Rapport Pôle IA

Robustesse de l'IA face aux attaques adverses

Grégoire Desjonqueres
Aymeric Palaric
Victor Fernando Lopes De Souza
Amadou Sékou Fofana

Groupe 2.01
30 janvier 2022

Table des matières

Introduction

Nos clients sont les représentants de l'Institut de Recherche Technologique IRT-SystemX. Créé en 2012, cet institut a pour objectif de soutenir l'innovation en France. C'est une fondation de coopération scientifique entre grands groupes qui se positionne comme un accélérateur de la transformation numérique de l'industrie, des services et des territoires au moyen d'une innovation flexible, ouverte et collective. En 2020, un programme basé sur l'intelligence artificielle est mis en place.

Un total de 345 travaux a été publiés par l'IRT-SystemX, menés par 100 ingénieurs chercheurs et 22 doctorants. Implanté au cœur de l'Université Paris-Saclay, un des dix pôles les plus importants mondialement en innovation (selon le classement du MIT Technology Review), l'IRT-SystemX s'est également exporté à Lyon et à Singapour en 2017.

Nos interlocuteurs et intervenants de l'IRT-SystemX sont Mohamed IBN KHEDHER et Lucas MATTIOLI.

Chapitre 1

Etat de l'art

Une première partie de ce projet a été consacrée à la recherche et à l'étude de l'état de l'art concernant les critères et attaques sur les réseaux de neurones. Les premières définitions et concepts liés aux attaques adversariales nous ont été données par l'article de Naveed Akhtar et Ajmal Mian (2018) [1].
L'article *Understanding Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems* de Xingjun Ma et al. [14] montre que le domaine médical est particulièrement sensible aux attaques adversariales, mais qu'il est aussi paradoxalement facile de détecter si une image médicale donnée est perturbée ou non. Cet article nous a donc permis de comprendre l'aspect critique de la vulnérabilité des modèles aux attaques adverses. Il nous a aussi aidé à comprendre la faiblesse de certaines attaques, notamment celles se basant sur les gradients, qui sont facilement détectables dans le domaine médical.
L'article *A Survey on Security Threats and DefensiveTechniques of Machine Learning :A Data Driven View* de Qiang Liu et al. [12] quant à lui présente des attaques sur de nombreux types de modèles de Machine Learning (SVM, régression logistic, naive Bayes,...) et compare différentes attaques et méthodes de défense appliquées à chacun des ces modèles. Cet article nous a donc permis d'approfondir les définitions présentées précédemment.
Nous avons donc beaucoup étudié les plateformes de benchmarking disponibles, telles que l'*Adversarial Robustess Toolbox* (ART) (Adversarial Robustness Toolbox) [17], *CleverHans* (`https://github.com/cleverhans-lab/cleverhans`) [18], ou encore la plateforme *RobustBench* (`https://github.com/RobustBench/robustbench`) [4]. La plateforme CleverHans met à disposition un nombre assez restreint d'attaques, qui sont par ailleurs toutes disponibles sur la plateforme ART. RobustBench permet quant à lui de benchmarker un modèle selon différents critères simples, et vis-à-vis de peu d'attaques, là encore disponibles sur l'ART. En particulier, nous avons utilisé les attaques implémentées sur l'ART pour les déployer sur notre propre modèle.

L'ART met en effet à disposition 55 attaques, dont 38 sont des attaques par évasion, qui est le type d'attaque que nous devions étudier. De plus, le lien vers chacun les articles présentant chacune des attaques est disponible sur l'ART, en plus de celles décrites mathématiquement dans l'article *hreat of adversarial attacks on deep learning in computer vision : A survey* [1].

