**ADTA-5550: Deep Learning with Big Data**

**Mid term**

**Part I**

**Q1:**

In 21st century, we are currently using artificial intelligence (AI) in enormous applications. For example, photo editing, YouTube video recommendations, healthcare, and product discounts. These applications are cumulative research developments in the field of AI. The history of AI inventions way backs to the year of 1950. In the same year, a British polymath, Alan Turing explored mathematical possibility of AI and first introduced the Turing test in his paper “Computer Machinery and Intelligence.” This paper explains the characteristics of AI logically, ways to build intelligent machines, and how to test them.

Five years down the line, the proof of concept named as “Logic Theorist” which is highly considered as first AI program was presented at the “Dartmouth Summer Research Project on Artificial Intelligence” by Allen Newell, Cliff Shaw, Herbert Simon. This is the event where the first Artificial Intelligence (AI) word is termed. However, the conference did not go well but it had become a crucial catalyst in the further endeavors in the research of AI.

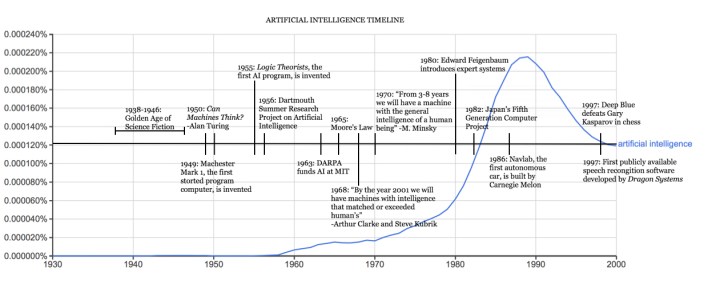


Figure 1: (Image source: <https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/>)

Later, in the next two decades, as the computers being invented, the research improvements in the field of AI embellished. Gradually, people started developing machine learning algorithms and tried to build solutions for their problems. Few early demonstrations have given assurance about the results and real feel about the machine solving the problems. This led to convince government to fund research in the areas of AI. Even the government keen to know about machines transcribing and translating spoken language. As the years go, the improvements in the research work have grown. During those days, the major fallback was to lack of computational power. This led to slow down of the developments for the next 10 years.

Later in 1980s, AI research work was fueled by expansions of AI toolkit and inventions of Deep Learning which allowed computers to learn by experience. Ironically, in the absence of government funding and public hype, AI thrived. During the 1990s and 2000s, many of the landmark goals of artificial intelligence had been achieved. AI also helped to build computer player in the games. Later it was also in news that Google AlphaGo have defeated the Chinese Go champion. In 2016, Sophia, the first robot citizen developed by Hong Kong based company introduced in United Nations Development Programme. In the recent years, we also seen number of chatbots which has become the live assistant feature in every business, also we did see Amazon Alexa, Google Voice and most trending tool known as ChatGPT. ChatGPT which is assist in answering almost all questions that a human being can ask. Other examples include Microsoft Bing and Google Bard.

*Machine Learning:*

Machine Learning (ML) is the subsection of AI which deals with learning and decision-making capabilities. In the late 1970s, ML has branched out from the research work on AI and grown separately and varied research implementations took place in developing machine learning algorithms. Machine Learning has become very important tool for cloud computing and ecommerce, and various technologies. Nowadays many large businesses are using machine learning to improve their sales and performance.

*Deep Learning:*

The concept of deep learning roots into the decade of 1940. Scientists Warren McCulloch and Walter Pitts in the year 1943 have proposed artificial version of neuron. Which is the computational element which can mimick the human brain’s neuron called Perceptron. This led a foundation for the further research work in the field of Deep learning. Around the same time, Frank Rosenblatt introduced the idea of a device called the Perceptron. In the year 1969, Marvin Minsky and Seymour Papert put forward the book named Perceptrons discussing the difficulty of training multi-layer neural networks.

Recent advances in machine learning have also led to breakthroughs in reinforcement learning, a technique for training machines to learn how to behave in an environment by trial and error. This has resulted in impressive results in game-playing, such as the systems developed by DeepMind. In 2014, Ian Goodfellow published a paper on generative adversarial networks (GANs), which have become a major focus of research in reinforcement learning alongside machine learning.

*References:*

1. Rockwell Anyoha (2017). The History of Artificial Intelligence ; https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/
2. Dilsizian, Matthew & Siegel, Eliot. (2018). Machine Meets Biology: a Primer on Artificial Intelligence in Cardiology and Cardiac Imaging. Current Cardiology Reports. 20. 10.1007/s11886-018-1074-8.
3. Few references include from: <https://www.livescience.com/47544-history-of-a-i-artificial-intelligence-infographic.html>.
4. An article from https://www.mckinsey.com/featured-insights/artificial-intelligence/deep-learnings-origins-and-pioneers.

