Enseigner l'IA frugale ?

Quelques réflexions de coin de table (qui n'engagent que leur auteur)

Sylvain Bouveret

JRAF'23, 13 décembre 2023

LIG, Ensimag-Grenoble INP, Ecoinfo

Position du problème

IA Frugale ?

Frugal

[En parlant d'une chose] Qui est empreint de simplicité, de sobriété

(source : *Trésor de la Langue Française Informatisé*)

IA Frugale ?

Frugal

[En parlant d'une chose] Qui est empreint de simplicité, de sobriété

(source : *Trésor de la Langue Française Informatisé*)

Sobriété

A. – Tempérance dans le boire et le manger. Synon. frugalité.

(source : *Trésor de la Langue Française Informatisé*)

IA Frugale ?

Frugal

[En parlant d'une chose] Qui est empreint de simplicité, de sobriété

(source : *Trésor de la Langue Française Informatisé*)

Sobriété

B. – Modération, mesure, discrétion. Anton. enflure, exagération, ostentation

(source : *Trésor de la Langue Française Informatisé*)

IA Frugale ?

Frugal

[En parlant d'une chose] Qui est empreint de simplicité, de sobriété

(source : *Trésor de la Langue Française Informatisé*)

Sobriété

B. – Modération, mesure, discrétion. Anton. enflure, exagération, ostentation

(source : *Trésor de la Langue Française Informatisé*)

Dans le cadre de l'IA, acception générale :

« Faire aussi bien (voire mieux) avec moins de ressources »

IA Frugale ?

Frugal

[En parlant d'une chose] Qui est empreint de simplicité, de sobriété

(source : *Trésor de la Langue Française Informatisé*)

Sobriété

B. – Modération, mesure, discrétion. Anton. enflure, exagération, ostentation

(source : *Trésor de la Langue Française Informatisé*)

Dans le cadre de l'IA, acception générale :

« Faire aussi bien (voire mieux) avec moins de ressources »

IA

???

Les deux dimensions de l'IA

Des systèmes qui pensent comme des humains

Des systèmes qui pensent rationnellement

Des systèmes qui agissent comme des humains

Des systèmes qui agissent rationnellement

Les deux dimensions de l'IA

Des systèmes qui pensent comme des humains	Des systèmes qui pensent rationnellement
Des systèmes qui agissent comme des humains	Des systèmes qui agissent rationnellement

Deux dimensions [Russell and Norvig, 2010]

- Qui est la référence ? Humain ou modèle idéal de rationalité ?
- À l'aune de quoi juge-t-on ? Pensée ou comportement ?


Russell, S. J. and Norvig, P. (2010).
Artificial intelligence : a modern approach.
Pearson Education, 3rd edition edition.

Les deux dimensions de l'IA

Des systèmes qui pensent comme des humains	Des systèmes qui pensent rationnellement
Des systèmes qui agissent comme des humains	Des systèmes qui agissent rationnellement

Deux dimensions [Russell and Norvig, 2010]

- Qui est la référence ? Humain ou modèle idéal de rationalité ?
- À l'aune de quoi juge-t-on ? Pensée ou comportement ?


Russell, S. J. and Norvig, P. (2010).
Artificial intelligence : a modern approach.
Pearson Education, 3rd edition edition.

Les deux dimensions de l'IA

IA forte

Des systèmes qui pensent comme des humains	Des systèmes qui pensent rationnellement
Des systèmes qui agissent comme des humains	Des systèmes qui agissent rationnellement

IA faible

Deux dimensions [Russell and Norvig, 2010]

- Qui est la référence ? Humain ou modèle idéal de rationalité ?
- À l'aune de quoi juge-t-on ? Pensée ou comportement ?


Russell, S. J. and Norvig, P. (2010).
Artificial intelligence : a modern approach.
Pearson Education, 3rd edition edition.

Une définition de l'IA

Une définition possible de l'IA [Alliot et al., 2002], d'après Minsky...

> L'intelligence artificielle a pour but de faire exécuter par l'ordinateur des tâches pour lesquelles l'homme, dans un contexte donné, est aujourd'hui meilleur que la machine.

Définition par nature floue et aux contours mouvants (chaque fois qu'un problème est résolu en IA, il sort par définition du domaine de l'IA...)


Alliot, J.-M., Schiex, T., Brisset, P., and Garcia, F. (2002). ***Intelligence Artificielle et Informatique Théorique.*** Cépaduès.

