Qu’est-ce que l’intelligence artificielle ?

Qu’est-ce que l’intelligence ? Est-ce la capacité à percevoir le monde, à prédire le futur immédiat ou lointain, ou à planifier une série d’actions pour atteindre un but ? Est-ce la capacité d’apprendre, ou celle d’appliquer son savoir à bon escient ? La définition est difficile à cerner.

On pourrait dire que l’intelligence artificielle (IA) est un ensemble de techniques permettant à des machines d’accomplir des tâches et de résoudre des problèmes normalement réservés aux humains et à certains animaux.

Les tâches relevant de l’IA sont parfois très simples pour les humains, comme par exemple reconnaître et localiser les objets dans une image, planifier les mouvements d’un robot pour attraper un objet, ou conduire une voiture. Elles requièrent parfois de la planification complexe, comme par exemple pour jouer aux échecs ou au Go. Les tâches les plus compliquées requièrent beaucoup de connaissances et de sens commun, par exemple pour traduire un texte ou conduire un dialogue.

Depuis quelques années, on associe presque toujours l’intelligence aux capacités d’apprentissage. C’est grâce à l’apprentissage qu’un système intelligent capable d’exécuter une tâche peut améliorer ses performances avec l’expérience. C’est grâce à l’apprentissage qu’il pourra apprendre à exécuter de nouvelles tâches et acquérir de nouvelles compétences.

Le domaine de l’IA n’a pas toujours considéré l’apprentissage comme essentiel à l’intelligence. Par le passé, construire un système intelligent consistait à écrire un programme « à la main » pour jouer aux échecs (par recherche arborescente), reconnaître des caractères imprimés (par comparaison avec des images prototypes), ou faire un diagnostic médical à partir des symptômes (par déduction logique à partir de règles écrites par des experts). Mais cette approche « manuelle » a ses limites.

L’apprentissage machine

Les méthodes manuelles se sont avérées très difficiles à appliquer pour des tâches en apparence très simples comme la reconnaissance d’objets dans les images ou la reconnaissance vocale. Les données venant du monde réel – les échantillons d’un son ou les pixels d’une image – sont complexes, variables et entachées de bruit.

Pour une machine, une image est un tableau de nombres indiquant la luminosité (ou la couleur) de chaque pixel, et un signal sonore une suite de nombres indiquant la pression de l’air à chaque instant.

Comment une machine peut-elle transcrire la suite de nombres d’un signal sonore en série de mots tout en ignorant le bruit ambiant, l’accent du locuteur et les particularités de sa voix ? Comment une machine peut-elle identifier un chien ou une chaise dans le tableau de nombre d’une image quand l’apparence d’un chien ou d’une chaise et des objets qui les entourent peuvent varier infiniment ?

Il est virtuellement impossible d'écrire un programme qui fonctionnera de manière robuste dans toutes les situations. C'est là qu'intervient l'apprentissage machine (que l'on appelle aussi apprentissage automatique). C'est l'apprentissage qui anime les systèmes de toutes les grandes entreprises d'Internet. Elles l'utilisent depuis longtemps pour filtrer les contenus indésirables, ordonner des réponses à une recherche, faire des recommandations, ou sélectionner les informations intéressantes pour chaque utilisateur.

Un système entraînable peut être vu comme une boite noire avec une entrée, par exemple une image, un son, ou un texte, et une sortie qui peut représenter la catégorie de l'objet dans l'image, le mot prononcé, ou le sujet dont parle le texte. On parle alors de systèmes de classification ou de reconnaissance des formes.

Dans sa forme la plus utilisée, l'apprentissage machine est *supervisé* : on montre en entrée de la machine une photo d'un objet, par exemple une voiture, et on lui donne la sortie désirée pour une voiture. Puis on lui montre la photo d'un chien avec la sortie désirée pour un chien. Après chaque exemple, la machine ajuste ses paramètres internes de manière à rapprocher sa sortie de la sortie désirée. Après avoir montré à la machine des milliers ou des millions d'exemples étiquetés avec leur catégorie, la machine devient capable de classifier correctement la plupart d'entre eux. Mais ce qui est plus intéressant, c'est qu'elle peut aussi classifier correctement des images de voiture ou de chien qu'elle n'a jamais vues durant la phase l'apprentissage. C'est ce qu'on appelle la *capacité de généralisation*.

