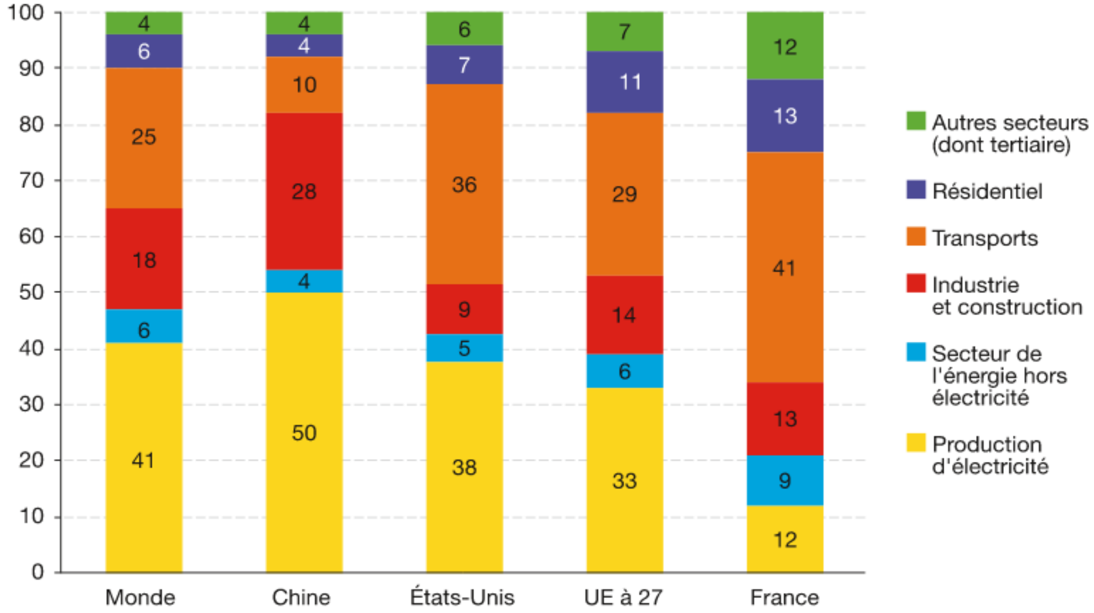
Pour un débat apaisé sur la consommation énergétique de l'IA

Conférence IA & éducation
Dock B – 8 juin 2023

Pierre Beyssac – Eriomem
@pbeyssac @pb@mast.eu.org

- Les centres de données
- Optimisations : Moore et algorithmes
- Libre / Open Source
- Sobriété / Temps gagné
- IA vs conventionnel



ORIGINE DES ÉMISSIONS DE CO_2 DUES À LA COMBUSTION D'ÉNERGIE EN 2018

En %

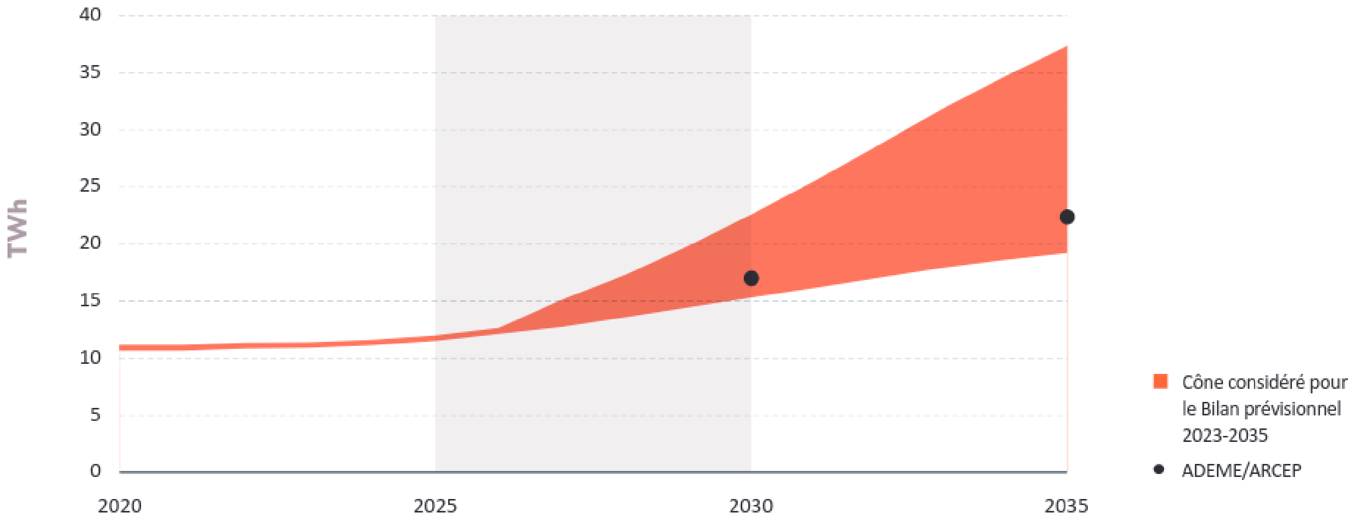
Source : *AIE, 2020*

https://www.statistiques.developpement-durable.gouv.fr/edition-numerique/chiffres-cles-du-climat/7-repartition-sectorielle-des-emissions-de

