Un outil pour explorer la généalogie des modèles d'IA en source ouverte

Qu'est-ce qu'un modèle d'IA ?

Son entraînement

Les champs d'action de l'IA sont vastes et semblent difficiles à circonscrire puisqu'ils s'étendent à de nombreux aspects du quotidien : que ce soit pour effectuer des recherches ou des achats en ligne, le ciblage publicitaire, la traduction automatique, les assistants numériques personnels, les villes connectées, mais aussi dans le domaine des transports, de la santé, etc.

D'après l'article 3 du règlement européen sur l'intelligence artificielle un système d'IA peut être défini comme *« un système basé sur une machine qui est conçu pour fonctionner avec différents niveaux d'autonomie et qui peut faire preuve d'adaptabilité après son déploiement, et qui, pour des objectifs explicites ou implicites, déduit, à partir des données qu'il reçoit, comment générer des résultats tels que des prédictions, du contenu, des recommandations ou des décisions qui peuvent influencer des environnements physiques ou virtuels. »*

Ces systèmes intègrent un ou plusieurs modèles d'IA qu'il est possible de définir comme des algorithmes, dont le fonctionnement est déterminé par un ensemble d'attributs, et qui sont conçus pour opérer, selon les cas, différentes tâches, telles que la prédiction, la classification, l'inférence ou la génération. Par exemple, les modèles de réseaux de neurones profonds (*deep neural networks*) sont constitués de nœuds (les neurones), répartis selon des couches, et reliés entre eux par des connections possédant chacune un paramètre ou « poids ». Ces paramètres sont ajustés durant la phase d'entraînement pour apprendre la distribution statistique des données d'entraînement. Concrètement, dans le cas d'un réseau de neurone simple, les attributs du modèle pourraient être :

(i) Le type et la taille de chaque couche (linéaire, convolutionnel, attention, etc.),

(ii) Les poids attribués à chaque arrête (parfois aussi appelés « paramètres »),

(iii) Les fonctions d'activations présentes entre chaque couche,

(iv) Et possiblement d'autres opérations qui peuvent être situées au sein ou entre les couches.

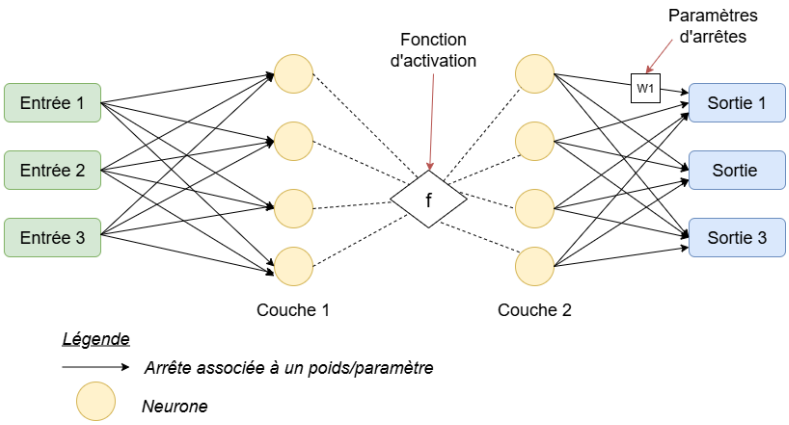

Figure 1 : Schéma d'un réseau neuronal (auteurs)

Par exemple, quand un réseau de neurones est entraîné pour reconnaître des images, il lui est fourni des exemples où les pixels de l'image sont associés à une annotation (ou « étiquette »). Le modèle ajuste alors ses paramètres, appelés « poids », pour apprendre à attribuer le bon label le plus souvent possible.

La principale différence entre un modèle d'apprentissage profond et un programme informatique classique est que le modèle apprend de manière autonome les règles d'inférence à partir des données.

Dans un programme classique, la résolution d'une tâche repose sur un ensemble de règles explicites définies à l'avance par le développeur. Par exemple, pour trier une liste de nombres, l'ordre dans lequel comparer les éléments est programmé précisément. Ce type d'approche fonctionne très bien pour des tâches précises et délimitées, ce qui permet d'établir des règles claires pour les résoudre.

