Sustainable AI: Assessing the environnemental impact of AI

M, Guillaumont*, T. Cuvilliers*, P. Greullet*, O. Matz*, B. Deguilhem*

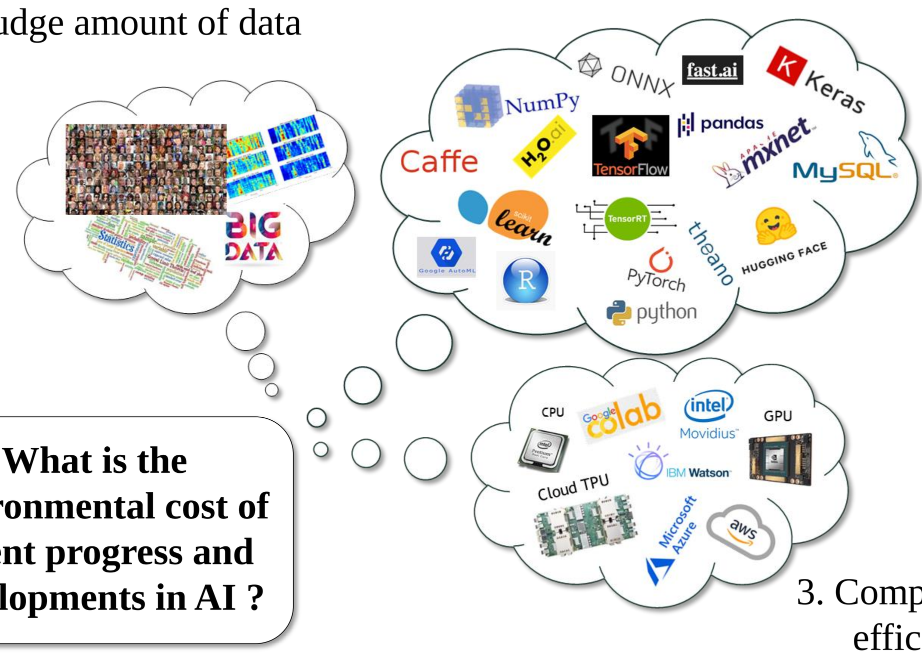
* Capgemini Engineering - DEMS, département R&I, Toulouse, France

General context overview, main motivation and use case

1. Hudge amount of data

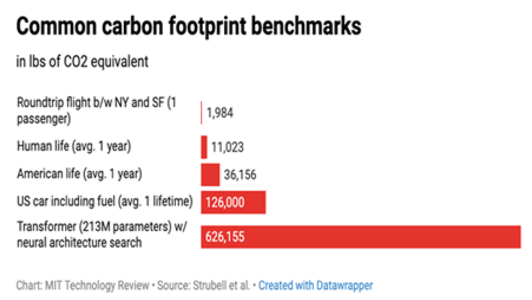
2. Many languages and framework

3. Computational efficiency



What is the environmental cost of recent progress and developments in AI ?

➢ **The carbon footprint and environmental impact of AI is not negligible:**



Training of Transformer algorithm (NAS) in 2019[1] has emitted more CO_2 than 5 cars in their entire life cycle (more than 300 000 kg eq CO_2).
Due to **increase** in the total number of optimized parameters (65M in 2017 vs 213M in 2019) and the total number of data used for training the CO_2 emission for this algorithm has been multiplicated **by 24 in two years**.

➢ **Recent developments in IA tend to worsen this impact:**

NLP algorithm	Parameters	Data (number of token)	Year
BERT [2]	213M	3 B	2019
GPT-2	1.5 B	40 B	2019
GPT-3 [3]	175B	500 B	2020

Wolff *et al.* has evaluated that the training of GPT-3 algorithm has emitted more than 80 000 kg eq. CO_2 based on average intensity carbon in USA in 2017 [3].