Rapport RTE 2023

11 Twh = Fessenheim = environ 2,2 % de la consommation fr

Figure 10 Trajectoires possibles de consommation électrique des data centers



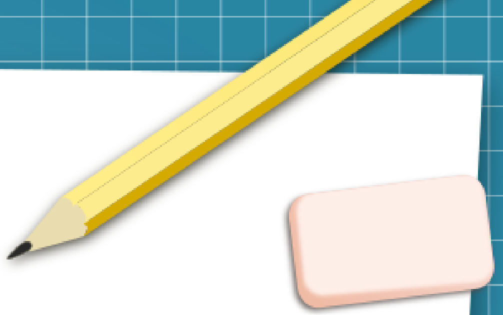
https://assets.rte-france.com/prod/public/2023-06/2023-06-07-bilan-previsionnel-points-etape.pdf

Les centres de données (DC)

Évaluation très difficile :

- De nombreux petits DC de PME sont « sous le radar »
  - peu efficaces (PUE) => cloud public
  - En décroissance mais généralement non comptés
- Les gros DC très visibles :
  - Privatifs type GAFAM (cloud public)
  - Colocation
  - En croissance
  - Forts gains de mutualisation, effets d’échelle etc

Ordres de grandeur



- Batterie de téléphone mobile ~15 Wh
- Ampoule basse consommation 7 W
- Voiture électrique 150 Wh/km
- Appareil à raclette 1500 W

1 recharge complète de téléphone mobile

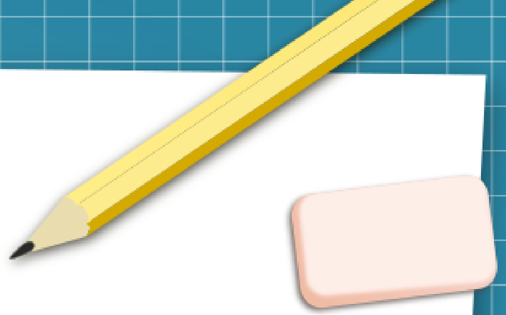
= 100 mètres en voiture

= 36 secondes de raclette

= 2 heures d’éclairage

TOTAL
KWh
0.00
MODE
UP

« oui mais l’effet rebond »



- Grosse consommation d’énergie

  => mauvais

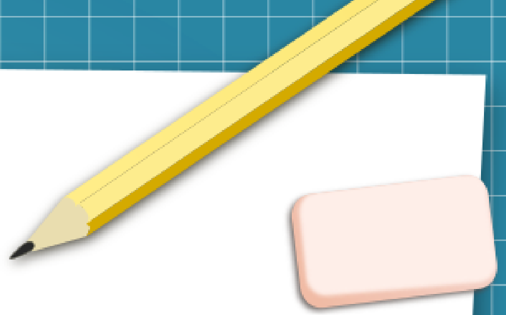
MAIS

- Optimisations

  => mauvais car provoque une augmentation de la demande (?)

Alors que fait-on ? On arrête tout ?

Sobriété



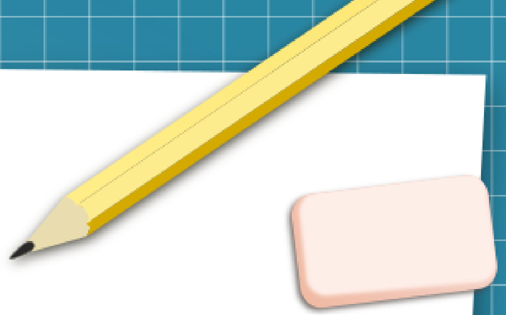
- Autonomie batterie
- Rapidité de réponse
- Ressources matérielles (CPU, mémoire, GPU...) terminaux & serveurs
- Énergie

  => coûts => facturation

  => un service ne va pas coûter plus que l’énergie+matériel

Critères qui vont freiner fortement l’utilisation d’IA « gourmandes »

Autres voies



- DC : valorisation chaleur fatale
- Preuve de travail ? Synergie cryptomonnaies <-> IA ?

Algorithmes

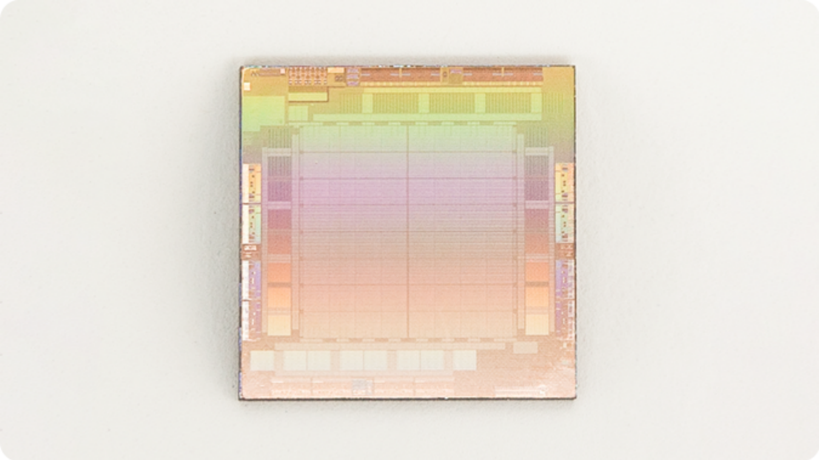
Progrès énormes en quelques mois :

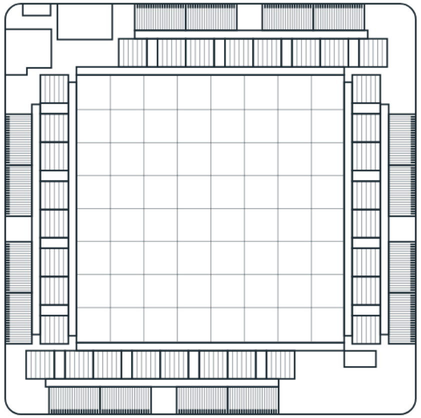
- Quantification
- Training
- Fine-tuning
- LoRA
- (cf demo, pas imaginable il y a seulement 6 mois)

