Module 4 Lesson 1 part 2- seungju lee

## 8. Training Data

Training Data is the currency of machine learning and is the basis of learning component in ML

Training Data is like teaching a young child that has never seen a car what a car is

Graphical user interface

Description automatically generated

Data in ML

* Data defines model behavior and performance
* Model parameters are updated based on training data
* A model will not learn if it’s not in the data
* Bad data = bad model

## 9. A pet model

Imagine a model that determines the type of pet in the image

Graphical user interface, text, application

Description automatically generated

If the model is not given a particular type of data, it will never be able to infer on other types of that data.

Only after being trained on a given type or class of data, will the model begin to make predictions of that type

## 10. Training Data is Key

When training a model, you must train it with all types of data that the model is likely to encounter in the future in order to build a robust model.

* Models will only learn about data that they are trained with
* Ensure that the data used to train a model reflects real-world data
* Use a diverse set of data to build a robust model
  + If model only trained with images of dogs in the snow, very unlikely to identify a dog pictured in a field of grass

### Example of training with the wrong data

: Want to train a model to detect shoes

A screenshot of a video game

Description automatically generated with low confidence

* Training data doesn’t represent all the types of shoes that we see in real-life



* Now model will be able to identify the type of shoes

### The right data… Diversify your data to understand all the potential scenarios

* Data that encompasses all likely scenarios
* Photos from the real world (not just a studio)
* Audio collected with background noise
* Text of various writing styles
* Equal amounts of the different types of data

### Common Issues with training data

1. Distribution of training data

Chart, bar chart

Description automatically generated

Model will learn significantly more about cats and dogs than other classes and will tend to bias towards those classes

To prevent this, we can

* Collect more training data for the classes that are lacking
* Reduce the amount of training data for those that have too much

More training data will often lead to a better model, balancing the data is as important if not more to produce an unbiased model.

1. Training Examples are different than the real-world examples

Graphical user interface, chart

Description automatically generated

Audio collected from mobile device has a great deal of noise as the cases real life situations.

If we were to train the model only on studio audio, our model would fail miserably at detecting audio collected from mobile devices.

Example where we want to be aware of what the real-world data will look like, so we can collect the appropriate data

1. Issue of mislabeled data

Timeline

Description automatically generated

Parsing entities out of text, may be unclear where any particular entities should be classified.

Need to ensure that the labels that are returned are accurately classified as expected.

1. Insufficient data

Chart, bar chart

Description automatically generated

The amount of data required will vary widely based on a number of factors, including the datatype, complexity of the data.

No clear-cut rule as to how much data will be needed!

Generally, we want to start with a few 100 examples of each target class and then scale up the amount of training data until we reach a desired accuracy.

## 12. Training Data Summary: Training data is key to building robust AI

* Training data will make or break a machine learning model.
* Ensure there is a variety of data and that the data reflects real world scenarios
* Unbalanced data will cause a model to skew towards a particular outcome
* Mislabeled, or dirty, data will significantly impact model performance