Introduction to Data Engineering

What is Data Engineering?

Data engineering is a critical component of supervised learning, and consists of defining requirements, recording data, processing it, and improving the dataset. The quality and quantity of collected data determines the tractability of a machine learning objective. Training examples need to include enough salient features, along with representative noise induced by the surrounding environment (for example, day or night in images, or quiet and loud background noise for audio), for an ML algorithm to accurately distinguish between classes when deployed into the real world.

Data engineers need a rigorous problem definition in order to know what data should be collected, and must identify the potential sources of data. Data might come from on-device sensors, product users, or paid or unpaid contributors, and each may introduce potential licensing or privacy restrictions. This data must be labeled, and this usually requires manual effort by individual workers. It may also require domain expertise, for example, when labeling medical images. Mislabeled or garbled data may also need to be filtered out through manual inspection. Data engineers must also manage changing needs for a dataset, for example, in order to support additional languages.

What's in this Module?

All machine learning tasks begin with datasets. Datasets are often handed to us, and many of us never think about where they came from. But that's a big mistake. Not being careful about dataset engineering can have serious consequences down the road. We are going to discover some of those pitfalls so that we don't fall subject to those same potholes. We will use existing datasets as examples to ground the principles you will learn, specifically around identifying dataset requirements, data collection, data refinement, and sustainability.

Specifically we will touch upon these items:

- The need for datasets as benchmark standards
- How to collect datasets for TinyML use cases
- How to be wary of bias and incorrect labeling in the datasets
- How to broaden the dataset collection process to accommodate diverse populations
- How can we repurpose existing (large) datasets for TinyML use cases
- How to be responsible when collecting datasets