- Not enough to close OOD gap

Google Research

Ablation: Self-supervised Learning



2 stage training

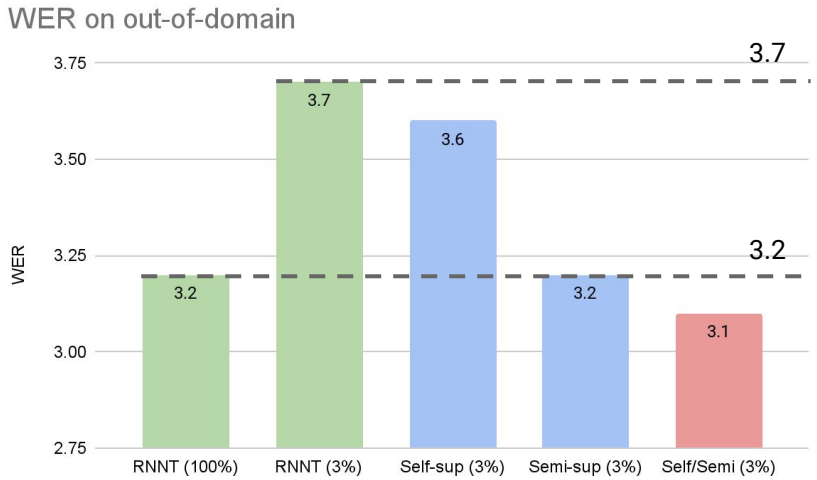
- Self-sup: all the data (400k hrs)
- Finetune: RNNT with 3% OOD (1k hrs) + in-domain data (374k hrs)

Self-sup contribution is minimal

- When label size is large (1k hrs), self-sup has little better WER than supervised learning.
- Self-sup is not good enough to solve domain mismatch problem.

Google Research

Results: Combine Self/Semi-sup



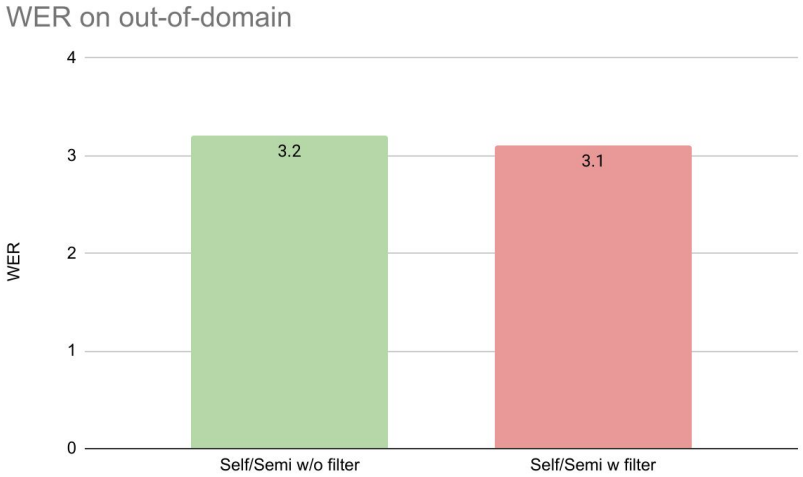
Self-sup and Semi-sup are complementary

- Semi-sup: close most of the OOD gap
- Self-sup: improve the last mile of performance

3% of human label with self/semi-sup has better WER than 100% labeled baseline.

Google Research

Results: Confidence filter



Confidence filter enhances WER

- All semi-supervised experiments use a confidence filter [14] that filters out target domain data when the utterance-level confidence score is less than 0.9.
- It drops 24% of utterances on target domain.
- It improves WER from 3.2 to 3.1.

Google Research

Conclusion

- With large resource data, semi-supervised learning is very efficient method to close out-of-domain mismatch.
- Self-supervised learning is complementary with semi-supervised learning.
- Combining both has the best WER, which is even better than baseline.

Google Research

Thank You!


Dongseong

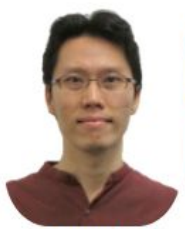

Ananya


Zhouyuan


Nikhil

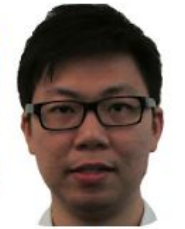

Shefali


David


Khe Chai


Trevor


Françoise


Yanzhang

Google Research