Enjeux d'éthique liés aux algorithmes d'apprentissage machine et mise en œuvre technique de recommandations

Mélanie Gornet - élève ingénieure 3^e année ISAE-SUPAERO
sous la supervision de
Catherine Tessier et Claude Kirchner - membres du CNPEN
Olivier Grisel - chercheur en apprentissage machine à l'INRIA-CEA

Contexte du stage

La popularité croissante des techniques d'apprentissage automatique appelle à une réflexion sur leur rôle dans la transformation de la société. Elle incite également à favoriser la prise en compte des enjeux d'éthique de ces technologies tant par les chercheur.e.s et acteurs.trices de l'intelligence artificielle que les autorités publiques

Le rôle du Comité National Pilote d'Éthique du Numérique (CNPEN), mis en place à la fin de l'année 2019 à la demande du Premier ministre et constitué de 27 membres issu.e.s d'horizons différents, est d'élaborer des avis sur les enjeux d'éthique du numérique.

De nombreux textes internationaux relatifs à l'« éthique de l'intelligence artificielle » (IA) ont vu le jour ces cinq dernières années. Ils énoncent des valeurs, principes et critères à étudier lors du développement et plus généralement du cycle de vie d'un « système d'IA » : explicabilité, équité, contrôle humain, etc. [5, 4, 2]

En accompagnant le travail d'élaboration d'un avis du CNPEN intitulé « Reconnaissance faciale, posturale et comportementale : entre questionnements et enjeux d'éthiques », ce stage a pour objectif d'analyser les critères préconisés dans les textes internationaux pour l'éthique de l'apprentissage machine et d'étudier la manière dont ils peuvent être traduits sur le plan technique. L'étude porte sur un code de reconnaissance faciale fondé sur de l'apprentissage machine.

Activités

Au sein du Comité, j'étais chargée d'organiser les réunions et auditions du groupe de travail, de rechercher des articles et documents pertinents et de constituer la bibliographie, de rédiger des comptes rendus ou des fiches de synthèse. J'ai participé à la rédaction d'un questionnaire de consultation [3] et d'un chapitre du livre du CNPEN [1]. Enfin, j'ai contribué à la réflexion du groupe à travers mon travail de recherche.

J'ai également pu assister à des séminaires extérieurs (PDIA'21, GDR IA, PFIA - dont CNIA...) ainsi qu'à des événements organisés par le Comité (colloque, séminaire annuel).

Le travail de recherche a comporté deux volets : une analyse des textes internationaux et de la littérature sur la conception orientée valeurs puis le développement d'un cas école sur l'authentification par reconnaissance faciale et l'analyse des choix réalisés lors de la conception du code.

Analyse des valeurs

Les textes internationaux parlent tour à tour de « valeurs », « principes » ou « exigences » de façon interchangeable et sans donner d'indication pour les prendre en compte de manière concrète au sein des systèmes numériques. La démarche du stage s'inscrit dans la volonté de converger vers des exigences opérationnelles pour les logiciels d'authentification par reconnaissance faciale. En effet, chaque choix de conception comporte des implications, des enjeux et est porteur de certaines valeurs. En choisissant de construire un système de telle ou telle façon, le concepteur privilégie des valeurs par rapport à d'autres. Notre travail consiste à identifier où ces choix s'effectuent et tenter d'identifier les directions à prendre pour respecter certaines valeurs.

En préambule, nous avons analysé les valeurs promues par le Groupe d'experts de Haut Niveau de la Commission européenne [5]. Nous avons ensuite décider de nous concentrer sur le critère d'équité.

Développement du cas école

Fonctionnement du code

Le système développé est un réseau de neurones convolutionnel, entraîné par *triplet loss* [6] pour l'authentification de personnes [1].

Lors de l'utilisation normale du système, deux images de visage sont données au réseau qui en crée des gabarits (vecteurs). La distance entre les deux gabarits est calculée : si elle est inférieure à un certain seuil, nous considèrons que les deux images représentent la même personne.

1. L'authentification permet de s'assurer qu'une personne correspond bien à ce qu'elle est censée être - par exemple que le porteur d'un passeport est bien celui que mentionne le passeport ou que la personne qui déverrouille un téléphone est bien le propriétaire de l'appareil. Du point de vue logique, cela correspond à un processus d'appariement 1 parmi 1, par opposition à l'identification qui est un processus en 1 parmi n pour repérer un individu dans un ensemble de personnes. Adapté de [1]

Pour entraîner ce modèle, nous formons des triplés avec une image ancre (A) et une image positive (P) de la même personne, ainsi qu'une image négative (N) d'une personne différente. La perte du triplé, à travers le réseau (f), avec une marge α, est donnée par la fonction :
$L : (A, P, N) \rightarrow L(A, P, N)$ avec
$L(A, P, N) = max(\|f(A) - f(P)\|^2 - \|f(A) - f(N)\|^2 + \alpha, 0)$

Les poids du réseau sont alors actualisés pour obtenir la distance entre A et P plus petite que la distance entre A et N.

