"A study on Deep learning applications and challenges in big data"

Vikas

Research Scholar, Sri Satya Sai University of Technology and Medical Sciences, Sehore

Abstract

Large Data Analytics and Deep Learning are two high-focal point of information science. Enormous Data has become significant the same number of associations both open and private have been gathering monstrous measures of area explicit data, which can contain helpful data about issues, for example, national knowledge, digital security, misrepresentation location, showcasing, and clinical informatics. Organizations, for example, Google and Microsoft are investigating enormous volumes of information for business examination and choices, affecting existing and future innovation. Profound Learning calculations separate elevated level, complex reflections as information portrayals through a progressive learning process.

Complex reflections are found out at a given level dependent on moderately less complex deliberations detailed in the first level in the progressive system. A key advantage of Deep Learning is the investigation and learning of huge measures of unaided information, making it an important device for Big Data Analytics where crude information is to a great extent unlabelled and un-sorted. In the current examination, we investigate how Deep Learning can be used for tending to a few significant issues in Big Data Analytics, including removing complex examples from enormous volumes of information, semantic ordering, information labelling, quick data recovery, what's more, improving discriminative undertakings. We likewise research a few parts of Deep Learning research that need further investigation to join explicit difficulties presented by Large Data Analytics, including gushing information, high-dimensional information, versatility of models, and dispersed processing. We finish up by introducing bits of knowledge into applicable future works by suggesting some conversation starters, including characterizing information examining models, area adjustment displaying, characterizing models for getting helpful information deliberations, improving semantic ordering, semi-directed learning, and dynamic learning.

Keywords *Deep learning; Big data.*

Introduction

Deep Learning algorithms are one promising avenue of research into the automated extraction of complex data representations (features) at high levels of abstraction. Such algorithms develop a layered, hierarchical architecture of learning and representing data, where higher-level (more abstract) features are defined in terms of lower-level (less abstract) features. The hierarchical learning architecture of Deep Learning algorithms is motivated by artificial intelligence emulating the deep, layered learning process of the primary sensorial areas of the neocortex in the human brain, which automatically extracts features and abstractions from the underlying data. Deep Learning algorithms are quite beneficial when dealing with learning from large amounts of unsupervised data, and typically learn data representations in a greedy layer-wise fashion. Empirical studies have demonstrated that data representations obtained from stacking up nonlinear feature extractors (as in Deep Learning) often yield better machine learning results, e.g., improved classification modelling, better quality of generated samples by generative probabilistic models, and the invariant property of data representations. Deep Learning solutions have yielded outstanding results in different machine learning applications, including speech recognition, computer vision, and natural language processing. A more detailed overview of Deep Learning is presented in Section "Deep learning in data mining and machine learning" Big Data represents the general realm of problems and techniques used for application domains that collect and maintain massive volumes of raw data for domain-specific data analysis. Modern data-intensive technologies as well as increased computational and data storage resources have contributed heavily to the development of Big Data science. Technology based companies such as Google, Yahoo, Microsoft, and Amazon have collected and maintained data that is measured in exabyte proportions or larger. Moreover, social media organizations such as Facebook, YouTube, and Twitter have billions of users that constantly generate a very large quantity of data. Various organizations have invested in developing products using Big Data Analytics to addressing their monitoring, experimentation, data analysis, simulations, and other knowledge and business needs, making it a central topic in data science research.

