NUANCE | WHAT'S NEXT BLOG

Innovation & Research

Reducing the human labeling effort for training end-to-end speech recognition

Innovation at Nuance | Employee Guest Blogger

October 23, 2020

The latest generation of Nuance's deep learning technology for speech recognition features a novel algorithm integrating data augmentation with semi-supervised learning, which results in state-of-the-art recognition accuracy with much less human labeled data.

About the Author: Felix Weninger is a senior tech lead (Senior Principal Research Scientist) at Nuance Communications. His research interests include deep learning, speech recognition, speech emotion recognition, and source separation. He received his PhD degree in computer science from Technical University of Munich (TUM), Germany, in 2015. Prior to joining Nuance, he worked at Mitsubishi Electric Research Labs (MERL), Cambridge, MA, USA and in the Machine Intelligence and Signal Processing Group at TUM's Institute for Human-Machine Communication. He has published more than 100 peer-reviewed papers in books, journals, and conference proceedings.

Editor's Note: Franco Mana, Roberto Gemello, Jesús Andrés-Ferrer, and Puming Zhan contributed to the paper and this blog post. The paper was presented at the INTERSPEECH conference, October 2020.

Deep learning technology has rapidly transformed the way that computers perform speech recognition. It has enabled us to build speech recognizers for very challenging applications such as Dragon Ambient eXperience (DAX), which transcribes conversations between doctor and patient. In particular, the *end-to-end* (E2E) speech recognition system has been a primary focus of research in recent years. Traditionally, automatic speech recognition (ASR) systems consisted of separate components for modeling the acoustic

pattern of the smallest spoken unit (i.e. phonemes) of language (acoustic model), the mapping between phonemes and words (pronunciation model), and the dependency of words in a sentence (language model).

In contrast, E2E ASR systems subsume the acoustic, pronunciation, and language models into a single deep neural network (DNN). While such E2E models have been shown to be superior in terms of simplicity and accuracy, they require a large amount of labeled speech to learn the hundreds of millions of parameters necessary to achieve state-of-the art performance. However, manually transcribing large amounts of speech data is a tedious and costly process.

In the paper titled "Semi-supervised learning with data augmentation for end-to-end ASR", we explored semi-supervised learning (SSL) and data augmentation (DA) for leveraging unlabeled speech data, reducing the amount of labeled speech required, while maintaining recognition accuracy. Starting from a state-of-the-art E2E ASR system for transcribing doctor-patient conversations trained on 1900 hours of manually labeled speech data, we show how to combine SSL with DA to achieve similar accuracy with only ¼ of the labeled training data. Our paper has been accepted at INTERSPEECH 2020, the world's largest conference on spoken language understanding.

SSL is a family of machine learning techniques for training with a small amount of labeled and a large amount of unlabeled data. In ASR, the most common form of SSL is to use a *seed* ASR system trained only on the labeled data to generate the (pseudo) transcriptions of the unlabeled data. Then, the labeled data (with manual transcription) and the unlabeled data (with automatically generated transcription) are used jointly to train a new ASR system which should perform better than the seed ASR system. Such a process can be iterated with more unlabeled data being included at each iteration.

DA refers to techniques that create copies of the training data by performing small perturbations (e.g. to the frequency spectrum) of the speech signal, while leaving the labels unchanged. In this way, an arbitrary amount of labeled training data can be generated. In our paper, we employ the SpecAugment technique for doing DA, which is depicted below.

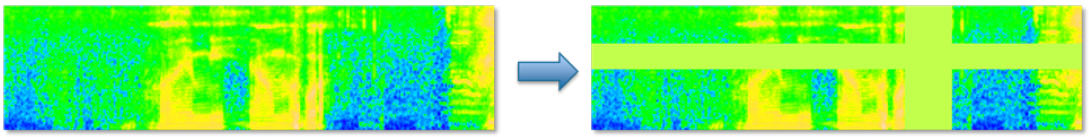

Figure 1: The SpecAugment approach modifies the spectrogram of a speech utterance by randomly masking some regions on the time (horizontal) and frequency (vertical) axes.
The most obvious approach to using SSL in combination with DA is to have the seed ASR system generate transcriptions for all the unlabeled data, and then apply DA. However, this simple approach has a drawback wherein erroneous transcriptions by the seed ASR system can be reinforced when using these pseudo transcriptions as training targets for the unlabeled data. In our paper, we proposed several techniques to help avoid this kind of error reinforcement.

