Graduate School of Science and Technology Master's Thesis Abstract

Laboratory name (Supervisor)	Augmented Human Communication (Satoshi Nakamura (Professor))		
Student ID	2111428	Submission date	2024 / 1 / 19
Name	QI HELI		
Thesis title	A Temporal-Averaged Teacher-Student Model for Semi-Supervised Automatic Speech Recognition with Consistency Regularization		

Abstract

Automatic speech recognition (ASR) is increasingly pivotal in a wide range of real-world applications, from voice assistants to automated transcription services. However, the development of effective ASR systems is often hampered by the necessity for a large amount of labeled speech data -- a process both time-consuming and resource-intensive. Semi-supervised learning (SSL) emerges as a crucial methodology to simultaneously boost the model's accuracy on par with fully supervised performances and alleviate the high demand for labeled speech data, leveraging a small amount of labeled data coupled with a large amount of unlabeled data. The conventional and classic SSL methods for ASR involve initially training a base model on labeled data, followed by the generation of pseudo-text from unlabeled speech, which is then utilized as a form of indirect supervision for further SSL training. However, this approach struggles with the inherent limitation of labeled data scarcity under the SSL setting, which can limit the reliability of the generated pseudo-texts.

This thesis introduces an SSL algorithm for ASR, especially for the underexplored attention-based sequence-to-sequence (S2S) ASR methods. We focus on applying the consistency regularization idea to the CTC-attention joint ASR model, constraining the model to generate similar outputs on differently perturbated views. Here, we adopt both Speech Chain Reconstruction and SpecAugment as the augmentations for perturbating the inputs. The overall structure is inspired by the mean teacher algorithm's temporal ensembling, employs the exponential moving average of the student model to update the teacher model, thereby stabilizing the training and improving the quality of pseudo ground truth generated by the teacher model and then enhancing the model's SSL performance effectively. Extensive experiments on different data settings exhibit the superior performance of our method, ranging from data-limited to data-abundant settings.

Overall, this thesis bridges a gap in SSL for S2S ASR models, offering a groundbreaking approach that combines the strengths of speech-chain-based augmentations and temporal-averaged consistency regularization for improved ASR performance.