DU TEXTE À L'IMAGE : CAP FRANCHI POUR L'IA

Les nouvelles technologies en intelligence artificielle génératrice offrent la possibilité de créer des images sur tout sujet et dans tout style. Quels sont les ressorts de ces capacités qui s'apparentent à de la créativité?

L'AUTEUR

JEAN-PAUL DELAHAYE
professeur émérite à l'université de Lille et chercheur au laboratoire Cristal (Centre de recherche en informatique, signal et automatique de Lille)

Qu'elles se nomment DALL-E, Midjourney ou Stable Diffusion, les intelligences artificielles (IA) qui engendrent des images à partir de textes offrent aujourd'hui des résultats d'une qualité étonnante. Le principe général pour obtenir ces outils est de les nourrir d'une multitude d'images annotées, issues par exemple d'internet. L'IA en déduit des liens entre images et suites de mots et, une fois acquise cette «compréhension» du rapport entre mots et images, la machine travaille sans aide pour transformer en images vos descriptions – appelées «invites» (ou *prompts*, en anglais).

Pendant la phase d'apprentissage, on nourrit l'IA d'un très grand nombre d'exemples de couples texte-image – les corpus varient de quelques centaines de milliers à quelques centaines de millions de couples. Mais bien sûr, ce n'est pas suffisant, et il a fallu élaborer des méthodes spéciales pour que le système devienne capable d'effectuer les tâches délicates qu'on attend de lui. On attend des nouvelles IA qu'elles analysent des descriptions complexes, comme «une ville sur une petite planète avec des avions au-dessus et un gros ballon» ou «un chat habillé en moine dans une forêt, style gravure ancienne», et qu'elles créent des images inédites répondant à la description (*voir l'encadré 2*). La version 6 de l'IA Midjourney, sortie en décembre dernier, est même en mesure de prendre en compte des *prompts* bien plus longs et littéraires, comme celui ayant permis d'obtenir l'image page ci-contre.

Ce n'est que très récemment que ces méthodes ont été perfectionnées. La technique de débruitage d'images, dont on verra plus loin qu'elle est centrale dans ces systèmes générateurs d'images, date de 2020. Le développement par la société OpenAI, en 2021, du réseau de neurones CLIP – pour *Contrastive language-image pre-training* («Pré-entraînement à la discrimination des couples texte-image») – a également marqué un tournant dans le domaine.

Venons-en aux idées qui ont conduit à cette réussite inattendue de l'intelligence artificielle.

COUPLES TEXTE-IMAGE

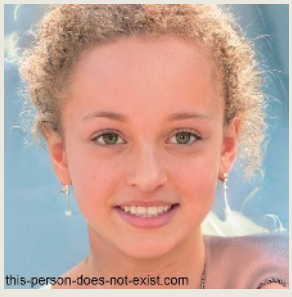
Les systèmes aujourd'hui capables de maîtriser les liens entre des mots et des images reposent sur des méthodes d'apprentissage profond (ou deep learning) utilisant plusieurs réseaux de neurones artificiels. Le réseau principal sur lequel s'appuie, par exemple, l'IA génératrice d'images DALL-E 2 (baptisé d'un jeu de mots formé avec le nom du peintre surréaliste Salvador Dalí et celui du personnage de dessin animé Wall-E) est le fameux CLIP. Son principe général, avec des variantes, est aussi utilisé par les autres systèmes donnant aujourd'hui les meilleurs résultats.

Jean-Paul Delahaye a récemment publié : **Au-delà du Bitcoin** (Dunod, 2022).

Image générée avec l'IA Midjourney par Benoît Raphaël, à l'aide du *prompt* suivant (initialement en anglais, ici traduit en français) : « Dans une rue bondée de Canggu, à Bali, un homme blanc de 40 ans aux cheveux châtains très courts, à la barbe d'un jour et aux yeux verts, marche aux côtés d'une superbe femme balinaise de 35 ans. Les cheveux de la femme sont légèrement ébouriffés, incarnant une beauté candide. Ils se frayent un chemin à travers un labyrinthe de scooters, avec des volutes de fumée d'un feu de rue s'élevant dans l'air et des vendeurs ambulants de *bakso* en arrière-plan. Ils ont une discussion animée, et ne sont pas encore remis d'une situation comique récente. Leurs expressions mêlent amusement et surprise, comme s'ils venaient tout juste de vivre un événement joyeux. La caméra capture ce moment spontané, se concentrant sur leurs visages, mettant en évidence les détails de leurs expressions sur fond de rue animée et enfumée. Cette scène, empreinte de vie et de rires inattendus, est capturée dans un style brut et candide. »
Source : https://generationia.flint.media/p/reinventez-vos-prompts-pour-exploiter-puissance-midjourney-6

