International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Review Article

A Review on Neural Network and Deep Learning

Eng. Nawaf Mohammad H Alamri*

Mechanics, Materials and Advanced Manufacturing, Cardiff University, Queen's Buildings, 14-17 The Parade, Cardiff, CF24 3AA, United Kingdom

Received 13 July 2020, Accepted 15 Sept 2020, Available online 16 Sept 2020, **Vol.10, No.5 (Sept/Oct 2020)**

Abstract

The idea of Neural Network (NN) derived from brain activities which consists of neurons connected with thousands of their neighbors forming a very complex local network. As a discipline of artificial intelligence, neural network attempts to bring the computers closer to brain capability by simulating certain aspects of brain information system so that the NN has the ability to learn and generalize with higher processing speed. As an extension to NN, Deep Learning (DL) can handle high dimensional data with between leaning capability as feature extraction process became automatic. The aim of this paper is to present a review about NN and DL showing a detailed explanation of different architectures and algorithms.

Keywords: *Artificial Intelligence, Machine Learning, Neural Network, Deep Leaning, Activation Function.*

1. Introduction

(Packianather, 2018) stated that the idea of Neural Network (NN) derived from brain activities which consists of neurons connected with thousands of their neighbours forming a very complex local network. As a discipline of artificial intelligence, neural network attempts to bring the computers closer to brain capability by simulating certain aspects of brain information system so that the NN has the ability to learn and generalize with higher processing speed than human brain as 100Hz is the processing speed for human brain and 109 Hz is the processing for the computer network.

(De Filippis *et al*., 2017) said that it can be used for process improvement, monitoring, controlling and optimization by extracting meaningful pattern and relationship between the parameters in order to predict the future state of the process. As an extension to NN, Deep Learning (DL) can handle high dimensional data with between leaning capability as feature extraction process became automatic.

The aim of this paper is to present a review about NN and DL showing a detailed explanation of different architectures and algorithms.

2. Neural Network

This section will present a detailed explanation of NN definition, input types, learning types, algorithms and activation functions.

2.1 Definition

(Packianather, 2018) stated that NN is a computing model derived from the brain activities to bring the computer to be closer to it, but with higher capabilities. It consists of the interaction of large number of processing units to mimic human brain.

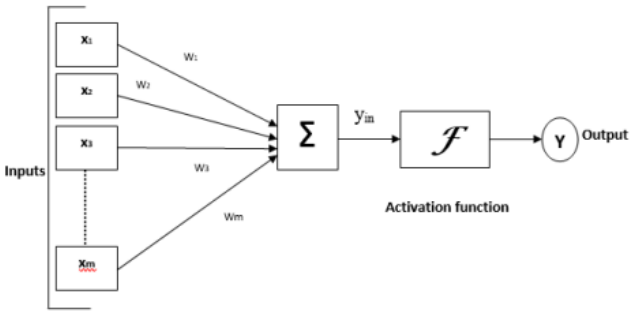

Fig. 1 The general model of artificial neural network (tutorialspoint, 2019)

The general model shown in figure 1 consists of set of inputs which represent local information (concepts, features, letter, words and patterns), a state of activation with different ranges for different models, an output function for each element which depends on the activation in a probabilistic manner, a connectivity pattern which can be represented by weight matrix, a propagation rule, an activation rule for each element which dictates how inputs are combined to produce a new activation, a learning rule which is changing the knowledge and originates from Hebb's law which says that if one neuron stimulates another one when the

*Corresponding author's ORCID ID: 0000-0002-5641-0178
DOI: https://doi.org/10.14741/ijcet/v.10.5.7

receiving neuron is firing, the strength of the connection between the two cells is strengthened. Finally, the last element is the environment in which the system operates.

