ALLISTENE
l'alliance des sciences et technologies du numérique

INFRASTRUCTURE DE RECHERCHE POUR L'INTELLIGENCE ARTIFICIELLE

JANVIER 2018

TABLE DES MATIERES

PREAMBULE

Après FranceIA, le rapport de l'OPECST sur l'intelligence artificielle ainsi que d'autres initiatives récentes, le Premier Ministre a officiellement lancé le 8 septembre 2017 la stratégie nationale sur l'intelligence artificielle (IA) et a confié au député Cédric Villani la charge de poser les fondements de cette stratégie. Tout cela témoigne de la volonté de l'état de maîtriser pleinement les enjeux de cette technologie clé, stratégique tant en termes de production des savoirs, d'innovation que de sécurité et de souveraineté nationale.

Dans ce contexte, la mission de ce groupe est de s'intéresser à l'un des leviers pour hisser la France parmi les premiers acteurs mondiaux de l'intelligence artificielle : la construction d'une grande infrastructure de calcul et de gestion des données. Un tel outil permettrait de soutenir et d'accentuer les efforts de recherche, de transfert, de formation et d'innovation en IA. Pour cela, le groupe a eu comme objectif d'apporter des réponses, argumentées scientifiquement, aux questions suivantes :

- quels sont les besoins de l'IA en terme de moyens de calcul et de stockage ?
- comment ces moyens spécifiques identifiés peuvent être mis à disposition de la communauté ? Soit par une plateforme indépendante ou est-ce que GENCI peut réorienter une partie de ses investissements pour ces besoins de l'IA ?

Résumé des recommandations du groupe

Le groupe considère que la communauté IA nationale – en particulier celle de l'apprentissage machine, dont les besoins sont dimensionnants – ne dispose pas des ressources nécessaires pour conduire des recherches au meilleur niveau mondial, au contraire de ses partenaires et néanmoins concurrents anglais, allemands ou japonais pour ne citer qu'eux. Les problèmes à résoudre, les volumes de données, les besoins de calculs correspondants, sont largement au-dessus des capacités fournies par les divers équipements auxquels la communauté de l'enseignement supérieur et de la recherche a accès à l'heure actuelle.

Nous préconisons donc la mise en place d'un Grand Equipement National pour l'Intelligence Artificielle (GENIAL), à destination de la communauté de recherche et d'enseignement supérieur, mais aussi utilisable par les acteurs du monde économique – petites et grandes entreprises.

Cet équipement, mono-localisé sur l'un des sites nationaux de calculs intensif, d'une puissance de 5 pétaflop/s à base de GPU et disposant de plus de 10 pétaoctets de stockage, devrait être réservable pour une part longtemps à l'avance par un système d'appels à propositions et pour une autre part à la volée avec une supervision visant à réguler les utilisations. L'accès serait gratuit pour les équipes de recherche, payant à coûts réels pour les entreprises ou pour les projets collaboratifs de type ANR, Europe ou PIA et payant à coût préférentiel pour les start-ups. Une politique de la donnée serait mise en place afin de constituer progressivement des bases de données de référence. Le coût global serait d'environ 10M€ sur cinq ans dont 5M€ pour l'acquisition initiale et 5M€ pour l'exploitation y compris une équipe technique d'une dizaine de personnes.

COMPOSITION DU GROUPE DE TRAVAIL

- Jamal Atif (Université Paris-Dauphine, CNRS-INS2I, animateur)
- Laurent Besacier (Université Grenoble Alpes)
- Olivier Boissier (IMT, Mines Saint-Etienne)
- Bertrand Braunschweig (Inria, Saclay, animateur)
- Stéphane Canu (INSA Rouen-Normandie)
- Julien Chiaroni (CEA-LIST, Saclay)
- Michel Dayde (CNRS, IRIT, Université de Toulouse, INP Toulouse)
- Elisa Fromont (Université de Rennes 1)
- Jean-Noël Patillon (CEA-LIST, Saclay)
- Christan Roux (IMT DRI, Paris)
- Laurent Simon (Bordeaux-INP, Talence)
- Anne-Sophie Taillander (IMT DRI, Paris)

INTRODUCTION

Depuis une dizaine d'années, la recherche en Intelligence Artificielle (IA) connait un véritable renouveau, avec d'indéniables succès technologiques (reconnaissance d'images, assistants personnels, jeu de go, médecine, véhicules autonomes, traduction...). Cela suscite un engouement et une médiatisation sans précédent qui en font l'une des technologies clé pour nombre d'acteurs, y compris les grandes puissances économiques et militaires.

Au cœur de ces succès se trouve une classe de méthodes d'apprentissage automatique connue sous le nom de réseaux de neurones profonds (*deep learning*). Outre la disponibilité de grandes masses de données et des moyens de leurs traitements, des avancées sur les modèles ainsi que le développement d'outils logiciels, ces méthodes ont pu profiter de la maturation de technologies de calcul dédiées, fondées principalement sur l'exploitation des capacités des accélérateurs graphiques (GPU[1]).

En effet, si l'on prend un exemple largement médiatisé ces derniers mois : l'algorithme Alpha Go, un programme joueur de Go développé par Google DeepMind[2], capable de battre les meilleurs joueurs humains, dans l'une de ses versions récentes, AlphaGo Zero, ne nécessitant que peu de supervision humaine, l'optimisation des paramètres du réseau profond, une fois la topologie du réseau est arrêtée, a mobilisé un cluster de 64 GPU et 19 CPU pendant 3 jours. A ces temps, il convient de rajouter la phase la plus coûteuse : la recherche de la topologie/architecture optimale pour le problème étudié. Celle-ci se fait en général de façon empirique, par essai/erreur : un réseau typique de 10 couches est mis au point en testant quelques milliards de configurations. Ainsi, les moyens de calcul constituent le chainon essentiel pour la production des modèles. A cet égard la recherche dans ce domaine s'apparente davantage aux sciences expérimentales, dont les instruments sont les moyens de calcul et les données.

Force est de constater qu'aucune équipe en France n'a les moyens matériels de reproduire de tels résultats, ou d'autres comme détaillé dans la section 1. Les chercheurs les plus actifs dans le domaine de l'IA disposent de quelques cartes GPU pour élaborer leurs modèles, recourent parfois à des méso-centres régionaux, à Grid'5000 ou aux services commerciaux (AWS par exemple).

[1] Graphics Process Unit est un circuit intégré aux cartes graphiques initialement conçu pour les tâches de visualisation. Il s'agit d'un processeur massivement parallèle particulièrement efficace pour les tâches de calcul matriciel. Les GPU présente par ailleurs l'avantage d'être plus économes en énergie en comparaison des processeurs classiques (CPU).

[2] Derrière l'aspect ludique et médiatique du jeu, se cachent des questions aux enjeux profonds de planification dans l'incertain et d'optimisation, sujets clé en robotique, en diagnostic médical, etc.

Les grands acteurs privés du numérique (i.e. GAFAM[3] et ensuite BATX[4]), se sont dotés très tôt de grands moyens de calcul et en étaient jusqu'à peu les propriétaires presque exclusifs. Cela participe à une stratégie de développement, et d'attractivité notamment à destination des jeunes chercheurs qui ne se trouvent pas freinés par des conditions matérielles pour le développement de leurs travaux. Cela explique aussi le nombre de résultats scientifiques et technologies émanant de ces mêmes groupes.

Les deux dernières années ont vu les grandes nations (EUA, Canada, Japon, Angleterre, Allemagne) s'emparer de cette question en se dotant de grands instruments de calcul dédiés à l'IA (Jade, RIKEN, etc ; plus d'éléments sont donnés dans la section 2).

Pour la communauté des chercheurs en IA en France et de ses nombreux domaines d'application (SHS, biologie, santé, physique…), ne pas disposer d'un accès à de telles infrastructures constitue un réel handicap, à la fois en termes de production des savoirs, d'attractivité de jeunes chercheurs et de positionnement international.

Une grande infrastructure de calcul et de données spécialisée en IA constitue un élément crucial pour la souveraineté nationale dans ce domaine critique. Elle permet à la recherche publique et privée de s'appuyer sur un équipement indispensable face à la compétition internationale et constitue un levier majeur pour le développement et la transformation des politiques publiques, tout en servant la compétitivité des entreprises et l'économie du pays. Chaînon manquant dans le transfert et la valorisation des résultats de la recherche, un tel équipement constitue une condition indispensable au positionnement de la France et de l'Europe comme leader de l'application des technologies de l'IA.

Par ailleurs, si les résultats les plus marquants aujourd'hui viennent du domaine de l'apprentissage automatique, l'on peut conjecturer que les avancées futures viendront en partie d'une hybridation de sous-domaines de l'IA : représentation des connaissances, raisonnement à partir de grandes masses données, apprentissage automatique, etc. L'exemple le plus illustratif étant les assistants conversationnels qui requièrent une intégration subtile entre ces sous-domaines. Les techniques SAT (satisfaisabilité de formules logiques), ou de raisonnement sur de grands graphes de connaissances (Knowledge graphs) ou de systèmes d'IA tel que Watson d'IBM, requièrent aujourd'hui moins des accélérateurs graphiques, que de grandes capacités mémoires et beaucoup de temps CPU.

De plus, si l'IA est encore largement dominée par des mécanismes et modèles pensés et développés à l'échelle individuelle : un système, un logiciel, un robot, il faudrait préparer les conditions de la maturation des mécanismes d'intelligence collective et sociale où les agents apprennent à coopérer, à communiquer, et à prendre des décisions collectives. Cela passe aussi par l'exploitation d'infrastructures de calcul. Les directions de recherche les plus prometteuses, combinant théorie de choix social, de jeux, et d'apprentissage par renforcement, nécessiteraient autant de ressources de calcul par agent que celles déployées pour AlphaGo Zero.

[3] Google, Amazon, Facebook, Apple, Microsoft.