En ce qui concerne les critères de robustesse, nous avons recensé dans certains articles scientifiques les différentes métriques utilisées, qui constituent des critères différents pour qualifier la robustesse d'un réseau de neurones face à des attaques adverses. Mais pourquoi doit-on disposer de plusieurs critères de robustesse distincts pour envisager notre problème ? Pourquoi donc existe-t-il différentes métriques pour quantifier la robustesse d'un réseau de neurone face à des attaques adverses, et parmi elles, y en a-t-il une meilleure ? Nous avons une approche en largeur de notre problème : nous appliquons un nombre important d'attaques existantes sur un même réseau de neurones. Or certaines attaques sont incompatibles avec des métriques, c'est pourquoi il est nécessaire d'en utiliser d'autres différentes. La première métrique que nous avons retenue est la courbe de l'accuracy en fonction de l'intensité de l'attaque [6]. C'est une métrique couramment utilisée, mais elle exige déjà que l'attaque dispose d'un paramètre qu'on peut relier à l'intensité de cette dernière. Ensuite, l'article [7] expose l'intensité minimale de bruit provoquant une mauvaise prise de décision par le réseau de neurones. C'est une métrique qui ne fonctionne que sur les attaques agissant sur les entrées, ce qui est le cas dans notre étude. Cette métrique a une utilité concrète et pratique, puisqu'elle renseigne sur le moment où le réseau de neurone va se tromper (ce qui in fine intéresse les ingénieurs), mais n'indique pas la manière dont le réseau est perturbé, si c'est linéaire en l'intensité de l'attaque ou une autre relation. Deux autres critères, "the pointwise robustness" et l'adverserial frequency [2] étudie la robustesse d'un réseau de neurone suivant chacune de ses données. C'est une métrique qu'on détaillera plus tard dans ce rapport.

Chapitre 2

Travail réalisé

L'objectif de notre travail est d'étudier la robustesse d'un réseau de neurones face à des perturbations à l'aide de critères pour évaluer la capacité du réseau de neurones à garantir des prédictions correctes malgré les perturbations qu'on lui fait subir ; il s'agit alors d'identifier ces métriques. Pour cela nous avons commencé par faire une étude de la documentation sur la robustesse de l'IA face aux attaques adverses et nous avons alors identifié les attaques et les métriques de la littérature.

2.1 Etude des attaques

2.1.1 Déroulement de la tâche

Pour cette partie, le but était de sélectionner quelques attaques que l'on appliquerait à un réseau déjà entraîné. Une des consignes du client était que ces attaques se fassent sur des images que l'on donnera ensuite au réseau attaqué.
Lors de l'étude des 4 différents types d'attaque, nous avons étudié chacun un type d'attaque. Puis, pour l'étude plus spécifique des attaques par évasion, Aymeric, Victor et Amadou ont chacun selectionné et étudié plusieurs attaques disponibles sur l'ART [17]. Cette tâche a duré 4 semaines.

2.1.2 Eléments techniques

Il y'a quatre types d'attaques sur les réseaux de neurones :

1. **Les attaques par inférence** : Avec l'inférence, l'attaquant teste plusieurs requêtes sur le modèle et étudie l'évolution de son comportement dans le but de s'approprier le modèle ou certains de ses paramètres. Ainsi en récupérant les sorties du modèle, l'attaquant constitue un nouveau modèle très semblable(voire identique) au premier qu'il entraine à l'aide de toutes les données

qu'il a recuperées.Ainsi il dispose des informations pouvant mener à n'importe quelle prédiction du modèle et peut donc déjouer le modèle avec des attaques qui ne seront pas détectées.

2. **Les attaques par poisoning** : Ces attaques prennent pour cible la phase d'apprentissage du modèle. Elles induisent des données empoisonnées dans le processus d'apprentissage du réseau de neurones afin de modifier son comportement et ses prédictions.
3. **Les attaques par extraction** : Ces attaques suivent le même principe que celles par inférence, on envoie plusieurs requêtes au modèle afin de recupérer les données de sortie, reproduire le comportement du modèle et enfin recupérer ses données d'entrainement, qui sont souvent privées et confidentielles.
4. **Les attaques par évasion** : Elles ont pour but de perturber l'IA en apportant des perturbations imperceptibles aux données d'entrée comme la modification d'un pixel sur une image d'entrée d'un modèle de classsification.

Dans ce projet, le réseau de neurones qu'on devra attaquer est déjà entrainé, alors nous ne considérerons pour la suite que les attques par évasion qui portent sur le données d'entrée du modèle.
Nous avons donc utilisé 11 attaques par évasion disponibles sur l'ART :

— Attaque Wasserstein [19]
— Virtual Adversarial Method [15]
— Shadow Attack [8]
— Attaque NewtonFool [11]
— Fast Gradient Sign Method [9]
— Attaque Carlini&Wagner [3]
— Attaque DeepFool [16]
— Iterative Frame Saliency Attack [10]
— Attaque Auto-Projected Gradient Descent [5]
— Adversarial Patch [13]

2.2 Etude des critères

On se pose ensuite la question de savoir : Comment savoir à quel point un réseau de neurones est robuste face à une attaque donnée ?