1 L'IA ET LES IMAGES : PREMIÈRES MÉTHODES

Deux types d'applications des réseaux de neurones ont été des précurseurs des réussites actuelles de l'IA génératrice d'images.
(a) La reconnaissance d'images pour en permettre le tri est devenue classique. Elle se fonde sur des réseaux de neurones qu'on entraîne en leur présentant un grand nombre d'images et en leur indiquant, pour chacune, à quelle catégorie elle appartient parmi un nombre fini de possibilités. Une fois cet apprentissage effectué, le réseau est capable de déterminer, pour une image qu'il n'a jamais vue, si elle représente « un chat », « une voiture », etc. Notez que, contrairement à ce qui est fait pour CLIP et qui est présenté dans l'article, cette méthode ne fait qu'identifier à quelle catégorie appartient une image, en se limitant à un petit nombre d'options fixées une fois pour toutes. Cela permet, par exemple, la lecture de caractères manuscrits ou le traitement automatique d'images médicales de la peau pour y détecter des mélanomes et évaluer leur gravité.
(b) Plus récemment, en 2014, une équipe menée par Ian Goodfellow, alors affilié à l'université de Montréal, au Canada, a développé la méthode GAN (pour *Generative adversarial network* ou *Generative adversarial nets*, en anglais), conçue pour créer des images. Cette technique met en compétition deux réseaux de neurones, et exploite une famille d'images de référence – par exemple, des photos de visages. L'un des réseaux génère une image en essayant d'inventer un visage, l'autre tente de distinguer à l'aveugle, entre une image de la base de référence et l'image inventée, laquelle est la plus crédible. À chaque étape, les réseaux sont informés du gagnant, et les paramètres sont mis à jour – on retrouve la classique mécanique de descente de gradient. L'objectif est de faire en sorte que les images produites égalent ou surpassent le corpus d'entraînement : une fois l'apprentissage simultané des deux réseaux terminés, celui qui, au départ, proposait des images médiocres de visages est devenu capable d'en « inventer » de très convaincantes, susceptibles de tromper le réseau de reconnaissance. Le site This person does not exist (https://this-person-does-not-exist.com/en) propose ainsi de créer des photos de personnes qui n'existent pas *(voir ci-dessous)*.



Pour entraîner le réseau de neurones CLIP, on lui présente une image et sa description – par exemple: «photographie d'un chien tenu en laisse par son maître» – et le réseau ajuste ses paramètres de sorte que, lorsqu'on lui remontrera le même couple texte-image, il le considèrera comme valide (il identifiera que le texte et l'image correspondent). L'ajustement des paramètres se fait selon le principe de la descente de gradient, que nous avons expliquée il y a deux mois dans l'article de cette rubrique sur les modèles massifs de langage, (*voir* Pour la Science *n° 555*). On recommence un très grand nombre de fois avec d'autres images associées à leurs descriptions. On présente aussi des couples incorrects, que le réseau apprend à considérer comme tels.

Le réseau CLIP a été conçu et décrit par Alec Radford et son équipe en 2021. Sa phase d'apprentissage, longue et coûteuse, s'est faite en grande partie de manière autosupervisée – c'est-à-dire sans intervention humaine. Cette façon de procéder a permis d'assimiler beaucoup de couples texte-image déjà disponibles en ligne ou dans des bases de données spécifiques. Pour entraîner CLIP, ce sont en effet 400 millions de couples texte-image issus «de différentes sources publiquement disponibles sur internet» qui ont été exploitées – une quantité de données gigantesque, quand on sait que le plus gros ensemble public d'images annotées utilisé dans le domaine, le corpus ImageNet, comporte «seulement» 14 millions de couples texte-image. Notez, en comparaison, que si vous regardez attentivement 1 000 images par jour pendant dix ans, vous n'en aurez même pas analysé 4 millions! Les données d'entraînement ne sont pas systématiquement fournies de manière explicite par les équipes qui développent les IA. Cette opacité fait débat, en particulier parce que certaines entreprises ont été soupçonnées d'y inclure des textes ou des

2 DU TEXTE À L'IMAGE

Pour bien mesurer la puissance des nouvelles IA génératrices d'images, voici quelques exemples de *prompts* et des résultats correspondants, obtenus avec le site https://www.bing.com/images/create, qui utilise DALL-E.

