Data Poisoning on Deep Learning Models

Charles Hu
Woodside High School &
Governor's School for Science and Technology
Newport News, Virginia, USA
computerscience@verizon.net

Yen-Hung (Frank) Hu
Department of Computer Science
Norfolk State University
Norfolk, Virginia, USA
yhu@nsu.edu

***Abstract*—Deep learning is a form of artificial intelligence (AI) that has seen rapid development and deployment in computer software as a means to implementing AI functionality with greater efficiency and ease as compared to other alternative AI solutions, with usage seen in systems varying from search and recommendation engines to autonomous vehicles. With the demand for deep learning algorithms that can perform increasingly complex tasks in a shorter time frame growing at an exponential pace, the developments in the efficiency and productivity of algorithms has far outpaced that of the security of such algorithms, drawing concerns over the many unaddressed vulnerabilities that may be exploited to compromise the integrity of these software. This study investigated the ability of poisoning attacks, a form of attack targeting the vulnerability of deep learning training data, to compromise the integrity of a deep learning model's classificational functionality. Experimentation involved the processing of training data sets with varying deep learning models and the incremental introduction of poisoned data sets to view the efficacy of a poisoning attack under multiple circumstances and correlate such with aspects of the model's design conditions. Analysis of results showed evidence of a decrease of classificational ability correlating with an increase of poison percentage in the training data sets, but the scale of which the decrease occurred varied with the specified parameters in the model design. Based on this, it was concluded that poisoning can provide varying levels of damage to deep learning classificational ability depending on the parameters utilized in the model design, and methods to countermeasure such were proposed, such as increasing epoch count, implementing mechanisms bolstering model fit, and integrating input level filtration systems.**

***Index Terms*—deep learning, machine learning, artificial intelligence, data poisoning**

I. Introduction

Deep learning is a subset of machine learning and a form of artificial intelligence that utilizes a structured layer of algorithms in order to process data and generate classificational capabilities. While steeped in the classical practices of artificial intelligence, such as teaching through example, deep learning sets itself apart from other alternatives by incorporating large sets of artificial neural networks and automated feature extraction from data samples, allowing deep learning to process and scale with larger sets of data with greater speed and efficiency [1]. These attributes, alongside the relative ease and reliability at which deep learning can be implemented, have allowed a multitude of organizations to implement and employ AI functionalities into their products and services. Through this, deep learning has spurned on the development and introduction of new and innovative technologies such as autonomous vehicles, assistance devices, and cancer cell detection systems [2].

The strong focus in the deep learning field on productivity and efficiency though has led to the overemphasis on the need to optimize the performance of the algorithms as opposed to the strengthening of the security of them, leading to growing concerns over exposed and unaddressed security threats. Lacking in any major countermeasures in response, these threats have led to the development of multiple attack methodologies that pose a significant danger to systems incorporating deep learning models. An instance of such an attack that proves a major threat, poisoning attacks, relies on the corruption of vulnerable training data to cause incorrect classificational learning, resulting in faulty and even harmful classificational capabilities that can be exploited by malevolent individuals and organizations to potentially devastating results [3].

In spite of these flaws, deep learning still remains the future of artificial intelligence. It is currently the only known viable solution to bridging the gap between big data and artificial intelligence and remains the most promising in terms of moving towards unsupervised learning [4]. It has an estimated market value at $4.4 billion as of 2020 and many major corporations, such as Google, Amazon, Samsung, and NVIDIA, have invested heavily in the advancement and integration of deep learning into many of their applications [5], making it a critical mainstay solution to AI problems. However, security threats like poisoning attacks still pose a significant hindrance to the development of the field.

Thus, this paper delved into the mechanics and processes involved in data poisoning on deep learning and analyzed methods by which poisoning attacks could be used to effect deep learning models and their classificational capabilities.

The remainder of this paper is organized as follows: Section II introduces background of the research. Section III discusses research methodology. Section IV demonstrates research results. Section V concludes the paper with some reflections on findings and suggestions for future work to build upon them.