The first is the *consistency training* principle: The seed ASR system is asked to transcribe the unlabeled data *after* it has been passed through DA, thus generating several copies of the data where both the speech and the transcriptions are slightly modified. In contrast to the simple SSL + DA approach, this avoids training on the same, potentially erroneous, transcription many times. The second technique is the usage of so-called *soft labels*, where the seed ASR system produces a probability distribution over all possible outputs, rather than being forced to make a hard decision. Finally, we found that E2E ASR systems are prone to repeating parts of sentences, which is why we introduced a heuristic technique to filter out transcriptions where the seed ASR system ran into a 'loop'. The proposed SSL + DA algorithm is sketched in the figure below.

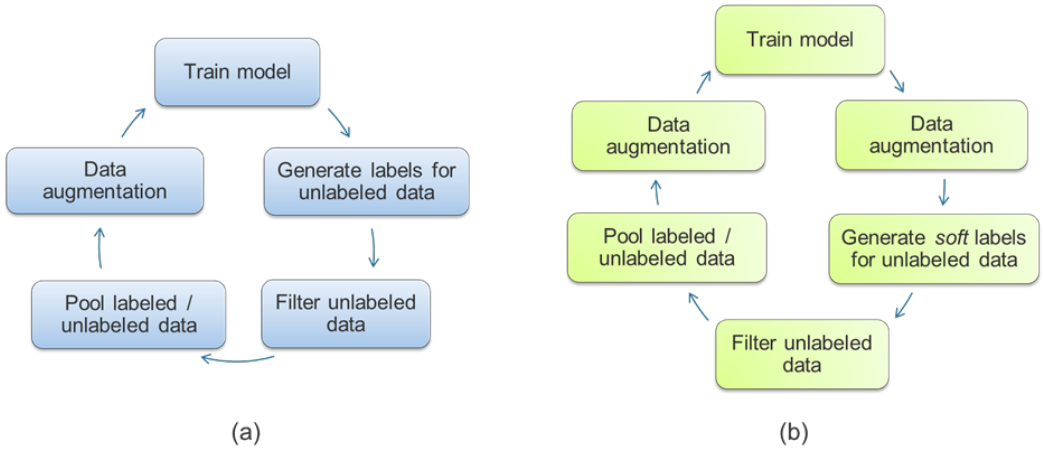

Figure 2: Traditional (a) and proposed (b) approach for combining SSL with DA. The proposed approach differs from the traditional one in the usage of consistency training (doing DA in the label generation process) and soft labels.
We applied this generic approach to two SSL algorithms known from the image classification literature, the *Noisy Student* and the *FixMatch* algorithm, and adapted them to the E2E ASR use case. In the Noisy Student algorithm, the seed ASR system is used as a teacher, while the model to be trained is treated as a student. By modifying the inputs to the student model via DA, the learning task becomes more difficult, requiring the student to generalize the teacher's knowledge to multiple variants of the data. The model trained this way can be expected to be more robust and perform better on unseen data.

In contrast, the FixMatch algorithm does not distinguish between teacher and student. It is a realization of the 'self-training' principle, where a model is trained with its own predictions for the unlabeled data. Since in the early stages of training, the predictions of the model are often wrong, it is necessary to have a way to measure the correctness of the predictions in absence of the ground truth. This can be achieved by computing the model's confidence in its output. As shown in the paper, when we only accept predictions with a high confidence, the convergence of the training process is accelerated. Moreover, since the model becomes more and more confident in its predictions over time, the training process iteratively includes more and more unlabeled data.

By applying the consistency training principle along with soft labels and heuristic loop filtering, we were able to outperform the simple approach of doing DA after SSL, achieving 4% relative improvement in word error rate (WER) on doctor-patient conversations. Furthermore, we found that both the Noisy Student and the FixMatch algorithms converged to similar WERs.

When training with 475h labeled data only, we achieved 16.8% WER. By adding 1425 hours of unlabeled data using the Noisy Student approach, we could reach 14.4% WER. This is in a similar ballpark as our best system trained on 1900 hours of labeled data (13.8% WER), while using only ¼ of the labeled training data.

In conclusion, our results put forward a promising avenue towards building state-of-the-art ASR systems with limited labeled data, which will be highly useful for specialized application domains and under-resourced languages in the future.

Tags: