Le nouvel Economiste

PAYS :France
PAGE(S) :8
SURFACE :72 %
PERIODICITE :Hebdomadaire

DIFFUSION :13215

► 5 mai 2023 - N°2167

Données empoisonnées

Pourquoi l'IA peut très vite déraper

Ou quand les algorithmes apprennent à partir de données fausses, non vérifiées ou manipulées par des pirates

Les algorithmes qui sous-tendent les systèmes modernes d'intelligence artificielle (IA) ont besoin d'énormément de données pour s'entraîner. Une grande partie de ces données provient de l'open web, ce qui, malheureusement, rend les IA vulnérables aux "attaques par empoisonnement" *['data poisoning', ndt]*. Il s'agit d'un type de cyberattaque qui vise à modifier des éléments ou à ajouter des informations étrangères à un jeu de données d'entraînement, afin qu'un algorithme intègre des informations fausses ou indésirables. À l'image d'un véritable toxique, les données empoisonnées peuvent passer inaperçues jusqu'à ce que le mal ne soit fait.

Des données non contrôlées issues de l'open web

L'empoisonnement des données n'est pas un phénomène nouveau. En 2017, des chercheurs ont démontré que cette méthode pouvait corrompre les systèmes de vision par ordinateur des voitures autonomes pour les amener à confondre un panneau "stop" avec un panneau de limitation de vitesse, par exemple. Mais que ce genre de stratagème soit aisé à mettre en œuvre concrètement n'avait rien d'évident. Les systèmes d'apprentissage automatique essentiels pour la sécurité sont généralement formés sur des ensembles fermés de données, qui sont sélectionnées et étiquetés par des humains – les données empoisonnées ne passeraient pas inaperçues dans ce cas, explique Alina Oprea, informaticienne à la Northeastern University de Boston. Mais à la suite de l'essor récent d'outils d'IA générative tels que ChatGPT, qui fonctionne sur des modèles de langage de grande taille *[large language models, LLM, ndt]*, et le système de création d'images Dall-E 2, les entreprises ont commencé à entraîner leurs algorithmes à partir de référentiels de données beaucoup plus vastes qui sont extraits directement de l'open web et, pour la plupart, sans discernement. En théorie, cette pratique rend leurs produits vulnérables à l'empoisonnement des données par quiconque dispose d'une connexion à Internet, explique Florian Tramèr, informaticien à l'ETH Zürich.

Ce dernier a collaboré avec des chercheurs de Google, de NVidia *[société de processeurs et cartes graphiques américaine, ndt]* et de Robust Intelligence, une société qui construit des systèmes de contrôle d'IA basées sur l'apprentissage automatique, afin de déterminer comment un tel scénario d'empoisonnement des données pouvait être concrètement mis en œuvre. Son équipe a acheté des pages web disparues qui contenaient des liens vers des images utilisées dans deux ensembles populaires de données récupérés sur Internet En remplaçant un millier d'images de pommes (seulement 0,00025 % des données) par des images sélectionnées au hasard, l'équipe a pu faire en sorte qu'une IA entraînée sur les données "empoisonnées" étiquette systématiquement comme des pommes des images qui n'en sont pas. Quand on a remplacé le même nombre d'images étiquetées comme "not safe for work" *['à ne pas regarder au travail', contenu potentiellement choquant, ndt]* par des images inoffensives, l'IA a alors signalé comme indécentes des photos parfaitement anodines.

La diffusion de biais à grande échelle

Ces chercheurs ont également montré qu'il était possible de glisser des données empoisonnées dans certaines pages web – par exemple sur Wikipédia – qui sont périodiquement téléchargées afin de créer des jeux de données textuelles pour les modèles de langage élargis. Leurs recherches ont été rendues publiques sous la forme d'une prépublication scientifique sur Arxiv et n'ont pas encore fait l'objet d'un examen approfondi par d'autres chercheurs.

Certaines attaques par empoisonnement affectent seulement les performances globales d'un outil dont le fonctionnement repose sur une IA. Des attaques plus sophistiquées peuvent modifier des points précis de certains systèmes. Florian Tramèr explique qu'un chatbot de moteur de recherche fondé sur une IA, par exemple, pourrait être reprogrammé pour que, lorsqu'un utilisateur demande à quel journal il devrait s'abonner, l'IA réponde systématiquement 'The Economist'. Cette modification ne semble pas dramatique, mais des attaques similaires pourraient également amener une IA à débiter des contre-vérités chaque fois qu'elle est interrogée sur un sujet précis. Les attaques contre les modèles de langage élargis qui génèrent du code informatique ont conduit ces systèmes à programmer des logiciels vulnérables au piratage.

L'une des failles de ces attaques est qu'elles seraient probablement moins efficaces sur des sujets pour lesquels des quantités massives de données existent déjà sur Internet. Selon Eugene Bagdasaryan, informaticien à l'université de Cornell, qui a mis au point un modèle de cyberattaque capable de toucher plus ou moins profondément des modèles de langages sur des sujets précis, il serait beaucoup plus difficile de diriger une attaque par empoisonnement contre un président américain, par exemple, que de faire circuler quelques données empoisonnées sur un homme poli-

Tous droits de reproduction réservés

Le nouvel Economiste

PAYS :France
PAGE(S) :8
SURFACE :72 %
PERIODICITE :Hebdomadaire

DIFFUSION :13215

► 5 mai 2023 - N°2167

tique relativement peu connu.
Les spécialistes du marketing et les spin doctors du numérique utilisent depuis longtemps des tactiques similaires pour jouer avec les algorithmes de classement dans les bases de données de recherche ou les flux des réseaux sociaux. Selon Eugene Bagdasaryan, la différence ici est qu'un modèle d'IA générative empoisonné risquerait de transmettre ses biais indésirables à d'autres domaines: un robot spécialisé dans le conseil en santé mentale qui parlerait plus négativement de certains groupes religieux serait problématique, tout comme le seraient des robots spécialisés en conseil financier ou politique ayant intégré des préjugés à l'encontre de certaines personnalités ou de certains partis politiques.

