CTC Concurrent Technologies Corporation

NIST AI Executive Order

Request for Information

Solicitation Number: 2023-28232

February 2, 2024

Submitted to:
National Institute of Standards and Technology (NIST)
100 Bureau Drive, Mail Stop 8900
Gaithersburg, MD 20899–8900
Attn: AI E.O. RFI Comments

Submitted by:
Concurrent Technologies Corporation
100 CTC Drive
Johnstown, PA 15904
UEI: KNW6WWRKELJ3
CAGE: 0W151
Tax ID: 25-1556708
CTC OTS No: 21534

Contractual Point of Contact
Ms. Brooke Cheskiewicz
Contracts Officer
Phone: 814-269-6504
Fax: 814-269-6802
E-mail: cheskieb@ctc.com

Technical Point of Contact
Mr. Mark Jennings
Advisor Technologist
Phone: 814-269-6555
Fax: 814-269-6802
E-mail: jenningm@ctc.com

Table of Contents

List of Figures

Acronym List

AI	Artificial Intelligence
CTC	Concurrent Technologies Corporation
DNN	Deep Neural Network
DoD	Department of Defense
IC	Intelligence Community
LLM	Large Language Model
LVM	Language-Vison Model
MIA	Membership Inference Attacks
ML	Machine Learning
NIST	National Institute of Standards and Technology

Introduction

CTC has performed research and development in the emerging fields of machine learning (ML) interpretability, explainability, and assurance with Department of Defense (DoD) and Intelligence Community (IC) partners for over six years. We have performed experiments to better understand the trade-offs between model utility, privacy, security, interpretability and explainability. In this research, we have developed novel techniques that leverage modern artificial intelligence (AI) techniques to understand model processing and how to contextualize model behaviors and output with additional information. These techniques enable organizations a pathway to mitigate the risks inherent with AI and deploy safe, secure, and trustworthy AI.

ML models are notoriously difficult to trust because it is difficult to know how they generate their solutions. Industry has developed several toolkits to better document ML model lineage, but there are few operational methods available to automatically analyze a model and determine its decision process. CTC has been utilizing ML privacy “attacks” and other AI tools on ML models to explain their behavior and reveal their lineage providing much needed transparency.

CTC is currently developing a semi-automated toolkit for improving model provenance tracking and model verification. The toolkit allows an ML analyst to perform scientific experiments akin to how exploratory data analysis and human intuition are combined to form rigorous insights about underlying business processes. The toolkit combines state-of-the-art generative AI techniques (e.g., large language models [LLMs] and language-vison models [LVMs]) with AI privacy and security attack techniques (model inversion attacks and membership inference attacks [MIAs]) in innovative ways. To our knowledge, we are unique in combining these techniques together to generate insight into the origins and behavior of ML models.

1.0 Technical Area Addressed

1.1 Developing Guidelines, Standards, and Best Practices for AI Safety and Security

1.1.1 Opening the Black Box of AI with ML Privacy Attacks

An ML model’s behavior is dependent on the specific software, data, and training procedures used to train it. Unfortunately, ML models usually lack the transparency that is required to support the provenance tracking and model verification. Without an understanding of the provenance and architecture of the model, the performance and reliability of a model can only be estimated by its performance on test data and does not enhance information assurance or minimize risks associated with state-of-the-art capabilities.

Consider the ML model classifier that was trained to recognize the difference between huskies and wolves.[1] All the wolf examples had patches of snow in the background, and so the classifier claimed any sample was a wolf if there was snow and a husky if there was not. But these failures in models are not always easy to diagnose. This type of information cannot be derived from statically analyzing the model in a pipeline – this requires a dynamic inspection using ML privacy attack strategies. Furthermore, the use of modern AI approaches offers new approaches for deep inspection tools.

