JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

UNLEASHING THE FACTS AND MYTHS ABOUT DEEP LEARNING

[1]Mr.Vishwas Victor , [2]Dr.Ragini Shukla
[1]Research Scholar, Dr. C. V. Raman University, Bilaspur,
[2]Professor, Dr. C. V.Raman University, Bilaspur, India

Abstract: Nowadays, Deep Learning is being popular among the computer scientists and Researchers. It is a stimulating field of Machine Learning. It is true to say that Deep Learning is the most effective, supervised, time as well as cost efficient machine learning approach. Deep Learning follows various procedures and topographies that can be applied to a group of complicated problems instead of restricted learning approach. In Deep Learning, multiple layers are used to represent the abstractions of data, just to build computational models. There are few Deep Learning Algorithms such as Generative Adversarial Network, Convolutional Neural Networks and Model Transfers which have changed our perception of information processing completely. Though, there exists an opening of understanding behind this tremendously fast paced domain. The only reason behind this is that it was never represented from a multi-scope perspective previously. Deep Learning methods have made a significant breakthrough with appreciable performance in a wide variety of applications with useful security tools. This led to a wide use of Deep Learning domains in Business, Science and Government. It is further used in adaptive testing, biological image classification, computer vision, cancer detection natural language processing, object detection, face recognition handwriting recognition speech recognition stock market analysis and many more. In this research paper, we focused on the evolution of Deep Learning, its comparison with Machine Learning, different Deep Learning concepts and applications. Also, we are going to address the facts and myths about it. Finally, the paper ends with the conclusion and future aspects.

Introduction: In last few years, Machine Learning has become more and more popular in research. It is used in a large number of applications, including multimedia concept retrieval, image classification, video recommendation, social network analysis, text mining etc. Among various Machine Learning Algorithms, Deep Learning is widely used in these applications. It may be interesting to know that Deep Learning is also known as “Representation Learning”. The explosive growth and convenience of knowledge and also the outstanding advancement in hardware technologies have led to the emergence of recent studies in distributed and deep learning. Deep Learning, that has its roots from typical neural networks, considerably outperforms its predecessors. It utilizes, graph technologies with transformations among neurons to develop many-layered learning models. Many of the newest Deep Learning techniques are given and have incontestable promising results across totally different types of applications like Natural Language Processing (NLP), Visual Data Processing, Speech and Audio Processing, and plenty of different well-known applications.

Traditionally, the efficiency of Machine Learning Algorithms highly depends upon the quality of representation of the data taken as input. A bad data representation often leads to lower performance as compared to a good data representation. That’s the reason why, feature engineering has been an important research direction in Machine Learning for a long time, which focuses on building features from raw data and has led top lots of research studies. Even, feature engineering is often very domain specific and requires and requires significant human efforts. For example, in Computer Vision, various typesof features have been proposed and compared such as Histogram of Oriented Gradients (HOG), Scale Invariant Feature Transform (SIFT), and Bag of Words (BoW). Once a new feature is proposed with high efficiency, it becomes a trend until another feature is developed and is found to be more efficient than that of previous one. Similar situations took place in other domains also such as Speech Recognition and Natural Language Processing (NLP).

Comparatively, Deep Learning Algorithms perform feature extraction in an automated way. This helps researchers to extract discriminative features with the requirement of minimal domain knowledge and human effort. In these algorithms, there exists layered architecture of data representation, where the high-level features can be extracted from the last layer of

networks while the low-level features are extracted from the lower layers. These architectures were originally influenced by Artificial Intelligence (AI) simulating its process of the key sensorial areas in the human brain. It may be interesting to know that our brain can automatically extract data representation from different scenes. The input is the scene information received from eyes, whereas the output is the classified object. This highlights the major advantage of using Deep Learning Algorithms as "it mimics how our brain works".

Deep Learning now became one of the hottest research directions in the Machine Learning Society just because of its success in various fields. This Research Paper will give an overview of Deep Learning from different perspectives, including history, challenges, few Deep Learning Algorithms and facts along with myths about Deep Learning.