Silicium pour les neurones

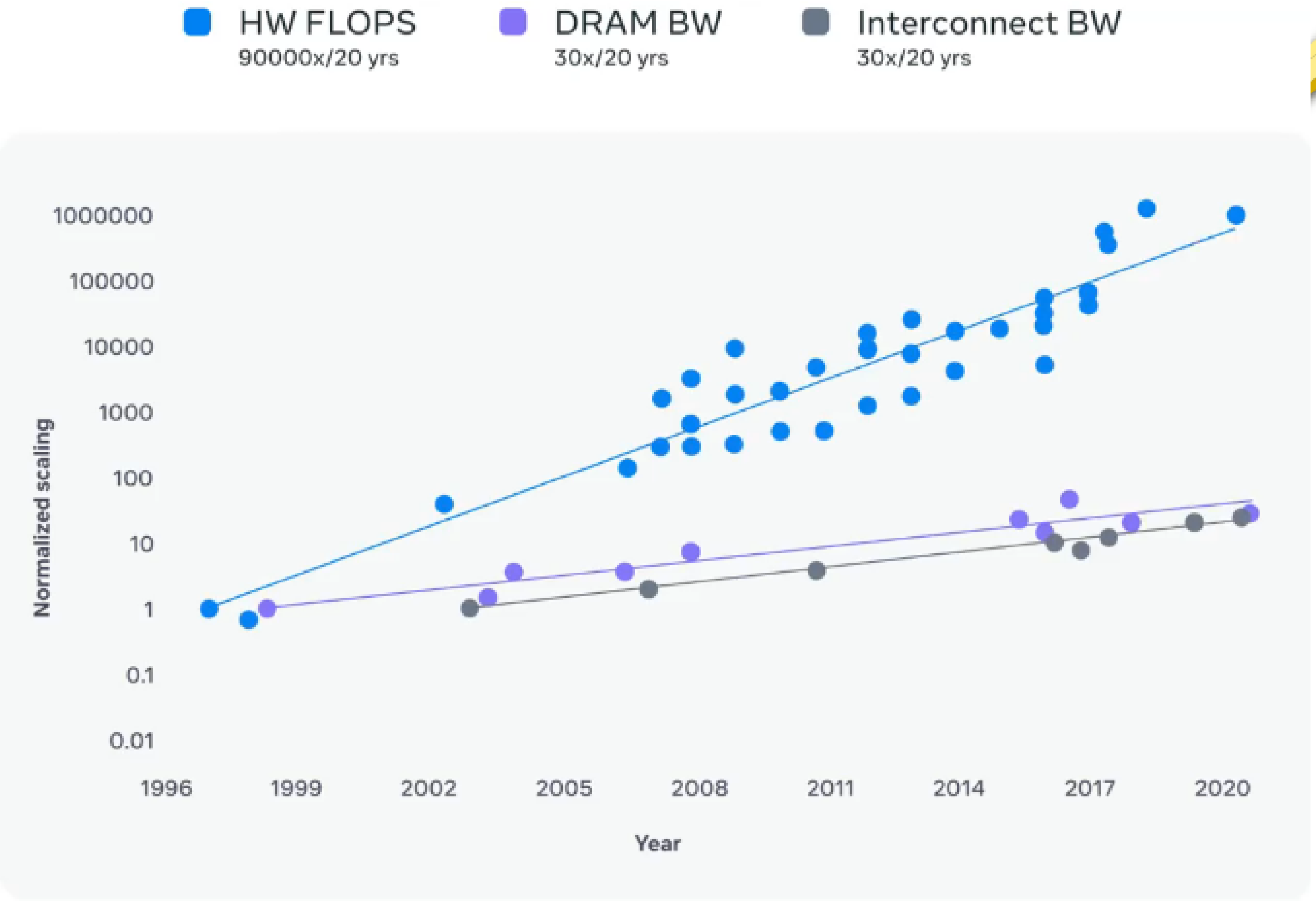
- Processeurs avec AVX, AVX2, AVX-512 (calcul vectoriel)
  - Ultra courant aujourd’hui
- FMA (fused multiply-add)
- GPU
- Circuits dédiés (ASIC) : MTIA (Facebook), ?? (Google)...
- Probable : apparition de circuits dédiés d’accélération dans les téléphones et ordinateurs

Meta Training and Inference Accelerator (MTIA)



