Machine Learning Security for Tactical Operations

Dr. DeNaria Fields  
*Virginia Tech*, denaria@vt.edu

Shakiya A. Friend  
*Norfolk State University*, s.a.friend@spartans.nsu.edu

Andrew Hermansen  
*Virginia Tech*, andrewhermansen@vt.edu

Dr. Tugba Erpek  
*Virginia Tech*, terpek@vt.edu

Dr. Yalin E. Sagduyu  
*Virginia Tech*, ysagduyu@vt.edu

Follow this and additional works at: [https://digitalcommons.usf.edu/mca](https://digitalcommons.usf.edu/mca)

Part of the Computer and Systems Architecture Commons, Digital Communications and Networking Commons, Military, War, and Peace Commons, Other Computer Engineering Commons, and the Systems Science Commons

**Recommended Citation**

Fields, Dr. DeNaria; Friend, Shakiya A.; Hermansen, Andrew; Erpek, Dr. Tugba; and Sagduyu, Dr. Yalin E. (2024) "Machine Learning Security for Tactical Operations," *Military Cyber Affairs*: Vol. 7 : Iss. 1 , Article 3. Available at: [https://digitalcommons.usf.edu/mca/vol7/iss1/3](https://digitalcommons.usf.edu/mca/vol7/iss1/3)

This Article is brought to you for free and open access by the Open Access Journals at Digital Commons @ University of South Florida. It has been accepted for inclusion in Military Cyber Affairs by an authorized editor of Digital Commons @ University of South Florida. For more information, please contact digitalcommons@usf.edu.
Machine Learning Security for Tactical Operations

DeNaria Fields, Shakiya Friend, Andrew Hermansen, Dr. Tugba Erpek, and Dr. Yalin E. Sagduyu

Introduction

Machine learning holds significant importance across a broad range of diverse fields due to its ability to analyze and interpret large volumes of data, identify patterns, and make high-fidelity decisions that may not have been feasible in the past. Machine learning also finds rich applications in the tactical domain, such as identifying anomalous patterns in data, corresponding to suspicious activities or behaviors, which can indicate potential security threats or breaches, and assisting real-time decisions by analyzing data from multiple sources, including sensors, intelligence reports, and situational awareness platforms.

Deep learning is a branch of machine learning that uses artificial neural networks to learn from large amounts of data and execute complex tasks [1]. As opposed to traditional machine learning algorithms where domain knowledge is typically leveraged to manually design relevant features to operate on, deep learning models are designed to automatically learn features or representations directly from raw data, eliminating the need for feature engineering. This is one of the key advantages of deep learning, especially for tasks involving high-dimensional data like images, audio, and text. To that end, deep learning applications span various domains such as computer vision, natural language processing, speech recognition, and cybersecurity.

Despite their benefits, deep learning models are susceptible to various attacks and exploits. Adversarial machine learning is an emerging field that studies the vulnerabilities of machine learning models and mechanisms to protect them [2]. In this context, adversaries can exploit weaknesses in the design, training process, or deployment of machine learning systems to manipulate their behavior or compromise their performance. One major security concern is centered around adversarial (evasion) attacks that involve malicious attempts to fool or degrade the performance of the models by adding small perturbations to the input data to introduce biases or vulnerabilities into models. In this paper, we first discuss application areas of deep learning in the tactical domain. Next, we present adversarial machine learning as an emerging attack vector and discuss the impact of adversarial attacks on the deep learning performance. Finally, we discuss potential defense methods that can be applied against these attacks.
Deep Learning for Tactical Networks

Deep learning is a subfield of machine learning where nonlinear transformations are applied on the input data to extract a pattern and perform a regression or classification task. Deep learning has various applications in the realm of tactical networks that include:

- **Object Detection and Recognition**: Deep learning algorithms can be used for real-time object detection and recognition in satellite imagery, aerial surveillance, and ground-based cameras [3]. This aids in identifying enemy vehicles, equipment, and personnel, supporting Intelligence, Surveillance, and Reconnaissance (ISR) efforts. Additionally, Automatic Target Recognition (ATR) systems can be implemented to automatically identify and track potential targets in real-time.