2.2 Input Types

There are two types of inputs, discrete input as the input will be either in the binary form of 0s or 1s or non-binary including -1, 0, 1, 2, 3, ……., 9. The second type is continuous valued input as the input will be in the form of continuous values range. It could be between 0 and 1, between -1 to 1 or between 0 to 100. A bigger range result in less reliable training as mentioned in (richardbowles, 2019)

2.3 Learning Types

(Packianather, 2018) mentioned that there are three learning types, the first one is supervised learning which learns from data in input-output patterns. The output is a teacher to supervise the process of learning. The second type is unsupervised learning without output in the data patterns. The aim of this type of learning is to analyse and discover patterns in the inputs without supervision of teacher. The third one is reinforcement learning defined as reinforcing an action by following it with satisfactory state of affairs. Otherwise, the system will not produce that action.

2.4 Network Architectures and Algorithms

There are different network architectures and algorithms as shown below:

Feedforward Network: it is a nonrecurrent network where the information flow moves in one direction from left to right. (Dr. Packianather, 2018) and (towardsdatascience, 2019) stated different types:

- Single Layer Feedforward Network: it has only one weighted layer which means that input layer is connected directly to the output layer. Its applications are limited in the real life.
  - Gradient Descent Algorithm: it is an algorithm used to train single layer perceptron.
- Multilayer Feedforward Network: it has more than one weighted layer between the input and output which called hidden layers. Its applications are speech recognition, character recognition, human face recognition and machine translation.
  - Backpropagation Algorithm: it is an algorithm used to train multi layers perceptron and use delta rule. The adjustment is controlled by two parameters, the first one is learning rate which is the step size in the error and the other one is momentum which is a proportion of the weight change. The values of both parameters are between 0 and 1, the training become better as their values 6 increase. However, choosing a large value could lead to oscillations in the output.
- Extreme Learning Machine: it is used to create spares hidden layers with random to reduce the complexity of feedforward network result in less computational power.
- Auto Encoder Network: it can be trained without supervision. Output cells is equal to input cells and hidden cells is less than input cells. Its applications are classification, clustering and feature compression.
  - Variational Auto Encoder Network: it is used for probability compression instead of feature compression.
  - Denoising Auto Encoder Network: it is used to vary the data by random bit and reconstruct output from a bit input making it more general helps to select more common features.
  - Sparse Auto Encoder Network: the number of hidden cells is bigger than input cells.

Recurrent Network: it is feedback network where the information flow moves in both directions using closed loop. It can generate novel sentences and document summaries, in addition to speech and connected handwriting recognition. (tutorialspoint, 2019), (towardsdatascience, 2019) and (medium, 2019) stated different types:

- Fully Recurrent Network: all nodes are connected to all other nodes, so the node works both as input and output.
  - Hopefield Network: it consists of single layer having one or more fully recurrent neurons. It is used for optimization and auto-association. It consists of neurons with one inverting and one non-inverting output. Input and output can be either binary between 0 and 1 or bipolar between -1 and +1.
  - Brain-State-in-a-Box Network: it is fully connected recurrent network having the maximum number of nodes that depend on the dimensions number of the input space. The weights are updated simultaneously, and neurons take values between -1 to +1.
- Boltzmann Machine: it is a stochastic learning process having recurrent structure. It consists of stochastic neuron which have two possible states either 0 or 1. 7
  - Restricted Boltzmann Machine: it belongs to energy-based models which aims to minimize the predefined energy function. It consists of one input and hidden layers and there is no output layer. Its applications are filtering and prediction.
- Jordan Network: the output will go to the input again.
- Long Short-Term Memory: it is used for modelling temporal sequences along with their long-term dependencies more accurate than traditional recursive neural network. It does not use activation function and does not modify the stored values. It is

implemented in blocks with three gates called input, forget and output gate helps in information flow controlling on a logistic function. It is applied in Part-of-Speech Tagging.

