Recap: Dataset Engineering

In this module, you will learn about the nuances of data engineering, and develop your own dataset which you will use to train and deploy a custom keyword spotting model!

Within these courses so far, we have used numerous datasets to train our algorithms. The existence of a dataset spared us from having to tackle the most time-consuming components of the machine learning pipeline: the collection and wrangling of data. In this section of the course, we are going to create our own dataset to perform keyword spotting, and expose you to some of the challenges that come along with the process, such as dataset bias, the need for edge cases, and the difficulty associated with procuring large quantities of data. This procedure is aptly named **dataset engineering**.

Data engineering is an important aspect of supervised learning, and entails the specification of the dataset, as well as the procedures for collecting and processing data. Naturally, this involves defining what features will be present, the structure of the data (e.g., images, time series, row tuples), and other characteristics.

One of the key features of data engineering is identifying potential sources of data, which may come from external sources (free or purchased), crowdsourcing, existing users of a service, sensors, or other means. When developing our dataset, it is imperative that we consider our use case at all times, since this is our ultimate goal.

A natural corollary of this is that we must determine the environment our device will be in once deployed. For example, if an algorithm processing sound data expects to be used on the street, it is necessary to make recordings in the presence of adversarial sounds such as construction work, loud vehicles, as well as other background noise. We may have to decide what time of day images are taken, and also make sure that each of the classes in our dataset have roughly equal representation to prevent biasing to the majority class.

Another important concern when performing data engineering is data provenance. We must be cognizant of the origin of our data, most notably when we are using information procured from an external source, such as images from the internet. These images may be subject to copyright laws and other restrictions related to privacy. To preclude the possibility of legal issues, the origin of data should also be assessed, consulting with domain experts where necessary.

Certain types of data may require labeling, which may be done using crowdsourcing platforms such as Amazon Mechanical Turk in some cases. However, in other cases, domain experts may be required to label data. For example, radiologists may be necessary to assess x-rays or mammograms for abnormalities that would be vague or unnoticeable to the layman.

A further concern is the shifting of our objectives. If a plan is not sufficiently detailed or implemented correctly, dataset creep may occur, wherein more features are added in an attempt to improve the utility of the dataset. Alternatively, certain feature variables may be omitted or inputted incorrectly, such as with incorrect labeling.

Clearly, there are many ways in which the engineering of a dataset can go awry, and it is our job as a data engineer to ensure that the dataset has been properly collected with consideration of the deployment environment, as well as considerations of labeling, dataset bias, copyright, and other aforementioned concerns.

In the remainder of this section, you will learn to design, collect and develop your own dataset for the creation of a custom keyword spotting model. Good luck!