[4] Baidu, Alibaba, Tencent, Xiaomi

L'IA suscite beaucoup d'enthousiasme à la fois en termes de champs d'application et de recherche fondamentale, mais soulève aussi chez nos concitoyens des inquiétudes à cause notamment de l'opacité des modèles et des décisions. Il est donc nécessaire, pour l'acceptabilité de cette technologie, de disposer d'une plateforme d'expérimentation, de test, de validation et de parangonnage des modèles et algorithmes, permettant de promouvoir la reproductibilité des approches, de mesurer les différents biais induits par les données et de contribuer ainsi aux grands enjeux transverses de l'IA (explicabilité, certification, performances, confiance…).

L'infrastructure doit donc être pensée pour satisfaire tous ces besoins. Elle doit en plus être suffisamment dimensionnée pour ne pas freiner les développements de nouvelles architectures neuronales : ceci demande à la fois de nombreuses machines, mais également de disposer d'une gamme suffisamment large de matériels (puissance, mémoire disque, taille de la mémoire GPU, capacité de traitement multi-GPU, CPU) pour servir tous les besoins. Il faut aussi associer à ce matériel un environnement offrant à la fois des capacités de développement et de production pour le traitement de jobs de grande taille, des garanties de service et l'appui d'une équipe compétente assurant la gestion efficace de ces ressources et un support de qualité aux utilisateurs. Une telle infrastructure devrait aussi venir en appui à de grands défis nationaux liés à la santé, au transport, à l'environnement, au secteur défense-sécurité, etc. La section 1 détaille quelques-uns de ces défis afin d'aider à mieux dimensionner une telle infrastructure.

STRUCTURE DE CE DOCUMENT

La suite de ce rapport est organisée ainsi : la première section donne une rapide présentation de la recherche en intelligence artificielle et de sa spécificité expérimentale, puis nous donnons un certain nombre d'illustrations des besoins de calcul et de stockage au travers d'exemples marquants et de défis sociétaux. La deuxième section contient un résumé des infrastructures à disposition de la communauté nationale et de la concurrence internationale, le détail figurant en annexe. La troisième section rappelle les recommandations émises par #FranceIA sur le sujet, et montre que l'Europe ne s'est pas saisie de la question. La quatrième section est consacrée aux recommandations du groupe de travail : dimension de l'équipement, budget, support, mode d'utilisation, modèle économique et quelques autres aspects. Les deux annexes de la dernière section contiennent les détails des équipements comme indiqué ci-dessus.

1. LA RECHERCHE EN INTELLIGENCE ARTIFICIELLE

L'intelligence artificielle (IA) est un domaine de recherche protéiforme. Sous cette appellation, on retrouve différents sous-domaines, constituant en soi des communautés de recherche indépendantes : apprentissage automatique, représentation des connaissances, raisonnement, décision collective, traitement automatique des langues, vision, robotique, etc. Pour une définition plus précise de ces sous-domaines, nous orientons le lecteur vers le rapport FranceIA, et en particulier le volet traitant de la « recherche amont [5] ».

Les avancées récentes viennent de ce qui est appelé l'IA faible ou étroite, en opposition à l'IA forte ou générale. L'IA faible s'attache à traiter des problématiques bien définies, de spectre étroit, souvent de reconnaissance de formes (reconnaissance d'activités dans les vidéo, traduction automatique, etc.). L'IA forte s'intéresse quant à elle aux mécanismes généraux de l'intelligence, et est à cet égard en parenté avec les sciences cognitives, la philosophie, les neurosciences.

L'IA faible repose largement sur l'apprentissage automatique, domaine à la croisée de l'informatique et des mathématiques appliquées (en particulier les statistiques et l'optimisation), qui doit ses succès pratiques en grande partie à l'explosion de masses de données et à la disponibilité de moyens de calcul pour les traiter. En particulier deux paradigmes de l'apprentissage automatique contribuent aux ruptures récentes : l'apprentissage profond et l'apprentissage par renforcement et leur combinaison.

L'apprentissage automatique est un champ de recherche qui consiste à développer des algorithmes capables (i) d'extraction automatique de "connaissances" à partir de masses de données à des fins de description ou de prédiction, et (ii) d'auto-amélioration à partir d'expérience.

L'apprentissage profond ou *deep learning*, sous-domaine de l'apprentissage automatique, désigne un ensemble d'algorithmes se fondant sur les réseaux de neurones artificiels : un modèle de calcul permettant sous certaines hypothèses d'approximer n'importe quelle fonction par la composition de fonctions élémentaires. Ces fonctions élémentaires dites neurones artificiels, puisqu'elles en étaient initialement inspirées, forment les nœuds d'un graphe ou réseau à plusieurs niveaux ou couches.

L'apprentissage par renforcement consiste, pour un agent évoluant dans un environnement incertain, d'apprendre une stratégie/politique permettant d'associer à chaque état complétement ou partiellement observé une action afin de maximiser une espérance du gain cumulé à horizon temporel fini ou infini.

La spécificité de la recherche dans ces domaines réside dans son caractère expérimental et empirique. Il s'agit en effet d'un domaine où les garanties théoriques sur l'apprentissage dans des contextes généraux manquent cruellement. La conception des systèmes fondés sur l'apprentissage profond s'appuie en réalité sur une forte expertise et un travail d'ingénierie conséquent [6]. Pour concevoir un réseau profond, en sus de l'optimisation des poids des

[5] https://www.economie.gouv.fr/files/files/PDF/2017/Conclusions_Groupes_Travail_France_IA.pdf, pp. 31-50

[6] p.ex. AlexNet à l'origine du retour en grâce des réseaux de neurones et des résultats impressionnants en vision par ordinateur.
Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

connexions faites souvent par rétro-propagation du gradient, la topologie du réseau est à adapter pour chaque application (nombre de couches, nombre de neurones, nature des connexions, choix de la fonction d'activation). On parle alors de l'optimisation des hyper-paramètres. Cette phase est la plus coûteuse. Les pratiques les plus répandues procèdent par une recherche par quadrillage (*grid search*) dans l'espace de ces hyper-paramètres. D'autres techniques s'appuient sur l'optimisation stochastique mais sont encore à un stade de recherche exploratoire. Pour une approche, comme pour l'autre, des moyens de calcul très conséquents doivent être mobilisés, puisque pour chaque configuration des hyper-paramètres, une procédure d'optimisation des poids des neurones est exécutée.

QUELQUES EXEMPLES DIMENSIONNANTS

Pour mieux définir les besoins en capacité de calcul, en architecture, et en stockage de l'infrastructure, nous détaillons ci-après quelques défis sociétaux et exemples de réalisations scientifiques parmi les plus marquantes de ces dernières années, qui sont actuellement inaccessibles à la communauté de recherche.

EXEMPLES DE REALISATIONS

DEEP SPEECH : LE SYSTEME DE TRANSCRIPTION AUTOMATIQUE DE LA PAROLE DE BAIDU

Afin d'évaluer les ressources nécessaires pour construire un système de transcription automatique de la parole tel que ceux développés par les géants industriels du domaine, nous avons analysé en détail l'article qui décrit le système *Deep Speech 2*, récemment déployé par Baidu. Cet article (publié à la conférence ICML 2016) est en ligne sur: https://arxiv.org/abs/1512.02595. Le système permet, d'après les résultats expérimentaux obtenus sur des jeux de données standard, des performances de transcription proches des performances humaines (*"…/… in several cases, our system is competitive with the transcription of human workers when benchmarked on standard datasets."*). En premier lieu, il est intéressant de noter que cet article de conférence rassemble **34 co-auteurs** (tous ne sont pas spécialistes de traitement de la parole et d'apprentissage profond car l'entraînement le déploiement opérationnel d'un tel système nécessite aussi des compétences en HPC, par exemple).

Deux systèmes de transcription automatique sont construits pour l'Anglais et le Mandarin à partir de 12000h et 9400h de parole transcrite, respectivement. Si on suppose une fréquence d'échantillonnage des signaux de 16khz, ceci représente des **données brutes de 2.5To** (hors "data augmentation") auxquelles il faut ajouter des grandes quantités de textes (transcriptions des signaux + données pour l'apprentissage des modèles de langue) qui représentent quelques dizaines de Go supplémentaires.

L'architecture du système *Deep Speech 2* présente jusqu'à 12 couches (réseaux convolutionnels dans les couches "basses" et réseaux récurrents dans les couches "hautes"). Le **nombre total de paramètres du réseau est entre 18 millions** pour le modèle le plus petit (taux d'erreur de mots sur l'anglais 10.59%) **et 100 millions** pour le modèle le plus grand évalué (taux d'erreur de mots sur l'anglais 7.73%).

L'**apprentissage** du modèle nécessite des dizaines d'**exaflops** (10^{18}) qui représenteraient **3 à 6 semaines pour s'exécuter sur un seul GPU**. Les auteurs décrivent un certain nombre d'optimisations qui permettent de revenir à un apprentissage qui nécessite environ 50 teraflop/s (10^{12}) lorsqu'il se déroule sur 16 GPU de type Titan X (3 à 5 jours d'entraînement). La figure ci-dessous indique le temps d'apprentissage pour 1 époque en fonction du nombre de processeurs GPU disponibles. On voit qu'une époque d'apprentissage représente 2^{18}s (3 jours) sur un seul GPU.

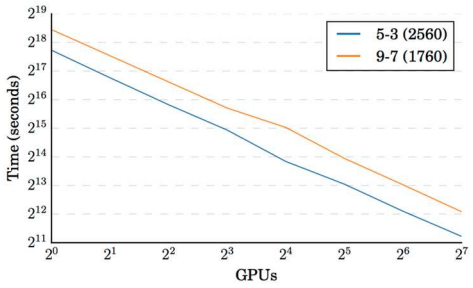

Figure 4: Scaling comparison of two networks—a 5 layer model with 3 recurrent layers containing 2560 hidden units in each layer and a 9 layer model with 7 recurrent layers containing 1760 hidden units in each layer. The times shown are to train 1 epoch. The 5 layer model trains faster because it uses larger matrices and is more computationally efficient.