2.2.1 Déroulement de la tâche

Le but était donc de trouver dans la littérature ou imaginer des critères pertinents pour l'étude de la robustesse d'un réseau de neurones face aux attaques précédemment implémentées.
Cette tâche a été effectuée par Grégoire et Amadou, et elle a duré 4 semaines, en parallèle de l'étude des attaques.

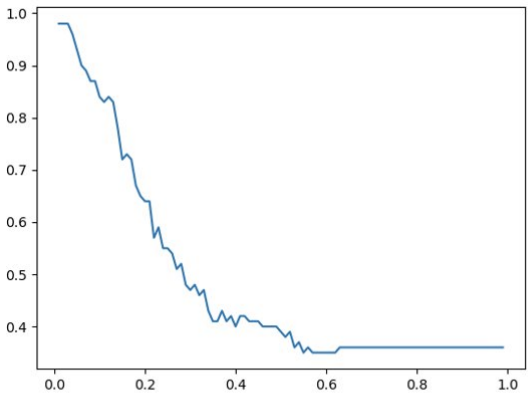
2.2.2 Eléments techniques

1. **Métriques de la littérature :**
   L'état de l'art nous a permis d'identifier plusieurs métriques :
   - **l'accuracy** : [6]
     Une métrique évidente, qu'on identifie facilement dans la littérature.Ainsi, on attaque le modèle considéré, et on regarde avec quelle précision il renvoie son nouveau résultat. C'est une métrique dont le calcul est implicement donné par le réseau de neurones mais elle ne prend pas en compte les paramètres de l'attaque tels que son intensité.
   - **l'intensité minimale de bruit faisant prendreau réseau de neurones une mauvaise décision** : [7]
     La littérature suggère également cette métrique en tant qu'indicateur assez fiable de l'évaluation de la robustesse et relativement simple à calculer surtout quand les attaques prennent en paramètre leurs intensités. On fait varier l'intensité de la perturbation et on regarde jusqu'où le réseau de neurones continue de faire des prédictions correctes. On peut alors se dire que plus cette métrique est grande pour un modèle plus ce dernier est robuste.
   - **The pointwise robustness** : [2]
     C'est intuitivement le fait d'étudier la robustesse en chaque donnée du modèle. Et un modèle est robuste en une donnée si une « une petite perturbation »n'affecte pas la classification de de ce dernier. Formellement on introduit une norme pour chaque écart entre la donnée réelle et la donnée perturbée et la robustesse ponctuelle correspond au minimum de ces normes pour chaque perturbation induite.

— **L'adversarial frequency :** [2]
Etant donné une perturbation, cette métrique mesure combien de fois le modèle échoue à être robuste en une donnée et plus cette fréquence plus le modèle est considéré comme non robuste.

2. **Métrique de la littérature mise en place :**
Après avoir identifié les attaques et les métriques de la litterature, on a attaqué plusieurs réseaux de neurones et on a essayé de mettre en place les métriques qui nous semblent pertinentes pour juger de la robustesse d'un réseau de neurones.

   (a) **La courbe précision/intensité perturbation** : [7]
   Cette courbe peut être assez indicative de la robustesse, on peut constater à partir d'elle comment la précision varie en fonction de l'intensité de l'attaque, de plus c'est une métrique assez simple à calculer pour les attaques ayant leur intensité comme par paramètre. C'est une métrique qu'on a mise en place pour certaines attaques : la figure suivante en donne une illustration pour l'attaque Carlini&Wagner.



Précision en fonction de l'intensité
(C&W)

   Avec ce tracé, on peut se faire une idée générale de la robustesse de ce réseau de neurones qui perd 38% en précison lorsque l'intensité du bruit passe de 0 à 0.2%.

2.3 Interface Graphique

2.3.1 Déroulement de la tâche

Un souhait ultérieur du client a été le déploiement des 2 tâches précédentes à l'aide d'une interface graphique. Celle-ci a été réalisée par Aymeric, pendant 2 semaines.