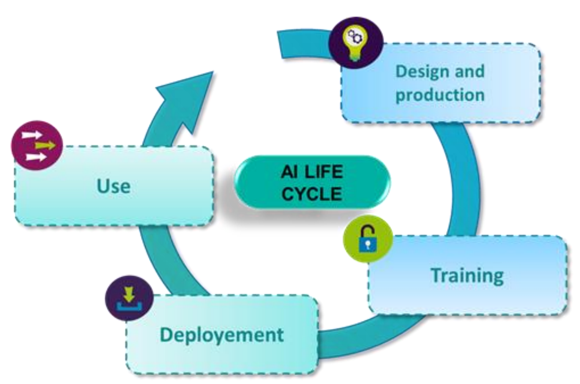
Our first objective: Recent studies report that training of NLP algorithms have a significant impact, what about this impact in the computer vision domain and for other processes than training ?
Project use case: Object detection in **computer vision** using convolutional neural network.

Evaluation of classification, segmentation and detection tasks. Deployment of IA on optimized embedded hardware.

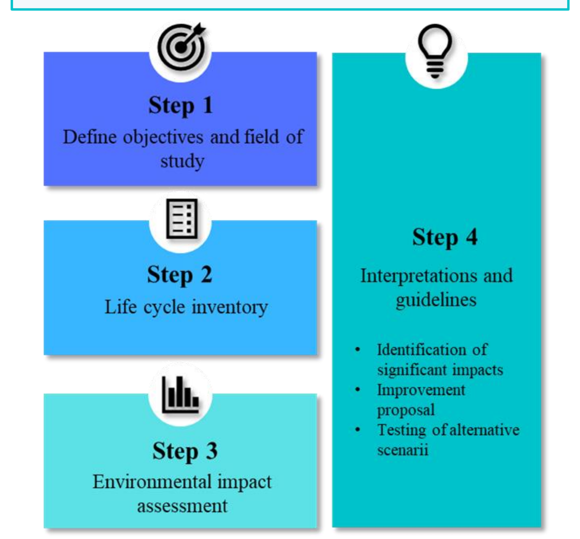
➢ **Yet environmental impact is not considered as valuable metric to evaluate performance of deep learning algorithms.**

Life Cycle Analysis

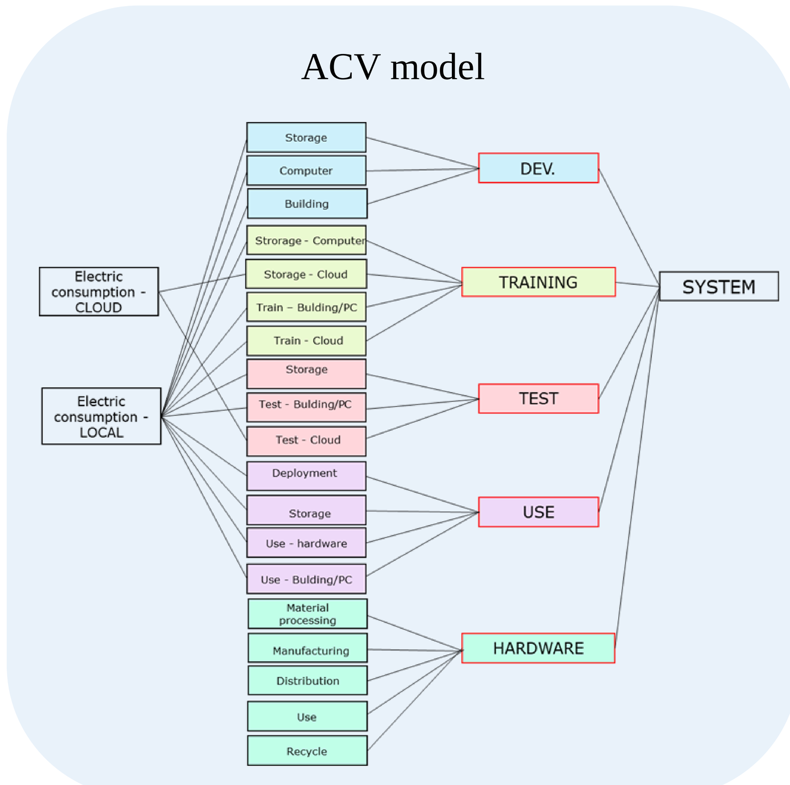
➢ Our goal is to adapt LCA methodology to the life of an AI taking in account Design and production, Training, Deployement and Use process.



ACV is a normalized methodology for **systemic quantificaton** of the environmental impacts of a product, a process or a service.