Evolution Of Deep Learning: It had been a dream of sages from centuries to develop a machine that can simulate Human Brains. The concept of Deep Learning came into existence in 300 B.C. when a Greek Philosopher, Aristotle proposed 'Associationism'. There are various definitions of Associationism which are given by different Scholars. As per Wikipedia, it refers to the idea that mental processes operate by the association of one mental state with its successor states.Associationism started the history of humans' ambition in trying to understand the brain, since such an idea requires the scientists to understand the mechanism of human recognition systems. If we talk about the modern history of Deep Learning, it started in 1943 when McCulloch Pitts (MCP) model was introduced and became popular as the prototype of Artificial Neural Models. They developed a computer model based on the neural networks functionally mimicking neocortex in Human Brains. Algorithms and Mathematics are combined together to form 'Threshold Logic'. This was used in their model to mimic the Human Thought Process but not to learn. Since then, Deep Learning evolved steadily.

After MCP Model, Hebbian Theory was implemented which was originally used for the biological systems in the natural environment. After implementation, 'Perceptron' was developed. It was the first electronic device within the context of the cognition system and was introduced in 1958. It is different from 'perceptron' which we are using currently. The emergence of 'back propagandist' became another Milestone at the end of first AI winter. Back Propagation was introduced by Werbos. According to him, the use of errors in coaching Deep Learning models which opened the gate to modern neural network. 'Neocogitron' was introduced in 1980 which was the base of convolutional neural network. Afterwards, Recurrent Neural Networks was introduced in 1986 whereas due to LeNet, Deep Learning (DNNs) started working practically in 1990s. Though it was not so popular and highly recognised. LeNet was quite native and cannot be applied to large datasets, just because of its hardware limitation. It was 2006, when Deep Belief Network (DBNs) and layer-wise frameworks were developed. It trains a simple two layer unsupervised model just like Restricted Boltzmann Machine (RBMs), freeze all the parameters, stick a new layer on top and train the parameters for the new layers. Due to Deep Learning, Researchers were able to train neural networks that were much deeper than the previous attempts. Originally from Artificial Neural Networks (ANNs) and after decades of development, Deep Learning now is one of the most important and efficient tools as compared to other machine learning algorithms with great performance.

Graphics Processing Unit (GPUs) are very much popular and the reason behind it is their performance in computing large - scale matrices in network architecture on a single machine, a number of distributed Deep Learning frameworks have been developed to speed up the training of deep learning models. Some research studies focus more on improving noise robustness of training modules using unsupervised or semi-supervised Deep Learning techniques and the reason behind it is simple, the vast amount of data come without labels or with noisy labels. Researchers are now paying more attention to a cross-modality structure, which may yield a huge step forward in Deep Learning. In present days, Google Alpha Go is an inspirational application of Deep Learning. It completely shocked the world at the beginning of year 2017. Alpha Go is able to defeat world champion Go players because it uses the modern Deep Learning Algorithms and sufficient hardware resources.

There are two key factors on which Deep Learning method is based on. These are as follows:

a. Non Linear processing in multiple layers or stages
b. Supervised or Unsupervised learning

Characteristics of Deep Learning:

There are several characteristics few of which are as follows

1. Extensively powerful tools which is used in many fields
2. It is purely based on neural networks along with more than two layers and so called deep
3. Have strong learning ability
4. Can make use of datasets more effectively
5. Learn features extraction methods from data
6. Surpass human ability to solve highly computational tasks
7. Very little engineering is required
8. Optimized results
9. Deep Learning Networks depend upon the nature of the network architecture, activation function and data representation
10. Solve highly computational tasks

