RadaЯ
ISSN : 2825-9696

4 | 2019
Le détail : un dispositif du regard

Les mots, les choses et les images. Apprendre à voir à une machine

Vivien Philizot

https://www.ouvroir.fr/radar/index.php?id=212

DOI : 10.57086/radar.212

Electronic reference
Vivien Philizot, « Les mots, les choses et les images. Apprendre à voir à une machine », *RadaЯ* [Online], 4 | 2019, Online since 01 janvier 2019, connection on 16 mars 2025. URL : https://www.ouvroir.fr/radar/index.php?id=212

Les mots, les choses et les images. Apprendre à voir à une machine

Vivien Philizot

OUTLINE

TEXT

Un problème de culture visuelle

1 Le 15 mars 2019, un homme armé de plusieurs fusils d'assaut se rend dans deux mosquées de la ville néo-zélandaise de Christchurch et ouvre le feu sur toutes les personnes qu'il croise. L'attentat est filmé par son auteur à l'aide d'une caméra GoPro qui lui permet d'en diffuser les images en direct sur Facebook. La diffusion dure 17 minutes, pendant lesquelles ni les 3000 modérateurs de l'entreprise, ni les intelligences artificielles censées repérer des « contenus inappropriés » ne réagiront[1]. Ces dernières semblent bien moins sensibles à des images d'attentat qu'à des représentations de tétons, qu'elles reconnaissent et suppriment avant même leur mise en ligne.

2 On peut s'interroger sur les choix discutables qui conduisent Facebook à censurer certains contenus plutôt que d'autres, et expliquer ces 17 minutes par la tolérance de l'entreprise à l'égard de nombreuses autres formes de représentations de scènes de violence tout à fait communes, comme les fictions cinématographiques. Mais dans le prolongement de cette question, le filtrage des contenus sur ce type de plateforme se présente comme une tâche bien compliquée. En 2018, 243 000 photos ont été ajoutées chaque minute en moyenne. Il y avait au total fin 2018, environ 240 milliards d'images sur Facebook[2]. Face à un tel volume, les yeux humains ne sont pas

assez nombreux ni assez rapides pour faire un travail de discrimination qui a donc été confié à des machines, plus précisément des intelligences artificielles, dont les performances, fruit d'un apprentissage profond (*deep learning*), sont tant vantées depuis quelques années. Les réseaux de neurones artificiels semblent capables de voir à notre place et de prendre des décisions sur le produit de leurs observations. Or les résistances que nous opposent depuis des siècles, à nous humains, les images et leur interprétation, ne semblent pas tomber si facilement. En quoi la vidéo d'un attentat se distingue-t-elle d'une séquence de jeu vidéo ? Et en quoi le téton qui pointe dans une image pornographique se distingue-t-il de celui d'une vénus allongée dans une peinture de la Renaissance ? Deux images semblables peuvent véhiculer des idéologies fondamentalement différentes. Et il n'y a aucune raison de croire que les machines s'en sortent mieux que les humains face à un problème qui, malgré les milliers de pages que la philosophie ou l'esthétique lui ont consacrées, reste entier pour nous. Pour le dire à cette étape introductive de manière simple, les images convoquent un voir qui est aussi un savoir, et qui nous empêche de décider positivement de leur sens sur une seule base phénoménologique ou perceptive. Le « contenu sémantique », ce Graal des réseaux sociaux, semble bien résister au regard machinique, pour lequel une image n'est, après tout, qu'une suite de chiffres. C'est pourtant cette question qui a constitué le moteur des recherches en intelligence artificielle dans le domaine visuel. Comment le sens vient-il à l'image ? Ou comment passer de l'image au sens, alors qu'une caméra numérique ne peut « voir » que des pixels – c'est-à-dire des chiffres ?

3 Avant de faire quelques remarques méthodologiques sur cette question, j'aimerais préciser ici qu'elle se formule simultanément de manière inverse : comment passer du sens à l'image ? Car il ne s'agit pas seulement pour les machines de « voir », mais aussi de concevoir, de produire du visuel. Il faut se rendre à l'évidence : depuis maintenant quelques années, notre environnement est occupé par des images totalement nouvelles, non pas dans leurs apparences ou leurs sujets – ces images sont on ne peut plus banales : chats, voitures, maisons – mais dans leur mode de production. Ces images sont inédites car elles ne sont pas produites par des humains mais par des réseaux profonds de neurones. Mis en ligne récemment, le site *ThisPersonDoesnotExist* présente à chaque chargement de la page

d'accueil une image d'une personne qui « n'existe pas », entièrement construite de toutes pièces par une intelligence artificielle. De la même manière, d'autres réseaux de neurones peuvent dorénavant produire des représentations plutôt crédibles de façades de maison, d'animaux ou d'objets divers à partir de quelques indications de départ. Mais à partir de quoi ces représentations sont-elles construites ? Comment une machine peut-elle connaître une catégorie d'objets au point d'être capable d'en produire une image ?

4 Apprendre à voir et à dessiner à une machine nous impose de redéfinir ce que le « voir », comme processus nécessairement imprégné par un savoir, peut bien vouloir dire. Si ce problème peut sembler ne se poser que depuis ses disciplines afférentes – les mathématiques, l'informatique ou les sciences cognitives – il s'avance avec insistance sur les terrains éthique et sociologique, car l'usage de la vision artificielle s'étend, au-delà de mes exemples introductifs, à des activités de plus en plus nombreuses et diversifiées, comme la conduite automatique, le diagnostic médical, les procédés de contrôle industriels, la surveillance. En portant sur les processus d'apprentissage et notre façon de construire le monde par le regard, ce problème relève par ailleurs de la psychologie de la perception, lorsqu'elle dialogue avec l'esthétique et la philosophie. Irréductible à un champ du savoir spécifique, le problème de la vision des machines est fondamentalement un problème de culture visuelle. Et c'est à ce titre qu'il nous concerne. Or comme le dit très bien l'artiste et géographe Trevor Paglen, ce que ce problème est en train de transformer, c'est la culture visuelle elle-même, lorsqu'elle se détache des yeux humains, pour aller s'adresser prioritairement à des machines[3] À l'image du cerveau, les réseaux de neurones sont souvent mobilisés comme des boîtes noires sur lesquelles nous projetons toutes sortes d'inquiétudes ou d'espérances et qui semblent résister à la moindre tentative de description. Pour tenter de comprendre ce que « voir » peut bien vouloir dire pour une machine, je vais interroger dans ce texte la manière dont les réseaux profonds de neurones apprennent à lier ensemble le langage, le monde et la pensée.

Du détail à la vue d'ensemble : la construction du regard machinique

5 C'est bien la trace d'une facture, d'une couture ou d'une brisure (dans la texture des pixels) que le regard cherche invariablement dans le détail des portraits de ces « personnes qui n'existent pas[4] », comme dans ces images d'objets ou d'animaux.

Portraits et détails générés par des réseaux antagonistes génératifs
https://thispersondoesnotexist.com/. Goodfellow, Ian J., Jean Pougeat-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, et Yoshua Bengio. 2014. « Generative Adversarial Networks ». arXiv:1406.2661 [cs, stat], juin.

http://arxiv.org/abs/1406.2661

6 La nature de ces images impossibles semble résider dans des « traces » : une oreille mal placée dans un visage, une dent curieusement orientée dans la gueule d'un chat, une fenêtre étrangement raccordée à un toit dans une architecture… ces détails dissonants s'offrent comme autant de raisons de douter de ce que nous voyons, de chercher à distinguer le vrai du faux.