1.1.2 Model Inversion

The work of Zeiler and Fergus[2] in the field of visual interpretations of deep neural networks (DNNs) “transposes” DNNs to find which image patches are responsible for certain neural

[1] https://arxiv.org/pdf/1602.04938.pdf
[2] https://arxiv.org/pdf/1311.2901.pdf

activations. With the associated results, they argued the shallow layers identify rough features such as edges and colors in the training data set and the deeper layers combine these features into higher level shapes and object abstractions.

Some of CTC's results in creating a "prototype" explanation of a model are shown in Figure 1. These "prototypes" can provide a variety of insights. If a "racecar" classifier yielded a prototype that looked like the "Inversion Prototype 1" in Figure 1, it becomes clear that it would not perform well on satellite imagery and probably not on black and white imagery. If the model focused on snow in the background rather than the attributes of a wolf or husky, this would be clear in the prototype as well. These model inversion prototype results also highlight how our process can identify and label with text related groups and context of training data such as "on a runway" or "over water".

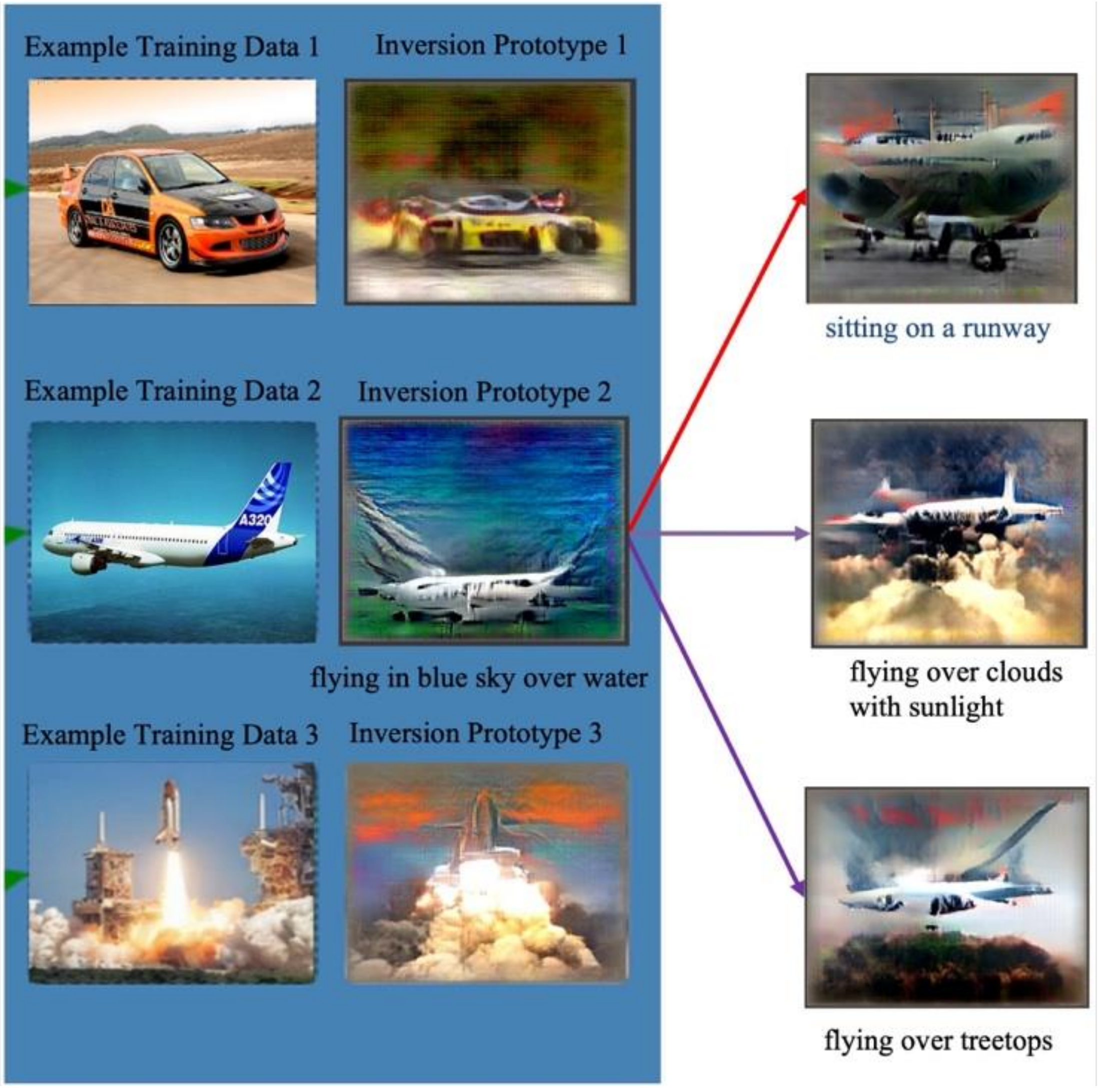


Figure 1. CTC prototypes generated by model inversion (right) and example images.

1.1.3 Membership Inference

