Neural Network Watermarking Technical Specification

2023/01/11 T08:00 UTC

http://nnw.mpai.community

MPAI. community

About MPAI

Moving Picture, Audio, and Data Coding by Artificial Intelligence.

International, unaffiliated, non-profit SDO.

Developing AI-based data coding standards.

With clear Intellectual Property Rights licensing frameworks.

MPAI’s AI standardisation is “component-based”

An AI application:

- Subdivided in smaller components: AI modules (AIM).

- Aggregated in one or more AI workflows (AIW).

- Executed in a standard environment (AIF).

MPAI’s results so far

1 foundational standard: AI Framework (MPAI-AIF)

3 application standards :
- Context-based Audio Enhancement (MPAI-CAE)
- Compression and Underst. of Financial Data (MPAI-CUI)
- Multimodal Conversation (MPAI-MMC)

1 system standard: Governance of the MPAI Ecosystem (MPAI-GME).

Current MPAI activities

About to approve

Neural Network Watermarking (MPAI-NNW).

Extending

- *MPAI-AIF*
- *MPAI-CAE*
- *MPAI-MMC*

Developing

Avatar Representation and Animation (MPAI-ARA)

Six exploratory activities

- *AI Health*
- *Connected Autonomous Vehicles*
- *Server-based Predictive Multiplayer Gaming*
- *AI-based End-to-End Video Coding*
- *AI-Enhanced Video Coding*
- *XR Venues*

Adopted as IEEE standards

- *MPAI-AIF – 3301-2022*
- *MPAI-CAE – 3302-2022*
- *MPAI-MMC – 3300-2022*
- *MPAI-CUI (on its way)*

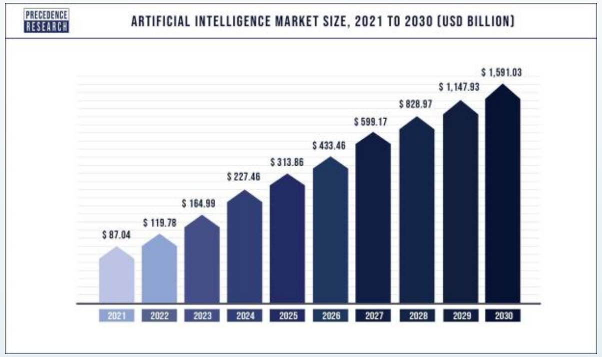
MPAI. community

Neural Networks

- Deployed in an increasing variety of domains
- Continuously renewed (industry & academia)
- At the heart of various autonomous systems:
  - Autonomous robots
  - Unmanned vehicles
- Deployed in more and more critical domains:
  - Medical decisions
  - Autonomous vehicle with passengers

Why watermarking is useful for Neural Networks

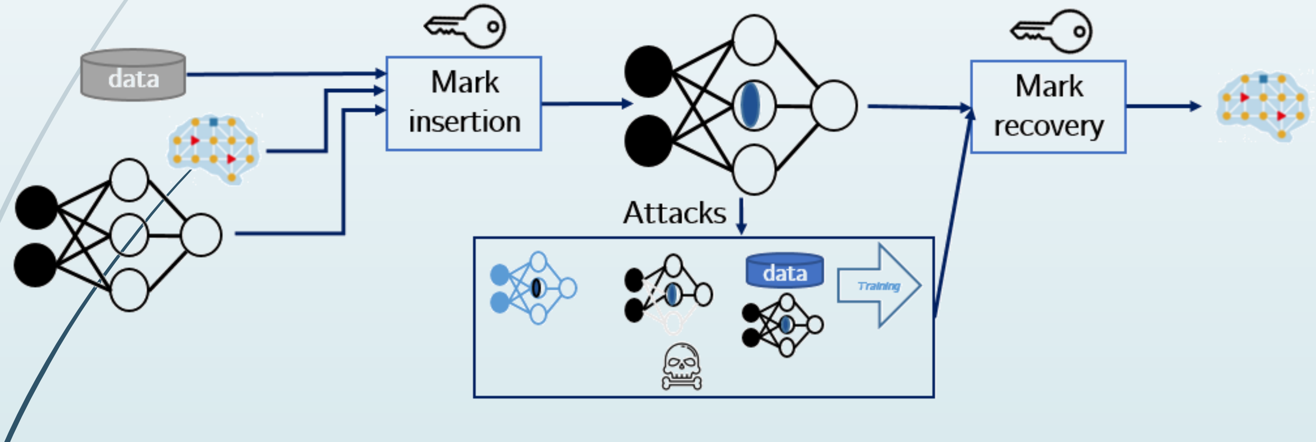
- Machine learning is a costly field:
  - Buying AI solution ranges from $ 6000 to $300.000
  - Renting a pre-built module costs around $ 40.000/year
- An AI solution could:
  - use multiple alternative Neural Networks to provide an inference ➡ identifying the one that actually produced the inference is important
  - be shared among multiple users ➡ keeping track of this process is useful
  - be altered or maliciously attacked ➡ identifying such modifications avoid faulty functioning