Remarque : l'IA selon l'UE

Définition d'un « Système d'intelligence artificielle » selon l'UE :

> [AI system] means software that is developed with one or more of the techniques and approaches listed in Annex I[1].

[1] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206

Remarque : l'IA selon l'UE

Définition d'un « Système d'intelligence artificielle » selon l'UE :

> [AI system] means software that is developed with one or more of the techniques and approaches listed in Annex I[1].

Annex I :

1. Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning;
2. Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems;
3. Statistical approaches, Bayesian estimation, search and optimization methods.

[1] *https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206*

Et l'apprentissage dans tout ça ?

L'apprentissage désigne la faculté à créer de la connaissance, permettant de résoudre des tâches pour lesquels l'algorithme n'a pas été explicitement programmé.

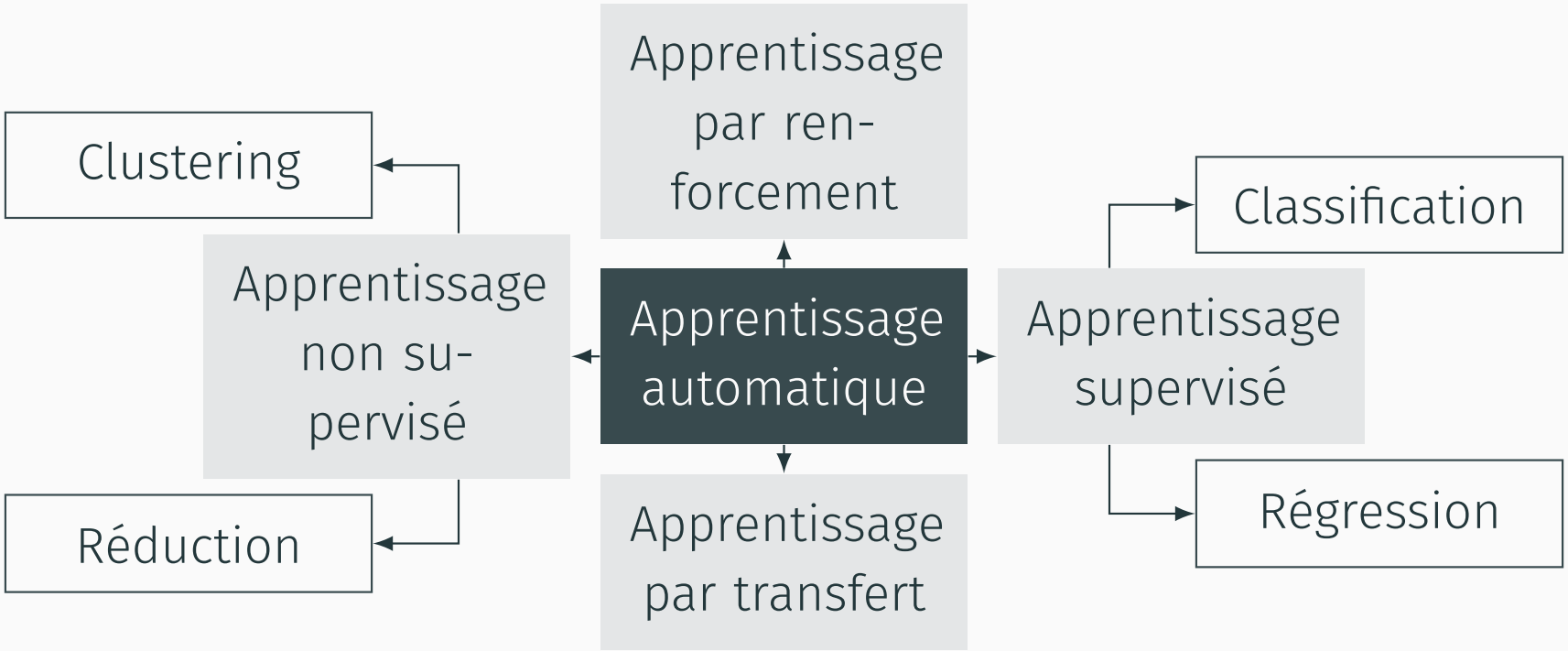
Et l'apprentissage dans tout ça ?

L'apprentissage désigne la faculté à créer de la connaissance, permettant de résoudre des tâches pour lesquels l'algorithme n'a pas été explicitement programmé.