Jusqu'à récemment, les systèmes de reconnaissance des formes classiques étaient composés de deux blocs : un extracteur de caractéristiques (*feature extractor* en anglais), suivi d'un classifieur entraînable simple. L'extracteur de caractéristiques est programmé « à la main », et transforme le tableau de nombres représentant l'image en une série de nombres, un *vecteur de caractéristiques*, dont chacun indique la présence ou l'absence d'un motif simple dans l'image. Ce vecteur est envoyé au classifieur, dont un type commun est le *classifieur linéaire*. Ce dernier calcule une somme pondérée des caractéristiques : chaque nombre est multiplié par un poids (positif ou négatif) avant d'être sommé. Si la somme est supérieure à un seuil, la classe est reconnue. Les poids forment une sorte de « prototype » pour la classe à laquelle le vecteur de caractéristiques est comparé. Les poids sont différents pour les classifieurs de chaque catégorie, et ce sont eux qui sont modifiés lors de l'apprentissage. Les premières méthodes de classification linéaire entraînable datent de la fin des années cinquante et sont toujours largement utilisées aujourd'hui. Elles prennent les doux noms de *perception* ou *régression logistique*.

Apprentissage profond et réseaux neuronaux

Le problème de l'approche classique de la reconnaissance des formes est qu'un bon extracteur de caractéristiques est très difficile à construire, et qu'il doit être repensé pour chaque nouvelle application.

C'est là qu'intervient l'apprentissage profond ou *deep learning* en anglais. C'est une classe de méthodes dont les principes sont connus depuis la fin des années 80, mais dont l'utilisation ne s'est vraiment généralisée que depuis 2012, environ.

L'idée est très simple : le système entraînable est constitué d'une série de modules, chacun représentant une étape de traitement. Chaque module est entraînable, comportant des paramètres ajustables similaires aux poids des classifieurs linéaires. Le système est entraîné de bout en bout : à chaque exemple, tous les paramètres de tous les modules sont ajustés de manière à rapprocher la sortie produite par le système de la sortie désirée. Le qualificatif *profond* vient de l'arrangement de ces modules en couches successives.

Pour pouvoir entraîner le système de cette manière, il faut savoir dans quelle direction et de combien ajuster chaque paramètre de chaque module. Pour cela il faut calculer un *gradient*, c'est-à-dire pour chaque paramètre ajustable, la quantité par laquelle l'erreur en sortie augmentera ou diminuera lorsqu'on modifiera le paramètre d'une quantité donnée. Le calcul de ce gradient se fait par la méthode de *rétropropagation*, pratiquée depuis le milieu des années 80.

Dans sa réalisation la plus commune, une architecture profonde peut être vue comme un réseau multicouche d'éléments simples, similaires aux classifieurs linéaires, interconnectés par des poids entraînables. C'est ce qu'on appelle un réseau neuronal multicouche.

Pourquoi neuronal ? Un modèle extrêmement simplifié des neurones du cerveau les voit comme calculant une somme pondérée et activant leur sortie lorsque celle-ci dépasse un seuil. L'apprentissage modifie les efficacités des *synapses*, les poids des connexions entre neurones. Un réseau neuronal n'est pas un modèle précis des circuits du cerveau, mais est plutôt vu comme un modèle conceptuel ou fonctionnel. Le réseau neuronal est inspiré du cerveau un peu comme l'avion est inspiré de l'oiseau.

Ce qui fait l'avantage des architectures profondes, *c'est leur capacité d'apprendre à représenter le monde de manière hiérarchique*. Comme toutes les couches sont entraînables, nul besoin de construire un extracteur de caractéristiques à la main. L'entraînement s'en chargera. De plus, les premières couches extrairont des caractéristiques simples (présence de contours) que les couches suivantes combineront pour former des concepts de plus en plus complexes et abstraits : assemblages de contours en motifs, de motifs en parties d'objets, de parties d'objets en objets, etc.

Réseaux convolutifs, réseaux récurrents

Une architecture profonde particulièrement répandue est le *réseau convolutif*. C'est un peu mon invention. J'ai développé les premières versions en 1988-1989 d'abord à l'Université de Toronto ou j'étais post doctorant avec Geoffrey Hinton (qui travaille maintenant chez Google), puis aux Bell Laboratories, qui était à l'époque le prestigieux labo de recherche de la compagnie de télécommunication AT&T.

