

CLAIMS

1. A method for securing trained machine learning models in a computing environment by one or more processors comprising:

providing one or more hardened machine learning models secured against adversarial attacks by adding adversarial protection to one or more previously trained machine learning models, wherein the one or more hardened machine learning models are generated by iteratively retraining, following a specified protocol with a projected gradient descent white-box attack with an infinity norm constraint on adversarial samples, the one or more previously trained machine learning models using a ratio of the adversarial examples to benign examples selected from mini-batches of training data not previously input into the one or more trained machine learning models prior to a current retraining iteration, the ratio of adversarial examples iteratively increased during each retraining iteration according to a cosine schedule indicative of an amount of the adversarial examples and benign examples to select from the mini-batches of the training data to use for each subsequent retraining iterations;

monitoring and tracking each state of the one or more previously trained machine learning models while being retrained;

detecting a training collapse to the one or more previously trained machine learning models during the retraining; and

enabling one or more roll-back strategies for the one or more previously trained machine learning models during the retraining.

2. The method of claim 1, further including: receiving the one or more previously trained machine learning models; and performing the retraining of the one or more previously trained machine learning models to include the adversarial protection based on one or more adversarial protection protocols.

3. The method of claim 1, further including adding one or more preprocessing layers to the one or more previously trained machine learning models, wherein a degree of adversarial protection strength is adjusted in each of the one or more preprocessing layers.

4. The method of claim 1, further including adding one or more neural network layers to the one or more previously trained machine learning models.

5. The method of claim 1, further including adding one or more post-processing output layers to the one or more previously trained machine learning models, the one or more post-processing output layers having been retrained with the adversarial protection based on one or more adversarial protection protocols.

6. The method of claim 1, further including: automatically implementing one or more adversarial protection protocols to be used for providing the one or more hardened machine learning models; or receiving, from a user, one or more adversarial protection protocols to be used for providing the one or more hardened machine learning models.

7. A system for securing trained machine learning models in a computing environment, comprising:

one or more computers with executable instructions that when executed cause the system to:

provide one or more hardened machine learning models secured against adversarial attacks by adding adversarial protection to one or more previously trained machine learning models, wherein the one or more hardened machine learning models are generated by iteratively retraining, following a specified protocol with a projected gradient descent white-box attack with an infinity norm constraint on adversarial samples, the one or more previously trained machine learning models using a ratio of the adversarial examples to benign examples selected from mini-batches of training data not previously input into the one or more trained machine learning models prior to a current retraining iteration, the ratio of adversarial examples iteratively increased during each retraining iteration according to a cosine schedule indicative of an amount of the adversarial examples and benign examples to select from the mini-batches of the training data to use for each subsequent retraining iteration;

monitor and track each state of the one or more previously trained machine learning models while being retrained;

detect a training collapse to the one or more previously trained machine learning models during the retraining; and

enable one or more roll-back strategies for the one or more previously trained machine learning models during the retraining.

8. The system of claim 7, wherein the executable instructions: receive the one or more previously trained machine learning models; and performing the retraining of the one or more previously trained machine learning models to include the adversarial protection based on one or more adversarial protection protocols.
9. The system of claim 7, wherein the executable instructions add one or more preprocessing layers to the one or more previously trained machine learning models, wherein a degree of adversarial protection strength is adjusted in each of the one or more preprocessing layers.
10. The system of claim 7, wherein the executable instructions add one or more neural network layers to the one or more previously trained machine learning models.
11. The system of claim 7, wherein the executable instructions add one or more post-processing output layers to the one or more previously trained machine learning models, the one or more post-processing output layers having been retrained with the adversarial protection based on one or more adversarial protection protocols.
12. The system of claim 7, wherein the executable instructions: automatically implement one or more adversarial protection protocols to be used for providing the one or more hardened machine learning models; or receive, from a user, one or more adversarial protection protocols to be used for providing the one or more hardened machine learning models.
13. A computer program product for, by a processor, securing trained machine learning models in a computing environment, the computer program product comprising a non-transitory computer-readable storage medium having computer-readable program code portions stored therein, the computer-readable program code portions comprising:

an executable portion that provides one or more hardened machine learning models secured against adversarial attacks by adding adversarial protection to one or more previously trained machine learning models, wherein the one or more hardened machine learning models are generated by iteratively retraining, following a specified protocol with a projected gradient descent white-box attack with an infinity norm constraint on adversarial samples, the one or more previously trained machine learning models using a ratio of the adversarial examples to benign examples selected from mini-batches of training data not previously input into the one or

more trained machine learning models prior to a current retraining iteration, the ratio of adversarial examples iteratively increased during each retraining iteration according to a cosine schedule indicative of an amount of the adversarial examples and benign examples to select from the mini-batches of the training data to use for each subsequent retraining iterations;

an executable portion that monitors and tracks each state of the one or more previously trained machine learning models while being retrained;

an executable portion that detects a training collapse to the one or more previously trained machine learning models during the retraining; and

an executable portion that enables one or more roll-back strategies for the one or more previously trained machine learning models during the retraining.

14. The computer program product of claim 13, further including an executable portion that: receives the one or more previously trained machine learning models; and performs the retraining of the one or more previously trained machine learning models to include the adversarial protection based on one or more adversarial protection protocols.

15. The computer program product of claim 13, further including an executable portion that adds one or more preprocessing layers to the one or more previously trained machine learning models, wherein a degree of adversarial protection strength is adjusted in each of the one or more preprocessing layers.

16. The computer program product of claim 13, further including an executable portion that: adds one or more neural network layers to the one or more previously trained machine learning models; or adds one or more post-processing output layers to the one or more previously trained machine learning models, the one or more post-processing output layers having been retrained with the adversarial protection based on one or more adversarial protection protocols.

17. The computer program product of claim 13, further including an executable portion that: automatically implements one or more adversarial protection protocols to be used for providing the one or more hardened machine learning models; or receives, from a user, one or more adversarial protection protocols to be used for providing the one or more hardened machine learning models.