- **Autonomous Navigation**: Deep learning algorithms facilitate autonomous navigation for unmanned vehicles, allowing them to navigate complex environments independently. This reduces the need for constant human intervention, enhancing the efficiency of reconnaissance and surveillance operations.

- **Autonomous Vehicles and Drones**: Deep learning enables the development of autonomous vehicles and drones for reconnaissance and surveillance purposes. These systems can navigate and make decisions based on visual data, minimizing human involvement and risk.

- **Natural Language Processing (NLP)**: NLP can be employed for analyzing and extracting valuable information from vast amounts of textual data, including intelligence reports, social media, and communication intercepts. This aids in understanding and predicting enemy intentions.

- **Predictive Analytics**: Deep learning models can analyze historical data to predict potential threats and suggest optimal strategies. This can include predicting enemy movements, identifying patterns in insurgent activities, and forecasting potential conflicts.

- **Cybersecurity**: Deep learning is used for anomaly detection and pattern recognition in network traffic [4]. It can help in identifying and mitigating cyber threats, securing communication channels, and protecting sensitive military information.

- **Speech Recognition**: Speech recognition technology is valuable for military communication and command systems. It allows for hands-free operation and efficient voice commands, enhancing communication in challenging environments.

- **Simulation and Training**: Deep learning can be used to create realistic simulations for training purposes. This includes simulating battlefield scenarios, enemy behaviors, and various environmental conditions to train military personnel effectively.
• **Health Monitoring and Predictive Maintenance:** Deep learning models can be utilized for monitoring the health and performance of military equipment. Predictive maintenance can help prevent unexpected failures, ensuring that critical assets are in optimal condition.

• **Decision Support Systems:** Deep learning can assist military commanders in decision-making by providing real-time analysis and actionable insights. This includes assessing risks, optimizing resource allocation, and developing strategic plans.

• **Radio Frequency (RF) Environment Understanding:** Deep learning can be employed to interpret the RF environment, including the detection and classification of friendly and adversarial RF emitters, such as radars, jammers, or communication devices [5]. This understanding can inform tactical decisions and enhance situational awareness on the battlefield.

By leveraging deep learning in these application areas, tactical systems can benefit from enhanced performance, resilience, and adaptability, ultimately supporting the mission objectives of military operations.

**Challenges of Deep Learning for the Tactical Domain**

One key step for the successful deployment of deep learning models in the tactical domain is to ensure the reliability, safety, fairness, transparency and accountability of the developed algorithms. In tactical operations, deep learning systems must perform reliably under challenging and dynamic conditions, such as harsh environments, limited communication bandwidth, and adversarial threats. AI assurance deals with supporting AI systems such as those based on deep learning models to meet stringent performance requirements, including accuracy, robustness, and responsiveness, to support mission-critical tasks such as target recognition, threat detection, and decision support. Figure 1 illustrates some of the factors that shape the trust in deep learning models, namely, environmental uncertainties, repeatability, bias, explainability, robustness and security.
Explainability is essential for building trust, enhancing transparency, mitigating biases, and facilitating informed decision-making in deep learning applications. Explainable machine learning models support regulatory compliance and accountability requirements by providing justification for model predictions and decisions. By understanding how models arrive at their predictions or recommendations, military personnel can gain valuable insights into the operational context, mission objectives, and potential threats, enhancing situational awareness and decision-making capabilities.

Bias refers to the systematic errors or inaccuracies in the predictions or decisions made by a deep learning model due to the model’s inability to represent the true underlying relationship between the input features and the target variable. Bias can arise from different sources such as algorithmic, label or feature bias. Biases in sensor data, intelligence reports, or situational awareness data may lead to erroneous model predictions or suboptimal decision-making in dynamic and uncertain battlefield conditions. By addressing bias in deep learning within the tactical domain, military organizations can enhance the fairness, effectiveness, and trustworthiness of automated decision support systems.