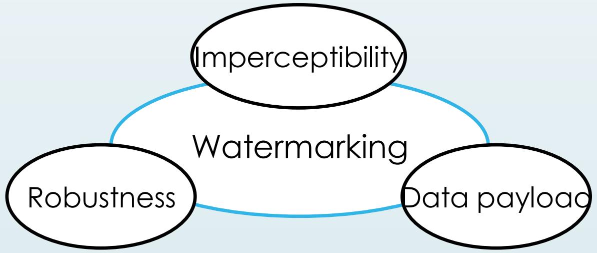


- **Ensuring traceability and integrity of Neural Networks becomes mandatory**

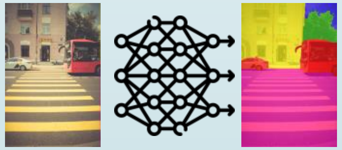
Neural Network Watermarking

- Watermarking provides tools allowing to **imperceptibly** and **persistently** insert some **data** into original content





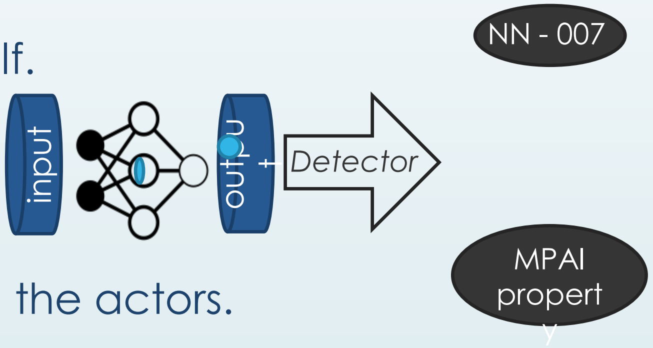
- Watermarking Neural Network is a new challenge:
  - Watermark insertion is no longer **static** but **dynamic**, as the watermark can also be:
    - *inserted during training*
    - *detected from inference*
  - Evaluation of watermark impact on inference is much more complex than on multimedia quality



Use cases

- *Identify an NN*

  The retrieved data conveys information about the NN itself.

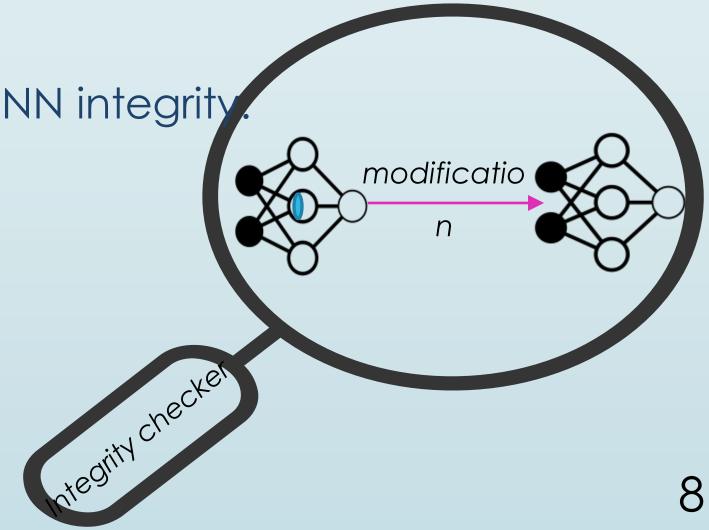


- *Identify the actors of an NN*

  The retrieved data conveys information about some or all the actors.

- *Verify the integrity of an NN*

  The retrieved data conveys information about NN integrity.