**Q2**

Artificial Intelligence is a vast and base section in the field of computers intelligence. There are further sections in inside AI. Machine Learning, Neural Networks and Deep Learning, all these are the subsets of Artificial Intelligence. However the terminologies, literatures, features and problem solving techniques differ among them. We can also understand that deep learning is the part of machine learning.

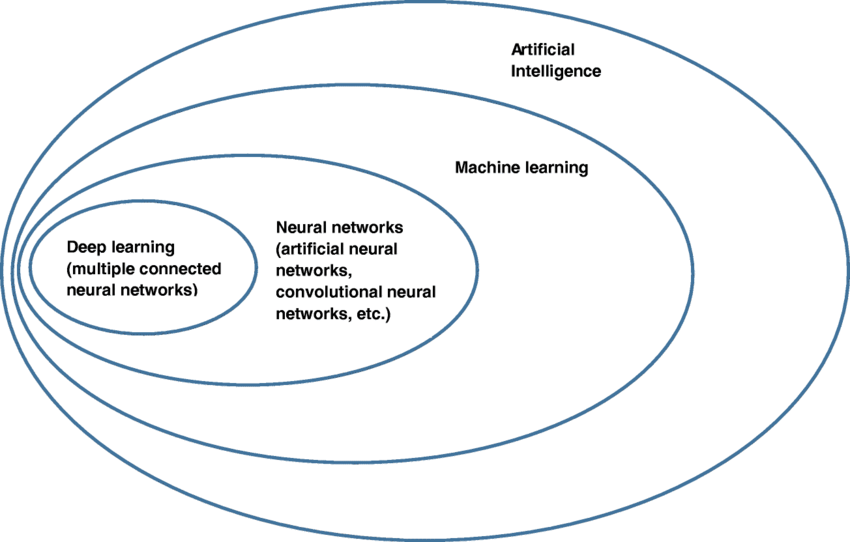


Figure 2: (taken from https://www.researchgate.net/figure/Deep-learning-and-machine-learning-as-subsections-of-artificial-intelligence\_fig1\_328367178)

Artificial Intelligence is the arching system. Artificial Intelligence is a broad field that encompasses deep learning and machine learning. Its aim is to solve problems, perform classifications, segmentations, and analysis. It works and assists in decision making and what-if analysis.

Machine learning is the subset of AI. It makes prediction about the data after training on data which of similar kind but not same and minimalize errors to produce accurate results. There will many steps involved in building the machine learning model right from choosing the data set, cleaning the data, preprocessing it to remove unwanted features and rows, understanding the data with the help of statistics and charts, building the model, testing and tuning. Its an iterative process. As the models train on data hence the model characteristics changes based on the quantity of data provided as training the model. Types of machine learning models are supervised, unsupervised and reinforcement learning models.

Deep learning which is subpart of machine learning also works on the same viewpoint which is about solving the problem there by producing more accurate results. As discussed earlier the difference is process and way of learning. Deep learning models have the underlying neural network as an algorithm. Each layer in the network takes the data from previous layer and cumulative the results and applies function whether it need to be forwarded to next layer or not. Finally the output layer receives all the results from nodes and gives as an output based on evaluation parameter like accuracy.

While there are few types of deep learning models as well. They are Convolutional Neural Network, Recurrent neural network and multilayer neural network.

*Relationship between Deep Learning, Machine Learning and Artificical Intelligence:*

Deep learning is subset of Machine Learning. Algorithms in both categories learn from data and they only differ in processing and ways of learning. The machine learning algorithms learn progressively and become better based on learning new data, these may need little human interventions to correct the model if it makes incorrect predictions. While deep learning models, can predict results accurately based on neural network. They do not need any kind of human interventions. They look similar to human brains.

Further, machine learning models use thousands of data points, and performs well with smaller data sizes. While deep learning models use millions of data points which makes models to work best for huge datasets. Secondly, machine learning models solves problems explicitly while deep learning models solves problems based on neural networks. Thirdly, ML models takes less time to train hence results appear quicker. DL models takes time from few hours to weeks based on data size, layers and nodes.

Reference:

1. Few points referred from Zendesk blog post at <https://www.zendesk.com/blog/machine-learning-and-deep-learning/>
2. Few points referred from IBM blog at <https://www.ibm.com/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks/>.

**Q3**

why Deep Learning is very popular in recent years.

As we discussed what is deep learning and how it differs from machine learning and artificial intelligence. Deep learning has many varied features than machine learning. Also, deep learning so popular nowadays and many projects hanging over deep learning in their projects. We can see the google trends curve in the below screenshot showing popularity gained about deep learning over the years.