S. No.	Statement	Myths / Reality	Clarification
1.	Deep Learning is a very recent technology	Myth	The first application based on Deep Learning was developed in 1952, when Arthur Samuel wrote a program which was capable enough to play checkers. Deep Learning requires a considerable computing capacity to achieve useful results in very less time, a power that has only be available in recent years.
2.	Data is used as fuel for Deep Learning	Reality	One of the important reasons why automatic learning has gained relevance in last few years is the availability of a massive amount of data (called Big Data) along with the greater computing power. Companies that are indulged in developing Deep Learning Solutions put their focus especially on importance of the quality of pre-selection of the data with which the Machine Learning Systems are 'fed'.
3.	Deep Learning only needs a large amount of data	Myth	Big Data is essential for Deep Learning Systems but the quality of the data supplied to it will be even more important. If the previous data 'pre-selection' is of poor quality an automatic learning systems will give unsatisfactory results.
4.	Deep Learning can be more 'Intelligent' than human beings	Myth	It is correct that the power of Deep Learning to find correlations can exceed as compared to that of human beings but it doesn't mean that Deep Learning can draw intelligent conclusions.
5.	Deep Learning is like Human Learning	Myth	Currently, we are not in a position to say that Deep Learning is like Human Learning, the only reason behind it is – we still don't know how the human brain works.
6.	Deep Learning is one of the technologies of the future	Reality	In recent years, the potential of Deep Learning has been demonstrated with increasingly promising results. Huge number of Companies are already using Deep Learning Solutions and the boom is expected to be even greater in the coming years.

There are many more myths about Deep Learning such as:

- Deep Learning requires expert knowledge
- Deep Learning is time consuming
- Deep Learning is very expensive
- Deep Learning has poor interpretability
- Deep Learning needs GPUs etc.
- Deep Learning Networks already have an understanding of the world similar to humans

Deep Learning Techniques

Different Deep Learning Algorithms not only helps in improving the learning performance and broaden the scopes of applications but also simplifies the calculation process. However, there's a major problem that many Researchers face and that is – the extremely long training time which is required for understanding Deep Learning Models. Increasing the size of training data and model parameters the classification accuracy can be drastically enhanced. We have to apply several enhanced Deep Learning Techniques that are proposed in literature to accelerate the Deep Learning processing. In Deep Learning Frameworks there exists implementation of modularized Deep Learning Algorithms, optimization techniques, distribution techniques and infrastructures support. The main intention developing Deep Learning Frameworks is to

simplify the implementation process and boost the system level Research & Development. In this section, we have introduced few of the representative techniques.

- **Unsupervised and Transfer Learning:** Very few studies have addressed the unsupervised learning problem in deep learning as compared to that of supervised learning in which a vast amount of work is done. Though, in last few years, the benefit of learning reusable features using unsupervised techniques has shown promising results in different applications. For unsupervised techniques, generative models such as GANs and VAEs have become dominant technique in recent years. GANs stands for Generative Adversarial Network whereas VAEs denotes Variational Auto-encoders. A GAN is a class of Machine Learning frameworks which was efficiently designed by Lan Goodfellow and his colleagues. The core idea of GAN is based on the 'Indirect' training through the discriminator, which itself is also being updated dynamically. From this statement, we can conclude that the generator is not trained to minimize the distance to a specific image, but rather to tool the discriminator. Though GAN was originally proposed in the form of generative model for unsupervised / unspecialized leaning. VAE are based on latent variable model. Instead of trying to develop a latent space (space of latent variables) explicitly and to sample from it in order to find samples that could actually generate proper outputs, we construct a Encoder-Decoder like network. This is the key idea behind VAE.

