Thesis: Data Selection For Speech Processing

Abstract: The rise of big data has led to scalable solutions for learning from large amounts of data to build state of the art machine learning models. However, in many real-world scenarios, with speech being our focus, we still contend with learning in a data-efficient manner. Beyond the amount of data, model performance depends on factors such as the quality of the annotations and whether the data are representative of conditions in which models will be deployed. In this work, we explore data selection methods that can identify highly informative subsets of speech examples to maximize the performance of a model while minimizing the amount of data required.

First, we present data selection methods in the context of unsupervised active learning for automatic speech recognition (ASR). We focus on low-resource languages, where the scarcity of data and annotation expertise makes it difficult to develop ASR models. We employ novel submodular functions for selecting highly informative and diverse examples that produce high performing ASR models with limited annotations.

Second, we present data selection methods in the context of speech enhancement. We focus on concatenative resynthesis systems, which produce noiseless and high-quality speech from noisy speech. The core component of these systems is a deep neural network (DNN), which learns a non-linear similarity metric between clean and noisy speech. We use linguistic and acoustic characteristics of the data to sample paired examples of clean and noisy speech that allow the DNN to learn a more generalizable similarity metric.

Finally, we revisit speech recognition with the goal of supervised active learning. We propose to investigate data valuation for ASR. Data valuation is concerned with determining the utility of examples to a supervised learning algorithm. We focus on Shapley valuation that arises from cooperative game theory. We propose to use Shapley values to determine the utility of speech examples for an end-to-end ASR system comprised of encoder and decoder recurrent neural networks. In preliminary work, we show that this is effective approach for appraising acoustic frames and identifying misleading data for an ASR acoustic model.

Committee:

- Professor Michael Mandel, Mentor, Brooklyn College
- Professor Rivka Levitan, Brooklyn College
- Professor Alla Rozovskaya, Queens College

Outside Member:

- Brian Kingsbury, Ibm, Thomas J. Watson Research Center