Apprentissage automatique

Et l'apprentissage dans tout ça ?

L'apprentissage désigne la faculté à créer de la connaissance, permettant de résoudre des tâches pour lesquels l'algorithme n'a pas été explicitement programmé.



IA = ML ?

Pour certains scientifiques, l'apprentissage constitue l'**essence-même** de l'intelligence.

IA = ML ?

Pour certains scientifiques, l'apprentissage constitue l'**essence-même** de l'intelligence.

D'où l'équation souvent lue (ou entendue) :

IA = ML

IA = ML ?

Pour certains scientifiques, l'apprentissage constitue l'**essence-même** de l'intelligence.

D'où l'équation souvent lue (ou entendue) :

IA = ML

Et son équation jumelle plus restrictive :

IA = DL

Enseigner l'IA frugale, message #1

Message #1 :
Il faut sortir du paradigme IA = ML (et encore plus IA = DL).

Paradigme délétère à plus d'un titre :

- Appauvrissant pour la discipline
- Mène parfois à utiliser des approches inadaptées
- Problématique d'un point de vue environnemental

DL et environnement

- Qui dit apprentissage dit besoin de données d'entrées
- Difficile de sortir de la formule $\nearrow$ données $\Rightarrow \searrow$ erreur
- $\rightsquigarrow$ inflation dans la taille des jeux de données d'entraînement (et des modèles)

DL et environnement

- Qui dit apprentissage dit besoin de données d'entrées
- Difficile de sortir de la formule $\nearrow$ données $\Rightarrow \searrow$ erreur
- $\rightsquigarrow$ inflation dans la taille des jeux de données d'entraînement (et des modèles)
- Ex. pour GPT-3 :
  - $1,75 \times 10^{11}$ paramètres, 96 couches
  - 32000 mots en entrée
  - 12000 dimensions dans l'espace des mots
  - 1287 MWh (552 tCO_2e) pour l'entraînement [Patterson et al., 2021]

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., and Dean, J. (2021).
Carbon Emissions and Large Neural Network Training.
arXiv :2104.10350 [cs].

De la frugalité comme efficacité

Mesurer des impacts directs

Si nos hypothèses sont (i) IA = DL et (ii) frugalité = « faire aussi bien avec moins », alors il nous faut donc des moyens de **mesurer**.

Mesurer des impacts directs

Si nos hypothèses sont (i) IA = DL et (ii) frugalité = « faire aussi bien avec moins », alors il nous faut donc des moyens de **mesurer**.

NB : dans le cas de l'apprentissage, la spécificité est qu'il faut analyser deux phases :

- Phase d'apprentissage (hors-ligne en général)
- Phase d'inférence (en ligne)

Des outils de mesure

Pour mesurer, plusieurs solutions :

1. utilisation d'une grosse infrastructure de calcul (e.g Grid5000) déjà instrumentée ou utiliser les outils des fournisseurs de cloud privés

Des outils de mesure

Pour mesurer, plusieurs solutions :

1. utilisation d'une grosse infrastructure de calcul (e.g Grid5000) déjà instrumentée ou utiliser les outils des fournisseurs de cloud privés
2. wattmètres (accès à la machine physique nécessaire, ou instrumentation préalable d'un serveur local)

Des outils de mesure

Pour mesurer, plusieurs solutions :

1. utilisation d'une grosse infrastructure de calcul (e.g Grid5000) déjà instrumentée ou utiliser les outils des fournisseurs de cloud privés
2. wattmètres (accès à la machine physique nécessaire, ou instrumentation préalable d'un serveur local)
3. sondes logicielles (e.g compteurs RAPL) :

Des outils de mesure

Pour mesurer, plusieurs solutions :

1. utilisation d'une grosse infrastructure de calcul (e.g Grid5000) déjà instrumentée ou utiliser les outils des fournisseurs de cloud privés
2. wattmètres (accès à la machine physique nécessaire, ou instrumentation préalable d'un serveur local)
3. sondes logicielles (e.g compteurs RAPL) :
    - nombreux outils / bibliothèques disponibles : codecarbon, pyjoules, powerAPI, perf...

