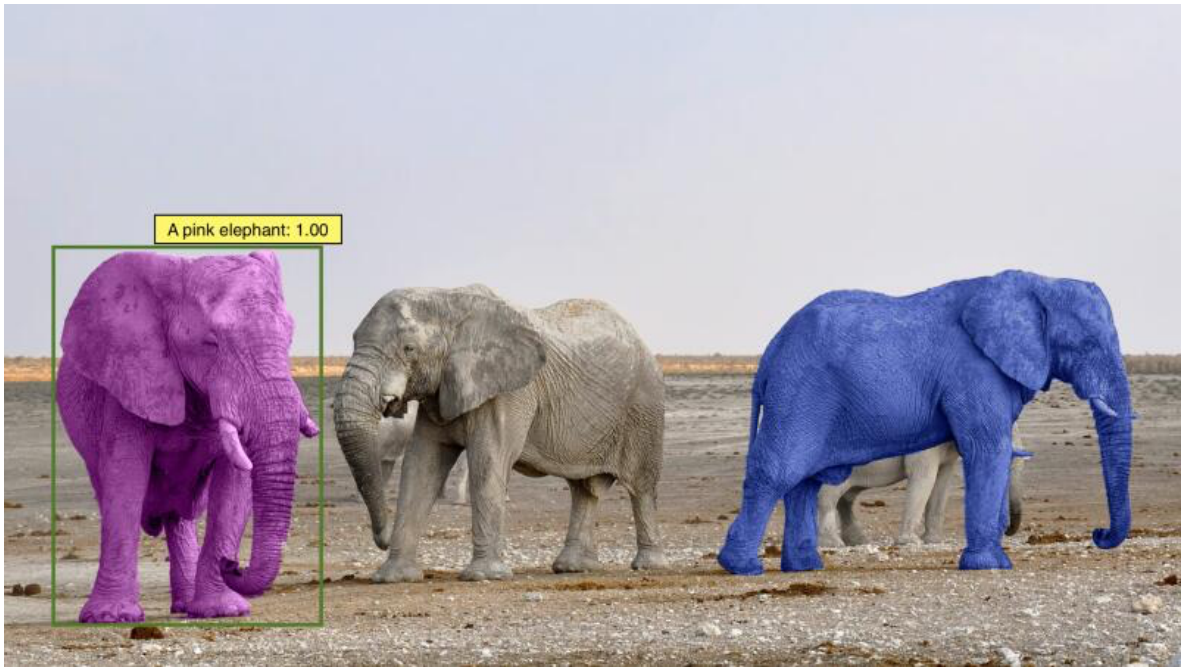
SCIENCES • INTELLIGENCE ARTIFICIELLE

La reconnaissance d'images par l'ordinateur fait un pas de géant

Après des années sans progrès majeurs, la vision par ordinateur à travers l'intelligence artificielle avance à nouveau. Les algorithmes ont de moins en moins besoin de la main humaine pour fonctionner et ouvrent la perspective de systèmes capables de réflexion.

Par David Larousserie

Publié le 20 juillet 2021 à 18h00, modifié le 22 juillet 2021 à 06h05 • Lecture 5 min.



Cet algorithme de Facebook (MDETR) reconnaît l'animal qualifié « d'éléphant rose » alors qu'il n'en a jamais vu lors de son entraînement. FACEBOOK AI RESEARCH

En 2012, une équipe de l'université de Toronto (Canada) surprenait le monde lors d'une compétition de reconnaissance d'images par ordinateur : 15 % d'erreurs seulement pour son logiciel contre 26 % pour le deuxième. C'était le début de la nouvelle vague de l'intelligence artificielle, dite « apprentissage profond » ou *deep learning,* car le programme, apparenté à un réseau de neurones artificiels connectés, trouve les bonnes « connexions » en s'entraînant sur des millions d'exemples.

Puis la vague s'est étendue aux jeux (go, échecs, poker), à l'automobile (conduite autonome), la voix (dans les assistants vocaux), la science (forme des protéines)… Mais, vedettes des premiers jours, les images ont vu passer les trains suivants des progrès, avec des performances qui plafonnaient. Jusqu'à ces derniers mois.

« Je dois dire que je n'ai pas été autant excité dans ce domaine depuis dix ou vingt ans ! », a expliqué Yann LeCun, responsable scientifique chez Facebook et pionnier du *deep learning* depuis trente ans, lors d'une présentation à la presse le 30 juin des dernières avancées de la recherche du géant californien. *« Ça va très vite. Il y a deux ans, il n'y avait rien de neuf »,* confirme Matthieu Cord, professeur à Sorbonne Université et chercheur chez Valeo.

Lire aussi | Comment le « deep learning » révolutionne l'intelligence artificielle

Le changement est lié à plusieurs innovations permettant de corriger les défauts des premières méthodes. *« La clé du succès des premières techniques est ce que l'on appelle l'apprentissage supervisé. C'est-à-dire que le programme apprend ses paramètres, grâce à des données annotées par des humains »,* précise Jean Ponce, professeur d'informatique à l'Ecole normale supérieure. Pour « reconnaître » un chat, un chien ou une voiture, des milliers d'images légendées « chat » ou « chien » ou « voiture » sont montrées au programme qui adapte ses paramètres afin de trouver la bonne réponse. Ensuite, même sur des images inconnues, il donne la bonne réponse.

