A study of machine learning in the context of big data with a focus on deep learning

Anthony Galtier

***Abstract*— Big data defines itself by the challenges it poses. Big data goes beyond traditional processing and analytics and this does not exclude the field of machine learning. This paper provides an overview of the approaches to performing machine learning in the context of big data according to it's four "V" dimensions. As deep learning is standing out as the most prominent field of machine learning and one of the most promising in tackling the challenges of big data, this paper will also provide an overview of how deep learning can deal with each of these four "V" dimensions of big data.**

I. INTRODUCTION

The progress made since the beginning of our digital era in means of storing and collecting information has given birth to this present context of Big Data. The advent of smart phones, social networks, sensor networks, web technologies have led to an explosion in the amount and rate at which all sorts of data are generated. A report from IBM states that 90% of the data in the world today has been created in the last two years alone, at 2.5 quintillion bytes of data a day. [1]

This context has given birth to the concept of Big Data. Often referred to as a revolution, Big Data carries among others promises of great knowledge and insight discovery that will change the way we make decisions for good. However, research firms tell us that most collected information is never used. It sits uncleaned, unanalyzed, unused in databases. [18]

The realization of these promises relies on the ability to extract value from Big Data using data analytics [10]. Machine learning is key player in the field of data analytics and like other approaches, most of the assumptions on which it relies are broken in the context of Big Data creating a variety of new challenges. As a result and in order to realize the potential of Big Data, a lot of research has focused on finding new approaches to performing machine learning in this new context of Big Data. Out of all machine learning approaches, the most promising one is arguably deep learning. After introducing the concepts of big data and machine learning, this survey will firstly present the latest approaches to performing machine learning with Big Data and secondly, will focus on the latest and more specific approaches to performing deep learning with Big Data.

II. CONCEPTS

A. Big Data

Big data is a recent term whose definition is not rigorously formalized. It is sometimes qualified as a buzzword. That said, there is consensus about the characteristics and implications of Big Data. Big data is characterized by large volumes, high velocity and great variety and furthermore implies a need for new processing paradigms to extract it's value. In other words, Big Data starts where traditional approaches to data management and processing fail.

The characteristics of Big Data are usually broken up into "V" dimensions. Although more can be defined [11], the four core V dimensions of Big Data are it's volume, variety, velocity and veracity. Volume refers to the amount of data generated and stored. Variety refers to the diversity in structure and meaning of the data as well as the diversity of its sources. Velocity refers both to the rate at which data is generated and the rate at which it needs to be processed and analyzed. Veracity refers to the quality, reliability, completeness and trustworthiness of the data and its sources.

For a better understanding of the issues, in the following sections, the approaches to performing machine learning and deep learning with big data will be presented according to these four "V" dimensions.

B. Machine Learning

Machine learning is a field of Artificial Intelligence that aims at giving computer systems the ability to "learn" from data to perform tasks such as classification, clustering, pattern recognition and more with minimal human intervention. Machine learning uses statistical techniques and optimization to automate analytical model building.

Machine learning algorithms can be classified according to their learning method. The two main classes are supervised and unsupervised learning methods. Other learning methods include semi-supervised learning, reinforcement learning and active learning.

In supervised learning both the inputs and desired outputs of the data are known. The aim of supervised learning is to encounter the best fitting mapping function between the input and desired output of the data. Supervised learning, among others, consist in regression and classification tasks. Some popular supervised learning algorithms are K-Nearest-Neighbours, Linear Discriminant Analysis and Support Vector Machines for classification.

In unsupervised learning, only the input data is specified. The aim in unsupervised learning is to find the model that will best fit the underlying structure of the data. Unsupervised learning algorithms perform clustering and association tasks. Popular unsupervised learning algorithms are K-Means, Mean-Shift, DBSCAN (density-based spatial clustering of applications with noise).