- Sequence to Sequence Model: it is a model with two recurrent networks, including encoder to process the input and decoder to produce the output. It is used in chatbots, system for answering question and machine translation.
- Recursive Network: the same weights are applied recursively.
- Gated Recurrent Network: it is similar to long short-term memory without output gate. It is commonly used in sound and speech synthesis.
- Liquid State Machin: it is not fully connected network where the activation functions are replaced by threshold level. It is commonly used in speech recognition.
- Echo State Network: it has special training since only weights between hidden cells are updated. There is no application yet in the real life, only theoretical benchmarks.
- Neural Turing Machine: it is an abstraction over long short-term memory. The memory is addressed by the contents, so the network can read from it and write to it as well.

Genetic Algorithm: it is used to simulate the natural selection and genetics mechanics where organisms mate and evolve in their development. A binary string is used to represent a chromosome consisting of a sequence of genes as mentioned in (Packianather, 2018).

Support Vector Machine Algorithm: it is used for data points classification based on a hyperplane in multiple features dimensional space. The objective is to choose a plane with the maximum distance between two classes. Its applications are handwriting recognition, recognizing the fraud in credit cards, speaker identifications and face detection. As stated in (towardsdatascience, 2019).

Bayes Algorithm: it is a classification technique used to calculate the conditional probability. It works well with categorical input variable compared to numerical variables. Its applications are prediction, text classification, spam faltering smart recommendations system as stated in (analyticsvidhya, 2019)

Marcov Chain: (towardsdatascience, 2019) mentioned that it is used to predict the likelihood of each possible next states knowing the current states. It is commonly used in speech recognition, classification based on probability and clustering.

The Kohonen Self-organising Map: (Packianather, 2018) stated that it is used for forming a miniature model of a much larger input space and preserving the topological order in the data set (Rectangular Grid Topology and Hexagonal Grid Topology). It has a clustering capability

Learning Vector Quantization: it is used to quantize input vectors into reference values to use them for pattern classification. It has full connection between the input and hidden layers and partial connection between the hidden and output layers with fix weight of 1 as mentioned in (Packianather, 2018).

Generative Adversarial Network It is used to lean a map from an input image to an output image. It is commonly used to generate images as mentioned in (towardsdatascience, 2019).

Radial Base Function: (Packianather, 2018) said that it is feed forward network with one hidden layer. It can be used for, classification, function approximation, modelling of time series and dynamic systems.

2.5 Activation Functions Types

(towardsdatascience, 2019) stated different activation functions types:

1. Linear Function (Identity Functions): it is used for non-classification problems such as regression in prediction problem. F(x) = x.
2. Sigmoid or Logistic Function: it is slow in convergence and used with classification problems. It has two types:
   - Binary Sigmoidal Function: it converts input between 0 and 1. F(x) = 1 / 1+e-x •
   - Bipolar Sigmoidal Function: it converts input between -1 and 1. F(x) = (2 / 1+e-x) - 1.
3. Hyperbolic Tangent Function (Tanh): its output is zero centered since the output fall in the range between -1 and 1 and used with optimization problems. F(x) = 1- e -2x / 1+e-2x.
4. Rectified Linear Units Function (ReLu): it has better convergence than Tanh by 6 times, but is limited to be used within Hidden layers. F(x) = max (0, x) i.e if x < 0, F(x) = 0 and if x >= 0, F(x) = x A special case of it is leaky rectified linear units (leaky ReLu) to avoid zero derivatives where F(x) = 0.01x if x < 0.
5. Softmax Function: it is used with output layer to compute the probability of the classes in classification problems.

3. Deep Learning

(Singh *et al*., 2018) stated that deep learning is a part of machine learning techniques that utilizes different processing layers where the previous layer output is used as an input for the following layer to learn data representations with more than one level of abstraction. He explained the difference between machine learning and deep learning that the process of feature extraction depends on data and model types in case of using traditional approach which leads to time consuming in using trial and error and the success depend on the user experience. Following deep learning approach will give an advantage of automatic feature extraction by learning large number of nonlinear filters before making decisions. So, deep learning combines the process of extracting features and making decisions in one model and avoids the suboptimal manual handcrafting. In addition, they said that the training of

deep learning models can be done in supervised and unsupervised ways. In the first way, the input data are mapped to the output so the learning from data will be with target labels while in the other way there is no target label in the learning process.

