Survey on Deep learning for Cyber security

Hamad Aslam[1], Noor Fathima[2], Arjun B C[3]
[1]dept. Of Information Science And Engineering, [2]dept. Of Computer Science and Engineering
[3]asst. Professor & Hod, Dept. Of Information Science and Engineering
Rajeev Institute of Technology, Hassan – 573201

Abstract- Security is the most challenging aspect in the internet and network applications. To ensure Cyber security of an organization, we can leverage NN-DL to act upon the system exploit, threats or breaches like intrusion, malware, spam and phishing that are taking place in the digital world. DL is derived from ML and ML is a sub version of artificial intelligence. Since DL is more advance and has higher forecasting value and the prediction can be improved continuously. The amount of data is enormous (big data) we can use that data to train our deep learning models. This paper gives the survey of various Cyber security issues which are solved using DL.

Keywords- Cyber security, Deep learning (DL), Machine Learning, Neural network, Big data, intrusion, malware, spam and phishing, Neural Networks and Deep learning (NN-DL).

I. INTRODUCTION

Artificial intelligence (AI) is similar to human intelligence which can be displayed by machines, especially computer systems. An example for this is ML. ML is broadly classified into two types: supervised (learning with labels) and unsupervised (self learning).
Labeled data is identification of any given input to a system by humans. Labeling commonly consider a set of unlabeled data and picks each individual piece and gives meaning full tags to make it more informative. For example, labels might show whether a photo contains a dog or a cat. Labels can be produced by humans by judging the given piece of unlabeled data and are significantly more expensive to obtain than the raw unlabeled data. Supervised machine learning requires labels in the data. Labels which can be expected outputs that allows Supervised machine learning by example. Deep learning is a form of supervised machine learning which needs improved education and assets. It was recognized in 2012.

Unsupervised machine learning are self learning and doesn't depend upon labels from the data. It is the only learning algorithm which identifies abnormal indication of deep threats. Using limited data sets of unsupervised ML can easily discover patterns typically without labels. Different DL techniques are very effective to various threats in different cases.

Cyber security is the collection of policies, processes, technologies & techniques that work together to secure the confidentiality, integrity and availability of computing resources, networks, software programs and data from attack. Cyber defence mechanisms exist at the application, network, host and data level. There are plenty of tools such as firewall, antivirus software, intrusion detection systems (IDSs), and intrusion protection systems (IPS), that works in silence to prevent attacks and detect security breaches. However, many adversaries are still at an advantage because they only need to find one vulnerability in systems needing protection. The increase in number of internet-connected systems will increase the attack surface, leading to greater risk of attacks. Furthermore, attackers are becoming more sophisticated developing zero-day exploits and malware that evade security measures, enabling them persist for one periods without notice. Zero-days exploits are attacks that have not been encountered previously but are often variations on a known attack.

To worsen the problem, attack mechanisms are being advantageous, allowing for rapid distribution without needing an understanding for developing exploits. In addition to defending against external threats, defenders must also guard against insider threats for individuals or entities within an organization that misuses their authorized access. There are massive quantity of data from applications, servers, smart devices and other Cyber-enabled resources generated by machine-to-machine and human-to-machine interactions. Cyber defense system are getting voluminous data, such as the Security Information Event management (SIEM) system, which often overwhelms the security analyst with event alerts. The user of data science in Cyber security can help to co-relate events, identify patterns, and detect abnormal behavior to improve the security posture of any defenceprogramme. We are starting to see an emergence of Cyber defence system leveraging data analytics. For instance, network intrusion detection systems (NIDSs) that inspect packets transmission are evolving from signature-based systems that detected well-known attacks anomaly-based systems that detect deviation from a “normal” Behavior profile.

This paper is deliberate for researchers who wish to use deep learning concepts for providing Cyber security. It is identified that DL may give significantly varieties of embracing techniques. Related surveys have also described that machine learning applications to Cyber problems but does not include DL methods.

II. NEURAL NETWORK

The Artificial Neural Network (ANN) is inspired from the biological neural network. Using this concept from the biological neural network, an ANN, also simply known as the neural network is formed. In the following, each neuron accomplishes two actions. One is gathering input from different neurons or it can gather inputs in a weighted form and second is adding up all input, weights from the neuronbased on the summed value, it calls an activation function.

A group of neuron layers can be used to create a neural network. The network architecture varies based on the objective it needs to achieve. In the development of unsupervised learning throughself-organized learning and supervised learning through the creation of perceptron (type of ANN) respectively. They are composed of few layers of neurons connected by adaptive weight (Figure.1), and the adjacent network layers are usually fully connected. The universal approximation theorem for ANNs states that using multi-layer perceptron with just one hidden layer, it is possible to map intervals of real numbers to some output and can be appropriated arbitrarily. This means that one hidden layer of ANN is capable of producing a non-linear continuous function and as such. Much of the early research on ANNs concentrated on networks with just single hidden layer, trained using back-propagation.

