A Review of Artificial Intelligence Methods for Data Science and Data Analytics: Applications and Research Challenges

Krishna C V, Rohit H R, Mohana
Telecommunication Engineering
R. V. College of Engineering
Bengaluru, India

***Abstract*— Artificial intelligence is a field which requires multidisciplinary expertise where the final goal is to automate all the human activities that presently require human intelligence. The major problem is to develop a method which works exactly the way how a human brain works. The architecture of artificial intelligence must emphasize on evaluation and redesign the nature of design process. Data science is also trending now and analytically deals to solve complex problems. Data is divided into smaller parts and its trends, behaviors are understood. The main problem in data science is to handle large quantities of data. Though there is significant increase in terms of research opportunities few challenges like lack of compute power, people power still remains a big challenge.**

Keywords—Deep Learning, Machine Learning, Artificial Intelligence, Data science and Data Analytics.

I. Introduction

Artificial intelligence (AI) is the intelligence exhibited by machines. Artificial intelligence is a method to simulate human intelligence using a set of algorithms and produce a new machine which can do similar work with human consciousness and also to perform parallel computing. Machine learning is subset of artificial intelligence that gives the path towards designing of computers that are intelligent. Deep learning is a subset in machine learning, used to represent data abstraction though predefined model architectures. Deep learning replicates the working of human brain in data processing and creates patterns, reduces it if possible and produces accurate results. This paper describes the methods of AI, applications, hardware and software resources used and some of the research challenges.

II. Evolution of Data Science and Data Analytics

The digital era of data science, was started in the mid-1900s, where the Statistical Analysis System (SAS) was the foundation of a project in North Carolina State University's (NCSU) agricultural department. When data science was first started in the industry level the main objective was to find more accurate and reliable solutions than those obtained using business analysis.· Data processing, predictive modeling, visualization, Sought Skillsets were those skills that an analyst must have to work on the data science field. In the present time python and R are the main technologies used for data processing [1]. However in the future Google's Go programming language may be used for data processing and analysis. Data science is evolving at an exponential space due to the availability of tools, technologies and resources. It addresses many real word problems with optimistic solution [2] [3] [4].

A. Machine Learning

Machine learning technology has been existence from the early 1950's. In 1990 data driven approach is transformed in to machine learning. During the time period of 1995 to 2005 there was a shift of focus towards the natural language search and information retrieval. The neural network made a comeback during 2005 where it was first tried in 1957 for first neural network computers. Machine Learning is among those technologies which have a number of successful and some failure stories but in the near future (2 to 5 years) there are possibilities of this technology to acquire mainstream. To keep up the growth in machine learning industry there should also be growth in some of the factors affecting it such as infrastructure and technical skill

B. Deep Learning

In 1965 Deep Learning is introduced by Alexey Grigoryevich Ivakhnenko and Valentin Grigor' evich Lapa [4]. They used some models with polynomial functions and complicated equations that was analyzed statically. In 1995, method for recognizing and mapping related or similar data was developed. During 1997 long short term memory for recurrent neural networks are established. During the late 1990's with the development of the processors with fast computational speeds, which eventually multiplied the computational speed by an approximate factor of 1000 times which made the processing of pictures by the GPU's efficient. Different layers of pre training and improvement of long short-term memory were used during the early 2000's and by 2011, with the growth of speed of GPUs,which made machines to work on convolutional neural networks which doesn't require the layer-by-layer pre-training. Currently the processing of Big Data is dependent on Deep Learning. Currently AI and Deep Learning growing and more advanced ideas are coming up.

C. Artificial Intelligance (AI)

In 1951 AI programs are written and run on Ferranti mark 1 machine. AI research field initiated in 1956 at Dartmouth College during summer workshop [5]. During those times there was the problem of computer hardware resources for computations. During the 1980's billions of dollars were provided by the industry and government for the development of AI after the insisting by the Government. During 2000 to 2010 after the funding and interest shown to develop the field of AI, there was a boom in the field. After the development of powerful computer hardware machine learning became a successful approach to a number of problems in industry and society.

III. Role of artificial intelligance in data analytics

Artificial intelligence and data analytics are hot research topics in many parts of the world. Developing countries are investing millions to gain prominence in world. Few of them are investing a small amount in non-classified artificial intelligence research. Artificial intelligence is a boon for industries in many sectors.

A. Machine Learning methods

Decision trees uses variables or decision nodes in a hierarchy and gives answers step by step. They are useful in evaluating lists of noticeable qualities, features. Support vector machine is a supervised learning model that analyzes data used for classification and regression analysis. They are good for binary classification of one variable versus other variables and there is no restriction for the relationship between variables to be linear. Naive Bayes classifier is based on concept probabilistic classifiers based on Bayes theorem with independence assumptions made between features. They compute the combined, conditional probabilities of multiple attributes.