II. Background

A. Deep Learning Overview

Deep learning is a type of artificial intelligence that is typically categorized as a sub-field of the general AI field of machine learning, and tackles the challenge of intelligent behavior in computers by mimicking the learning process

978-1-7281-7624-6/20/$31.00 ©2020 IEEE
DOI 10.1109/CSCI51800.2020.00111

found in living organisms through the employment of artificial neural networks.

Artificial neural networks, or neural networks, is a design approach to artificial intelligence that emphasizes the processing of data and information in a manner similar to how a biological brain would function. Neural networks are comprised of a collective of nodes, known as neurons, that are structured to form a multitude of layers. These layers are connected to one another through the neurons, in which each neuron both receives and transmits signals to and from adjacent layers in a forward manner (typically dubbed 'feed-forward').

Each neuron contains a specialized value, or weight, that it uses in conjunction with outputs from neurons in the previous layer to calculate a value. If the value reaches a certain threshold, it will output the value to the next layer's neurons. If the value fails to reach the threshold, the neuron will not output the value to the next layer. Through multiple instances of 'training', in which data performs a full run through every layer (known as an 'epoch'), these weights will be gradually adjusted by validation checks until the weights are optimized to perform accurate classification through its training data set [6]. As this process is scalable, the greater the amount of data inputted into the model, the more optimized and accurate the classificational ability will become throughout the training process [7]. This continuous cycle of training and reinforcement, similar to biological processes of learning, forms the backbone of the deep learning process and allows it to produce efficient and productive classificational models.

Deep learning also incorporates the function of automated feature extraction. Feature extraction itself is the process in which the initial raw input data is procedurally reduced in a manner in which the resulting data is far more manageable for the neural networks to process, essentially summarizing the data through selective combination in order to create groups, or 'features', that retain the traits and attributes of the original data set but are now much more computationally efficient to process [8].

Traditional approaches regarding feature extraction for artificial intelligence usually entailed manual feature identification and extraction under the direction of a data scientist, which was typically greatly consuming in both human labor and resources as well subject to human error at times. Deep learning, on the other hand, automated the feature extraction process, allowing the immediate input of raw data into the model instead of requiring the pre-processing of such data [9]. This greatly improved the ease and efficiency at which one could train deep learning models with raw training data sets as compared to other forms of machine learning and artificial intelligence.

B. Data Poisoning Overview

Data poisoning is a type of attack in which a bad actor attempts to alter, or 'poison', either the internal machine learning model algorithm or the machine learning training data.

The general goal of this form of attack is to render a poisoned machine learning model unreliable or incapable of producing the intended output that the system is designed to perform by interfering with the internal learning process occurring within the model.

Poisoning attacks typically vary from factors that influence the required design philosophy, such as intended goal or environmental restrictions.

Intended goals are typically categorized as either targeted attacks, specifically attempting to alter a certain item or process in the model, or non-targeted attacks, not specifically attempting to alter any certain thing but instead anything available to alter; environmental restrictions typically come from how much access an attacker has to the targeted model and is categorized as a white-box, gray-box, or black-box scenario corresponding to full access, partial access, and no access to the model respectively [10].

The table below (see TABLE I) describes the four primary approaches used in poisoning attacks [11].

TABLE I
POISONING ATTACH APPROACHES

Poisoning Approach	Description
Logic Corruption	An attacker has access to the internal algorithms of the model, allowing them the ability to alter internal processes and mechanisms.
Data Manipulation	An attacker has access to the training data set used to train the model, allowing them the ability to alter the training data itself or labels used to aid the classification of the training data.
Data Injection	An attacker has access to the training data set but is limited to only adding data to the pre-existing set.
Transfer Learning	A new model is poisoned by the transferring of an old poisoned training data set to the new one.

Data poisoning is traditionally associated with machine learning, with some of the earliest studies on data poisoning being performed on machine learning based spam filters [12].