The authors also mentioned that selecting hyperparameters in the training process may determines a successful deep learning model. The important parameters are network structure which is number of layers and number of units in each one, learning rate, momentum and choice of activation functions. However, there is a potential problem may happen during training which is overfitting, it is a case when a model learns excessive details about training data that leads to poor ability of generalization about unused data in that model. Using a large dataset may reduce the problem, so it is necessary to augment the dataset in case of a large dataset is not available. In addition to traditional approach, there are novel techniques for solving such a problem called dropout and batch normalization.

Furthermore, (Najafabadi *et al*., 2015) discussed deep learning in data mining as the main concept in deep learning algorithms is automated extraction of representation from a huge amount of dataset with a large motivation of artificial intelligence field. The extracted features from the set of data are distributed allowing for a large number of possible configurations which leads to better generalization. The relation between the number of possible configuration and the number of extracted features is exponential, generation of the observed data was based on the interaction of different factors so that obtaining a pattern through some configurations can help to obtain 6 additional patterns of unseen data through new configurations. However, there are certain challenges associated with big data can inhibit adopting deep learning. The first aspect is dealing with non-stationary data. It is important to adapt deep leaning models to deal with fast moving and streaming input data. The second challenge is dealing with high dimensional data such as images that contribute heavily to the volume of the data which may lead to slow learning process. In addition, it is important to overcome the challenge of scaling the recent successful deep learning models to a large scale with a massive of set of data as the empirical results demonstrated the effectiveness of a large scale models with a very large number of parameters that have the ability to extract very complex representations and patterns. Finally, exploring the volume of training input data is necessary in order to get good representation about the used data which can then be generalized for new one in specific domain. It is important to consider the shift between training and generalizing the representations for input and target data sources. Furthermore, the focus on domain adaptation during learning process is essential in deep learning, the distribution of training data for learning representation is different from test data for deploying the learnt representation. Another key area is to explore defining criteria for allowing the extracted patterns to give beneficial semantic meaning about the set of data. Following active learning methods could be helpful for obtaining improved data patterns where the input from the experts can be used to get labels for samples to better improve and tune the learnt representation.

3.1 Convolutional Neural Network

It is a deep learning network that has many applications, mainly it is applied image recognition cases. (Singh *et al*., 2018) showed major architectures for this type of application which are AlexNet, GoogLeNet, ZFNet, RESNet and VGGNet. Among these networks, the most common one is AlexNet and it is defined by (MathWorks) as a pretrained convolutional neural network on more than millions of images taken from image database called ImageNet. It consists of 8 deep layers and has the ability to classify images in 1000 object categories resulting in a model that learns rich features representation for a wide range of images.

(Buda *et al*., 2018) investigated the performance of classification in convolutional neural networks by showing the impact of class imbalance on it which is a common problem in classical machine learning and address this issue by comparing frequently used methods. The authors used three different benchmark datasets MNIST, 7 CIFAR-10 and ImageNet to make the investigation and then perform an extensive comparison between different methods which are under-sampling, oversampling, two-phase training and thresholding that will be used for compensating previous class probabilities. The results of the experiments revealed that the effect of class imbalance on the classification performance is detrimental and the method of addressing it in most of the scenarios was oversampling that should be applied in a way that ensure imbalance elimination, whereas the optimal ratio for under-sampling is dependent on the extent of imbalance. The resulted oversampling did not cause overfitting of convolutional neural network opposed to classical machine learning models. Finally, the thresholding used for compensating previous class probabilities when overall number of cases classified properly is of interest.

3.2 Deep Recurrent Network

(Singh *et al*., 2018) said that it is a deep learning network used to model sequential data such as time series data, but it is limited to model this type of data with short-range order or memory. In other cases, deep long short-term memory architectures are used to build a deep learning model suitable for such time series data.