Scope of the MPAI-NNW Technical Specification

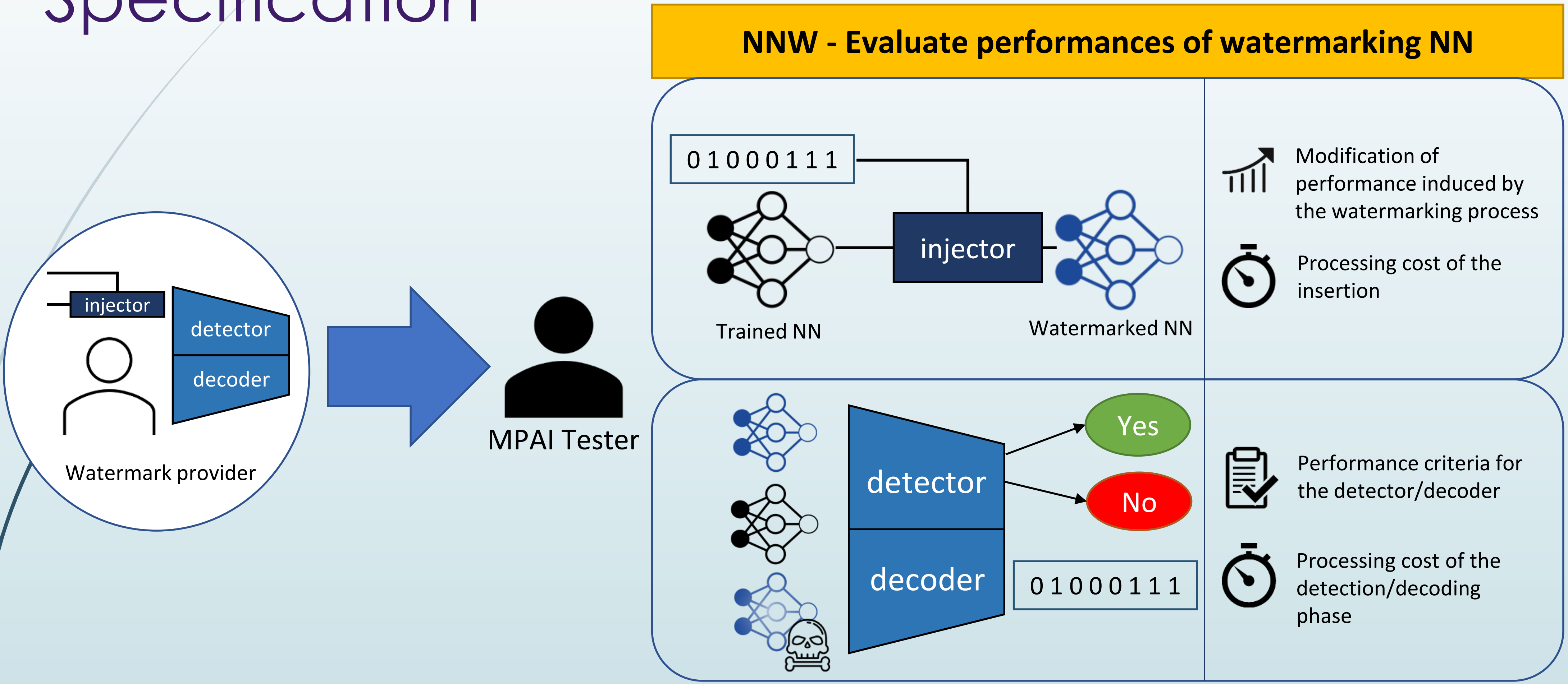
MPAI-NNW specifies methodologies to evaluate the following aspects of a neural network watermarking technology:

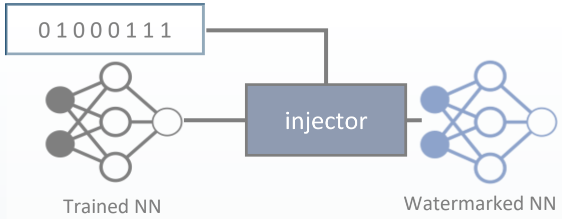
1. **The impact on the performance of a watermarked neural network and/or on its inference.**

2. **The ability of a neural network watermarking detector/decoder to detect/decode a payload when the watermarked neural network has been modified.**

3. **The computational cost of injecting, detecting or decoding a payload in the watermarked neural network.**

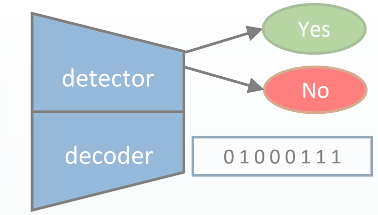
Overview of MPAI-NNW Technical Specification


injector
detector
decoder
Watermark provider
MPAI Tester
NNW - Evaluate performances of watermarking NN
01000111
injector
Trained NN
Watermarked NN
Modification of performance induced by the watermarking process
Processing cost of the insertion
detector
decoder
Yes
No
01000111
Performance criteria for the detector/decoder
Processing cost of the detection/decoding phase



1) Imperceptibility evaluation

- Define a pair of training and testing datasets, with a size at least 10 times larger than the number of trainable parameters.
- Apply the watermark to a set of unwatermarked NNs trained on the task.
- Feed the unwatermarked and watermarked NNs on the test dataset.
- Measure the task-dependent quality of the produced inference.



2) Robustness evaluation

- Define a pair of training and testing datasets, with a size at least 10 times larger than the number of trainable parameters.
- Apply the watermark to a set of unwatermarked NNs trained on the task
- Select and apply one modification (attack):
  - Gaussian noise addition, L1 pruning, random pruning, quantization, fine-tuning, knowledge distillation or watermark overwriting
- Evaluate the Robustness of the detector or decoder

Processing cost of the insertion

Processing cost of the detection/decoding phase

3) Computational cost evaluation

The following four elements shall be used to characterize the injection process:

- The memory footprint
- The time to execute the operation required by one epoch normalized according to the number of batches processed in one epoch
- In case injection is done concurrently with network training, the number of epochs required to insert the watermark
- The time for the watermarked neural network to compute an inference

Two elements shall be used to characterize the detection/decoding process:

- The memory footprint
- The total duration

Next steps

- Please send comments on the Technical Specification document the secretariat [mailto:secretariat@mpai.community] until 2023/01/23 23:59UTC
- We expect the NNW standard to be approved on 2023/01/25

Join the fun,
build the future!

MPAI.
community

https://www.mpai.community/

more about NNW at http://nnw.mpai.community/