Machine learning originated in the 1950s and since then has relied on a certain number of assumptions. Machine learning algorithms, for the most part, rely on the assumptions, among others, that the data sets fit entirely into memory, that the data be uniformly distributed across all classes of the data sets and that statistical properties be similar across the data sets. These assumptions made sense at time, however, in the present context of big data, many of these assumptions are broken. New challenges for machine learning have arisen from these broken assumptions. Thus, in the following section, the approaches to overcoming these specific challenges of big data will be presented according to the four "V" dimensions of big data mentioned earlier.

III. Approaches

When it comes to finding solutions to performing machine learning in the context of big data, there are two paths researchers can explore. The first is to develop entirely new machine learning algorithms that rely on assumptions that are consistent with the context of big data. The second, is to adapt already existing machine learning techniques to this new context of big data. It is this latter path that most research has preferred to focus on. The following subsections will provide an overview of this research organized accordingly to the "V" dimensions of big data.

A. Volume

Volume is the most obvious characteristic of big data and also a very challenging one for machine learning to deal with. As data set sizes increase both vertically, in terms of the number of data points collected, but also horizontally, in terms of the number of features, both the processing time and performance of machine learning algorithms suffer. The simple fact that increasing the scale of the data increases the computational complexity of task becomes a major concern when it comes to machine learning. Indeed, machine learning algorithms usually have a relatively high time and space complexity. Running machine learning tasks on voluminous data can become very costly if not unfeasible in some cases. Furthermore, data sets can be so large that they cannot fit entirely into memory. This issue is referred to as the curse of modularity. Another notable curse is the curse of dimensionality, as the horizontal size of a data set increases, the prediction performance of machine learning algorithms decrease. This is also known as the Hughes effect. [8]

To overcome these challenges of volume, research has brought forward a variety of solutions. A first approach is to improve the pre-processing with the idea of trying to mimic "small data" with "big data". Data reduction, although not new, has had renewed attention to deal with the volume.Data reduction consists in techniques to reduce the size of a data set with minimal loss of information, both

vertically with instance selection and horizontally with dimensionality reduction.
Instance selection consists in selecting the most representative subset of data to reduce the height of the data set. This selection can be based on statistical sampling, random, based knowledge of the domain. Clustering and genetic-based algorithms can also be used for this purpose.
Dimensionality reduction consists in mapping high dimensional spaces onto a lower dimensional space. A variety of techniques exist, there are linear mapping techniques like Principal Components Analysis dating back to 1901 that consists in finding orthogonal linear combinations of features, known as principal components, that minimize overall variance. Rather that using as input all the features of the data, one can use only a inferior number of principal components that will sufficiently represent the data. Other dimsensionality reduction techniques include non-linear techniques, random projections or the use of auto-encoders.
Beyond improving the pre-processing of the data, solutions have been proposed to tackle the processing of of machine learning algorithms. These solutions are based on distributed computing. Several machine learning algorithms like K-Means or Mean-Shift have been adapted to work in parallelization environments inspired by Google's MapReduce [4] like Hadoop. Dispersing processing over a network of computer like so is known as horizontal scaling or scaling out. Additionally to this, vertical scaling or scaling up has come as a solution to improving the processing of machine learning algorithms. This consists in increasing the capacity of hardware and software to accelerate the execution of the algorithms. Graphical Processing Units (GPUs) have proven their efficiency for inherently parallel machine learning algorithms. The use of GPUs has in particular become a must in the field of deep learning to train neural networks.

B. Variety

In the context of big data machine learning must adapt to the both the structural and semantic variety of the data that characterizes big data. Machine learning algorithms were not developed to handle data that is structurally and semantically diverse. Dealing with such variety is an active research topic. Approaches to solve this issues up to now consist in algorithm modifications for specific applications such as Pegasos [16] for text processing. The challenges of variety do not stop at the structural and semantic diversity. The variety of big data also refers to the variety of its sources. Indeed, the curse of modularity, implied by the fact that the entire data set cannot fit entirely into memory, means that to work with big data, machine learning algorithms must be able to deal with data coming from a variety of different sources. The approaches to deal with this variety of sources are the same as those mentioned earlier to improve the processing of voluminous data. Rather than bringing data to computation, research has focused on bringing the computation to the data with distributed computing and parallelisation of machine learning algorithms.
Among these distributed computing environments, batch oriented systems are particularly interesting to perform machine learning. These process large batches of data in parallel and thus can work with data in different physical location. Batch oriented-systems deal well with variety in the locality and allow for faster processing of voluminous data.

