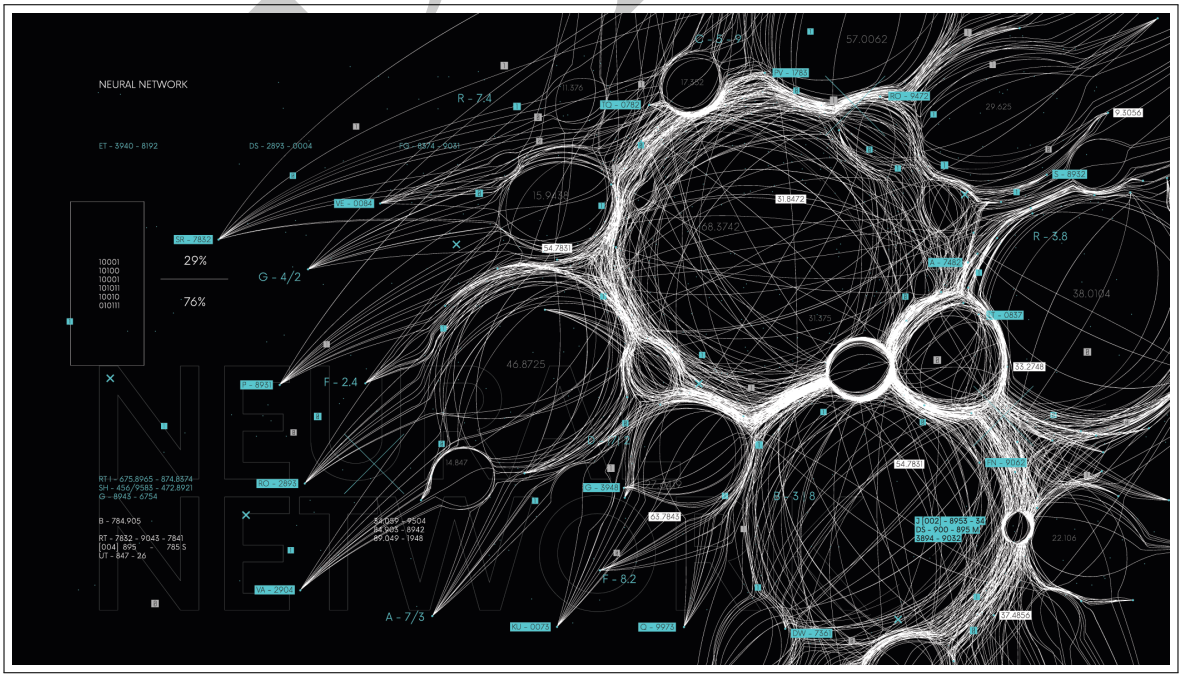
Institut National des Sciences Appliquées
Université de Rennes

Étude d'Attaques par Inférence d'Appartenance (MIA)

Une première approche avec le concours Snake Strikes Back

Auteurs:
Thomas Aubin
Selyan Da Silva
Moussa Ouassou
Émile Peltier

Responsable de projet :
Cédric Eichler
Créateurs de la compétition :
Tristan Allard
Mathias Bernard



Résumé en quelques lignes du projet
Mots-clés : mot1, mot2, mot3, mot4, mot5

November 27, 2024

Abstract

Machine Learning could be correctly defined as the Science of Artificial Intelligence (in the sense of algorithms capable of imitating some behaviour until now considered as strictly human). This activity-growing topic is in fact composed of numerous and complex scientific domains. Among them can be found of course statistics, but also other mathematics subdomain like probabilities, linear algebra, numerical analysis and optimization, as well as Computer Science skills, to cite Data Science, high-level programmation and complexity theory. Plenty of Machine Learning algorithms have been designed and implemented for the last two decades, and they all have but one thing in common : they use tremendous amounts of data to be operational. Data for Machine Learning Models can basically be seen as fuel for cars : nothing happens until you have provided enough of it to your machine. In regard to the surprisingly fast ascent of AI-usage in civil and industrial activities, usage of personal data have also risen up as a global concern, since many models are accused of using data they're not allowed to, or in the contrary, to leak the data used in the training phases to their users.
As a consequence of this problem, a whole new research field in Machine Learning is born somewhere around 2015, called Adversarial Machine Learning. It basically consists in attacking Machine Learning Algorithms to understand their behavior, and especially which datasets they are training on.