À l'inverse, dans le cas d'un modèle d'apprentissage profond, les règles ne sont pas spécifiées directement. Il est plutôt fourni au modèle un grand volume de données d'exemples (dites « données d'entraînement ») pour permettre, dans la phase dite d'apprentissage, de trouver les régularités statistiques ou les stratégies qui permettent de résoudre la tâche. Cette approche permet d'automatiser des tâches beaucoup plus complexes pour lesquelles il serait extrêmement difficile, voire impossible, de définir toutes les règles à la main.

Son utilisation

Nous allons maintenant explorer certaines des tâches complexes que les modèles d'IA peuvent résoudre. Une fois qu'un modèle a été entraîné, il peut être utilisé tel quel, sans modification supplémentaire, pour effectuer automatiquement des tâches spécifiques. C'est ce qu'on appelle la phase d'inférence. À ce moment-là, le modèle reçoit une entrée (par exemple une image, un texte, ou un signal audio) et produit une sortie en fonction de ce qu'il a appris lors de l'entraînement. Il agit alors comme une « boîte noire » : il applique les régularités qu'il a intégrées, sans modifier sa structure interne ni apprendre de nouvelles choses.

Prenons l'exemple de la traduction automatique. Un modèle basé sur un réseau de neurones entraîné sur des millions de paires de phrases en espagnol et en anglais peut être utilisé, à l'inférence, pour traduire automatiquement un nouveau texte de l'anglais vers l'espagnol. Les règles linguistiques ne sont pas explicitement implémentées dans le modèle, mais il a appris à faire correspondre des séquences de mots en s'appuyant sur des régularités statistiques présentes dans les données d'entraînement.

Autre exemple, les modèles « *image-to-text* » permettant la génération de légendes d'images. Il est possible d'entraîner un modèle à associer des images avec des descriptions textuelles. Une fois entraîné, il

est capable de recevoir une nouvelle image, par exemple, une photo de chien courant dans un parc, et de générer automatiquement une phrase du type : « Un chien court sur l'herbe dans un parc ».

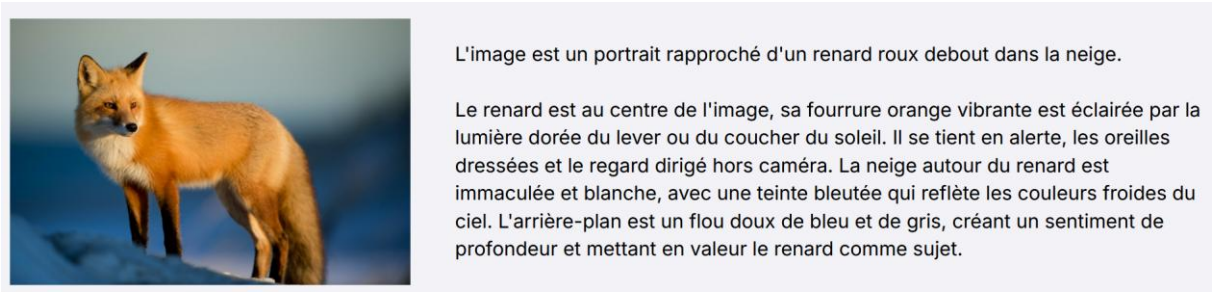

Figure 2 – Exemple de description textuelle d'une image (source : https://imagedescriber.online/fr)

Ces cas d'usage illustrent la puissance des avancées de l'apprentissage profond dans l'automatisation de tâches complexes, souvent subjectives ou ambiguës, pour lesquelles il serait difficile voire impossible d'écrire des règles explicites à la main.

Ses dérivés : *finetuning, merge, quantization...*

Toutefois, pour adapter plus finement un réseau de neurones à une tâche spécifique, optimiser ses performances, ou encore réduire ses coûts d'exécution, plusieurs transformations peuvent être effectuées à partir d'un modèle initial pré-entraîné.
Ces modifications sont très fréquentes dans l'écosystème en source ouverte (*open source*) et permettent, à partir d'un ou plusieurs modèles initiaux, et possiblement de données supplémentaires, de créer de nouveaux modèles. Parmi ces transformations qui apparaissent entre un modèle cible et un modèle source, nous pouvons en mentionner quatre types :

- Ajustement (*finetune*) : le modèle d'origine est un modèle général qui parfait son entrainement sur un jeu de données spécifique afin d'améliorer ses performances pour une tâche plus précise (par exemple : un grand modèle de langage (LLM), initialement entraîné sur des sources internet librement accessibles, est ajusté sur les données métiers d'une entreprise, afin de mieux maîtriser le vocabulaire et les expressions métiers de celle-ci).