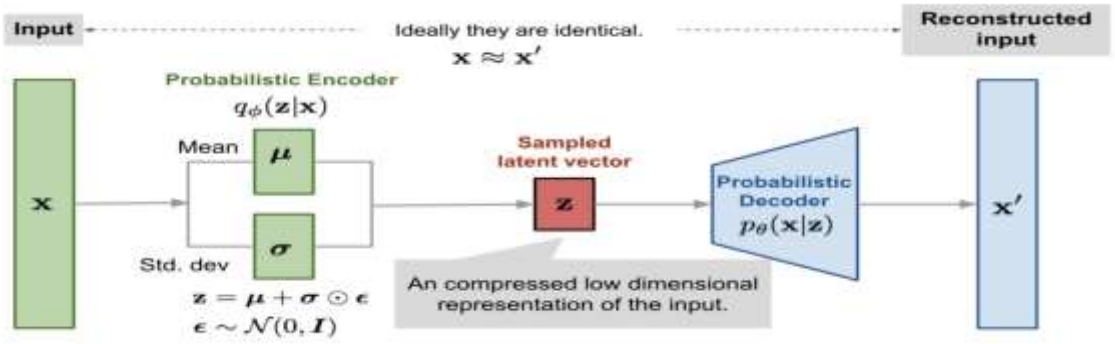

Figure 1: Global Architecture of Variational Auto Encoder (VAE)

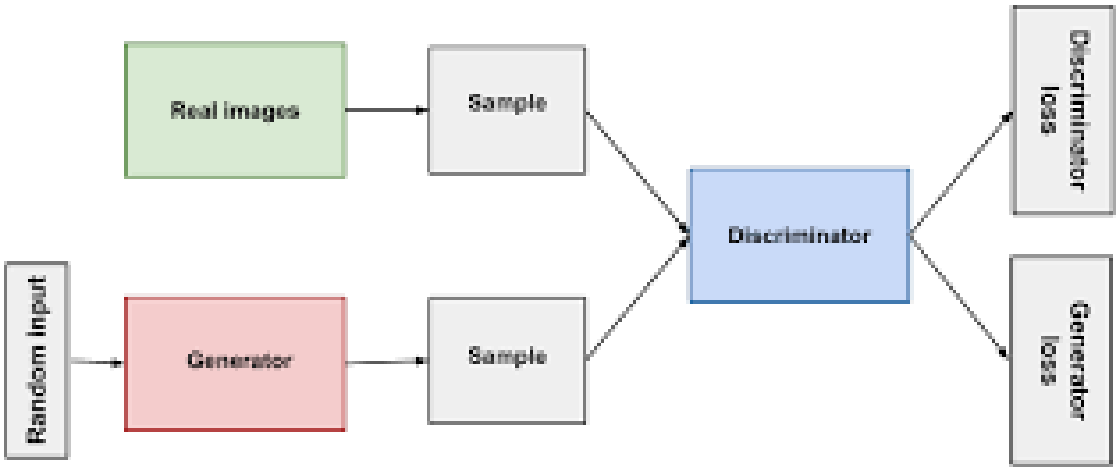

Figure 2: Overview of Generative Adversarial Networks (GAN)

GANs are based on Convolutional Neural Networks (i.e. CNNs). It has shown its supremacy as Unsupervised Learning in visual data analysis. It's a fact that the number of people is very less who have the luxury to access very high speed GPUs and using powerful hardware to train a very deep network from scratch in a reasonable period of time. Hence pertaining a Deep Network on a large scale datasets is very common. This technique is also known as Transfer Learning. It can be done by using the pertained network as fixed feature extractors or fine tuning the weights of the pertained model.

- **Online Learning:** Generally, the network topologies and architectures used in Deep Learning are time static and also time inherent. In another words, these are predefined before the learning starts. When the data is streamed online this restriction on time complexity leads to a serious challenge. Previously, online learning came into mainstream research but if we talk about development, a modest advancement is observer in online Deep Learning. Deep Neural Networks are built upon the Stochastic Gradient Descent (SGD) approach in which the training samples are used individually to update the model parameters with a known label. The need is that the updates should be applied as batch processing rather than sequential processing of each sample. Another challenge that is being faced on the issue of online learning is high-velocity data along with time varying distributions. It represents the retail and banking data pipelines that holds tremendous business values. In current situation, the data is largely close in time to safely assuming piecewise stationarity thus having a similar distribution. This assumption characterizes data with a certain degree of correlation and develops the model accordingly. Unfortunately, these non-stationary data streams are not IID and are often longitudinal data streams.