Contents

List of Figures

List of Tables

Liste des Équations

Table des éléments de code

Introduction

Bien que le projet ait pour coeur la participation à la compétition, celui-ci a nécessité un important travail de montée en compétences et de documentation en Machine Learning pour l'ensemble du groupe, ce domaine n'étant que peu abordé à ce stade de la formation. C'est pourquoi la partie opérationnelle et technique du projet est précédée d'une part d'un court travail de bibliographie ayant pour visée la synthèse des connaissances mathématiques et algorithmiques indispensables à la participation au concours, et d'autre part par une présentation des tenants et aboutissants du concours, laquelle prend soin d'expliquer le plus finement possible les données sur lesquelles nous nous entraînons ainsi que le modèle attaqué.

Part I

Notions d'*Adversarial Machine Learning*

Chapter I

Introduction aux concepts utiles d'Intelligence Artificielle

I.1 Réseaux de neurones et Deep Learning

I.1.1 Sous-section 1.3.1

I.1.1.1 Sous-sous-section 1.3.1.1

I.1.1.2 Sous-sous-section 1.3.1.2

I.1.1.3 Sous-sous-section 1.3.1.3

I.1.2 Sous-section 1.3.2

I.1.2.1 Sous-sous-section 1.3.2.1

I.1.2.2 Sous-sous-section 1.3.2.2

I.1.2.3 Sous-sous-section 1.3.2.3

I.1.3 Un modèle à deux réseaux : le *Generative Adversarial Network* (GAN)