Figure 1 *(image issue de : https://arxiv.org/pdf/1512.02595.pdf)*

Il est cependant important de noter que **les chiffres donnés ci-dessus concernent un seul apprentissage tandis que l'exploration des diverses architectures et hyper-paramètres nécessite des allers-retours permanents entre apprentissage et évaluation**. Un modèle comme celui présenté par Baidu dans cet article a sans doute nécessité de tester plus de 500 modèles.

ALPHAGO ZERO

AlphaGo Zero[7] est un programme joueur de Go capable d'apprendre une stratégie optimale de jeu sans avoir recours à un corpus de parties jouées par les humains, développé récemment par Google DeepMind. Il s'agit d'une évolution du programme de jeu AlphaGo qui reposait d'une part sur un apprentissage supervisé, par réseaux de neurones profonds, à partir de parties jouées par des humains, et d'autre part sur un apprentissage par renforcement où l'algorithme jouait contre lui-même. AlphaGo Zero repose entièrement sur la seconde partie, dite de self-play, qui se fonde exclusivement sur un apprentissage par renforcement et une recherche arborescente Monte Carlo. Notons que derrière l'aspect ludique du jeu se cache des enjeux scientifiques et technologiques fondamentaux de planification dans l'incertain et d'optimisation, sujets clé dans nombre de champs d'application : robotique, aide au diagnostic médical, finance, etc.
AlphaGo Zero, et son désormais ancêtre AlphaGo, sont l'exemple type de ruptures scientifiques qui n'auraient pas vues le jour sans la disponibilité de moyens de calcul conséquent mis à disposition du géant Google. Ceux-ci sont comme suit :

[7] *Mastering the game of Go without human knowledge,* David Silver et al. *Nature* **550**, 354–359 (19 October 2017). doi:10.1038/nature24270

- **Phase de jeu :** 4 *tensor processing unit* (TPU[8])
- **Phase d'entrainement :** cluster de 64 GPU et 19 CPU pendant 3 jours pour un réseau profond de 20 blocks résiduels. Pendant cet entraînement le programme a joué 4.9 millions de parties[9] contre lui-même, ce qui lui a permis de battre la version supervisée AlphaGo Lee. Cette version du programme est appelée AlphaGo Zero Master. Une autre version du programme a été testée, avec un réseau de 40 blocks résiduels. Le temps d'entrainement nécessaire pour battre AlphaGo Zero Master (89-11) est de 40 jours (voir Figure ci-dessous).
- *A ces temps, il convient d'ajouter des calculs supplémentaires nécessaires pour régler initialement les hyper-paramètres du réseau de neurones.*

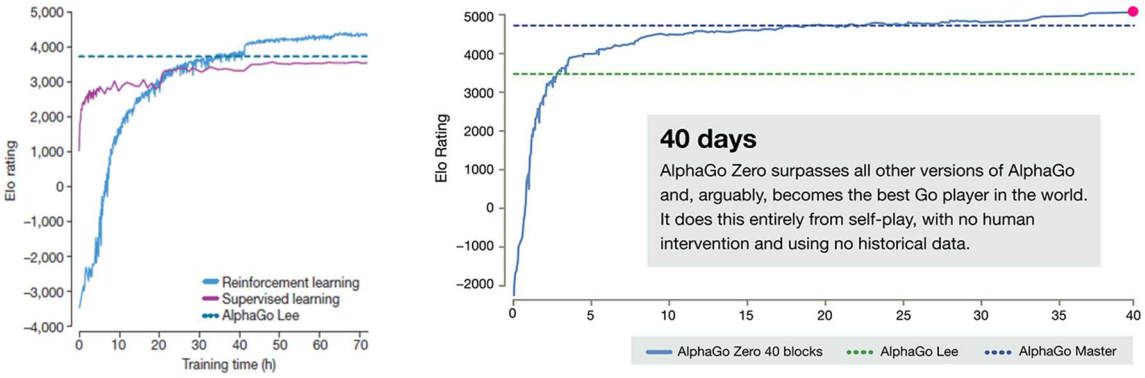

FIGURE 2 à gauche : comparaison des performances de AlphaGo Zero (Reinforcement Learning) avec la version AlphaGo ayant battu le champion du Go Lee Sedol (AlphaGo Lee) et une version de l'algorithme basée exclusivement sur un apprentissage supervisé à partir du corpus KGS de parties jouées par des humains (supervised learning). Les performances sont données selon le classement Elo.

A droite : Performance de l'algorithme après 40 jours d'entrainement. Source : note 7.

[8] TPU : architecture développée par Google comme alternative au GPU, et optimisée pour la bibliothèque logicielle TensorFLow. Les TPU deuxième génération sont d'une mémoire à haute bande passante de 600GB/s et peuvent atteindre des performances en calcul jusqu'à 45 teraflop/s.

[9] Le programme Golois, développé par T. Cazenave (Université Paris-Dauphine) qui se base aussi sur des réseaux de neurones profonds résiduels requière 10 secondes pour explorer 1600 nœuds dans l'arbre du jeu, sur une machine avec 4 GPU. Pour simuler 4.9 millions parties, avec 4 GPU, le temps d'attente serait de … 466 ans.
Residual Networks for Computer Go, Tristan Cazenave. IEEE Transactions on Computational Intelligence and AI in Games, 2017.

QUELQUES DEFIS SOCIETAUX

MAITRISE DE L'ENERGIE

Ce cas d'usage est un parmi les nombreuses applications IA envisageables pour les opérateurs de systèmes énergétiques, dont en premier lieu la distribution d'électricité.

LE PROBLEME

Prenons l'exemple d'un compteur connecté installé chez les ménages par un opérateur, capable de produire des données toutes les secondes si on le souhaite.
Pour utiliser l'apprentissage profond pour des prévisions de consommations, il est courant de réduire la taille du problème en comprimant les données par auto-encodage avant même de les utiliser dans une phase de prévision.

PRATIQUE ACTUELLE

La première solution pour faire de l'auto-encodage en préparation des prévisions de consommation consiste à prendre un sous-ensemble des données : par exemple un million de foyers, un seul point par jour représentant la consommation quotidienne. Un réseau auto-encodeur comprimant les données par 10 aura environ 500 valeurs en entrée et en sortie pour une seule variable sur un peu plus d'un an. Pour une profondeur de 3 ou 4 couches internes, le nombre de poids est de l'ordre de 100.000. Un apprentissage d'un réseau avec une convergence assez rapide en 20 itérations sur le million d'exemples peut être opéré avec une carte GPU en une journée. Pour traiter les dix variables, une semaine est nécessaire.
Pour faire du *gridsearch*, c'est-à-dire essayer des configurations diverses de réseaux, on multiplie par le nombre de configurations à tester.

Pour une prévision de consommation beaucoup plus fine à l'intérieur d'une journée, les données moyennes à la journée ne conviennent pas ; une solution est de prendre des données à la seconde avec une fenêtre glissante de quelques heures (typiquement trois heures, correspondant au cycle d'un appareil électroménager : lave-linge, lave-vaisselle). Pour un foyer, on compte 1000 exemples pour un jour, soit un milliard d'exemples pour une base d'un million de foyers.

Chaque exemple contient 10.000 secondes environ en entrée pour une variable. Le réseau auto-encodeur divisant la taille par 10 comporte typiquement 10 millions de poids. Le temps de calcul pour un apprentissage en 200 itérations (donc 400 calculs : feedforward et feedback) est approximativement 10^7x10^9x400 soit 4x10^{18} opérations, ou encore 4 millions de secondes de GPU à 1Téraflop/s.

Un *gridsearch* d'un millier de configurations aboutit à environ 4 milliards de secondes de GPU, de l'ordre de grandeur d'une centaine d'années. Pour dix variables, multiplier encore par 10. Dans ce cas la taille des données est 10 variables x 4 octets x 10000 valeurs x 10^9 exemples soit 400 téraoctets pour un jour d'un million de foyers.

CLIMAT

LE PROBLEME

Comprendre le changement climatique est l'un des défis scientifiques actuels parmi les plus importants. Dans ce but, les simulations peuvent fournir d'immenses jeux de données permettant d'évaluer l'impact de divers scénarios d'émission de carbone et des stratégies d'intervention. Les codes modernes de simulation du climat produisent des données massives : un seul cycle de 30 ans du modèle de résolution de 25 km produit 100 To de données multi-variables[10].

Dans un article récent[11], les auteurs cherchent à retrouver des événements météorologiques extrêmes à partir de ce type de données produites par un simulateur : 400 000 images haute résolution 768x768 pixels sur 16 canaux pour un volume total de 15 To. Leur but est de fournir un outil quantitatif permettant d'identifier les conditions météorologiques extrêmes pour aider les climatologues à comprendre l'évolution de ces conditions météorologiques (par exemple l'évolution de la fréquence des ouragans de catégorie 4/5) et leurs causes (c.-à-d. la probabilité qu'une activité cyclonique tropicale soit due aux émissions anthropiques, plutôt que d'être une propriété intrinsèque du système climatique).

PRATIQUE ACTUELLE

L'état de l'art dans ce domaine de la science du climat repose principalement sur des heuristiques associées à des seuils de déclenchement multi variées pour la détection des valeurs extrêmes, spécifiées par les experts[12]. Le volume des données à assimiler par un modèle pour qu'il soit performant est jugé trop important.

Un réseau de neurones profond semi-supervisé est utilisé pour produire des boîtes englobantes sur les images et des étiquettes des classes d'ouragans à reconnaître à partir des images disponibles (15 To de données). Ce réseau est composé d'un sous réseau convolutionnel entièrement supervisé pour la prédiction de la position des boîtes englobantes et d'un auto-encodeur convolutionnel non supervisé (avec deux couches cachées à convolutions avec des masques de tailles 9 et 5). La taille totale du réseau de neurones est de 302,1 Mb.