2.3.2 Eléments techniques

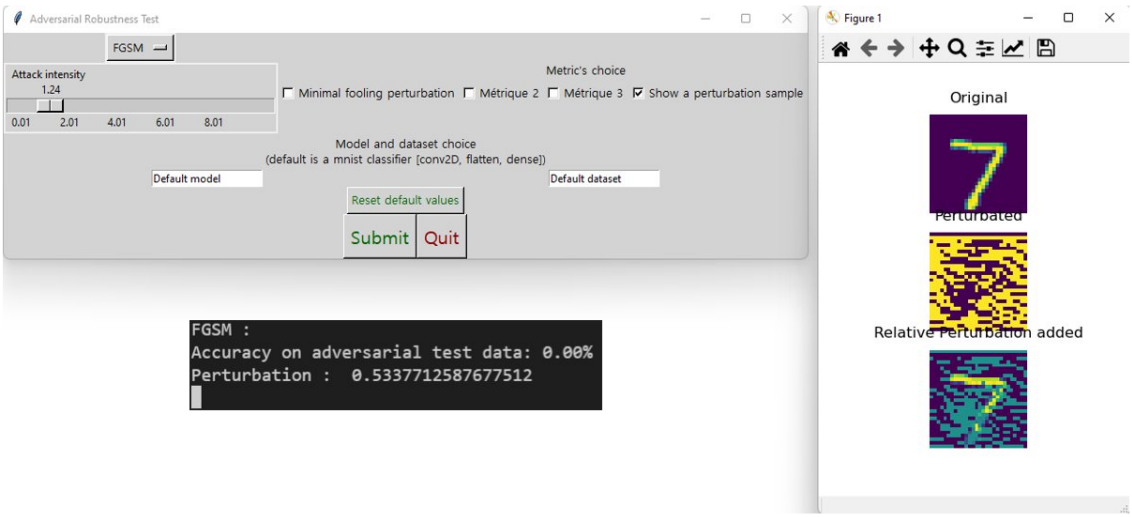
L'interface graphique a été réalisée sous TKinter, et elle permet le choix de nombreux paramètres :

— L'attaque à utiliser ;

— L'intensité de cette attaque ;

— Le ou les critères à évaluer

— Le choix du modèle et du dataset à utiliser, tout en fournissant un modèle et un dataset par défaut.



2.4 Livrables

L'étude des attaques et l'étude des critères ont tous deux été utilisées dans l'interface graphique, qui constitue un des livrables de ce projet. Le second livrable est un notebook détaillé présentant chacune des attaques et chacun des critères présentés précédemment. Ces deux livrables seront envoyés en même temps que ce rapport. Le recensement des critères de robustesse présents dans la littérature constitue également une partie des livrables attendus par le client. Le notebook, et l'interface graphique sont joints à ce rapport.

Chapitre 3

Conclusion et perspectives

Pour le client, ce projet a pu lui apporter un état de l'art en largeur des différentes attaques étudiées et pouvant être utilisées dans le cadre de la perturbation d'un réseau de neurones classifiant des images, et de leur plateformes de déploiement. Il lui apporte aussi une étude des différents critères de robustesse que l'on peut trouver dans la littérature, ainsi que leur facilité de mise en place et leur portée. Il leur a aussi été rendu possible une rapide étude de ces attaques et de ces critères à travers une interface graphique.

De notre côté, ce projet a été intéressant de par l'aspect "en largeur" de l'étude : en effet, dans tous nos projets précédents respectifs, nous ne nous étions confrontés qu'à des études en profondeur, et ce projet nous a permis de toucher du doigt la diversité du domaine des attaques adverses. Nous avons donc trouvé stimulant le fait de changer souvent de cadre d'étude, en étudiant tour à tour des attaques et des critères différents, plutôt que de se focaliser sur un seul élément. En outre, pour certains d'entre nous, ce projet a été d'autant plus intéressant parce qu'il a permis d'introduire le domaine des attaques adverses, que nous ne connaissions pas, avec tous les enjeux que celui-ci apporte.

En ce qui concerne le projet en lui-même, des perspectives d'approfondissement de l'étude peuvent être imaginées. On peut par exemple élargir le domaine des attaques, en étudiant des attaques appartenant aux 4 groupes (evasion, poisoning, inference, extraction), ce qui peut aussi s'accompagner d'une ouverture des modèles à d'autres applications, d'autres entrées, car le projet actuel ne s'intéresse qu'aux réseaux de neurones classifiant des images. Il serait aussi envisageable de développer davantage l'interface graphique en y incorporant toutes ces précédentes perspectives, et en y a joutant une dimension de benchmarking, pour permettre à l'utilisateur de comparer plusieurs modèles ou attaques selon davantage de critères, voire même

de transformer cette interface en plateforme de robustification, en utilisant des méthodes de défense qui auraient été étudiées au préalable.