Deep learning in data mining and machine learning

The main concept in deep leaning algorithms is automating the extraction of representations (abstractions) from the data. Deep learning algorithms use a huge amount of unsupervised data to automatically extract complex representation. These algorithms are largely motivated by the field of artificial intelligence, which has the general goal of emulating the human brain's ability to observe, analyze, learn, and make decisions, especially for extremely complex problems. Work pertaining to these complex challenges has been a key motivation behind Deep Learning algorithms which strive to emulate the hierarchical learning approach of the human brain. Models based on shallow learning architectures such as decision trees, support vector machines, and case-based reasoning may fall short when attempting to extract useful information from complex structures and relationships in the input corpus. In contrast, Deep Learning architectures have the capability to generalize in non-local and global ways, generating learning patterns and relationships beyond immediate neighbors in the data. Deep learning is in fact an important step toward artificial intelligence. It not only provides complex representations of data which are suitable for AI tasks but also makes the machines independent of human knowledge which is the ultimate goal of AI. It extracts representations directly from unsupervised data without human interference. A key concept underlying Deep Learning methods is distributed representations of the data, in which a large number of possible configurations of the abstract features of the input data are feasible, allowing for a compact representation of each sample and leading to a richer generalization. The number of possible configurations is exponentially related to the number of extracted abstract features. Noting that the observed data was generated through interactions of several known/unknown factors, and thus when a data pattern is obtained through some configurations of learnt factors, additional (unseen) data patterns can likely be described through new configurations of the learnt factors and patterns. Compared to learning based on local generalizations, the number of patterns that can be obtained using a distributed representation scales quickly with the number of learnt factors. Deep learning algorithms lead to abstract representations because more abstract representations are often constructed based on less abstract ones. An important advantage of more abstract representations is that they can be invariant to the local changes in the input data. Learning such invariant features is an ongoing major goal in pattern recognition (for example learning features that are invariant to the face orientation in a face recognition task). Beyond being invariant such representations can also disentangle the factors of variation in data. The real data used in AI-related tasks mostly arise from complicated interactions of many sources. For example an image is composed of different sources of variations such a light, object shapes, and object materials. The abstract representations provided by deep learning algorithms can separate the different sources of variations in data. Deep learning algorithms are actually Deep architectures of consecutive layers. Each layer applies a nonlinear transformation on its input and provides a representation in its output. The objective is to learn a complicated and abstract representation of the data in a hierarchical manner by passing the data through multiple transformation layers. The sensory data (for example pixels in an image) is fed to the first layer. Consequently the output of each layer is provided as input to its next layer.

Big data analytics

Big Data generally refers to data that exceeds the typical storage, processing, and computing capacity of conventional databases and data analysis techniques. As a resource, Big Data requires tools and methods that can be applied to analyse and extract patterns from large-scale data. The rise of Big Data has been caused by increased data storage capabilities, increased computational processing power, and availability of increased volumes of data, which give organization more data than they have computing resources and technologies to process. In addition to the obvious great volumes of data, Big Data is also associated with other specific complexities, often referred to as the four Vs: Volume, Variety, Velocity, and Veracity. We note that the aim of this section is not to extensively cover Big Data, but present a brief overview of its key concepts and challenges while keeping in mind that the use of Deep Learning in Big Data Analytics is the focus of this paper.

The unmanageable large Volume of data poses an immediate challenge to conventional computing environments and requires scalable storage and a distributed strategy to data querying and analysis. However, this large Volume of data is also a major positive feature of Big Data. Many companies, such as Facebook, Yahoo, Google, already have large amounts of data and have recently begun tapping into its benefits. A general theme in Big Data systems is that the raw data is increasingly diverse and complex, consisting of largely un-categorized/unsupervised data along with perhaps a small quantity of categorized/supervised data. Working with the Variety among different data representations in a given repository poses unique challenges with Big Data, which requires Big Data preprocessing of unstructured data in order to extract structured/ordered representations of the data for human and/or downstream consumption. In today's data-intensive technology era, data Velocity – the increasing rate at which data is collected and obtained – is just as important as the Volume and Variety characteristics of Big Data. While the possibility of data loss exists with streaming data if it is generally not immediately processed and analyzed, there is the option to save fast-moving data into bulk storage for batch processing at a later time. However, the practical importance of dealing with Velocity associated with Big Data is the quickness of the feedback loop, that is, process of translating data input into useable information. This is especially important in the case of time-sensitive information processing. Some companies such as Twitter, Yahoo, and IBM have developed products that address the analysis of streaming data. Veracity in Big Data deals with the trustworthiness or usefulness of results obtained from data analysis, and brings to light the old adage "Garbage-In-Garbage-Out" for decision making based on Big Data Analytics. As the number of data sources and types increases, sustaining trust in Big Data Analytics presents a practical challenge. Big Data Analytics faces a number of challenges beyond those implied by the four Vs. While not meant to be an exhaustive list, some key problem areas include: data quality and validation, data cleansing, feature engineering, high-dimensionality and data reduction, data representations and distributed data sources, data sampling, scalability of algorithms, data visualization, parallel and distributed data processing, real-time analysis and decision making, crowdsourcing and semantic input for improved data analysis, tracing and analyzing data provenance, data discovery and integration, parallel and distributed computing, exploratory data analysis and interpretation, integrating heterogenous data, and developing new models for massive data computation.

