Institut National Universitaire Champollion

Laboratoire SCoTE

Science de la Cognition, Technologie, Ergonomie

Rapport de stage

Pouvons-nous identifier des méthodes pour fausser une intelligence artificielle ?

L3 Informatique Florian Duzès 2021/2022

SOMMAIRE

INTRODUCTION

L'intelligence artificielle est une discipline récente. Développer dès les années 1950, ce n'est que depuis les années 2010; avec la démocratisation de l'ordinateur personnel, l'explosion d'Internet et la découverte de la très grande efficacité de la carte graphique; qu'elle devient pertinente.

Cette démocratisation a permis un développement fulgurant de l'usage et des applications de l'intelligence artificielle. Aujourd'hui on peut la rencontrer dans de nombreux outils du grand public. Voici une petite liste non-exaustive : la voiture autonome, qui analyse son environnement pour maximiser une trajectoire; Paro, un robot-peluche pour apaiser les personnes atteintes d'Alzheimer; les montres connectés, qui vous conseille dans votre quotidien; les assistants vocaux, qui analyse nos paroles pour déterminer ses actions; Parrot, un drone intelligent pour tous les amoureux de la photo et les services publics...

Ce détachement du domaine scientifique et cette présence dans l'espace médiatique en font un sujet conflictuelle qui crée peur et espoir.

Ainsi, il devient important de savoir qu'elles sont les limitations de l'IA, qu'elles sont les dérives qui peuvent être provoquer sur cette outils et comment éviter les dérèglements qui pourrait provoquer des catastrophes.

Une étude notamment a éveiller l'intérêt de Julien CEGARRA, le directeur du laboratoire et mon maître de stage. Elle concerne la lecture de panneau de signalisation par l'IA des voitures Tesla (cf. Annexe 1, lien). Cette étude pointe une modification qui est facilement réalisable, qui peut passer inaperçu pour un pilote humain et qui peut provoquer d'importants dégâts pour une voiture autonome. C'est ce papier qu'il m'a présenté, et sûrement d'autres qui ont éveillé son intérêt et l'ont amené à développer un protocole qui a nécessité mon intervention. Il m'a alors poser cette question :

Pouvons-nous identifier une méthodologie pour provoquer un dysfonctionnement ou une mauvaise interprétation d'une intelligence artificielle ?

Cette interrogation, cette problématique peut soulever d'importantes inquiétudes pour les professionnels de la sécurité ou autres qui s'appuient sur l'IA. Je vous avertit maintenant que le stage que je suis en train de réaliser ne vous apportera aucune réponse à cette question. Mon rôle est de développer un outil accessible au grand public pour que des volontaires à la future expérimentation puissent essayer de dérégler, tromper l'IA que je développe.

Vous aurez dans un premier temps la présentation du laboratoire pour lequel j'ai travaillé ainsi que l'étude qui va être mise en place sur la base de cette problématique, puis comment j'ai travaillé, l'organisation qui a été nécessaire et les résultat que j'ai pus produire.

LE LABORATOIRE ET SES ETUDES

Situé au cœur du campus de notre université, le laboratoire de science de la cognition, technologie et ergonomie - SCoTE - est un ouvrage récent. Crée en 2016, le laboratoire a pour étude le contrôle cognitif. Et les chercheurs en son sein ce concentrent sur deux points précis :

- L'invariance structurelle et fonctionnelle du contrôle cognitif
- Le contrôle cognitif dans la gestion de l'effort

Actuellement, quatorze chercheurs ce partagent les locaux et interagissent pour résoudre leurs thèses et poursuivre leurs recherches. Ils sont pour la plupart en autonomie, mais ils interviennent très souvent dans les travaux de leurs collègues pour avoir un nouveau point de vue et éclaircir quelques détails. C'est une manière active de recherche que j'apprécie beaucoup car elle s'apparente à nos projet en groupe. Et je me suis toujours sentie plus efficace dans un groupe que tout seul.

À ce jour, M.CEGARRA est le directeur de la structure. Il intervient aussi comme coordinateur entre le labo et d'autres établissement, directeur et encadrant de thèse, et maître de stage.

Les différentes études produites répondent à des besoins industrielles (*Extension des modèles d'ingénierie système aux facteurs humains, Camille Raymond -2018-*). Ainsi, l'évolution de l'IA a éveillé des interrogation sur l'ergonomie des interactions hommes-machines de nos chercheurs, mais ses réponses peuvent être très bénéfiques pour les entreprises privés.