C. Velocity

In the context of big data, data is constantly generated and at an increasingly fast rate. The value of data is also more and more dependent on how fast it is processed and analyzed. Traditional machine learning algorithms usually rely on the entire data set being available at the time of training. This is no longer true in this context. Research has thus looked into incremental machine learning. [6] Incremental learning or sequential learning consists in algorithms that adapt their learning to constantly arriving new data. These algorithms, as opposed to traditional machine learning algorithms, do not need to retrain on the entire data set every time they are presented with new data. Support Vector Machines have in particular been well adapted to incremental learning. [20] Another challenge posed by the velocity dimension of big data is to adapt machine learning to real-time processing. Here again, distributed computing environments have proved the effectiveness, in particular stream-oriented systems. As opposed to batch-oriented

systems, stream-oriented systems operate only on one element or small data set in real-time or near-real time. These systems are also inspired by the MapReduce paradigm but without in-memory dependence. Streaming systems like Apache Storm and Yahoo S4 rely on a graph-based topology whereas others like Spark Streaming deal with micro-batches of data, making it more similar to batch-oriented systems. Although these systems allow for real-time or near-real time processing of data, the operations they can perform are less complex, and they are thus only suited for a few very simple machine learning algorithms. Adapting more machine learning algorithms to streaming systems is a promising field of research to manage the velocity dimension of big data.

D. Veracity

The veracity of a data set cannot be taken for granted in the context of big data. One must deal with measurement errors, outliers and missing values. Furthermore, a new challenge that has emerged is dealing with inherently imprecise data such as very subjective sentiment data from social networks. Traditional machine learning was not developed for complete, precise and trustworthy data sets. In this challenge of adapting machine learning to such imprecise and incomplete data, a first step is to collect contextual information of the data to evaluate and establish the veracity of the data set. To do so, a solution is to trace and record the origin and movements of data as it is processed to identify the sources of errors. RAMP (Reduce and Map Provenance) was developed for this purpose as an extension to Hadoop for MapReduce based workflows. [13] A downside of collecting contextual information in the context of big data is that it adds great overhead computation. This contextual information itself may also become too voluminous to be processed effectively. A challenge of research in this regard is to strike the right balance between overhead cost and contextual information to establish the veracity of a data set.

Beyond collecting contextual information, research has looked into improving the pre-processing of the data using smoothing filters, wavelet transforms or even auto-encoders to remove noise and outliers from the data. [2] Further research into improved imputation techniques for big data however is needed to deal with the incompleteness of data sets. The downside of these pre-processing modifications is again the additional computational overhead it adds. Tackling the veracity dimension of big data is arguably the least explored research field compared to the other ”V” dimensions considered in this paper.

IV. Deep Learning

Deep learning is field of machine learning methods, somewhat inspired by biological nervous systems. It automatically extracts features from the data by attempting to model high level abstractions of data in a hierarchical, layered architecture of non-linear transformations. Deep learning models can be trained in supervised or unsupervised manner. They thus can perform all sorts of machine learning tasks from classification to clustering. Deep learning models are extensively used in image recognition, natural language processing, text processing, fraud detection and more.

The idea of neural networks, fundamental to deep learning, dates back to the 1940s, however, it really took off in 2006 when Hinton et al. presented a two-stage training strategy consisting in an unsupervised ”pre-training” followed by a supervised ”fine-tuning” of the model. [7] Since then, deep learning has become one of the most prominent and promising research topics in the field of machine learning. Deep learning has in particular proven to be a promising paradigm to tackle the challenges of big data. In the following subsections some deep learning approaches tackling each of the ”V” dimensions of big data will be presented.