- **Optimization Techniques in Deep Learning:** To train a DNN is a optimization process i.e. to find the parameters in the network that minimized the loss function. Practically, the SGD method is a fundamental algorithm applied to deep learning that adjusts the parameters iteratively based on the gradient for each training sample. The computational complexity of the original Gradient Descent Method is higher / more than that of SGD. In the original gradient descent method, the whole dataset is considered every time the parameter is updated.

  The updating speed is controlled by hyper-parameter learning rate in the learning process. Lower learning rates will eventually lead to an optimal state after a long time. On the other hand, higher learning rates decay the loss faster but it's quite possible that fluctuations may occur during the training. The idea of using momentum is introduced in order to control the oscillation of SGD. This technique gets a faster convergence and a proper momentum that can improve the optimization result of SGD as it is inspired by Newton's first law of motion. On the other hand, to determine the proper learning rate, several techniques are proposed. To adjust the learning rate and accelerate the convergence, weight decay and learning rate are introduced primitively. A weight decay works as a penalty coefficient in the cost function to avoid overfitting whereas a learning rate decay can reduce the learning rate dynamically to improve the performance. Also, to avoid the fluctuation, adapting the learning rate with respect to the gradient of previous stages is found helpful. The first successful adaptive algorithm that is used in deep learning is Adagrad. It amplifies the learning rate for infrequently updated parameters and supresses the learning rate for the frequently updated parameters by recording the accumulated squared gradients. The learning rate of Adagrad can become extremely small and does not optimize the model anymore just because the squared gradients are always positive. Adadelta is introduced to solve this issue where a decay fraction β_2 is introduced to limit the accumulation of the squared gradients just as follows:

  $\mathbf{E[g^2]_t = \beta_2 E[g^2]_{t-1} + (1- \beta_2)(g_t)^2}$

  Where, $\mathbf{E[g^2]_t}$ is the accumulated squared gradient at stage t and $\mathbf{(g_t)^2}$ is the squared gradient at stage t. Ananother decay fraction $\boldsymbol{\beta_1}$ is introduced to record the accumulation of the gradients just to improve Adadelta furthermore. It is shown that Adam performs better as compared to other algorithms with an adaptive learning rate.
- **Deep Learning in Distributed Systems:** Model training's efficiency is limited to a single-machine system and the distributed deep learning techniques have been developed to further accelerate the training process. Data parallelism and model parallelism are the two main approaches to train the model in a distributed system. The model is replicated to all the computational nodes and each model is trained with the data assigned subset of data to achieve data parallelism. The weight update needs to be synchronized among the nodes after a certain period of time in case of Data Parallelism. Comparatively, all the data is processed with one model where each node is responsible for the partial estimation of the parameters in the model to achieve Model Parallelism.

  The most straight forward algorithm to combine results from the slave nodes is parameter averaging among data-parallel approaches. Let $\mathbf{W_{t,i}}$ be the parameter in the neural network at node i at time t with N slave nodes used for training. At time t, the weight on the master node is $\mathbf{W_t}$. Thenafter, a copy of the current parameters is distributed to the slave node. Once the updated parameters are sent back to the master node, the weight at time t + 1 on the master node will be

$$W_{t+1} = (1/N) \sum_{i=1}^{N} \mathbf{W_{t+1, i}}$$

  If parameters are averaged after each mini batch or each worker processes the same number of data copies, parameters averaging would be identical to single machine training. However, the network communication and synchronization costs can nullify the benefits of extra machines. This is the reason why, the averaging process is usually applied after a certain number of mini batches, fed to each slave node. The frequency of training and the model performance need to be balanced as required. Update-based data parallelism is a more popular approach for data parallelism that uses SGD where the updates of the learning rate decay and momentum are transferred. However, the synchronous weight update is not scalable for a larger cluster. The overhead of communication increases exponentially with respect to the number of nodes. That's why, a parameter server framework is proposed by Google to process the training asynchronously. The asynchronous update allows each node to spend more time on computation instead of waiting for the parameter to be updated on the master node. Meanwhile, the network communication cost can be reduced significantly, by de-centralization i.e. transmitting the updates in peer-to-peer mode instead of master-slave node.