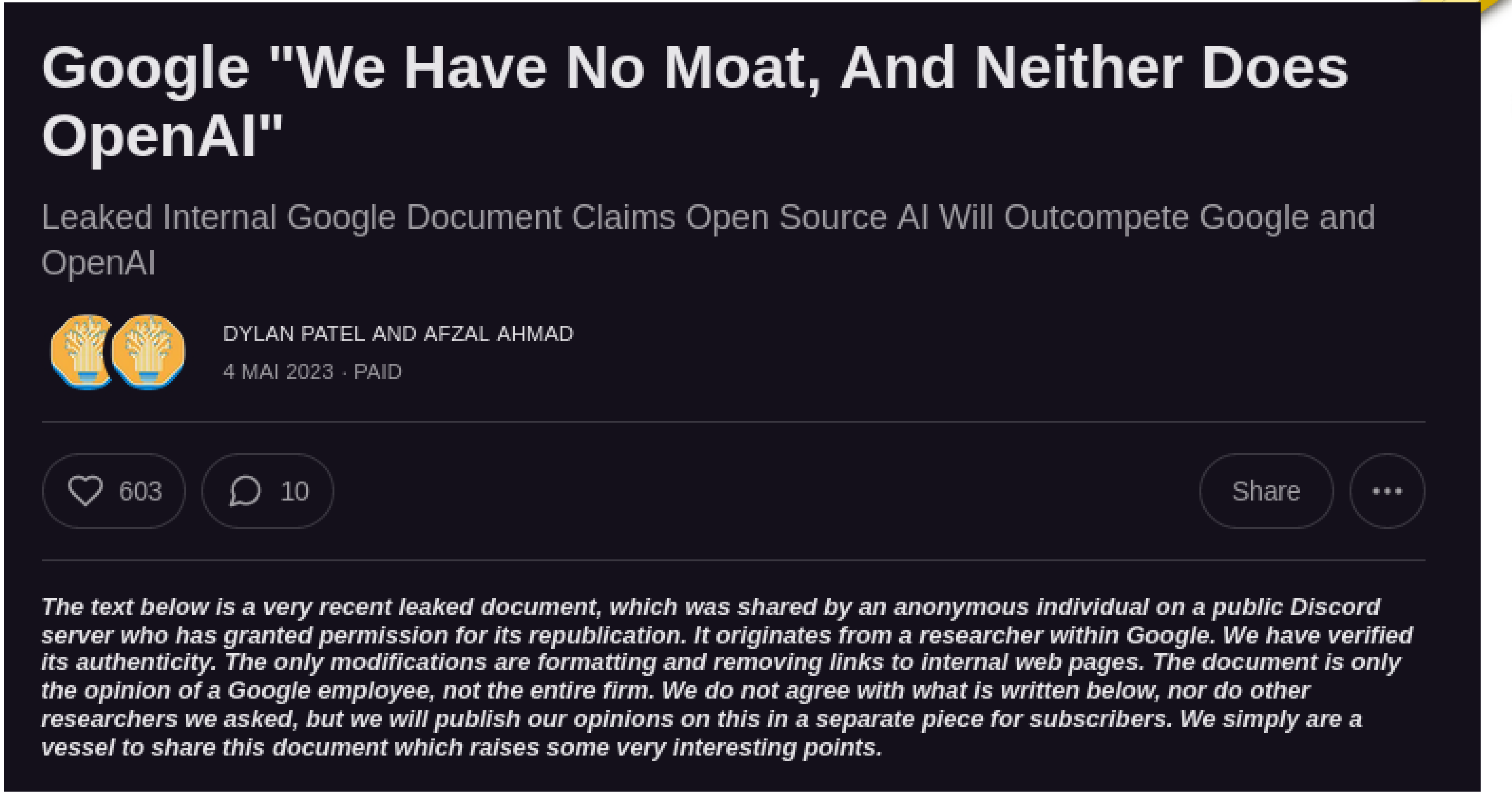


https://ai.facebook.com/blog/meta-training-inference-accelerator-AI-MTIA/


● HW FLOPS
90000x/20 yrs
● DRAM BW
30x/20 yrs
● Interconnect BW
30x/20 yrs
1000000
100000
10000
1000
100
10
1
0.1
0.01
Normalized scaling
1996
1999
2002
2005
2008
2011
2014
2017
2020
Year

Free / Libre / open source (FLOSS)

- Monde de la recherche
- Google => PyTorch
- OpenAI a changé son fusil d’épaule
- Nombreuses initiatives libres en réaction :
  - Meta => LlaMa (PyTorch)
    - Code libre mais pas le modèle
  - Nombreux dérivés de Llama
  - Nombreuses autres initiatives



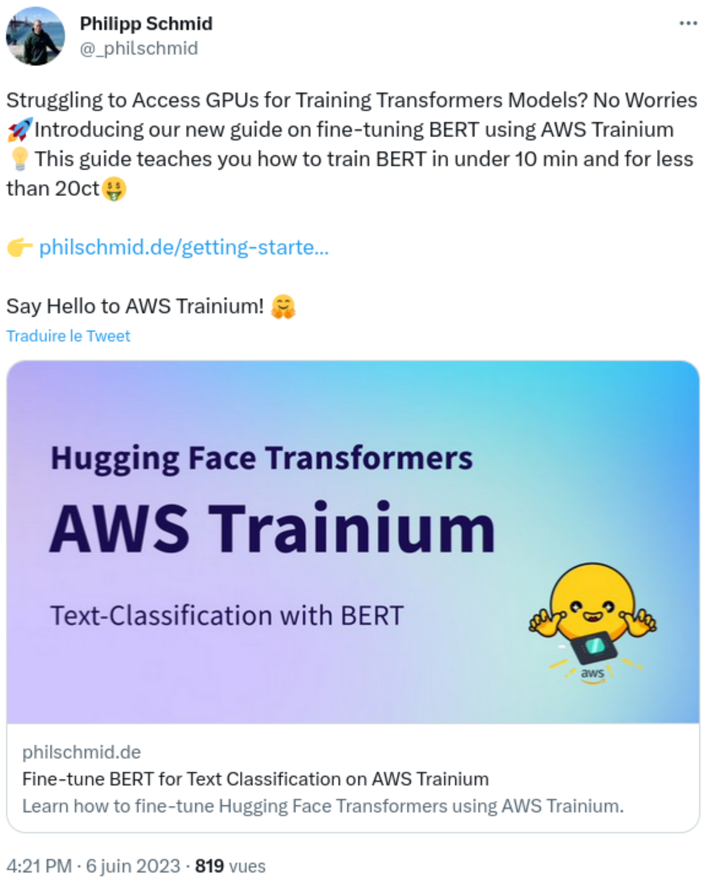
Google "We Have No Moat, And Neither Does OpenAI"

