**LITERATURE REVIEW**

**INTRODUCTION:**

Neural Network Watermarking is a technology that has arisen as a viable method for ensuring the ownership of neural network models, which have become a vital asset in disciplines such as image recognition, natural language processing, and speech recognition. As these devices have become more common and accessible, the fear of unauthorized access or theft has intensified. Researchers have proposed different watermarking approaches in recent years, with the goal of embedding a unique signature into the model, which can be used to validate ownership or detect unauthorized modifications. This study provides a review of the literature on some of the most famous neural network watermarking algorithms published in recent years, as well as their benefits and drawbacks.

**LITERATURE REVIEW:**

The virus that is embedded in neurons in this paper's proposal is a novel way to transmit malware evasively and surreptitiously through a neural network model. The research emphasizes the significance of evasive and covert malware delivery in sophisticated malware campaigns and suggests the rapid substitution strategy to do so by changing neurons. The suggested method alters the entire neuron, not just the insignificant bits of a parameter, to embed malware. The study demonstrates that alterations to some neurons have no effect on the functioning of the neural network, which can withstand an antivirus engine's security check thanks to redundant neurons in the network layers. By inserting 36.9MB of malware into a 178MB-AlexNet model with 1% accuracy loss, the paper tests the viability of this approach, and antivirus engines in VirusTotal express no concerns. The report also emphasizes how using neural networks for attacks has become increasingly popular because of the extensive use of artificial intelligence. By outlining a reference scenario for the defense against neural network-aided attacks, the paper adds to the body of literature. The report emphasizes that the receiver can check the integrity of the embedded malware and extract it from the model using established procedures. The article also offers results on malware embedding and assesses how well the malware performs when it is integrated into various models and malware samples. Overall, the paper provides a new approach to delivering malware covertly and evasively through a neural network model, which can evade detection from antivirus engines.[1]

The study examines copyright threats to deep neural networks (DNNs) in an attack-defense scenario with the model owner as the victim and the adversary. Model extraction, model pruning, and model finetuning are three frequent challenges to DNN copyright. When the adversary has complete knowledge of the victim model and a limited dataset to fine-tune the model, model tuning takes place. Model pruning opponents first refine the victim model using a few data sets, then further refine the model using several pruning techniques. When the adversary can only make predictions from the victim model and wants to use the victim model's capabilities, this is known as model extraction. A duplicate of the victim model is then trained on the annotated dataset after the adversary first receives an annotated dataset by asking the victim model for a set of auxiliary samples. By comparing similarities between the victim model and a suspect model, the framework may identify stolen machine learning models. The study employs DEEPJUDGE on four datasets, including the voice recognition dataset SpeechCommands and the picture classification datasets MNIST, CIFAR-10, and ImageNet. Positive suspect models are created via fine-tuning, pruning, or model extraction from the victim models, whereas negative suspect models are trained independently. With much lower values of RobD and JSD for positive suspect models compared to negative suspect models, the results demonstrate that DEEPJUDGE efficiently detects stolen models. The ROC curve used in the study also demonstrates the usefulness of the suggested measures. [2]

This study analyses the development of StegoNet, a DNN-based stegomalware that embeds harmful code in a DNN model to create a covert channel for triggering physical-world events. The study presents a collection of payload injection and triggering algorithms that take advantage of neural networks' unique properties. Extensive experimental findings and debates on the evasiveness and integrity of payload injection techniques, as well as the dependability and sensitivity of triggering mechanisms, indicate StegoNet's feasibility and practicality. While this approach is unique and effective, it underlines the potential pitfalls of the rising use of DNNs in real-world applications, including the possibility of malevolent use. The report emphasizes the importance of further research into protecting DNN services from potential security risks. StegoNet is an attack procedure that embeds harmful binary malwares into DNN models utilizing specified payload injection techniques and "chessboard" triggers. They alter the Softmax function used in DNN testing in order to extract and run inline routines for their payload within the function via dynamic execution. (Python Exec). The accuracy of the updated DNN models is assessed before and after incorporating malwares using approaches such as LSB substitution/resilience training and value/sign mapping. The authors emphasize that accuracy degradation following payload injection can be relatively minimal for big DNN models because to ample redundant space, while accuracy can be severely deteriorated for smaller models.[3]

The research provides an optimization methodology for building covert and distributed triggers to increase neuron activity to conduct "invisible" backdoor assaults on deep neural networks (DNNs). The suggested technique defines invisibility for human users using Perceptual Adversarial Similarity Score (PASS) and hides the trigger inside the input data using L2 and L0 regularization. By analyzing attack success rates and invisibility scores, the study demonstrates that the suggested invisible backdoors are successful across several DNN models as well as three datasets. The study's main contribution is to give an optimization framework for the development of invisible backdoor assaults that are less obvious by human inspection while still allowing neural networks to recognize the backdoor triggers. The problem of building an invisible backdoor is balancing the efficiency of the trigger in deceiving the ML system with its invisibility to escape detection by human examination. On three datasets, L2 and L0 attackers were evaluated based on Attack Success Rate and Functionality. The results demonstrated that, even with minor changes, both types of attackers may provide good outcomes. L0-norm regularization was discovered to help neural networks to memorize triggers more easily, resulting in faster convergence and a higher Attack Success Rate. The validation accuracy of clean photos was only slightly reduced in functionality. Larger L2-norms were required on GTSRB to achieve an Attack Success Rate comparable to CIFAR while preserving comparable Functionality. The ramifications of these discoveries for the creation of resilient and secure neural network models are significant. [4]

The capsulation attack is a new threat to existing backdoor-based watermarking systems for deep neural networks (DNNs), with the ability to invalidate triggers and ownership queries at little cost. The authors offer CAScore, a new security metric for assessing the security of DNN watermarking techniques against this attack. In addition, they suggest a novel black-box DNN watermarking approach that is resistant to capsulation attack. The paper explains the suggested attack and the requirements for a secure watermarking technique in detail. The evaluation criteria employed in the trials, on the other hand, are not described in length, and the work may benefit from a more thorough examination of the suggested watermarking system. The suggested reverse-backdoor watermarking system is evaluated by comparing it to four baseline watermarking schemes: Noise, Wonder Filter, Stamp, and Steganography. The security of the investigated schemes was assessed using the CAScore, which can only be increased by exhausting a finite set of candidate classifiers. The findings indicated that the Reverse method had the best CAScore since no classifier, even theoretically, can distinguish triggers from normal queries. The effectiveness of the capsulation attack was also assessed, revealing that schemes with a low CAScore may be invalidated with little effort using the capsulation attack. Finally, the functionality-preservation evaluation revealed that, for current deep models with suitable redundancy, the performance loss induced by using the reverse-backdoor was no more than that of other options. [5]

**References:**

[1] EvilModel: Hiding Malware Inside of Neural Network Models

[2] Copy, Right? A Testing Framework for Copyright Protection of Deep Learning Models

[3] StegoNet: Turn Deep Neural Network into a Stegomalware

[4] Invisible Backdoor Attacks Against Deep Neural Networks

[5] Solving The Capsulation Attack Against Backdoor-Based Deep Neural Network Watermarks By

Reversing Triggers