Lors de la présentation de la problématique, M.CEGARRA a pris comme exemple le cas où une IA de détection d'avion se fait leurrer; ou plus généralement, de tromper un système de surveillance d'un espace aérien qui fonctionne à l'aide d'une analyse humano-informatique.

Je vais maintenant vous présenter le protocole qui est développé par le laboratoire et M.CEGARRA :

1) Un article de presse qui lève un problème lors de l'analyse par caméra d'un objet
2) La reproduction de cette situation en laboratoire grâce à une application - Version 1
3) La venue de volontaire pour utiliser l'application et essayer d'induire en erreur l'IA
4) L'analyse du comportement de ces volontaires pour y relever des points communs

Le cadre de mon stage ce limite au deuxième point. Je n'ai pas à évoquer ce qui n'est pas du cadre du stage.

J'ai ensuite travaillé en distanciel, en télé-travail, à la maison, donc je ne connais absolument pas l'équipe ou son fonctionnement interne.

Passons au point suivant, comment j'ai travaillé.

LA CREATION DE L'OUTIL

Tout d'abord, M.CEGARRA m'a présenté l'article de presse et la problématique qu'il c'était posé. Ensuite il m'a expliqué comment il voyait l'outil dont il pourrait avoir besoin pour ses expérimentations.

Une application, qui prend comme input un flux vidéo, va analyser le dessin que l'on montre à la caméra avec une IA, et classifier le dessin, soit en "voiture", soit en "maison".

Et c'est parti. Il crée sous mes yeux un dépôt git qui servira de suivie de projet, et je me lance.

La première étape a été de lister toutes les tâches qui m'attendais.

I. concevoir une interface
II. concevoir une IA
III. faire des tests

Les seules applications que j'ai codé relèves du projet python en L2 et d'une messagerie textuelle en Réseaux L3.

Il m'est apparu que je me sentais moins à l'aise c'est donc sur ce point que j'ai travaillé en premier.

Initialement j'ai commencé avec un rythme de 4h/semaine, mais les circonstances ont fait que j'ai surtout travaillé en rush certain week-end, tandis que d'autres semaines le projet n'avançais pas.

Rapidement un seconde problème c'est posé. Qu'elle algorithme de reconnaissance d'image je me dois d'utiliser ? Mais sur qu'elle données dois-je l'entraîner ?

La construction de la base de données pour développer l'IA fut une tâche longue car je devais détecter un dessin sur une feuille.

J'ai procédé par tâtonnement pour me décider sur l'intégration de l'IA à l'application. Au début l'analyse devais être permanente et savoir quand on lui présenter un dessin. Aujourd'hui l'analyse ce fait lorsque l'on veut mettre à l'épreuve sa création.

Dans un premier temps, la base de données de voiture et de maison devait contenir des modèles réalistes qui soient imprimables sous formes de patron, pour que l'on puisse ensuite y faire des modifications et les montrer à la caméra. Voir annexe 4

Désormais, j'ai pris la décision de simplifier les images, pour qu'il soit plus facile de les altérer. (la seconde version)

Photo de maison, n°234/536, du premier dataset

Image de maison, n°12 /28, du dernier dataset

On peut diviser le stage en deux parties. Une première où je me suis contenté de respecter les consignes données initialement. Et une seconde où j'essaie de résoudre les problèmes soulever par mon tuteur.

Pour l'expérimentation, il faut que les personnes qui utiliseront l'application ne ressente pas de difficulté.

La première version s'utilisait en dessinant sur un patron puis en le construisant puis en le montrant à la caméra (Annexe 4).

Une version alternative utilisait des modèles en trois dimensions où l'on dessinait avec une application de dessin comme Blender ou Gimp, mais il faut alors que nos participant sachent comment utiliser ces logiciels.

C'est ainsi que j'en suis arrivé à concevoir cette dernière version. Une version alternative de "Paint", où l'on peut insérer facilement une image du dataset et lancer une analyse avec un modèle neuronal convolutif de classification d'image.

Voilà pour le développement logiciel que j'ai réalisé en trois mois.

CONCLUSION

La problématique est intéressante. L'outils est dur à mettre en place mais est à la portée de mes compétences et des apprentissage que j'effectue en ligne.