Série de représentations de chats produites par des réseaux antagonistes génératifs
Karras, Tero, Samuli Laine, et Timo Aila. 2018. « A Style-Based Generator Architecture for Generative Adversarial Networks ». arXiv:1812.04948 [cs, stat], décembre.

http://arxiv.org/abs/1812.04948

7 Ces images sont toutes réalisées à l'aide de réseaux antagonistes génératifs (GAN[5]), constitués en fait de deux réseaux de neurones profonds confrontés l'un à l'autre. Le premier (génératif) est entraîné à produire des images les plus réalistes possibles, qui sont alors soumises au second (discriminatif), dont l'unique tâche est de discerner ce qui lui semble vrai de ce qui lui semble faux[6]. Les auteurs l'expliquent de cette manière : « Le modèle génératif peut être considéré comme une équipe de faussaires qui tentent de produire de la fausse monnaie et de l'utiliser sans se faire repérer,

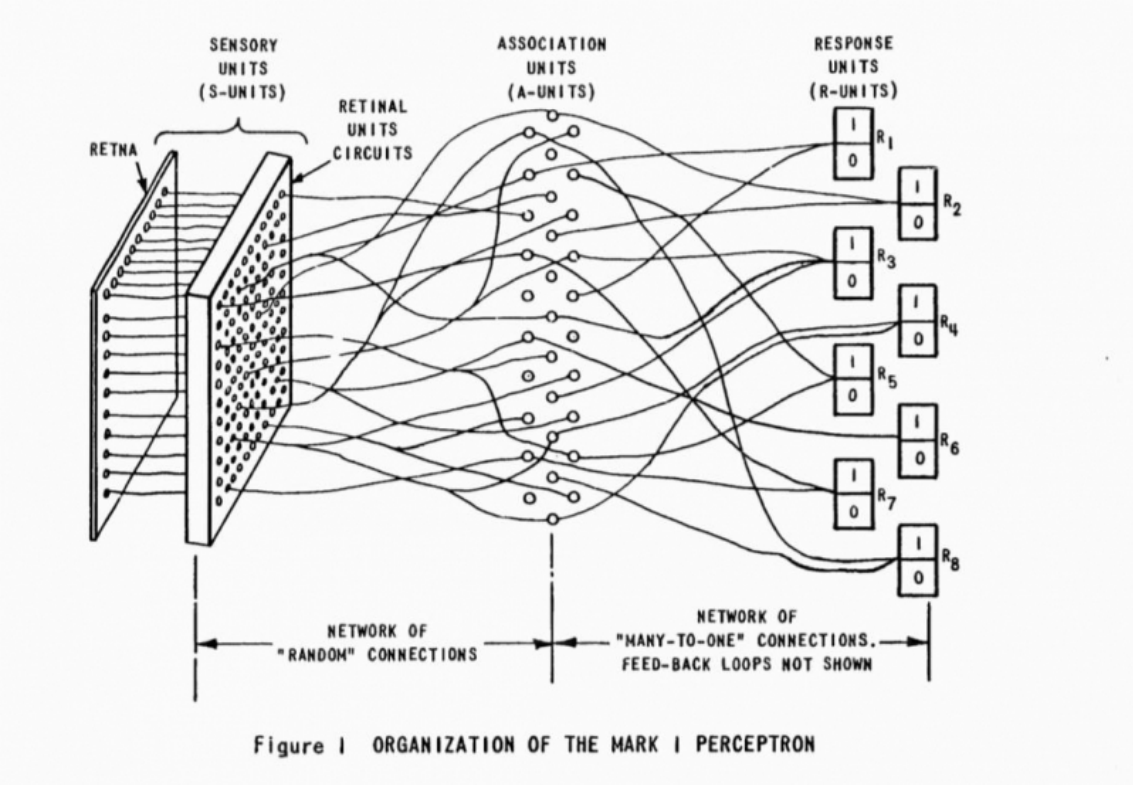
tandis que le modèle discriminatif est un peu comme la police qui cherche à détecter la contrefaçon. La concurrence dans ce jeu pousse les deux équipes à améliorer leurs méthodes jusqu'à ce que les contrefaçons soient indiscernables des originaux [7] ».

8 Puisqu'il est ici question d'image, cette compétition doit plutôt nous évoquer le jeu auquel se livrent le faussaire de tableaux [8] et l'amateur d'art au regard aiguisé, sensibilisé à ces « traces » dont l'usage avait si bien été décrit, au croisement de l'histoire de l'art, de la psychanalyse et de l'enquête criminelle, par Carlo Ginsburg [9]. Si c'est par le détail que notre regard entre dans ces portraits, c'est aussi par là que les réseaux de neurones artificiels ont appris à les réaliser. Pour s'en rendre compte, il n'est pas inutile de remonter à la naissance de l'intelligence artificielle. Celle-ci est couramment située en 1956, au moment de la conférence du Dartmouth College, université privée nord-américaine qui réunit alors à peu près tous les scientifiques qui feront, dans les vingt années suivantes, des avancées significatives en la matière. Dans cette association de termes inventée par John McCarthy et Marvin Minsky, l'intelligence est entendue comme une capacité à « manipuler les symboles », déjà préfigurée par Leibniz 250 ans plus tôt [10], et entérinée par les sciences cognitives alors en plein développement. Il s'agit de doter les machines d'un raisonnement autonome qui réside essentiellement dans une capacité à calculer des symboles selon des règles programmées.

9 Cette idée pose le cadre des ambitions de la branche historique de l'intelligence artificielle, dite symbolique [11] ou cognitive [12], associée aux noms d'Alan Newell et Herbert Simon, qui avaient développé en 1955 le premier programme destiné à modéliser le fonctionnement de l'esprit à l'aide du traitement symbolique [13]. Or les « intelligences artificielles » qui connaissent à partir des années 2000 de spectaculaires avancées dans le domaine de la vision sont le produit d'une conception radicalement différente de ce cadre inaugural. L'apprentissage profond qui caractérise ces modèles récents va plutôt chercher ses sources dans la cybernétique et le behaviorisme, qui évacuent de la machine le problème du raisonnement au profit de son comportement. Les réseaux profonds de neurones ne suivent pas des règles logiques préétablies pour en déduire des cas particuliers ; ils sont plutôt conçus pour découvrir d'eux-mêmes des représentations appropriées, en optimisant leur marge d'erreur au cours d'un

processus d'apprentissage. Les systèmes que nous croiserons plus loin sont tous conçus sur ce même modèle, dit « connexioniste », car il fait reposer les représentations qu'il construit sur la nature même des connexions entre les éléments dont il est composé. Étroitement associée à la reconnaissance visuelle, son histoire est marquée par les travaux fondateurs de Frank Rosenblatt au laboratoire d'aéronautique de l'Université Cornell à partir de 1957. Rosenblatt conçoit ce qu'il appelle le « perceptron[14] », un système qui met en réseau des neurones formels, dont le fonctionnement avait été modélisé dès les années 1940 par Warren McCulloch et Walter Pitts[15].

Diagramme du Mark 1 Perceptron



Mark I Perceptron Operators' Manual. Buffalo, NY: Cornell Aeronautical Laboratory, 1960.

Frank Rosenblatt et Charles Wightman, 1957

10 Un neurone de ce type est tout simplement un opérateur qui fait une somme pondérée de ses entrées, c'est-à-dire une somme des informations qu'il reçoit – chacune de ces informations étant minimisée ou maximisée selon l'importance relative (le poids) de l'entrée à laquelle elle se présente[16]. La nouveauté du Perceptron consiste à

introduire un principe de feedback dans ce modèle, afin de modifier progressivement les poids des entrées au cours d'une phase initiale d'apprentissage[17]. Le perceptron pouvait donc être entraîné à « reconnaître » des échantillons visuels en incorporant sa marge d'erreur pour améliorer ses réponses, et peut à ce titre être considéré comme la première machine apprenante. Il faut cependant nuancer la performance de ce système, qui, selon Rosenblatt, « peut faire la différence entre un chat et un chien, bien qu'il ne serait pas capable de dire si le chien est à la gauche ou à la droite du chat. Pour le moment, il n'a pas d'utilité pratique [...] mais un jour, il pourrait être utile d'en envoyer un dans l'espace pour voir à notre place (*to take in impressions for us*[18]). Rosenblatt ne se doutait pas que les principes de l'apprentissage profond en plein essor de nos jours seraient directement empruntés à ce réseau primitif. Après avoir fait l'objet de recherches intenses dans les années 1960, les réseaux de neurones artificiels seront cependant relativement délaissés dans la décennie suivante[19], au profit d'approches symboliques, qui, dans le domaine de la vision, vont plutôt chercher à « instruire la machine » pour lui permettre d'appréhender des univers clos et simplifiés. C'est du moins, au début des années 1970, la démarche de Marvin Minsky, qui développe avec le mathématicien Seymour Papert des « micromondes[20] » de blocs au MIT, dont les structures pouvaient être appréhendées par un ordinateur muni d'une caméra.

11 Le problème de la vision reçoit ici une réponse bien différente du comportement adaptatif du Perceptron. Il s'agit de modéliser le raisonnement au cœur de la machine, en s'appuyant sur les avancées des sciences cognitives. Il faut rappeler que Papert avait collaboré avec Jean Piaget entre 1958 et 1963 au sein du Centre international d'épistémologie génétique de Genève. Papert s'intéresse alors de près à l'apprentissage fondé sur l'expérience, qu'il aborde dans un texte de 1963 en décrivant les principes par lesquels il serait possible d'en reproduire les structures sur ordinateur[21]. L'expérience des micromondes a indéniablement bénéficié de l'apport des notions piagéennes de régulation, d'assimilation[22], mais surtout de schème, une notion qui peut trouver son équivalent dans les « cadres » (*frames*), théorisées par Minsky en 1975[23].