Bibliographie

[1] N. Akhtar and A. Mian. Threat of adversarial attacks on deep learning in computer vision : A survey. *CoRR*, abs/1801.00553, 2018.

[2] O. Bastani, Y. Ioannou, L. Lampropoulos, D. Vytiniotis, A. V. Nori, and A. Criminisi. Measuring neural net robustness with constraints. *CoRR*, abs/1605.07262, 2017.

[3] N. Carlini and D. A. Wagner. Towards evaluating the robustness of neural networks. *CoRR*, abs/1608.04644, 2016.

[4] F. Croce, M. Andriushchenko, V. Sehwag, E. Debenedetti, N. Flammarion, M. Chiang, P. Mittal, and M. Hein. Robustbench : a standardized adversarial robustness benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

[5] F. Croce and M. Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. *CoRR*, abs/2003.01690, 2020.

[6] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller. A causal view on robustness of neural networks. *NeurIPS Proceedings*.

[7] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller. Adversarial attacks on deep neural networks for time series classifications. *CoRR*, abs/1903.07054, 2019.

[8] A. Ghiasi, A. Shafahi, and T. Goldstein. Breaking certified defenses : Semantic adversarial examples with spoofed robustness certificates. *CoRR*, abs/2003.08937, 2020.

[9] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples, 2015.

[10] N. Inkawhich, M. Inkawhich, Y. Chen, and H. Li. Adversarial attacks for optical flow-based action recognition classifiers. *CoRR*, abs/1811.11875, 2018.

[11] U. Jang, X. Wu, and S. Jha. Objective metrics and gradient descent algorithms for adversarial examples in machine learning. In *Proceedings of the 33rd Annual Computer Security Applications Conference*, ACSAC 2017, page 262–277, New York, NY, USA, 2017. Association for Computing Machinery.

[12] Q. Liu, P. Li, W. Zhao, W. Cai, and S. Yu. A survey on security threats and defensive techniques of machine learning : A data driven view. *IEEE Access*, 6 :12103–12117, 02 2018.

[13] X. Liu, H. Yang, L. Song, H. Li, and Y. Chen. Dpatch : Attacking object detectors with adversarial patches. *CoRR*, abs/1806.02299, 2018.

[14] X. Ma, Y. Niu, L. Gu, Y. Wang, Y. Zhao, J. Bailey, and F. Lu. Understanding adversarial attacks on deep learning based medical image analysis systems. *CoRR*, abs/1907.10456, 2019.

[15] T. Miyato, S.-i. Maeda, M. Koyama, K. Nakae, and S. Ishii. Distributional smoothing with virtual adversarial training. *arXiv preprint arXiv :1507.00677*, 2015.

[16] S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool : a simple and accurate method to fool deep neural networks. *CoRR*, abs/1511.04599, 2015.

[17] M.-I. Nicolae, M. Sinn, M. N. Tran, B. Buesser, A. Rawat, M. Wistuba, V. Zantedeschi, N. Baracaldo, B. Chen, H. Ludwig, I. Molloy, and B. Edwards. Adversarial robustness toolbox v1.2.0. *CoRR*, 1807.01069, 2018.

[18] N. Papernot, F. Faghri, N. Carlini, I. Goodfellow, R. Feinman, A. Kurakin, C. Xie, Y. Sharma, T. Brown, A. Roy, A. Matyasko, V. Behzadan, K. Hambardzumyan, Z. Zhang, Y.-L. Juang, Z. Li, R. Sheatsley, A. Garg, J. Uesato, W. Gierke, Y. Dong, D. Berthelot, P. Hendricks, J. Rauber, and R. Long. Technical report on the cleverhans v2.1.0 adversarial examples library. *arXiv preprint arXiv :1610.00768*, 2018.

[19] E. Wong, F. R. Schmidt, and J. Z. Kolter. Wasserstein adversarial examples via projected sinkhorn iterations. *CoRR*, abs/1902.07906, 2019.