L'apprentissage du réseau de neurones profond est réalisé par un algorithme de gradient stochastique en simple précision, avec de nombreux hyper-paramètres à régler. Pour les déterminer, une grille permettant de tester différentes combinaisons est utilisée. Cette grille comprend : 2 pas différents, 3 moments, 3 tailles de mini-batch, chacune répétées 3 fois à cause de la randomisation de l'algorithme due à l'initialisation et à l'ordre de présentation des exemples. Cela donne un total de 54 réseaux de neurones à entrainer pour identifier la bonne architecture à évaluer. Les apprentissages ont été réalisés avec la distribution Intel de Caffe et

[10] M. Wehner et al., "Resolution dependence of future tropical cyclone projections of cam5.1 in the US clivar hurricane working group idealized configurations," Journal of Climate, ol. 28, no. 10, pp. 3905–3925, 2015.

[11] Thorsten Kurth et al. *Deep Learning at 15PF: Supervised and Semi-Supervised Classification for Scientific data*, International Conference for High Performance Computing, Networking, Storage and Analysis, 2017.

[12] U. Neu and et al., "Imilast: A community effort to intercompare extratropical cyclone detection and tracking algorithms," Bulletin of the American Meteorological Society, vol. 94, no. 4, pp. 529–547, 2013.

Intel Machine Learning Scalability Library (MLSL) mis en œuvre sur un système basé sur une approche hybride asynchrone avec 9600 nœuds Xeon Phi, avec une performance maximale de 15.07 pétaflop/s.

L'architecture semi-supervisée obtenue permet avec succès de détecter les événements climatiques extrêmes. En outre, elle a permis de découvrir de nouvelles caractéristiques météorologiques associées aux ouragans. Les auteurs concluent en soulignant que cette approche de reconnaissance des formes est fondamentalement nouvelle pour la communauté des sciences du climat et que, plus généralement les domaines scientifiques pouvant générer de très grandes quantités de données peuvent bénéficier dès aujourd'hui des progrès réalisés dans le domaine de l'apprentissage profond pour traiter ces données.

VEHICULE AUTONOME/VISION PAR ORDINATEUR

Cette section est un court résumé de l'article de Nvidia « training AI for self-driving vehicles : the challenge of scale » par Adam Grzywaczewski, Octobre 2017[13]

LE PROBLEME

L'apprentissage de réseaux neuronaux profonds pour les véhicules autonomes est basé sur des données obtenues à partir de véhicules réels instrumentés, parcourant des millions de kilomètres en tout, combiné avec des simulations réalistes qui produisent encore bien plus de données d'entraînement. L'auteur fait des estimations très conservatives (i.e. minimalistes) des besoins et des capacités de calcul nécessaires.

PRATIQUE ACTUELLE

Comparaison est d'abord faite avec les besoins de trois applications dans d'autres domaines :

- Reconnaissance d'images, Microsoft ResNet (2015) : 60 millions de paramètres, 7 exaflops ;
- Reconnaissance de la parole, Baidu Deep Speech (2016) : 300 millions de paramètres, 20 exaflops ;
- Traduction de texte, Google Neural Machine Translation (2017) : 8,7 milliards de paramètres, 100 exaflops.

Les estimations des besoins de données d'apprentissage : cent véhicules, roulant pendant un an (2080 heures), produisant au minimum 1 téraoctets de données par heure avec deux caméras de 30 mégapixels de résolution : 204 pétaoctets de données brutes, réduites par échantillonnage (1/30) et compression (1/70) à 104 téraoctets.
Le temps d'apprentissage pour un réseau de type Inception-v3[14] en 50 itérations, avec un seul GPU traitant 19 mégaoctets par seconde, est de 9,1 années, réduit à 1,13 années avec huit Tesla P100, le matériel actuel Nvidia.
Pour obtenir des temps de calcul plus raisonnables, la solution immédiate serait de distribuer le calcul sur plusieurs GPU.

MEDECINE GENOMIQUE

[13] https://devblogs.nvidia.com/parallelforall/training-self-driving-vehicles-challenge-scale/

[14] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2818-2826).

LE PROBLEME

La médecine actuelle se trouve confrontée à un changement de paradigme. D'une médecine fondée sur les référentiels de diagnostic et de traitement optimisés pour un patient « moyen », elle se dirige vers une personnalisation des actes médicaux prenant en charge les caractéristiques biologiques et cliniques propres à chaque patient (médecine personnalisée et médecine de précision). La quantité et la complexité de ces données deviennent telles, que l'utilisation de ressources informatiques significatives (stockage, traitement et calcul[15]) est nécessaire pour les interpréter.

L'avatar du patient ou « jumeau numérique » contient toutes les données du modèle qui découlent des multiples méthodes ou technologies permettant d'explorer la physiologie de la personne incluant l'interrogatoire médical, l'examen clinique et tous les examens complémentaires faisant appel à la biologie et à l'imagerie. L'interprétation de l'avatar nécessite le développement d'algorithmes basés sur la biologie des systèmes, la modélisation et l'intégration multi-échelle des processus biologiques et la simulation des altérations des modèles reflétant les données privées de chaque patient et les perturbations entrainées par la thérapeutique. Pour être efficace et significative, cette démarche demande que les données de très nombreux patients puissent être comparées, impliquant de les rendre interopérables après analyse du signal ou minage de texte et de repenser les systèmes d'informations médicales pour les mettre à la disposition du patient et du système de soins et s'adapter aux méthodes d'intelligence artificielle.

Parmi les sources de données, le génome de l'individu joue un rôle considérable compte tenu du poids que ses variations jouent dans la susceptibilité aux maladies et leur évolution naturelle ou sous l'effet d'une thérapeutique. Les progrès considérables des technologies de séquençage font que très prochainement le séquençage complet du génome fera partie de la pratique médicale quotidienne pour une fraction notable (si ce n'est 100%) de la population.

PRATIQUE ACTUELLE

Stephens et al.[16], ont évalué en 2015 les capacités de calcul et de traitement nécessaire pour la génomique et les ont comparés à différentes sources majeures de données (astronomie et réseaux sociaux twitter et youtube). Ils montrent que la génomique est parmi les plus exigeantes en termes de capacité de stockage (de 2 à 40 exabytes/ans) soit dix fois plus que l'astronomie et en termes de CPU. Les projections suivant la pente actuelle indiquent un doublement tous les 7 mois - à comparer avec la loi de Moore pour les semi-conducteurs qui projette un doublement tous les 18 mois.

[15] Il est difficile d'évaluer précisément les ressources nécessaires, cependant l'ordre de grandeur est supérieur à l'exascale (exabyte et exaflops).

[16] Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. (2015) Big Data: Astronomical or Genomical? PLoS Biol 13(7): e1002195. https://doi.org/10.1371/journal.pbio.1002195

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

doi:10.1371/journal.pbio.1002195.t001

La plupart des outils d'analyse du génome est à base de statistiques et d'apprentissage automatique (P.ex. GATK, 16GT). En décembre 2017, Google a annoncé la mise à disposition de DeepVariant[17] un algorithme basé sur des réseaux de neurones profonds capable de construire une image précise du génome complet sans trop d'effort. Il s'agit d'un pipeline d'analyse pour identifier des variants génétiques issus de données de séquençage d'ADN de nouvelle génération. L'une des difficultés des programmes de séquençage réside dans les parties difficiles du génome, où chacun des outils a ses forces et faiblesses[18]. Ces régions difficiles sont de plus en plus importantes pour le séquençage clinique, et il est important d'avoir plusieurs méthodes. Il est encore trop tôt pour évaluer les performances de DeepVariant, mais il confirme l'intérêt de ces approches pour la médecine génomique.

AUTRES DEFIS EN IA, NON DIMENSIONNANTS

TEST DE LA SATISFAISABILITE (SAT)

À la croisée des mathématiques discrètes et de l'informatique théorique se trouve le problème SAT, considéré comme l'archétype des problèmes difficiles. Il est ainsi le représentant des problèmes combinatoires pour lesquels il n'est pas connu d'algorithmique efficace pour leur résolution dans le cas général. Malgré cette difficulté intrinsèque, d'impressionnants progrès ont été réalisés ces dernières années pour la résolution pratique de problèmes industriels représentés dans ce formalisme (vérification de processeurs, preuve formelle, raisonnement, planification, bio-Informatique…). Cela explique pourquoi Edmund Clarke, l'un des Turing Awards 2007 ait ainsi déclaré « la résolution pratique du problème SAT est une technologie clé pour l'informatique du 21ème siècle » [SAT 2009].

La recherche SAT est rythmée par des compétitions internationales annuelles (http://www.satcompetitions.org) Elles monopolisent de nombreuses équipes de recherche et d'importants moyens en temps CPU (en 2017 un an de temps CPU a été nécessaire pour établir le classement dans la catégorie principale, en séquentiel). Lorsqu'il s'agit d'établir les classements sur les solveurs parallèles (pouvant être lancés sur des machines de 64 cœurs), ces temps s'accroissent très rapidement.

La France est bien placée sur la scène internationale SAT et plus généralement en

[17] https://github.com/google/deepvariant/

[18] Rapporté par Brad Chapman, chercheur à l'école de santé publique de Harvard, au MIT Technology Review

programmation par contraintes mais la compétition est rude et un effort semble nécessaire pour maintenir cette place. Les publications dans le domaine nécessitent de se conformer aux standards des compétitions : intégrer de nombreux problèmes (généralement plusieurs centaines), comparer aux meilleurs autres solveurs des dernières compétitions, intégrer des variantes des méthodes proposées (pour comprendre l'impact d'éventuels effets de bords de la méthode proposée), tester différents paramètres (hyper tuning des méthodes). Le cycle de développement des meilleurs SAT solveurs intègre de plus, en amont, une importante composante de « génération et test » permettant de tester différentes idées, pour ne publier que celles qui offrent de bons résultats expérimentaux. Au final, il est quasiment impossible aujourd'hui de mener une recherche de premier plan en SAT et contraintes sans avoir a disposition un cluster de calcul.

Les demandes en calcul sont principalement divisées en deux : (1) de nombreux nœuds de calculs identiques (pour comparer les performances sur de nombreux problèmes de différentes méthodes) et (2) de nombreux nœuds massivement parallèles, la recherche s'orientant vers des déploiements sur des machines parallèles à mémoire partagées (avec typiquement de 32 à 64 cœurs).