- Quantification (*quantized*) : la précision des poids du modèle d'origine est réduite afin de diminuer son empreinte en mémoire (par exemple : le poids sont initialement encodés sur 32 bits, et sont arrondis au nombre codé sur 16 bits le plus proche).

- Adaptation (*adapter*) : le modèle d'origine est ajusté pour pouvoir être utilisée avec peu de ressources de calcul (par exemple pour pouvoir être utilisé sur téléphone portable), la plupart du temps basée sur la technique de *Low Rank Adaptation* (LoRA).

- Fusion (*merge) :* les couches de différents modèles sont mélangées afin d'améliorer leur performance. Par exemple : un LLM A et un LLM B ont tous les deux été entraînés sur des corpus généraux de texte. La moyenne des poids qui se situent dans la 12^{e} couche de A et B est effectué : il apparaît que le LLM C obtenu en remplaçant dans A la 12^{e} couche par la moyenne des couches de A et B a des meilleures performances que A et B.

Une plateforme pour l'IA en source ouverte : *HuggingFace*

Afin de permettre le partage et la mise à disposition de modèles d'IA pour et par le plus grand nombre, l'entreprise franco-américaine HuggingFace, créée en 2016, a développé une plateforme *de* centralisation des modèles et jeux de données. Elle propose aussi des outils logiciels pour déployer des modèles d'IA. C'est la plateforme qui recense aujourd'hui le plus de modèles en source ouverte disponibles (de deux millions de modèles disponibles en septembre 2025) et qui joue un rôle de catalyseur de l'écosystème de l'IA en source ouverte.

Voici un exemple concret pour comprendre comment fonctionne cette plateforme :

- L'utilisateur C a pour objectif la création d'un modèle permettant la détection automatique de courriels frauduleux.
- L'utilisateur A a publié sur HuggingFace un modèle de traitement du langage automatique (exemple : Google a publié google/gemma-3-27b-it · Hugging Face) et l'utilisateur B a publié un jeu de données contenant des millions de courriels classés comme frauduleux ou non.
- L'utilisateur C peut télécharger ce modèle et ce jeu de données. Ensuite, il entraîne le modèle sur ce jeu de données afin de le spécifier pour sa tâche de classification. Une fois qu'il obtient un modèle de classification automatique avec de bons résultats, il peut publier son nouveau modèle sur HuggingFace afin que n'importe quel utilisateur puisse l'utiliser tel quel ou l'entraîner de nouveau sur d'autres jeux de données pour améliorer ses performances.

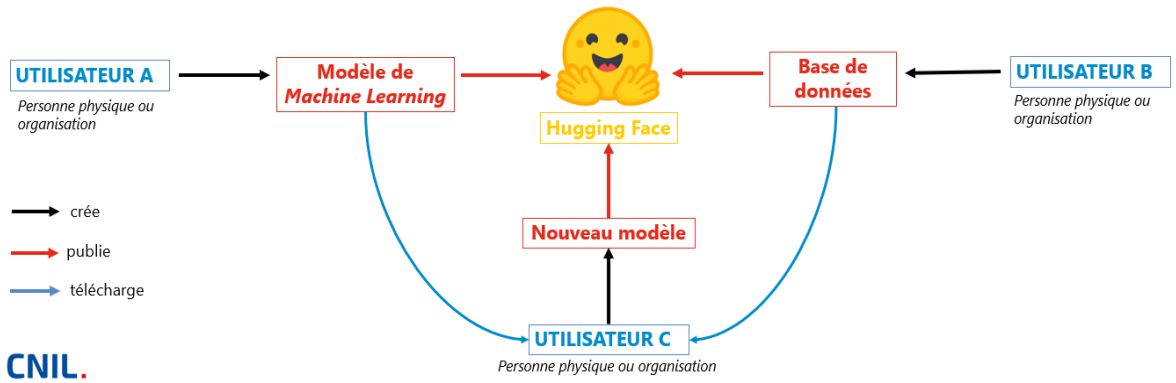

Figure 3 : Exemple d'utilisation de HuggingFace

En somme, *HuggingFace* est une plateforme qui fournit des outils pour construire, entraîner et déployer des modèles d'apprentissage profond basés sur des technologies et du code en source ouverte. Il offre également un espace où chercheurs, ingénieurs et amateurs peuvent se réunir pour échanger des idées, obtenir du soutien et contribuer à des projets en source ouverte.