Repeatability is fundamental to integrity, trustworthiness, and advancement of deep learning research and applications. When models and experiments are repeatable, different stakeholders in the tactical field can assess the methodology, assumptions, and multi-domain data used in the research process, fostering trust in the reported findings and conclusions.

Robustness in deep learning refers to the ability of a model to maintain performance and stability across diverse and challenging conditions, including variations in data distributions, environmental changes, and attacks, as often expected in the modern battlefield.
Environmental uncertainties can degrade the quality of data collected from a variety of sources. Furthermore, machine learning models trained on data from a specific environment may struggle to generalize to new or unseen environments with different characteristics. This poses a major challenge for tactical operations that are subject to dynamic battlefield conditions and unknown adversarial effects.

Security in machine learning is critically important due to the potential vulnerabilities and risks associated with data-driven systems. Adversaries may attempt to exploit vulnerabilities in deep learning models to deceive or manipulate them, leading to compromised decision-making and mission failure. Security measures should be implemented to help safeguard this data against unauthorized access, ensuring the confidentiality and integrity of critical information conveyed in tactical operations.

In the remainder of this article, we will focus on machine learning security, explain different attack types, discuss how they affect the performance of deep learning algorithms, and review potential defense techniques.

Adversarial Machine Learning

Adversarial machine learning refers to the study of understanding and mitigating vulnerabilities in machine/deep learning models where adversaries aim to manipulate or deceive models by exploiting weaknesses in their design, training process, or input data in test time. Adversarial machine learning includes a variety of attacks including exploratory, adversarial (evasion), poisoning and backdoor (Trojan) attacks [6]-[10].

- In exploratory attacks, adversaries use probe-based strategies to interrogate the deep learning model by submitting carefully crafted queries or input samples. These queries are designed to reveal information about the model’s internal workings, such as its decision boundaries, feature representations, or vulnerabilities to adversarial manipulation.
- Adversarial (evasion) attacks craft input samples that evade detection or mislead the model’s predictions during the operation. Adversaries may exploit vulnerabilities in model architectures or feature representations to generate inputs that are eventually misclassified. Adversarial inputs often involve making small, imperceptible modifications to input data that lead to significant changes in model outputs.
- Model poisoning attacks involve injecting malicious data or manipulating training datasets to compromise the performance or integrity of deep learning models. Adversaries may insert biased or misleading examples into training data to influence model behavior or introduce backdoors that enable unauthorized access or control over the model.
- Backdoor (Trojan) attacks involve the insertion of maliciously crafted inputs into the training data to manipulate the behavior of machine learning models.
These inputs contain subtle alterations that are inconspicuous during normal operation but trigger specific responses or misclassifications by the model when encountered during inference. This manipulation can compromise the integrity and reliability of the model.

Privacy attacks can also be applied on deep learning algorithms through membership inference and model extraction attacks.

- Membership inference attacks aim to determine whether a particular data sample was used in the training dataset of a deep learning model. Adversaries exploit information leakage in model outputs to infer the presence or absence of specific samples in the training data, compromising the privacy and confidentiality of sensitive information.
- Model extraction attacks involve reverse-engineering deep learning models to extract sensitive information or proprietary knowledge embedded in model parameters. Adversaries may use black-box queries, model inversion techniques, or membership inference attacks to infer details about the model's architecture, training data, or decision boundaries.

In deep learning, attacks can also be categorized based on the level of knowledge the attacker has about the target model and its internal workings. The three main categories are white-box, black-box, and gray-box attacks.

- In white-box attacks, the attacker has complete knowledge of the target model, including its architecture, parameters, and training data. The attackers can directly access the model's internal representations, gradients, and decision boundaries, making them highly effective at crafting adversarial examples.
- In black-box attacks, the attacker has limited or no access to the target model's internal parameters or gradients. The attackers can only interact with the target model by querying it with input samples and observing the corresponding outputs. Despite the lack of detailed knowledge about the target model, black-box attackers aim to craft adversarial examples using techniques such as transferability, where adversarial examples generated on a substitute model are used to attack the target model.
- Gray-box attacks lie between white-box and black-box attacks, where the attacker has partial knowledge about the target model. In gray-box attacks, the attacker may have access to some information about the target model, such as its architecture or output predictions, but not its internal parameters or gradients.