Une course de pères Noël à vélo dans un désert.

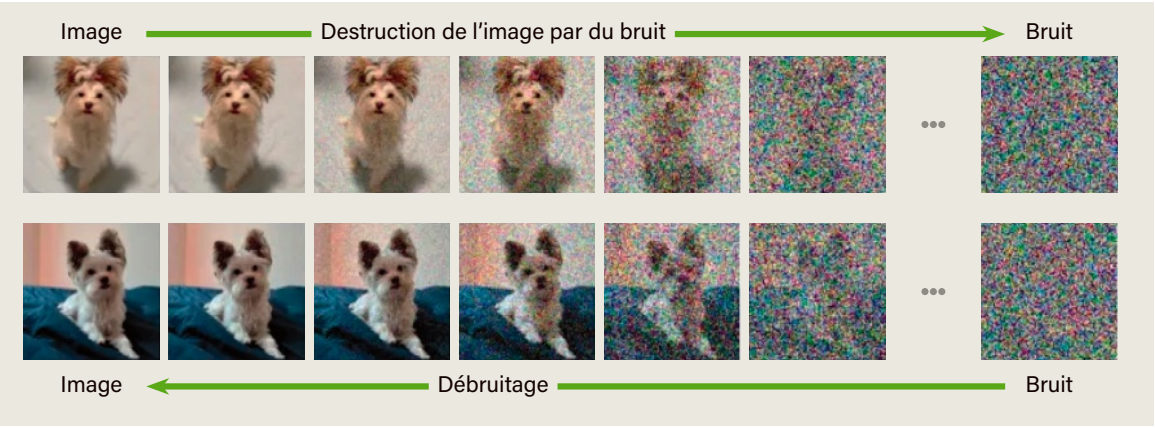
Un chat habillé en moine dans une forêt, style gravure ancienne.

Une petite église dans une montagne enneigée avec des skieurs, peinture à l'huile.

L'explosion d'un avion de ligne au-dessus d'une plage.

Une ville sur une petite planète avec des avions au-dessus et un gros ballon.

Une tour de Babel avec des livres et des personnes humaines.



En ajoutant progressivement du bruit à une image, on apprend à un réseau de neurones à reconnaître un couple image d'origine-image bruitée. Il peut ensuite utiliser cette connaissance pour « débruiter » un mélange aléatoire de pixels afin de générer une image.

images qu'elles n'avaient pas le droit d'exploiter pour cet usage – typiquement, des œuvres artistiques protégées par le droit d'auteur, dont les IA se révèlent pourtant capables d'imiter le style, une fois entraînées.

Pour fonctionner, CLIP utilise un « encodeur » de texte qui associe à chaque mot une série de nombres constituant un point dans un espace de grande dimension, appelé « espace sémantique ». On parle de plongement du texte et de l'image dans l'espace sémantique de CLIP, ou d'*embedding*, en anglais. Le passage par cet espace où des mots ayant des sens proches sont associés à des points proches de l'espace offre au système une forme de compréhension générale du langage. Sans apprentissage supplémentaire, cela permet au réseau de traiter des textes et des images qui ne lui ont jamais été présentés : on le nomme « apprentissage *zero shot* ».

APPRENDRE DES CONCEPTS

Le travail d'apprentissage permet d'établir un lien entre des images et des textes, ce qui revient à disposer d'un espace sémantique commun pour les deux. C'est cet espace sémantique, qui est structuré conformément à ce qu'un être humain comprend, qui donne au système la possibilité d'estimer si un texte quelconque correspond bien à une image. Lorsqu'un couple non utilisé pendant l'apprentissage lui est présenté, il est capable d'évaluer, sous la forme d'une probabilité de pertinence, la qualité de ce texte en tant que descripteur de l'image. Le système a en quelque sorte appris des concepts liés aux mots et aux images, et le lien entre les deux. Malheureusement, comme souvent avec les réseaux de neurones, il est impossible en pratique d'aller voir quels concepts ont été appris et ce que cela signifie pour les paramètres du réseau. CLIP reste donc en partie une boîte noire, qui fait assez bien ce qu'on attend – évaluer des couples texte-image quelconques – mais dont on ignore pourquoi précisément.

Ce que nous avons présenté jusqu'à présent ne crée pas d'image : il faut donc compléter le système. Pour cela, on exploite l'espace sémantique commun aux textes et aux images, qu'on associe à une méthode de synthèse d'images employant la technique de « diffusion probabiliste », introduite en 2020 par une équipe autour de Jonathan Ho, alors à l'université Berkeley, en Californie.