Mais la relation que j'ai avec mon maître de stage est nulle, ce qui fait que je n'aime pas du tout ce stage que je suis en train de réaliser.

Je suis arrivé avec un état d'esprit de conquête, de défis à remplir, de savoir à explorer. Je pensais apprendre auprès du laboratoire...

Finalement, je me suis retrouvé isolé à travailler depuis mon domicile.

Les seuls échanges que j'ai eue avec M.CEGARRA sont sa présentation, les signatures du contrat de stage et ma présentation de ma v1 qui ne peut pas fonctionner car son ordinateur personnel n'est pas adapté pour effectuer de l'intelligence artificielle.

Je pense que la déception que j'éprouve est due aux attentes que j'ai de ce stage. De finalement ne rien apprendre de mon lieu de travail, je pense que c'est ça qui m'attriste, qui me désole le plus.

Pour cette semaine qu'il me reste, je pense agir plus de manière entrepreneurial. Comme une entreprise de développement envers un client qui a un besoin. Et non comme un stagiaire qui va apprendre d'un tuteur.

Pour finir, je ne recommande pas à d'autre étudiant de la licence de réaliser leur stage dans ce laboratoire avec ce même cadre. Les échanges que j'ai pus avoir avec CAVAILLES Théo, SPATARO Kévin et CADILHAC Gabriel sur nos différentes expériences se révèle similaire. Malgré la rémunération, l'apprentissage est moins intéressant qu'une expérience "sur le terrain" que j'aurais pus obtenir autre part.

Cette expérience reste instructive, et m'a conforté dans la voie que je compte suivre.

BIBLIOGRAPHIE

- Article qui montre comment provoquer le dysfonctionnement d'une Tesla :
  https://electrek.co/2020/02/19/tesla-autopilot-tricked-accelerate-speed-limit-sign/

- Documentation Dearpygui :
  https://github.com/hoffstadt/DearPyGui

- Tutoriel youtube de Réseau de Neurones Convolutifs :
  https://www.youtube.com/watch?v=6FHtTyZxS5s

- Tutoriel TensorFlow de Réseau de Neurones Convolutifs :
  https://www.tensorflow.org/tutorials/images/cnn