On notera que l'utilisation de GPU n'est à ce jour pas pertinente pour les approches SAT et contraintes. L'algorithmique actuelle la plus efficace (propagation des contraintes) n'étant pas parallélisable via ce type d'architecture.

On recommandera donc l'accès à des clusters de calculs homogènes ayant un grand nombre de nœuds. La recherche en SAT parallèle nécessite aussi l'existence de nœuds de calculs ayant un grand nombre de cœurs et une mémoire partagée importante.

SYSTEMES MULTI-AGENTS ET WEB SEMANTIQUE

Le domaine des systèmes multi-agents s'intéresse au développement d'applications distribuées et décentralisées, potentiellement à large échelle, qui intègrent différentes technologies du domaine de l'Intelligence Artificielle au sein d'agents autonomes coopérant entre eux. Dans cette optique, les besoins en capacités de calcul et en stockage découlent des technologies intégrées, de la complexité des problèmes abordés (p.ex. théorie des jeux algorithmiques, théorie du choix social, processus décisionnels de Markov avec leurs différentes extensions découlant de l'introduction d'incertitude, d'observabilité partielle ou d'absence de communication). Les technologies des systèmes multi-agents sont également appliquées dans le cadre de la simulation multi-agent, approche permettant de répondre à des défis sociétaux importants au travers de simulations centrées individus pour l'évaluation, par exemple, de politiques énergétiques, de gestion de crises, de transport, d'aménagement de territoire à l'échelle d'un territoire. De telles simulations requièrent des ressources importantes de calcul pour l'étude de la sensibilité des modèles impliquant des milliers d'agents en interaction et spatialisés, leur calibration et leur validation. Ces dernières années ont vu le développement d'outils de simulation multi-agent pour pouvoir fonctionner sur des grilles de calcul.

Le développement de modèles et de technologies issues du domaine de la représentation des connaissances et du raisonnement nécessite des ressources de calcul importantes. Ainsi par

exemple, le développement des nouvelles générations de moteur de recherche sur le Web sémantique et le Web de données (p. ex travaux actuels et futurs autour de QWANT) nécessite des capacités de stockage et de calcul. Il s'agit en effet d'optimiser les parcours des sites et de la toile, d'extraire et analyser les contenus et la structure du graphe de données obtenus afin de calculer des indicateurs et annotations. Pour aller plus loin que les moteurs de recherche classiques il faut des capacités de raisonnement et des approches qui passent à l'échelle pour appliquer des raisonnements enrichissant les données et leurs liens. Enfin pour répondre en temps réel aux requêtes des utilisateurs il faut une infrastructure robuste, redondante, etc. d'analyse et de résolution de requêtes. Actuellement, l'hébergement du SPARQL endpoint de DBpedia.fr demande une VM avec 48 GO de RAM, 8 processeurs et 2TO de disque SSD. Ceci n'est rien comparé aux besoins nécessaires pour crawler, indexer, faire des raisonnements de base et requêter les volumes du Web de données d'aujourd'hui et de demain.

SYNTHESE

Les exemples et défis discutés ci-dessus permettent de dégager un ensemble de constats, dont :

- L'entraînement d'un système de transcription automatique tel que *Deep Speech* de Baidu n'est pas possible pour la majorité des laboratoires académiques en France, sans parler du déploiement pour traiter plusieurs dizaines de requêtes utilisateurs en parallèle.
- L'optimisation de l'architecture d'un réseau de neurones pour des applications réelles (p.ex. par des méthodes bayésiennes) est quasiment impossible avec les infrastructures Française existantes (Google parle de l'utilisation de 800 GPUs par expérience). Un réseau typique de 10 couches est mis au point en testant des milliards de réseaux différents (https://research.googleblog.com/2017/05/using-machine-learning-to-explore.html). La plupart des chercheurs Français qui travaillent sur le « deep learning » utilisent donc des réseaux existants pré-entrainés.
- Les méthodes de « Deep Reinforcement Learning » utilisées par exemple pour entrainer AlphaGo Zero ne sont pas reproductibles par les chercheurs des universités françaises. C'est d'autant plus vrai pour des jeux encore plus complexes (avec bien plus de situations et de mouvements possibles) tels que StarCraft ([StarCraft II: A New Challenge for Reinforcement Learning, Aout 2017]). Pouvoir créer des systèmes pouvant concurrencer un joueur humain sur ce type de jeu offrirait des perspectives économiques importantes dans le monde du jeu vidéo ainsi que des problèmes de recherche stimulants en planification et en optimisation. La thématique émergente d'apprentissage par renforcement profond multi-agents (Multi-Agent Deep Reinforcement Learning) nécessiterait par ailleurs le déploiement d'infrastructures de calcul encore plus conséquentes que celles utilisées pour les cas d'usage actuels.
- S'attaquer à de grands défis sociétaux tels que la maîtrise de l'énergie, le climat, la santé, à partir du prisme de l'apprentissage automatique est porteur de ruptures scientifiques et technologiques, mais cela requiert aussi des moyens en calcul de l'ordre de pétaoctets.

2. LES INFRASTRUCTURES A DISPOSITION DE LA COMMUNAUTE

TECHNOLOGIES ACTUELLES EN HPDA

Les besoins en puissance de traitement requis par le Deep Learning ou certaines applications de l'IA se traduisent par l'utilisation d'architectures parallèles du type HPC allant du petit cluster à des configurations très puissantes, mixant souvent CPU et GPU. On peut, par exemple, citer le supercalculateur ATOS-BULL récemment installé à l'Université Oxford dans le cadre du projet national « JADE[19]». Cette machine est équipée de 22 nœuds NVIDIA DGX-1 composés chacun de 8 Tesla P100 pour serveurs optimisés NVLink, soit un total de 176 GPU P100. Chaque P100 affiche une performance de crête de 5,5 teraflop/s sur 64 bits, 10,6 Teraflop/s sur 32 bits et 21,2 teraflops sur 16 bits, fournissant donc une puissance maximale pour JADE de de 968 teraflop/s sur 64 bits, 1,86 petaflop/s en 32 bits et 3,7 petaflop/s en 16 bits.

Les architectures pour la HPDA et l'IA sont proches de celles communément utilisées en HPC avec quelques spécificités :

- Utilisation de GPU de par leur efficacité et la disponibilité de codes publics de Deep Learning permettant d'exploiter leur potentiel (performance élevée pour une consommation énergétique relativement faible). Des travaux récents démontrent aussi l'efficacité des processeurs du type « manycores » (Intel KNL par exemple) sur le Deep Learning.
- La possibilité d'utiliser des calculs sur 32 voire 16 bits permettant d'obtenir des performances bien plus élevées sur certains cas d'apprentissage
- Utilisation d'opérateurs processeurs spécialisés (e.g. TPU de Google ou tensor cores sur le NVIDIA Tesla V100).

Quasiment tous les constructeurs, à commencer par les constructeurs historiques du HPC, se sont positionnés sur les marché du Deep Learning et de l'Intelligence artificielle. Les fournisseurs de processeurs autant CPU que GPU (AMD, Intel, NVIDIA, …) affichent tous un intérêt profond pour l'Intelligence artificielle que ce soit avec des environnements orientés IA, des librairies logicielles spécialisées, voire avec l'annonces de jeux d'instructions spécifiques et d'extensions matérielles spécifiques (opérateurs, accélérateurs, …). Ainsi chez Intel on note plusieurs annonces : Knights Mill pour les réseaux neuronaux profonds à base d'une arithmétique 32 bits, accélérateurs FPGA, versions optimisées de Caffe et TensorFlow, optimisations annoncées du Skylake pour l'IA, …. Chez NIVIDIA et AMD les gammes GPU continuent à progresser en performance et à développer des librairies spécifiques pour le Deep Learning. Le modèle de GPU le plus récent de NIVIDIA, le Tesla V100 intègre des composants appelé « Tensor Cores » qui ciblent spécifiquement le Deep Learning. Les 640 tensor cores du V100 fonctionnement en mixant du calcul en 16 et 32 bits. Ainsi, le V100 avec NVLInk atteint 7,8 teraflop/s en 64 bits, 15,7 teraflop/s en 32 bits et surtout la performance impressionnante de 125 teraflop/s en Deep Learning. CRAY propose depuis quelques années des machines

[19] *Joint Academic Data science Endeavour* : projet, porté par l'Université d'Oxford, avec le soutien de l'Alan Turing Institute et des universités de Bristol, Édimbourg, Sheffield, Southampton et King's College, Queen Mary et UCL (Londres). Le projet s'est traduit par le déploiement d'un système GPU national pour soutenir les sciences multidisciplinaires en misant sur le machine Learning et la dynamique moléculaire.

orienté IA ou traitement de graphes de grande taille à base de GPU avec des suites logicielles spécifiques. HP s'est orienté vers des architectures à grand volume mémoire s'appuyant sur le savoir-faire de SGI qu'il a racheté, en explorant aussi des solutions innovantes autour de la mémoire (concept de « memory-driven computing », « Machine User Group », mémoire non volatiles, …). IBM n'est pas en reste avec depuis quelques années le programme Watson qui s'exécute sur un cluster à base de processeurs Power7 avec des applications en finance, jeux, secteur médical, assistance juridique ….. Google a conçu un processeur dédié à l'apprentissage automatique nommé TPU (TensorFlow Processing Unit). La première génération de TPU à base d'une puce 8 bits affichait 92 teraflop/s, alors que la seconde génération de TPU arrive quasiment au double (environ 180 teraflop/s par TPU). Google utilise ce composant depuis plus d'un an dans ses centres de données et affirme avoir gagné sept ans en performance comparativement à une architecture classique.

