Volume 8, No. 2, March 2017 (Special Issue)
International Journal of Advanced Research in Computer Science
REVIEW ARTICLE
Available Online at www.ijarcs.info
ISSN No. 0976-5697

Cloud computing for deep learning analytics:A survey of current trends and challenges

Anam Saiyeda
Department of CSE, SEST
Jamia Hamdard, New Delhi, India
anam.7sd@gmail.com

Mansoor Ahmad Mir
Department of CSE, SEST
Jamia Hamdard, New Delhi, India
mansoormir500@gmail.com

Abstract: Deep learning, a sub-field of machine learning is inspired by the principle of information processing in the human brain. Its applications are autonomous driving, robotics control, machine translation etc. It needs multiple training examples of the task, specialized GPU hardware, capital investment and its libraries evolve quickly, so frequent updates are needed. Cloud computing is a type of computing in which computing resources are provided on demand. Cloud is an apt choice for a platform for deep learning analytics as it provides servers, storage and networking resources. It provides scalability, processing, storage and analytics resources. Deep learning algorithms like CNN are computationally intensive for a commercial computer for large larger datasets. Cloud computing allows prevailing the processing, memory constraints of average computers, allowing computations on larger datasets. This paper discusses how cloud computing allows us to overcome the constraints of deep learning analytics on average systems and the various platforms offered by various providers. Google, NVIDIA, IBM provide us platforms for deep learning. IBM's Rescale, Nervana Cloud , Google deep learning cloud provide full-stack hosted platform for deep learning.

Keywords: Deep learning, Artificial intelligence, scalability, virtualization

I. INTRODUCTION

Artificial intelligence is ubiquitous. From daily transactional tasks like online shopping to bank transactions to robotics every field is affected by it. Deep learning a part of machine learning has made its presence felt in the machine learning world. Major players like Facebook, Microsoft and Google are all using it. However, for deep learning to be effective it requires huge amounts of data. Deep learning architecture ensures many layers of the neural network. "Deep" will be useful when the depth i.e. number of layers are more in number. This requires more storage for this large amount of data needed for training. The power requirements also increase as the tasks become computationally intensive. So the traditional computers may not work very effectively. Also this leads to more capital investments by the company. So an easier and effective way would be the use of the services provided by the cloud for performing deep learning. analytics are discussed followed by the conclusion.

II. RESEARCH METHOD

2.1 Need for review

A thorough and exhaustive review is needed to find the gaps in existing methodologies and techniques. This paper intends to summarize the existing techniques and platforms available for performing analytics using deep learning on the cloud platform. One of the purposes of the review is to study all available techniques, challenges to traditional systems, available platforms and a justification of the use of the cloud platform for deep learning analytics.

2.2 Sources of information

In order to carry out the review, quality journals were searched and the filtering was done as per the universally available guidelines. The papers were taken from the databases of IEEE, Springer and ACM

2.3 Search Criteria

The search convention included all the papers associated to analytics on the cloud. The keyword <deep learning analytics on the cloud > and <big data analytics on the cloud> were used. The search was further refined with the keywords <Computer Science>. Irrelevant papers related to other disciplines were filtered out and the results are summarized below.

Search with the keyword in the Springer database gave 286 results. The papers were ordered on the basis of relevance. After refining the search to include only the papers from computer science, 178 results were obtained. These contained 152 articles, 26 Chapters, 6 conference papers.

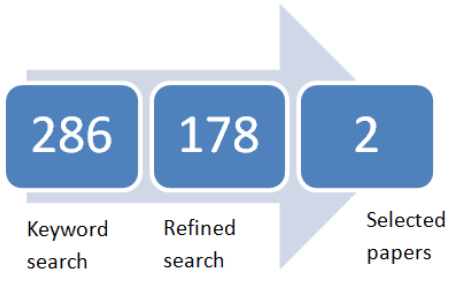

Fig 1.1: Selection of papers from Springer database

In the ACM digital library, the keyword gave 289,237 results. Refining the search andstudying abstracts gave us 5 relevant papers finally.

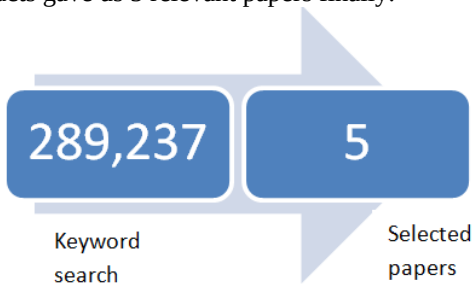

Fig 1.2: Selection of papers from the ACM database

The IEEE database the keyword gave 566 results. Refining the search and studying abstracts gave us 9 relevant papers finally.

