

CLAIMS

WHAT IS CLAIMED IS:

1. A method for determining an attack on a neural network, the method comprising:
receiving, a data sample at a first classifier neural network and at a watermark classifier neural network, wherein the first classifier neural network is trained using a first dataset and a watermark dataset;
determining, using the first classifier neural network and the data sample, a classification label for the data sample;
determining, using the watermark classifier neural network and the data sample, a classification label for the watermark; and
determining that the data sample is an adversarial data sample associated with the attack on the neural network based on the classification label for the data sample and the classification label for the watermark.
2. The method of claim 1, wherein determining that the data sample is the adversarial data sample further comprises:
determining that a value of the classification label for the data sample is coupled to a value of the classification label for the watermark.
3. The method of claim 1, wherein determining that the data sample is the adversarial data sample further comprises:
determining, using a separator classifier neural network and the classification label for the watermark that the data sample is the adversarial data sample.
4. The method of claim 3, further comprising:
training the watermark classifier neural network by passing at the first dataset through the watermark classifier neural network and generating the watermark classification labels based on the first dataset; and

training the separator classifier neural network using the watermark classification labels generated by the watermark classifier neural network.

5. The method of claim 1, wherein the first classifier neural network includes multiple first classifier sub-neural networks and the watermark classifier neural network includes multiple watermark classifier sub-neural networks, wherein each watermark classifier sub-neural network is trained using a different watermark dataset set from other watermark classifier sub-neural networks; and

further comprising:

generating pairs from the multiple first classifier sub-neural networks and the multiple watermark classifier sub-neural networks, each pair including a first classifier sub-neural network and a watermark classifier sub-neural network;

determining, using the multiple watermark classifier sub-neural networks and the data sample, watermark classification labels for the watermark; and

wherein determining the data sample as the adversarial data sample further comprises identifying that the classification labels for the watermark from the multiple watermark classifier sub-neural networks are above a first configurable threshold or below a watermark configurable threshold.

6. The method of claim 1, further comprising:

training the watermark classification neural network using the watermark dataset, wherein the watermark dataset includes watermarks and watermark classification labels, one classification label for one watermark.

7. The method of claim 1, further comprising:

training the first classifier neural network using the first dataset wherein the first dataset includes non-adversarial data, and the watermark dataset.

8. The method of claim 1, further comprising generating the watermark dataset using multiple non-overlapping gaussian distributions, each distribution associated with a unique class identifier.

9. The method of claim 8, further comprising:
 - identifying a vector from a non-overlapping gaussian distribution in the non-overlapping gaussian distributions;
 - generating, using a generator neural network, the watermark from the vector; and
 - generating, using the watermark classifier neural network, the classification label for the watermark, wherein the classification label is associated with a unique class identifier of the vector.
10. The method of claim 8, further comprising:
 - initializing the first classifier neural network with weights from the watermark classifier neural network.
11. The method of claim 8, wherein a structure of the first classifier neural network is the same as a structure of the watermark classifier neural network.
12. A system for determining an attack on a neural network, the system comprising:
 - a non-transitory memory storing instructions; and
 - one or more hardware processors coupled to the non-transitory memory and configured to read the instructions from the non-transitory memory to cause the system to perform operations comprising:
 - receive, a data sample at a first classifier neural network and at a watermark classifier neural network, wherein the first classifier neural network is trained using a first dataset and a watermark dataset;
 - determine, using the first classifier neural network and the data sample, a classification label for the data sample;
 - determine, using the watermark classifier neural network and the data sample, a classification label for the watermark; and
 - determine that the data sample is an adversarial data sample associated with the attack on the neural network based on the classification label for the data sample and the classification label for the watermark.

13. The system of claim 12, wherein to determine that the data sample is the adversarial data sample the system further performs operations comprising:
- determine that a value of the classification label for the data sample is coupled to a value of the classification label for the watermark.
14. The system of claim 12, wherein to determine that the data sample is the adversarial data sample the system further performs operations comprising:
- determine, using a separator classifier neural network and the classification label for the watermark that the data sample is the adversarial data sample.
15. The system of claim 14, wherein the system further performs operations comprising:
- train the watermark classifier neural network by passing at the first dataset through the watermark classifier neural network and generating the watermark classification labels based on the first dataset; and
 - train the separator classifier neural network using the watermark classification labels generated by the watermark classifier neural network.
16. The system of claim 12, wherein the first classifier neural network includes multiple first classifier sub-neural networks and the watermark classifier neural network includes multiple watermark classifier sub-neural networks, wherein each watermark classifier sub-neural network is trained using a different watermark dataset set from other watermark classifier sub-neural networks; and
- further comprising:
 - generate pairs from the multiple first classifier sub-neural networks and the multiple watermark classifier sub-neural networks, each pair including a first classifier sub-neural network and a watermark classifier sub-neural network;
 - determine, using the multiple watermark classifier sub-neural networks and the data sample, watermark classification labels for the watermark; and
 - wherein determining the data sample as the adversarial data sample further comprises identifying that the classification labels for the watermark from the multiple

watermark classifier sub-neural networks are above a first configurable threshold or below a watermark configurable threshold.

17. The system of claim 12, system to perform operations comprising:
train the watermark classification neural network using the watermark dataset, wherein the watermark dataset includes watermarks and watermark classification labels, one classification label for one watermark.
18. The system of claim 12, system to perform operations comprising:
train the first classifier neural network using the first dataset wherein the first dataset includes non-adversarial data, and the watermark dataset.
19. The system of claim 12, system to perform operations comprising:
identify a vector from a non-overlapping gaussian distribution in multiple non-overlapping gaussian distributions;
generate, using a generator neural network, the watermark from the vector; and
generate, using the watermark classifier neural network, the classification label for the watermark, wherein the classification label is associated with a unique class identifier of the vector.
20. A non-transitory machine-readable medium having stored thereon machine-readable instructions executable to cause a machine to perform operations for determining an attack on a neural network, the operations comprising:
receiving, a data sample at a first classifier neural network and at a watermark classifier neural network, wherein the first classifier neural network and the watermark classifier neural network are trained using a first dataset and a watermark dataset;
determining, using the watermark classifier neural network and the data sample, a classification label for the watermark; and
determining, using the separate classifier neural network that the classification label for the watermark is associated with the data sample that is an adversarial data sample.