  On the other hand, a model parallelism approach splits the training step across multiple GPUs. Each GPUs computes only a subset of the model in a straightforward modal-parallel strategy. For instance, the system with two GPUs can use each of them to calculate one LSTM Layer for a model with two LSTM layers. Model-parallel strategy makes training and prediction with massive deep neural networks possible and this is the biggest advantage of it. There is one drawback associated with model parallelism which is – "each node can only compute a subset of results and synchronization is thus needed to get the full results." The synchronization loss and communication overhead of model-parallel strategies are more than those of data-parallel strategies since each node in the former must synchronize gradients and parameter values both on every update step.

  Challenges with Deep Learning: There is no doubt that Deep Learning Techniques are efficient and having very good performance in solving various complicated applications with multiple layers and high level of abstraction. It's not wrong to say that the accuracy, acuteness, receptiveness and precision of Deep Learning Systems are almost or may surpass human experts. In today's scenario, the Deep Learning Technology has to accept any challenges to feel the exhilaration of victory. Hence following is the list of challenges that Deep Learning has to overcome.
- Deep Learning Algorithms are supposed to manage the input data continuously.
- Deep Learning Algorithms must have to ensure transparency of the conclusion.
- High performance GPUs and storage requirements are required by the resources.
- Improved methods for Big Data Analytics.

- There exists complex designs and hyper parameters.
- In the processing / implementation of Deep Learning algorithms, very high computation power is required.
- It requires a tremendous amount of data.
- Deep Learning Systems are expensive for the complex problems and computation.
- It suffers from local minima.
- Deep Learning Systems are computationally intractable.
- It has no strong theoretical foundation.
- It is found to be difficult to find the topology, training parameters for deep learning.
- It provides new tools for infrastructures for the computation of the data and enables computer to learn objects and representations.

Conclusions: Deep Learning is a popular and fast growing application of machine learning. The rapid use of Deep Learning Algorithms in various fields really showcase its success and versatility. To build Deep Learning Systems in near future, high performance, computing infrastructure-based systems together with theoretically sound parallel learning algorithms and novel architectures are needed. Currently, continuous growth is taking place in computer memory and computational power through parallel or distributed computing environments. Further Research will focus on finding solution for the issues related to computation and communication. After few years, solutions for addressing the scalability, reliability, adaptability of the unsupervised learning models will take the central stage.

References:

[1] Samira Pouyanfar, SaadSadiq and Yilin Yan, HaimanTian,Yudong Tao, "A Survey on Deep Learning: Algorithms, Techniques, and Applications".
[2] V. Pream Sudha1, R. Kowsalya2 (Department of Computer Science)" A SURVEY ON DEEP LEARNING TECHNIQUES, APPLICATIONS and CHALLENGES ", PSGR Krishnammal College for Women, India.
[3] Shaveta Dargan · Munish Kumar · MaruthiRohitAyyagari · GulshanKumar ,"A Survey of Deep Learning andIts Applications: A New Paradigm to Machine Learning", Archives of Computational Methods in Engineering.
[4] Weibo Liua, ZidongWanga.∗, XiaohuiLiua, NianyinZengb, YurongLiuc,d and Fuad E. Alsaadid, "A Survey of Deep Neural Network Architectures and Their Applications"
[5] MD. Zakir Hossain, FerdousSohel, Mohd Fairuz, Shiratuddin, Hamid Laga, "A Comprehensive Survey of Deep Learning for Image Captioning".
[6] J.Pamina, J.Beschi Raja, "SURVEY ON DEEP LEARNING ALGORITHMS"International Journal of Emerging Technology and Innovative Engineering.