12 L'approche cognitiviste va ainsi tenter de donner ces cadres à la machine, là où le connexionisme, héritier du perceptron, va plutôt

faire en sorte qu'ils soient construits par la machine à partir des exemples d'apprentissage. On peut reprendre ici la distinction faite par Dominique Cardon et al. entre ces deux démarches opposées : la première, hypothético-déductive, offre une place centrale au programme, qui permet à la machine d'appréhender par déduction des cas particuliers ; la seconde, inductive, laisse le soin à la machine de trouver son programme elle-même, pour l'appliquer ensuite à des situations nouvelles [24]. Mais si les expériences de Minsky et Papert au MIT sont ambitieuses, elles déçoivent cependant les attentes et ne trouvèrent pas de prolongements marquants dans le domaine de la reconnaissance visuelle, qui va plutôt devenir le terrain de jeu des réseaux de neurones. Les approches connexionistes vont ainsi ressurgir au cours des années 1980 et 1990, défendues notamment dans un ouvrage marquant publié en 1986 en deux volumes par le groupe de recherche PDP [25]. Les ordinateurs ont alors gagné en puissance, et les réseaux peuvent additionner des couches de neurones pour s'organiser en structures bien plus complexes.

Diagramme d'un réseau de neurones artificiels avec deux couches intermédiaires

13 Le chercheur en intelligence artificielle Yann Lecun travaille à partir de 1988 sur la reconnaissance des images par ces techniques de manière indirecte : en mettant au point, au sein du laboratoire d'AT&T, une technique d'apprentissage et de reconnaissance de caractères [26]. Le véritable apport de cette technique réside dans un algorithme qui permet, au cours d'un apprentissage supervisé, de renvoyer au travers du réseau la marge d'erreur mesurée en sortie, pour modifier les poids synaptiques de chaque entrée de neurone et progressivement converger vers la réponse attendue. Si cet algorithme dit de « rétropropagation de gradient stochastique » est le fruit de plusieurs recherches dont les plus anciennes remontent aux années 1970, son efficacité sur les images est véritablement démontrée au milieu des années 1980 [27]. C'est aussi l'augmentation de la taille des bases de données d'images étiquetées qui a véritablement permis de faire ses preuves à cette approche, qui consomme un grand nombre d'exemples au cours de la phase d'apprentissage. Apprendre à un réseau de neurones à « reconnaître » un chat consiste à lui présenter des milliers d'images différentes de chats

(photographiques, dessinées, de différentes tailles, formes, races, dans différentes postures et selon différents cadrages [28]. Progressivement, un tel réseau modifie les filtres par lesquels il perçoit ce qui lui est présenté, pour se faire une « idée » de ce qu'est un chat, découvrant par lui-même ce que mille images de chats ont en commun [29].

14 Dans cette curieuse opération prédictive, c'est la force des connexions du réseau qui se transforme d'elle-même de manière statistique. À l'inverse des modèles symboliques des premières heures de l'intelligence artificielle, les réseaux de neurones fonctionnent de manière non séquentielle, non centralisée, non hiérarchisée [30]. Ils remontent des cas particuliers vers une représentation générale qui n'est pas donnée à l'avance ni stockée dans un programme à un endroit bien identifié du réseau. On peut aussi se représenter l'activité de ce genre de réseaux comme une recherche d'invariants dans un ensemble incroyablement diversifié. Sur un plan purement descriptif, apprendre à voir à une machine reviendrait ainsi à modéliser les « paramètres explicatifs » des objets du monde, et rejoindrait le vieux rêve positiviste d'une mise en données ou d'une mathématisation du monde [31]. Mais en quoi consistent ces paramètres explicatifs ? Qu'est-ce que ces réseaux parviennent à modéliser ? Quelles représentations se font-ils des objets qu'ils semblent arriver, au terme d'un apprentissage, à « connaître » ? Ces questions semblent bien s'imposer comme préalable dès lors que nous voulons confier à des machines le soin de voir à notre place.

Rêves profonds

15 En 2012, un algorithme réussit, lors du concours « Large Scale Visual Recognition Challenge », à faire descendre sa marge d'erreur à 16%, bien en dessous de tout ce qui avait été réalisé lors des années précédentes. Le réseau qui obtient cette performance a été conçu par Alex Krizhevsky, Ilya Sutskever, et Geoffrey E. Hinton de l'Université de Toronto [32]. Il a la particularité d'être profond et d'être construit de manière convolutive [33]. Le principe de la convolution consiste à analyser des images à l'aide de filtres plus petits, sensibles à un motif spécifique (arrêtes, contours orientés), et conçus pour activer ou inhiber des neurones artificiels de la couche suivante selon que ce motif est détecté ou non. Cette nouvelle couche est ensuite analysée

de la même manière par d'autres filtres, sensibles à d'autres motifs un peu moins élémentaires (formes primaires, motifs) et ainsi de suite, de couche en couche.

Diagramme de l'architecture d'un réseau à convolution

Cet exemple est celui de LeNet-5, conçu par l'équipe de Yann LeCun en 1998 pour la reconnaissance de caractères. LeCun, Yann, L.Bottou, Yoshua Bengio, et P. Haffner. 1998. « Gradient-Based Learning Applied to Document Recognition ». Proceedings of the IEEE 86 (11): 2278-2324

16 La profondeur du réseau permet de progresser par niveaux croissants de complexité, depuis les premières couches sensibles à d'infimes détails, jusqu'aux dernières, dédiées à des formes complexes et composées [34].

Représentation des motifs auxquels sont sensibles les différents filtres d'un réseau à convolution

Du plus simple dans les couches basses, au plus complexe dans les couches supérieures. Zeiler, Matthew D., et Rob Fergus. 2013. « Visualizing and Understanding Convolutional Networks ». arXiv:1311.2901 [cs], novembre.

http://arxiv.org/abs/1311.2901

Exemple de reconnaissance d'image par des légendes, réalisé à l'aide du réseau à convolutions DenseCap

Johnson, Justin, Andrej Karpathy et Li Fei-Fei. 2015. « DenseCap: Fully Convolutional Localization Networks for Dense Captioning ». arXiv:1511.07571 [cs], novembre.

http://arxiv.org/abs/1511.07571

17 Cette progression hiérarchique permet au réseau de se construire une idée du monde à partir de ses détails. Le monde visuel serait « compositionnel », comme le terme de Le Cun [35], c'est-à-dire qu'il peut être envisagé comme un assemblage d'éléments de complexité croissante. Les travaux des neurobiologistes David Hunter Hubel et Torsten Nils Wiesel sur le cortex visuel des chats dans les années 1960, souvent mobilisés dans la recherche en intelligence artificielle, avaient d'ailleurs mis en évidence le même principe de composition progressive des formes à partir de bribes élémentaires au sein des aires visuelles [36]. Appréhender le monde à partir du détail permettrait à ces réseaux de produire de la généralité, c'est-à-dire, de se construire, par abstraction progressive, des représentations qui

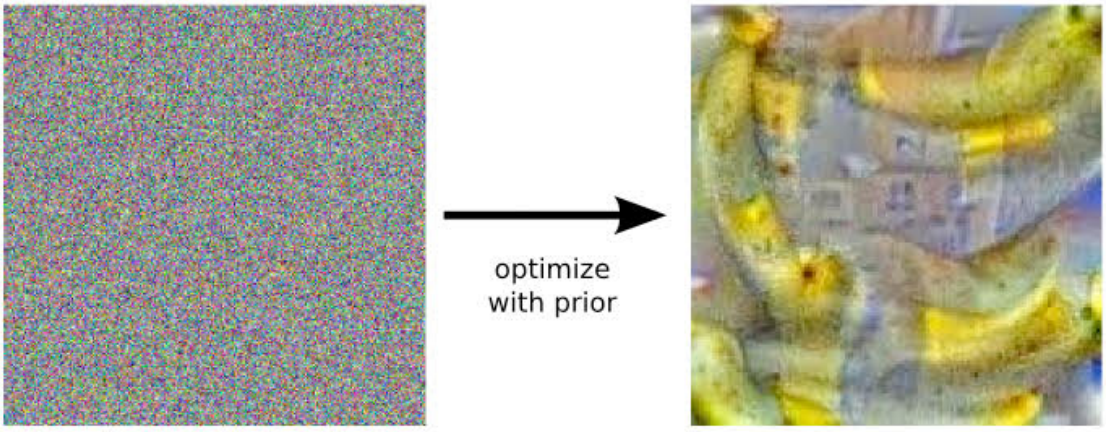
tiennent dans l'apprentissage un rôle semblable à celui que Piaget faisait tenir aux schèmes. Les schèmes sont des structures qui se construisent au fur et à mesure de la répétition d'une action, et qui permettent ensuite d'appréhender des situations nouvelles [37].