(Rahman *et al*., 2018) used deep recurrent neural network to make prediction in medium to long term of electricity consumption in residential and commercial buildings responsible for most of the overall energy consumption at resolution of one hour. This will help in

decision making related to operations, strategies for demand response and distributed generation systems installation. The drawbacks of the models are missing information about schedules and equipment of the buildings and having energy consumption missing data making which makes time series predictions difficult. The authors developed and optimized deep recurrent neural network models in order to make prediction of electricity consumption in medium to long term and then they used different types of patterns for electricity consumption to analyse the relative performance of the models. They performed amputation of electricity consumption dataset that contain missing data. The proposed recurrent neural network sequence to sequence models relatively have lower error when compared to convolutional neural network with multi-layered perceptron, but both models are similar in terms of the accuracy of predicting electricity consumption.

3.3 Long Short-term Memory

Deep neural network can deal with complex problems and achieve excellent performance from the models making it powerful machine learning tool. However, (Sutskever *et al*., 2014) mentioned that there is a significant limitation that it can be applied only to cases whose inputs and targets that sensibly encoded with vectors having fixed dimensionality. It is considered as an important disadvantage because there are many problems are best formulated with sequences of previously unknow length like speech recognition, machine translation and question answering.

The authors showed an application of sequence to sequence learning problem using long short-term memory architecture. The idea is to use one for reading the input sequence, one as timestep for each time for obtaining vector representation of large fixed dimensionality and then to use another one for extracting the sequence of the output from the obtained vector. The second used long short-term memory is important to be a recurrent neural network except that it is conditioned to the sequence of input. This type pf network is considered as a suitable choice for this application due to the fact that there is a lag time between the inputs and its corresponding outputs. The model is applied to machine translation tasks for short and long sentences without having any difficulties. In addition, the paper showed that reversing the sequence of the words in all source sentences (not including the target sentences) improved the overall performance of the long short-term memory considerably because it introduces many dependencies in the short term between the target and source sentences making the optimization problem much easier.

3.4 Unsupervised Deep Neural Networks

They are part of deep neural networks which used when a large amount of data is not available and there is a need for extracting features from set of data without specific target variable. (Singh *et al*., 2018) stated an example about this type of network which is deep belief network which is built layer by layer using an old popular unsupervised model called restricted Boltzmann machines which is defined by (Shen *et al*., 2017) as a single layer undirected graphical model consisting of one visible layer and one hidden layer with symmetric connectivity between them without connection between the units in the same layer. Having this advantage helps to generate input observations from the hidden layers as well. The parameters of model are trained using contrastive divergence algorithm. However, restricted Boltzmann machines are stacked 9 to build a deep architecture which result in a single stochastic model named as deep belief network which consists of one visible layer and many series hidden layers.

In addition, (Singh *et al*., 2018) also mentioned that the other most popular unsupervised deep network is deep autoencoder that used in painting, denoising and hashing cases. (Shen *et al*,2017) said that it is a special type of two-layer neural network which reduces the reconstruction error between the values of the input and output leading to learning the representation of the input in latent or compressed way. It is a simple network and its structure is shallow in a single layer autoencoder which has very limited applications, but when stacked autoencoder configuration is used, multiple autoencoders are stacked leading to improve the representational power significantly by using the activation values of the hidden units of an autoencoder as input to the next higher one. An important advantage of stacked autoencoders is the ability to discover or learns highly complicated or nonlinear patters such as the interaction and correlation between different inputs. For each input, they are different layers of the network representing different levels of information, as the layer is lower the patterns are simple, but they become more complicated when the layers are higher.

