Importance of Data Collection in Artificial Intelligence

Why do we need datasets?

Artificial Intelligence(AI) depends heavily on data. Without data, an AI system can't learn. It is the most crucial aspect that makes algorithm training possible. No matter how great your AI team is or the size of your data set, if your data set is not good enough, your entire AI project will fail.

Machine learning algorithms require large amounts of data to work. When managing millions or even billions of data samples, it's truly difficult to pinpoint what precisely causes a framework to perform seriously. Along these lines, when aggregating your information, it's insufficient to accumulate an immense stream of data, feed it to your model, and anticipate great outcomes. The cycle should be substantially more finely tuned.

Data assortment is a significant bottleneck in AI and a hot subject in different networks. There are generally two reasons for the criticality of data collection. To start with, as AI is turning out to be all the more generally utilized, we are seeing new applications that do not really have enough tagged data. Second, unlike conventional AI, deep learning techniques automatically generate features, which saves feature engineering costs, but in return may require larger amounts of labeled data. Interestingly, recent research in data collection comes not only from the machine learning, natural language, and computer vision communities but also from the data management community due to the importance of handling large amounts of data. Data collection largely

consists of data acquisition, data labeling, and improvement of existing data or models.

Data collection is the single most important step in solving any machine learning problem. Teams that dive headfirst into projects without considering the right data collection process often don’t get the results they want. Fortunately, there are many data collection tools to help prepare training datasets quickly and at scale.

The best data collection tools are easy to use, support a range of functionalities and file types, and preserve the overall integrity of data. They are:

- Synthetic Data Generators

Synthetic data can also be programmatically generated to obtain large sample sizes of data. This data can then be used to train neural networks. There are a variety of tools for generating synthetic datasets. Various Python libraries can be used to generate a large synthetic database as specified by the user. Data generator tools let users create custom CSV, SQL, JSON, and Excel datasets to test and demo software.

- Data Augmentation Tools

Data Augmentation can be used to expand the size of an existing dataset without gathering more data. For instance, an image dataset can be augmented by rotating, cropping, or altering the lighting conditions in the original files. OpenCV and Scikit python libraries include image augmentation functions and features for bounding boxes, scaling, cropping, rotation, filters, blur, translation, etc.

- Open-Source Datasets

Another way to obtain raw data for machine learning is to obtain pre-built, publicly available datasets on the internet. There are thousands of publicly available datasets spanning a wide range of industries and use cases.

- Data Collection Tools & Services

The majority of algorithms require data to be formatted in a very specific way. As such, datasets usually require some amount of preparation before they can yield useful insights. After you've collected enough raw data, you'll still need to preprocess it before it's useful for training a model. There are a hundred of Data Collection service providers around the world.

Quality, Scope, and Quantity

Machine Learning is not only about large data sets. Indeed, you don't feed the system with every known data point in any related field. We want to feed the system with carefully curated data, hoping it can learn, and perhaps extend, at the margins, the knowledge that people already have.

I have a data set, what now?

Not so fast! You should know that all data sets are inaccurate. At this moment of the project, we need to do some data preparation, a very important step in the machine learning process. Basically, data preparation is about making your data set more suitable for machine learning. It is a set of procedures that consume most of the time spent on machine learning projects. The collected data needs to be processed as per the algorithm and application type before it is used for the training.

Data Preprocessing

At this step, you have gathered your data that you judge essential, diverse, and representative for your AI project. Preprocessing includes the selection of the right data from the complete data set and building a training set. The process of putting together the data in this optimal format is known as feature transformation.

1. Format: The data might be spread in different files. For example, sales results from different countries with a different currency, languages, etc. which needs to be gathered together to form a data set.

2. Data Cleaning: In this step, our goal is to deal with missing values and remove unwanted characters from the data.

3. Feature Extraction: In this step, we focus on the analysis and optimization of the number of features. Usually, a member of the team has to find out which features are important for prediction and select them for faster computations and low memory consumption.