Each type of attack presents unique challenges and considerations for defending against adversarial threats in deep learning systems. In the next section,
we will delve into the adversarial attacks and discuss how they impact the deep learning performance.

**Adversarial Attacks**

Adversarial attacks aim to manipulate the input data in test time to fool deep learning models into making wrong decisions. Many methods have been proposed in the literature to launch adversarial attacks [11]. We focus on Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) for our analysis.

- **FGSM** [12] is one of the simplest and most effective techniques for crafting adversarial examples. It works by perturbing input data in the direction of the gradient of the model’s loss function with respect to the input features while staying under a certain perturbation value, epsilon. The FGSM attack is fast and computationally efficient, requiring only a single gradient computation to generate adversarial examples. However, FGSM-generated adversarial examples may lack robustness and transferability to other models or defenses.

- **PGD** [13] is an iterative method to determine the perturbations that can mislead the model. In other words, the PGD attack slightly manipulates input data and its features to cause the model to make a mistake. If the input data is sensitive enough, especially to small changes, then the PGD attack will have a higher chance of causing the model to misclassify. The PGD attack is more computationally intensive compared to FGSM since it involves multiple iterations of gradient descent. However, PGD-generated adversarial examples are generally more effective and robust, as the iterative optimization process explores the local geometry of the loss landscape more thoroughly.

To illustrate the impact of adversarial attacks, we consider the FGSM and PGD attacks against ResNet-50 model on ImageNet dataset [14]. This dataset spans 1000 object classes and contains 1,281,167 training images, 50,000 validation images and 100,000 test images [15]. ResNet-50 is a convolutional neural network (CNN) architecture consisting of 50 layers, including convolutional layers, pooling layers, batch normalization layers, and fully connected layers. ResNet-50 has achieved state-of-the-art performance on various benchmark datasets, including ImageNet and is widely used as a feature extractor or backbone architecture in various computer vision tasks, including image classification, object detection, image segmentation, and image generation. Figure 2 shows the attack performance results in terms of the classifier accuracy with increasing perturbation level that we refer to as epsilon, for FGSM method. Adversarial machine learning attacks can be categorized as *untargeted* and *targeted attacks*. *Untargeted attacks* are characterized by an attempt to perturb input data in such a way that it causes the machine learning model to make an incorrect prediction, without specifying any particular target class or outcome. On the other hand, *targeted attacks* aim to manipulate the model's behavior towards a specific class or decision. We focus on
untargeted attacks in our implementation. Initially, the classifier accuracy is 87.5% before any attack is launched. The accuracy drops to 0% with increasing epsilon value. PGD method is more effective in decreasing the performance of the classifier compared to the FGSM. On the other hand, as an iterative method, PGD takes longer time to compute. Figure 3 shows the classifier accuracy under PGD attack with number of iterations. We observe that the classifier accuracy continues to decrease with increasing number of iterations and drops to 0% after 13 iterations.

![Figure 2. Classifier accuracy vs. epsilon under FGSM attack](https://digitalcommons.usf.edu/mca/vol7/iss1/3)
Defense Against Adversarial Machine Learning Attacks

With the increased use of deep learning in diverse set of applications in tactical operations, adversarial machine learning attacks will continue to be a challenge. However, it is important to think about defensive strategies and take proactive measures to mitigate these problems. Mitigating adversarial attacks in deep learning requires robust and resilient defense mechanisms. One particular defense mechanism is adversarial training that involves augmenting the training dataset with adversarial examples to improve model robustness and resilience to adversarial attacks [12]. By exposing models to adversarial perturbations during training, deep learning systems learn to generalize better and become more robust to adversarial manipulations. Ensemble learning is another method that can be leveraged as a defense mechanism. Ensemble methods combine multiple deep learning models to make predictions, improving model robustness and resilience to adversarial attacks. By leveraging diverse model architectures and training strategies, ensemble methods can help detect and mitigate adversarial examples more effectively.