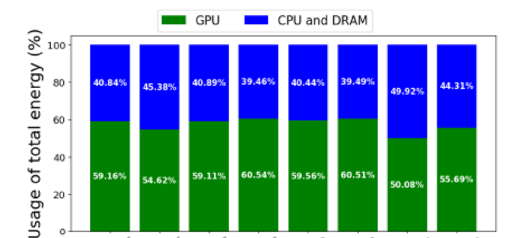


ACV model



Asses energy consumption during training

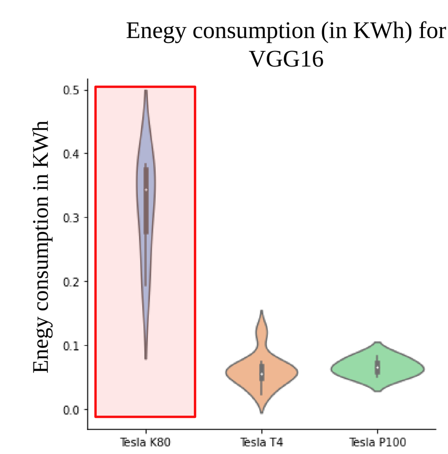
➢ On the training process, the nature of the **hardware**, algorithm **architecture** and the **localization** of the datacenter are depicted as key parameters to assess energy use and carbon footprint. CarbonTracker [4] and CodeCarbon libraries are used to monitor energy use during training.



GPU usage is accounting for more than a half of the total energy [4]

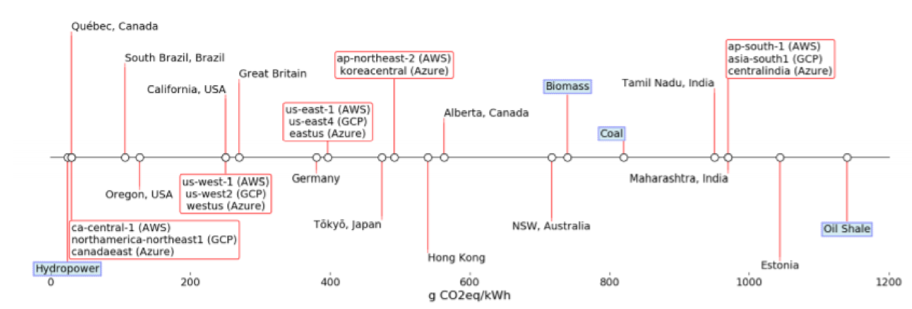
	Efficacité (GOPS/W)	Précision de calcul
T4	804,49	Int 8
P100	63,85	Float 16
K80	23,89	Float 32

Nature of GPU can lead to **a five times increase** in energy consumption.



Ratio CO_2 Emissions/Energy consomption ($kgCO_2/kWh$)
1. USA, Oregon (**0,139**)
2. USA, District of Columbia (**0,219**)
3. USA, South Carolina (**0,286**)
4. USA, Nevada (**0,349**)
5. USA, Iowa (**0,453**)
6. USA, Maryland (**0,459**)
7. Belgium, Brussel capital (**0,627**)
8. Netherlands, Groningen (**0,770**)
9. Taiwan, New Taipei (**0,790**)

Depending on the localization of the datacenter CO_2 emissions can be **reduced up to 90%**.



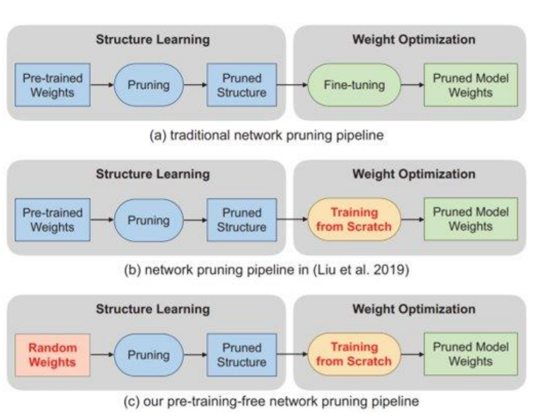
How to reduce the environnemental impact of AI ?