B. Deep Learning Methods

Back propagation algorithm is used to calculate gradients which are needed for weight calculation to be used by a network. It has two modes of learning namely stochastic and batch. Dropout is a technique used for dropping the units randomly during deep learning training. It is basically used for reducing the number of parameters. Skip-gram is a method where if two vocabulary terms are similar only if they share similar context. For example given a word in middle of any sentence it looks at words nearby and picks out one word at random. The network will tell us the use of nearby word in our vocabulary in terms of probability.

IV. Software and Hardware Resources

A. Hardware resources

For AI Technology many of researchers, data scientists and research organizations are using NVIDIA or INTEL hardware GPU components for training, testing and to accelerate their deep learning algorithms. GPU performs parallel computing tasks for data, images, videos and graphics efficiently with less duration. Nvidia Ai Chip-Nvidia's new Jetson Xavier computer; it is compact size hardware contains various processing components. It includes Volta tensor core GPU, 8 core ARM 64 CPU, 2 NVDLA accelerators and processors for images and videos. It contains nine billion transistors and computes 30 trillion operations per second (TOPS) and consumes 30 watt of power. Intel AI Chip-Neural network processor provides flexible support for all deep learning primitives by making efficient core hardware components. It maximizes utilization of computation and scales multiple compute nodes with less power.

B. Software Resources

Softwares and tools used for artificial intelligence and deep learning are explained.

Pylearn2: designed for machine learning algorithms, flexible and extensible and it provides a library consisting of GPU and CNN.

Torch: It is AI and DL tool provides machine learning open source library. Scripting language used is Lua. Because of its flexibility and computational efficiency one of the most popular tool. But the drawback is reimplementation logic is not supported. Torchnet is the new open source framework, it is reusable and efficient.

Theano: It is scripting language python library allows evaluating mathematical expressions efficiently. It involves multi-dimensional arrays. Theano Faster than CPU in data-intensive calculations because it uses mathematical expressions. For large data sets, using to train deep learning networks [6].

Caffe: It is an open source framework supports various libraries such as C++, python, MATLAB, CUDA for AI deployment architecture.

Cuda-convnet: Fast neural network supports C++, CUDA and Python. It implements convolutional neural networks.

Deeplearning4j: It is an open source framework supports various libraries such as C, C++, Java, Scala. It has a GPU support for distributed frame work library. Supports all most all deep neural networks.

TensorFlow: It is an open source framework for numerical computation. It is used where computations can be modeled as data flow graph. It runs faster written in python. It has a CUDA support, interfaces in C++, and also available on embedded platforms.

Table I describes the comparision of various AI hardware and software resources with respect to hardware utilization, architecture, performance, platform and interfaces. Table II explains advantages and disadvantages of various AI methods. Table III depicts the usage of AI in various applications [9] [10] [11].

V. Trends and Market Analysis

Global AI market is expected to grow at a robust pace over the next five years, owing to its widespread implementation in numerous industries, such as automobile, finance, healthcare, consumer electronics, etc. Moreover, growing investments coupled with increase in number of acquisitions of start-ups by leading players such as Google, IBM, Microsoft and others companies are playing a huge role in pushing the global artificial intelligence market. During 2017 to 2022 the CAGR is expected to grow over 60% by the survey done by global AI market 2012-2022 [7]. As per the growth of global AI market data analysis, NLP, audio and video processing, image recognition and gesture control has a major share. Among these categories, image recognition dominated global artificial intelligence market in 2016. And it is expected to continue the dominance in next 5 years as well.

VI. Applications of data science and data analytics

Today's world has many applications of artificial intelligence and research rate is increasing to make human life as simple as possible. Most of the search engines use artificial intelligence algorithms to predict the search by user. E-commerce sites uses artificial intelligence to recommend the users about products based on their recent searched products. Also the delivery logistics companies use data science to improve their efficiency in operation. Websites like Junglee, Trivago and many more are being driven by huge amount of data which is fetched using RSS Feeds and APIs. Apart from the above few examples data science and artificial intelligence are used in every possible industry where data is generated and also where data is analyzed.

VII. research Challenges of data science and data analytics

A. Reasoning, problem solving

During the late 1980's and early 1990's the algorithm used to provide solutions for large reasoning problems were not sufficient because there were a lot of combinations possible for a problem. This reduced the computational speeds to decrease exponentially as the problems grew larger and larger. The concepts of probability and economics were the solutions developed by the researchers of AI to deal with uncertainty of information or incomplete information.

B. Security

Security faults have been detected and studied in AI.An image classifiers can be easily fooled by providing bad images or any moving images or any kind of images which are perturbed

C. Motion and manipulation

AI robots when given a small static environment then it can easily locate the place and map its environment but on the other hand in a dynamic environment ,or where the movement requires physical contact with the object then recognition of those objects become difficult to program [5].

VIII. Conclusion

In current world we are advancing more towards artificial intelligence and we are also witnessing the wide spread of it. It is also important to realize the capacity of these technologies in many ways that minimizes the risks safely. Machine learning algorithms are now being used to predict analytics. Deep learning is considered to have more advantages than the conventional machine learning techniques in producing accurate results. The more progress we make in this field, the complexity and the challenges associated with it increases.