L'apprentissage autosupervisé

Le principal problème est que la technique nécessite énormément d'images légendées. En outre, la diversité des situations réelles est telle qu'il est impossible de la représenter avec des bases de données d'images validées par des petites mains. *« Les performances des systèmes de vision des voitures autonomes s'effondrent si on montre des images de nuit ou bien de chiens mouillés »,* constate Matthieu Cord.

Quelques parades ont bien été tentées, comme la génération d'images de synthèse plus vraies que nature grâce à… l'intelligence artificielle. En 2014, un autre pionnier du *deep learning,* Yoshua Bengio, et ses collègues de l'université de Montréal proposent ainsi un algorithme fabriquant de fausses images. Le système utilise en fait deux sous-systèmes, l'un proposant des images et l'autre essayant de savoir si elles sont « bonnes » ou « mauvaises », et en cas de mauvaise réponse, le système artiste modifie sa proposition, etc. A l'issue de la « compétition », les images (visages, objets, tableaux de maître…) trompent même des humains. Mais une fois encore, la diversité du monde est trop grande.

Lire aussi | GPT-3, l'intelligence artificielle qui a appris presque toute seule à presque tout faire

Arrive alors, en 2017, une idée majeure, toujours cosignée de Yoshua Bengio. Il s'agit d'une nouvelle organisation des paramètres des réseaux de neurones, appelée *transformer*. Elle va ouvrir la voie aux succès de l'apprentissage dit « autosupervisé », c'est-à-dire que le système apprend sur de vastes bases de données non légendées. La « magie » de l'idée est que l'algorithme apprend sur des tâches un peu idiotes, comme remplir des phrases à trous, mais que cela construit une « représentation » abstraite

des textes, ce qui permet ensuite de réaliser des tâches plus utiles comme répondre à des questions, identifier la tonalité d'un texte, écrire des textes originaux…

Cela a fini par marcher, même pour les images. En octobre 2020, Google fabrique un *transformer* meilleur que ceux de ses concurrents de l'apprentissage supervisé, mais au prix d'une digestion de 300 millions d'images et de plus de 600 millions de paramètres (dix fois plus qu'en 2012). Deux mois plus tard, une équipe de Facebook et de Matthieu Cord égale ces performances, mais avec sept fois moins de paramètres. De ce travail est sorti un outil de « segmentation », Dino, c'est-à-dire un logiciel qui suit un objet dans une vidéo.

Course à l'innovation

Néanmoins, les *transformers* ne sont pas les seuls à avoir réussi des percées et de nouvelles méthodes sont apparues, plus rapides et sans doute moins gourmandes en données. Deux familles puisent leurs racines dans de « vieux » travaux des années 1990 de Yann LeCun et du troisième pionnier du domaine, Geoffrey Hinton. En février 2020, ce dernier, désormais chez Google, dévoile SimCLR, qui égale les méthodes classiques supervisées grâce à un apprentissage futé, mais avec 100 fois moins d'images annotées. L'entraînement consiste à apprendre à l'algorithme à classer comme semblables des images modifiées (par floutage, par zoom, par déformation des couleurs). Dans cette famille, le concurrent Facebook a aussi proposé des logiciels comme MOCO ou PIRL.

Yann LeCun a en outre lui aussi ressorti de ses cartons une vieille idée dite « des réseaux siamois ». Deux réseaux de neurones en parallèle se comparent et servent à construire une représentation abstraite des objets, là aussi à partir de transformations contrôlées des images. Le résultat, SEER, égale les meilleurs modèles supervisés et les dépasse dès lors que des images « atypiques » sont proposées. D'autres entreprises ont également récemment publié des algorithmes apparentés à cette technique, comme BYOL de Deepmind (filiale d'Alphabet, la maison mère de Google) ou OBOW de Valeo. En quelques mois, alors qu'il n'y avait rien, plusieurs architectures sont donc en concurrence et prêtes à sortir des laboratoires.

Lire aussi | « Désormais, face aux avancées de l'intelligence artificielle, l'avis des philosophes compte »

« Jusqu'à présent, les programmes étaient en réalité des idiots savants. Ils mettaient une étiquette sur un chat, mais sans comprendre qu'il s'agit d'un chat », rappelle Jean Ponce. *« Ils ont l'esprit étroit et peuvent faire des erreurs stupides. Par exemple, ils ne reconnaissent pas les vaches sur une plage car ils ont souvent vu des vaches dans un pré vert et associent en réalité la couleur à l'animal »,* note Yann LeCun. *« Nous faisons le pari que ces nouveaux systèmes feront mieux que les précédents dans des situations complexes »,* espère Jean Ponce. Voire qu'ils se rapprocheront d'une véritable compréhension du monde réel en élaborant une sorte de sens commun, qui fait par exemple que très vite un enfant devine qu'un objet en équilibre va tomber.

« Le but est maintenant de doter ces algorithmes de capacités à raisonner », estime Matthieu Cord, titulaire d'une chaire récente intitulée « Raisonnement visuel ». *« Une question est de savoir quel paradigme d'apprentissage utilise le cerveau pour construire son sens commun. Je pense que ça doit ressembler à l'autoapprentissage »,* rappelle Yann LeCun, qui a baptisé ce paradigme « matière noire de l'intelligence artificielle », pour signifier qu'il reste mystérieux mais qu'il est probablement majoritaire, comme l'est la matière noire dans l'Univers.

David Larousserie