Les technologies qui se profilent pour le futur affichent plusieurs orientations avec :

- Des processeurs avec des composants, jeux d'instructions ou opérateurs dédiés, du FPGA (TPU) ou autre.
- Des tailles de mémoires très importantes.
- Des capacités de calcul dans la mémoire. Par exemple, les memristors (ou autres technologies permettant de réaliser des mémoires non volatiles comme STT-RAM), étudiés entre autres par HP, permettent la conception de mémoire non volatiles autorisant donc le stockage de données à grande échelles pour une faible consommation avec un accès bien plus rapide qu'un disque.
- L'utilisation de la photonique pour les communications (bientôt les opérations flottantes vont consommer moins que les mouvements de données dans les composants).
- D-Wave développe une offre de calculateur quantique avec 2000 qubits. Cette architecture se révèle performante sur des problèmes spécifiques (d'optimisation ou de classification binaire) mais les calculateurs quantiques auront encore besoin de longues années de développement.

INFRASTRUCTURES EXISTANTES

Le tableau ci-dessous recense quelques infrastructures et ressources publiques utilisées par la communauté de recherche française pour subvenir à ses besoins en calcul pour l'IA, notamment pour l'apprentissage automatique. Le détail est fourni en annexe 1.

Machine	Tflop/s	GPU	Stockage	commentaire
CRIANN	600	170 Tflop/s	2,5 Po	
ROMEO	254	260 (NVIDIA Tesla K20X)	247 To	260 accélérateurs NVIDIA Tesla K20X + 260 proc. INTEL IvyBridge
ICI-CENTRALE NANTES	281	14 (NVIDIA Kepler K80)	900 To	
CALMIP	274	4 (NVIDIA QUADRO 6000)	111 To	39 To de RAM
TERALAB		4 Tflop/s		
GRID 5000		72 (NVIDIA GTX , TESLA Titan)		
SACLAY-IA IDRIS		10 NVIDIA K80		
Inria Saclay		54 TFlop/s (NVIDIA GTX 1080 ti et Titan X)	50 To	

TABLEAU 1 : SYNTHESE DES PRINCIPALES CARACTERISTIQUES DE SUPERCALCULATEURS IA EN FRANCE

COMPETITION INTERNATIONALE

Cette partie présente des exemples de supercalculateurs spécialisés pour l'intelligence artificielle (IA). La table 2 montre une synthèse d'architectures recensées à travers le monde. On constate que de nombreux pays se sont dotés de ce type de calculateurs pour leur recherche académique. La plupart de ces calculateurs sont construits à partir de GPU. Des pays comme le Japon ou le Royaume Uni disposent de moyens de calculs comprenant de l'ordre de 500 GPU pour une puissance de l'ordre de 2 à 4 petaflops et 15 à 30 teraflops de mémoire. Soulignons que, question consommation énergétique, le cout des GPU est aujourd'hui de l'ordre de 2 à 3 fois plus faible que celui des CPU (http://www.nvidia.com/object/gcr-energy-efficiency.html). Au moins deux machines de grandes envergures sont annoncées pour 2018 (AI institues au Japon et Summit aux Etats Unis).
L'Allemagne est aujourd'hui absente de cette liste. Elle dispose d'un centre de recherche sur l'intelligence artificielle (DFKI, https://www.dfki.de/) qui est en train de se doter d'une telle infrastructure. Ce centre de recherche est composé de neuf centres de compétences dont un sur l'apprentissage profond (deep learning competence center, http://dl.dfki.de/).

Machine	Pflop/s	#GPU	#CPU	SSD (TB)	Disk	commentaire
Jade (UK)	0,89	176		4 – 11	1 PB	4 sur le site
Wilkes2 (UK)	1,75	360		8,6	8,6 PB	
Riken (Japon)	0,94	196		12,2		
Hélios (Canada)	0,41	216		8,5		
Surf (NL)	1,84	132	47 k			Non spécifique IA
Cloud TPU (EUA)	11,53	64				TPU (Google)
Big Basin - fb (EUA)	4,89	>256		63,0		#GPUinconnu
SaturnV Volta (EUA)	40	4800		84		Nvidia
AI institutes (Japon)	37	4 352	2 176			Annoncé 2018
Summit (EUA)		27600			10 PB	Annoncé 2018

TABLEAU 2 : SYNTHESE DES PRINCIPALES CARACTERISTIQUES DE SUPERCALCULATEURS IA A TRAVERS LE MONDE

Sources : TOP 500 et les site web des supercalculateurs concernés.
Pflops = pic théorique (Rpeak du TOP 500) ; SSD = mémoire (telle qu'indiquée dans le TOP 500).

3. ACTION NATIONALE ET ACTION EUROPEENNE

RECOMMANDATIONS DE #FRANCEIA

Les recommandations #FranceIA ont été remises le 21 mars 2017 au Président de la République, suite au travail de plus de cinq cents experts mobilisés par les Secrétariats d'Etat au Numérique et à l'Enseignement Supérieur et la Recherche entre janvier et mars 2017. Plusieurs recommandations parmi les cinquante-neuf émises par la dizaine de groupes de travail concernent la mise à disposition de ressources de calcul et de stockage au service de la communauté de recherche en intelligence artificielle : nous les synthétisons ci-après. Elles proviennent essentiellement de cinq GT. On observe une grande convergence sur la volonté de disposer de moyens de calcul, de grandes bases de données et de logiciels.

Le GT « recherche amont » propose une très grande infrastructure pour l'IA avec capacités de calcul en particulier GPU, bases de données étiquetées et de corpus y compris en français, accès aux logiciels développés par les équipes de recherche, et des personnels pour opérer l'infrastructure.
Le GT « formation » recommande la mise en place d'une ressource nationale de données de toutes natures (voix, images, textes, signaux, vidéos etc.) non agrégées, structurées et labellisées en français.
Le GT « transfert de technologies » recommande la mise en place de plusieurs plateformes pour le parangonnage, l'expérimentation et la démonstration des technologies ; parmi les plateformes proposées, il y a notamment : plateforme de données, logiciels, ressources de calcul pour l'apprentissage ; plateforme de ressources pour le traitement automatique de la langue naturelle, l'interaction homme-machine et les agents conversationnels.
Le GT « véhicule autonome » recommande la mutualisation des bases de données d'enregistrements capteurs (caméra, radar, GPS, accéléromètre, etc.) pour la compréhension, l'apprentissage statistique et l'évaluation.
Le GT « Souveraineté et Sécurité Nationale » recommande la création d'une plateforme logicielle intégrative d'intelligence artificielle, qui vise à fédérer un écosystème entre recherche et industrie en vue de développer une solution technologique nationale concurrente aux solutions extra-européennes.

PROGRAMME EUROPEEN

La Commission Européenne met l'accent sur la création d'une plateforme « AI-on-demand » destinée à recevoir des briques technologiques provenant de tous les états membres, et à permettre la composition de solutions à partir de ces briques. Cette plateforme fait l'objet d'un appel à propositions dans le programme 2018-2019, doté de 20M€ pour un projet unique, avec la promesse de renouvellement en 2020 pour un montant identique. On voit qu'il ne s'agit pas de ressources de calcul et de données – il ne s'agit pas d'une plateforme d'exécution - mais de moyens pour intégrer des technologies au sein d'applications tournées vers les utilisateurs. Il s'agit donc d'une action très complémentaire à la proposition que nous soutenons.

4. RECOMMANDATIONS DU GROUPE DE TRAVAIL

Les sections précédentes justifient amplement la demande de mise en place d'une grande infrastructure nationale pour l'IA, en particulier pour les besoins de l'apprentissage machine dont notamment le *deep learning* et l'apprentissage par renforcement. Non que l'infrastructure ne soit utilisable que pour ces seules technologies, puisque l'on a vu des besoins significatifs pour le raisonnement SAT à grande échelle et pour le web sémantique, par exemple, mais ce sont celles qui demandent le plus de ressources (notamment capacité de calcul, mémoire, stockage) et sont donc dimensionnantes pour l'infrastructure que nous appellerons « GENIAL[20] » (Grand Equipement National pour l'Intelligence Artificielle) dans la suite de ce document.

Nous recommandons la mise en place **d'un équipement mono-localisé hébergé dans l'un des sites nationaux de calcul intensif**, ceci afin de maximiser l'efficacité de l'équipement, de maîtriser sa maintenance et sa fiabilité, d'optimiser les coûts, et de pouvoir y associer un centre de compétences. La qualité des accès réseaux rend cette mono-localisation transparente pour les utilisateurs. L'autre hypothèse que nous avons envisagée, une infrastructure distribuée à la Grid'5000, ne correspond pas au besoin de la communauté. Il faudra cependant accorder une attention particulière au transport des données nécessaires aux applications : RENATER ne permet pas, dans sa capacité actuelle, de transférer des masses de données de centaines de téraoctets avec un délai raisonnable. Les données, une fois localisées dans GENIAL, ne voyageront pas aisément.

DIMENSIONNEMENT

Afin d'être présent dans la compétition internationale, **une machine d'une puissance utile de 5 pétaflops environ, principalement à base de GPU (entre 500 et 1000 GPU), est le minimum envisageable**. Cela correspond, par exemple, au double de la capacité de calcul en IA de l'Université d'Oxford pour son propre usage. C'est aussi dans l'ordre de grandeur des principaux équipements universitaires mondiaux existants ou en préparation, comme celui de RIKEN au Japon et du DFKI en Allemagne.
Il est important que GENIAL dispose également d'une **grande capacité mémoire**, indispensable pour les applications envisagées, et d'un **espace de stockage de très grande capacité, tant en SSD qu'en disques durs** (de 15 à 30 teraoctets de SSD et de 10 à 15 pétaoctets) afin d'accueillir les données d'applications.

UN CENTRE DE COMPETENCES DEDIE

Nous recommandons d'associer à GENIAL un centre de compétences, d'une dizaine de personnes chargées de :

- Faciliter l'accès au supercalculateur avec une aide au portage des codes,
- Se maintenir à l'état de l'art pour évaluer et comparer les méthodes existantes pour l'aide à l'innovation en IA,
- La formation d'utilisateurs non-experts de technologies de l'IA,

[20] Nom provisoire

- Mettre à jour les logiciels utiles en IA,
- Gestion et mise à disposition des données pour l'IA,
- Soutien à des hackathons et des écoles d'été en IA,
- Développer des cas d'usage et des démonstrateurs des dernières techniques en IA (en association avec un réseau de living labs ?),
- Réaliser des études de marché, des livres blancs et des études de faisabilité en IA.