Fig 1.3: Selection of papers from the IEEE database

2.4 *Study selection*

Initial filtering of papers at the first step was done on the basis of titles. Then on the basis of keywords the search was refined and further selection was done. The abstract ofthe remaining papers were studied. After reading the complete text the final set of papers was selected. 16 papers were included in the final set.

2.5 Data extraction

The quality of the papers was the prime benchmark considered for the purpose of paper selection. For the study 16 papers were finalized.

2.6 *Research questions*

The purpose of this review is to justify the use of cloud platform for deep learning and present the various techniques available for deep learning analytics on the cloud. The selection of papers was done according to the following research questions. The questions are as follows.

RQ1. What are the problems and challenges faced by traditional systems while performing deep learning analytics?

RQ2. How will the use of cloud platform improve the analytics done using deep learning algorithms?

RQ3. What are the existing techniques available for performing deep learning on the cloud?

III. LITERATURE REVIEW

Research has been carried out to justify the performing of big data analytics on the cloud. The datasets in deep learning can be thought of as big data as it involves large sets of images, videos, audios. The relevant papers have attempted to justify the use of cloud platform for this.

The focus of a paper by Tsai, Chun-Wei, et al[12] was on the question that how to develop a high performance platform to efficiently analyze big data. It also attempted to design an appropriate algorithm which can find useful information from big data. Salloum et al's[13] paper studies the analytics on the platform Apache Spark.It is a framework for big data analytics with its advanced in-memory programming model, upper-level libraries for scalable machine learning, graph analysis, streaming and structured data processing. Another paper by Middelfart, Morten[14] performed cloud based deployment of analytics and business intelligence using two different approaches i.e an analyst specialist platform and a social platform. Mohammad, Atif et al [15] developed a Big Data Architecture, considering its relation with Analytics, Cloud Services and Business Intelligence. Fiore, Sandro, et al [16] proposed a cloud infrastructure for big data analytics for climate change and biodiversity. Hamdaqa, Mohammad, et al.[17] proposed a MapReduce framework for ad-hoc cloud computing. For analytics of big data new frameworks on the cloud have been proposed. Analytics on healthcare big data is a task for getting insights into a very large data set for improving the health services.Farheen [] discussed about distributed cloud computing approaches for big data.

This enormous amount of data puts a great deal of stress on the write performance, scalability, requirement of efficient storage and meaningful processing of this data. The traditional relational databases are insufficient for this. So a new big data storage architecture was proposed consisting of application cluster, storage cluster to facilitate read/write/update speedup and data optimization. [19] A License Plate Recognition System (LPRS) was proposed using deep convolutional neural network on a cloud platform for plate localization, character detection and segmentation. Use of bare-metal cloud servers with kernels optimized for NVIDIA GPUs, led to an increase in the speed of training phase of the CNN LPDS algorithm. The paper shows the superiority of the performance in recall, precision and accuracy compared to traditional LP detecting systems. [20] A paper designed deep Convolutional Neural Networks (CNNs) to mine the deep features of cloud. [21] Work has been done to predict VM workload in the cloud using deep learning [22], develop efficient mobile cloud system for deep learning, [23]feature extraction for 3d point cloud data using autoencoder[24] Deep computation models have been designed to offload the expensive operations to the cloud. [25]Another research shows the benefits of hardware acceleration and the high performance gains on the cloud. [26]Thus a lot of research has been done in the field of performing analytics on the cloud and the various applications of big data analytics via cloud[27]. Thus it is a field with a lot of research potential.

IV. DEEP LEARNING

Deep learning is an approach to AI enabling computer systems to improve with experience and data. It represents the world as a nested hierarchy of concepts, where every relation is defined in terms of simpler concepts and more abstract representations are computed in terms of simpler ones. [1]
It is the latest trend in machine learning being used by the major players like Facebook, Google, Microsoft etc. Google uses it in the Android Operating System's speech recognition system, photo search for Google+ and video recommendations in YouTube. Google's deep learning research project is Google Brain. Facebook uses deep learning for Textual analysis, facial recognition, targeted advertising and **designing AI applications**. Facebook developed a tool DeepText to extract meaning from words posted on the social networking site. It does this by learning to analyze them contextually. Relationship between words is analyzed using neural networks to understand how their meaning changes depending on other words around them. It is a semi-unsupervised learning. Reference data like a dictionary with meaning of words is not available. Instead, it learns for itself. Based on people's conversations the tool directs them towards products they may want to purchase. DeepFace[2] is another deep learning application by Facebook to recognize people in photos. It is being used in various fields like Robotics, Affective Computing (sense human emotion), healthcare, cybersecurity, genomics, computer Vision, Conversational Interfaces.