Recent work though has shown that deep learning models too are susceptible to data poisoning attacks targeted at neural networks through methods such as training data manipulation [13] and reverse-engineering the computational process of neural networks [14]. These works implicate the transferability of data poisoning from a machine learning level to a deep learning level and demonstrate the potential threats that deep learning models face should they be attacked through this methodology.

III. RESEARCH METHODOLOGY

A. Approach Overview

To view the efficacy of data poisoning on deep learning models and the potential impact possible on its classification capacity, an empirical experiment was performed on a deep learning model with a data manipulation method gradual poisoning of its training sample pool to view the effect on the classification ability of the resultant models.

Two deep learning training models with differing amounts of dropout layers were used for testing. A comparative baseline was established with a clean training set using these two models trained on both 50 and 100 training epochs to view the classification and validation data for an un-poisoned set.

Following such, poison would be incrementally introduced into the training data pool with the deep learning model retrained to view the resulting change in classification and validation data. The resulting classificational output was then tabulated and compared against the baseline to view general trends in change from the incremental data poisoning.

B. Model Design

The deep learning training models were written in Python and built upon the Keras [15] library with a TensorFlow [16] framework, and the code was hosted on a Windows 10 machine with a 4 core 8GB RAM processor.

Model 1 consisted of 10 layers (*i.e.,* 3 convolution layers, 3 polling layers, 1 flatten layer, 1 full connected layer, 1 dropout layer, and 1 final layer) and the Model 2 consisted of 13 layers (*i.e.,* 3 convolution layers, 3 polling layers, 1 flatten layer, 1 full connected layer, 4 dropout layers, and 1 final layer). The two models primary differed in the amount of dropout layers used for model fitting.

Model 1 had a single dropout layer at 0.5 dropout, while Model 2 had 4 dropout layers at 0.3 dropout each; this differentiation allowed the observance of a situation in which a poorly fit model and properly fit model are poisoned, and how such affected their classificational capabilities.

Both models were further differentiated through training on both 50 epochs and 100 epochs with 16 overall trainings per model type through a poisoned training data set range of 0% to 35%.

The deep learning models were trained with a training data set consisting of 400 labeled training images separated into two classes and validated by a pool of 100 labeled validation images similarly separated into two classes.

IV. Research Results

A. Results Overview

Across the testing phase, 64 total accuracy reports were generated by the two deep learning models, where each accuracy report detailed the classification accuracy, classification loss, validation accuracy, and validation loss for each model when trained through the training data set.

These accuracy reports described the general classificational trends across the models as they were retrained with an increasing amount of poisoned data, as well as how the classification trends were affected by factors such as epoch amount and dropout integration.

B. Model 1 Results

Model 1 was a deep learning training model utilizing only one dropout layer at 0.5 dropout value and went through two runs through the poisoned training data set, once at 50 epochs and another at 100 epochs.

At the baseline of 0% poisoned, the model had already shown signs of over-fitting. At 50 epochs, the classification accuracy at 0.9775 (see Fig. 1) and classification loss at 0.1007 (see Fig. 2) showed reasonable rates, but the validation accuracy at 0.9160 (see Fig. 3) and validation loss at 0.3049 (see Fig. 4) showed poor fitting between the model and training data. Similar trends were seen at 100 epochs, with reasonable rates with the classification accuracy and loss, but abnormal validation accuracy and loss rates at 0.9271 and 0.6733, respectively.

Following the introduction of poisoning into the training sample, the classification accuracy and loss faced minor negative alterations in their values (an accuracy drop of about 0.06 for 50 epochs and 0.2 for 100 epochs, and a loss increase of about 0.1 for both epoch sizes), but validation accuracy and loss experienced dramatic alterations (an accuracy drop of about 0.2 across both epoch sizes, and a loss increase by 1.0 and 0.8 for 50 and 100 epochs respectively); thus, indicating the poison had drastically altered the models' ability to classify images outside of its own training sample pool in a greatly negative manner, making it virtually unreliable in classification.

