Mis à jour 10 / 2019

Carte des Menaces sur l’IA

Introduction

En quoi le contrôle de l'IA est important et ce que cela vous coûtera si vous ne maîtrisez pas ces sujets.

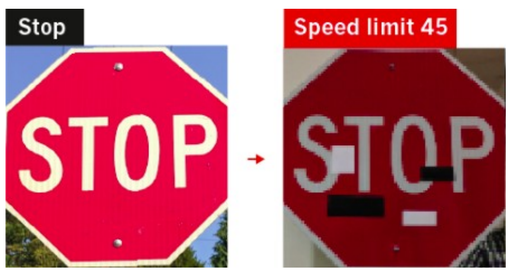
Le *Machine Learning* (*ML*), le *Deep Learning* (*DL*) et plus généralement l’*Intelligence Artificielle* (*IA*) sont en passe de devenir un rouage dans les process de la plupart des domaines technologiques. L'utilité de l'IA n'est plus à démontrer. Ces technologies sont actuellement des algorithmes en *boîte grise* du point de vue du propriétaire, car il est très difficile d’expliquer leur fonctionnement interne, d’en prévoir les limites et les cas extrêmes. Lors de la mise en production, ces systèmes peuvent, comme d’autres technologies analogues, souffrir des problèmes suivants :

- Mauvais usage volontaire (hacking) ou involontaire
- Fuite de données et problèmes de gouvernance des données
- Biais non-intentionnels et injustes (traitement de données relatives à l'homme, par exemple)
- Incapacité de traiter des données réelles imprévues

De plus, contrairement à la plupart des technologies, l’IA souffre souvent de la méfiance des utilisateurs. Ainsi, pour que notre société adopte cette technologie sans frictions, nous devons faire preuve de plus de maîtrise, de responsabilité et d’une plus grande capacité à rendre des comptes avec l’utilisation de l’IA. L'intelligence artificielle peut être un outil de grande prospérité, mais elle peut aussi devenir rapidement la source d'un désastre industriel et médiatique. En 2018, les scandales liés à l'IA n'ont jamais été aussi élevé en termes de nombre et de magnitude. Plus qu'un coût direct, ces scandales minent la confiance du public et des clients potentiels dans l’IA, ce qui ralentit sa progression, son développement et son adoption. C'est un coût d'opportunité énorme.

Hacking

Comme tout logiciel doté d'une interface publique, l'intelligence artificielle est soumise aux menaces de sécurité habituelles telles que le déni de service et l'exploitation de vulnérabilités. De plus, il peut être la cible de menaces sur son cœur algorithmique. Un échantillon d’entrée soigneusement conçu peut être soumis afin de contrôler la prédiction de l’IA. C'est ce qu'on appelle un **sample adversarial**.



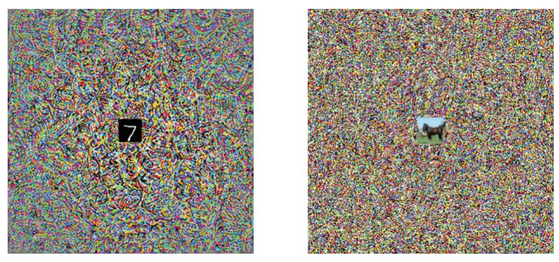
Un sample adversarial qui trompe la prédiction d'un algorithme de Deep Learning entraîné pour reconnaître les panneaux de signalisation.

Les algorithmes de *Deep Learning* possèdent une propriété intéressante du point de vue d'un attaquant: la **transférabilité**. C'est le phénomène que deux modèles de *DL* différents entrainés sur le même problème présentent des bornes de décision similaires. Un attaquant peut utiliser cette propriété pour entraîner un modèle de *DL* sur un problème similaire, puis trouver une attaque active sur son propre modèle. Cette attaque est susceptible de fonctionner sur le modèle cible sans connaître sa configuration ou ses paramètres.

Le **reprograming adversarial** est une nouvelle menace. Cette attaque détecte une perturbation unique, qui peut être ajoutée à toutes les entrées d'un modèle afin de permettre au modèle d'exécuter une tâche choisie par l'adversaire. Il s'agit d'un vol de puissance GPU / CPU.

Points clefs

- Votre IA a-t-elle une interface publique?
  - Il est sensible à des samples adversarials, à un adversarial reprograming, à un vol de modèle et à l’extraction de données.
- Avez-vous un contrôle parfait (accès / contenu) de votre jeu de données d’entraînement ?
  - Votre IA est sujette à un empoisonnement des données.
- Votre IA gère-t-elle les données relatives à l'humain?
  - Votre AI peut fonder ses décisions sur des préjugés injustes et peut entraîner des discriminations involontaires.



Adversarial reprograming qui transforme le modèle ImageNet en un classificateur MNIST ou un classificateur CIFAR-10.

Dans les deux cas précédents, nous avons expliqué ce que peut faire un attaquant externe. Dans le scénario suivant, l'attaquant a accès aux exemples d'apprentissage et modifie ou ajoute des exemples pour contrôler partiellement l'apprentissage du modèle. Il peut s’agir d’un sabotage interne ou l’attaquant a eu accès un accès au process interne de l’entreprise. On appelle cette attaque **empoisonnement de données**. L’attaquant a généralement pour objectif de déplacer la borne de décision du modèle à son avantage (détecteur de programme malveillant, par exemple).

