On the Devastating E□ects of Single–Task Data Poisoning in Continual Learning

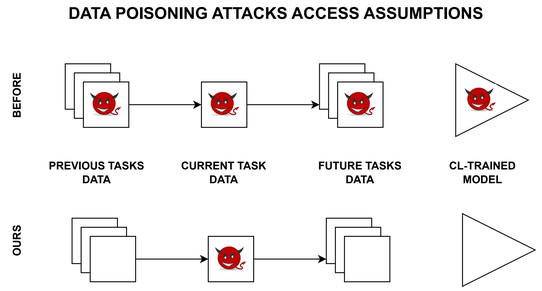
Stanisław Pawlak [1] Bartłomiej Twardowski [2,3] Tomasz Trzciński [1,2,4] Joost van de Weijer [3]

[1]Warsaw University of Technology [2]IDEAS NCBR [3]Computer Vision Center, Universitat Autonoma de Barcelona [4]Tooploox

Intro & Motivation

- Security threats are an under-explored field within Continual Learning.
- Data poisoning attacks are based on the manipulation of the training data aimed at degradation of model performance at inference time.
- In CL training is performed in multiple stages, possibly on the stream of data from various sources and tasks are usually learned in isolation. These factors make data poisoning in CL an important threat to address.

Most of the attacks use unrestricted threat model:



Contributions

- We propose a more realistic adversarial setup for CL, where adversary data access is restricted to one task in the sequence, with no knowledge of previous tasks or the trained model.
- We show that even under this restricted threat model data poisoning attacks can have devastating e□ects.
- To our best knowledge, we are first to investigate the consequences of poisoning on future tasks performance, and show an attack that increase catastrophic forgetting and impede future training at the same time.

TLDR

We introduce a Single-Task Poison (STP) setup to investigate the e□ect of data poisoning attacks on CL exemplar-free methods. It is more restrictive regarding adversary access to data and model. STP attacks cause a performance drop for non-poisoned tasks within the CL sequence.

Figure 1: **The Single-Task Poison (STP) setting.** We propose a new, more realistic CL setup for data poisoning attacks, where an adversary has only access to a single task with no knowledge about other tasks in the sequence and no access to the trained model. Despite this constraints, even simple STP attacks may strongly a□ect the stability and the plasticity of the model.

Data Poisoning Attacks

Unlike the joint training, in CL poisoning one task data may a□ect previous and future tasks performance:

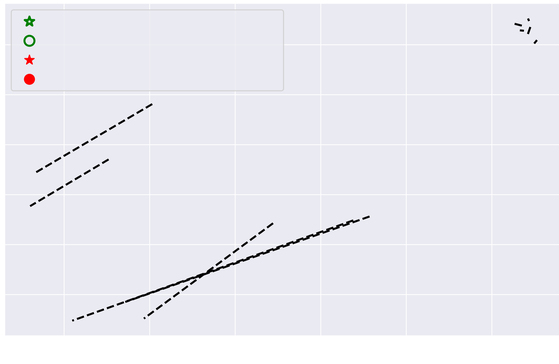


Figure 2: **Joint vs CL poisoning.**

Figure 3: **Base and Bait attacks.**

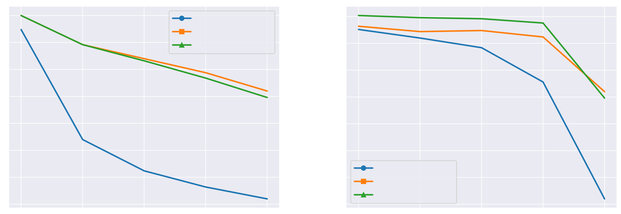


Figure 4: **Impact of corruption severity and poison data rate.**