Several methods can be used to reduce environmental impact of IA:

1. During Training and inference: choose the best parameters among nature of **hardware** and **software**, hardware's **localization**, algorithms **architectures**…
2. During inference: Use of **model compression of model reduction methods**

➢ **Pruning**

This methods is efficient to develop smaller network by eliminating unnecessary values in the weight tensor

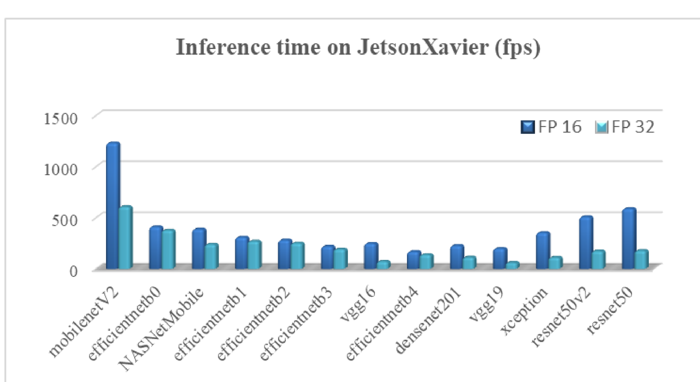


This figure is extract from Y. Wang *et al.*, « Pruning from Scratch », 2019 http://arxiv.org/abs/1909.12579

Winning Tickets method: Pruned subnetworks can reach test accuracy comparable to the original network in the same number of iterations [5].

➢ **Quantization**

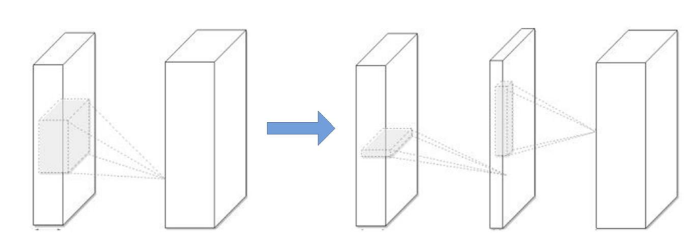
- This method is suitable for **convolutional** and **dense** layers. It consist in reducing the bit size of the model weights.
- Several levels of quantization: FP32, FP16, INT8.
- Hardware specific method: Nvidia Maxwell GPU supports FP16 but Nvidia Volta GPU supports INT8
- Integer quantization may require dataset calibration.



Going from FP32 to FP16 can increase the inference time (fps) up to 70%. The efficiency of quantization method highly depends on algorithm nature.

➢ **Low-rank Factorization [6]**

- This method is suitable for **convolutional** and **dense** layers. It consist in using tensor/matrix decomposition to estimate the informative parameters. (*e.g.* Truncated SVD)
- Cheng *et al.* report results achieved with no loss in accuracy [6]
- Compression: less parameters to store (2-5x less)
- Acceleration: architecture-dependant (1-2x faster)



On the left : original convolutional layer. The figure on the right show application of low-rank constraints to the convolutional layers with rank-K.

Next steps and bibliography

1. Generation of a **large dataset** in Computer Vision domain for Classification, Segmentation and Detection tasks : fist focus on **carbon footprint**
2. Run ACV models to evaluate the **global environmental impact** of IA with an ensemble of criteria
3. Complexification of **ACV models** by adding other processes than training and hardware life cycle
4. Reduction of environmental impact during training and inference using **model compression of model reduction methods** on specific hardware

1. E. Strubell *et al.*, « Energy and Policy Considerations for Deep Learning in NLP », **2019**, http://arxiv.org/abs/1906.02243
2. J. Devlin, *et al.*, « BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding», **2019**, https://arxiv.org/pdf/1810.04805.pdf
3. A. Brown *et al.*, « Language Models are Few-Shot Learners », **2020**, https://arxiv.org/abs/2005.14165
4. A. Wolff *et al.* « CarbonTracker : Tracking and Predicting the Carbon Footprint of Training Deep Learning Models », **2020**, https://arxiv.org/abs/2007.03051
5. J. Frankle and M. Carbin, « The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks », **2018**, http://arxiv.org/abs/1803.03635
6. Y. Cheng, *et al.*, « A Survey of Model Compression and Acceleration for Deep Neural Networks », **2020**, http://arxiv.org/abs/1710.09282