The Importance and Key Points of Data

Baobab, Inc.

What is Machine Learning?

Machine learning is what occurs when a computer recognises a large number of patterns between questions and answers given to it by humans (called training data) and goes on to autonomously infer regularities, rules and laws so that it can make its own predictions and decisions.
Machine learning is a technology applied in the field of artificial intelligence, in areas such as machine translation, natural language processing, chat systems, speech synthesis, speech recognition, and image recognition.

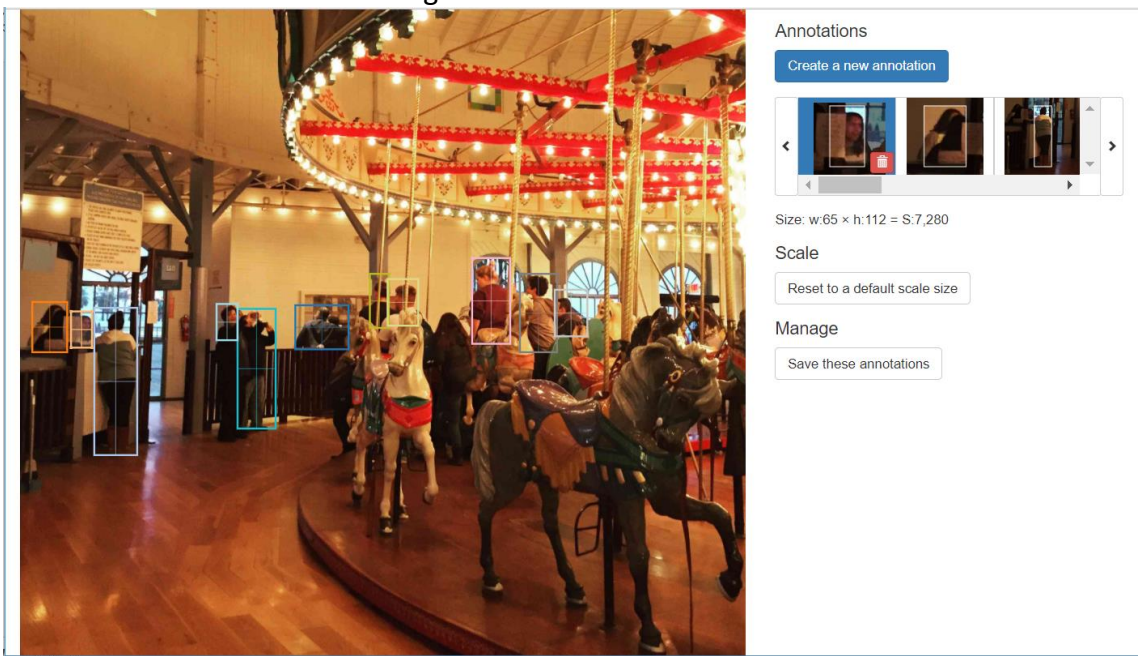
Machine learning is an essential mechanism for artificial intelligence.

What is Training Data?

It's a set of questions and answers that is given to a computer to learn.
Did you ever answer exercises in elementary school?
Every question (or equation) has an answer associated with it.
Imagine a similar kind of exercise.

For example, if you want to create an image recognition model able to recognize a person from an image, you will need a dataset containing a large number of high quality images of people to use as the training data.
People come in all shapes and sizes: they can be male or female, young or old, have different skin tones, different coloured hair, they can be sitting or running, standing in profile or facing away from the camera, in a wheelchair, using an umbrella or a walking stick, and so on.
So it is important to collect images containing a diverse range of people to reflect this, but we also need to make sure that the images are appropriate training data.
Below is an example of an annotation task, in which images are manually annotated to make them suitable for use as training data (in this case, for the purpose of building an

image recognition model able to recognize people). Can you see that all the people have been marked with a rectangle?



All images containing people marked with rectangles are then tagged with the label “people”.

In order to develop a good image recognition model, it is necessary to have a set of images that have been carefully and accurately annotated in this way as the training data.

What are the Key Points to Keep in Mind When Creating Training Data?

(1) Having a sufficient amount of data

In general, a large amount of data is required to infer rules with high accuracy. Using apples as an example, simply having images of each apple variety (Golden Delicious, Gala, Braeburn, etc.) is not enough, we also need to prepare images taken from different angles and in different light conditions.

(2) Having correctly labelled data

Machine learning looks for rules based on the assumption that everything labelled in the training data is actually correct. The computer does not consider that the data might be wrong. Therefore, the training data must be of high quality.
Data containing errors not only hinders the discovery of the correct rules, but also unnecessarily increases the computational cost of learning and the cost of creating the data.

For example, imagine you show a computer the following:
(1) Golden Delicious
(2) Gala
(3) Braeburn
(4) Cantaloupe,
and you tell it that they are all apples. Since it has now mistakenly learned that a cantaloupe is an apple, the next time the computer is shown an image of a cantaloupe or other variety of melon, it will recognize it as an apple.

(3) Having sufficiently comprehensive data

In order to create a machine translation engine capable of translating Japanese newspaper articles into English, the training data would need to contain sentence pairs such as the one shown below, taken from an actual newspaper article:

> *平昌五輪スピードスケート女子５００メートルで小平奈緒が金メダルに輝いた。*
>
> *Speed skater Nao Kodaira won the gold medal in the women's 500 meters at the Pyeongchang Games on Sunday.*

Newspaper articles contain many expressions that are not found in everyday speech. If such expressions are not adequately covered in the training data, then machine learning will not be able to uncover the appropriate translation rules.

In addition, returning to the apple example above, if the training data only contains images of red apples, and does not include any green apples, then when the computer is presented with an image of a green apple, it will judge it to be not an apple, because they are missing from the training data.

Do I Have to Create My Own Training Data?

Some datasets are publicly available. For instance, a well-known repository of image datasets is ImageNet (http://www.image-net.org/). However, it does have some drawbacks, the amount of data is limited, use is permitted for research purposes only, and the data tends to come mainly from the US and Europe. For example, ImageNet does not include images of ramen noodles. Therefore, if you show a picture of ramen noodles to an image recognition model trained on ImageNet data, it will recognize it as something else, such as "carbonara" or "mashed potatoes"*. If you wanted to use it at your company for commercial purposes, you would need to consider whether it is suitable for that purpose.
In most cases, training data needs to be created in-house by the company.
*From “2.3.5.1 Shared Data Sets” (p237) in a 2017 white paper on AI

Finally,

Data is the key to success in machine learning.
This success hinges upon the creation of data.
Thus, training data is an essential part of machine learning.