The broad categories of tasks done using deep neural network are classification and finding patterns. Unsupervised learning tasks use auto encoders and Restricted Boltzman machine(RBM). For supervised learning tasks like image recognition DBN(deep belief networks), convolution nets are used, for object recognition CNN and RNTN used and speech recognition utilizes recurrent nets. For general classification MLP and deep belief nets are used. Time series uses Recurrent nets. The purely supervised learning algorithms are Logistic Regression, Multilayer perceptron and Deep Convolutional Network. The unsupervised and semi-supervised learning algorithms include Auto Encoders, Denoising Autoencoders, Stacked Denoising Auto-Encoders, Restricted Boltzmann Machines and Deep Belief Networks.

In a deep learning algorithm the composition of a layer of nonlinear processing units depends on the problem to be solved. Layers in deep learning include hidden layers of an artificial neural network. The inputs of these algorithms are transformed through more layers than shallow learning algorithms. Like an artificial neuron whose parameters are 'learned' through training at each layer, the signal is transformed by a processing unit. In deep learning higher level more abstract concepts are learned from the lower level ones. It helps to disentangle these abstractions and pick out which features are useful for learning. A deep neural network (DNN) is an artificial neural network (ANN) with multiple hidden layers of units between the input and output layers.[3] DNNs are designed as feed forward networks. Convolutional deep neural networks (CNNs) are used in computer vision and in acoustic modeling for automatic speech recognition (ASR). [4]

A. *Requirements*

Deep Learning is a computationally intensive task. It may involve lots of operations and tasks like matrix multiplications on a large scale. Thus a GPU i.e. Graphics Processing Unit is needed for it. The GPU is the heart of deep learning applications. Apart from the GPU in order to run deep learning tasks other requirements need to be taken care of like Needed CPU clock rate (frequency), needed RAM clock rate RAM size, Power supply unit (PSU). The CPU does little computation when you run your deep nets on a GPU, but it does still performs the task of writing and reading variables of code, executing instructions , Initiating function calls on your GPU, Creating mini-batches from data and initiating transfers to the GPU. Apart from hardware the system running deep learning algorithms needs the capacity to store huge amounts of data which could be in the form of images, audio or video.

B. *Challenges*

Deep learning requires specialized GPU hardware which means that organizations need capital investment to have appropriate GPU resources on-premise. Also deep learning libraries are evolving very quickly, resulting in the need for frequent updates to stay current. Deep learning involves large datasets. Managing these datasets is not a trivial task. Storage of this huge amount of data is another challenge.

V. CONCEPT

Cloud computing is Internet-based computing. Clouds are distributed technology platforms that leverage technology innovations to provide highly scalable and resilient environments. [5] Shared resources, software and information are provided to computers and other devices on-demand, like the electricity grid. Hardware, systems software, applications are delivered as a service over the internet. The cloud provides service in the form of IaaS, PaaS and SaaS[6]

Cloud computing is the apt platform for deep learning analytics as the architecture provides support for Scalability, Virtualization, Storage for huge amounts of data-structured & unstructured and unlimited resources on demand.

Scalability is the ability of a system, network, or process to handle a growing amount of work in a capable manner. It is the most important factor for analysis of huge datasets. The traditional models spend a large amount of time in designing scale-up and scale-out solutions. Significant investments are

made on the hardware platform. Cloud Computing removes this overhead for the Architect/Organization by providing on-demand (elastic) computing resources on the fly. Role of the architect is then reduced to finding right cloud vendor.

The cloud model offers database scalability. It is able to handle very large amounts of data and provide the input/output operations per second (IOPS) necessary to deliver data to analytics tools. It offers storage for both structured and unstructured data. Thus database scalability, distributed computing and virtualization ensure there is never shortage of storage space The Cloud Computing model provides unlimited resources on demand. Big data environments require clusters of servers to support the tools that process large volumes high velocity and varied formats of data. Clouds are already deployed on pools of server storage and networking resources. Thus they offer cost effective way to support big data technologies. Use of cloud computing for big data implementation lowers the in-house processing power commitment by shifting the data processing to the cloud. Provide sufficient benefit to a small to medium sized companies.