A. Volume

The number of parameters to train in deep learning is usually counted in millions, sometimes even in billions. To process huge amounts of data, some large scale deep learning models have been developed. Parallel deep learning models have been researched such as Deep Stacking Networks. [9] These models are faster to process and allow for training in distributed computing environments. Distbelief is the most notable of these environments optimized for large scale distributed deep network processing. [5]

Another approach to processing these large scale deep learning models has been to compress the models. Low rank factorization has been used to compress the parameter matrices of a deep learning model for speech recognition. [15] A technique called the Hashing Trick has also been used to compress large scale deep learning models. This technique consists in gathering the network connections sharing the same weights into a same hash group. Doing so, only a fraction of the parameter information needs to be retained in memory thus significantly speeding up the training process. [21] Last but not least, scaling up vertically has been most arguably the most effective solution to processing large scale deep learning models. Graphical Processing Units with programming interfaces such as NVIDIA's CUDA have extensively been used to speed up the training process of deep learning models. Custom commodity high performance computer systems for the specific purpose of training deep learning models now exist as off-the-shelf products. More recently, there have been some attempts at FPGA based implementations of large scale deep learning processing. [12]

B. Variety

Two interesting approaches to dealing with heterogeneous data in deep learning are multi-model deep learning models and deep computation models. Objects considered in the context of big data like multimedia clips or web pages are multi-model. Deep learning has proven to be an effective solution in tackling these multi-model objects. Several multi-model deep learning models have been proposed for specific tasks. They all adopt the same approach, the first layers extract the features of each modality separately before merging them in a deeper layer into an abstract joint-representation to be used as input to the specific model responsible for the task of classification or recognition. A downside of this multi-model approach is that the potential correlations between the different modalities are ignored in the separate module feature extraction stage which is a loss of potential significant information. Deep computation models, on the other hand, rely on tensor-based big data representation methods that take these correlations into account. The downside of deep computation networks is that they are very complex and require great computing capacity to be processed.

C. Velocity

Unlike the for the volume and variety dimensions of big data, only limited progress has been made in dealing with the velocity dimension of big data. Even performing a task classification or recognition with an already trained deep learning model can be difficult to do in real-time or near-real-time. In image recognition for example. Specific object detection in an input image with some already trained convolutional neural networks (CNNs) can take more than 30 seconds. Specific implementations of CNNs however have been optimized to allow for real-time detection like the You Only Look Once neural network. [14] Incremental learning models have also been researched to deal with constantly arriving data. These rely either on parameter updating or structure updating to account for arriving data. [3] These techniques have proven to be effective for single hidden layer models but are yet to be applied to deeper models. A downside of the structural updating incremental model is furthermore that it introduces redundancy in the model as it continuously adds new neurons to the model as data arrives.

D. Veracity

Most deep learning models are designed for high quality data, however, some researchers have come up with approaches to adapt deep learning models to the data that may be noisy or incomplete. Denoising autoencoders have been proposed as a solution to dealing with noisy data. [19] With denoising autoencoders, a parameter is trained to reconstruct the original output from corrupted data. Imputation autoencoders have also been developed to deal with incomplete data and have proven to be efficient when stacked in deep imputation networks. [17]
Although solutions have been developed to deal with imprecise and incomplete data, no efficient solutions exist up to now to deal with redundancy, outdated data and imprecise data.

V. CONCLUSIONS

After a brief presentation of the context of big data and the concepts of machine learning this paper provided an overview of the approaches to performing machine learning with big data. As deep learning is standing out as the most prominent machine learning paradigms and one of the most promising to deal with big data, this study focused in its last section on providing an overview of how deep learning can tackle each "V" dimension of big data.

Big data defines itself by the challenges it poses. Big data goes beyond traditional processing and analytics and this does not exclude the field of machine learning. Hopefully, this study will provide the reader with a good understanding of the specific challenges to performing machine learning with big data and a proper overview of the approaches, solutions and research going on to tackle them.

REFERENCES

[1] *10 Key Marketing Trends for 2017*. 2017. URL: https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=WRL12345USEN.

[2] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. "Non-Local Means Denoising". In: *Image Processing On Line* 1 (2011). DOI: 10.5201/ipol.2011.bcm_nlm.