Leaked Internal Google Document Claims Open Source AI Will Outcompete Google and OpenAI

DYLAN PATEL AND AFZAL AHMAD
4 MAI 2023 · PAID

The text below is a very recent leaked document, which was shared by an anonymous individual on a public Discord server who has granted permission for its republication. It originates from a researcher within Google. We have verified its authenticity. The only modifications are formatting and removing links to internal web pages. The document is only the opinion of a Google employee, not the entire firm. We do not agree with what is written below, nor do other researchers we asked, but we will publish our opinions on this in a separate piece for subscribers. We simply are a vessel to share this document which raises some very interesting points.

https://www.semianalysis.com/p/google-we-have-no-moat-and-neither



Philipp Schmid
@_philschmid

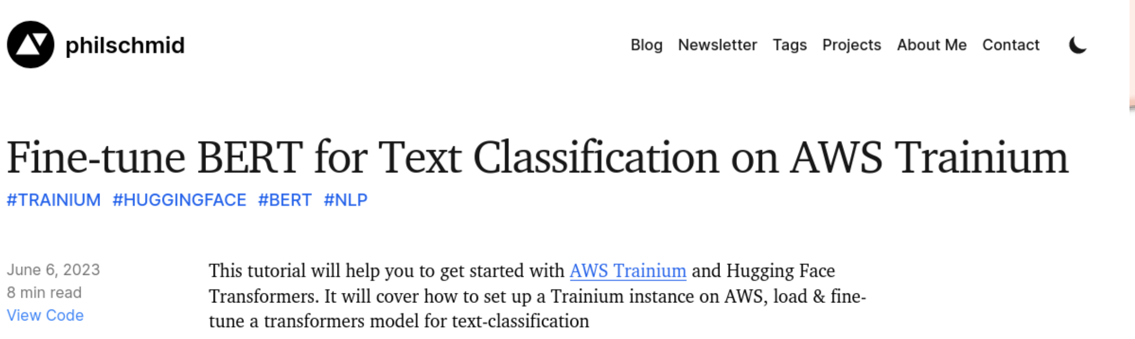
Struggling to Access GPUs for Training Transformers Models? No Worries
🚀Introducing our new guide on fine-tuning BERT using AWS Trainium
💡This guide teaches you how to train BERT in under 10 min and for less than 20ct🤑

👉 philschmid.de/getting-starte...

Say Hello to AWS Trainium! 🤗

philschmid.de
Fine-tune BERT for Text Classification on AWS Trainium
Learn how to fine-tune Hugging Face Transformers using AWS Trainium.

4:21 PM · 6 juin 2023 · **819** vues



Fine-tune BERT for Text Classification on AWS Trainium

#TRAINIUM #HUGGINGFACE #BERT #NLP

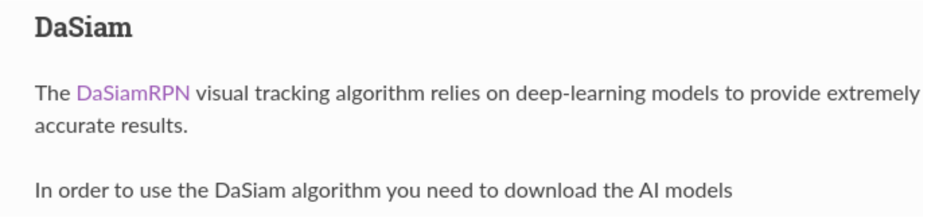
June 6, 2023
8 min read
View Code

This tutorial will help you to get started with AWS Trainium and Hugging Face Transformers. It will cover how to set up a Trainium instance on AWS, load & fine-tune a transformers model for text-classification

https://www.philschmid.de/getting-started-trainium

IA vs conventionnel

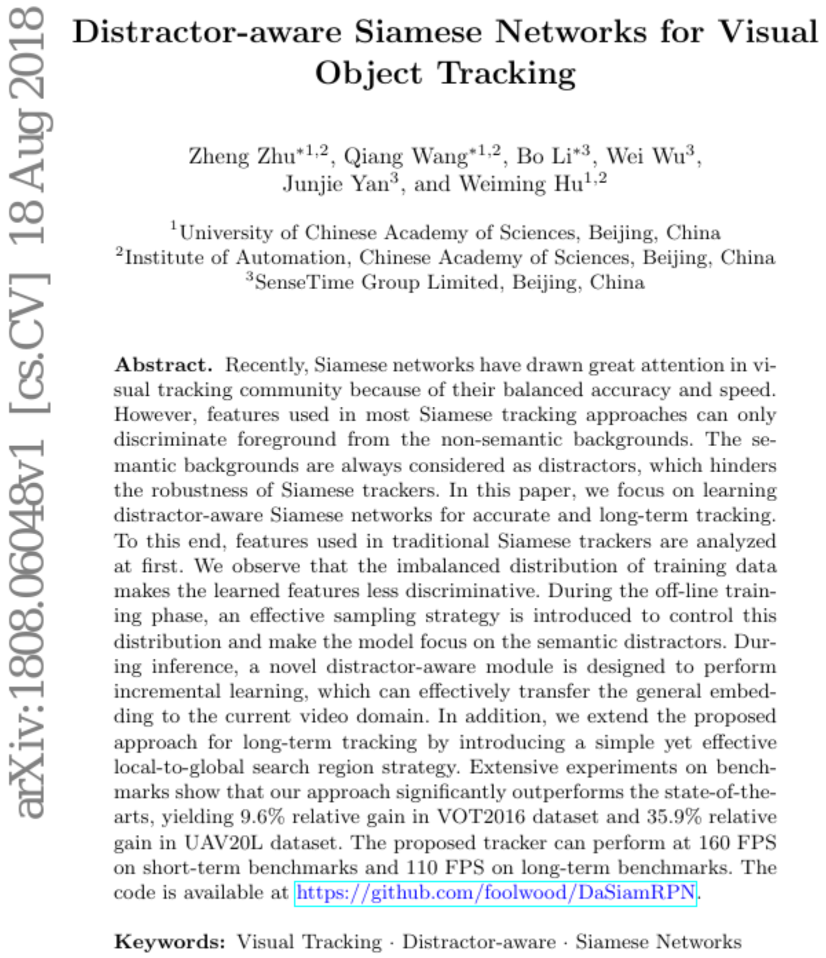
Utilisé dans kdenlive (montage vidéo)



