Interview

Questions répondues par Cyril Poulet :

1. Comment définiriez-vous le Machine Learning ?

C'est un ensemble de méthodes et de techniques permettant de modéliser et/ou résoudre des problèmes complexes en utilisant des données plutôt que des algorithmes explicites. Le domaine et vaste (et pas limité aux réseaux de neurones, mais je vous invite à regarder Wikipédia pour un aperçu meilleur que ce que je pourrai vous faire).

2. Quel est selon vous le plus grand avantage à utiliser le Machine Learning ?

Lorsque l'on s'intéresse à des problèmes complexes, mettre aux points des algorithmes y répondant peut s'avérer très/trop compliqué, voire impossible :

- Parce que le problème ne s'y prête pas (ex un filtre de spam : comment définir un spam ?)
- Parce que trop de paramètres entrent en jeu (comment trouver les principaux facteurs dans un ensemble de données ?)
- Parce qu'il est possible de parvenir par un algorithme « classique » à traiter 60 ou 70% des cas, mais que les 30% de cas restants seront trop difficiles, ou simplement pas prévus lors de l'écriture de l'algorithme (assez classique en traitement de l'image/vision par ordinateur)
- Parce qu'un algorithme exact est connu, mais trop couteux en calcul

Le Machine Learning permet de modéliser le problème en définissant un type de modèle puis en l'ajustant aux données connues. Le modèle résultant ne sera qu'une approximation, mais s'il est choisi convenablement et que l'apprentissage est bon, cette approximation doit être suffisante pour le problème que l'on souhaite traiter.

3. Aujourd'hui les réseaux de neurones sont utilisés notamment pour faire de la reconnaissance d'image. Pensez-vous qu'une autre méthode présentant moins d'erreurs va voir le jour, ou bien voyez-vous plus un perfectionnement de cette technologie ?

Les réseaux de neurones profonds (Deep Learning) sont effectivement très performants pour les tâches de reconnaissance d'image (en tout cas pour une certaine catégorie d'entre eux, les réseaux convolutionnels). Ils sont d'ailleurs déjà déployés par Google pour l'indexation d'image, et par Facebook pour l'analyse automatique de photo.

Il y a peu de chance pour que l'on arrive à mettre au point des algorithmes plus classiques pour cette tâche (c'est d'ailleurs pourquoi les réseaux de neurones sont autant à la mode). En revanche, toute tâche d'apprentissage est fondamentalement limitée par les données à disposition lors de l'apprentissage : un réseau peut apprendre à reconnaître un cheval d'après des images de chevaux, un zèbre d'après des images de zèbres, mais pas un zèbre à partir d'images de chevaux et de rayures.

Il est donc très probable que les prochains systèmes soient des mélanges de différentes méthodes d'apprentissages (arbres de décision, réseaux profonds, etc) et d'apprentissages trans-media (c'est à dire mélangeant image et texte, par exemple). Il y a d'ailleurs des travaux de recherche dans ce sens qui commencent à être publiés.

4. A quelles limites aujourd'hui l'apprentissage automatique est-il confronté selon vous ?

Un problème d'apprentissage automatique demande plusieurs étapes qui sont toutes complexes, et qui doivent toutes être correctement menées :

- Définition du problème :
  - Que veut-on faire / apprendre ?
  - A-t-on une idée claire de ce que l'on cherche à réaliser ?
  - L'apprentissage est-il vraiment l'outil le plus adapté ?

- Comment évaluer la ou les solutions que l'on mettra au point ?

- Acquisition de données :
  - De quoi a-t-on besoin comme données ?
  - Si on en a peu, peut-on en acquérir de nouvelles, ou augmenter artificiellement celles que l'on a ?
  - Si on en a beaucoup, peut-on dégrossir en essayant d'enlever celles qui ne seront probablement pas pertinentes pour le problème en cours ?

- Choix du modèle :
  - Comment choisir un modèle / technique d'apprentissage adapté à mon problème ?

Bien souvent ces dernières années les entreprises disposent de données, ont entendu le mot magique « apprentissage » et sont persuadées que ces données vont donc pouvoir être exploitées pour résoudre des problèmes qu'elles ne savent pas formuler…

Les données sont aussi dans certains cas le nerf de la guerre : plus un problème est complexe, plus la quantité de données nécessaire est grande (par exemple l'analyse automatique d'image, ou la traduction automatique).

Enfin, plus un modèle est gros (en nombre de paramètres à apprendre), plus la puissance de calcul nécessaire à l'apprentissage est important. Le boom du calcul sur GPU a d'ailleurs fortement contribué à l'essor du Deep Learning.

Ces problématiques de quantité de données et de puissance de calcul sont un véritable frein aux nouveaux entrants, et expliquent la prépondérance actuelle de très gros acteurs tels que Google et Facebook, tous les 2 très en avance dans la recherche actuelle.

5. Pensez-vous que dans les années à venir des progrès majeurs vont avoir lieu dans ce domaine ?

L'intelligence artificielle au sens large et l'apprentissage en particulier sont en plein essor depuis que la puissance calculatoire a augmenté (en particulier par les GPUs), et une très grosse communauté de chercheurs contribue à son développement.