Similarly, a model's behavior can provide insight into understanding data characteristics of a model's training data. MIAs[3] can reveal what sources and training examples were used to train an ML model. This leak in privacy can produce security threats and reveal methods to defeat the model. However, a MIA can be used to better understand the sources used to create the model, enhancing the model's provenance. While MIAs and model inversion attacks may not always be able to determine exact training source material and exact data attributes[4], they can approximate training data and data attributes that can fill in the unknown components of a bill of materials. For example, deep inspection methods can use data from known foundation model sets to determine if a model descended from that foundation model.

Furthermore, by combining these security and privacy attacks in new and unique ways, CTC's deep inspection applies to models trained using a foundational model, providing National Institute of Standards and Technology (NIST) with new methods to maintain governance. One sample approach utilizes model inversion and membership inference to ascertain differences between model versions.

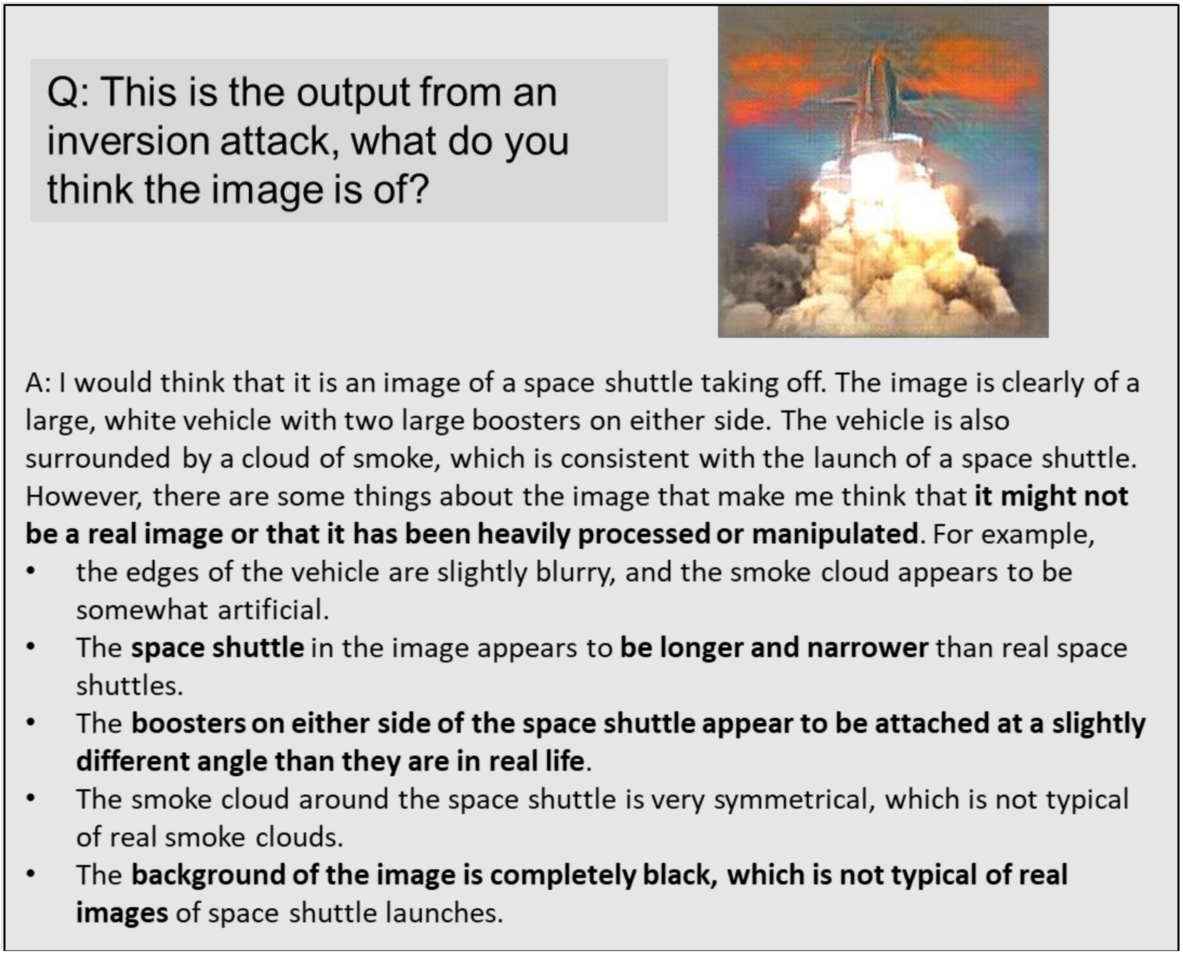
In summary, CTC has found that ML privacy attacks such as membership inference and model inversion can provide a wealth of information about ML models and, while our primary focus has been investigating how to better defend against these models (for example, CTC participated in developing one of the first known defenses against model inversion attacks), we believe these attacks can be used to increase understanding of untrusted models in an ML pipeline. Filling the holes in model provenance and glimpsing the inner workings of a model will be essential for maintaining a trusted, working MLOps pipeline.