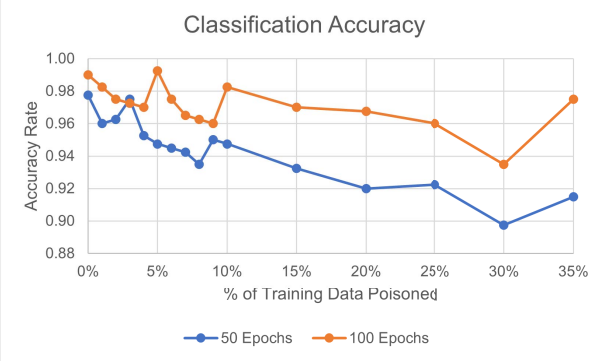

Fig. 1. Classification Accuracy for Model 1

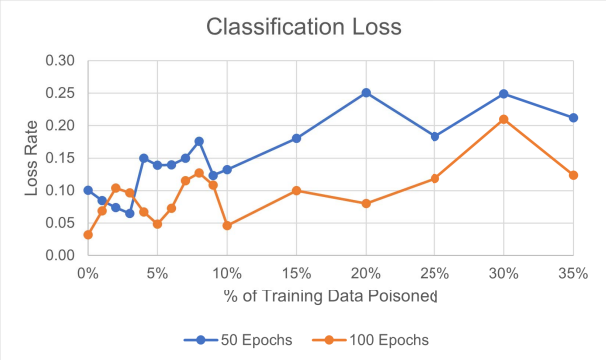

Fig. 2. Classification Loss for Model 1

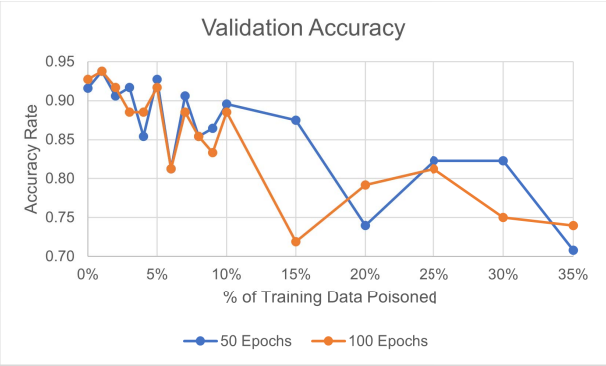

Fig. 3. Validation Accuracy for Model 1

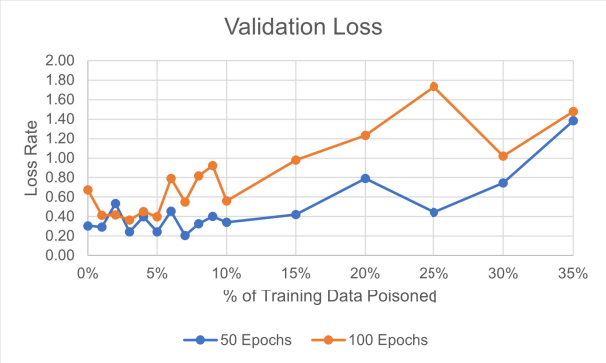

Fig. 4. Validation Loss for Model 1

C. Model 2 Results

Model 2 was a deep learning training model utilizing 4 dropout layers with 0.3 dropout value each and similarly ran through two runs of poisoned training data sets at 50 and 100 epochs.

As it has adequate dropout rates, the baseline classification accuracy and loss as well as the validation accuracy and loss maintain reasonable rates that do not suggest poor model fitting.

After the introduction of poison, classification rates suffered minor alteration in value, while validation rates suffered greater alterations in value but lacked severity in comparison to the results in Model 1.

In classification accuracy (see Fig. 5), the 50 epoch run saw a rate decrease of 0.03 and the 100 epoch run saw a decrease of about 0.02 in value. Classification loss (see Fig. 6) at 50 epochs saw a net increase of 0.02 but experienced fluctuations that had increased the value by 0.1 at times; classification loss at 100 epochs also saw a similar trend in value fluctuation but overall had a net decrease by 0.03 in value.

