Research Passion

Everyone should have access to speech technology in their native language.

This is the conviction that drives my research. I want the world to be a place where blind children don't have to settle for a lesser education because of lack of audiobooks. Unfortunately, this is the case for most of the world's 6,000 languages. Working with people who faced this issue, I helped build a free Kyrgyz speech synthesizer. In this way, I've played a small role to make speech technologies accessible, but there is a hard limit as to what I can accomplish on my own.

The Google AI residency is the next step towards my goals. I want to publish influential work and get more people caring about low-resource languages. With my knowledge of linguistics and passion for machine learning, the Google Brain team is the perfect place for me to learn, collaborate, and share my research with others. In particular, I have many ideas for multilingual speech technologies, extracting information from large datasets to transfer knowledge to smaller domains. Working with Google researchers, I am confident I can tackle these problems.

The competitive (yet collaborative) research atmosphere at Google excites me, and I can offer my enthusiasm and dedication to research.

Challenges of Open-endedness

The hardest part of research is not finding answers, but finding questions. I push myself to find questions of real import - questions whose answers can *scale*.

My thesis topic is Multi-Task Learning for Acoustic Modeling in Automatic Speech Recognition (ASR), and at first I only answered questions relevant to ASR. Now I'm looking for questions that scale to all of machine learning.

Acoustic models in ASR are classifiers which accept a speech signal and return probabilities over phonetic transcriptions. I use a Multi-Task feedforward neural network to perform this classification. Learning related tasks in parallel can improve performance on the main task, because the weights in the net will be biased towards general, task-independent representations of the data (Caruana 1997).

The Multi-Task approach offers an elegant way to exploit small datasets, if you can come up with the right tasks. In my earlier experiments, I created auxiliary tasks using linguistic theory, lowering Word Error Rates for small data sets.

Standard labels for acoustic modeling are speech sounds such as `[b]` or `[f]`. However, these sounds are actually a bundle of linguistic features (eg. vocal cord vibration, tongue placement, and lip rounding). Past research created multiple tasks by predicting each linguistic feature in isolation (Stadermann 2005). This is like taking a 3-dimensional label `[x,y,z]`, and predicting `[x]` and `[y]` and `[z]` as additional tasks. The problem with this approach for human speech is that the predictive power of any one dimension is very weak.

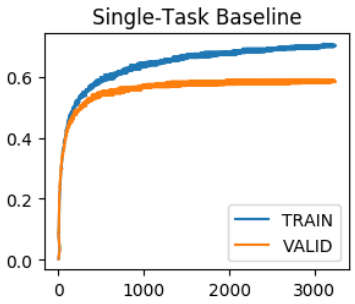
Reformulating the problem, I created tasks by *removing* one dimension for each task, projecting into `N-1`-dimensional spaces. That is, instead of `[x,y,z]` $\rightarrow$ `[x]` and `[y]` and `[z]`, I projected `[x,y,z]` $\rightarrow$ `[x,y]` and `[y,z]` and `[x,z]`. These `N-1`-dimensional spaces have more predictive power, and also force the net to learn the importance of each dimension.

Questions of Import

While the above tasks showed improvement over the baseline, defining each task by hand is not a scalable approach. Since then, I've devoted my research to finding scalable, automatic solutions to task creation.

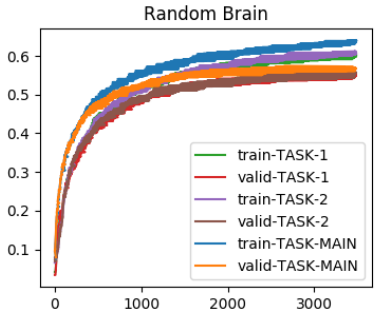
A Multi-Task classifier is a special case of an ensemble model, where the feature extractors are trained in parallel. Looking for inspiration from *scalable* ensemble approaches, I came across the Random Forest.

A Random Forest is created by training multiple trees on random re-samples of the data. Each subsample has its own decision plane, specific to its particular data and noise. By voting among all trees, noise is ignored but data generalities remain. Extending this approach, my current research investigates Multi-Task neural nets, where the tasks are random subsets of the data.



The figure to the left is my current baseline model performance (on Train and Validation) with a 5-layer Time-Delay Neural Network, with 200 dimensional hidden layers. Training data is a 5-hour section of the LibriSpeech corpus. The x-axis is number of iterations, and the y-axis is audio-frame classification accuracy. This model has overfit the training data. Over time, training data accuracy increases, but accuracy on the validation data plateaus and eventually decreases.

To the right is the performance of my Multi-Task model (I dub the 'Random Brain'). This model is currently training on an Amazon instance with 30 auxiliary tasks (I am only showing 2 auxiliary tasks here). The tasks are random 50% subsets of the dataset. The main task is the entire dataset. My impressions from this early data are the following: the Random Brain achieves lower accuracy, but less plateauing compared to the Single-Task model (given the same amount of training iterations). This could mean the extra tasks are a hindrance (making the task harder to learn), or the tasks are harshly penalizing overfitting, and given enough iterations the Random Brain will outperform the baseline on the held-out test data. If this approach works, then the Random Brain could be used to train *any neural net* on *any dataset*. This excites me, because it's potentially a solution which could impact the field of machine learning as a whole, not just speech recognition.



Scientific work is most valuable when it impacts both the research community and people's everyday lives. I want to contribute knowledge to the wider machine learning community, and play my part in creating accessible language technologies. Otherwise, speakers of the world's 6,000 languages will be waiting a long time for the resources we English speakers currently take for granted.

Thank you for your time and consideration.