Applications of deep learning in big data

As stated previously, Deep Learning algorithms extract meaningful abstract representations of the raw data through the use of an hierarchical multi-level learning approach, where in a higher-level more abstract and complex representations are learnt based on the less abstract concepts and representations in the lower level(s) of the learning hierarchy. While Deep Learning can be applied to learn from labeled data if it is available in sufficiently large amounts, it is primarily attractive for learning from large amounts of unlabeled/unsupervised data, making it attractive for extracting meaningful representations and patterns from Big Data. Once the hierarchical data abstractions are learnt from unsupervised data with Deep Learning, more conventional discriminative models can be trained with the aid of relatively fewer supervised/labeled data points, where the labeled data is typically obtained through human/expert input. Deep Learning algorithms are shown to perform better at extracting non-local and global relationships and patterns in the data, compared to relatively shallow learning architectures. Other useful characteristics of the learnt abstract representations by Deep Learning include: (1) relatively simple linear models can work effectively with the knowledge obtained from the more complex and more abstract data representations, (2) increased automation of data representation extraction from unsupervised data enables its broad application to different data types, such as image, textural, audio, etc., and (3) relational and semantic knowledge can be obtained at the higher levels of abstraction and representation of the raw data. While there are other useful aspects of Deep Learning based representations of data, the specific characteristics mentioned above are particularly important for Big Data Analytics. Considering each of the four Vs of Big Data characteristics, i.e., Volume, Variety, Velocity, and Veracity, Deep Learning algorithms and architectures are more aptly suited to address issues related to Volume and Variety of Big Data Analytics. Deep Learning inherently exploits the availability of massive amounts of data, i.e. Volume in Big Data, where algorithms with shallow learning hierarchies fail to explore and understand the higher complexities of data patterns. Moreover, since Deep Learning deals with data abstraction and representations, it is quite likely suited for analyzing raw data presented in different formats and/or from different sources, i.e. Variety in Big Data, and may minimize need for input from human experts to extract features from every new data type observed in Big Data. While presenting different challenges for more conventional data analysis approaches, Big Data Analytics presents an important opportunity for developing novel algorithms and models to address specific issues related to Big Data. Deep Learning concepts provide one such solution venue for data analytics experts and practitioners. For example, the extracted representations by Deep Learning can be considered as a practical source of knowledge for decision-making, semantic indexing, information retrieval, and for other purposes in Big Data Analytics, and in addition, simple linear modeling techniques can be considered for Big Data Analytics when complex data is represented in higher forms of abstraction. In the remainder of this section, we summarize some important works that have been performed in the field of Deep Learning algorithms and architectures, including semantic indexing, discriminative tasks, and data tagging. Our focus is that by presenting these works in Deep Learning, experts can observe the novel applicability of Deep Learning techniques in Big Data Analytics, particularly since some of the application domains in the works presented involve large scale data. Deep Learning algorithms are applicable to different kinds of input data; however, in this section we focus on its application on image, textual, and audio data.

Deep learning challenges in big data

The prior section focused on emphasizing the applicability and benefits of Deep Learning algorithms for Big Data Analytics. However, certain characteristics associated with Big Data pose challenges for modifying and adapting Deep Learning to address those issues. This section presents some areas of Big Data where Deep Learning needs further exploration, specifically, learning with streaming data, dealing with high-dimensional data, scalability of models, and distributed computing. Incremental learning for non-stationary data One of the challenging aspects in Big Data Analytics is dealing with streaming and fast-moving input data. Such data analysis is useful in monitoring tasks, such as fraud detection. It is important to adapt Deep Learning to handle streaming data, as there is a need for algorithms that can deal with large amounts of continuous input data. In this section, we discuss some works associated with Deep Learning and streaming data, including incremental feature learning and extraction [49], denoising autoencoders [50], and deep belief networks [51]. Zhou et al. [49] describe how a Deep Learning algorithm can be used for incremental feature learning on very large datasets, employing denoising autoencoders [50]. Denoising autoencoders are a variant of autoencoders which extract features from corrupted input, where the extracted features are robust to noisy data and good for classification purposes. Deep Learning algorithms in general use hidden layers to contribute towards the extraction of features or data representations. In a denoising autoencoder, there is one hidden layer which extracts features, with the number of nodes in this hidden layer initially being the same as the number of features that would be extracted. Incrementally, the samples that do not conform to the given objective function (for example, their classification error is more than a threshold, or their reconstruction error is high) are collected and are used for adding new nodes to the hidden layer, with these new nodes being initialized based on those samples. Subsequently, incoming new data samples are used to jointly retrain all the features. This incremental feature learning and mapping can improve the discriminative or generative objective function; however, monotonically adding features can lead to having a lot of redundant features and overfitting of data. Consequently, similar features are merged to produce a more compact set of features. Zhou et al. [49] demonstrate that the incremental feature learning method quickly converges to the optimal number of features in a large-scale online setting. This kind of incremental feature extraction is useful in applications where the distribution of data changes with respect to time in massive online data streams. Incremental feature learning and extraction can be generalized for other Deep Learning algorithms, such as RBM [7], and makes it possible to adapt to new incoming stream of an online large-scale data. Moreover, it avoids expensive cross-validation analysis in selecting the number of features in large-scale datasets