However, the output can be used as features for another model. The model is trained by taking binary input data and feeding it towards the model. Then it feeds backward through the model to reconstruct the input data. The energy of the system is then calculated and used to update the weights. This process is continued until the model converges.

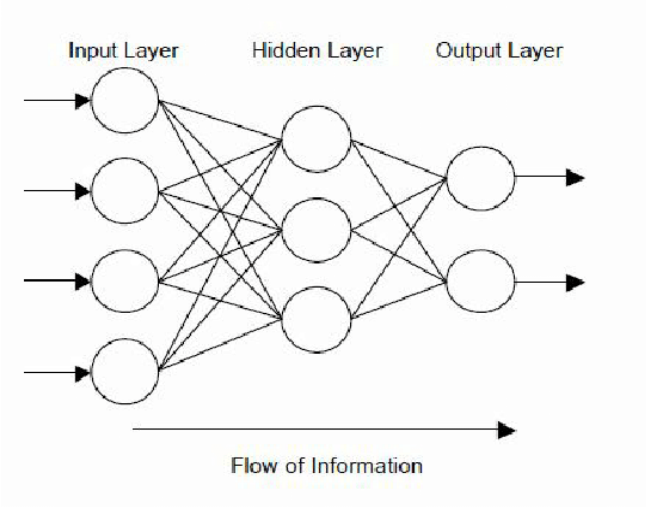

Fig.1: Artificial neural network.

III. DEEP LEARNING

DL is a branch of ML works on a group of algorithms which try to create high levelextraction ondata. DL is one of the family member from different ML techniques which is used in data representation, as against to task-specific algorithm. Deep neural networks and recurrent neural networks which have architectures of deep learning have been applied to fields like speech recognition, natural language processing, social network filtering, audio recognition, image recognition, machine translation, drug design, computer vision, medical, material inspection and board game programs, where they have originated results equivalent to and in some cases superior to human experts.

As we build large NN and train models with more and more data, their performance simultaneously increases. This generally varies for other ML techniques that reach an another level in achievement. This is the reason why DL has become so trending topic today. Major parts of DL applications that we see in our daily life are usually generated towards fields like marketing, sales, finance, etc. We barely go through, read articles or find resources on DL being used to secure these products, and the business, from malware and cracker attacks. While the biggest tech-giant companies like alphabet, apple and Microsoft have already started adapting DL into their products, the Cyber security industry has not yet started using advance technologies like DL. It is a challenging field but one that needs our full attention. And has relatively high promising results.

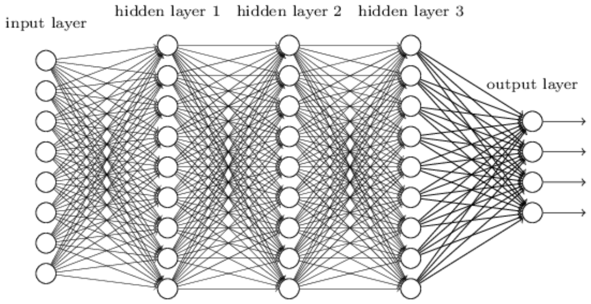

Fig.2: Deep neural network.

IV. DEEP LEARNING METHODS FOR CYBER SECURITY

A. Recurrent neural network (R-NN)

A R-NN Figure 3, increases the capabilities of a traditional neural network, which can only take fixed-length data inputs, to handle input sequence of variable lengths. The R-NN operate inputs one element at a time, using the output of the hidden units as additional inputs for next element. Therefore, the R-NN can address speech and language problems as well as time series problems.

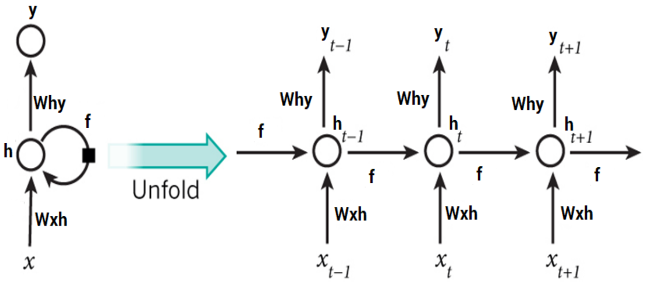

Fig.3: Recurrent neural network.

Typically, R-NNs are more tough to train because the gradients can easily disappear or explode. However, improvement in training and architecture have produced a variety of R-NNs that are easier to train. As a result, R-NNs have shown success in next-word-in-a-sentence prediction, speech recognition, image capturing, language translation and other time series prediction tasks. The invisible units of R-NN are capable of maintaining a “state vector” that hold a memory of the past events in the sequence. The length of this "memory" can be modified based on the type of R-NN note that is used. The longer the memory, the longer term the dependencies the R-NN is capable of learning.