De nombreuses entreprises s'y intéressent aussi (en particulier les gros acteurs de la Silicon Valley), et elles ont souvent l'avantage de pouvoir acquérir rapidement énormément de données (surtout par les applis sur smartphone, que les gens sont souvent ravis d'utiliser gratuitement sans trop se demander quelles sont les informations de leur vie privée qui sont récoltées en échange).

Il est intéressant d'ailleurs de noter que Facebook, Google, etc. participent fortement à la communauté open-source, en particulier en Deep Learning, et vont jusqu'à mettre à disposition gratuitement des modèles pré-entrainés sur une quantité de données inaccessible au commun des mortels.

Donc je pense que de nombreux progrès vont arriver ces prochaines années, certains étant déjà bien avancés (smart home, smart car, à terme smart city, etc).

6. Sur quel(s) type(s) de projet(s) en lien avec le Machine Learning avez-vous travaillé ?

J'ai commencé lors de mon stage de fin d'école d'ingénieur, en entrainant un réseau profond pour la détection de visage dans les images et en le portant sur FPGA (l' "ancêtre du GPU" en termes de calcul parallèles) pour avoir une carte dédiée à la reconnaissance de visage à 100 images / sec (c'était en 2008, donc avant que l'arrivée des GPU sur le marché ne lance l'intérêt public sur le Deep Learning).

Je n'en ai plus fait pendant quelques années : ma thèse portait sur les systèmes multi-agents, ce qui est un sous domaine de l'IA mais ne comprend en général pas d'apprentissage, et mon premier emploi était dans une startup trop petite pour que nous puissions acquérir des données nécessaires à l'apprentissage.

J'en fait maintenant, principalement dans le traitement de l'image, mais je ne peux pas en dire plus (accord de confidentialité envers mon employeur).

7. Quelle est selon vous la meilleure application de ce type de technologie (un exemple qui vous a marqué) ?

Je ne saurais pas tellement dire quelle est la meilleure, mais de nombreuses applications s'avèreront à terme très pratiques :

- Voiture intelligente (aide à la conduite, puis conduite automatique)
- Maison intelligente (régulation de la consommation électrique, gestion de la lumière/son, contrôle des visiteurs, etc).
- Résumés automatique (image, vidéo, texte) -> on voit déjà Google et Facebook s'y essayer pour les images
- Traduction automatique (texte, image, et à termes vidéo live)
- Aides au diagnostic médical
- Robotique
- Etc.

8. Dans quoi travaillez-vous précisément aujourd'hui ?

Cf question 6, je ne peux malheureusement pas en parler ici.

9. Question additionnelle : outre les limites, quelles sont les dangers de cette technologie ?

L'un des gros problèmes actuels est la façon dont les entreprises récupèrent les informations, c'est à dire sans que les utilisateurs finaux ne comprennent bien souvent tous les tenants et aboutissants du volume de données collectées. En effet, ces données peuvent être utilisées à des fins de surveillance : bien que les multinationales se défendent en disant qu'elles anonymisent les données, de nombreux travaux ont montré que l'on peut néanmoins suivre quelqu'un malgré les mesures mises en place (en particulier, on peut identifier quelqu'un à partir d'une dizaine de point heure-position de cette personne (même anonyme) et d'un ensemble de données d'apprentissage anonymisées).

On peut imaginer sans peine ce que l'on pourrait faire avec les données d'une maison connectée…

Il y a aussi de nombreux sujets intéressants liés aux méthodes d'apprentissage (on s'est aperçu récemment qu'un réseau pouvait se mettre à reproduire des biais de race, de sexe ou d'âge qui étaient cachés dans la masse de donnée d'apprentissage). Ce genre de biais pourrait rapidement apparaître dans les résumés automatiques, par exemple.

Enfin, l'autonomie future des systèmes intelligents posent de vraies questions applicatives, telles que :

- En cas d'accident de voitures autonomes, qui est responsable ?
- Si une voiture autonome doit choisir entre écraser 10 personnes ou s'aplatir dans un mur en tuant ses passagers, comment choisir ?
- Les drones et avions de chasses actuels sont maintenant limités par leurs pilotes humains, qu'on garde pour qu'il puisse prendre la décision de tirer. Si un jour on délègue cette responsabilité à la machine, qui sera responsable en cas de bavure ?
- Etc.

L'IA est un domaine en plein développement, et conjugue à la fois une énorme dynamique, des possibilités assez gigantesques, et un secteur privé autant à la pointe de l'état de l'art que la recherche publique (alors qu'en général le privé ne s'empare d'une technologie qu'une fois qu'elle a atteint une certaine maturité), avec plus de ressources à la fois pécuniaires et en acquisition de données, et moins d'intérêt pour les conséquences de leurs pratiques.

Il va donc falloir être vigilant ces prochaines années envers les pratiques des différents acteurs, à la fois en termes de recherche et en termes d'usage.