Pipeline of Machine Learning

Chris Cornwell

Aug 26, 2025

Steps in an ML Project

Project Pipeline

0. Define the problem.
1. Collect data.
2. Design the features in the data.
3. Training of the model.
4. Test the model.

Discussion on the ML Project Pipeline

0. Define the problem.
    - Example: From a given image, determine if it is an image of a dog or not.

Discussion on the ML Project Pipeline

0. Define the problem.
    - Example: From a given image, determine if it is an image of a dog or not.
1. Collect data.
    - Example: Put together a large collection of images, some having dogs in them, others having a different animal, or no animal. Have a label (your output, "y") for each image. Split into training set and test set.

Discussion on the ML Project Pipeline

0. Define the problem.
    - Example: From a given image, determine if it is an image of a dog or not.
1. Collect data.
    - Example: Put together a large collection of images, some having dogs in them, others having a different animal, or no animal. Have a label (your output, "y") for each image. Split into training set and test set.
2. Design the features in the data.
    - Not one thing that you always do here. Sometimes use experience/knowledge of what the data represents, sometimes use another learning algorithm to *learn* good features.

Discussion on the ML Project Pipeline

3. Training of the model.
    - Model is determined by set of parameters. In training, you alter the parameters iteratively – “tune” them – using optimization techniques (on the loss function).

Discussion on the ML Project Pipeline

3. Training of the model.
    - Model is determined by set of parameters. In training, you alter the parameters iteratively – “tune” them – using optimization techniques (on the loss function).
4. Test the model.
    - Evaluate the trained model’s performance on test data, measured by the same loss function.

Poll question

Difficulty in Defining the Problem

Examples worked with when first learning ML have a easily defined problem (*e.g.*, here are images of handwritten numbers; determine which number is written); defining problem harder in real practice.

Difficulty in Defining the Problem

Examples worked with when first learning ML have a easily defined problem (*e.g.*, here are images of handwritten numbers; determine which number is written); defining problem harder in real practice.

> Example. Have database of Tweets (from X / Twitter) about news events; interested in using machine learning to determine which are giving misinformation.
> Is it a simple classification problem, `'misinformation'` vs. `'not'`? If not, what alternative is there?

Difficulty in Defining the Problem

Examples worked with when first learning ML have a easily defined problem (*e.g.*, here are images of handwritten numbers; determine which number is written); defining problem harder in real practice.

> Example. Have database of Tweets (from X / Twitter) about news events; interested in using machine learning to determine which are giving misinformation.
> Is it a simple classification problem, `'misinformation'` vs. `'not'`? If not, what alternative is there?

Sometimes the issue is in the data.

> Example. Attempting to use crime data in Baltimore to model how crimes occur by location (*e.g.*, reoccurrence of crime at same location shortly after).

Difficulty in Designing Features

Very often, data is gathered in format or coordinates that make it hard to achieve ML task or, at least, do not help.

Will discuss some techniques that identify coordinates that “do not help” (Feature selection).

Difficulty in Designing Features

Very often, data is gathered in format or coordinates that make it hard to achieve ML task or, at least, do not help.

Will discuss some techniques that identify coordinates that “do not help” (Feature selection).

Domain knowledge is important for designing (computing) better features from given data.

Difficulty in Designing Features

Very often, data is gathered in format or coordinates that make it hard to achieve ML task or, at least, do not help.

Will discuss some techniques that identify coordinates that "do not help" (Feature selection).

Domain knowledge is important for designing (computing) better features from given data.

Example. In textbook, reconstructed Galileo experiment for objects falling.

$$\text{Force of gravity is constant } (g)$$
$$\Downarrow$$
$$\text{height change is } \frac{g}{2}t^2 \text{ (from Calculus)}$$

Poll question

Difficulty in Designing Features

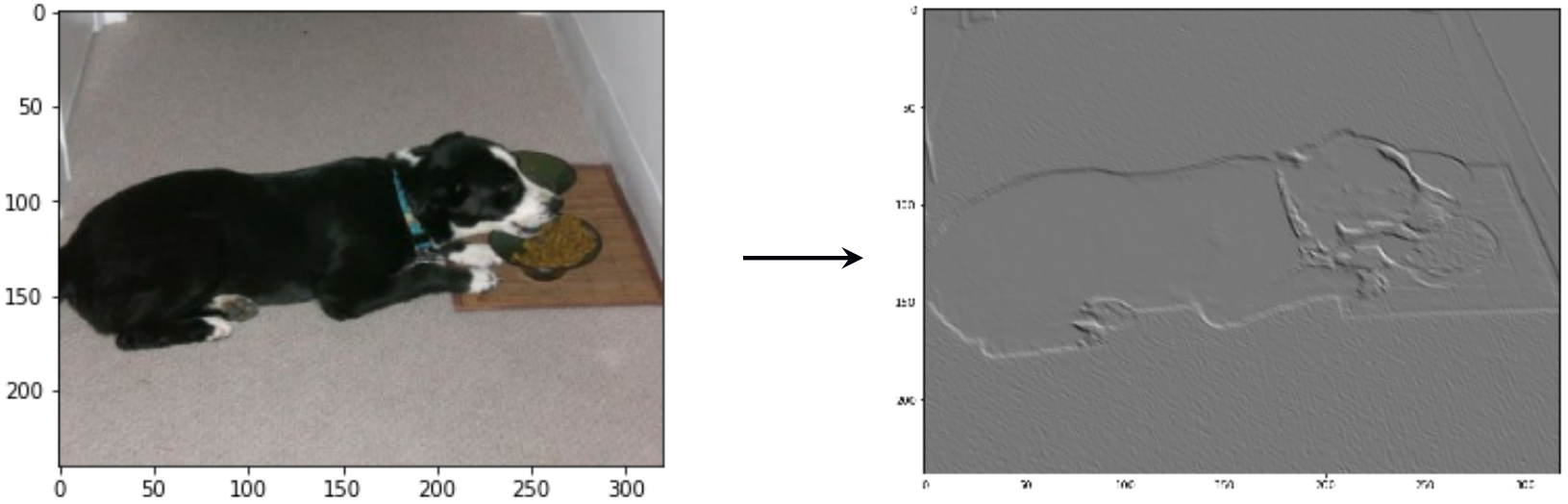
Some studies on brain function suggest that visually recognizing something is correlated with identifying edges in an image.

Computer vision learning algorithms are built to compute such edge features from input image.

Difficulty in Designing Features

Some studies on brain function suggest that visually recognizing something is correlated with identifying edges in an image.

Computer vision learning algorithms are built to compute such edge features from input image.



Questions?