BUDGET

Le coût d'acquisition initial sera aux environs de **5M€** - coût précis à définir au moment de la rédaction du cahier des charges – ce à quoi il convient d'ajouter, sur une période initiale de cinq ans :

- Un coût de **mise en place** logiciel et matériel de l'ordre de 100K€ ;
- Un coût **d'exploitation annuel** (fluides : électricité, climatisation) de 500K€ environ
- Un **coût humain d'exploitation** annuel de 600K€ environ (techniciens, ingénieurs) ;

Le modèle économique, détaillé plus loin, permettrait d'envisager une **contribution significative au coût de fonctionnement annuel à hauteur de 20%** provenant d'entreprises pour des projets propres ou des projets collaboratifs comme FET européens, ANR ou PIA.

MODALITES D'UTILISATION

Un point essentiel est que **l'accès à GENIAL soit très souple**, car le mode de recherche particulier de l'IA fait qu'il est difficile voire impossible de prévoir longtemps à l'avance de quelle quantité de calcul on aura besoin ; la recherche (et les sessions expérimentales en enseignement) se fait souvent par essai/erreur, analyse d'une exécution y compris en temps réel, modification des paramètres et nouvel essai. Par ailleurs, des acteurs privés proposent des plateformes de calcul en IA accessibles sur le web, attractives pour la communauté de l'ESR, avec une très grande qualité de service. Nous prônons donc une certaine flexibilité dans l'allocation de la ressource, au moins pour une partie significative, et une qualité de service au moins aussi comparable aux offres privées.
En complément, nous considérons qu'il est également possible qu'une partie de la ressource soit réservable à l'avance, afin d'obtenir la garantie de pouvoir faire des expérimentations en particulier mobilisant une grande partie du calculateur.
Pour ces raisons, nous proposons une utilisation basée sur **deux partitions** de l'équipement GENIAL, partitions dans l'espace (un sous-ensemble des ressources) voire éventuellement dans le temps (périodes consacrées à l'une ou l'autre des modalités) :

- Une **charte d'utilisation** de l'équipement devra être signée par tout utilisateur afin d'obtenir un compte ; cette charte précisera notamment les modes d'accès et de contrôle tels que décrits ci-dessous. Le non-respect de la charte entraînera le bannissement de l'utilisateur.
- Une **partition en mode ouvert**, « premier entrant, premier servi , une fois passée l'étape d'accréditation pour obtenir un compte. Ce mode de fonctionnement est faisable puisque déjà expérimenté dans GRID'5000. De plus, dans cette partition, chaque utilisateur peut réquisitionner les ressources disponibles – en partie et même en totalité – pour ses expérimentations. Un comité opérationnel examine périodiquement

(p.ex. hebdomadairement, bimensuellement, mensuellement) l'utilisation de la partition afin d'éliminer les squatteurs qui mobiliseraient trop de ressources pendant un temps trop long sans bonnes raisons.

- Une **partition en mode réservation**, avec des heures attribuées annuellement par un comité scientifique sur la base de propositions de projets ; à cette partition ne peuvent accéder que les utilisateurs qui disposent d'un crédit d'heures attribué, seul le contrôle de la quantité d'heures est effectué puisque les projets ont fait l'objet d'une évaluation scientifique et technique en amont.
- Le dimensionnement temporel (périodes) et spatial (ressources) de ces partitions devrait être **adaptable à la demande**. Une possibilité simple serait de revoir annuellement le partitionnement au vu de la demande de réservation ; mais nous n'excluons pas des méthodes d'adaptation plus fréquentes si cela se révèle praticable.
- Enfin, nous recommandons qu'un **accord soit établi entre GENIAL et d'autres centres de calculs nationaux ou régionaux**, afin de reverser le trop-plein éventuel vers ces autres centres (par exemple, le cluster Gulliver d'Inria Saclay, ou Grid'5000, ou ...) et inversement. Pour cela il faudra bien entendu organiser l'interopérabilité et la portabilité des applications, ainsi que la transportabilité des données d'entrée – en quantité raisonnable. Cette option ne peut fonctionner que pour des applications ne nécessitant pas de déplacer des très grandes quantités de données, et ne demandant pas de configuration spécifique seulement disponible sur la machine GENIAL.

COMPARTIMENT SECURISE

Afin de favoriser l'accès aux données à fortes valeurs (propriété industrielle ou intellectuelle) ainsi que les collaborations public-privé, un compartiment sécurisé est nécessaire au sein de l'infrastructure GENIAL. En effet il est nécessaire d'apporter les garanties de sécurité, de souveraineté et de neutralité pour accompagner et permettre aux industriels de mettre à disposition leurs données pour des projets de recherche ou d'innovation avec des laboratoires de recherches, des entreprises innovantes ou encore des étudiants.

Plusieurs niveaux de sécurité seront nécessaires :

- Un premier niveau permettant d'isoler l'espace de travail aussi bien d'un point de vue performance de calcul que cloisonnement des données ;
- Un second niveau au-dessus de cet espace de travail doit pouvoir s'adapter en fonction du besoin du projet en partant d'un niveau ouvert à un niveau très restreint, ne permettant pas la sortie de données par les chercheurs autorisés à travailler sur le projet.

Cet accès sécurisé a un cout et devra être pris en compte par le fournisseur des données.

MODELE ECONOMIQUE

Le principe de base est la **gratuité de l'utilisation pour la communauté nationale** de recherche et d'enseignement supérieur : chercheurs, enseignants-chercheurs, étudiants de l'enseignement supérieur (Masters Recherche et doctorants), pour leurs besoins propres, hors projets collaboratifs. Que ce soit en mode ouvert ou en mode réservation, l'accès à la ressource sera gratuit.

Nous avons aussi la volonté d'**ouvrir la ressource à des entreprises**, d'une part pour soutenir la compétitivité et l'économie nationale ; d'autre part, si les conditions le permettent, pour aider à constituer des grandes bases de données de référence dans les domaines applicatifs.
Nos recommandations sur ce point comportent quatre items :

- Pour **une entreprise (grand groupe, PME, ETI) demandant un accès pour ses propres besoins**, sans relations avec un établissement de recherche ou d'enseignement supérieur[21] : accès facturé au coût réel, que l'on imagine équivalent à celui pratiqué par les grands fournisseurs de cloud sur le marché, tels que Amazon (AWS), OVH etc. Cet accès facturé n'entraîne aucune contrainte de quelque sorte, aucun contrôle de la validité scientifique ou technique de l'utilisation ;
- Pour **une entreprise (grand groupe, PME, ETI) ou un établissement de recherche et d'enseignement supérieur demandant un accès dans le cadre d'un projet collaboratif** financé par des fonds publics : accès facturé au coût réel, dans les mêmes conditions que ci-dessus, prise en charge par le projet. Il faudra donc veiller à ce que les propositions de projets collaboratifs (ANR, PIA, Europe, régions ...) incluent les coûts prévisionnels correspondants.
- Nous recommandons, en complément, la mise en place d'un **dispositif spécifique pour les start-ups**, afin de leur fournir la possibilité de faire des expérimentations de faisabilité technologique (« galops d'essai »). Le modèle tarifaire pourrait être analogue à celui pratiqué dans le programme SIMSEO piloté par TERATEC et GENCI, à savoir cofinancement à 50-50 par la startup et par les fonds publics. A priori ce dispositif, puisque déjà en place dans d'autres cadres, est compatible avec les contraintes d'encadrement communautaire européennes.
- Nous souhaitons impulser une **politique de la donnée** qui incite les utilisateurs de l'équipement, en particulier industriels, à déposer des données qualifiées sur GENIAL. Ainsi, comme le pratique la commission européenne, et moyennant des contraintes éventuelles de confidentialité ou de propriété intellectuelle spécifiques au projet, le dépôt des données applicatives d'un projet sera fortement encouragé, en particulier des données d'entraînement nécessaires aux algorithmes d'apprentissage.

EVOLUTIVITE

Il est important de prévoir des financements ultérieurs (p.ex. annuels grâce à la facturation aux entreprises et projets collaboratifs) afin d'assurer l'évolutivité de GENIAL qui doit rester au meilleur niveau international et à l'état de l'art des outils et technologies de calcul. Des investissements supplémentaires après cinq ans sont à considérer.

[21] Ceci s'applique également au cas du laboratoire académique qui demanderait à accéder à GENIAL dans le cadre d'un projet bilatéral avec une entreprise, financé par celle-ci.

5. ANNEXES

ANNEXE 1 : INFRASTRUCTURES NATIONALES

CRIANN

Le Centre Régional Informatique et d'Applications Numériques de Normandie *(http://www.criann.fr/le-criann/) « a pour mission d'aider les organismes publics et privés normands à développer des activités d'enseignement, de recherche et de développement basées sur l'utilisation des nouvelles technologies de communication et sur l'informatique. Pour cela, le CRIANN déploie des infrastructures informatiques à haut niveau de performance au service de l'enseignement supérieur, de la recherche et de l'innovation en Normandie. ».*
Le CRIANN dispose d'une solution ATOS BULL, nommée Myria, dont les caractéristiques sont comme suit :

- Puissance crête théorique totale : 600 TFlops (403 TFlops Xeon, 170 TFlops GPU et 27 TFlops Xeon Phi KNL)
- Capacité de stockage : espace disque de 2,5 Po.
- GPU : 170 TFlops GPU
- Réseau d'interconnexion : réseau d'interconnexion à faible latence et haut débit (100 Gbit/s)

ROMEO

https://romeo.univ-reims.fr/pages/presentation_hardware
ROMEO (https://romeo.univ-reims.fr) est une plateforme de calcul de l'université de Reims Champagne-Ardenne soutenue par la région Champagne-Ardenne.