En partant d'un mélange aléatoire de pixels, le système génère une image

Celle-ci consiste à apprendre à un réseau de neurones comment se transforme une image quand on la brouille en y introduisant du bruit. Là encore, cet apprentissage se fait sur un grand nombre d'exemples créés automatiquement de couples image d'origine-image bruitée. On peut ensuite utiliser ce que le réseau a appris pour lui faire retirer du bruit d'une image quelconque : c'est cette capacité de débruitage qui va engendrer des images. En partant d'un mélange aléatoire de pixels, qui ne représente donc rien, le système met en œuvre le débruitage plusieurs fois de suite : cela génère des images de plus en plus précises, qui finissent par représenter quelque chose. Lorsque ce débruitage est fait en

exploitant le point de l'espace sémantique que CLIP associe au texte donné décrivant l'image qu'on veut obtenir, il compose des images en lien avec la description: on parle de «débruitage conditionnel». Une fois arrivé à une image débruitée, le système la retravaille encore pour la rendre plus fine et détaillée, à la manière de ce qui est présenté dans l'image en haut de la page 83.

En résumé: à partir d'une grande quantité couples texte-image, on constitue un espace sémantique établissant de manière pertinente une relation entre images et textes. Puis, une technique de débruitage conditionnel se servant du plongement du texte donné par l'utilisateur dans l'espace sémantique produit une ou plusieurs images qui, après traitements complémentaires, correspondent au texte soumis.

Dans une étude, les meilleures idées humaines ont souvent dépassé celles des IA

Précisons encore que, comme cela est fait pour les grands modèles de langage (à l'origine des chatbots ChatGPT, Bard, etc.), la construction décrite ci-dessus est suivie d'une étape de réglages fins (ou *fine tuning*, en anglais) et d'un «renforcement de l'apprentissage par réactions humaines» (*reinforcement learning from human feedback*, RLHF), qui améliorent la capacité de l'IA à effectuer les tâches que l'on attend d'elle et s'assurent que les réponses proposées soient acceptables au regard de nos normes – qu'elles ne contiennent pas d'éléments mettant en cause des personnes particulières ni d'éléments trop violents ou à caractères sexuels, par exemple. Par ailleurs, on évite aussi de présenter au réseau de neurones, pendant la phase d'apprentissage, certaines images qui pourraient amener l'IA à en produire d'inadaptées, et on refuse certains textes que l'IA repère comme risquant d'engendrer des images non désirées. Le monde mis dans la tête de l'IA est en quelque sorte soumis à la censure, comme celui d'un enfant à qui on cache les réalités dont on ne veut pas qu'il ait connaissance. Il faut noter que les conditions dans lesquelles sont réalisées les phases nécessitant des interventions humaines font débat: en janvier 2023, le *Time Magazine* publiait ainsi une grande enquête dénonçant l'exploitation de travailleurs kényans sous-payés à qui l'on demandait de lire des contenus d'une extrême violence afin de les étiqueter comme «violents», «sexuels», «haineux», etc., et d'en nourrir ensuite ChatGPT.

DÉFORMATION CONTINUE

On remarque que dans les textes donnés à l'IA, on peut préciser des styles d'images souhaités: «gravure ancienne», «aquarelle», «art nouveau», etc. Il est aussi possible de lui demander de remplir une partie effacée d'une image. Bien sûr, le système ne fera pas réapparaître le contenu précis de l'image d'origine, mais il comblera le vide avec quelque chose en bonne harmonie avec le reste de l'image. De même, on peut utiliser ces IA génératrices pour élargir une image sur la droite, sur la gauche, au-dessus et en dessous. C'est ce qu'on nomme l'«extension d'image» (ou *out-painting*, en anglais): l'IA propose un univers plus grand que celui de l'image qu'on lui confie, mais plausible et compatible avec cette dernière. Récemment, cette fonctionnalité a d'ailleurs fait polémique quand une IA a été utilisée pour compléter un tableau de l'artiste américain Keith Haring... que ce dernier avait volontairement laissé inachevé.

Une autre utilisation des IA génératrices d'images est l'interpolation. Le «morphing» est une technique qui permet de passer continument d'une première image à une seconde – par exemple, de transformer un visage en un autre. Ce procédé existait déjà pour les films

« Morphing » entre une maison victorienne et une maison moderne, réalisé avec l'aide du réseau de neurones CLIP.

en argentique, mais il s'est considérablement amélioré avec le développement de l'informatique : ce qui est obtenu avec les nouvelles IA génératrices d'images dépasse largement tout ce qu'on savait faire avant. La transformation d'une image de maison victorienne en une image de maison moderne, réalisée par OpenAI avec l'aide de CLIP et présentée ci-dessus, l'illustre parfaitement : c'est toute la structure en trois dimensions qui évolue d'une image à l'autre, en donnant à chaque étape l'image cohérente d'une maison possible.