Validation accuracy (see Fig. 7) generally decreased all around, with the 50 epoch run experiencing a 0.1 drop in value and the 100 epoch run experiencing a 0.3 drop in value. Validation loss (see Fig. 8) increased by 0.3 for the 50 epoch run and by 0.5 for the 100 epoch run.

Generally, the Model 2 runs experienced similar but more subdued alteration in rates compared to the Model 1 runs. The 50 epoch run experienced worse classification rates as a result of the poisoning compared to the 100 epoch run, but the 100 epoch run faced far more drastic an effect on its validation rates compared to the 50 epoch run once poison was introduced.

Overall, the poisoning runs of Model 2 proved damaging to the ability of the model to classify images during the training period for a standard fitting model, as seen with the drop in classification accuracy and increase in classification loss, but even more so demonstrated how damaging poisoning can be on a model's ability to classify new images outside of its training set, as seen with how drastic the validation rates were negatively altered throughout the poisoning process.

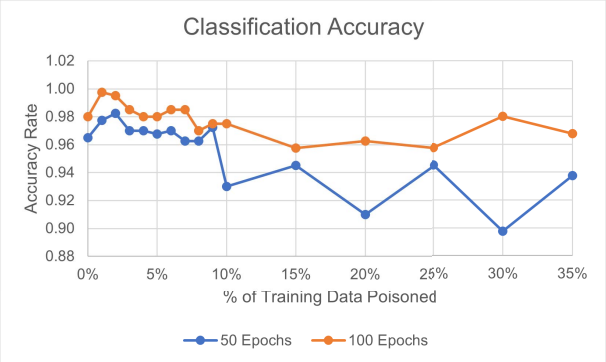

Fig. 5. Classification Accuracy for Model 2

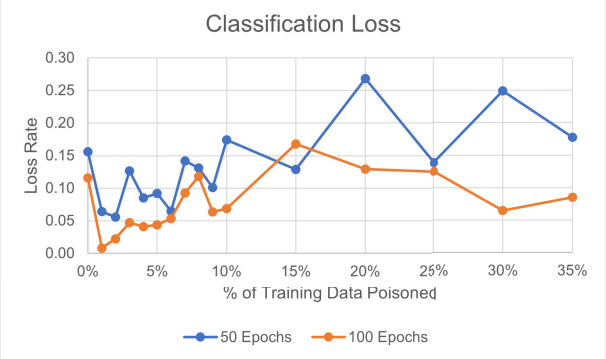

Fig. 6. Classification Loss for Model 2

V. Conclusion and Future Work

As AI functionality becomes increasingly integrated into our technology, the need for deep learning will only grow. Unfortunately, deep learning is not infallible, and the existence of vulnerabilities poses a major threat to the integrity and reliability of deep learning-based software; thus, highlighting the need for research exploring the means by which deep learning can be exploited.

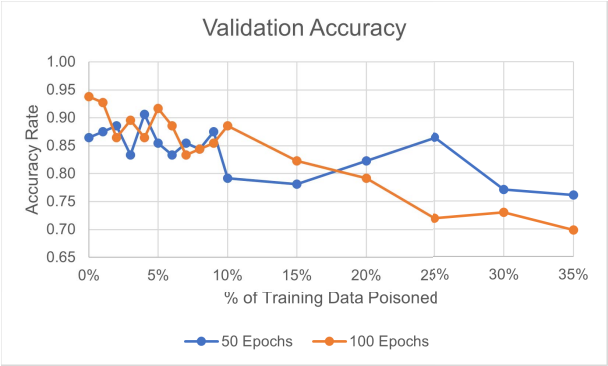

Fig. 7. Validation Accuracy for Model 2

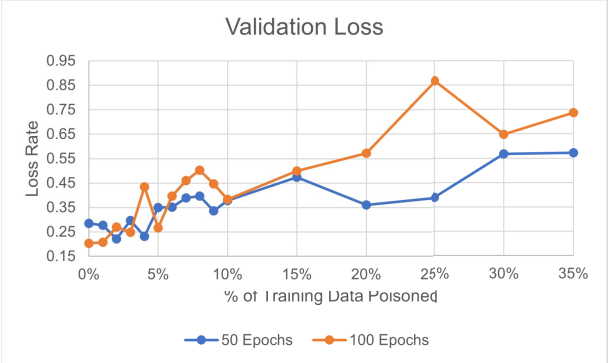

Fig. 8. Validation Loss for Model 2

This paper overviewed the mechanics and processes behind data poisoning for deep learning models and demonstrated and analyzed a method by which poisoning could be applied to a deep learning model, showing how data poisoning can cripple a model's classificational capacity and render it unreliable.