The long short-term memory (LSTM) units have been introduced to permit R-NNs to handle problems that require long-term memories. LSTM units contain a structure called a memory cell that gather information, which connects to itself in the next time step. The values of the memory cell are augmented by new input and a forget gate that weights newer and older information higher of lower depending on what is needed.

V. CYBER APPLICATIONS OF DEEP LEARNING METHODS

Deep learning can be applicable in these following attacks.

A. Malware:

The number and variety of malware attacks are continually increasing, making it more complicated to defend against them using standard methods. DL offers an opportunity to construct generalized models to detect and classify malware autonomously. This can provide difference against small-scale actors using known malware and large-scale actors using new types of malware to attack organizations or individuals. In recent time however, deep learning techniques have shown impressive results in self learning feature representations for complex problem domains like images, speech,voice and text. We take advantage of these advances in deep learning to self learn how to detect malware without costly feature engineering.

As it turns out, deep learning architectures and in particular R-NN can do a good job of detecting malware. Over few years experimenting with deep learning architectures for malware classification, as well as methods to avoid them. Experiments have demonstrated amazing levels of accuracy that are competitive with traditional DL-based solutions, while avoiding the price of manual feature engineering. It is also evident that we can obtain results with accuracy upwards of 96%.

B. Intrusion Detection**:**

There have been relatively high researches on intrusion detection but the most changes occur in the data set collected which contains many samples of intrusion techniques such as brute force, denial of service (D.O.S) or even an infiltration from within a network. Variation in network actions and its patterns and intrusions evolve, it has very much become required to move away from static and one-time data file toward more dynamically generated data file which not only reflect the traffic formation and intrusions for that time, but are also modifiable, extensible, and reproducible. So we can train a deep learning model to identify anomaly from a given data set.

With successful identifications on our model, we can confidently proclaim that our model can be used as a back end engine to an intrusion detection system application that can be mounted on the border of any computer network. However it is wise to further test, and if need to be re-train, the model on the new data set.

C. Spam and Phishing:

While deep learning can be essential to numerous safe systems including spam and phishing detection, the road to bring in its benefits is paved with numerous safety challenges. Drawing from solid examples from fraud and abuse problems.

VI. CONCLUSION

DL can provide new approaches for addressing Cyber security problems. It has proven significant development over traditional signature-based and rule-based systems as well as classic machine-learning-based solutions. Most efforts have focused on applying DL to malware detection and network intrusion detection. R-NNs were popular method and used to address the broadest set of Cyber security threats (ie. malware malicious,domain names, network intrusions, host intrusions and Cyber physical intrusions). R-NN are likely popular because many Cyber security related tasks or the data can be treated as a time series problem. This lends itself well to R-NN.

It is difficult to draw conclusions about the performance of any particular approach because there are different data sets and different metrics however, some trends are not worthy. Performance vary greatly across different security domains. The ability to detect network intrusions was highly dependent on the type of attack and number of classes and performing attack classification. Another critical factor that impacted performance of across all domains was the ratio of benign data to malicious data in the training set. This problem arises from the fact that it is difficult to obtain legitimately malicious data. Often data is created from stipulations and reverse engineering of malware because real data can be hard to obtain.

The Cyber domain has large volumes of data from a variety of sources to which DL can be applied. However research in this space is hampered by the limited publicly available data sets that are either small, old or internally generated and not shared among researches. To develop meaningful trust in DL methods, large, regularly updated, benchmark data sets will be critical to advancing Cyber security solutions. More over the ability to test proposed DL methods in real operational scenarios is needed to compare detection rates, speed, memory usage, and other performance metrics. The Cyber security industry has just began to appreciate the value of DL and new data sets are emerging.

VII. REFERENCES

[1]. https://en.wikipedia.org/wiki/Artificial_neural_network

[2]. https://en.wikipedia.org/wiki/Deep_learning

[3]. Cryptography, Network Security and Cyber Laws, Bernard L.Menezes and Ravinder Kumar, 2018 Cengage Learning India Pvt.Ltd.

[4]. https://en.wikipedia.org/wiki/Cyberattack

[5]. AvneetPannu "Artificial Intelligence and its Application in Different Areas", 2015

[6]. Daniel S.Berman, Anna L.Buczak *, Jeffrey S.chavis and charitL.corbett "A survey of Deep Learning methods for Cyber security"

[7]. IEEE.org reference papers

[8]. https://Interset.AI

[9]. https://towardsdatascience.com/building-an-intrusion-detection-system-using-deep-learning-b9488332b312

[10]. https://www.fireeye.com/blog/threat-research/2018/12/what-are-deep-neural-networks-learning-about-malware.html