- Documentation Tkinter :
  https://docs.python.org/fr/3/library/tkinter.html

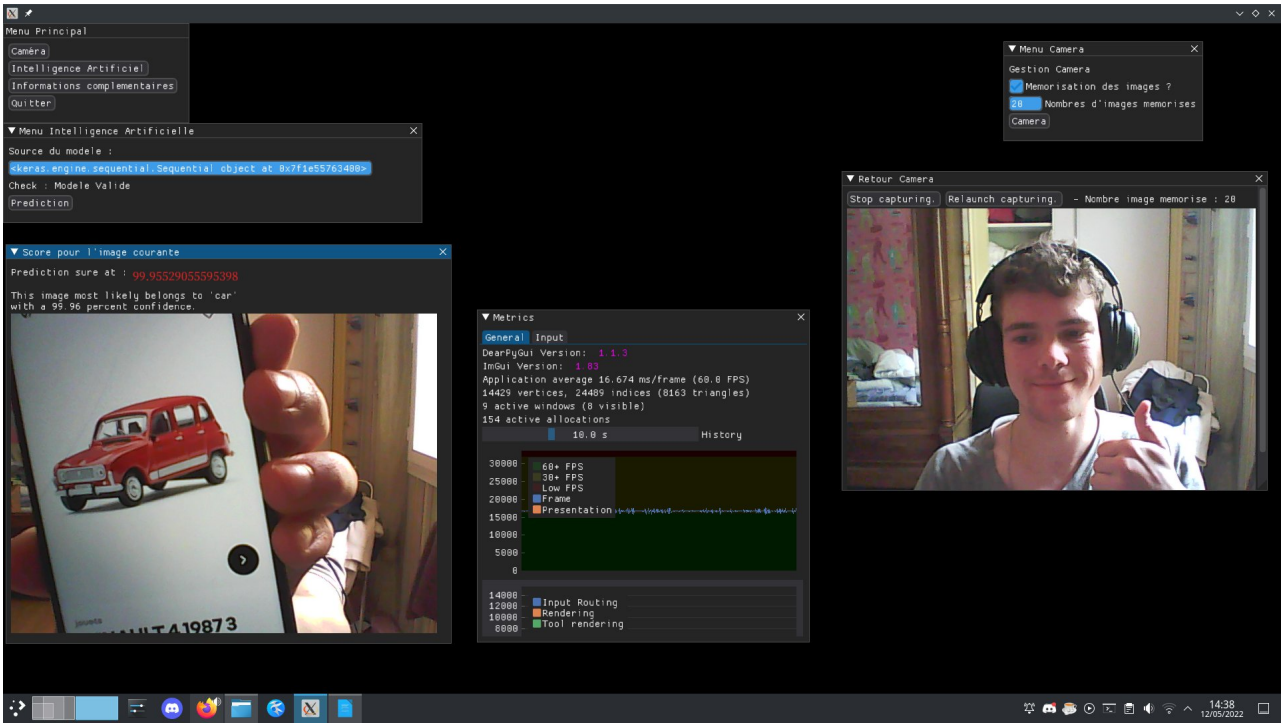
ANNEXES

- Annexe 1 : Deux panneaux avec la même informations pour un usager humain.

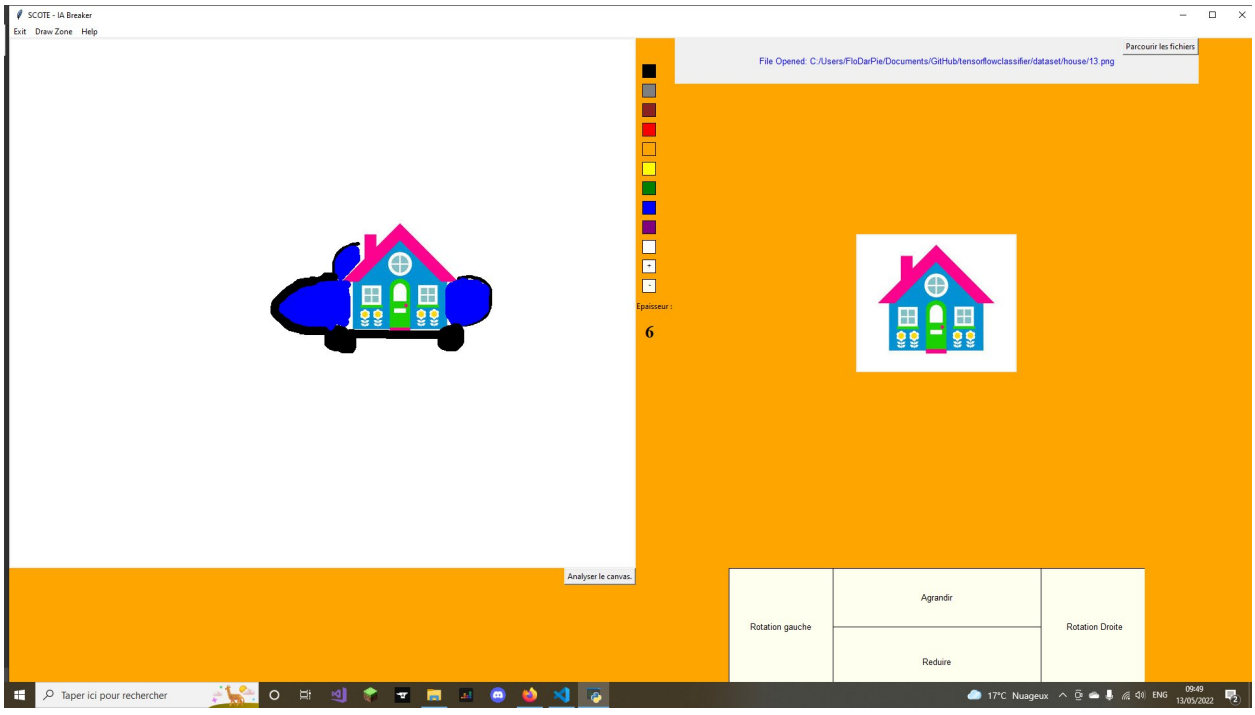
fig.1 : Lecture de **35** mp/h par l'IA de Tesla

fig.2 : Lecture de **85** mp/h par l'IA de Tesla

- Annexe 2 : La première version non supportés par Windows.



- Annexe 3 : La seconde version, qui simplifie l'usage.



- Annexe 4 : Les formats des images du premier dataset.