Pour réaliser cette prouesse, on exploite à nouveau l'espace sémantique créé pendant la phase d'apprentissage : on considère les points associés aux images de départ et d'arrivée souhaitées, et on construit dans l'espace sémantique un chemin continu entre les deux extrémités. On applique ensuite à chacun des points du chemin la méthode de débruitage conditionnel, qui permet de produire des images intermédiaires cohérentes, et le tour est joué.

L'HUMAIN VS LA MACHINE

Qu'il s'agisse de l'aptitude à comprendre un petit texte pour en tirer une image, de la capacité d'agrandir une image donnée en y introduisant des éléments nouveaux ou de transformer progressivement une image en une autre, dans toutes ces compétences la génération d'images semble bien égaler, voire dépasser l'humain, comme elle l'a fait précédemment dans ces autres domaines :

(a) Le calcul numérique (depuis la première moitié du XX^{e} siècle) ;

(b) La mémorisation de grandes quantités de données (également depuis la première moitié du XX^{e} siècle) ;

(c) La manipulation de formules mathématiques par les systèmes de calcul formel (plus récemment : on peut par exemple citer le logiciel Maxima, développé à la fin des années 1960) ;

(d) L'aptitude à certains jeux comme les échecs et le go (depuis 1978 pour les échecs et 2016 pour le go) ;

(e) La possibilité de produire, résumer, traduire à peu près convenablement toutes sortes de textes (depuis deux ou trois années avec l'explosion des capacités des modèles massifs de langage).

OÙ EST LA CRÉATIVITÉ ?

Au-delà des questions éthiques que leur utilisation soulève, ces systèmes d'IA nous amènent à nous interroger sur ce qu'est la créativité. Certains prennent le parti d'exploiter les nouveaux outils offerts par les IA pour tenter de générer ce qu'on peut, peut-être, qualifier de nouvelles formes d'art – comme cela a été le cas, historiquement, avec la photographie et le cinéma. La potentielle « créativité » des nouvelles IA elles-mêmes est un sujet controversé. Pour tenter de verser des éléments objectifs aux pièces du débat, une belle étude mettant en compétition la créativité humaine et celle des IA a été menée par Mika Koivisto, du département de Psychologie de l'université de Turku, en Finlande, et Simone Grassini, du département de Psychologie sociale de l'université de Bergen, en Norvège. Ils ont conçu un protocole pour mesurer la créativité de 256 sujets humains et la comparer à celle de trois robots conversationnels récents, dont ChatGPT-4. Ce type de test, assez classique en psychologie, consiste à demander aux participants d'imaginer des usages inhabituels, humoristiques ou étranges, d'objets courants : des crayons, des ficelles, des boîtes, etc. L'étude atteste d'une certaine réussite des nouvelles IA : en moyenne, les robots conversationnels ont obtenu de meilleurs résultats que les sujets humains. Souvent, les réponses humaines s'appuyaient sur des idées assez classiques, alors que les IA formulaient des propositions plus inattendues. Toutefois, les meilleures idées humaines ont souvent dépassé celles des IA. Ouf !

La conclusion proposée par les deux chercheurs ne veut pas nous désespérer : « Cette étude met donc en évidence l'indubitable potentiel des IA en tant qu'outil d'amélioration de la créativité, mais souligne également la nature unique et complexe de la créativité humaine, qu'il semble difficile de reproduire ou de dépasser avec la technologie de l'IA. » ■

BIBLIOGRAPHIE

D. Louapre, *Comment ces IA inventent-elles des images ?*, vidéo YouTube, 2023.

M. Koivisto et S. Grassini, Best humans still outperform artificial intelligence in a creative divergent thinking task, *Scientific Reports*, 2023.

L. Yang *et al.*, Diffusion models : A comprehensive survey of methods and applications, sur *arXiv*, 2023.

A. Ramesh *et al.*, Hierarchical text-conditional image generation with CLIP latents, sur *arXiv*, 2022.

A. Radford *et al.*, Learning transferable visual models from natural language supervision, Proceedings of the 38th International Conference on Machine Learning, 2021.

J. Ho *et al.*, Denoising diffusion probabilistic models, *Advances in Neural Information Processing Systems*, 2020.