PS643 Notes - Oct 18, 2024

MNIST (handwritten digits recognition) - a nontrivial problem long back, but trivial now

Even a 12-layer NN model was complicated in the early 2000s. Slowly the hardware required and datasets were improvised

IMAGENET: a very large image dataset (14M images with 21K labels)
- Flicker was another large image dataset before IMAGENET (not as big as imagenet)
- It was aimed at accomplishing large-scale image recognition
- Images were annotated with texts as well apart from just classifying them
- Google was interested in this dataset as well for multi-modal tasks

VQA Dataset: Visual Question Answer dataset involving multi-modal tasks
- You can give a photo and ask questions regarding it. Eg. What is the color of the background?

Deep Learning: (basically a big neural network !)
- It has layers apart from fully connected layers. For example convoluted layers
- A simple neural network with fully connected layers contains input signals and weights that propagate signals that combine to give an output signal. The final output is obtained by using a final layer like SOFTMAX to get a discrete output.
- A convoluted layer is used to extract certain useful and special layers
- Flatten layers are layers that convert complicated signals like 2D ones to a simple (1D) signal
- Convolution + Relu can be thought of as a differentiation. For example, one Conv + Relu layer does edge detection, and two do further refinement.

Convolution layer:
- Canny edge detector works on the same principle
- A sliding window sums up the values of neighboring pixels present in the window to give a new convoluted image.

Relu layer:
- Anything below 0 is 0, otherwise it's the number itself.

Pooling layer:
- It divides the image into a bunch of clumps based on filter size and stride, and the maximum of each clump is found.

- Object detection is a simple use case of this. We would like to detect every object when we see an image (like all the persons, cars, etc)
    - A small error can prove to be very very dangerous, eg. in the case of automated cars.
- From a public policy POV:

- A car is much less efficient when compared to public transport like train and bus
- People own a car only to show off
- An automated car is a nuisance in public policy POV because even a normal car itself is a hindrance in this regard

A complex Use case:

- Side note: Many datasets especially the ones in the computer vision field are being publicly extracted unethically for financial benefits
- One interesting project is to try using comic books involving pictures and text in a multi-modal modal. Can we predict the text in the last panel if we read the ones before it? Humans can do it with 85 percent accuracy (are good at comic reading!).

Solving the MNIST problem:

- Make use of Keras and Lasagne
- These are libraries that simplify the coding of large machine-learning models
- The code is available on Moodle
-