Contribution ID: **66** Type: **not specified**

AI Friendly Hacker : when an AI reveals more than it should...

Tuesday, 17 September 2024 12:00 (30 minutes)

The aim of AI based on machine learning is to generalize information about individuals to an entire population. And yet...
- Can an AI leak information about its training data?
- Since the answer to the first question is yes, what kind of information can it leak?
- How can it be attacked to retrieve this information?

To emphasize AI vulnerability issues, Direction Générale de l'Armement (DGA, member of MoD in France) proposed a challenge on confidentiality attacks based on two tasks:
- Membership Attack: An image classification model has been trained on part of the FGVC-Aircraft open-access dataset. The aim of this challenge is to find, from a set of 1,600 images, those used for training the model and those used for testing.
- Forgetting attack: The model supplied, also known as the "export" model, was refined from a so-called "sovereign" model. The sovereign model has certain sensitive aircraft classes (families) which have been removed and replaced by new classes. The aim is to find which of a given set of classes have been used to train the sovereign model, using only the weights of the export model.

Friendly Hackers team of ThereSIS win the two tasks. During this presentation, we will present how we did it and what lessons we learned during this fascinating challenge.

Type of presentation

Talk

Classification

Both methodology and application

Keywords

Machine Learning, Privacy, AI vulnerability, Machine Unlearning

Primary authors: Dr HÉLIOU, Alice (Thales SIX GTS France); Dr MORISSE, Baptiste; Dr HUYNH, Cong Bang (Thales SIX GTS France); Dr LAMPE, Rodolphe (Thales SIX GTS France); THOUVENOT, Vincent (Thales SIX GTS France)

Presenter: THOUVENOT, Vincent (Thales SIX GTS France)

Session Classification: frEnbis invited session: Deep learning in industry

Track Classification: Other/ Special/ Invited