- Puissance crête théorique totale *:* 254,0 TFlop/s
- Capacité de stockage : 247 To d'espace disque
- GPU : deux accélérateurs NVIDIA TESLA K20X

ICI-CENTRALE NANTES

Source : http://calcul.math.cnrs.fr/spip.php?article276
Le mésocentre Centrale Nantes de l'Institut de Calcul Intensif

- Puissance crête théorique totale *:* 281 TFlop/s
- Stockage : 900 To
- GPU : 14 nœuds sont équipés chacun de 256Go de RAM et d'une carte accélératrice Nvidia (Kepler K80) avec 2 unités de traitement GPU par nœud.

CALMIP

https://www.calmip.univ-toulouse.fr/spip.php?article388&lang=fr

- Puissance crête théorique totale *:* 274 TF
- Stockage : *39 To RAM +* 111 To de disque
- GPU : 4 GPU (nvidia Quadro 6000)

TERALAB

Teralab (www.teralab-datascience.fr) est une plate-forme Big Data et IA sécurisée, sûre & souveraine pour la recherche, l'enseignement et l'innovation. Elle permet aux industriels de partager leurs données avec la recherche et l'innovation afin de lever des verrous scientifiques et technologiques au plus près de leurs valeurs et fournit les outils, architectures clé en main, et le support pour que les acteurs puissent se concentrer le plus rapidement possible sur leurs problématiques.

- Serveurs classiques :
    - VCPU = 1100
    - stockage = *500TB*
    - RAM : 14TB
- Serveurs GPU :
    - Cores : 13500
    - RAM : 60GB

50 projets ont pu bénéficier de la plateforme depuis début 2014. La plateforme est mise à jour régulièrement pour proposer des outils et machines à l'état de l'art.
TeraLab a un taux d'utilisation moyen de 70%.

GRID 5000

Grid 5000 (https://www.grid5000.fr/) est un banc d'essai à grande échelle et polyvalent pour la recherche expérimentale dans tous les domaines de l'informatique, avec un accent sur l'informatique parallèle et distribuée, y compris le Cloud, HPC et Big Data. Grid'5000 s'est équipé récemment d'accélérateurs graphiques et dispose désormais de 76 accélérateurs (72 NVIDIA GPU et 4 Intel Xeon Phi) répartis entre Lille, Lyon et Nancy.

SACLAY-IA IDRIS

Le Center for Data Science de l'université Paris-Saclay dispose d'une grappe de calcul GPU (5 serveurs équipés chacun de 2 cartes K80) installée à l'IDRIS.

ANNEXE 2 : INFRASTRUCTURES INTERNATIONALES

Nous détaillons ci-après, pays par pays, les différentes infrastructures publiques dédiées à l'IA recensées.

GRANDE BRETAGNE

Organisme : Oxford et autres univ UK. (the Jade Cluster)
Machine : 492.2 Tflops - 22 DGX1 (= 176 P100), 4 TB of SSD 1PB of Seagate cluster storage
Constructeur : Atos (opéré par STFC Hartree)
mise en service : nov 2017 ?
http://people.maths.ox.ac.uk/~gilesm/JADE/
ranked #11 on the TOP20 of the Green500 - https://www.top500.org/green500/list/2017/06/

Organisme : Cambridge – The Wilkes2 cluster
Machine : 1,193.0 Tflops , 360 P100, in 90 Dell EMC server nodes. 8,640 GB,
Constructeur : Dell EMC
mise en service : upgrades nov 2017
https://www.hpc.cam.ac.uk/services/wilkes
ranked #9 on the TOP20 of the Green500 - https://www.top500.org/green500/lists/2017/11/

SUISSE

Organisme : Swiss National Supercomputing Centre (CSCS, non spécifique AI)
Machine : Piz Daint Cray XC40/XC50, 169 TB DDR3 32 TB non-ECC GDDR5, 2.5 PB
Constructeur : Cray
mise en service : upgrade 2016 (multi usage)
http://www.cscs.ch/computers/piz_daint_piz_dora/index.html
ranked #10 on the TOP20 of the Green500 -
https://www.top500.org/green500/lists/2017/11/

Organisme : Swiss National Supercomputing Centre (CSCS)
Machine : Blue Brain 4, 839 TFlops 65 TB of RAM 128TB
Constructeur : IBM

JAPON

Organisme : Riken
Machine : 635.1 Tflops , 24 DGX1,
Constructeur : NVIDIA
Le centre de Calcul AI du Riken est composé de 24 DGX1 NVIDIA. Il est classé numéro 8 du top 500 des « super green computers »
http://www.riken.jp/en/research/labs/aip/
ranked #8 on the TOP20 of the Green500 - https://www.top500.org/green500/lists/2017/11/

Organisme : AI research center at the University of Tokyo (National Institute of Advanced Industrial Science)
Machine : 1,088 servers, 2,176 Intel Xeon processors, and 4,352 NVIDIA GPUs (37 Petaflop)
Constructeur : Fujitsu
mise en service : annoncée en 2018

37 Petaflop supercomputer. “Targeted at Deep Learning workloads, the machine will power the AI research center at the University of Tokyo’s Chiba Prefecture campus. The new Fujitsu system feature will comprise 1,088 servers, 2,176 Intel Xeon processors, and 4,352 NVIDIA GPUs.”
https://asia.nikkei.com/Business/Companies/Fujitsu-to-build-world-class-AI-supercomputer
http://www.ai.u-tokyo.ac.jp/index-e.html

EUA

Organisme : Google Cloud TPU
Machine : 64 TPU de seconde génération, jusqu’à 11.53 PF,
Constructeur : Google
mise en service : Mai 2017
https://www.blog.google/topics/google-cloud/google-cloud-offer-tpus-machine-learning/

Organisme : NERSC and Stanford (CORI)
Machine (software): Machine : Cori ranked as the 5th most powerful supercomputer in the world on the November 2016 list of Top 500 supercomputers in the world. Cori is a unique among supercomputers of its size with two different kinds of nodes, 2,388 Intel Xeon "Haswell" processor nodes 9,688 Intel Xeon Phi "Knight's Landing" nodes. Cori also features a 1.8 PB Cray Data Warp Burst Buffer with I/O operating at a world's-best 1.7 TB/sec.
Constructeur : INTEL/CRAY
mise en service : Aout 2017
http://www.nersc.gov/users/computational-systems/cori/
https://www.hpcwire.com/2017/08/28/nersc-scales-deep-learning15-pflops/

Organisme : IBM TrueNorth Neuromorphic System
Machine : 256 NVIDIA GPUs in 64 IBM Power systems
Constructeur : IBM
mise en service : Aout 2017 ? https://www-03.ibm.com/press/us/en/pressrelease/52657.wss
https://www.ibm.com/blogs/research/2017/08/distributed-deep-learning/

Organisme : NVIDIA DGX SaturnV Volta
Machine : 660 Pflop-AI (soit 40 Pflop FP 64), 528 V100 GPU, Xeon E5-2698v2, 16,896 GB
Constructeur : NVIDIA
mise en service : 2017 ?
https://www.nvidia.com/en-us/data-center/dgx-saturnv/
ranked #4 on the TOP20 of the Green500 - https://www.top500.org/green500/lists/2017/11/
usage recherche notamment en health care

Organisme : Big Basin (FaceBook)
Machine : ??? 256 GPU (Y. LeCun)
Constructeur : NVIDIA
mise en service : Juin 2017 ?
https://code.facebook.com/posts/1835166200089399/introducing-big-basin-our-next-generation-ai-hardware/
https://research.fb.com/publications/accurate-large-minibatch-sgd-training-imagenet-in-1-hour/

ResNet-50 gets to 42K img/s on 352 GPUs, resulting 30s per epoch and hence <1hr for 90 epochs. For ResNet-50, the total flops for training (FP+BP+WU) is ~12.3 Gflops per image, so a throughput of 42K img/s means that achieved performance is 525 TeraFlops, on 325 GPUS (p100). https://www.top500.org/system/179068.

Organisme : Summit (as a science solver) Oak Ridge National Laboratory
Machine : 4,600 noeuds avec 6 GPU volta chacun (135 Petaflop) et 800 gb NVRAM
Constructeur : IBM/NVIDIA ???
mise en service : annoncée en 2018 https://www.olcf.ornl.gov/summit/

CANADA

Organisme : Mc Gill Accelerator Technologies Cluster (Guillimin - ATC)
Machine CPU + Dual nVidia K20+ Dual Intel Phi 5110P, 1 TB memory 5PB sur disk.
+ Cloud Computing Services (with IBM) + Data storage et big data analytics
+ HPC consulting
Constructeur : INTEL + NVIDIA
http://www.hpc.mcgill.ca/index.php/starthere/81-doc-pages/215-guillimin-hardware#atc
http://www.hpc.mcgill.ca/index.php/services
Institut de valorisation des données - https://ivado.ca/

Organisme : Calcul Quebec (cluster Hélios)
Machine : 216 GPU, logiciels disponibles Cuda, Python, Java
Constructeur : Bull, CentOS 6.6 (NVIDIA GPU K20 et K80)
Helios est composé de 15 nœuds de calcul disposant chacun de huit GPU K20 de nVidia, et de 6 nœuds de calcul disposant chacun de huit cartes K80 nVidia. Chaque carte K80 contient deux GPU portant le total de GPUs à 216 GPUs.
https://wiki.calculquebec.ca/w/Helios/fr
utilise les technologies Globus Connect (transfert rapide et efficace de jeux de données volumineux) et une version personnalisée de Globus Publication
https://www.globus.org/, https://www.top500.org/system/178466
https://wiki.calculquebec.ca/w/Tableau_r%C3%A9sum%C3%A9_des_propri%C3%A9t%C3%A9s_des_serveurs_de_Calcul_Qu%C3%A9bec_(acc%C3%A9l%C3%A9rateurs)

HOLLANDE

Organisme : Surf
Machine : 47,776 cores + 132 GPUs: 1.843 Pflop/s (peak performance)
Constructeur : Atos/Bull
Plateforme big data
https://www.surf.nl/en/services-and-products/big-data-services/index.html
avec jupyter notebook

Pour mémoire : Top AI plateform : https://www.predictiveanalyticstoday.com/artificial-intelligence-platforms/
Azure, Google cloud API, Amazon web service, IBM.