I.1.3.1 Sous-sous-section 1.3.3.1

I.1.3.2 Sous-sous-section 1.3.3.2

I.1.3.3 Sous-sous-section 1.3.3.3

Chapter II

Chapitre II

II.1 Section II.1

II.1.1 Sous-section II.1.1

II.1.1.1 Sous-sous-section II.1.1.1

II.1.1.2 Sous-sous-section II.1.1.2

II.1.1.3 Sous-sous-section II.1.1.3

II.1.2 Sous-section II.1.2

II.1.2.1 Sous-sous-section II.1.2.1

II.1.2.2 Sous-sous-section II.1.2.2

II.1.2.3 Sous-sous-section II.1.2.3

II.1.3 Sous-section II.1.3

II.1.3.1 Sous-sous-section II.1.3.1

II.1.3.2 Sous-sous-section II.1.3.2

II.1.3.3 Sous-sous-section II.1.3.3

II.2 Section II.2

II.2.1 Sous-section II.2.1

II.2.1.1 Sous-sous-section II.2.1.1

II.2.1.2 Sous-sous-section II.2.1.2

II.2.1.3 Sous-sous-section II.2.1.3

II.2.2 Sous-section II.2.2

II.2.2.1 Sous-sous-section II.2.2.1

II.2.2.2 Sous-sous-section II.2.2.2

II.2.2.3 Sous-sous-section II.2.2.3

II.2.3 Sous-section II.2.3

II.2.3.1 Sous-sous-section II.2.3.1

II.2.3.2 Sous-sous-section II.2.3.2

II.2.3.3 Sous-sous-section II.2.3.3

II.3 Section II.3

II.3.1 Sous-section II.3.1

II.3.1.1 Sous-sous-section II.3.1.1

II.3.1.2 Sous-sous-section II.3.1.2

II.3.1.3 Sous-sous-section II.3.1.3

II.3.2 Sous-section II.3.2

II.3.2.1 Sous-sous-section II.3.2.1

II.3.2.2 Sous-sous-section II.3.2.2

II.3.2.3 Sous-sous-section II.3.2.3

II.3.3 Sous-section II.3.3

II.3.3.1 Sous-sous-section II.3.3.1

II.3.3.2 Sous-sous-section II.3.3.2

II.3.3.3 Sous-sous-section II.3.3.3

Chapter III

Attaques par Inférences d'Appartenance : contextualisation du projet

Part II

Le concours *Snake Strikes Back* : position du problème

Chapter IV

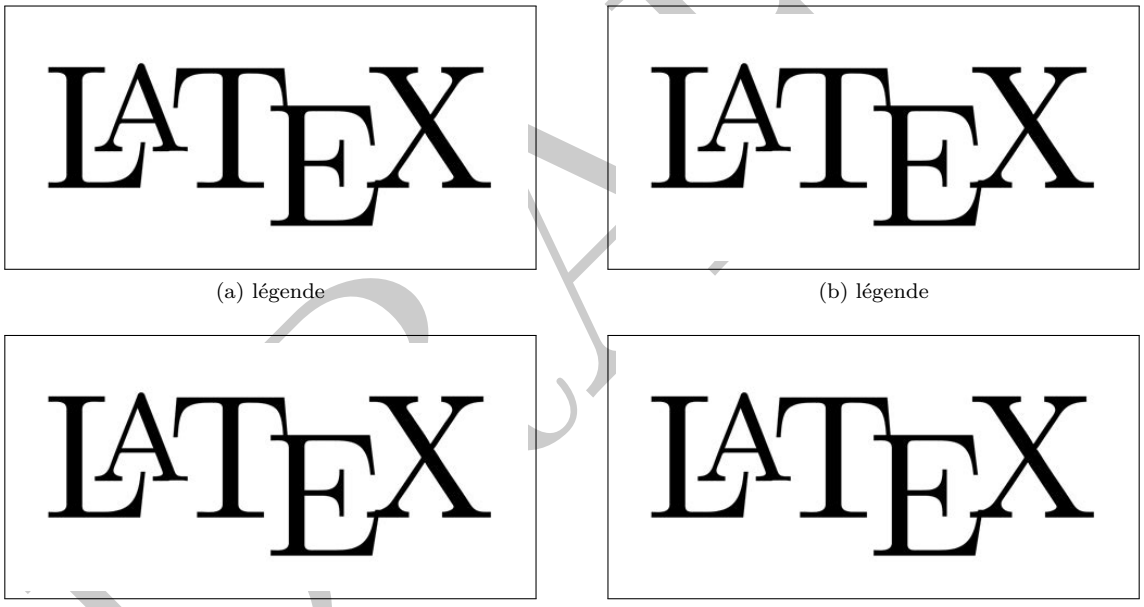
Contexte et enjeux de la compétition

IV.1 Section IV.1

IV.1.1 Sous-section IV.1.1

IV.1.1.1 Sous-sous-section IV.1.1.1

IV.1.1.2 Sous-sous-section IV.1.1.2



(a) légende (b) légende

(c) légende (d) légende

Figure IV.1: Exemple de figure avec plusieurs images

IV.1.1.3 Sous-sous-section IV.1.1.3

Chapter V

Parcours des ressources fournies

V.1 Processus d'installation : un peu de beta-testing et de documentation d'erreurs