DaSiam

The DaSiamRPN visual tracking algorithm relies on deep-learning models to provide extremely accurate results.

In order to use the DaSiam algorithm you need to download the AI models



arXiv:1808.06048v1 [cs.CV] 18 Aug 2018

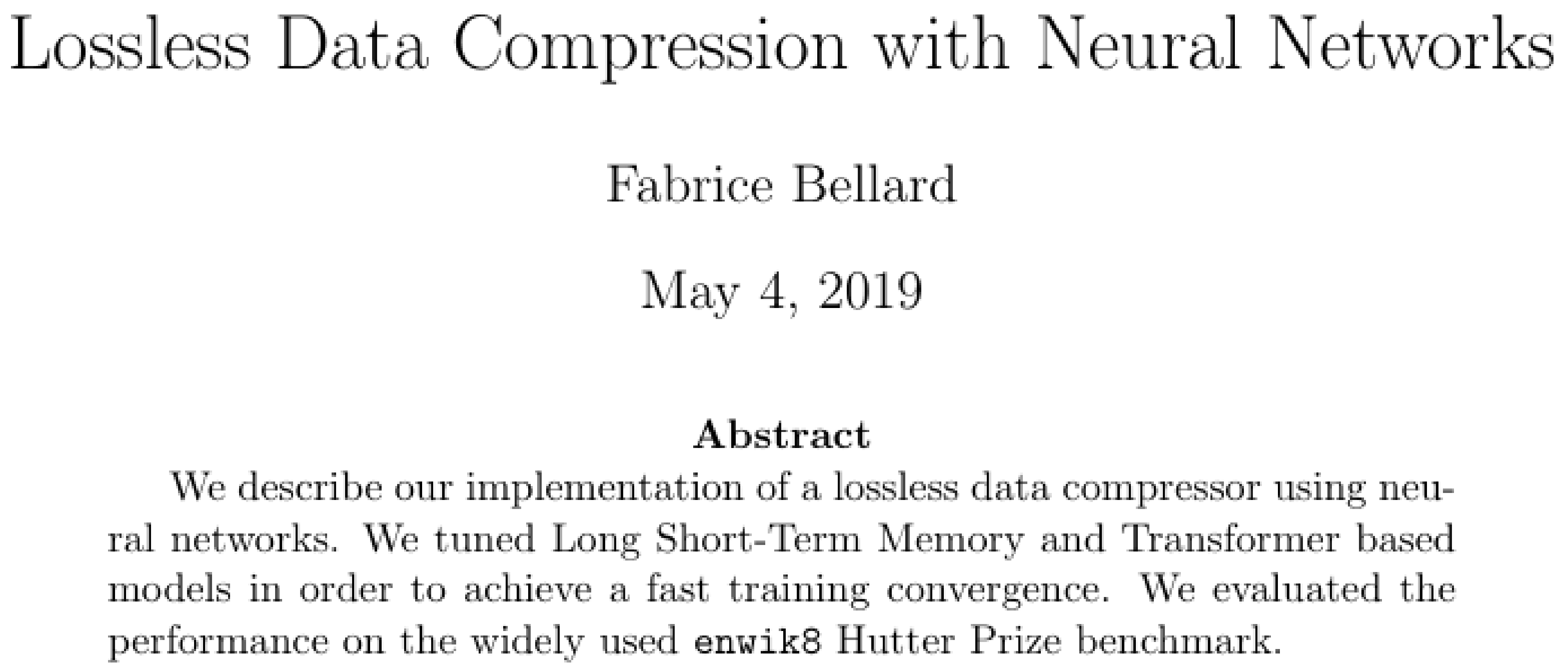
Distractor-aware Siamese Networks for Visual Object Tracking

Zheng Zhu*[1,2], Qiang Wang*[1,2], Bo Li*[3], Wei Wu[3],
Junjie Yan[3], and Weiming Hu[1,2]

[1]University of Chinese Academy of Sciences, Beijing, China
[2]Institute of Automation, Chinese Academy of Sciences, Beijing, China
[3]SenseTime Group Limited, Beijing, China