Bienfaits de l'IA en source ouverte

L'essor de l'IA en source ouverte montre que des modèles puissants et en partie transparents peuvent rivaliser avec les solutions propriétaires tout en stimulant l'innovation collective. Le modèle BLOOM (176 milliards de paramètres, 2022) développé par le consortium BigScience (voir l'entrevue avec le LINC) illustre cette dynamique : entraîné sur 46 langues, il a permis le développement d'assistants conversationnels multilingues en Afrique, en Amérique latine et dans le monde arabe, là où les modèles commerciaux restaient peu adaptés aux langues locales.

De même, GPT-J et GPT-NeoX (**EleutherAI**), Vicuna (**LMSYS**) ont servi de base à des projets en source ouverte qui ont permis à des universités et startups de créer des *chatbots* spécialisés sans dépendre de services fermés. Ces modèles ont également rendu possible la recherche sur la détection de biais et la robustesse des grands modèles de langage.

Dans le domaine de la vision, Stable Diffusion (**Stability AI**) a bouleversé la création visuelle : les poids du modèle ont été mis à disposition gratuitement, il a ouvert la voie à des applications dans le jeu vidéo, la publicité et la production audiovisuelle (génération d'images de concept, storyboards, design rapide). Son code en source ouverte a permis la création d'outils comme Automatic1111 ou ComfyUI, utilisés par des centaines de milliers d'artistes et de chercheurs.

L'impact est aussi industriel : LLaMA (Meta), initialement diffusé à la communauté de recherche, a donné naissance à toute une génération de modèles dérivés (Zephyr, Nous-Hermes, OpenChat, etc.) utilisés aujourd'hui pour des tâches concrètes de support client, de résumé de documents juridiques ou médicaux, et même de prototypage de code.

Enfin, dans le domaine scientifique, des projets comme BioGPT (Microsoft Research) ou OpenFold (inspiré d'AlphaFold) démontrent comment l'ouverture du code et des poids accélère la recherche biomédicale, en permettant à des laboratoires indépendants de reproduire et d'améliorer des résultats sur la prédiction de structures protéiques ou la recherche de molécules.

Ces réussites montrent que la source ouverte ne se limite pas à la réutilisation de modèles : elle permet une appropriation technologique, une adaptation locale et une innovation ouverte dans des domaines aussi variés que la création artistique, la santé, l'éducation, la science des données et les industries culturelles.
Néanmoins, un certain niveau d'opacité peut demeurer dans la façon dont ces modèles sont constitués : avec quelles données ? Avec quel algorithme d'entraînement ? Pour aller plus loin, une note de la CNIL ainsi qu'une note du PEReN sont disponibles sur le sujet.

Enjeux pour la vie privée

La mémorisation des modèles d'IA

La communauté scientifique a établi de longue date, qu'il est souvent possible d'extraire des informations sur les données à partir desquelles un modèle d'IA est entraîné, à partir d'un accès même partiel au modèle (voire à ce sujet l'article LINC sur la taxonomie des attaques). Dans le cas de l'IA générative, un modèle peut par exemple reproduire du texte ou des images qui sont très proches de données qui étaient présentent dans son jeu de données d'entraînement. Dans la figure ci-dessous, nous voyons que quand il est demandé au modèle de *stable diffusion* de générer une image correspondant à la légende « Emma Watson to play Belle in Disneys Beauty and the Beast », l'image générée en sortie correspond de très près à une image qui était présente dans la base d'entraînement. Il s'agit de **régurgitation** (voir figure ci-dessous), qui n'est qu'un des avatars de la mémorisation. Il est par exemple parfois possible d'obtenir d'autres types d'informations, telle que l'appartenance d'une donnée particulière au jeu d'entraînement, à l'aide de méthodes statistiques (attaques par inférence d'appartenance).