Iaas involves taking the physical hardware and going completely virtual. e.g. all servers, networks, storage, and system management all existing in the cloud. This is the equivalent to infrastructure and hardware in the non-cloud computing method running in the cloud. This will mitigate the need for a data center, heating, cooling, and maintaining hardware at the local level. This service model is the one which can be used for big data storage. IaaS technology ups processing capabilities by rapidly deploying additional computing nodes. IaaS enables you to allocate or buy time on shared server resources. These are virtualized, to handle the computing and storage needs for big data analytics. Cloud operating systems manage high-performance servers, network, and storage resources.

Flexibility is allowing resources to be deployed rapidly and only as needed. Cloud computing puts big data within the reach of companies that would otherwise not be able to afford the high costs or invest the time associated with buying sufficient hardware capacity to store and analyze large data sets. Cloud computing hosts a pool of shared resources: provided to consumers. Resources such as compute, memory, network and disk (storage) are allocated to consumers of the service from a shared pool. Rapid elasticity provided by allowing rapidly provisioning and release of resources as demand for the cloud service increases and decreases. This is done automatically. The rapid elasticity capability enables you to save money on compute, memory, network and storage resources. This is because these resources scale up and are released when needed. When projects start, the resources are allocated to consumers of the cloud service, and when the project comes to an end, these resources are released back into the cloud infrastructure's resource pool.

Computation: The cloud computing platform Amazon EC2 has provided an API for instantiating computing instances with any of the operating systems supported. Facilitates computations through Amazon Machine Images (AMIs) for various other models. Google Big query is an example of how computation of big data can be made easy by cloud service providers. It allows Query services for very large datasets and provides accurate results quickly.

VI. SERVICES PROVIDED FOR DEEP LEARNING

Various companies are providing facilities to perform deep learning analytics on the cloud. Google Cloud Platform added support for NVIDIA Tesla K80 GPUs, providing new capabilities for deep learning processing for users. The NVIDIA GPUs have been integrated with Google Cloud Machine Learning and TensorFlow to help reduce the time taken to train machine learning models at scale. [7]

AWS and Microsoft Azure, the two leaders in the cloud IaaS space, have been working to provide GPU integrations.

Cirrascale is a company which is a specialist in designing and hosting compute infrastructure for deep learning. The company uses its data center near San Diego to provide this infrastructure as a service. It has similarities with Amazon Web Services in the way it provides its cloud servers. Unlike AWS, which provides virtual server instances, Cirrascale's deep learning cloud is a bare-metal cloud service. It provides a dedicated high-performance box to run the requisite software. [8]

Customers doing machine learning development work are new to the world of high-performance computing. Setting up, managing, and cooling an HPC cluster is not an easy and trivial task. Thus researchers will be happy to offload that problem to someone who understands it and focus on the analysis part.

A solution for hosting and distributing trained deep learning models on Algorithmia using GPUs in the cloud. Researchers and developers can train their neural nets locally, and deploy them to Algorithmia's scalable, cloud infrastructure. There they become smart API endpoints for other developers to use.

Native support for the Caffe, Theano, and TensorFlow frameworks, and 15 open source deep learning models that run as microservices to start. [9]

GPUs on-demand and running in the cloud, eliminate the manual work required for teams and organizations to set up and experiment with cutting-edge deep learning algorithms and models, which allows them to get started for a fraction of the cost.

Nervana Cloud is a full-stack hosted platform for deep learning. It allows developing and deploying high-accuracy deep learning solutions at a fraction of the cost of building your own infrastructure and data science team. The company's open source deep learning framework is called neon. Nervana Cloud is optimized down to the silicon to handle the most complex machine learning training at scale. Nervana has achieved training speeds 10x faster than conventional GPU-based systems and frameworks. [10]

Rescale's Deep Learning Cloud, in partnership with IBM, provide you with powerful hardware to visualize and explore large datasets and design deep neural network models. [11] When the need arises to scale up and train on large datasets, instant access to GPU compute clusters is provided and payment is done only for what you use with hourly pricing. Deep Learning Cloud comes configured with the latest versions of popular deep learning software like TensorFlow and Torch, as well as IBM-licensed analytics products like SPSS. All software is already configured to take full advantage of NVIDIA GPUs

VII. CONCLUSION

The combination of advanced analytics software and the availability of cheap processing power makes the cloud a perfect place to perform analytics using deep learning. Machine Learning Is Everywhere and deep learning is the phrase de jour. The Cloud's power is inescapable. Analysis, computation and statistics are made easier on the cloud and the workloads are highly variable. Deep learning requires heavy computing resources. It is cost prohibitive to build the infrastructure yourself and power it locally. Deep learning in the cloud can utilize the massive infrastructure available online thus the combination of these two will be feasible.