Developing techniques for detecting and defending against adversarial attacks in real-time can help protect deep learning-based tactical systems from malicious manipulation. Adversarial detection methods such as anomaly detection, outlier detection, and adversarial robustness certification help identify and mitigate adversarial inputs before they can cause harm to the system. Overall, adversarial machine learning is an important area of research focused on understanding and mitigating vulnerabilities in deep learning systems. By developing robust defense

Figure 3. Classifier accuracy vs. number of iterations under PGD attack
mechanisms and resilient model architectures, researchers and practitioners can enhance the security, reliability, and trustworthiness of deep learning-driven systems in the face of adversarial threats.

Conclusion

Deep learning offers critical benefits for tactical operations across various military and defense applications. In this paper, we first explored some of the application areas that deep learning can be used to enhance the mission performance. Then, we highlighted that deep learning algorithms are susceptible to adversarial threats which can modify the input data to force the deep learning models to make an incorrect decision. We discussed different attack types that can be implemented on deep learning models and showed how these attacks can lead to a large degradation in the model accuracy even with small perturbations on the data. Finally, we highlighted some of the defense methods that can be applied to detect and minimize the impacts of these attacks on the deep learning algorithms.

About the Authors

DeNaria Fields

DeNaria Fields is a senior at Virginia Tech pursuing a degree in Computer Science. She is currently an undergraduate researcher for the Hume Center and over the past two years she has been given the opportunity to research and study adversarial machine learning. She has focused on studying machine learning/deep learning and their vulnerabilities and ways to make them more robust.

Shakiya Friend

Shakiya Friend is a senior computer science major at Norfolk State University. Shakiya has experience in various computer science fields including cloud computing, quantum computing, machine learning, artificial intelligence, and software development. Shakiya primarily works in backend software development. Shakiya has previously done Quantum Computing research with IBM and Brookhaven National Labs. Shakiya currently does Adversarial Machine Learning research alongside the VICEROY program.

Andrew Hermansen

Andrew Hermansen is a Senior at Virginia Polytechnic Institute and State University studying Computer Engineering with a focus in Networking and Cybersecurity. He has been working with Hume Center for the past 2 years on understanding adversarial and defense mechanisms for machine learning and deep learning. He has experience with Python and C++ as his primary coding languages. He spent a summer working with the Air Force Research Lab through the
VICEROY program where he had hands-on experience understanding model-based engineering.

**Dr. Tugba Erpek**

Dr. Tugba Erpek is a Research Associate Professor in the Intelligent Systems Division of the Virginia Tech National Security Institute. Her R&D work covers machine learning, adversarial machine learning, wireless communications, and physical layer security. Prior to joining Virginia Tech, she was a Lead Scientist and Network Communications Technical Area Lead at the Intelligent Automation, a BlueHalo Company and a Senior Communications Systems Engineer at the Shared Spectrum Company where she developed deep learning algorithms to improve the performance, situational awareness and security of next-generation communications systems.

**Dr. Yalin E. Sagduyu**

Dr. Yalin E. Sagduyu is a research professor in the Intelligent Systems Division of the Virginia Tech National Security Institute. His research interests include wireless communications, networks, security, 5G/6G systems, machine learning, adversarial machine learning, and data analytics. Prior to joining Virginia Tech, he was the director of Networks and Security with Intelligent Automation, a BlueHalo Company, where he directed a broad portfolio of R&D projects and product development efforts related to networks and security.

**Acknowledgments**

Effort sponsored by the AFRL under Agreement FA8750-19-3-1000. The U.S. Government is authorized to reproduce and distribute copies for Governmental purposes notwithstanding any copyright or other restrictive legends.

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force or the U.S. Government.

**References**