Abstract. Recently, Siamese networks have drawn great attention in visual tracking community because of their balanced accuracy and speed. However, features used in most Siamese tracking approaches can only discriminate foreground from the non-semantic backgrounds. The semantic backgrounds are always considered as distractors, which hinders the robustness of Siamese trackers. In this paper, we focus on learning distractor-aware Siamese networks for accurate and long-term tracking. To this end, features used in traditional Siamese trackers are analyzed at first. We observe that the imbalanced distribution of training data makes the learned features less discriminative. During the off-line training phase, an effective sampling strategy is introduced to control this distribution and make the model focus on the semantic distractors. During inference, a novel distractor-aware module is designed to perform incremental learning, which can effectively transfer the general embedding to the current video domain. In addition, we extend the proposed approach for long-term tracking by introducing a simple yet effective local-to-global search region strategy. Extensive experiments on benchmarks show that our approach significantly outperforms the state-of-the-arts, yielding 9.6% relative gain in VOT2016 dataset and 35.9% relative gain in UAV20L dataset. The proposed tracker can perform at 160 FPS on short-term benchmarks and 110 FPS on long-term benchmarks. The code is available at https://github.com/foolwood/DaSiamRPN.

Keywords: Visual Tracking · Distractor-aware · Siamese Networks



Lossless Data Compression with Neural Networks

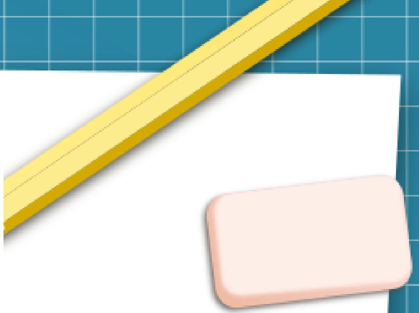
Fabrice Bellard

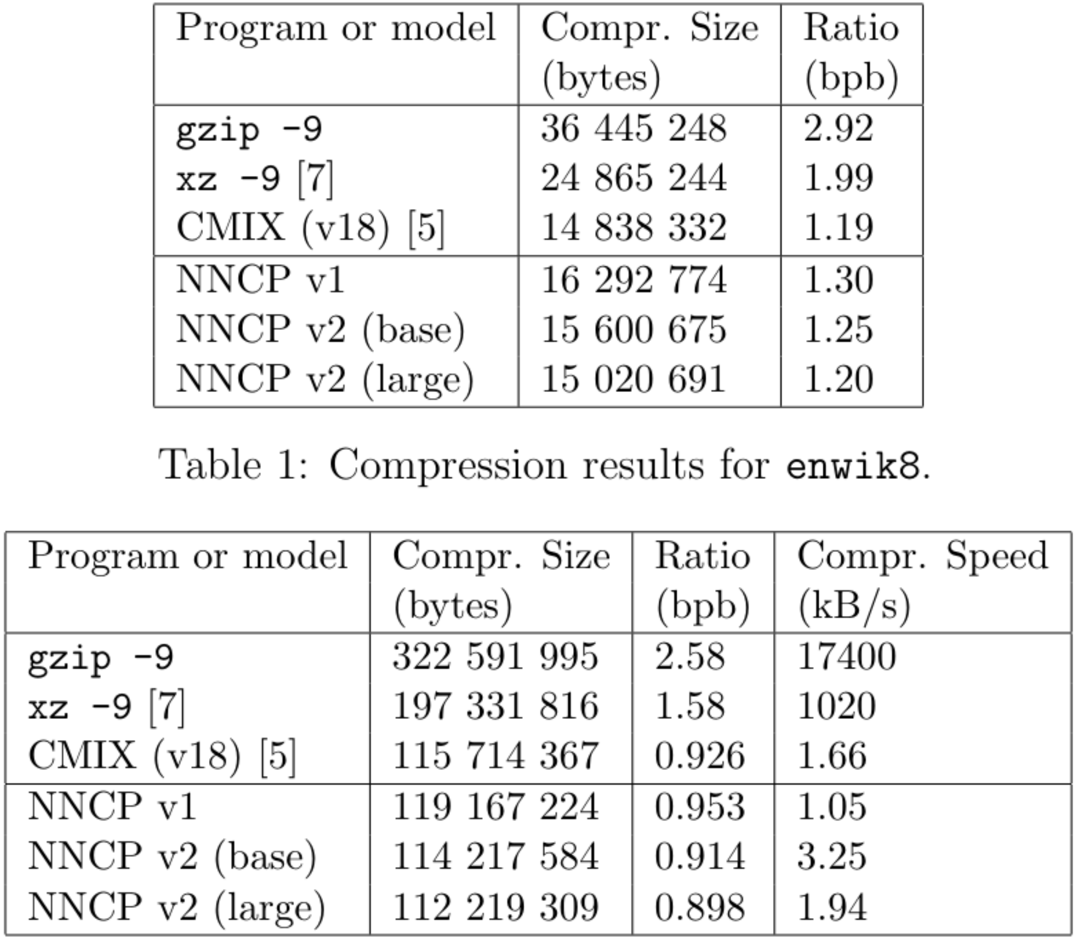
May 4, 2019

Abstract

We describe our implementation of a lossless data compressor using neural networks. We tuned Long Short-Term Memory and Transformer based models in order to achieve a fast training convergence. We evaluated the performance on the widely used `enwik8` Hutter Prize benchmark.

