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

Dongseong Hwang (dongseong@google.com), Ananya Misra, Zhouyuan Huo, Nikhil Siddhartha, Shefali Garg, David Qiu, Khe Chai Sim, Trevor Strohman, Françoise Beaufays, Yanzhang He

**Background**

**source domain:** short-form and YouTube (374k hrs)

**target domain:** long-form (26k hrs)

**filter utterances.** The target sequence \( c \) contains a 1 when the prediction word is between the estimated confidence \( p \) and the binary target sequence \( c \).

**RNN-T model.** The CEM is trained to minimise the binary cross entropy when the teacher model is trained by RNN-T loss, we add CEM whose confidence utterances by Confidence Estimation Module (CEM) [14].

**generated for the target domain data is error-prone, which is harmful data (same for the student model). As a result, the pseudo labels target domain data. The teacher model is trained with source domain data (unlabeled data) [13]. NST produces pseudo labels for target domain data.

**Semi-supervised Learning**

We use Noisy student training (NST). NST is very effective to close OOD gap. We use both source domain and target domain to train models by RNN-T loss. Bi-directional teacher model produces pseudo label for target domain data. When we use 100% pseudo label for target domain, the model has 3.4 WER on target domain. When we mix 3% human label with 97% pseudo label, the model has 3.2 WER, which is same to 100% baseline. Semi-sup can close all the OOD gap by 3% of target domain data.

**Self-sup contribution is minimal**

We pretrain the audio encoder with both source and target domain data and finetune the RNN-T model with 3% target domain data and 100% source domain data. It improves target domain WER by 0.1%. Self-sup enhances overall model generalization, but cannot reduce gap of out-of-domain (OOD) generalization.

**Combined Self/Semi-sup**

Combined both self- and semi-sup are complementary. Self + Semi-sup show even better WER. We are actually surprised that that Self + Semi-sup with 3% target domain has better WERs than supervised learning with 100% target domain. Semi-sup plays a much more critical role to close the OOD gap, and self-sup enhances WER last mile.

**Model & Dataset**

We use the RNN-T architecture, which is a 137 million parameter end-to-end neural ASR model predicting target labels based on acoustic input. The audio encoder has 17 Conformer blocks with model dimension 512. As the model is online ASR, we restrict the model from using any future information. We use large multi-domain (MD) datasets in English. MD utterances include multi domain data such as search, farfield, telephony and YouTube. Total size is 450k hours.

**Introduction**

Self- and semi-supervised learning methods have been actively investigated to reduce labeled training data or enhance model performance. However, these approaches mostly focus on in-domain performance for public datasets. In this study, we utilize the combination of self- and semi-supervised learning methods to solve unseen domain adaptation problems in a large-scale production setting for online ASR model. This approach demonstrates that using the source domain data with a small fraction of the target domain data (3%) can recover the performance gap compared to a full data baseline: 13.5% relative WER improvement for target domain data.

**Problem**

There are 2 datasets:

- source domain: short-form and YouTube (374k hrs)
- target domain: long-form (26k hrs)

The ASR model trained with source domain data has poor WER on target domain data. This study minimizes domain mismatch gap by self and semi-supervised learning.

**Method**

**Self-supervised learning**

All of the self-supervised methods are used to pre-train the audio encoder of the RNN-T model [12] using all the source and target domain data. This is followed by supervised training of the entire model using only the labeled source domain data. We use the three popular self-supervised learning methods in this work: Wav2vec [1, 2], and Wav2vec2.0 [3], APC [4].

**Domain adaptation approach**

First of all, self-sup trains the audio encoder with source and target domain data. Then, RNN-T model is trained by RNN-T loss. Bi-directional teacher model produces pseudo label for target domain data. When we use 100% pseudo label for target domain, the model has 3.4 WER on target domain. When we mix 3% human label with 97% pseudo label, the model has 3.2 WER, which is same to 100% baseline. Semi-sup can close all the OOD gap by 3% of target domain data.

**Baseline**

The model trained with both source domain and target domain data has 3.2 WER on target domain. The model with only source domain has 6.2 WER. We mix full source domain data and 3% target domain data, the model has 3.7 WER. We want to minimize the gap between 3.2 and 3.7 with 3% of target data.

**Experiments**

**Compare W2V, W2V2 and APC**

First, we compare the three popular self-supervised learning methods: Wav2vec [1, 2], and Wav2vec2.0 [3], APC [4]. Wav2vec and APC have better WER than Wav2vec2.0, unlike what Wav2vec2.0 paper reported [13]. The downstream ASR model is online RNN-T, which is a causal model. Wav2vec and APC are causal models like GPT-3, but Wav2vec2.0 is full context (non-causal) model like BERT. It shows causal self-sup has better performance for causal downstream task. Even though Wav2vec and APC have the same WERs, we use Warv2vec for rest of experiments. In our experience, APC is more sensitive to checkpoint fluctuations. When we choose a pre-trained checkpoint, Wav2vec works between 50k and 1.2M steps, but APC works only near 100k steps. In addition, APC requires total variation auxiliary loss to stabilise it [20].