Vol d’information confidentielle

L'intelligence artificielle étant étroitement liée au Big Data, dans le cas où ces données sont confidentielles ou privées, l’IA entraînée peut être un vecteur d'attaques. Les modèles d'apprentissage automatique «se souviennent» partiellement de leurs échantillons d'apprentissage. Cette capacité à mémoriser les données d'apprentissage peut être exploitée en testant si un échantillon fait partie de l'ensemble d'apprentissage, cette attaque est appelée **membership inference attack**. Dans le cas où des échantillons sont liés à une personne, tels que des données médicales ou financières, déterminer si les échantillons proviennent de l’entraînement du modèle ML constitue une menace pour la confidentialité. Des méta-données peuvent être divulguées à l'attaquant.

L'extraction de données est une autre attaque causée par ce phénomène de mémoire du modèles. Dans ce scénario, l'attaquant connaît une partie d'un échantillon d'apprentissage et il est capable de reconstruire les valeurs manquantes en testant plusieurs d'entre elles sur le modèle entraîné. Plus le modèle a confiance dans le résultat, plus les chances que les valeurs manquantes soient proches des valeurs d'origine sont élevées. Cette attaque est même possible lorsque l'attaquant doit deviner toutes les valeurs de l'échantillon d'apprentissage :

Une image récupérée avec une attaque par extraction de données (à gauche) par rapport à l'original (à droite). L’attaquant n’a que le nom de la personne, peut entrer un échantillon et recueille la sortie d’un système de reconnaissance faciale.

La dernière attaque connue est le **vol de modèle**. Dans ce scénario, le modèle est une technologie prioritaire et ce modèle a une interface publique. Un attaquant peut forger ou rassembler des échantillons similaires aux échantillons d'apprentissage et créer un modèle secondaire imitant les prédictions du modèle cible, en le volant effectivement.

Biais indésirables

Le terme biais dans l'apprentissage automatique a différentes significations pour différents contextes. Ici, un biais est l'importance qu'une caractéristique a dans la prédiction finale du modèle. Lorsque le problème concerne des données relatives aux personnes, il est possible qu’on ne veuille pas que le modèle base ses prédictions selon certaines caractéristiques. Cela peut être des biais *injustes*. Ce qui est considéré comme *juste* et *injuste* dépend de la culture et du pays de l’utilisateur et est lui-même un sujet de débat, mais par souci de simplicité, nous parlons ici principalement de préjugés *sexistes* et *raciaux*.

Plusieurs scandales ont montré que l'apprentissage automatique reproduisait ou amplifiait les biais existants dans le jeu de données d'apprentissage. Par exemple, Amazon utilise une AI de sélection de CV pour le recrutement. En 2018 un **article Reuters** expose un grave problème de préjugés sexistes dans ce processus. En fait, le modèle attachait une grande importance aux mots utilisés plus souvent par les hommes et une importance négative aux mots utilisés plus souvent par les femmes. Ceci est le résultat de données inégalement partagées entre les hommes et les femmes car il y en a beaucoup plus pour les premiers que pour les dernières chez Amazon. Le modèle a reproduit ces inégalités dans ses prédictions.

En 2016, le logiciel d'IA d'évaluation des risques nommé COMPAS a été analysé et un biais racial grave a été découvert. Cet outil a été utilisé pour prédire le risque de récidive. La couleur de peau intervient fortement dans le risque prédit de récidive du criminel.

Prédictions biaisée du logiciel COMPAS

Le problème général des biais indésirables n'est pas uniquement technologique. Il contient plusieurs sujets qui doivent être discutés au cas par cas. La liste des biais injustes est locale au problème. Par exemple, dans un cas, le biais d'âge sera juste, mais dans un autre cas, ce ne sera pas le cas. Il y a aussi la question de la façon dont l'équité est mesurée qui doit être discutée; Différentes méthodes existent, qui n'auront pas la même pertinence dans tous les cas. Cela fait de l'équité un sujet délicat sur lequel il existe des solutions à la fois technologiques et politiques.

Conclusion

Bien que l’Intelligence Artificielle soit une solution technique puissante, permettant des innovations incroyables, sa mise en œuvre nécessite la maîtrise d’un panel de sujets. Sans cette maîtrise, l'IA est une épée de Damoclès. La sécurité, la confidentialité et l'équité de l'IA sont des sujets en construction qui seront standardisés pour la mise en œuvre de l'IA. Dans ce contexte, des initiatives émergent dans cette direction, telles que le **Guide d'éthique pour une IA de confiance** à la commission Européenne, les **Principes de l'IA par l'OCDE** ou encore l’adoption des **Principes de l'IA par le G20**. Être en avance sur ces sujets, c'est être en avance sur les besoins, les normes et les menaces futures.

À propos

Disaitek a été fondée avec une seule mission: utiliser l'IA pour apporter des connaissances et pour apporter des connaissances sur l'IA. Nous travaillons à la construction d'une IA digne de confiance. Visitez notre site Internet **https://www.mlsecurity.ai/** pour voir ce que nous pouvons faire pour votre organisation ou contactez-nous directement à **contact@disaitek.ai**.