Referencess

[1] Data technology Landscape and evolution of data science . https://www.digitalvidya.com.

[2] Evolution of data analytics: then, now and later-Affineblog.

[3] The evolution of machine learning synectics for management decisions. http://www.smdi.com.

[4] A brief history of deep learning- Data versity.

[5] J Vincent , " Nvidia launches AI computer to give autonomous robots better brains" the verge, 2018.

[6] Welcome – Theano 1.0.0 documentation. http://deeplearning.net/software/theano/.

[7] Global Artificial intelligance market size and trends. http://www.techsciresearch.com.

[8] V. Kumar and M. L. Garg, "Deep learning in predictive analytics: A survey," *2017 International Conference on Emerging Trends in Computing and Communication Technologies (ICETCCT)*, Dehradun, 2017, pp. 1-6.

[9] P. Sherkhane and D. Vora, "Survey of deep learning software tools," *2017 International Conference on Data Management, Analytics and Innovation (ICDMAI)*, Pune, 2017, pp. 236-238.

[10] J. Liu *et al.*, "Applications of deep learning to MRI images: A survey," in *Big Data Mining and Analytics*, vol. 1, no. 1, pp. 1-18, March 2018.

[11] F. M. Hohman, M. Kahng, R. Pienta and D. H. Chau, "Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers," in *IEEE Transactions on Visualization and Computer Graphics*.

[12] Manjunath Jogin, Mohana, Madhulika, Meghana R K "Feature extraction using Convolution Neural Networks (CNN) and Deep Learning" *2018 IEEE International Conference On Recent Trends In Electronics Information Communication Technology,(RTEICT)* 2018, India.

[13] Arka Prava Jana, Abhiraj Biswas, Mohana, "YOLO based Detection and Classification of Objects in video records" *2018 IEEE International Conference On Recent Trends In Electronics Information Communication Technology,(RTEICT)* 2018, India.

[14] Chandan G, Ayush Jain, Harsh Jain, Mohana, "Real Time object detection and Tracking using Deep Learning and openCV" *International Conference On Inventive research and computing applications ,(ICIRCA)* 2018, Coimbture,India.

TABLE I. COMPARISION OF ARTIFICIAL INTELLIGENCE HARDWARE AND SOFTWARE RESOURCES [9][12][13][14]

	Pylearn2	Torch	Theano	Caffe	Cuda-convnet	Deeplearning4j	Tensor Flow
Extensibility	More Easy	More Easy	More Easy	Not Easy	Not Easy	Easy	More Easy
Hardware Utilization	CPU or GPU	CPU,GPU	CPU,GPU	CPU,GPU	GPU	CPU or GPU	CPU
Performance	Good	Good	Best	Good	Best	Good	Good
Architecture	Components: Dataset, Model, and Training Algorithm classes	Well design with modular interface	Fairly Hacky	Standard layer wise design	Convolutional feedforward neural networks		Components: Training libraries, Inference libraries, Python client, CAPI, kernel implementation, Networking layer, design layer
Ecosystem	Python	Lua	Python	C++	C++/CUDA	Java	Python API
Platform	Cross Platform	Linux based	Cross Platform	Cross Platform	Cross Platform	Cross Platform	Cross Platform
Open source	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Interfaces	Github	LuaJIT	Python	Pycaffe	CUDA		
Modeling Capability	Contrast	Excellence	Good	Excellence	Contrast	Contrast	Contrast

TABLE II COMPARISION OF ARTIFICIAL INTELLIGENCE HARDWARE AND SOFTWARE RESOURCES: ADVANTAGES AND DISADVANTAGES [9][12][13][14]

Advantages	Pylearn2 is designed for flexibility and extensibility of machine learning algorithms	Easy to write your own layer types and run on GPU, Lots of pre trained models	Quickest, Computational graph is nice abstraction, High level wrappers ease the pain	Train models without writing any code, feed forward networks and image processing	Efficient implementation of convolution in CUDA	Library is written for java and java virtual machine, Library can run on both Scala and Clojure, portable and platform neutral, parallelism is automatic	Has CUDA support, also provides interface to C++, available across all operating system platforms, implementation of data parallelism
Disadvantages	Load all data set to main memory	Not good for recurrent neural networks	Error messages can be unhelpful, Large models can have long compile times	Need to write C++ / CUDA for new GPU layers, Not extensible, bit of a hairball	It requires High performance GPU.	No clearly compelling features and no vibrant community around it.	No GPU support other than Nvidia and only language support, less computation speed, missing symbolic loops

TABLE III COMPARISION OF ARTIFICIAL INTELLIGENCE HARDWARE AND SOFTWARE RESOURCES ADVANTAGES AND ITS USAGE IN APPLICATIONS [9][12][13][14]

Application	State of the art on MNIST, CIFAR- 10, CIFAR- 100, and SVHN.	Face-book, Twitter	Computational graph	Image classification with convolutional nets	Object recognition in images, CIFAR-10	Easy to build an image classification web application	Speech recognition, object tagging videos, detection of flaws, air-sea-land drones