18 Si les recherches de Piaget n'ont jamais directement porté sur l'intelligence artificielle, elles ont trouvé de nombreux échos en la matière. Ainsi le schème s'apparente au « cadre » (*frame*) de Minsky que j'évoquais plus haut [38]. De l'aveu de l'auteur, l'idée de cadre n'est pas une nouveauté, mais trouve son équivalent dans les schémas du psychologue F.C. Bartlett, ou encore les paradigmes chers à Thomas Kuhn [39] Si le schème, le cadre, le paradigme, peuvent s'apparenter à des représentations mentales, nous pourrions même en faire remonter les sources à Kant. Mais dans le contexte de la vision, ces notions ont une histoire qui mérite par ailleurs d'être éclairée par les réflexions d'Ernst Gombrich. Dans ses travaux sur la représentation picturale, Gombrich insistait sur la fonction directrice des schémas structurant l'expérience. Pour concevoir une représentation, l'artiste ne part jamais de rien. Ce que j'ai appelé ici schéma ou cadre, est pour Gombrich un « formulaire de base », que l'on peut, d'une certaine manière considérer comme un ensemble de préconceptions au filtre desquels passe la perception [40].

19 Représenter, pour l'artiste, c'est compléter un « formulaire » existant. « Ce qui est connu et familier restera toujours le point de départ de la représentation de l'inhabituel [41] ». Cette idée est très proche de ce que Thomas Kuhn écrivait à quelques années d'écart : « [...] quelque chose qui ressemble à un paradigme est indispensable à la perception elle-même. Ce que voit un sujet dépend à la fois de ce qu'il regarde et de ce que son expérience antérieure, visuelle et conceptuelle, lui a appris à voir [42] ». Il n'est pas surprenant de constater que les recherches sur la vision artificielle aient eu à faire intervenir ces mêmes concepts pour modéliser la perception. Voir le monde et donner du sens aux choses, c'est faire entrer ce qu'appréhende le regard dans un ensemble de schémas familiers, dont la structure est elle-même le produit de l'expérience de l'observateur – humain ou non humain. Or nous avons vu que si de tels schémas étaient délibérément insérés dans les machines symboliques, ils ne trouvent aucun équivalent clair dans les réseaux de neurones, dont l'organisation homogène ne permet de produire que des représentations

non symboliques[43]. C'est d'ailleurs l'une des plus vives critiques qui leur est adressée : ces représentations ne sont après tout « que » le produit mathématique de régularités statistiques[44]. Ces observateurs non-humains que sont les réseaux peuvent cependant donner l'impression qu'ils possèdent une véritable expérience visuelle. C'est du moins ce qu'a tenté de faire apparaître le programme DeepDream[45], conçu en 2014 par Alexander Mordvintsev. DeepDream est construit de manière à parcourir le type de réseau que je viens de décrire dans le sens inverse. Alors qu'une image de bruit aléatoire est présentée en entrée, des instructions sont données à un neurone de sortie consacré au classement d'un type particulier d'image, afin qu'il obtienne un score plus élevé que les autres – et qu'il force le réseau, par conséquent, à produire ce type d'image.

Les premiers exemples d'images obtenues à partir d'un bruit blanc présenté en entrée d'un réseau, par Alexander Mordvintsev et son équipe

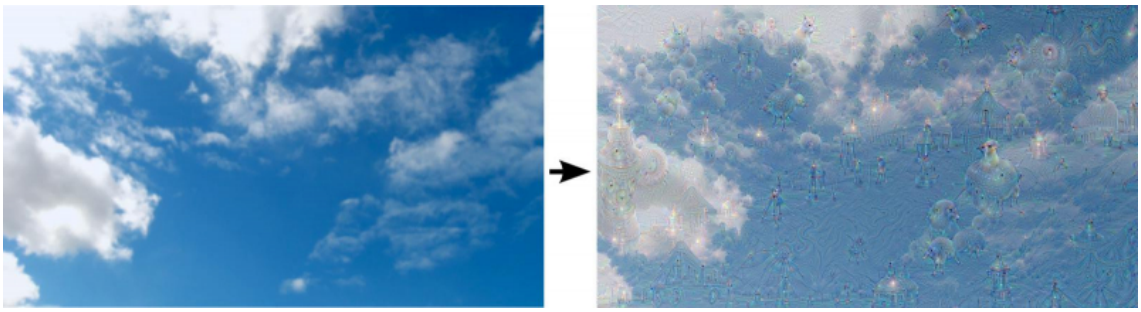


Mordvintsev, Alexander, Christopher Olah et Mike Tyka. 2015. « Inceptionism: Going Deeper into Neural Networks ». Google AI Blog (blog). 17 juillet 2015. http://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html.

20 Le réseau va alors modifier l'image de départ pour la faire tendre vers cette catégorie, en faisant apparaître les objets afférents. Les premiers exemples pris par les auteurs nous montrent des images de bananes, de fourmis, de vis, reconstruites ex-nihilo par le réseau. Mais les trois ingénieurs ont aussi essayé de partir d'une image existante en demandant à l'une des couches du réseau d'accentuer les détails auxquels elle est sensible. Largement diffusées ces dernières années, les images ainsi produites donnent à voir des paysages

improbables composés d'objets, de visages ou de têtes d'animaux, que le programme fait littéralement apparaître dans les visuels qu'on lui soumet : un plat de spaghettis, une voiture, un portrait de groupe, une reproduction de la Joconde, un ciel nuageux… se compliquent alors de formes ou d'animaux impossibles, saturés de couleurs et parsemés d'yeux, de pattes ou de museaux. Aussi surprenants soient-ils, ces détails doivent cependant rappeler quelque chose à quiconque a déjà observé des nuages en laissant aller son imagination. L'opération qui consiste à déceler au sein des nuées, des visages ou des animaux, ou essayer d'y découvrir un motif signifiant est communément appelée paréidolie, du grec *para*, « à côté de » et *eidôlon*, « image, apparence ». Si la paréidolie peut naître de tout type de choses, le nuage semble particulièrement approprié pour cet exercice de perception.

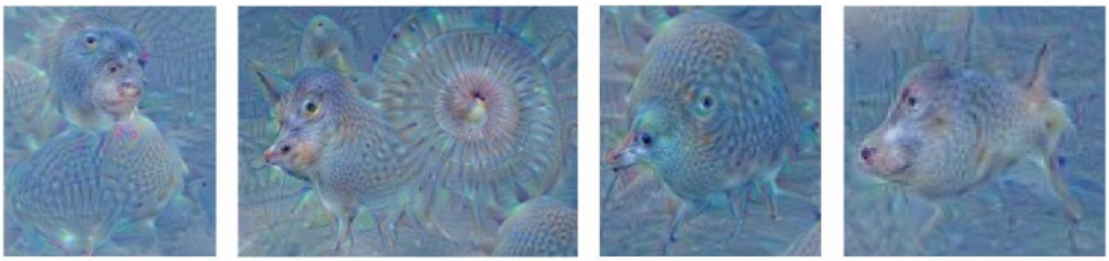
Images de ciel avant et après traitement par DeepDream



Alexander Mordvintsev et son équipe, 2015. Mordvintsev, Alexander, Christopher Olah et Mike Tyka. 2015. « Inceptionism: Going Deeper into Neural Networks ». Google AI Blog (blog). 17 juillet 2015.

http://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html.

Les détails apparaissant dans l'image précédente



Alexander Mordvintsev et son équipe, 2015. Mordvintsev, Alexander, Christopher Olah et Mike Tyka. 2015. « Inceptionism: Going Deeper into Neural Networks ». Google AI Blog (blog). 17 juillet 2015.

http://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html.

21 Car un nuage a rarement la forme d'un nuage. Il a la forme de ce que nous avons en tête lorsque nous l'observons. Mais qu'est-ce que les réseaux convolutifs ont-ils en tête ? En se laissant aller à cet exercice de « projection dirigée [46] », les réseaux de neurones empruntent une voie déjà décrite par Léonard de Vinci [47], mais aussi par Hermann Rorschach [48], pour nous donner à voir les schémas qui font sens à leurs yeux artificiels.

Image générée par DeepDream à partir d'un bruit blanc
Leonid Berov, 2016. Berov, Leonid. 2016. « Visual Hallucination For Computational Creation »

22 Ici encore, la modélisation connexioniste engendre sans surprise des phénomènes largement décrits par la psychologie, qu'il faut cependant rapporter au processus d'apprentissage dont ils sont le produit. À partir de quoi ces réseaux ont-ils appris ce qu'ils connaissent ?