https://bellard.org/nncp/nncp.pdf



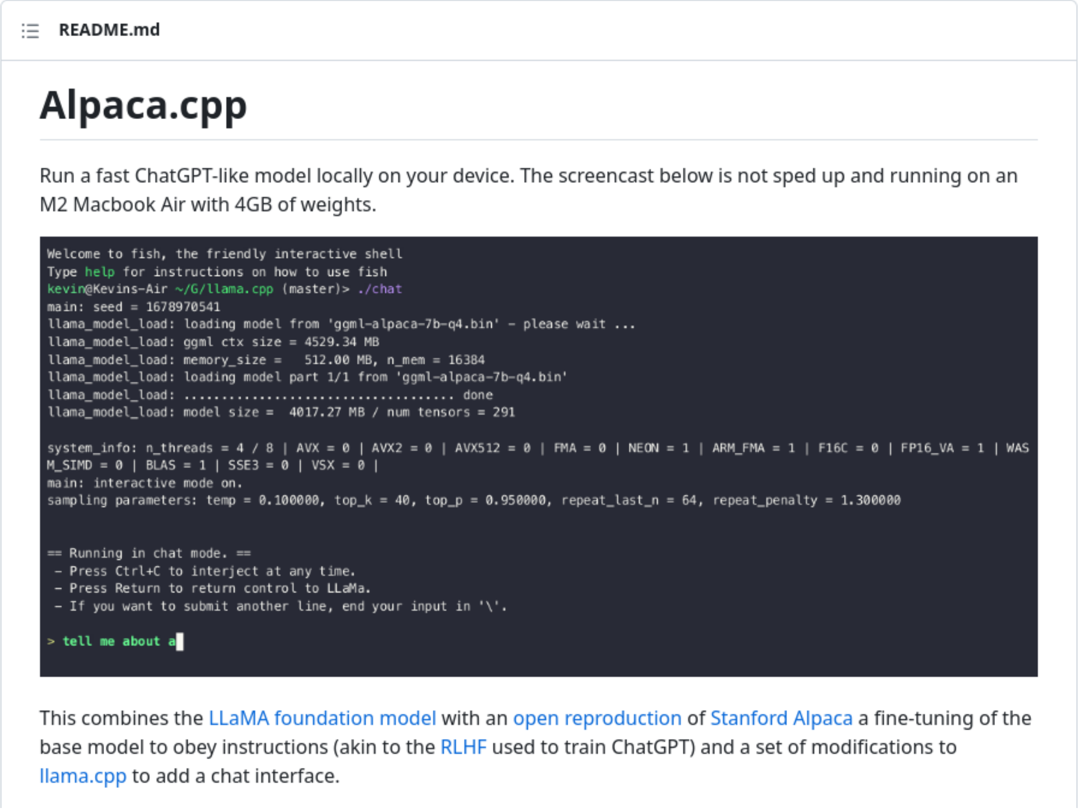


Program or model	Compr. Size (bytes)	Ratio (bpb)
`gzip -9`	36 445 248	2.92
`xz -9` [7]	24 865 244	1.99
CMIX (v18) [5]	14 838 332	1.19
NNCP v1	16 292 774	1.30
NNCP v2 (base)	15 600 675	1.25
NNCP v2 (large)	15 020 691	1.20

Table 1: Compression results for `enwik8`.

Program or model	Compr. Size (bytes)	Ratio (bpb)	Compr. Speed (kB/s)
`gzip -9`	322 591 995	2.58	17400
`xz -9` [7]	197 331 816	1.58	1020
CMIX (v18) [5]	115 714 367	0.926	1.66
NNCP v1	119 167 224	0.953	1.05
NNCP v2 (base)	114 217 584	0.914	3.25
NNCP v2 (large)	112 219 309	0.898	1.94

Démonstration avec alpaca.cpp



Alpaca.cpp

Run a fast ChatGPT-like model locally on your device. The screencast below is not sped up and running on an M2 Macbook Air with 4GB of weights.

```
Welcome to fish, the friendly interactive shell
Type help for instructions on how to use fish
kevin@Kevins-Air ~/G/llama.cpp (master)> ./chat
main: seed = 1678970541
llama_model_load: loading model from 'ggml-alpaca-7b-q4.bin' - please wait ...
llama_model_load: ggml ctx size = 4529.34 MB
llama_model_load: memory_size =   512.00 MB, n_mem = 16384
llama_model_load: loading model part 1/1 from 'ggml-alpaca-7b-q4.bin'
llama_model_load: .................................. done
llama_model_load: model size =  4017.27 MB / num tensors = 291

system_info: n_threads = 4 / 8 | AVX = 0 | AVX2 = 0 | AVX512 = 0 | FMA = 0 | NEON = 1 | ARM_FMA = 1 | F16C = 0 | FP16_VA = 1 | WAS
M_SIMD = 0 | BLAS = 1 | SSE3 = 0 | VSX = 0 |
main: interactive mode on.
sampling parameters: temp = 0.100000, top_k = 40, top_p = 0.950000, repeat_last_n = 64, repeat_penalty = 1.300000


== Running in chat mode. ==
 - Press Ctrl+C to interject at any time.
 - Press Return to return control to LLaMa.
 - If you want to submit another line, end your input in '\'.

> tell me about a
```

This combines the LLaMA foundation model with an open reproduction of Stanford Alpaca a fine-tuning of the base model to obey instructions (akin to the RLHF used to train ChatGPT) and a set of modifications to llama.cpp to add a chat interface.

https://github.com/antimatter15/alpaca.cpp

Merci !

Questions ?

Mastodon @pb@mast.eu.org

Twitter @pbeyssac

Merci !

Questions ?

Mastodon @pb@mast.eu.org

Twitter @pbeyssac