Concours en Beta

V.1.1 Sous-section

V.1.1.1 Sous-sous-section

V.1.1.2 Sous-sous-section

V.1.1.3 Sous-sous-section

V.1.2 Sous-section

V.1.2.1 Sous-sous-section

V.1.2.2 Sous-sous-section

V.1.2.3 Sous-sous-section

V.1.3 Sous-section

V.1.3.1 Sous-sous-section

V.1.3.2 Sous-sous-section

V.1.3.3 Sous-sous-section

Chapter VI

DoppelGANger : un générateur de séries temporelles puissant ... mais attaquable

VI.1 Les hyperparamètres du modèle

VI.1.1 Sous-section

VI.1.1.1 Sous-sous-section

VI.1.1.2 Sous-sous-section

VI.1.1.3 Sous-sous-section

VI.1.2 Sous-section

VI.1.2.1 Sous-sous-section

VI.1.2.2 Sous-sous-section

VI.1.2.3 Sous-sous-section

VI.1.3 Sous-section

VI.1.3.1 Sous-sous-section

VI.1.3.2 Sous-sous-section

VI.1.3.3 Sous-sous-section

Part III

Attaque d'un modèle de Machine Learning : méthodologie et raisonnements

Chapter VII

Création de *Shadow Models* pour reproduire le comportement étudié

VII.1 Critères déterminants dans la construction du modèle

VII.1.1 Le problème du surapprentissage ou *overfitting*

VII.1.1.1 Sous-sous-section

VII.1.1.2 Sous-sous-section

VII.1.1.3 Choix de la métrique

VII.1.2 Sélection des données d'entraînement

VII.1.2.1 Données synthétiques générées par le modèle attaqué

VII.1.2.2 Données basées sur des interprétations statistiques des données d'entrée

VII.1.2.3 Données considérées comme similaires aux données d'entrée

VII.1.3 Sous-section

VII.1.3.1 Sous-sous-section

VII.1.3.2 Sous-sous-section

VII.1.3.3 Sous-sous-section

VII.2 Phase d'entraînement des modèles

VII.3 Comparaison des comportements entre les modèles

Chapter VIII

Quelques pistes pour aller plus loin

VIII.1 Envisager un entraînement non supervisé ?

