1. Données, biais, équilibrage

Introduction aux concepts fondamentaux de données pour l'entrainement des IA. Notions de biais. Équilibrage des bases de données.

#B1C - Cours. Prérequis : **#A2C** (Comment fonctionne une IA ?)

Comprendre

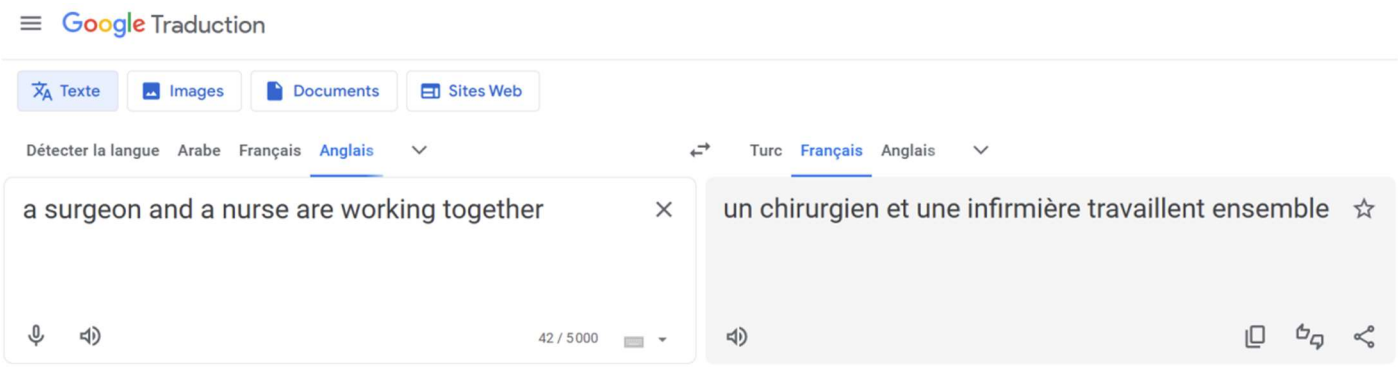
Introduction

La création et le développement d'une IA passe par une phase d'entraînement. Cet entraînement utilise une base de données. Si les données sont fausses, incomplètes ou non représentatives, alors l'apprentissage ne sera pas optimal et l'IA risque de développer des préjugés, des distorsions ou des erreurs dans ses décisions et prédictions. C'est le concept du « garbage in, garbage out » : si les données sont biaisées en entrée, alors elles le seront en sortie. C'est pourquoi il est essentiel de veiller à ce que les données utilisées soient fiables, diverses et représentatives afin de garantir que l'IA puisse apprendre de manière juste et impartiale.

1. Biais algorithmique

Exemple Google Traduction

Observez la traduction automatique de la phrase anglaise suivante vers le français : « a surgeon and a nurse are working together ». Notez que « surgeon » est traduit au masculin (un chirurgien) et « nurse » est traduit au féminin (une infirmière). Cet algorithme n'a pas été programmé pour intégrer des préjugés de genre : il s'agit d'un biais algorithmique.



Dans le cas de l'algorithme de traduction, puisque 70 % des chirurgiens sont des hommes contre 30 % de femmes, et que 85 % des infirmières sont des femmes contre 15 % d'homme, le modèle juge plus probable (d'un point de vue purement statistique) que le chirurgien soit un homme et l'infirmière, une femme, d'où sa traduction !

Définition

Biais algorithmique Un biais algorithmique se produit lorsque des données non représentatives ou distordues sont employées pour entraîner des algorithmes d'intelligence artificielle, entraînant des résultats injustes ou inexacts.

Exemple Algorithme de recrutement Amazon

En 2014, Amazon a lancé un algorithme de recrutement entraîné sur la base des recrutements passés. Les données d'entrées étaient les profils des candidats avec leurs CV et lettres de motivations, et la sortie indiquait si le candidat avait été recruté ou non. Puisque la majorité des employés étaient des hommes, l'algorithme s'est entraîné en apprenant que les hommes étaient les profils recherchés, et a écarté les profils féminins des candidatures ! Cet algorithme n'est plus utilisé aujourd'hui.

2. Les bases de données mal équilibrées

Les bases de données d'entraînement sont souvent mal équilibrées, causant des distorsions et des erreurs dans les prédictions des IA. Les données représentent souvent des photos, idées et opinions d'hommes blancs et occidentaux. Cependant, une minorité ne doit pas être représentée en minorité dans les bases de données ! Pour qu'une base de données soit bien équilibrée, il faut suffisamment de données sur tous les cas qui se présenteront à l'IA.

Exemple Joy Buolamwini et l'algorithme de reconnaissance faciale

C'est ainsi qu'une étudiante du MIT, Joy Buolamwini, a décelé une faille dans les systèmes de reconnaissance faciale dans son entreprise. Elle s'est aperçue que le système reconnaissait mieux les hommes que les femmes et mieux les personnes à la peau claire que les personnes à la peau foncée. En effet : les bases de données sur lesquelles les IA se sont entraînées sont composées à 70 % d'hommes et parmi eux, 80 % ont la peau claire. L'algorithme est par conséquent plus efficace sur ce type de profils, et beaucoup moins sur les profils de personnes afro-américaines comme Joy. Si les chercheurs avaient entraîné leurs algorithmes sur un plus grand nombre de femmes et de personnes de couleur de peau noire, ces algorithmes auraient sûrement été plus efficaces

Exemple Un chien ou un loup ?