As work on data poisoning for deep learning algorithms is limited, further study is required for a better understanding to expand our ability to diagnose and combat it. Research should be extended to explore other methods of data poisoning for deep learning. As adversarial examples are another major attack approach for deep learning models, further research should also go into viewing the synergistic capabilities of both data poisoning and adversarial examples, and countermeasures to combat both.

Acknowledgment

"This work was supported [in part] by the Commonwealth Cyber Initiative (CCI), an investment in the advancement of cyber R&D, innovation and workforce development. For more information about CCI, visit cyberinitiative.org."

References

[1] Jason Brownlee, "What is Deep Learning?" August 16, 2019, https://machinelearningmastery.com/what-is-deep-learning/
[2] Mathworks, "What Is Deep Learning?" https://www.mathworks.com/discovery/deep-learning.html
[3] Computer Science, University of Maryland, "Poison Frogs! Targeted Poisoning Attacks on Neural Networks," https://www.cs.umd.edu/~tomg/projects/poison/
[4] Keith D. Foote, "A Brief History of Deep Learning," Feburary 7, 2017, https://www.dataversity.net/brief-history-deep-learning/
[5] Reportlinker, "Global Deeping Learning Industry," July 2020, https://www.reportlinker.com/p05798338/Global-Deep-Learning-Industry.html?utm_source=GNW
[6] Larry Hardesty, "Explained: Neural networks," April 14, 2017, https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414
[7] Jeff Dean, "Large-Scale Deep Learning for Intelligent Computer Systems," https://static.googleusercontent.com/media/research.google.com/en//people/jeff/BayLearn2015.pdf
[8] DeepAI, "Feature Extraction," https://deepai.org/machine-learning-glossary-and-terms/feature-extraction
[9] Artem Oppermann, "Artificial Intelligence vs. Machine Learning vs. Deep Learning," October 29, 2019, https://towardsdatascience.com/artificial-intelligence-vs-machine-learning-vs-deep-learning-2210ba8cc4ac
[10] Alexander Polyakov, "How to attack Machine Learning (Evasion, Poisoning, Inference, Trojans, Backdoors)," August 6, 2019, https://towardsdatascience.com/how-to-attack-machine-learning-evasion-poisoning-inference-trojans-backdoors-a7cb5832595c
[11] Ilja Moisejevs, "Poisoning attacks on Machine Learning," July 14, 2019, https://towardsdatascience.com/poisoning-attacks-on-machine-learning-1ff247c254db
[12] Daniel Lowd, Christopher Meek, "Good Word Attacks on Statistical Spam Filters," Semantic Scholar, https://www.semanticscholar.org/paper/Good-Word-Attacks-on-Statistical-Spam-Filters-Lowd-Meek/16358a75a3a6561d042e6874d128d82f5b0bd4b3
[13] Ali Shafahi, W. Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, Tom Goldstein, "Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks," in Proceedings of 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canad, https://papers.nips.cc/paper/7849-poison-frogs-targeted-clean-label-poisoning-attacks-on-neural-networks.pdf
[14] Luis Muñoz-González, Battista Biggio, Ambra Demontis, Andrea Paudice, Vasin Wongrassamee, Emil C. Lupu, Fabio Roli, "Towards Poisoning of Deep Learning Algorithms with Back-gradient Optimization," August 29, 2017, https://arxiv.org/abs/1708.08689
[15] Keras, https://keras.io/
[16] TensorFlow, https://www.tensorflow.org/