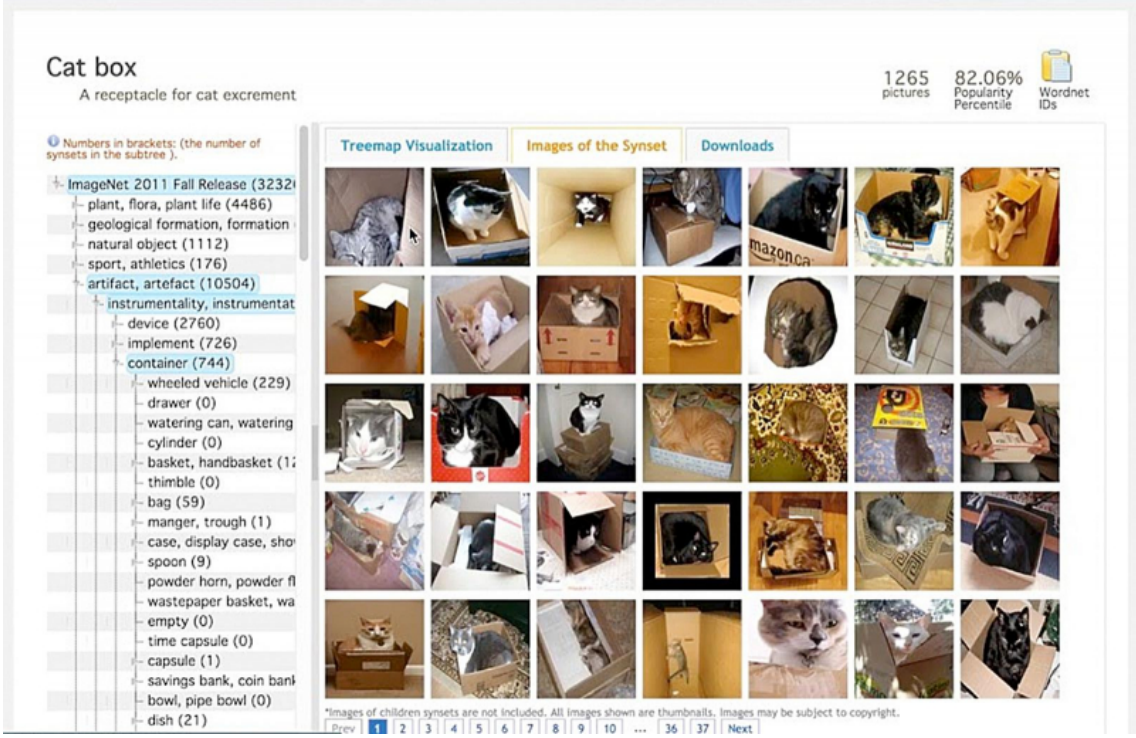
Pourquoi regarder les machines ?

23 Comme je l'écrivais plus haut, un certain nombre d'analogies entre l'homme et la machine ont inspiré les recherches sur l'apprentissage visuel par les intelligences artificielles. L'une d'entre elles a consisté à poser que pour connaître le monde, il faut en faire l'expérience. Apprendre à voir est étroitement lié au fait d'apprendre à nommer ce que l'on voit. Dans le cas des machines, c'est aussi par le langage que passe l'apprentissage du regard, qui consiste dès lors à pouvoir classifier, donc à nommer des objets. L'apprentissage supervisé vise à établir et à renforcer une capacité à généraliser, une capacité à pouvoir reconnaître l'image d'un type d'objet après avoir vu un grand nombre de représentations de ses occurrences. La première asymétrie entre l'homme et la machine apparaît d'emblée ici : alors que les images n'entrent que partiellement en jeu dans l'apprentissage humain, elles constituent la principale source employée à ce jour pour les machines. Loin d'être anecdotique, cette différence doit attirer notre attention sur le fait que les intelligences artificielles construisent une connaissance non pas du monde lui-même, mais du monde des images, qui a ses distorsions, ses limitations, et ses biais. Le plus large corpus d'images annotées destinées à l'entrainement à

ce jour est la base ImageNet, constituée à partir de 2009 par des chercheurs des départements de science informatique des Universités de Princeton et Stanford sous la direction de Fei Fei Li, professeur d'informatique à Stanford et co-directrice des instituts d'intelligence artificielle ainsi que de vision et d'apprentissage de Stanford [49].

24 Avec près de 15 millions d'images réparties en plus de 20 000 catégories [50], ImageNet, nous dit-on, « vise à fournir la couverture la plus complète et la plus diversifiée du monde de l'image [51] ». On peut cependant s'interroger sur la moyenne visuelle que produit le protocole suivi par l'équipe de Fei Fei Li, qui, sur le fond, exclut *a priori* les images gênantes, ambigües, offensantes, et sur la forme, se limite à des choses « dicibles », c'est-à-dire susceptibles de faire l'objet, non seulement d'une description textuelle, mais également d'un consensus sur cette description [52]. Car pour annoter manuellement cet ensemble monstrueux d'images, l'équipe s'est tournée vers le *crowdsourcing*, en déléguant ce travail à des milliers de personnes à travers le monde grâce à la plateforme Amazon Mechanical Turk [53]. Et pour organiser les images entre elles, Li a décalqué la structure modélisée dans la célèbre base lexicale Wordnet [54], en partant du principe que l'organisation sémantique du langage est « très proche » de la vision humaine, car, selon elle, « les mots (*labels*) du langage reflètent le monde visuel [55] »

L'entrée « Catbox » sur le site ImageNet



http://www.image-net.org/

https://doi.org/10.1109/CVPR.2009.5206848.

Deux branches de la structure d'ImageNet

Deng, Jia, Wei Dong, Richard Socher, Lia-Jia Li, Kai Li et Li Fei-Fei. 2009. « ImageNet: A large-scale hierarchical image database ». In 2009 IEEE Conference on Computer Vision and Pattern Recognition, 248-55. Miami, FL: IEEE.

25 Les intelligences artificielles ne sont peut-être qu'une nouvelle manière de poser de vieux problèmes, bien souvent posés dès l'antiquité [56]. Une telle remarque montre à quel point la vision reste tributaire, au cours de l'apprentissage, d'une structure linguistique. Et si au mieux, nous pourrions interpréter la proposition de Li à la lumière des tentatives positivistes pour faire correspondre la réalité sensible à l'espace logique du langage [57], nous savons à quel point ces tentatives de mise en ordre sont partielles et contingentes.

26 « L'ordre, c'est à la fois ce qui se donne dans les choses comme leur loi intérieure, le réseau secret selon lequel elles se regardent en quelque sorte les unes les autres et ce qui n'existe qu'à travers la grille d'un regard, d'une attention, d'un langage [58] ». Cette remarque de

Michel Foucault souligne les glissements qui s'opèrent entre le langage et le monde, dans un rapport qui, jamais fixé de manière définitive, doit forcément résister aux mises en chiffres et en fonctions mathématiques. Les cadres de l'expérience que construisent les intelligences artificielles pour voir le monde sont donc tributaires de la manière dont les bases d'images d'apprentissage comme ImageNet lient ensemble les mots et les choses. Et au-delà d'ImageNet, ce sont l'ensemble des réseaux sociaux qui servent à présent de bases d'apprentissage. Après tout, les images y sont annotées par leurs auteurs. Or en aspirant le contenu visuel d'internet, ces bases en reproduisent les lacunes et les excès, et bien sûr les obsessions. On ne s'étonnera pas de trouver autant d'images d'animaux dans l'histoire récente de la vision artificielle : le monde numérique en est peuplé. En témoignent les exemples qui ponctuent mon texte depuis le début. Il est à ce titre probable que les plus fines connaissances en matière de représentation de chatons soient de nos jours détenues par un algorithme. Et nous pourrons certainement dire rétrospectivement que les chats et les chiens, et plus généralement, les animaux, auront joué un rôle non négligeable dans la construction de la relation homme-machine [59]. Abreuvés d'innombrables images humaines d'animaux, les réseaux de neurones rejouent d'ailleurs, dans leurs réussites et dans leurs échecs, quelque chose du regard animal. Ainsi, les zèbres par exemple semblent poser aux machines les mêmes problèmes de reconnaissance qu'à leurs prédateurs, leurrés par des motifs naturellement prévus pour le camouflage et la dissimulation. En retour, ce sont des stratégies semblables qui sont imaginées par les humains eux-mêmes pour ne pas se faire reconnaître sur les réseaux sociaux qu'ils fréquentent [60]. Le regard machinique consiste ici à « prendre » et à éviter « d'être pris », selon ce jeu du chat et de la souris auquel s'adonnaient les réseaux antagonistes du début de mon texte. Un tel jeu nous renvoie à ce rapport paradoxal de la vérité et de l'illusion que WJT Mitchell a très bien décrit à partir de l'altérité du regard animal [61]. Bien-sûr, Mitchell historicise l'asymétrie entre les regards humain et animal pour interroger la fonction politique et sociale de l'illusion esthétique. Or si pour ce faire, l'auteur préconise, dans le prolongement de John Berger [62], d'apprendre à nouveau à regarder les animaux, nous pourrions ajouter qu'il faut aussi regarder les machines.

