DATA ANNOTATION & ITS ROLE IN MACHINE LEARNING

Data annotation plays an essential role in the world of machine learning. It is a core ingredient to the success of any AI model because the only way for an image detection AI to detect a face in a photo is if many photos already labelled as “face” exist.

If there is no annotated data, there is no machine learning model.

What is data annotation?

The core function of annotating data is to label data. Labeling data is among the first steps in any data pipeline. Plus, the act of labeling data often results in cleaner data and additional areas of opportunity.

Labeling data

It is necessary to have two things when annotating data:

1. Data
2. A consistent naming convention

As labeling projects grow more mature, the labeling conventions likely increase in complexity.

Sometimes, too, after training a model on the data, you might discover that the naming convention was not sufficient to create the kind of predictions or ML model you intended. Now you need to get back to the drawing board and redesign the tags for the dataset.

Clean data

Clean data builds more reliable ML models. To measure if the data is clean:

1. Test the data for outliers.
2. Test data for missing values or null values.
3. Ensure labels are consistent with conventions.

Annotation can help make a dataset cleaner. It can fill in gaps where there are some. When exploring the dataset, it might be possible to find bad data and data outliers. Data annotation can both:

- Salvage poorly tagged data or data with missing labels
- Create new data for the ML model to use

Automated vs human annotation

Data annotation can be costly, depending on the method.

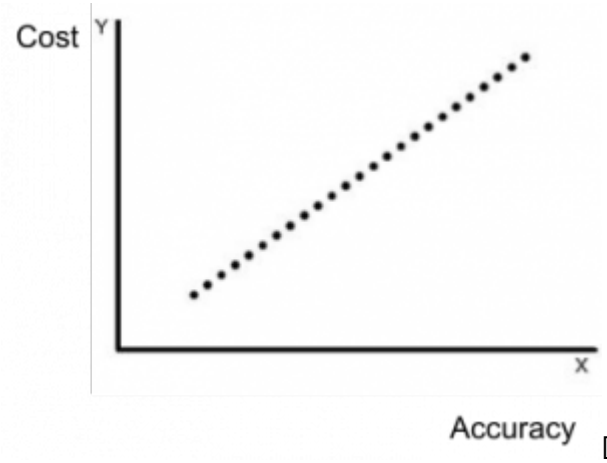
Some data can be automatically annotated, or, at least annotated through automated means with a degree of accuracy. For example, here are simple forms of annotation:

- Googling an image of a horse and downloading the top 1,000 photos into a horse file.

- Scraping a media site for all its sports content and labeling all the articles as sports articles.

You've automatically collected horse and sports data, but the degree of accuracy of that data is unknown until investigated. It's possible some horse photos downloaded are not actual photos of horses, after all.

Automation saves costs but risks accuracy. In contrast, human annotation can be much more costly, but it's more accurate.



Data annotators can annotate data to the specificity of their collected knowledge. If it is a horse photo, the human can confirm it. If the person is an expert in horse breeds, the data can be further annotated to the specific breed of the horse. It's even possible for the person to draw a polygon around the horse in the picture to annotate exactly which pixels are the horse.

For the sports articles, the article could be broken down to which sport, a game report, player analysis, or game predictions. If the data is tagged only as sports, the annotation has less specificity.

In the end, data is annotated to both:

- A degree of specificity
- A degree of accuracy

Which is more necessary, however, always depends on how the machine learning problem is defined.

Human-in-the-loop learning

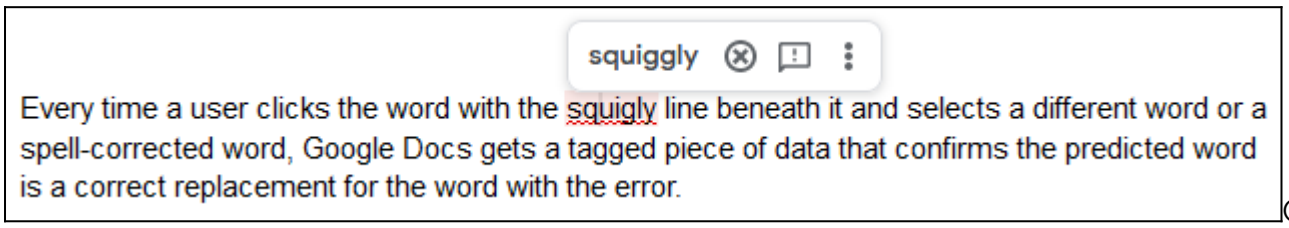
In IT, the "distributed" mentality is the idea of pushing workloads to a single instance to eliminate huge amounts of work piling in a single location. This is true of the Kubernetes architecture, computer processing infrastructure, edge AI concepts, microservices architecture—and it is true of data annotation.

Data annotation can be cheaper, even free, if the annotation can occur in the user's process.

It is a boring and unfulfilling job to offer someone to sit and tag data all day. But, if the labelling can occur naturally within the user experience or once in a while from many people instead of one, then the job becomes a lot more approachable and the possibility of getting annotations is even achievable.

This is known as human-in-the-loop (HITL), and it's often one function of mature machine learning models.

For example, Google has included HITL and data annotation in its Google Docs software. Every time a user clicks the word with the squiggly line beneath it and selects a different word or a spell-corrected word, Google Docs gets a tagged piece of data that confirms the predicted word is a correct replacement for the word with the error.



Google Docs has included its user as a part of the process by making a simple feature in its app to get annotated data, real-world data form its users.

In this way, Google sort of crowd-sources its data annotation problem and doesn't have to hire teams of people to sit at a desk all day reading misspelled words.

Tools for data annotation

Annotation tools are tools designed to help annotate pieces of data. The data they accept are:

- Text
- Image
- Audio

The tools generally have a UI that allows you to make annotations simply and export the data in various forms. The exported data can be returned as a .CSV file, text document, file of marked photos, or they can even format the annotated data into a JSON format specific to the convention for training that data in a Machine Learning model.

These are two well-known annotation tools:

- Prodigy
- Label Studio

But that's not nearly all of them. Awesome-data-annotation is a GitHub repository with an excellent list of data annotation tools to use.

Data annotation is an industry

Data annotation is essential to AI and machine learning, and both have added immense value to the world.

To continue growing the AI industry, data annotators are needed, so the job is sticking around. Data annotation is already an industry and will only continue to grow as more and more nuanced datasets are required to build out some of machine learning's more nuanced problems.

Related reading

- BMC Machine Learning & Big Data Blog
- Data Ethics for Companies

- 3 Simple Data Strategies for Companies
- Data Analytics vs Data Analysis: What's The Difference?
- Top Machine Learning Algorithms & How To Get Started
- Anomaly Detection with Machine Learning