VIII. REFERENCES

[1] Ian Goodfellow, Yoshua Bengio, Aaron Courville, "Deep Learning", MIT Press, 2016

[2] Taigman, Yaniv, et al. "Deepface: Closing the gap to human-level performance in face verification." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014.

[3] Deng, Li, Geoffrey Hinton, and Brian Kingsbury. "New types of deep neural network learning for speech recognition and related applications: An overview." Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013.

[4] Ciresan, Dan C., et al. "Flexible, high performance convolutional neural networks for image classification." IJCAI Proceedings-International Joint Conference on Artificial Intelligence. Vol. 22. No. 1. 2011.

[5] Thomas Erl, Ricardo Puttini, Zaigham Mahmood, "Cloud Computing: Concepts, Technology & Architecture", Prentice Hall, May 2, 2013

[6] Jin, Hai, et al. "Tools and technologies for building clouds." Cloud Computing. Springer London, 2010. 3-20.

[7] Google Cloud Machine Learning at Scale, Google Cloud Platform, Available: https://cloud.google.com/products/machine-learning/

[8] Cirrascale Press Release". Retrieved 2017-02-25,available:http://www.cirrascale.com/documents/datasheets/Cirrascale_Storage_SG_CM018C.pdf

[9] Charlie Crawford, Algorithmia, Retrieved 2017-02-24, Available :http://blog.algorithmia.com/introduction-to-deep-learning-2016/

[10] "Nervana Systems Puts Deep Learning AI in the Cloud". IEEE Spectrum: Technology, Engineering, and Science News. Retrieved 2016-06-22.

[11] Mark Whitney, "How to Run Deep Learning on a Cloud that's Right for You", December 1, 2016

[12] Tsai, Chun-Wei, et al. "Big data analytics: a survey." Journal of Big Data 2.1 (2015): 21.Springer

[13]Salloum, S., Dautov, R., Chen, X. et al., "Big data analytics on Apache Spark", Int J Data Sci Anal (2016) Springer 1: 145. doi:10.1007/s41060-016-0027-9

[14] Middelfart, Morten. "Analytic lessons: in the cloud, about the cloud." Proceedings of the 1st International Workshop on Cloud Intelligence. ACM, 2012.

[15] Mohammad, Atif, Hamid Mcheick, and Emanuel Grant. "Big data architecture evolution: 2014 and beyond." Proceedings of the fourth ACM international symposium on Development and analysis of intelligent vehicular networks and applications. ACM, 2014.

[16] Fiore, Sandro, et al. "Big data analytics for climate change and biodiversity in the EUBrazilCC federated cloud infrastructure." Proceedings of the 12th ACM International Conference on Computing Frontiers. ACM, 2015.

[17] Hamdaqa, Mohammad, et al. "Adoop: MapReduce for ad-hoc cloud computing." Proceedings of the 25th Annual International Conference on Computer Science and Software Engineering. IBM Corp., 2015.

[18] Chaudhuri, Surajit. "What next?: a half-dozen data management research goals for big data and the cloud." Proceedings of the 31st ACM SIGMOD-SIGACT-SIGAI symposium on Principles of Database Systems. ACM, 2012.

[19] Manish Kumar Pandey, Karthikeyan Subbiah, "A Novel Storage Architecture for Facilitating Efficient Analytics of Health Informatics Big Data in Cloud", IEEE International Conference on Computer and Information Technology (CIT) Year: 2016 Pages: 578 - 585, DOI: 10.1109/CIT.2016.86

[20] Polishetty, Rohith, Mehdi Roopaei, and Paul Rad. "A Next-Generation Secure Cloud-Based Deep Learning License Plate.

[21]Farheen Siddiqui "Applications for Big Data in of Intelligent Distributed Processing "IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 6, Ver. I (Nov. - Dec. 2016), PP 61-64 DOI: 10.9790/0661-1806016164