C'est aussi ce qu'il s'est passé lorsque des chercheurs ont voulu classer grâce une IA des photos de chiens et de loups. Le modèle avait de bonnes prédictions et semblait bien fonctionner mais lorsque les développeurs ont analysé les cas d'échecs, ils se sont aperçus que le modèle avait fait un raccourci. L'IA a en fait remarqué que la plupart des photos de loups sont prises dehors et souvent dans la neige contrairement à celles des chiens : ce modèle associe donc l'extérieur et la neige au loup. Quand on lui présente alors l'image d'un husky dans la neige : l'IA identifie la variable neige en arrière-plan et l'associe au loup… sans donc reconnaître qu'il s'agit bien d'un chien ! Les données d'entraînement étaient effectivement biaisées, puisque les photos de loup sont généralement prises dans la neige, en extérieur et rarement en intérieur !

3. Corrélation, causalité

Définitions

Corrélation
Deux grandeurs sont dites corrélées si leurs évolutions partagent des similarités. La corrélation contredit alors leur indépendance.

Causalité
La causalité est l'influence par laquelle un évènement, un processus, un état ou un objet (la cause) contribue à la production d'un autre (l'effet).

La corrélation n'implique pas la causalité. Deux variables peuvent montrer des évolutions similaires sans pour autant avoir un lien de cause à effet.

Exemple

Le nombre de glace vendues corrèle fortement avec le nombre de coups de soleil enregistrés. C'est-à-dire que lorsque l'on observe l'évolution positive des ventes de glace, on observe également un accroissement du nombre de coups de soleil.

Est-ce à dire que les glaces provoquent les coups de soleil ? Non ! Il faut étudier les grandeurs dans leur contexte et trouver le facteur de confusion. Ici, le lien de causalité se trouve dans la forte chaleur (ensoleillement important) qui impacte directement les ventes de glaces, mais également le nombre de coups de soleil.

➢ Pour aller plus loin : fiche d'exercice dirigé sur le paradoxe de Simpson **#B2D**.

Cet exemple avec des glaces est simpliste et anodin, mais cette erreur de raisonnement est fréquente et peut concerner des sujets sensibles.

4. Réduction des biais

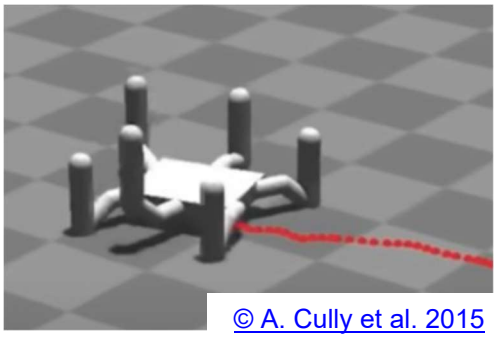
La place de l'IA dans notre monde ne fait qu'augmenter, notamment en entreprise. Elle permet de gagner du temps et faciliter les tâches à faible valeur ajoutée, mais est également employée en soutient à toute sorte d'activité. A terme, il est possible que l'IA prenne des décisions comme des attributions de prêts bancaires par exemple. Il est donc primordial de minimiser les biais algorithmiques et garantir un traitement équitable :

- En encourageant la diversité des équipes de développement pour inciter à la diversité des bases de données utilisées ;
- En mettant en place des audits sur les algorithmes et les bases de données ;
- En encourageant les initiatives qui visent à imposer aux développeurs de rendre leurs codes publics pour garantir une plus grande transparence.

Dans tous les cas, il est important d'essayer d'expliquer les cas d'échecs. C'est-à-dire que lorsqu'un disfonctionnement est constaté, il faut reconstruire la chaine des évènements afin de pouvoir apporter des corrections pertinentes.

Exemple Apprentissage de la marche

Des chercheurs, A. Cully et al. [1], ont modélisé une petite bête à 6 pattes, en lui demandant de marcher en minimisant le temps de contact au sol d'une des pattes. Cette modélisation fonctionnait grâce à un apprentissage par renforcement : c'est-à-dire que la petite bête devait apprendre par elle-même à marcher. Quand ils lui demandent qu'une de ses pattes ne touche pas le sol plus de 10 % du temps, cette petite bête apprend à boiter. Puis quand ils lui demandent qu'une de ses pattes ne touche plus du tout le sol, la bête décide de se retourner, et de marcher avec ses coudes ! Les chercheurs ont donc expliqué le cas d'échec et ont amélioré leur modélisation en interdisant le contact avec le sol de toute la patte (et non plus uniquement le pied), pour que la petite bête ne cherche pas à se retourner. En expliquant ce cas d'échec, les chercheurs ont pu améliorer leur modèle, puis continuer les simulations sur lesquelles ils travaillaient.



© A. Cully et al. 2015

Conclusion

En une phrase : *la qualité des données fait la qualité des modèles*.

En particulier, il faut être attentif à équilibrer les jeux de données en augmentant la quantité de données peu représentées. Il y a fort à parier que les minorités, quelles qu'elles soient, sont toutes traitées de manière biaisée par les IA.

Bibliographie

[1] Cully, A., Clune, J., Tarapore, D., & Mouret, J.-B. (2015). Robots that can adapt like animals. Nature, 521(7553), 503–507.