(Mehdiyev *et al*., 2017) proposed an approach for multi-stage deep learning to address the classification problem for time series data. The study focuses on production processes predictive monitoring on the basis of product quality enabled by Internet of Things (IoT) networks. As the nature of the problem is complicated, this reduce the effectiveness of handcraft feature extraction and increase the need for implementing models that have the ability to extract patterns from unlabelled data. The first stage of the proposed approach is to adopt stacked long short-term memory autoencoders in order to extract representations and patterns from time series dataset on the basis of unsupervised model. After completing all stages of data processing including zero padding, the second stage of the modelling starts by applying deep feedforward neural networks to perform the classification for the set of data.

(Erhan *et al*., 2010) stated that pre-training unsupervised deep neural network can result in

regularization effect that establish initial point for fine tuning procedure in small volume parameters space region. This pre-training procedure leads to increase the weight magnitude in unsupervised deep neural networks with sigmoidal nonlinear function. However, unsupervised pretraining help to make a restriction to the network parameters to be in a particular region.

Conclusion

Neural Network can be used for process improvement, monitoring, controlling and optimization by extracting meaningful pattern and relationship between the parameters in order to predict the future state of the process. As an extension to NN, DL can handle high dimensional data with between leaning capability as feature extraction process became automatic. The paper presented a review about NN and DL which showed a detailed explanation of different architectures and algorithms.

Acknowledgement

I would like to express my deepest gratitude to my parents and my wife for giving me the composure and the strength to complete this work. I would like to offer my thanks and genuine appreciation to Dr. Michael Packianather for guidance, invaluable advice and unlimited support to accomplish this work.

References

Dr. Packianather (2018), Lecture notes of EN4902/ENT633 course in school of engineering at cardiff university.

De Filippis, L. A. C., Serio, L. M., Facchini, F., & Mummolo, G. (2017). ANN Modelling to Optimize Manufacturing Process. In Advanced Applications for Artificial Neural Networks. IntechOpen.

Artificial Neural Network - Building Blocks. [online] Available at https://www.tutorialspoint.com/artificial_neural_network /artificial_neural_networ k_building_blocks.htm [Accessed: 26 February 2019].

A Neural Network Family Tree. [online] Available at http://www.richardbowles.co.uk/resources/neural/neural7.html [Accessed: 26 February 2019]

The mostly complete chart of Neural Networks, explained. [online] Available at https://towardsdatascience.com/the-mostly-complete-chart-of-neural-networksexplained-3fb6f2367464 [Accessed: 5 March 2019]

7 types of Artificial Neural Networks for Natural Language. [online] Available at Processinghttps://medium.com/@datamonsters/artificial-neural-networks-fornatural-language-processing-part-1-64ca9ebfa3b2 [Accessed: 18 February 2019].

6 Easy Steps to Learn Naive Bayes Algorithm. [online] Available at https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/ [Accessed: 2 March 2019]

Singh, A. K., Ganapathysubramanian, B., Sarkar, S., & Singh, A. (2018). Deep learning for plant stress phenotyping: trends and future perspectives. Trends in plant science.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. Journal of Big Data, 2(1), 1.

AlexNet. [online] Available at https://uk.mathworks.com/help/deeplearning/ref/alexnet.html [Accessed: 3 July 2019].

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks, 106, 249-259.

Rahman, A., Srikumar, V., & Smith, A. D. (2018). Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks. Applied energy, 212, 372-385.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Advances in neural information processing systems (pp. 3104-3112).

Shen, D., Wu, G., & Suk, H. I. (2017). Deep learning in medical image analysis. Annual review of biomedical engineering, 19, 221-248. 49

Mehdiyev, N., Lahann, J., Emrich, A., Enke, D., Fettke, P., & Loos, P. (2017). Time series classification using deep learning for process planning: a case from the process industry. Procedia Computer Science, 114, 242- 249.

Erhan, D., Bengio, Y., Courville, A., Manzagol, P. A., Vincent, P., & Bengio, S. (2010). Why does unsupervised pre-training help deep learning?. Journal of Machine Learning Research, 11(Feb), 625-660.