Les réseaux convolutifs sont une forme particulière de réseau neuronal multicouche dont l'architecture des connexions est inspirée de celle du cortex visuel des mammifères. Par exemple, chaque élément n'est connecté qu'à un petit nombre d'éléments voisins dans la couche précédente. J'ai d'abord utilisé les réseaux convolutifs pour la reconnaissance de caractères. Mes collègues et moi avons développé un système automatique de lecture de chèques qui a été déployé largement dans le monde dès 1996, y compris en France dans les distributeurs de billets du Crédit Mutuel de Bretagne. À la fin des années 90, ce système lisait entre 10 et 20 % de tous

les chèques émis aux États-Unis. Mais ces méthodes étaient plutôt difficiles à mettre en œuvre avec les ordinateurs de l'époque, et malgré ce succès, les réseaux convolutifs – et les réseaux neuronaux plus généralement – ont été délaissés par la communauté de la recherche entre 1997 et 2012.

En 2003, Geoffrey Hinton (de l'Université de Toronto), Yoshua Bengio (de l'Université de Montréal) et moi-même à NYU (l'Université de New York), décidions de démarrer un programme de recherche pour remettre au goût du jour les réseaux neuronaux, et pour améliorer leurs performances afin de raviver l'intérêt de la communauté. Ce programme a été financé en partie par la fondation CIFAR (l'Institut Canadien de Recherches Avancées), je l'appelle parfois « la conspiration de l'apprentissage profond ».

En 2011-2012 trois événements ont soudainement changé la donne. Tout d'abord, les GPUs (*Graphical Processing Units*) capables de plus de mille milliards d'opérations par seconde sont devenus disponibles pour moins de 1000 euros la carte. Ces puissants processeurs spécialisés, initialement conçus pour le rendu graphique des jeux vidéo, se sont avérés être très performants pour les calculs des réseaux neuronaux. Deuxièmement, des expériences menées simultanément à Microsoft, Google et IBM avec l'aide du laboratoire de Geoff Hinton ont montré que les réseaux profonds pouvaient diminuer de moitié les taux d'erreurs des systèmes de reconnaissance vocale. Troisièmement plusieurs records en reconnaissance d'image ont été battus par des réseaux convolutifs. L'événement le plus marquant a été la victoire éclatante de l'équipe de Toronto dans la compétition de reconnaissance d'objets « ImageNet ». La diminution des taux d'erreurs était telle qu'une véritable révolution d'une rapidité inouïe s'est déroulée. Du jour au lendemain, la majorité des équipes de recherche en parole et en vision ont abandonné leurs méthodes préférées et sont passées aux réseaux convolutifs et autres réseaux neuronaux.

L'industrie d'Internet a immédiatement saisi l'opportunité et a commencé à investir massivement dans des équipes de recherche et développements en apprentissage profond. L'apprentissage profond ouvre une porte vers des progrès significatifs en intelligence artificielle. C'est la cause première du récent renouveau d'intérêt pour l'IA.

Une autre classe d'architecture, les *réseaux récurrents* sont aussi revenus au goût du jour. Ces architectures sont particulièrement adaptées au traitement de signaux séquentiels, tels que le texte. Les progrès sont rapides, mais il y a encore du chemin à parcourir pour produire des systèmes de compréhension de texte et de traduction.

L'intelligence artificielle aujourd'hui. Ses enjeux

Les opportunités sont telles que l'IA, particulièrement l'apprentissage profond, est vue comme des technologies d'importance stratégique pour l'avenir.

Les progrès en vision par ordinateur ouvrent la voie aux voitures sans chauffeur, et à des systèmes automatisés d'analyse d'imagerie médicale. D'ores et déjà, certaines voitures haut de gamme utilisent le système de vision de la compagnie Israelienne MobilEye qui utilise un réseau

convolutif pour l'assistance à la conduite. Des systèmes d'analyse d'images médicales détectent des mélanomes et autres tumeurs de manière plus fiable que des radiologues expérimentés. Chez Facebook, Google et Microsoft, des systèmes de reconnaissance d'image permettent la recherche et l'organisation des photos et le filtrage d'images violentes ou pornographiques.

Depuis plusieurs années déjà, tous les moteurs de reconnaissance vocale sur smartphone utilisent l'apprentissage profond.