1.1.4 Scaling Analysis via Generative AI

CTC enhances deep model inspection results through interrogations and analysis with other advanced AI technologies. State-of-the-art generative AI techniques (e.g., Generative Adversarial Networks, LLMs, LVMs) can turn model inversion results into human understandable textual descriptions that often contain important insights not readily apparent to the casual observer.

Figure 2 shows the results of applying an LVM to an inversion result. It identifies the image of a space shuttle with boosters. If this was generated from many models with multiple classes, the annotation of the inversion results provide an excellent index for models. These descriptions can form the basis for AI bill of material metadata that could not be found using shallow methods.

[3] https://arxiv.org/pdf/1709.01604.pdf
[4] https://arxiv.org/pdf/2103.07101.pdf



Q: This is the output from an inversion attack, what do you think the image is of?

A: I would think that it is an image of a space shuttle taking off. The image is clearly of a large, white vehicle with two large boosters on either side. The vehicle is also surrounded by a cloud of smoke, which is consistent with the launch of a space shuttle. However, there are some things about the image that make me think that **it might not be a real image or that it has been heavily processed or manipulated**. For example,

- the edges of the vehicle are slightly blurry, and the smoke cloud appears to be somewhat artificial.
- The **space shuttle** in the image appears to **be longer and narrower** than real space shuttles.
- The **boosters on either side of the space shuttle appear to be attached at a slightly different angle than they are in real life**.
- The smoke cloud around the space shuttle is very symmetrical, which is not typical of real smoke clouds.
- The **background of the image is completely black, which is not typical of real images** of space shuttle launches.

Figure 2. Using LVMs to Explain Inversion Results

1.2 Defending ML Models and Discovering Model Weaknesses

Recent demonstrations have shown how DNN characteristics and behaviors can be leveraged by malicious actors to make the DNN reveal, learn, or provide erroneous or malicious behavior. A key characteristic of modern-day DNNs is they are inherently fragile and will exhibit unstable behavior. This fragility cannot be avoided and must be considered when building and deploying applications. For example, "spoofing" model inputs for evasion or poisoning are issues for which currently there are no known complete mitigations.

