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

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, 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

Introduction:

- 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 Stree t 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, 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

Literature Review:

SpeechStew

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

Feature Extraction:

Explain how audio features (e.g., MFCCs) are extracted from the audio signal.

Describe the process of extracting visual features, such as lip movement trajectories or facial landmarks.

Discuss synchronization techniques to align audio and visual data.

Model Architecture:

In our implementation, SpeechStew uses the Conformer [28] RNN-T [29] architecture. We experiment with both the 100M parameter [28] and the 1B parameter configuration [4]. We find that wav2vec pre-training [14] is needed to train the 1B parameter model [4]. We apply the default hyperparameters from prior work [28, 4] 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 reweighting or resampling to construct the training set for SpeechStew:

- AMI [30]. AMI is approximately 100 hours of meeting recordings.
-

iii. Common Voice [31]. Common Voice is a crowd-sourced open licensed speech dataset. We use the version 5.1 (June 22 2020) snapshot with approximately 1500 hours. The data was collected at 48 KHz, and we resampled it to 16 KHz.

iii. English Broadcast News (LDC97S44, LDC97T22, LDC98S71, LDC98T28). English Broadcast News is approximately 50 hours of television news.

iv. LibriSpeech [32]. LibriSpeech is approximately 960 hours of speech from audiobooks.

v. Switchboard/Fisher (LDC2004T19, LDC2005T19, LDC2004S13, LDC2005S13, LDC97S62). Switchboard/Fisher is approximately 2000 hours of telephone conversations. The data was collected at 8 KHz, and we upsampled it to 16 KHz.

vi. TED-LIUM v3 [33, 34]. TED-LIUM is approximately 450 hours of TED talks.

vii. Wall Street Journal (LDC93S6B, LDC94S13B). WSJ is approximately 80 hours of clean speech.

Transfer Learning

We demonstrate the transfer learning capabilities of SpeechStew. Once we have a general purpose SpeechStew model (trained on the datasets mentioned in Section 2.2), we can finetune and adapt SpeechStew onto a new task. CHiME-6 [6] is a noisy low resource dataset set, which contains approximately 40 hours of distant microphone conversational speech recognition in everyday home environments. CHiME-6 is difficult for end-to-end speech recognition models to train directly due to over-fitting issues [7]. We fine-tune SpeechStew on CHiME-6 to demonstrate the transfer learning capabilities. The transfer learning capabilities of SpeechStew are extremely practical. It implies we can train a general purpose model once, then fine-tune to specific low resource tasks. This can be done at a very low cost, since fine-tuning typically requires only a few thousand steps, compared to ≈100k steps needed to train a model from scratch.

Methodology:

Transfer Learning and CHiME-6

CHiME-6 [6] is a low resource noisy speech dataset. It is especially challenging due to the small size and noisy audio conditions. We perform front end enhancement following the official recipe [6]. We use BeamformIt to enhance our front end during training. We use guided source separation [45] with 12 channels to enhance our front end for evaluation. 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 pretraining) with CHiME-6. Table 2 summarizes our results. We have not recorded results for the 1B-parameter baseline model, as we have failed to train the model to exhibit non-trivial performance. We have adjusted the dropout rate, the weight-decay parameter, augmentation parameters, learning rate and warmup steps with respect to the default values from [28] and [4] to optimize the dev-set WER

Results:

- The official CHiME-6 HMM baseline [6] achieves 51.8 and 51.3 WER on the dev and eval set respectively. A strong HMM model with a strong language model [8] achieves 36.9 and 38.6 WER on the dev and eval set respectively. The previous best end-to-end model [7] achieves 49.0 on the dev set.
- The SpeechStew 1B parameter model in the zero-shot setting (having never seen any CHiME-6 data before), achieves 39.2 and 53.7 WER on the dev and eval set respectively.
- Further fine-tuning our SpeechStew model on the CHiME-6 data results in 31.9 and 38.9 WER on the dev and eval set respectively.
- We note that our results, obtained without using any external language model, compares to prior work that have utilized strong language models. This demonstrates that the SpeechStew model already possesses a significant amount of knowledge that is useful for the CHiME-6 task.
- We make two observations regarding the effectiveness of SpeechStew on ChiME-6 in particular. The first is that the amount of training time spent on the low-resource CHiME-6 data is greatly reduced.
- This is a natural product of the finetuning process. However, SpeechStew's construction allows us to skip to that stage of the training process, thus reducing the risk of overfitting.

- Secondly, SpeechStew appears to be robust enough to handle noisy data. SpeechStew spends more time learning the salient features from a larger pool of less noisy data

Conclusion:

Deep learning models have made significant progress from two simple principles:

Train on more data [22, 25].

Train larger and deeper neural networks [23, 13]. Supervised data is expensive to acquire. SpeechStew takles this problem by simply mixing all publicly available speech recognition data. Our approach leverages on currently available resources, labelled and unlabelled. We hope our work will encourage future research to leverage on all training data available, as opposed to training on only task specific datasets.

References:

- [1] A. Graves and N. Jaitly, "Towards End-to-End Speech Recognition with Recurrent Neural Networks," in ICML, 2014.
- [2]W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, Attend and Spell: A Neural Network for Large Vocabulary Conversational Speech Recognition," in ICASSP, 2016.
- [3] D. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. Cubuk, and Q. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in INTERSPEECH, 2019.
- [4] Y. Zhang, J. Qin, D. S. Park, W. Han, C.-C. Chiu, R. Pang, Q. V. Le, and Y. Wu, "Pushing the Limits of Semi-Supervised Learning for Automatic Speech Recognition," in arXiv:2010.10504, 2020.
- [5] C.-C. Chiu, T. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, N. Jaitly, B. Li, J. Chorowski, and M. Bacchiani, "State-of-the-art Speech Recognition With Sequence-to-Sequence Models," in ICASSP,2018.