Travail effectué du 16 au 25 mai 2022 :

1. Continuer Recherche sur le cycle de vie des sources d’énergie avec l’énergie solaire
2. Chercher et comprendre les termes suivants : économie transactionnelle et régulée, les libertariens, concept de volatilité en économie et la valeur réelle
3. Faire un tableau comparatif des différents articles étudiés et améliorer leurs analyses surtout les critiques
4. Faire la même étude pour les articles lus en mineur et les ajouter au tableau
5. Commencer à rédiger une synthèse en anglais de ces articles
6. Résoudre l’exercice de machine learning : classification

Energy and Policy Considerations for Deep Learning in NLP

La problématique :

Quantifier les coûts financiers et environnementaux approximatifs de la formation d'une variété de modèles de réseaux neuronaux ayant connu un succès récent pour le traitement automatique des langues, afin de sensibiliser les chercheurs.

La méthode :

* Les auteurs de l’article effectuent une analyse de l'énergie nécessaire à l'entraînement d'une variété de modèles populaires de NLP (Transformer, ELMo, BERT, GPT-2), ainsi qu'une étude de cas de la somme complète des ressources nécessaires au développement de LISA y compris tous les réglages (tuning) et l'expérimentation. Pour mesurer l'utilisation de l'énergie les auteurs entraînent les modèles utilisant les paramètres par défaut fournis, et échantillonnent la consommation d'énergie du GPU et du CPU pendant l'entraînement. Chaque modèle a été entraîné pendant un jour maximum. Tous les modèles sont formés sur un seul GPU NVIDIA Titan X, à l'exception d'ELMo qui a été formé sur 3 GPU NVIDIA GTX 1080 Ti. Pendant l'entraînement, un échantillonnage de la consommation d'énergie du GPU est effectué de manière répétée. Les auteurs ont ensuite calculé la moyenne de tous les échantillons. Pour échantillonner la consommation d'énergie du CPU, l’interface Running Average Power Limit d'Intel a été utilisée. Le temps total prévu pour que les modèles s'entraînent jusqu'à la fin a été estimé en utilisant les temps d'entraînement et le matériel indiqués dans les articles originaux. Ensuite, la consommation d'énergie en kilowattheures (kWh) est calculée comme suit :

Soit pc la consommation moyenne (en watts) de tous les supports de CPU pendant la formation, soit pr la consommation moyenne de tous les supports de DRAM (mémoire principale), soit pg la consommation moyenne d'un GPU pendant la formation, et soit g le nombre de GPU utilisés pour la formation. La consommation d'énergie totale est estimée en combinant la consommation du GPU, du CPU et de la DRAM, puis en la multipliant par le PUE (Power Usage Effectiveness), qui tient compte de l'énergie supplémentaire requise pour soutenir l'infrastructure informatique (principalement le refroidissement). Le coefficient PUE utilisé dans les calculs des auteurs et de 1,58, soit la moyenne mondiale de 2018 pour les centres de données (Ascierto). Ensuite, la puissance totale pt requise à une instance donnée pendant l'entraînement est donnée par l’équation suivante :



Le résultat obtenu est ensuite multiplié par le CO2 moyen produit (en livres par kilowattheure) pour l'énergie consommée aux États-Unis (EPA, 2018). D'après les auteurs comme la répartition de l'énergie aux États-Unis est comparable à celle du service cloud le plus populaire, Amazon Web Services, cette conversion fournit une estimation raisonnable des émissions de CO2 par kilowattheure d'énergie de calcul utilisée.

Une image contenant table

Description générée automatiquement

Résultats :

Comparaison des différents modèles :

Une image contenant table

Description générée automatiquement

Les TPU sont plus rentables que les GPU pour les charges de travail, les modèles émettent des émissions de carbone substantielles ; l'entraînement de BERT sur GPU est à peu près équivalent à un vol transaméricain. So et al. (2019) rapportent que le NAS atteint améliore sa performance de traduction automatique de l'anglais vers l'allemand de seulement 0,1 BLEU mais avec un coût d'au moins 150 000 $ en temps de calcul à la demande et des émissions de carbone non négligeables.

Etude de cas :

Une image contenant texte

Description générée automatiquement

Coût estimé en termes de cloud computing et d'électricité pour la formation : d'un seul modèle, d’un seul tuning et de tous les modèles formés pendant la R&D.

L'entraînement d'un seul modèle est relativement peu coûteux, cependant le coût de la mise au point (tuning) d'un modèle pour un nouvel ensemble de données, estimé par les auteurs ici à 24 emplois, ou la réalisation de l'ensemble de la R&D nécessaire pour développer ce modèle, devient rapidement extrêmement coûteux.