[3] Hown-Wen Chen and Von-Wun Soo. "An adaptive back-propagation learning method: A preliminary study for incremental neural networks". In: *[Proceedings 1992] IJCNN International Joint Conference on Neural Networks* (). DOI: 10.1109/ijcnn.1992.287103.

[4] Jeffrey Dean and Sanjay Ghemawat. "MapReduce: Simplified Data Processing on Large Clusters". In: *OSDI'04: Sixth Symposium on Operating System Design and Implementation*. San Francisco, CA, 2004, pp. 137–150.

[5] Jeffrey Dean et al. "Large Scale Distributed Deep Networks". In: *NIPS*. 2012.

[6] Paul E.utgoff. "Incremental Learning". In: *Encyclopedia of Machine Learning and Data Mining* (2014), 1–5. DOI: 10.1007/978-1-4899-7502-7_130-1.

[7] G. E. Hinton. "Reducing the Dimensionality of Data with Neural Networks". In: *Science* 313.5786 (2006), 504–507. DOI: 10.1126/science.1127647.

[8] G. Hughes. "On the mean accuracy of statistical pattern recognizers". In: *IEEE Transactions on Information Theory* 14.1 (1968), 55–63. DOI: 10.1109/tit.1968.1054102.

[9] Brian Hutchinson, Li Deng, and Dong Yu. "Tensor Deep Stacking Networks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8 (2013), 1944–1957. DOI: 10.1109/tpami.2012.268.

[10] H. V. Jagadish et al. "Big data and its technical challenges". In: *Communications of the ACM* 57.7 (2014), 86–94. DOI: 10.1145/2611567.

[11] M. Ali-Ud-Din Khan, Muhammad Fahim Uddin, and Navarun Gupta. "Seven Vs of Big Data understanding Big Data to extract value". In: *Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education* (2014). DOI: 10.1109/aseezone1.2014.6820689.

[12] Yufei Ma et al. "End-to-end scalable FPGA accelerator for deep residual networks". In: *2017 IEEE International Symposium on Circuits and Systems (ISCAS)* (2017). DOI: 10.1109/iscas.2017.8050344.

[13] Hyunjung Park, Robert Ikeda, and Jennifer Widom. "RAMP: A System for Capturing and Tracing Provenance in MapReduce Workflows". In: *37th International Conference on Very Large Data Bases (VLDB)*. Stanford InfoLab, 2011. URL: http://ilpubs.stanford.edu:8090/995/.

[14] Joseph Redmon et al. "You Only Look Once: Unified, Real-Time Object Detection". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016). DOI: 10.1109/cvpr.2016.91.

[15] Tara N. Sainath et al. “Low-rank matrix factorization for Deep Neural Network training with high-dimensional output targets”. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (2013). DOI: 10.1109/icassp.2013.6638949.

[16] Shai Shalev-Shwartz et al. “Pegasos: primal estimated sub-gradient solver for SVM”. In: *Mathematical Programming* 127.1 (2011), pp. 3–30. ISSN: 1436-4646. DOI: 10.1007/s10107-010-0420-4. URL: https://doi.org/10.1007/s10107-010-0420-4.

[17] H Shen and E Zhang. “Incomplete big data imputation algorithm using optimized possibilistic c-means and deep learning”. In: *Informatics, Networking and Intelligent Computing* (2015), 43–48. DOI: 10.1201/b18413-10.

[18] *The Promises and Limitations of Big Data*. 2017. URL: https://hbswk.hbs.edu/item/the-promises-and-limitations-of-big-data.

[19] Pascal Vincent et al. “Extracting and composing robust features with denoising autoencoders”. In: *Proceedings of the 25th international conference on Machine learning - ICML 08* (2008). DOI: 10.1145/1390156.1390294.

[20] Wenjian Wang. “An Incremental Learning Strategy for Support Vector Regression”. In: *Neural Processing Letters* 21.3 (2005), 175–188. DOI: 10.1007/s11063-004-5714-1.

[21] Zhiyong Zeng et al. “Deep hashing using an extreme learning machine with convolutional networks”. In: *Communications in Information and Systems* 17.3 (2017), 133–146. DOI: 10.4310/cis.2017.v17.n3.a1.