High-dimensional data

Some Deep Learning algorithms can become prohibitively computationally-expensive when dealing with high-dimensional data, such as images, likely due to the often slow learning process associated with a deep layered hierarchy of learning data abstractions and representations from a lower-level layer to a higher-level layer. That is to say, these Deep Learning algorithms can be stymied when working with Big Data that exhibits large Volume, one of the four Vs associated with Big Data Analytics. A high-dimensional data source contributes heavily to the volume of the raw data, in addition to complicating learning from the data. Chen et al. [52] introduce marginalized stacked denoising auto encoders (mSDAs) which scale effectively for high-dimensional data and is computationally faster than regular stacked demonising auto encoders (SDAs). Their approach marginalizes noise in SDA training and thus does not require stochastic gradient descent or other optimization algorithms to learn parameters. The marginalized denoising auto encoder layers to have hidden nodes, thus allowing a closed-form solution with substantial speed-ups. Moreover, marginalized SDA only has two free meta-parameters, controlling the amount of noise as well as the number of layers to be stacked, which greatly simplifies the model selection process. The fast training time, the capability to scale to large-scale and high dimensional data, and implementation simplicity make mSDA a promising method with appeal to a large audience in data mining and machine learning.

Conclusion

In contrast to more conventional machine learning and feature engineering algorithms, Deep Learning has an advantage of potentially providing a solution to address the data analysis and learning problems found in massive volumes of input data. More specifically, it aids in automatically extracting complex data representations from large volumes of unsupervised data. This makes it a valuable tool for Big Data Analytics, which involves data analysis from very large collections of raw data that is generally unsupervised and un-categorized. The hierarchical learning and extraction of different levels of complex, data abstractions in Deep Learning provides a certain degree of simplification for Big Data Analytics tasks, especially for analyzing massive volumes of data, semantic indexing, data tagging, information retrieval, and discriminative tasks such a classification and prediction. In the context of discussing key works in the literature and providing our insights on those specific topics, this study focused on two important areas related to Deep Learning and Big Data: (1) the application of Deep Learning algorithms and architectures for Big Data Analytics, and (2) how certain characteristics and issues of Big Data Analytics pose unique challenges towards adapting Deep Learning algorithms for those problems. A targeted survey of important literature in Deep Learning research and application to different domains is presented in the paper as a means to identify how Deep Learning can be used for different purposes in Big Data Analytics. The low-maturity of the Deep Learning field warrants extensive further research. In particular, more work is necessary on how we can adapt Deep Learning algorithms for problems associated with Big Data, including high dimensionality, streaming data analysis, scalability of Deep Learning models, improved formulation of data abstractions, distributed computing, semantic indexing, data tagging, information retrieval, criteria for extracting good data representations, and domain adaptation. Future works should focus on addressing one or more of these problems often seen in Big Data, thus contributing to the Deep Learning and Big Data Analytics research corpus.

References

Domingos P (2012) A few useful things to know about machine learning. Commun ACM 55(10)

2. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference On. IEEE Vol. 1. pp 886–893

3. Lowe DG (1999) Object recognition from local scale-invariant features. In: Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference On. IEEE Computer Society Vol. 2. pp 1150–1157

4. Bengio Y, LeCun Y (2007) Scaling learning algorithms towards, AI. In: Bottou L, Chapelle O, DeCoste D, Weston J (eds). Large Scale Kernel Machines. MIT Press, Cambridge, MA Vol. 34. pp 321–360. http://www.iro.umontreal.ca/~ lisa/pointeurs/bengio+lecun_chapter2007.pdf

5. Bengio Y, Courville A, Vincent P (2013) Representation learning: A review and new perspectives. Pattern Analysis and Machine Intelligence, IEEE Transactions on 35(8):1798–1828. doi:10.1109/TPAMI.2013.50

6. Arel I, Rose DC, Karnowski TP (2010) Deep machine learning-a new frontier in artificial intelligence research [research frontier]. IEEE Comput Intell 5:13–18

7. Hinton GE, Osindero S, Teh Y-W (2006) A fast learning algorithm for deep belief nets. Neural Comput 18(7):1527–1554

8. Bengio Y, Lamblin P, Popovici D, Larochelle H (2007) Greedy layer-wise training of deep networks, Vol. 19

9. Larochelle H, Bengio Y, Louradour J, Lamblin P (2009) Exploring strategies for training deep neural networks. J Mach Learn Res 10:1–40

10. Salakhutdinov R, Hinton GE (2009) Deep Boltzmann machines. In: International Conference on, Artificial Intelligence and Statistics. JMLR.org. pp 448–455.