Contribution personnelle :

3 recommandations :

* Les auteurs doivent indiquer et partager le temps de formation et sensibilité aux hyperparamètres :

Permettre de comparer directement différents modèles afin de réaliser une analyse coûts-avantages (précision).

* Les chercheurs universitaires ont besoin d'un accès équitable aux ressources informatiques : empêcher le monopole des grands groupes sur les algorithmes complexes pour permettre à plus de chercheurs d’étudier ces algorithmes et limiter la pollution de réentraîner des algorithmes simples ;
* Les chercheurs doivent privilégier les hardwares et les algorithmes efficaces en termes de calcul :

Développer un algorithme capable de résoudre plusieurs problèmes, plutôt que d’implémenter un algorithme en réponse à chaque problème.

L'impact des résultats :

Critique personnelle :

Axe d’amélioration :

Implémentation :

Quantifying the Carbon Emissions of Machine Learning

La problématique :

Comment quantifier les émissions de carbone de l'apprentissage automatique ?

Et développement d’un outil appelé le « Machine Learning Emissions Calculator »

La méthode :

* L’outil proposé par les auteurs prend en entrée plusieurs détails concernant l'entraînement du modèle de Machine Learning qui sont les suivants : le cloud utilisé, la zone géographique du serveur, le type de GPU utilisé ainsi que le temps d'entraînement de l'algorithmes. Après avoir analysé tous ces facteurs, l'outil donne en sortie la quantité approximative de CO2eq produite. Pour ces 4 variables, les chercheurs ont utilisé des données publiquement disponibles.
* Afin de faire une estimation de la quantité de CO2eq émis par les serveurs cloud selon leur localisation. Les auteurs supposent que tous les serveurs sont connectés aux réseaux locaux à leur emplacement physique. Ainsi, des données concernant les émissions de CO2eq de différents emplacements de réseau ont été rassemblées et croisées avec les emplacements connus des serveurs GPU des trois principaux fournisseurs de cloud : Google Cloud Platform, Microsoft Azure et Amazon Web Services.

Résultats :

Contribution personnelle :

Les auteurs ont développé un outil capable de faire une approximation des émissions de CO2eq

4 recommandations :

* Choisir son cloud de façon intelligente :

Depuis quelques années, les fournisseurs de cloud ont commencé à mettre en place des objectifs pour compenser leurs émissions grâce à des certificats d’énergie renouvelable (REC) dans le but de devenir neutre en carbone. Un REC atteste que 1MWh d’énergie renouvelable a été utilisée. Par exemple, Google Cloud Platform est certifiée neutre en carbone et finance des parcs solaires et éoliens grâce aux REC, selon une estimation faite en 2016 Microsoft Azure est également certifié neutre en carbone et 44 % de sa consommation d'électricité provient directement d'énergies renouvelables. Enfin, d’après les auteurs de cet article, Amazon Web Services finance également des projets d'énergie renouvelable et certains de ses centres de données sont alimentés par des énergies renouvelables.

De plus il est aussi important de prendre en compte l’indicateur d’efficacité énergétique (pue) des centres où sont hébergés les GPU, qui représente le pourcentage de la consommation d'énergie utilisé pour le refroidissement, la conversion d'énergie et d'autres tâches auxiliaires, et qui peut varier énormément. Par exemple, Google Cloud Services a un PUE moyen de 1,1, ce qui signifie que seulement 11 % de sa consommation totale d'énergie n'est pas utilisée pour les serveurs.

* Choisir le lieu de son data center :

Bien que plusieurs fournisseurs de cloud soient neutres en carbone, certains de leurs data center peuvent tout de même émettre beaucoup de carbone dû au lieu où ils se trouvent. Il est donc très important de bien choisir le lieu du data center où l’algorithme s’entrainera. Ce seul choix peut faire varier les émissions directes d'un algorithme d'un algorithme par un facteur de 40, et on peut donc passer de 20g CO2eq/kWh dans un lieu utilisant des sources d'énergie renouvelables à 820g CO2eq/kWh dans un lieu dépendant uniquement des combustibles fossiles. Pour un modèle qui sont entraînés sur plusieurs GPU pendant plusieurs semaines, cela peut permette d'éviter d'émettre plusieurs centaines de kilogrammes de CO2eq.