Des outils de mesure

Pour mesurer, plusieurs solutions :

1. utilisation d'une grosse infrastructure de calcul (e.g Grid5000) déjà instrumentée ou utiliser les outils des fournisseurs de cloud privés
2. wattmètres (accès à la machine physique nécessaire, ou instrumentation préalable d'un serveur local)
3. sondes logicielles (e.g compteurs RAPL) :
   - nombreux outils / bibliothèques disponibles : codecarbon, pyjoules, powerAPI, perf...
   - nécessite une architecture particulière (Intel $\geq$ SandyBridge)

Des outils de mesure

Pour mesurer, plusieurs solutions :

1. utilisation d'une grosse infrastructure de calcul (e.g Grid5000) déjà instrumentée ou utiliser les outils des fournisseurs de cloud privés
2. wattmètres (accès à la machine physique nécessaire, ou instrumentation préalable d'un serveur local)
3. sondes logicielles (e.g compteurs RAPL) :
   - nombreux outils / bibliothèques disponibles : codecarbon, pyjoules, powerAPI, perf...
   - nécessite une architecture particulière (Intel $\geq$ SandyBridge)
   - nécessite une autorisation ou un accès superutilisateur (Linux $\geq$ 5.4.77)

Des outils de mesure

Pour mesurer, plusieurs solutions :

1. utilisation d'une grosse infrastructure de calcul (e.g Grid5000) déjà instrumentée ou utiliser les outils des fournisseurs de cloud privés
2. wattmètres (accès à la machine physique nécessaire, ou instrumentation préalable d'un serveur local)
3. sondes logicielles (e.g compteurs RAPL) :
   - nombreux outils / bibliothèques disponibles : codecarbon, pyjoules, powerAPI, perf...
   - nécessite une architecture particulière (Intel $\geq$ SandyBridge)
   - nécessite une autorisation ou un accès superutilisateur (Linux $\geq$ 5.4.77)
   - ne donne qu'une vue partielle de la consommation (socket CPU + DRAM + éventuellement GPU intégrés)

Mesurer des impacts directs

Question : a-t-on réellement besoin de ces mesures de consommation énergétique ?

Mesurer des impacts directs

Question : a-t-on réellement besoin de ces mesures de consommation énergétique ?

Des proxies pour la consommation énergétique :

- Quantités physiques :
  - Temps CPU
  - Empreinte mémoire
  - Nombre de défauts de cache
  - …
- Mesures formelles de coût algorithmique :
  - Complexité temporelle
  - Complexité spatiale
  - …

Réduire l'impact direct

- Réduction de la taille des modèles
  - Nombre de paramètres
  - Quantification
- Réduction de la taille des jeux de données d'entraînement
- Réutilisation de modèles pré-entraînés (apprentissage par transfert)
- Utiliser des matériels dédiés
- ...

Réduire l'impact direct

- Réduction de la taille des modèles
  - Nombre de paramètres
  - Quantification
- Réduction de la taille des jeux de données d'entraînement
- Réutilisation de modèles pré-entraînés (apprentissage par transfert)
- Utiliser des matériels dédiés
- ...

De beaux problèmes mathématiques et informatiques à la clef !

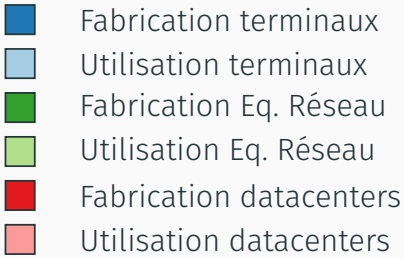
Impacts directs, les angles morts

La production des données

- La plupart de ces modèles ne fonctionnent bien que parce qu'ils sont entraînés sur des jeux de données de **grande taille**
- Comment ces données ont-elles été produites ?
  - Jeux de données *ad hoc*
  - Web (cf GPT)