Choix effectués

Les choix se répartissent en quatre catégories : ceux qui sont effectués initialement, comme le choix de la base de données, d'un modèle par triplés ou d'un réseau convolutionnel ; ceux qui sont effectués pour atteindre l'*overfitting* comme la structure du réseau, la marge, le *learning rate* et son *scheduling*, la structure d'une epoch [2] et le nombre d'epochs, etc. ; ceux qui servent à généraliser le modèle comme l'augmentation des données, le choix des images négatives ou le *dropout* ; et ceux qui sont liés à la détermination du seuil d'authentification.

Pour chaque choix, nous avons, autant que possible, analysé son impact et argumenté la solution retenue.

Parmi ces choix, nous retenons par exemple la structure d'une epoch, construite pour que chaque identité de l'ensemble de données soit représentée au moins une fois, l'augmentation des données ou le choix de la valeur seuil, qui ont un impact important sur les résultats, à la fois du point de vue de la performance de l'authentification et celui de l'équité des résultats.

Analyse de l'équité

Une fois un modèle obtenu et la valeur seuil définie, nous avons mené une étude des résultats du point de vue de l'équité. Pour cela, nous avons comparé plusieurs métriques (précision, taux d'apprentissage, taux de faux positifs et faux négatifs, etc.) entre différents sous-groupes délimités par des attributs sensibles (genre, couleur de peau).

Au premier abord, la précision semble meilleure pour les images étiquetées « hommes blancs » que pour les autres. Néanmoins, le groupe « autres » est beaucoup plus hétérogène, ce qui peut expliquer la facilité de comparaison. Si nous regardons des groupes plus homogènes (« hommes » vs « non hommes » et « blancs » vs « non blancs ») la précision est équivalente. Cela ne permet pas de conclure à l'absence de biais de notre système, mais prouve que malgré des données très déséquilibrées, les biais peuvent être évités.

Enfin, les métriques que nous avons analysées ne peuvent pas être considérées en dehors de leur contexte. En effet, dans certains cas, un faux positif est beaucoup plus dérangeant qu'un faux négatif et les paramètres doivent alors être ajustés.

Pour aller plus loin

Pour compléter le travail, il conviendrait de vérifier chacun des choix effectués à l'aide d'une *ablation study* [3]. De plus, l'étude pourrait être poursuivie en évaluant quantitativement l'impact de certains paramètres (augmentation des données, marge...) sur l'équité des résultats, ou en se rapprochant davantage d'un cas industriel.

Conclusion

Le cas école développé n'avait pas pour but de créer un système de reconnaissance faciale parfait ou « non biasé », ceci étant très certainement impossible à obtenir. Notre objectif était plutôt d'éclairer la manière dont nous développons un tel système, de souligner que les choix que nous faisons sont porteurs de valeurs. La dépendance entre ces choix rend très difficile la rédaction d'exigences, celles-ci dépendant beaucoup du cas particulier étudié.

Références

[1] R. Chatila, L. Devillers, K. Dognin-Sauze, J.-G. Ganascia, M. Gornet, A. Pronesti, and C. Tessier. Pourquoi la reconnaissance faciale, posturale et comportementale soulève-t-elle des questionnements éthiques ? In E Germain, Cl. Kirchner, and C. Tessier, editors, *L'éthique du numérique : pour quoi faire ?* PUF, 2022.

[2] European Commission. White Paper On Artificial Intelligence - A European approach to excellence and trust, February 2020. URL : https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

[3] Comité National Pilote d'Ethique du Numérique (CNPEN). Reconnaissance faciale, posturale et comportementale - Appel à contributions. Á paraître, 2021.

[4] Ad Hoc Expert Group (AHEG) for the Preparation of a Draft text of a Recommendation the Ethics of Artificial Intelligence. Avant-projet de Recommandation sur l'éthique de l'intelligence artificielle, 2020. URL : https://unesdoc.unesco.org/ark:/48223/pf0000373434_fre/PDF/373434fre.pdf.multi.

[5] Independent Hight-Level Expert Group on Artificial Intelligence set up by the European Commission (HLEG). Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment, July 2020. URL : https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

[6] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet : A Unified Embedding for Face Recognition and Clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, June 2015. arXiv : 1503.03832 version : 1. URL : http://arxiv.org/abs/1503.03832, doi:10.1109/CVPR.2015.7298682.

2. Une epoch indique le nombre de passages dans l'ensemble des données d'apprentissage que l'algorithme a effectués.

3. Une *ablation study* étudie les performances d'un algorithme en retirant certains composants, afin de comprendre la contribution du composant au système global.