* Limiter la perte de ressources :

Il est très important de faire les bons choix lorsque des tests et des expériences sont faites car ils peuvent être la raison d’énergies utilisées à perte. Par exemple la recherche par grille est encore fortement utilisée alors que son efficacité en termes de performances du modèle ainsi que son impact environnemental sont mauvais. Il a été aussi démontré que la recherche aléatoire est remplacement simple mais a également le potentiel d’accélérer de manière significative la recherche d'hyperparamètres réduisant ainsi les émissions de carbone. De plus, on peut réduire le nombre d'expériences ratées grâce à une conception soignée, telle que des tests unitaires, des tests d'intégration et un débogage précoce et étendu. Il faut prendre le temps d'effectuer une analyse documentaire.

* Bien choisir son Hardware :

Le choix du matériel peut aussi influer sur les émissions du Machine Learning. Pour effectuer une comparaison entre différents dispositifs, leur efficacité de calcul peut être estimée en FLOPS/W. Cette estimation est basée sur leur performance maximale théorique par rapport à leur puissance de conception thermique (TDP). En utilisant cette approche, on constate que les CPU peuvent être 10 fois moins efficaces que les GPU, tandis que les TPU 3 peuvent être 4 à 8 fois plus efficaces que les GPU.

L'impact des résultats :

Critique personnelle :

Axe d’amélioration :

Implémentation :

Fiche économie :

* Économie transactionnelle : pas d’interactions humaine que du machine-machine
* Economie régulée : Une économie avec des réglementations. Il s'agit de l'application de régulation par le gouvernement ou les organismes de réglementation à des fins diverses, notamment pour remédier aux défaillances du marché, protéger l'environnement et gérer l'économie.
* Libertariens : Le libertarianisme prône une société fondée sur la liberté d'expression des individus dans le respect du droit, du pluralisme et du libre-échange des idées. Le libertarisme échappe à un positionnement politique classique de par ses thèses qui le situent à la fois à gauche au plan des libertés individuelles : légalisation des stupéfiants, liberté d'expression, liberté de circulation, liberté sexuelle... Et à droite au plan des libertés économiques : respect de la propriété privée, libre-échange, suppression des impôts, diminution drastique de la fiscalité, justice réhabilitative...).

Mais critiquée :

* Accusé le libertarianisme d'être une liberté faussée en particulier par l'argent
* Une société qui fonctionnerait selon les principes libertariens américains s'autodétruirait en quelques secondes
* Concept de volatilité : est un indicateur de la dispersion de ses rendements par rapport à une moyenne sur une période donnée. Elle est mesurée à l’aide d’un indicateur statistique : l’écart-type. Plus les marchés sont instables plus la volatilité est importante. C’est un moyen de calculer le risque
* Valeur réelle : permet d’exprimer la valeur d’une variable en prix constants, c’est‑à‑dire en neutralisant les effets de l’inflation (augmentation du niveau général des prix), grâce à l’utilisation des prix d’une année de référence, appelée année de base. La valeur réelle se lit en monnaie constante et il faut ajouter l’année de base.

A ne pas confondre avec la valeur nominale qui est le prix en de l’année en court prenant en compte l’inflation, la devise etc…

* L’inflation : On parle d’inflation lorsque les prix augmentent globalement, et non uniquement les prix de quelques biens et services. Cette situation correspond à une baisse du pouvoir d’achat de la monnaie. En clair, avec la même somme d’argent, on peut acheter moins de choses qu’auparavant.
* Amortissement : selon le contexte, remboursement d’une partie ou de la totalité d’un crédit bancaire ou d’un emprunt, ou quote-part de la valeur d’un actif correspondant à sa décote (liée notamment à son obsolescence ou usure) par rapport à son coût d’acquisition.
* Actualisation : calcul qui permet de transformer un flux financier futur en une valeur équivalente présente. L’actualisation sert à mesurer la pertinence d’un investissement. L’inverse de l’actualisation est la capitalisation.
* Benchmark : indice de référence. Il permet d’effectuer des comparaisons, en mesurant la performance d’un acteur ou d’un titre.

Travail à faire pour les 2 semaines suivantes :

* Etablir la liste des parcs éoliens en France ainsi que leur puissance et le nombre de turbines installées
* Faire un tableau des mix énergétiques des 12 pays du CEE
* Lister tous les coûts associés au cycle de vie de l’énergie nucléaire (initial + maintenance/utilisation/gestion des déchets + démantèlement)
* Continuer les résumés des articles vus en mineur
* Continuer le tableau
* Rédaction de la synthèse
* Continuer l’exercice de machine learning