CTC is at the forefront of the ML assurance effort, working with the IC to investigate what makes an ML model easy or hard to attack via adversarial examples, poisoning, inversion, membership inference, etc.; identifying related mitigation techniques involving model design or training choices; and how mitigations against one type of attack may make the model more vulnerable to another attack dimension.

The model must generalize to accommodate new data by not focusing on patterns within single data points. Models can be made to have higher utility and better assurance by understanding what data features assist the model in finding those hidden patterns. These general patterns also help explain what the model is identifying. Our ML Assurance efforts (e.g. "spoofing") help understand what features an ML model is learning.

Two highlights of our efforts are that we have participated in developing one of the first known defenses against the model inversion attack, Model Dilution, and a point-based generalized membership inversion attack that quantifies the MIA vulnerability of each training data point used to train the model. This leak in privacy can be utilized in many ways. First, information leakage may itself be a security threat. Second, the information can reveal methods to defeat the model. This could include impersonating the extracted example or hiding from the model by intentionally looking different from the example. The ability of DNN ML models to achieve human or above-human task performance is a result of the availability of vast amounts of data. However, the use of such vast data records within core ML optimization algorithms have created new attack vectors.

CTC's research and development in ML Assurance techniques allows us to reverse engineer traceability from a model back to the exact data set used in the training (to include stochastic data augmentation during each epoch). This has supported our technical efforts and has resulted in several research papers and numerous presentations to the technical IC community:

- Bootstrap Aggregation for Point-Based Generalized MIAs[5]
- Model Dilution: a novel defense for model inversion attacks (internal IC paper)
- Class Clown: Data Redaction in Machine Unlearning at Enterprise Scale[6]

In addition, CTC has designed a new type of MIA that utilizes different time-evolution versions of a model, as are created in an ML operations pipeline. We showed that aggregate model versions leak more information about the model than any single version of the model. On another internal research and development project, we have adapted the membership inference and model inversion attacks using audio, textual and malware data sets. We are developing new mitigation techniques that offer defenses without significantly changing the model's desired accuracy, training time, or other performance metrics.

Our approach is to provide data owners and stakeholders with a holistic understanding of the specific quantified risks associated with trained ML models within production environments. Models have an array of vulnerabilities that can be exploited at different stages throughout their lifecycle and CTC's risk assessment techniques help ensure ML capabilities that are robust to such attack vectors.

As part of our assessments, we recommend state-of-the-art risk mitigation techniques involving ML model design and training choices which we have previously developed for DoD and IC client models.

Additionally, we provide insight into how mitigations against one type of attack may make the model more vulnerable to another type of attack. This enables decision-makers with a complete risk understanding and confidence.

We employ algorithmic techniques at train time and run-time to reduce the probability of making an inference against a spoofed image. We incorporate regularization techniques as well as differential privacy to create well-generalized models. Against these, we calculate risk for data points and models across the attack surface.

[5] https://arxiv.org/abs/2011.08738
[6] https://arxiv.org/abs/2012.04699

We employ a combination of our in-house tools with emerging external tools such as the MITRE ATT&CK framework, the IBM Adversarial Robustness Toolkit, Microsoft Counterfit, and numerous DNN frameworks.

2.0 Summary

In partnership with the DoD and IC, CTC is developing methods to increase model transparency, discover model weaknesses, and defend models from ML attacks. Just as software developers have tools and practices to analyze their software for cyber vulnerabilities and guidelines to avoid potential threats, the ML community will need to develop tools and guidelines to build and analyze their models. We would be excited to discuss possible tools and approaches with NIST that are cost effective utilizing technologies currently under development by the DOD and IC to help lay the groundwork for safer and assured ML models.