System and Method for Protecting Artificial Intelligence (AI) Systems (#8261, 8263)

A task-generalizable system and method for detecting deceptive inputs and classifying anomalies in AI systems

Inventors at Georgia Tech have developed a system and method to improve AI systems. The invention can be used for detecting and protecting against deceptive inputs for AI systems and measuring vulnerabilities of existing AI systems. This system and method can detect distorted data classification, rectify distorted data classification, detect adversarial (malicious) data, rectify adversarial (malicious) data, and provide score for the vulnerability/safety of any pre-trained AI system. The inventors have also developed a mechanism/apparatus that is capable of differentiating between normal and abnormal data for any AI algorithm, effectively detecting anomalies for pre-trained AI algorithms, working on any type of data, and performing the task for any AI system as an add on.

Benefits/Advantages

- **Scalable**: can be scalable to very large datasets.
- **Real-time**: real-time detection of deceitful, anomalous, and noisy data that do not promote safe usage of AI algorithms.
- **Time-saving**: system provides a representation of the same data that is more conducive to the AI system without requiring retraining.

Potential Commercial Applications

- Any system that uses a neural network to perform a task such as classification, tracking, detection, segmentation, translation, and mapping can use the invention to boost performance. Applications include but are not limited to:
  - Adversarial Image Detection and Defense
  - Noisy Image Recognition
  - Domain adaptation
  - Anomalous Input Classification
  - Aberrant Event Recognition
  - Robust Real Time Image Detection
  - ML system stress test
  - Autonomous Vehicles
  - Assisted driving
  - Subsurface analysis
  - Medical scans analysis
  - Materials analysis
  - Satellite images analysis
Robust feature representation (Audio, Image, Video)

**Background/Context for This Invention**

Recent advancements in artificial intelligence (AI) achieve generalizable performance across large datasets, but most AI algorithms cannot properly handle input data not presented during their training. Therefore, the characterization of anomalies, which have classes or attributes that cannot be inferred from training data, is essential to ensure the robustness of AI algorithms. Additionally, AI systems trained on pristine images have shown a vulnerability to distortions. Visual distortions including but not limited to blur, noise, and exposure can cause neural nets to operate incorrectly if relying on the absolute position of a data point in its respective manifold.

**Ghassan AlRegib**  
Professor – Georgia Tech School of Electrical and Computer Engineering

**Dogancan Temel**  
Postdoctoral Fellow - Georgia Tech School of Electrical and Computer Engineering

**Gukyeong Kwon**  
Graduate Student – Georgia Tech School of Electrical and Computer Engineering

**Mohit Prabhushankar**  
Graduate Student – Georgia Tech School of Electrical and Computer Engineering

For more information about this technology, please visit:  