La production du matériel

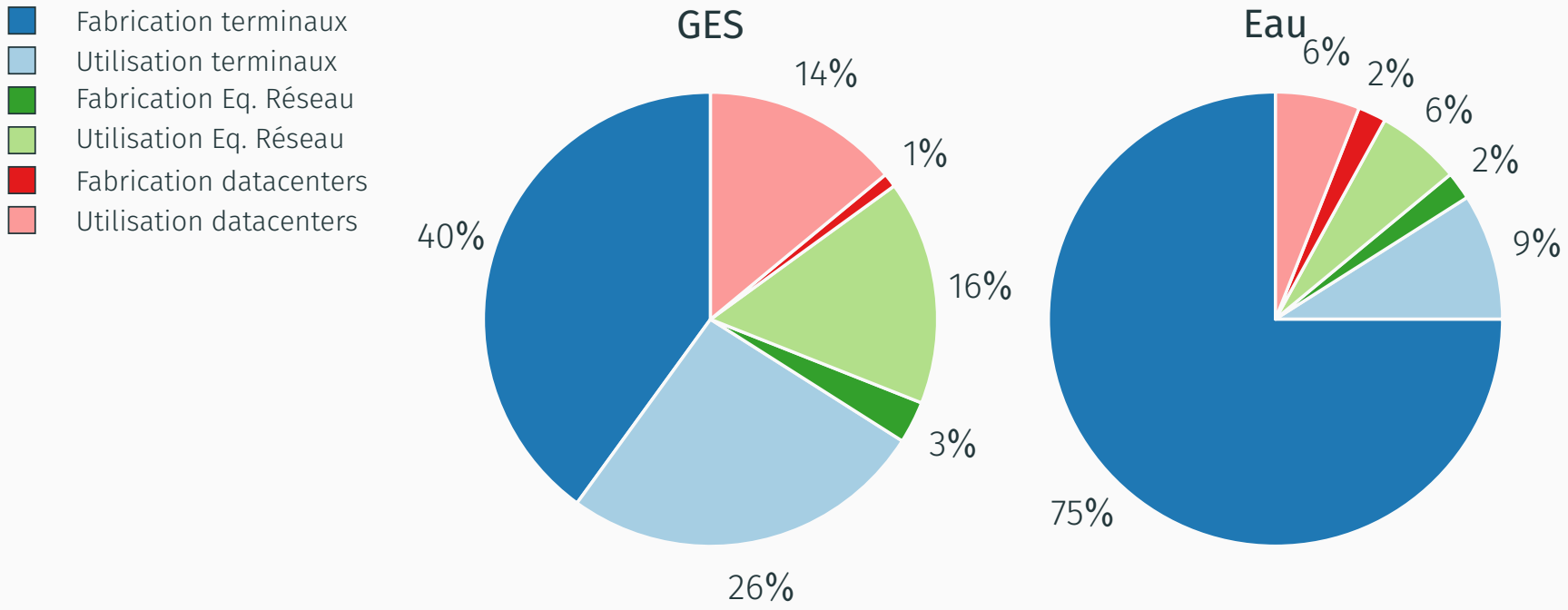
La figure classique [Bordage, 2019]



Bordage, F. (2019).
Empreinte environnementale du numérique mondial.
Technical report, GreenIT.fr.

La production du matériel

La figure classique [Bordage, 2019]



Bordage, F. (2019).
Empreinte environnementale du numérique mondial.
Technical report, GreenIT.fr.

Enseigner l'IA frugale, message #2

Message #2 :
Il faut sortir du paradigme impact environnemental = impact en phase d'usage (apprentissage + inférence)

Paradigme délétère à plus d'un titre :

- Oublie la phase de production des données, qui peut être massivement impactante
- Oublie la phase de production du matériel, qui dans le cas de l'informatique, est souvent la phase la plus impactante