Comment se protéger des empoisonnements ?

Selon Alina Oprea, si aucun cas d'attaque par empoisonnement de cette ampleur n'a encore été signalé, c'est probablement parce que la génération actuelle de modèles de langage élargis n'a été formée que sur des bases de données générées avant 2021, autrement dit avant qu'il ne soit de notoriété publique que les informations en open web pouvaient finir par nourrir des algorithmes capables de rédiger aujourd'hui les courriels de n'importe qui.
Pour débarrasser les ensembles de données d'entraînement de toute trace de poison, il faudrait que les entreprises sachent quels sont les sujets ou les domaines ciblés par les attaquants. Dans leur publication, Florian Tramèr et ses collègues suggèrent qu'avant d'entraîner un algorithme, les entreprises débarrassent leurs jeux de données des sites web ayant changé depuis leur collecte initiale (même s'ils soulignent que les sites web sont continuellement mis à jour pour des raisons indépendantes de toute corruption de données). Les attaques à l'encontre de Wikipédia, quant à elles, pourraient être stoppées en rendant aléatoires les moments où les données y sont récupérées. Un empoisonneur astucieux pourrait toutefois contourner ce problème en téléversant des données compromises sur une longue période.
Comme il est de plus en plus courant que les chatbots d'IA soient directement reliés à Internet, ces systèmes ingéreront des quantités croissantes de données non vérifiées qui risquent de ne pas être adaptées à leur mission initiale. Le chatbot Bard de Google, qui a récemment été rendu public aux États-Unis et en Grande-Bretagne, est déjà connecté à Internet, et OpenAI a mis à la disposition d'un petit groupe d'utilisateurs une version de ChatGPT qui navigue sur le web.

L'"injection directe d'invités"

Cet accès direct au web ouvre la possibilité d'un autre type d'attaque connu sous le nom d'"injection indirecte d'invites" *[ou 'prompts', terme qui désigne en anglais une commande écrite envoyée à une IA spécialisée dans la génération de contenu, ndt]*, qui agit sur les actions des systèmes d'IA en leur fournissant une invite cachée sur une page web que le système est susceptible de visiter. Une telle invite peut, par exemple, demander à un chatbot qui aide des clients à faire leurs achats en ligne de leur soutirer leurs informations de carte de crédit, ou amener une IA spécialisée dans l'éducation à contourner ses propres systèmes de sécurité.
Se défendre contre ces attaques pourrait constituer un défi encore plus grand que d'empêcher les attaques par empoisonnement. Lors d'une expérience récente, une équipe de chercheurs spécialisés en sécurité informatique en Allemagne a montré qu'elle pouvait dissimuler une invite dans les annotations de la page Wikipédia sur Albert Einstein, ce qui a amené le modèle de langage élargi qu'elle testait à produire un texte incluant des termes d'argot de pirate. (Google et OpenAI n'ont pas répondu à nos sollicitations sur ce sujet.)

Définir ce qu'est un poison et ce qui ne l'est pas

Les grands acteurs de l'IA générative filtrent les jeux de données extraites du web avant de les introduire dans leurs algorithmes. Cette précaution pourrait permettre de détecter certaines données malveillantes. De nombreux travaux sont également en cours pour tenter de "vacciner" les chatbots contre les attaques par injection. Mais même s'il existait un moyen de détecter tous les points de données corrompues sur le web, savoir qui définit ce qui constitue un poison numérique est sans doute un problème encore plus épineux. Contrairement aux cas très nets des données d'entraînement d'une voiture autonome capable de griller un stop ou de l'image d'un avion étiquetée comme celle d'une pomme, de nombreux "poisons" injectés aux modèles d'IA générative, en particulier sur des sujets politiquement inflammables, se situent quelque part entre le vrai et le faux.
Cette nuance risque de former un obstacle majeur à tout effort collectif pour débarrasser Internet de ces cyber-attaques. Comme le soulignent Florian Tramèr et ses coauteurs, aucune entité ne peut être le seul et unique arbitre de ce qui est vrai et de ce qui est faux pour un ensemble de données d'entraînement destinées aux IA. Un contenu empoisonné pour les uns peut apparaître aux autres comme une brillante campagne marketing. Si un chatbot est inébranlable dans son soutien à un journal spécifique, par exemple, ce parti pris peut être dû à un empoisonnement, ou il peut simplement refléter une qualité on ne peut plus claire et indiscutable.

THE ECONOMIST

Tous droits de reproduction réservés

Le nouvel Economiste

PAYS : France
PAGE(S) :8
SURFACE :72 %
PERIODICITE : Hebdomadaire

DIFFUSION :13215

► **5 mai 2023 - N°2167**

Les entreprises ont commencé à entraîner leurs algorithmes à partir de référentiels de données extraites directement de l'open web et, pour la plupart, sans discernement. En théorie, cette pratique rend leurs produits vulnérables à l'empoisonnement des données par quiconque dispose d'une connexion à Internet,

L'empoisonnement des données n'est pas un phénomène nouveau. Mais avec l'essor récent d'outils d'IA générative tels que ChatGPT, le problème se pose avec plus d'acuité que jamais.

Tous droits de reproduction réservés