Formes de vie

27 Les modèles de réseaux présentés au fil de mon récit semblent nous entraîner dans un trajet circulaire qui peut être décrit en ces termes : la machine apprend à voir de manière inductive, en se construisant une représentation non symbolique d'un objet après avoir dégagé ce qu'un grand nombre d'images de cet objet ont en commun, pour être ensuite capable de reconnaître mais aussi de générer un nouvel objet dans cette série. La critique symboliste des réseaux connexionistes vise le sens de ces représentations, très éloignées des représentations mentales que le philosophe Jerry Fodor avait mises au cœur des processus cognitifs [63]. L'esprit était envisagé par Fodor selon le modèle computationnel du langage de la pensée, qui mobilise des symboles dans des opérations logico-mathématiques. Or l'apprentissage profond, difficilement compatible avec ce type de description, s'accorde plus facilement avec des modèles qui rejettent le mentalisme, ou du moins, qui s'en tiennent à une description pragmatique des phénomènes.

28 Contre le modèle computationnel, le comportement des réseaux profonds face aux images nous invite alors à renouer avec la seconde philosophie de Wittgenstein, qui rapportait l'ensemble des opérations cognitives à l'horizon de nos pratiques quotidiennes et du langage, en refusant d'en donner des explications faisant intervenir des représentations mentales. La description des réseaux profonds par ce paradigme non mentaliste nous permet dès lors de comprendre les limites que leur impose le mouvement circulaire que je viens d'esquisser ici : s'il n'y a rien en dehors de la pratique, s'il n'y a pas d'opération obscure qui préside à l'élaboration des cadres de l'expérience, alors nous comprenons qu'un réseau, aussi profond soit-il, ne peut découvrir que ce qu'il a appris. Les réseaux apprennent à voir selon ce que Wittgenstein appelle un « enseignement ostensif » qui consiste à fixer des relations d'association entre les mots et les choses, en suivant l'idée que « les mots du langage dénomment des objets – les phrases sont des combinaisons de telles dénominations [64] ». Wittgenstein ne rejette pas cette conception, mais il insiste sur le fait qu'elle ne recouvre qu'un des aspects du fonctionnement du langage. Car ainsi décrite, elle ne saurait partir « de rien », mais présuppose au contraire déjà le langage [65]. Comme le

démontre Wittgenstein dans l'ensemble des *Recherches*, il n'y a là qu'une forme « primitive » de la façon dont le langage fonctionne, limitée à un jeu de langage particulier. Or ce que requiert plutôt notre langage de tous les jours, c'est ce que Wittgenstein appelle une « forme de vie[66] », c'est-à-dire un arrière-plan communément partagé, qui est précisément constitué de tout ce que l'auteur avait délaissé dans sa première philosophie[67] : les valeurs, les émotions, la culture. C'est cet arrière-plan qui donne du sens à ce que nous disons, et non un mode d'emploi des mots que nous aurions entièrement assimilés avant de nous mettre à parler.

29 Entre le *Tractatus* et les *Recherches philosophiques*, publiées 30 ans plus tard, Wittgenstein est passé d'une conception du langage comme image du monde à une conception fondée dans la pratique, les habitudes et la culture. Ce que nous enseignent les *Recherches*, c'est que le sens que nous donnons aux images – par l'intermédiaire du langage et de la pensée – est profondément tributaire de nos formes de vie, c'est-à-dire d'expériences qui se construisent dans une longue durée en associant étroitement voir et savoir. Si l'apprentissage ne semble pas encore permettre aux réseaux de neurones de distinguer une vidéo d'attentat d'une séquence de jeu vidéo, c'est que les cadres qu'ils construisent ne saisissent le monde que dans ses apparences phénoménales, tel qu'il se donne en x et en y, en dehors du z de la pratique. Et c'est certainement pour dépasser cette limite que se développe l'informatique ubiquitaire[68], prophétisée par Mark Wieser au début des années 1990 : pour atteindre qualitativement le regard humain, la machine doit tenter de disparaître pour se fondre au cœur même de nos formes de vie. Si les « images invisibles[69] », ou les « photographies sans hommes[70] » suivent ce chemin, il faudra que nous conservions notre sens critique en maintenant ouvertes ces boites noires que sont les réseaux de neurones artificiels lorsqu'ils sont systématiquement insérés dans les différents aspects de nos existences.

NOTES

1 De nombreux autres exemples auraient pu figurer ici, y compris sous leur forme inverse : en septembre 2016, Facebook avait momentanément

supprimé de son réseau la célèbre photographie d'une jeune fille brûlée au napalm pendant la guerre du Vietnam en 1972, la nudité de la jeune fille ayant alors pris le pas sur la valeur historique et documentaire de cette image.

2 Thomas Coëffé, « Chiffres Facebook – 2018 », BDM Media, publié le 4 juillet 2018, [en ligne], http://www.blogdumoderateur.com/chiffres-facebook

3 Trevor Paglen, « Invisible Images [Your Pictures Are Looking at You] », The New Inquiry, publié le 8 décembre 2016, [en ligne], https://thenewinquiry.com/invisible-images-your-pictures-are-looking-at-you/.

4 Voir https://thispersondoesnotexist.com/.

5 En anglais, *Generative adversarial networks*. Voir Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, et Yoshua Bengio, « Generative Adversarial Networks », juin 2014, *arXiv:1406.2661 [cs, stat]*, [en ligne], http://arxiv.org/abs/1406.2661.

6 Pour être plus précis, il faut dire que les GAN optimisent les paramètres par descente de gradient pour réduire progressivement la marge d'erreur entre ce qu'ils produisent et ce que le réseau discriminatif considère comme une « vraie » image. Je détaille bien entendu le fonctionnement des réseaux de neurones dans les pages qui suivent.

7 *Generative adversarial networks*. Voir Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, et Yoshua Bengio, *op. cit.*, 2014, p. 1.

8 Il n'est à ce titre pas surprenant qu'un collectif ait immédiatement trouvé dans les réseaux antagonistes génératifs l'opportunité d'interroger une énième fois le statut de l'artiste, en produisant par ce moyen inédit, des images imprimées sur toile et vendues aux enchères. Voir https://obvious-art.com.

9 Carlo Ginzburg, « Traces. Racines d'un paradigme indiciaire » [1979] in *Mythes, emblèmes et traces. Morphologie et histoire*, par Carlo Ginzburg, Paris, Éditions Verdier, 2010, p. 218-294.

10 Sur les origines historiques de l'intelligence artificielle, voir Stuart J. Russell, et Peter Norvig, *Artificial Intelligence: A Modern Approach* [2010], Third edition, Global edition. Prentice Hall Series in Artificial Intelligence, Boston, Pearson, 2016. Voir également Daniel Crevier, *À la recherche de*

l'intelligence artificielle [1993], traduit par Nathalie Bucsek, Paris, Flammarion, 1997.

11 Dominique Cardon, Jean-Philippe Cointet, et Antoine Mazières, « La revanche des neurones: L'invention des machines inductives et la controverse de l'intelligence artificielle » in Réseaux 211 [5]: 173, 2018, https://doi.org/10.3917/res.211.0173.

12 L'approche symbolique est aussi appelée « cognitive » ou « classique », pour bien faire entendre qu'elle incarne le paradigme dominant des débuts de l'intelligence artificielle, dans ses liens avec les sciences cognitives. C'est aussi ce qui est couramment désigné par « intelligence artificielle forte ». Voir Daniel Andler, « Connexionnisme et cognition : À la recherche des bonnes questions » in *Revue de Synthèse*, 111 [1-2]: 95-127, 1990, https://doi.org/10.1007/BF03181031.

13 Ce premier programme est appelé Logic Theorist [Théoricien de la logique]. Voir Joanna Pomian, s. d. » Aux origines de l'Intelligence Artificielle : H. A. Simon en père fondateur » in *Quaderni*, n°1:9-25, 1987, https://doi.org/10.3406/quad.1987.2093.

14 Frank Rosenblatt, « The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain » in *Psychological Review*, 65 [6]: 386-408, 1958, https://doi.org/10.1037/h0042519.

15 Warren S. McCulloch, et Walter Pitts, « A Logical Calculus of the Ideas Immanent in Nervous Activity » in *The Bulletin of Mathematical Biophysics*, 5 [4]: 115-33, 1943, https://doi.org/10.1007/BF02478259.

16 Il faudrait aussi mentionner ici l'influence des travaux du neuropsychologue Donald Hebb, qui explique les phénomènes d'apprentissage en mettant en évidence quelques années plus tôt les effets des sollicitations répétées sur la consolidation du lien entre deux neurones. Sur un plan neurologique, l'apprentissage est souvent décrit comme la modification durable ou non de l'efficacité des synapses, qui se renforcent ou s'affaiblissent au gré de sollicitations spécifiques. Voir Donald Olding Hebb, *The Organization of Behavior: A Neuropsychological Theory*, New York, Wiley, 1949, n. p.

