Automated Feature and Model Optimization for Task-Specific Acoustic Models

Akshay Chandrashekaran
Department of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
akshayc@cmu.edu

Ian R. Lane
Department of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
lane@cmu.edu

Abstract

The usage of hybrid systems of deep neural networks (DNN) for acoustic modelling and hidden Markov models (HMM) for decoding has enabled state-of-the-art performance for ASR. Though these systems are trained in a statistical approach to learn the parameters of the models, they are governed by a set of tunable parameters, called hyper-parameters, that determine the structure and generalization performance of the models. Here, we focus on keyword detection as the task of the system. The performance of the system is determined by metrics such as the area under the Receiver Operating Characteristic (ROC) curve and the keyword accuracy of the system. When developing speech-based applications such as keyword search or spoken dialog systems, there is no guarantee that the acoustic model and acoustic features used within the system are optimal for the end-to-end application task. We look at the optimization of the aforementioned hyper-parameters using statistical optimization techniques to find a feature extraction and acoustic model structure that produces the best result for the desired performance metric in an end-to-end speech application.

State-of-the-art ASR systems use a deep neural network(DNN) acoustic model (AM) as a statistical representation of the sounds that make up a word. A DNN is a feed-forward artificial neural network with multiple hidden layers between the input and the output layer. They are combined with a Hidden Markov Model (HMM) decoding graph to give a DNN-HMM hybrid ASR system. In the ASR framework, these DNNs are given spliced consecutive frames of features extracted from a raw audio signal as input. They return the likelihood of the HMM states modelled by a soft-max output layer.

A typical feature used in DNN-HMM hybrid models is the log-mel feature. In sound processing, a log-mel feature is a representation of the short time power spectrum of a sound, based on it's log of the power spectrum mapped to a nonlinear Mel scale of frequency. Though these features are not as de-correlated as Mel Frequency Cepstral Coefficient (MFCC) features, they still find widespread use as the multi-layer structure of a DNN allows it to extract rich internal representations robust to variability in the source.

Typically, the selection of model-space and feature-space hyper-parameters is done manually by a speech recognition expert. However, this is affected by the expert's bias, and human fatigue, resulting in an incomplete and inefficient search over the hyper-parameter space. To mitigate this, there has been increasing interest in the usage of data-driven approaches for the optimization of performance of systems by tuning the algorithmic hyper-parameters. Some research has looked at the optimization of algorithmic hyper-parameters like the learning rate, dropout, momentum using Bayesian optimization to optimize the WER, but a fixed model structure was used. To the best of our knowledge, no paper has explored the joint automatic optimization of the structure of the models and the feature space.

In the feature space, we look at window size, window shift and number of mel-bins as the hyper-parameters. In the model space, we look at the number of frames at input, the number of hidden layers and the number of neurons per layer as the hyper-parameters.

We perform a comparison of expert manual optimization, random sampling and Bayesian optimization using Gaussian process priors as the hyper-parameter optimization techniques for this task.

We perform all experiments on a small custom data-set of speech where users were asked to speak a single keyword per utterance. The data was split into 3 second chunks. 25% of the utterances contained the keyword while the rest were background noise. The data is split into independent random chunks of training, development and test sets in the ratio of 2:1:2. All hyper-parameter optimization techniques are performed on the development set. The results are reported on the test set.

We compared the end-to-end system performance when the optimization metric is either keyword accuracy, or area under the ROC curve (AUC) and show that the Bayesian optimization method performs better than randomized search, both of which are better than a baseline expert manually optimized system. Using the proposed optimization we are able to improve keyword accuracy by 8.5% relative (from 82% to 89%) or reduce AUC by 22% (from 0.161 to 0.125) compared to the manually tuned baseline system.

The experimental results show that both methods show large improvement over the baseline expert manual optimization method, with Bayesian optimization showing slightly better performance than random sampling for this task. We see that optimizing the structure of the DNN in the acoustic model as well as exploring the feature space leads to large gains in performance. Instead of tuning all the hyper-parameters by hand, human intervention is needed only for specifying the ranges of the hyper-parameters and for making the decision of termination of the optimization process. Finally, we also see that optimizing directly towards the desired metric, though more expensive, will give better generalization performance than optimizing over an assumed equivalent metric.