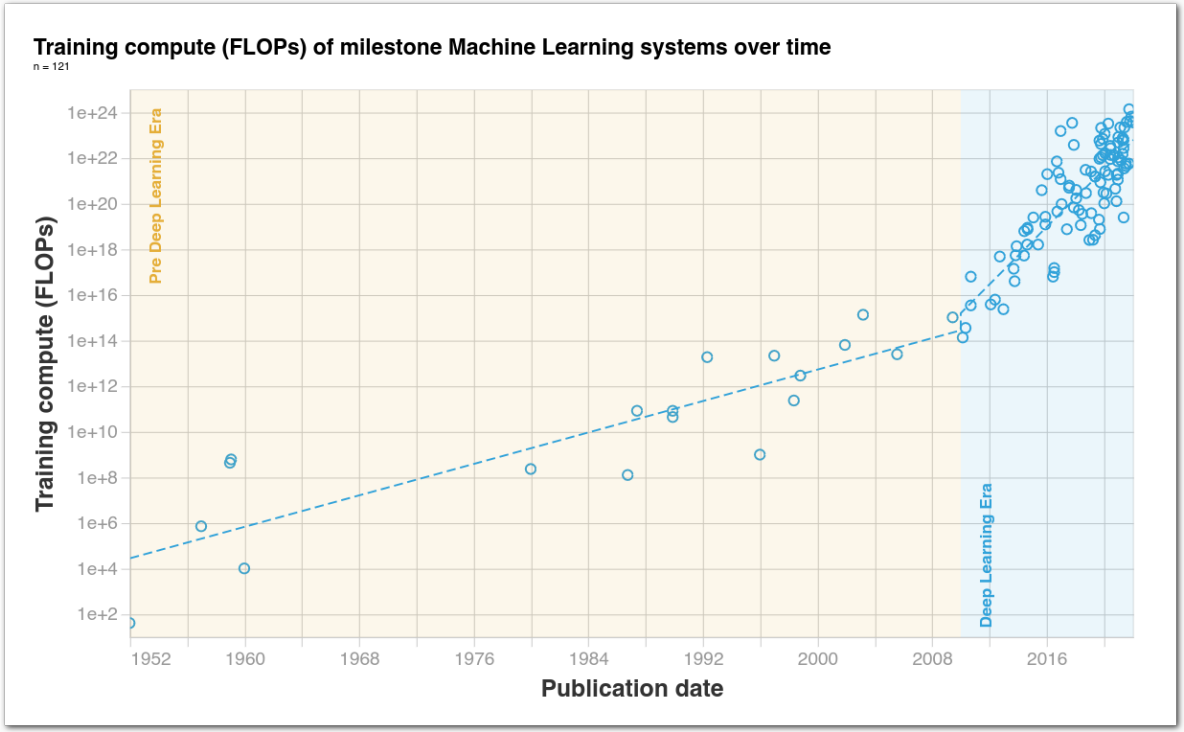
Pourquoi / doit-on enseigner l'IA frugale ?

Pourquoi enseigner l'IA frugale ?

Trois raisons possibles :

- De beaux problèmes mathématiques et informatiques
- IA embarquée
- **Réduction de l'empreinte environnementale de l'IA**

Tendances observées dans le ML



Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., and Villalobos, P. (2022).
Compute Trends Across Three Eras of Machine Learning.
arXiv :2202.05924 [cs].

Efficacité vs exponentielle

Un petit exercice de pensée :

- chaque FLOP consomme x joules
- à l'année N, le nombre moyen de FLOPS effectué par l'entraînement d'une machine est de C (donc consommation totale Cx joules)
- cette même année, une innovation de rupture permet de diviser par 10 la consommation unitaire d'une FLOP

Efficacité vs exponentielle

Un petit exercice de pensée :

- chaque FLOP consomme x joules
- à l'année N, le nombre moyen de FLOPS effectué par l'entraînement d'une machine est de C (donc consommation totale Cx joules)
- cette même année, une innovation de rupture permet de diviser par 10 la consommation unitaire d'une FLOP

Question : En combien de temps après l'année N aurai-je dépassé la consommation totale Cx joules si les tendances se poursuivent ?

Efficacité vs exponentielle

Un petit exercice de pensée :

- chaque FLOP consomme x joules
- à l'année N, le nombre moyen de FLOPS effectué par l'entraînement d'une machine est de C (donc consommation totale Cx joules)
- cette même année, une innovation de rupture permet de diviser par 10 la consommation unitaire d'une FLOP

Question : En combien de temps après l'année N aurai-je dépassé la consommation totale Cx joules si les tendances se poursuivent ?

Réponse : En 2 ans dans le meilleur des cas.

Efficacité vs exponentielle

Un petit exercice de pensée :

- chaque FLOP consomme x joules
- à l'année N, le nombre moyen de FLOPS effectué par l'entraînement d'une machine est de C (donc consommation totale Cx joules)
- cette même année, une innovation de rupture permet de diviser par 10 la consommation unitaire d'une FLOP

Question : En combien de temps après l'année N aurai-je dépassé la consommation totale Cx joules si les tendances se poursuivent ?

Réponse : En 2 ans dans le meilleur des cas.

Morale de l'histoire : l'augmentation d'efficacité ne fait que repousser des limites mais n'induit pas à moyen terme de diminution d'impact (et encore, on ne parle pas des effets systémiques)

Enseigner l'IA frugale, message #3

Message #3 :
Il faut sortir du paradigme frugalité = réduction d'impact environnemental.