17 Il est tout d'abord implémenté sur un IBM 704, avant d'être réalisé sous une forme matérielle du nom de Mark I Perceptron. La couche d'entrée est constituée de 400 cellules photoélectriques ; la sortie est un classifieur linéaire composé d'unités qui s'activent ou non selon les sorties de la couche intermédiaire de neurones.

18 Mason Harding, D. Stewart et Brendan Gill, « Rival » in *The New Yorker*, 6 décembre 1958.

19 Souvent décrit comme « l'hiver de l'intelligence artificielle », le passage à vide des années 1970 et début 1980 s'explique autant par les modestes avancées de la recherche et l'assèchement des soutiens financiers et académiques, que par la publication en 1969 d'un ouvrage de Marvin Minsky et Seymour Papert très critique sur le modèle du Perceptron. Cet ouvrage va contribuer à freiner fortement les projets connexionistes. Voir Marvin Minsky, et Seymour A. Papert, *Perceptrons: An Introduction to Computational Geometry* [1969], Cambridge/Mass., The MIT Press, 1972.

20 Au-delà de cet exemple précis, le terme « micromonde » est employé dans le domaine de l'informatique éducative, en faisant référence au premier langage de programmation développé par Seymour Papert en 1967, « Logo »

21 Seymour Papert, « Étude comparée de l'intelligence chez l'enfant et le robot » in *La Filiation des structures*, par Léo Apostel, Jean-Blaise Grize, Seymour Papert et Jean Piaget, Paris, Presses Universitaires de France, 1963, n. p.

22 Jean-Jacques Ducret, « Jean Piaget et les sciences cognitives » in *Intellectica. Revue de l'Association pour la Recherche Cognitive*, 33 [2]: 209-29, 2001, p. 216. Voir Jean Piaget et Bärbel Inhelder, *La psychologie de l'enfant* [1966], Paris, PUF, 2015. Voir également Daniel Crevier, [1993] 1997, *op. cit.*, p.108-109.

23 Marvin Minsky, 1975. « A Framework for Representing Knowledge », in *The Psychology of Computer Vision*, par P. H. Winston, p. 211-279. New York, McGraw-Hill. p. 213.

24 Dominique Cardon, Jean-Philippe Cointet et Antoine Mazières, « La revanche des neurones : L'invention des machines inductives et la controverse de l'intelligence artificielle » in *Réseaux*, 211 [5]: 173, 2018, https://doi.org/10.3917/res.211.0173

25 David E. Rumelhart et James L. McClelland, *Parallel distributed processing: explorations in the microstructure of cognition. Computational models of cognition and perception*. Cambridge/Mass, MIT Press, 1986.

26 Yann LeCun, L. Bottou, Yoshua Bengio et P Haffner, « Gradient-Based Learning Applied to Document Recognition » in *Proceedings of the IEEE 86 [11]*: 2278-2324, n.m, s.l, 1998, n. p.

27 David E. Rumelhart, Geoffrey E. Hinton et Ronald J. Williams, « Learning representations by back-propagating errors » in *Nature 323 [6088]* : 533-36, n.m, n.l, 1986, n. p.

28 Une documentation abondante existe sur le fonctionnement détaillé du *deep learning*, que j'aborderai plus loin. On peut se rapporter à un texte très synthétique publié dans *Nature* par les trois spécialistes de l'apprentissage profond, qui se sont partagés en 2019 le prestigieux prix Turing : LeCun, Yann, Yoshua Bengio, et Geoffrey Hinton. 2015. « Deep learning » in *Nature*, 521 [mai]: 436. Pour une introduction plus longue, voir Ben Krose et Patrick van der Smagt, *An introduction to Neural Networks*, Amsterdam, The University of Amsterdam, 1996, n. p. Sur le fonctionnement des réseaux profonds, voir Jürgen Schmidhuber, « Deep Learning in Neural Networks: An Overview » in *Neural Networks 61 [Janvier]*, 2015, p. 85-117.

29 Voir la conférence de Yoshua Bengio sur les réseaux de neurones multi-couches lors de l'école d'été IVADO/MILA 2017. Yoshua Bengio, « Réseaux de neurones multi-couches », présenté à École d'été en apprentissage profond IVADO/MILA, 2017, [en ligne], https://www.youtube.com/watch?v=ImmQVrap1Uc. 6'07".

30 Voir à ce sujet Daniel Andler, *op. cit.*, 1990, p. 98.

31 Cette idée est évoquée par Yann LeCun dans son cours au collège de France. Yann LeCun, « Pourquoi l'apprentissage profond ? » in *Cours, Collège de France*, 12 février 2016, [en ligne] https://www.college-de-france.fr/site/yann-lecun/course-2016-02-12-14h30.htm. 58'.

32 Alex Krizhevsky, Ilya Sutskever, et Geoffrey E. Hinton, « ImageNet Classification with Deep Convolutional Neural Networks » in *Communications of the ACM* 60, [6], p. 84-90, 2017, n. p

33 La première expérimentation sur des réseaux à convolution remonte à 1980 avec le travail de Kunihiko Fukushima sur le « néocogitron », repris ensuite par Yann Le Cun pour la reconnaissance de chiffres au milieu des années 1990. Kunihiko Fukushima, « Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position » in *Biological Cybernetics 36 [4]*, 1980, p. 193-202.

34 Pour une explication détaillée des convolutions, voir Matthew D. Zeiler et Rob Fergus, « Visualizing and Understanding Convolutional Networks », 2013, [en ligne], http://arxiv.org/abs/1311.2901.

35 Yann LeCun, « Pourquoi l'apprentissage profond ? » 2016, *op. cit.*, 58'.

36 Les représentations visuelles complexes sont le produit de formes plus simples de reconnaissance au sein des aires visuelles du cortex. Voir Hubel, D. H. et T. N. Wiesel, « Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex » in *The Journal of Physiology 160 [1]*, 1962, p. 106-54, [en ligne], https://doi.org/10.1113/jphysiol.1962.sp006837 et David H. Hubel et Torsten N. Wiesel, *Brain and visual perception: the story of a 25-year collaboration* in *New York*, New York, Oxford University Press, 2005.

37 « Un schème est la structure ou l'organisation des actions telles qu'elles se transfèrent ou se généralisent lors de la répétition de cette action en des circonstances semblables ou analogues ». Jean Piaget et Bärbel Inhelder, *La psychologie de l'enfant* [1966] Paris, PUF. 2015, p. 11.

38 Marvin Minsky, « A Framework for Representing Knowledge » in *The Psychology of Computer Vision*, par P. H. Winston, 211-79, New York, McGraw-Hill. 1975, p. 213.

39 Thomas Kuhn, *La Structure des révolutions scientifiques* [1962], traduit par Laure Meyer, Paris, Flammarion, 2008, n. p.

40 Gombrich rappelle que la comparaison entre ces stéréotypes et les formulaires administratifs était courante dans le langage médiéval, le même terme « simile » étant appliqué « aussi bien aux formulaires procédurier qu'aux épures de l'art pictural ». Ernst Gombrich, *L'Art et l'illusion : Psychologie de la représentation picturale* [1960], traduit par Guy Durand, Paris, Gallimard. 1996, p. 63.

41 Ernst Gombrich, [1960] *op. cit.*, 1996, p. 72.

42 Thomas Kuhn, [1962], *op. cit.*, 2008, p. 160. Sept ans après la parution de son ouvrage, Kuhn reconnaît que le terme paradigme revêt au fil du texte pas moins de vingt-deux acceptions différentes. Voir à ce sujet la préface écrite en 1969 et figurant dans cette même édition.

43 On parle généralement de représentations « sous-symboliques ». Voir Paul Smolensky, « On the Proper Treatment of Connectionism » in *Behavioral and Brain Sciences 11 [01]*: 1, 1988, [en ligne], https://doi.org/10.1017/S0140525X00052432.

44 La critique du connexionisme est notamment portée par le philosophe américain Jerry Fodor, qui reproche aux représentations des réseaux de ne pas avoir de consistance syntaxique et sémantique. Voir Jerry A. Fodor et Zenon W. Pylyshyn, « Connectionism and Cognitive Architecture: A Critical

Analysis » in *Cognition 28 [1-2]*: 3-71, 1988, [en ligne] https://doi.org/10.1016/0010-0277(88)90031-5. Fodor a fait récemment évoluer certaines de ses positions. Voir Jerry Alan. Fodor, *L'esprit, ça ne marche pas comme ça: portées et limites de la psychologie computationnelle* [2000], traduit par Claudine Tiercelin, Paris, O. Jacob, 2003.

45 Alexander Mordvintsev, Christopher Olah et Mike Tyka, « DeepDream – a code example for visualizing Neural Networks », Google AI Blog [blog], 1 juillet 2015, [en ligne], https://ai.googleblog.com/2015/07/deepdream-code-example-for-visualizing.html.