Des efforts considérables de R&D sont consacrés au traitement du langage naturel : la compréhension de texte, les systèmes de question-réponse, les systèmes de dialogue pour les agents virtuels, et la traduction automatique. Dans ce domaine, la révolution de l'apprentissage profond a été annoncée, mais n'est pas encore achevée. Néanmoins, on assiste à des progrès rapides. Dans la dernière compétition internationale de traduction automatique, le gagnant utilisait un réseau récurrent.

La recherche en intelligence artificielle et les obstacles au progrès

Malgré tous ces progrès, nous sommes encore bien loin de produire des machines aussi intelligentes que l'humain, ni même aussi intelligentes qu'un rat.

Bien sûr, nous avons des systèmes qui peuvent conduire une voiture, jouer aux échecs et au Go, et accomplir d'autres tâches difficiles de manière plus fiable et rapide que la plupart des humains (sans parler des rats). Mais ces systèmes sont très spécialisés. Un gadget à 30 euros nous bat à plate couture aux échecs, mais il ne peut faire rien d'autre.

Ce qui manque aussi aux machines, c'est la capacité à apprendre des tâches qui impliquent non seulement d'apprendre à représenter le monde, mais aussi à se remémorer, à raisonner, à prédire, et à planifier. Beaucoup de travaux actuels à Facebook AI Research et à DeepMind sont focalisés sur cette question. Une nouvelle classe de réseaux neuronaux, les *Memory-Augmented Recurrent Neural Nets* (réseaux récurrents à mémoire) est utilisée de manière expérimentale pour la traduction, la production de légendes pour les images, et les systèmes de dialogues.

Mais ce qui manque principalement aux machines, c'est le sens commun, et la capacité à *l'intelligence générale* qui permet d'acquérir de nouvelles compétences, quel qu'en soit le domaine. Mon opinion, qui n'est partagée que par certains de mes collègues, est que l'acquisition du sens commun passe par *l'apprentissage non supervisé*.

Qu'il soit naturel ou artificiel, il y a trois formes principales d'apprentissage. Nous avons déjà parlé de l'apprentissage supervisé. Les deux autres formes sont l'apprentissage par renforcement, et l'apprentissage non supervisé.

L'apprentissage par renforcement désigne la situation où la machine ne reçoit qu'un simple signal, une sorte de récompense, indiquant si la réponse produite était correcte ou pas. Le scénario est similaire à l'entraînement d'un animal de cirque à qui l'on donne une friandise lorsqu'il exécute l'action désirée. Cette forme d'apprentissage nécessite de très nombreux essais, et est utilisée principalement pour entraîner les machines à jouer à des jeux (par exemple les jeux

vidéo ou le jeu de Go), ou à opérer dans des environnements simulés. On a assisté à un succès éclatant de l'apprentissage par renforcement combiné à l'apprentissage profond lors de la victoire récente du programme de Go AlphaGo de DeepMind face au champion européen.

L'apprentissage non supervisé, quant à lui, est le mode principal d'apprentissage des animaux et des humains. C'est l'apprentissage que nous faisons par nous même en observant le monde et en agissant. C'est en observant le monde que nous apprenons qu'il a trois dimensions, que des objets peuvent en cacher d'autres, que certains objets peuvent être déplacés, qu'un objet sans support tombe, qu'un objet ne peut pas être à deux endroits en même temps, etc.

C'est grâce à l'apprentissage non supervisé que nous pouvons interpréter une phrase simple comme « Jean prend son portable et sort de la pièce ». On peut inférer que Jean et son portable ne sont plus dans la pièce, que le portable en question est un téléphone, que Jean s'est levé, qu'il a étendu sa main pour attraper son portable, qu'il a marché vers la porte. Il n'a pas volé, il n'est pas passé à travers le mur. Nous pouvons faire cette inférence, car nous savons comment le monde fonctionne. C'est le sens commun.

Comment acquérir ce sens commun ? Une hypothèse possible est *l'apprentissage prédictif*. Si l'on entraîne une machine à prédire le futur, elle ne peut y arriver qu'en élaborant une bonne représentation du monde et de ses contraintes physiques. Dans un scénario d'apprentissage prédictif, on montre à la machine un segment de vidéo, et on lui demande de prédire quelques images suivantes. Malheureusement, le futur est impossible à prédire exactement et la machine s'en tient à produire une image floue, une mixture de tous les futurs possibles.

