SpeechStew: Simply Mix All Available Speech Recognition Data to Train One Large Neural Network

William Chan, Daniel S. Park, Chris A. Lee, Yu Zhang, Quoc V. Le, Mohammad Norouzi

Google Research, Brain Team
{williamchan,danielspark,chrisalee,ngyuzh,qvl,mnorouzi}@google.com

Abstract

We present SpeechStew, a speech recognition model that is trained on a combination of various publicly available speech recognition datasets: AMI, Broadcast News, Common Voice, LibriSpeech, Switchboard/Fisher, Tedlium, and Wall Street Journal. SpeechStew simply mixes all of these datasets together, without any special re-weighting or re-balancing of the datasets. SpeechStew achieves SoTA or near SoTA results across a variety of tasks, without the use of an external language model. Our results include 9.0% WER on AMI-IHM, 4.7% WER on Switchboard, and 8.3% WER on CallHome, and 1.3% on WSJ, which significantly outperforms prior work with strong external language models. We also demonstrate that SpeechStew learns powerful transfer learning representations. We fine-tune SpeechStew on a noisy low resource speech dataset, CHiME-6. We achieve 38.9% WER without a language model, which compares to 38.6% WER to a strong HMM baseline with a language model.

Index Terms: end-to-end speech recognition, multi-domain speech recognition

1. Introduction

End-to-end speech recognition models [1, 2, 3] have seen remarkable success in recent years [4]. The success of these methods has often been attributed to the abundance of training data [5] and the use of large deep models [6]. However, on noisy, low resource speech recognition datasets, such as CHiME-6 [7], where overfitting is a significant problem, end-to-end methods tend to struggle relative to HMM-based baselines [8]. For example, the best previously published end-to-end model achieved 49.0% WER on the CHiME-6 dev set [8], while the best HMM model achieves 36.9% WER [9].

Multi-lingual training [10, 11], multi-domain training [12, 13], unsupervised pre-training [14, 15], semi-supervised learning [16, 17] and transfer learning [18, 19] are some techniques proposed in the literature to enhance generalization. These methods optimize speech recognition models on data from related tasks (typically of high resource), to help the specific task of interest (typically of low resource). For example, in multi-lingual training, the knowledge from a high resource language may transfer to a low resource language [20]. In multi-domain training, combining different domain datasets of the same language, could facilitate the cross-sharing of knowledge across the domains [13]. In unsupervised pre-training, the knowledge from the pre-training task may transfer to the supervised task [21]. In transfer learning, a general model is trained with a large amount of data. Subsequently, its knowledge is transferred via fine-tuning on training data from a downstream task that is typically of low resource [22].

This paper presents SpeechStew. SpeechStew is a simple approach to end-to-end speech recognition, which leverages both multi-domain training and transfer learning. SpeechStew follows the following simple recipe:

1. Combine all available speech recognition data without any domain-dependent re-balancing or re-weighting.
2. Train a single large neural network (a 1000M or 1B parameter model) on the combined data.

Our method does not utilize any domain labels, or introduce any additional hyperparameters for combining the data. We do not incorporate an external language model during inference, yet our result compares favourably to prior work that utilize strong language models, achieving SoTA or near SoTA results across various tasks (AMI, Common Voice, LibriSpeech, Switchboard, Tedlium, and WSJ).

We also demonstrate that SpeechStew has strong transfer learning capabilities. When presented with a new unseen low resource dataset (CHiME-6 in our setup), we merely:


We find that this straightforward pre-training and fine-tuning procedure yields near-SoTA results on CHiME-6. This is encouraging since CHiME-6 is a particularly challenging task [7] for end-to-end speech recognition models, which suffer from over-fitting issues [8]. We also demonstrate that our method is complementary to other pre-training methods, in particular unsupervised wav2vec pre-training [21] which we use in conjunction with SpeechStew training.

2. SpeechStew

In this section, we describe the model and training data setup of SpeechStew. We also describe our transfer learning setup for fine-tuning on new unseen tasks.

2.1. Model

In our implementation, SpeechStew uses the Conformer [31] RNN-T [32] architecture. We experiment with both the 100M parameter [31] and the 1B parameter configuration [6]. We find that wav2vec pre-training [15] is needed to train the 1B parameter model [6]. We apply the default hyperparameters from prior work [31, 6] including the learning rate schedule. We do not incorporate an external language model.

2.2. Multi-domain Training

We combine the following datasets without any form of re-weighting or resampling to construct the training set for SpeechStew:

1. AMI [33]. AMI is approximately 100 hours of meeting recordings.
2. Common Voice [34]. Common Voice is a crowd-sourced open licensed speech dataset. We use the version 5.1
2. Prior Work

2.1. Single Domain

Prior Work (no LM)


Prior Work (with LM)


Our Work (no LM)

Single Domain Baseline (100M) 26.1 40.5 16.3 (13.8†) 2.1 4.4 5.6 9.7 7.6 28.2
SpeechStew (100M) 9.0 21.7 12.1 (9.7†) 2.0 4.0 4.7 8.3 5.3 1.3
SpeechStew (1B) 9.5 22.7 10.8 (8.4†) 1.7 3.3 4.8 10.6 5.7 1.3

Our Work (with LM)

Official HMM Baseline [7] 51.8 51.3
HMM [9] 36.9 38.6
RNN-T [8] 49.0

Table 1: Speech recognition word error rates (%) across multiple tasks. SpeechStew achieves SoTA or near SoTA across many tasks.

3. Experiments

We build single task mode baselines, where the models are trained only on their respective domains. We use the Conformer 100M architecture [31] for these baselines; we found the 1B model to overfit and perform dramatically worse. We perform model selection via the development sets per baseline task.

Our SpeechStew model uses the 100M parameter and 1B parameter Conformer architecture [31, 6]. We used the default experimental settings of these references to train the models. We train all our SpeechStew models for exactly 100k steps, without any model selection.

Table 2 summarizes our results. We emphasize that the reported performance of SpeechStew is obtained without the use of an external language model. SpeechStew outperforms almost all prior work, including those using strong language models.

3.1. Transfer Learning and CHiME-6

CHiME-6 [7] is a low resource noisy speech dataset. We fine-tune our 100M and 1B SpeechStew models with the CHiME-6 data, and compare these results against baselines obtained by training a 100M parameter Conformer model and a 1B-parameter Conformer model (with LibriLight-only pre-training) with CHiME-6. Table 2 summarizes our results.

4. Conclusion

We presented SpeechStew. We achieve SoTA or near SoTA results across a variety of speech tasks. Our approach simply mixes all available training data to train one large neural network. We can apply transfer learning and finetune SpeechStew to new tasks.
5. References