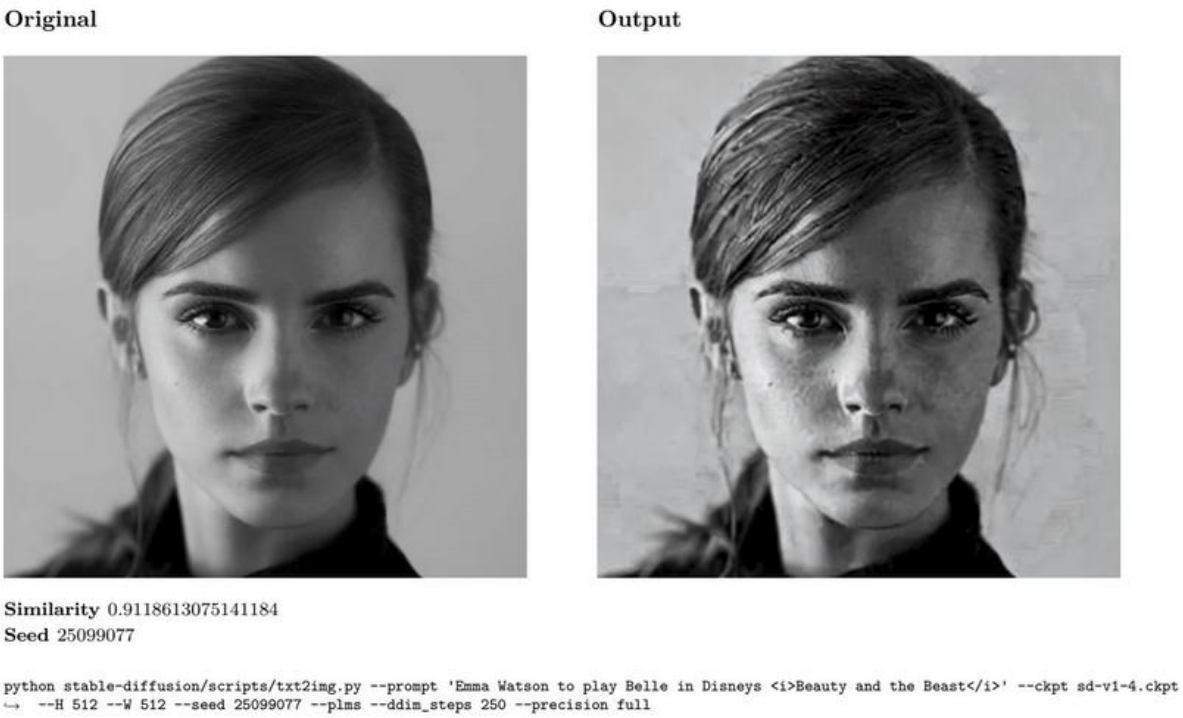

Figure 1 - Source Louis Hunt (Linkedin)
Source photo originale : ONU Femmes

Pour les modèles de texte comme les chatbots, des cas emblématiques de régurgitations sont déjà largement documentés, comme l'observation qu'une version de ChatGPT a été capable de générer presque à l'identique des articles du New York Times, ou bien de fournir des informations personnelles telles que le nom, l'adresse et le numéro de téléphone d'une personne (voir notre article sur le sujet).

Le RGPD et les modèles d'IA

Dès lors qu'il est en général possible d'extraire des informations concernant la base d'entraînement d'un modèle d'IA à partir de celui-ci, quel régime juridique doit s'appliquer à celui-ci si le jeu d'entraînement contient des données personnelles ? C'est ce qu'a clarifié le Comité européen de protection des données dans son avis 28/2024 sur les modèles d'IA, sur lesquelles se basent les dernières recommandation de la CNIL. En particulier, l'avis conclut que le RGPD doit s'appliquer dans de nombreux cas aux modèles d'IA lorsqu'ils ont été entraînés sur des données personnelles, en raison de leur capacité de mémorisation.

Exercer ses droits sur des modèles d'IA

Pour les modèles d'IA soumis au RGPD, les personnes concernées par la mémorisation ont des droits sur leurs données, tels que le droit d'opposition, d'accès ou d'effacement. Il faut noter que ces droits ne sont pas absolus, et qu'un responsable de traitement peut y déroger dans plusieurs situations, par exemple lorsque la demande est manifestement infondée ou excessive (article 12), ou bien que celui-ci n'est pas en mesure d'identifier la personne concernée (pour plus de détail, voir fiche sur l'exercice des droits).

Dans un contexte où les instances européennes confirment que le droit à la protection des données s'applique également aux modèles d'IA, la CNIL souhaite étudier les conditions dans lesquelles ceux-ci pourraient s'appliquer au sein de l'écosystème très dynamique de l'IA en source ouverte.