Si l'intelligence est un gâteau au chocolat, le gâteau lui-même est l'apprentissage non supervisé, le glaçage est l'apprentissage supervisé, et la cerise sur le gâteau est l'apprentissage par renforcement. Les chercheurs en IA sont dans la même situation embarrassante que les physiciens : 95 % de la masse de l'univers est de nature complètement inconnue : matière noire et énergie noire. La matière noire de l'AI est la génoise au chocolat de l'apprentissage non supervisé.

Tant que le problème de l'apprentissage non supervisé ne sera pas résolu, nous n'aurons pas de machine vraiment intelligente. C'est une question fondamentale scientifique et mathématique, pas une question de technologie. Résoudre ce problème pourra prendre de nombreuses années ou plusieurs décennies. En vérité, nous n'en savons rien.

À quoi ressembleront les machines intelligentes de demain ?

Si nous arrivons à concevoir des techniques d'apprentissage machine aussi générales et performantes que celle de la nature, à quoi ressembleront les machines intelligentes de demain ?

Il est très difficile d'imaginer une entité intelligente qui n'ait pas toutes les qualités et les défauts des humains, car l'humain est notre seul exemple d'entité intelligente. Comme tous les animaux, les humains ont des pulsions et des instincts gravés dans notre cerveau reptilien par l'évolution pour la survie de l'espèce. Nous avons l'instinct de préservation, nous pouvons

devenir violents lorsque nous sommes menacés, nous désirons l'accès aux ressources pour ne pas mourir de faim, ce qui peut nous rendre jaloux, etc. Nos instincts d'animaux sociaux nous conduisent aussi à rechercher la compagnie d'autres humains. Mais les machines intelligentes n'auront aucune raison de posséder ces pulsions et instincts. Pour qu'elles les aient, il faudrait que leurs concepteurs les construisent explicitement.

Les machines intelligentes du futur auront des sentiments, des plaisirs, des peurs, et des valeurs morales. Ces valeurs seront une combinaison de comportements, d'instinct et de pulsions programmés avec des comportements appris.

Dans quelques décennies, quand nous pourrons peut-être penser à concevoir des machines réellement intelligentes, nous devrons répondre à la question de comment aligner les valeurs des machines avec les valeurs morales humaines.

Mais c'est un futur lointain où l'on pourra donner de l'autonomie aux machines. D'ici là, les machines seront certes intelligentes, mais pas autonomes. Elles ne seront pas à même de définir leurs propres buts et motivations. L'ordinateur de votre voiture s'en tiendra à conduire votre voiture en toute sécurité. L'IA sera un amplificateur de notre intelligence, et non un substitut pour celle-ci.

Malgré les déclarations de certaines personnalités, le scénario à la Terminator est immensément improbable. Tout d'abord, il faut garder à l'esprit que l'apparition de l'IA ne sera pas un événement singulier, ni le fait d'un groupe isolé. Le progrès de l'IA sera progressif et ouvert. Comprendre l'intelligence est une des grandes questions scientifiques de notre temps. Aucune organisation, si puissante soit-elle, ne peut résoudre ce problème en isolation. La conception de machines intelligentes nécessitera la collaboration ouverte de la communauté de la recherche entière.

Faut-il avoir peur de l'intelligence artificielle ?

L'IA n'éliminera donc pas l'humanité de sa propre initiative.

Mais comme toute technologie puissante, l'IA peut être utilisée pour le bénéfice de l'humanité entière ou pour le bénéfice d'un petit nombre aux dépens du plus grand nombre.

L'émergence de l'AI va sans doute déplacer des métiers. Mais elle va aussi sauver des vies (par la sécurité routière et la médecine). Elle va très probablement s'accompagner d'une croissance de la production de richesses par habitant. La question pour les instances dirigeantes est comment distribuer ces nouvelles richesses, et comment former les travailleurs déplacés aux nouveaux métiers créés par le progrès technologique. C'est une question politique et non technologique. C'est une question qui n'est pas nouvelle : l'effet du progrès technologique sur le marché du travail existe depuis la révolution industrielle. L'émergence de l'IA n'est qu'un symptôme de l'accélération du progrès technologique.

Yann LeCun