VIII.1.1 Sous-section

VIII.1.1.1 Sous-sous-section

VIII.1.1.2 Sous-sous-section

$$a + b = c \tag{VIII.1}$$

VIII.1.1.3 Sous-sous-section

VIII.1.2 Sous-section

VIII.1.2.1 Sous-sous-section

VIII.1.2.2 Sous-sous-section

VIII.1.2.3 Sous-sous-section

VIII.1.3 Sous-section

VIII.1.3.1 Sous-sous-section

VIII.1.3.2 Sous-sous-section

VIII.1.3.3 Sous-sous-section

VIII.2 Section

VIII.2.1 Sous-section

VIII.2.1.1 Sous-sous-section

VIII.2.1.2 Sous-sous-section

VIII.2.1.3 Sous-sous-section

VIII.2.2 Sous-section

VIII.2.2.1 Sous-sous-section

VIII.2.2.2 Sous-sous-section

VIII.2.2.3 Sous-sous-section

VIII.2.3 Sous-section

VIII.2.3.1 Sous-sous-section

VIII.2.3.2 Sous-sous-section

VIII.2.3.3 Sous-sous-section

VIII.3 Section

VIII.3.1 Sous-section

VIII.3.1.1 Sous-sous-section

VIII.3.1.2 Sous-sous-section

VIII.3.1.3 Sous-sous-section

VIII.3.2 Sous-section

VIII.3.2.1 Sous-sous-section

VIII.3.2.2 Sous-sous-section

VIII.3.2.3 Sous-sous-section

VIII.3.3 Sous-section

VIII.3.3.1 Sous-sous-section

VIII.3.3.2 Sous-sous-section

VIII.3.3.3 Sous-sous-section

Chapter IX

Synthèse des résultats

IX.1 Tâche 1

IX.1.1 Sous-section

IX.1.1.1 Sous-sous-section

IX.1.1.2 Sous-sous-section

IX.1.1.3 Sous-sous-section

IX.1.2 Sous-section

IX.1.2.1 Sous-sous-section

IX.1.2.2 Sous-sous-section

IX.1.2.3 Sous-sous-section

IX.1.3 Sous-section

IX.1.3.1 Sous-sous-section

IX.1.3.2 Sous-sous-section

IX.1.3.3 Sous-sous-section

IX.2 Tâche 2

IX.2.1 Sous-section

IX.2.1.1 Sous-sous-section

IX.2.1.2 Sous-sous-section

IX.2.1.3 Sous-sous-section

IX.2.2 Sous-section

IX.2.2.1 Sous-sous-section

IX.2.2.2 Sous-sous-section

IX.2.2.3 Sous-sous-section

IX.2.3 Sous-section

IX.2.3.1 Sous-sous-section

IX.2.3.2 Sous-sous-section

IX.2.3.3 Sous-sous-section

IX.3 Tâche 3

IX.3.1 Sous-section

IX.3.1.1 Sous-sous-section

IX.3.1.2 Sous-sous-section

IX.3.1.3 Sous-sous-section

IX.3.2 Sous-section

IX.3.2.1 Sous-sous-section

IX.3.2.2 Sous-sous-section

IX.3.2.3 Sous-sous-section

IX.3.3 Sous-section

IX.3.3.1 Sous-sous-section

IX.3.3.2 Sous-sous-section

IX.3.3.3 Sous-sous-section

IX.4 Tâche 4

IX.4.1 Sous-section

IX.4.1.1 Sous-sous-section

IX.4.1.2 Sous-sous-section

IX.4.1.3 Sous-sous-section

IX.4.2 Sous-section

IX.4.2.1 Sous-sous-section

IX.4.2.2 Sous-sous-section

IX.4.2.3 Sous-sous-section

IX.4.3 Sous-section

IX.4.3.1 Sous-sous-section

IX.4.3.2 Sous-sous-section

IX.4.3.3 Sous-sous-section

Conclusion

Part IV

Annexes

Annexe 1 : Programmes conçus par l'équipe

Annexe 2 : Retour d'expérience et chronologie du projet

Annexe 3 : Framework utilisé

Bibliography

[1] *Adversarial Machine Learning*. Page Wikipedia de l'Adversarial Machine Learning. Nov. 2024. URL: https://en.wikipedia.org/wiki/Adversarial_machine_learning#Adversarial_attacks_and_training_in_linear_models.

[2] Tristan Allard and Mathias Bernard. "Snakes Strikes Back". In: (Oct. 2024).

[3] Tatev Aslanyan. *Machine Learning in 2024 – Beginner's Course*. Feb. 2024. URL: https://www.youtube.com/watch?v=bmmQA8A-yUA&t=1769s.

[4] Author. *Membership inference attacks from first principles*. How published. Some note. Month Year. URL: https://www.youtube.com/watch?v=1CNxfhMlk-A.

[5] author. *Comparing and Evaluating Datasets: A Simplified Guide*. Nov. 24, 2024. URL: https://www.markovml.com/blog/compare-datasets.

[6] Chloé-Agathe Azencott. *Introduction au Machine Learning*. (2nd). InfoSup. Dunod, Feb. 2022.

[7] *Generative adversarial network*. Page Wikipedia du modèle GAN. Nov. 2024. URL: https://en.wikipedia.org/wiki/Generative_adversarial_network.

[8] Benjamin JOURDAIN. *Probabilités et statistiques pour l'ingénieur*. Jan. 2018.

[9] Zinan Lin et al. "Using GANs for Sharing Networked Time Series Data : Challenges, Initial Promise, and Open Questions". In: (Jan. 2021). Présentation du modèle DoppelGANger. URL: https://arxiv.org/abs/1909.13403.

[10] *Machine Learning*. Page Wikipedia du Machine Learning. Nov. 2024. URL: https://en.wikipedia.org/wiki/Machine_learning.

[11] Boris Meinardus. *How I'd learn ML in 2024 (if I could start over)*. Youtube. 2024. URL: https://www.youtube.com/watch?v=gUmagAluXpk.

[12] *Overfitting*. Page Wikipedia de l'Overfitting. Nov. 2024. URL: https://en.wikipedia.org/wiki/Overfitting#Machine_learning.

[13] Reza Shokri. *Membership Inference Attacks against Machine Learning Models*. Vidéo de vulgarisation du papier du même nom. May 2017. URL: https://www.youtube.com/watch?v=rDm1n2gceJY&t=53s.

[14] Reza Shokri et al. "Membership Inference Attacks Against Machine Learning Models". In: ().