Paradigme délétère à plus d'un titre :

- S'inscrit dans un imaginaire dans lequel il existe des solutions technologiques miraculeuses aux problèmes environnementaux
- Détourne de manière trompeuse les étudiants des vrais problèmes (et des vraies solutions) : faire vivre une informatique dans le cadre des limites planétaires

À quoi rêvent les étudiants (ingénieurs)?

Retour d'expérience sur un projet IA

- **Contexte** : Un projet IA avec sujet choisi par les étudiants eux-mêmes.
- **Quelques constats** :
  - 100 % des projets impliquent de l'apprentissage (dont environ 80-90 % du DL)

Retour d’expérience sur un projet IA

- **Contexte** : Un projet IA avec sujet choisi par les étudiants eux-mêmes.
- **Quelques constats** :
  - 100 % des projets impliquent de l’apprentissage (dont environ 80-90 % du DL)
  - Aucun n’est appliqué à des problématiques environnementales (ni même scientifiques au sens large)

Retour d’expérience sur un projet IA

- **Contexte** : Un projet IA avec sujet choisi par les étudiants eux-mêmes.
- **Quelques constats** :
  - 100 % des projets impliquent de l’apprentissage (dont environ 80-90 % du DL)
  - Aucun n’est appliqué à des problématiques environnementales (ni même scientifiques au sens large)
  - Aucune prise en compte sérieuse de l’efficacité des modèles : on fait du fonctionnel d’abord et on mesure à la fin parce qu’on nous oblige à le faire.

Et sur un projet d'innovation ?

- **Contexte** : Un projet d'innovation orienté explicitement sur les enjeux environnementaux et sociétaux
- **Quelques constats** :
  - Quelques projets farfelus (toujours techno-solutionnistes)

Et sur un projet d'innovation ?

- **Contexte** : Un projet d'innovation orienté explicitement sur les enjeux environnementaux et sociétaux
- **Quelques constats** :
  - Quelques projets farfelus (toujours techno-solutionnistes)
  - Des projets sérieux sur le plan efficacité...

Et sur un projet d'innovation ?

- **Contexte** : Un projet d'innovation orienté explicitement sur les enjeux environnementaux et sociétaux
- **Quelques constats** :
  - Quelques projets farfelus (toujours techno-solutionnistes)
  - Des projets sérieux sur le plan efficacité...
  - ...Mais qui ne remettent jamais en cause les hypothèses de départ (*e.g* « le monde a besoin de plus de datacenters »)

Et sur un projet d'innovation ?

- **Contexte** : Un projet d'innovation orienté explicitement sur les enjeux environnementaux et sociétaux
- **Quelques constats** :
  - Quelques projets farfelus (toujours techno-solutionnistes)
  - Des projets sérieux sur le plan efficacité...
  - ...Mais qui ne remettent jamais en cause les hypothèses de départ (*e.g* « le monde a besoin de plus de datacenters »)
  - Une difficulté réelle à discuter de ces questions (terrain politique)

Enseigner l'IA frugale, message #4

Message #4 :
Il faut faire évoluer les imaginaires des étudiants

Conclusion

Doit-on enseigner l'IA frugale ?

Le domaine est victime d'une confusion majeure :
frugalité = vertu environnementale

Doit-on enseigner l'IA frugale ?

Le domaine est victime d'une confusion majeure :
frugalité = vertu environnementale

Enseigner l'IA frugale ? Oui mais...

Doit-on enseigner l'IA frugale ?

Le domaine est victime d'une confusion majeure :
frugalité = vertu environnementale

Enseigner l'IA frugale ? Oui mais...

1. Si frugal = « faire aussi bien avec moins de calcul », alors renommer le cours en **IA efficace**[2]

[2] Et par pitié, arrêter les images de feuilles d'arbres qui poussent dans des GPU...

Doit-on enseigner l'IA frugale ?

Le domaine est victime d'une confusion majeure :
frugalité = vertu environnementale

Enseigner l'IA frugale ? Oui mais...

1. Si frugal = « faire aussi bien avec moins de calcul », alors renommer le cours en **IA efficace**[2]
2. Sinon, alors parler des angles morts (et remettre en cause les hypothèses implicites) :
   - autres domaines de l'IA (symbolique par exemple)
   - fabrication des machines
   - usage
   - modèle productiviste / extractiviste de la société

[2] Et par pitié, arrêter les images de feuilles d'arbres qui poussent dans des GPU...