46 Selon l'expression employée par Gombrich. Voir Ernst Gombrich, *op. cit.* [1960], 1996, p. 170.

47 Je renvoie ici au *Traité de la Peinture* de Léonard de Vinci, mais aussi aux nombreux exemples donnés par Gombrich au chapitre VI « L'image dans les nuées » de *L'Art et l'illusion*. Voir Leonard De Vinci, *Traité de la peinture* [1651], Paris, 2003, Calmann-Lévy et Ernst Gombricht, *op. cit.* [1960], 1996, p. 154-169.

48 En 2016, l'artiste turc Osman Koç a fait passer un test de Rorschach à DenseCap, un système d'attribution de légendes basé sur un réseau de neurones à convolution développé par l'Université de Stanford.

49 Respectivement *Stanford University's Human-Centered AI Institute* et *Stanford Vision and Learning Lab*.

50 Pour le détail de la constitution d'ImageNet, voir Jia Deng, Wei Dong, Richard Socher, et *al*, « ImageNet: A large-scale hierarchical image database » in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248-55, Miami, FL: IEEE, 2009 et Hao Su, Jia Deng et Li Fei-Fei, « Crowdsourcing Annotations for Visual Object Detection » in *AAAI Publications*, Toronto, s.n, 2012.

51 Jia Deng, Wei Dong, Richard Socher et *al*, *Ibid.*, 2009, p. 1.

52 Fei-Fei Li confie d'ailleurs dans une conférence que certains termes ne produisaient jamais de consensus dans le processus de labélisation : « Il y a certains synsets [groupes de synonymes] sur lesquels les gens ne pourront jamais s'entendre » Fei-Fei Li, Conférence à *Google Tech* le 24 mai 2011, [en ligne], https://www.youtube.com/watch?v=qdDHp29QVdw [19'40"].

53 Il faut rappeler ici que l'un des points de repères dans l'histoire de l'intelligence artificielle est un canular bien connu du XVIII^{e} siècle. L'écrivain hongrois Johann Wolfgang von Kempelen conçoit en 1769 un automate

capable de jouer aux échecs qu'il nomme « Turc mécanique », qui s'avère rapidement n'être qu'une marionnette dirigée par un individu caché sous la table. C'est ce nom que choisira Amazon pour sa plateforme de *crowdsourcing* créée en 2005 : *Amazon Mechanical Turk*, sur laquelle des ouvriers humains aux quatre coins du monde réalisent des tâches dématérialisées contre une rémunération. Et c'est à cette plateforme que fera appel l'équipe d'ImageNet pour attribuer des labels à ses images, vérifiées à la chaîne par des humains « turkers » pour 4 centimes les 300 images. Voir https://www.mturk.com/.

54 Wordnet est une « ontologie », c'est-à-dire une base de donnée conçue pour inventorier le contenu lexical et sémantique de la langue anglaise. Le projet est développé par le laboratoire des sciences cognitives de l'Université de Princeton depuis 1985. Les concepts y sont classés selon une hiérarchie qui reflète leur degré de généralité et leurs rapports de dépendance sémantique les uns par rapport aux autres, en formant des chaines de type : *house / home / housing / structure / artifact / whole / object / physical entity / entity*. Chaque concept de cette hiérarchie est considéré comme un « synset », un élément caractérisé par ses relations spécifiques de dépendance sémantique. Voir le site hébergé par l'Université de Princeton : https://wordnet.princeton.edu/

55 Fei-Fei Li, Conférence à *Google Tech* le 24 mai 2011, [en ligne], https://www.youtube.com/watch?v=qdDHp29QVdw [18'01"].

56 Si les mots peuvent être liés d'une manière ou d'une autre au monde visuel, ce lien n'est bien évidemment pas nécessaire, comme le faisait déjà remarquer Platon dans le *Cratyle.* Platon, [386-385 av J.-C.], *Cratyle*, traduit par Catherine Dalimier. Paris, Flammarion, 1998, n. p.

57 Je pense notamment à la première philosophie de Wittgenstein. Voir Ludwig Wittgenstein, *Tractatus logico-philosophicus* [1921], traduit par Gilles Gaston Granger, Paris, Gallimard, 1993, n. p.

58 Michel Foucault, *Les Mots et les choses. Archéologie des sciences humaines*, Paris, Gallimard, 1966, p. 11.

59 Comme l'a bien décrit John Berger, les animaux sont liés dès l'origine à l'imagination humaine, qu'ils peuplèrent sous forme de symboles, de métaphores, de toutes sortes de manières de négocier un rapport au monde. John Berger, *Pourquoi regarder les animaux ?* [1977], Paris, Éditions Héros-Limite, 2011, n. p.

60 Je pense ici au projet CV Dazzle de l'artiste Adam Harvey, qui consiste à imaginer des transformations du visage pour échapper à la détection.

61 « le Soi est celui qui voit non seulement la vérité dans une illusion, mais [vue] comme illusion ; l'Autre est celui qui est pris par l'illusion, qui manque de la voir [vraiment] comme une illusion et qui se fourvoie en la prenant pour la réalité qu'elle représente [vraiment]. » William John Thomas Mitchell, « L'illusion, voir le voir animal », traduit par Maxime Boidy, Inframince, nº12: 31-41, 2018, p. 34.

62 John Berger, *op. cit.* [1977], 2011.

63 Jerry A. Fodor, *The Language of Thought. The Language & Thought Series*, Cambridge/Mass, Harvard Univ. Press, 1975, n. p.

64 Ludwig Wittgenstein, *Recherches philosophiques* [1953], traduit par Françoise Dastur, Maurice Élie, Jean-Luc Gautero et al, Paris, Gallimard, 2004, §1.

65 « On pourrait dire que la définition ostensive explique l'emploi – la signification – d'un mot si le rôle que ce mot doit généralement jouer dans le langage est déjà clair. » *Ibid.*, §30. Voir à ce sujet Francis Danvers et Joseph Saint-Fleur, « Ludwig Wittgenstein : une pédagogie en acte » in *Recherches & éducations*, nº3 [janvier], 2012, [en ligne], http://journals.openedition.org/rechercheseducations/575., §7 : « ne faut-il pas déjà posséder un certain langage pour pouvoir apprendre le langage ? »

66 Ludwig Wittgenstein, *op. cit.* [1953], 2004, §23 : « L'expression "jeu de langage" doit faire ici ressortir que parler un langage fait partie d'une activité, ou d'une forme de vie ».

67 Ludwig Wittgenstein, *op. cit.* [1921], 1993.

68 Mark Weiser, « The Computer for the 21st Century » in *Scientific American 265 [3]*: 94-104, 1991, [en ligne], https://doi.org/10.1038/scientificamerican0991-94.

69 Trevor Paglen, *op. cit.*, 2016.

70 Joanna Zylinska, *Nonhuman photography*, Cambridge, Massachusetts, The MIT Press, 2017, n. p.

ABSTRACT

Français
Ces dernières années, les recherches sur l'intelligence artificielle ont fait des avancées spectaculaires dans le domaine de la vision. Des réseaux profonds de neurones semblent désormais capables de voir à notre place et de prendre des décisions sur le produit de leurs observations. Or les résistances que nous opposent depuis des siècles, à nous humains, les images et leur interprétation, ne semblent pas tomber pour autant. Apprendre à voir et à dessiner à une machine nous impose de redéfinir ce que le « voir », comme processus nécessairement imprégné par un savoir, peut bien vouloir dire. Irréductible à un champ du savoir spécifique, le problème de la vision des machines est fondamentalement un problème de culture visuelle. À l'image du cerveau, les réseaux de neurones sont souvent mobilisés comme des boîtes noires sur lesquelles nous projetons toutes sortes d'inquiétudes ou d'espérances et qui semblent résister à la moindre tentative de description. Pour tenter de comprendre ce que « voir » peut bien vouloir dire pour une machine, ce texte interroge la manière dont les réseaux profonds de neurones apprennent à lier ensemble le langage, le monde et la pensée, en examinant les détails à partir desquels la vision machinique semble s'établir.

INDEX

Mots-clés
apprentissage, connexionisme, convolution, ImageNet, image, intelligence artificielle, perception, réseau profond de neurones, schème, vision

AUTHOR

Vivien Philizot
Vivien Philizot est docteur en arts visuels et designer graphique. Sa thèse et ses activités de recherche se situent au croisement des champs du design graphique, des études visuelles, et de l'épistémologie. Il a enseigné à la Haute école des arts du Rhin de Strasbourg, à la Haute école d'art et de design de Genève et a travaillé pendant plusieurs années au sein du collectif Atelier Poste 4. Il est actuellement maître de conférences à l'université de Strasbourg, où il enseigne la culture et la communication visuelle, le design graphique et la didactique par l'image.
IDREF : https://www.idref.fr/197567266