**A graph on a white background

Description automatically generated**

(Image taken from google trends website)

Deep learning models can process large amounts of data. Andrew Ng, a well-known scientist says “*The analogy to deep learning is that the rocket engine is the deep learning models and the fuel is the huge amounts of data we can feed to these algorithms.*” In the recent times, we can see customizable processor which can support heavy workloads. Almost all cloud providers, provide services on CPU and GPU supported virtual machines, which are very essential for deep learning models.

Moreover, deep learning does not require lots of human intervention in building the models like machine learning. They can learn and select required components in the models and build the model. For example, we perform preprocessing steps like feature extraction and feature selection in the machine learning models but this is not the case in deep learning, they work by themselves during preprocessing steps.

A diagram of a performance curve

Description automatically generated

(Image taken from professors slides)

Performance of the models have improved in deep learning over machine learning models when we use similar tests. Hence although deep learning models take huge memory and cost, deep learning are very efficient in producing better results with less interaction.

Reference:

1. Sambit Mahapatra (2018); blog post on towardsdatascience.com; https://towardsdatascience.com/why-deep-learning-is-needed-over-traditional-machine-learning-1b6a99177063.

**Part II**

In this section, I have read the Pima diabetes data from csv file. Data consists of various parameters of about whether a patient have chance of getting diabetes. Data consists of various patient characteristics like blood pressure, pregnancy, skin, body mass index, age, insulin etc. data has fields 'preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'class'. Here, the class field we use as target variable to predict whether patient have chance to get diabetes.

I have performed basic preprocess of the data like understanding about data in terms of rows and columns, datatypes, column names and number of nulls present in the dataset. I got zero nulls in all the columns in the dataset. Later I have performed exploratory data analysis. In this subsection, I have plotted histograms for each column to see the distribution about data. Few columns exhibit normal distribution while few others show uniform distribution. Secondly, I have plotted box plots for all the columns in the dataset. I could see there are few outliers in some of the columns in the dataset. Later, I have divided the predictors and target variables and represented them as X and y respectively. Later using this data, I have split them into train, test and validation sets in 0.6:0.2:0.2 (train:test:validation). I have not defined any encoding algorithms on target variable since data is already in 1’s and 0’s (however I tested with LabelBinarizer, but it showed up similar results).

I started building model using Keras sequential modelling. I have taken the 2 layers as in below screenshot. In the first model (I named it as sequential\_model() ), I have defined 16 nodes in the first layer. I have used Relu activation function in the first layer and sigmoid function in the second layer with one node as I have only one column for output. For compiling the model, I used binary\_crossentropy as a loss function since my output has single binary output column. I have used Adam optimizier and evaluation metric as accuracy.

A screenshot of a computer program

Description automatically generated

I have trained the above model with X and y train datasets. I have used 150 as epoch and batch size of 32. We can see the below results using the first model.

A screenshot of a computer

Description automatically generated

training accuracy = 73.26%

validation accuracy = 68.83%

we can see that the loss plot has dropped to lower values suddenly at the lower epochs only, and kept same till 150 epochs. When we see accuracy over epochs plot, we can interpret as training accuracy (green) is way higher than the validation accuracy (red). Which means, the data model is exhibits overfitting. I tried to resolve overfitting problem in the second model which will be discussed in part III.

A graph of loss and loss

Description automatically generated

A diagram of a network

Description automatically generated

**Keras Classifier**

KerasClassifier has been imported and used to evaluate above model. I have used 150 epochs and batch size as 32. I have used sklearn’s KFold to build 10-fold cross validation and I used accuracy as scoring measurement.

A screenshot of a computer

Description automatically generated

I got highest evaluation accuracy as 79.22% after performing 10-fold cross validation. In the above

**Part III**

As we discussed in the previous section, we have seen the model is getting overfitted for the data we provided. Which is a critical problem. I tried to resolve it by performing hyper parameter tuning on above model.

I have created another model under function named improved\_sequential\_model(). In this model, I have created 2-layer sequential model.

A screenshot of a computer code

Description automatically generated

Using this model, I could see the results showing up the gap between training and validation curve has reduced drastically. We can also see that loss plot has less noise when compared to previous model built. It seems like accuracy have reduced after improvement.

Training accuracy = 67.83%

Validation Accuracy = 68.18%

These results are observed by reducing the nodes in the first layer from 16 to 7 keeping all parameters in place. Doing this will make model to reduce overfitting. Also, we have less expensive power, complexity and reduced loss of information.

A screenshot of a computer

Description automatically generated

A screenshot of a graph

Description automatically generated

A diagram of a model

Description automatically generated