PS643: October 18th 2024

- MNIST (hand-written number images) is a major dataset for Computer Vision currently
  - A non-trivial task a few decades ago, but quite trivial now
- Some form of computer vision has existed in some form since decades. Technology took its time to catch up with the ideas.
- Hardware required to implement significantly large neural networks (as required by computer visionists at their time), also coincided with the evolution of dataset large and varied enough to solve the problems in hand.

  For example,
  - *Flickr* – "Instagram before Instagram became Instagram" – one of the early large dataset
  - *MSCOCO*

Significantly larger sized datasets --
1) "**IMAGENET**" dataset: 14M+ images, 21K+ labels (link)
[discussion regarding the publications on imagenet dataset, on the about page]

- Aimed at accomplishing large scale image recognition
- Images were annotated with texts as well apart from just classifying them
- Google was very interested in this dataset (for multi-modal tasks – for example)

2] Visual Question Answering [**VQA**] dataset (link)
- Ask questions based on provided image

What is Deep learning again?

- *(in short)* big neural network – can be called a convolutional network as well?
- another way to look at it is neural network with other than fully connected layers
- A normal (average) fully connected neural network:
  - Has input signals, and based on output finds error and updates weights to get the output closer to the expected. The final output is also a signal.
- Deep learning has multiple layers of fully-connected network, and also special layers which perform convolution to extract features from the input.
  - Each special layer comprises convolution-ReLu-pooling, which extract specific features
  - Every extra layer provides an extra degree of differentiation (like edge detection)

What is a convolution layer?

- We have a sliding window: which runs through the entire image, and at each step the sub-image is multiplied with the weights in the window, and summed up to get the output pixel.
- A 3d input is converted to a convoluted 3d output

What is a RELU (Rectified Linear Unit) layer?

- RELU(x) = max (0, x) : negative values are trimmed to a 0-value
- Works closely with the convolution layer

What is a Pooling layer?

- Divides the image into a number of clumps (or sub-image), and from each clump a representation element is chosen (e.g. maximum of all pixel values in *max-pool*)
- The final image at the end of last pooling layer is generally flattened into a 1d vector to be fed into the later fully connected layers.

A Simple use case !!!

- A headset which indicates to blind people of the objects in front of and around them, thus helping them to navigate the world with ease
    - *A simple error can be very detrimental !!*
    - *[Computer Vision is in a bubble which might explode anytime soon – no matter how much accuracy is reached, the world is ever changing and need not be the same and understandable by the vision model]*

Automated cars: From a public policy point of view –

- Buses are more efficient than cars
- Trains are more efficient than buses
- In total, trains are better than cars in all ways, and in an ideal world, there would only be buses and not cars, let alone automated cars
- Investing intelligence and money on automated cars does not make sense, but it's just market demand – society has more money than science
- None of the arguments to justify automated cars (or even cars) have proper basis

[[[To Think!!! – Especially in the field of computer vision, much of the training datasets (like flickr), are unethically publicly extracted with (or without) consent, but without the financial effects in the future]]]

A Complex Use Case !!!

- Identification and analysis or artificial images like paintings, arts, or comics (non-natural images)
- Building a multi-modal model to convert comic books to a different language
    - Extract the images using convolution
    - Extract the text in the bubbles using LSTM
    - Combine both the above features to predict the next page of the comic book
    - VERY LESS ACCURACY (~60%), whereas humans can do it at ~85% accuracy

Solving the MNIST Problem:

- *Keras, Lasagne* – libraries which simplify the coding part in making a large enough neural network
- *(A simple code will be provided on MOODLE)* – which implements a deep neural network in very few lines to solve the MNIST dataset problem quite effectively