Towards resource aware AI development: AI simulation

Danilo Carastan dos Santos[1]

[2]Université Grenoble Alpes, Grenoble INP, Inria, LIG, France
email:danilo.carastan-dos-santos@inria.fr

May 12, 2022

Neural Network development has become too costly

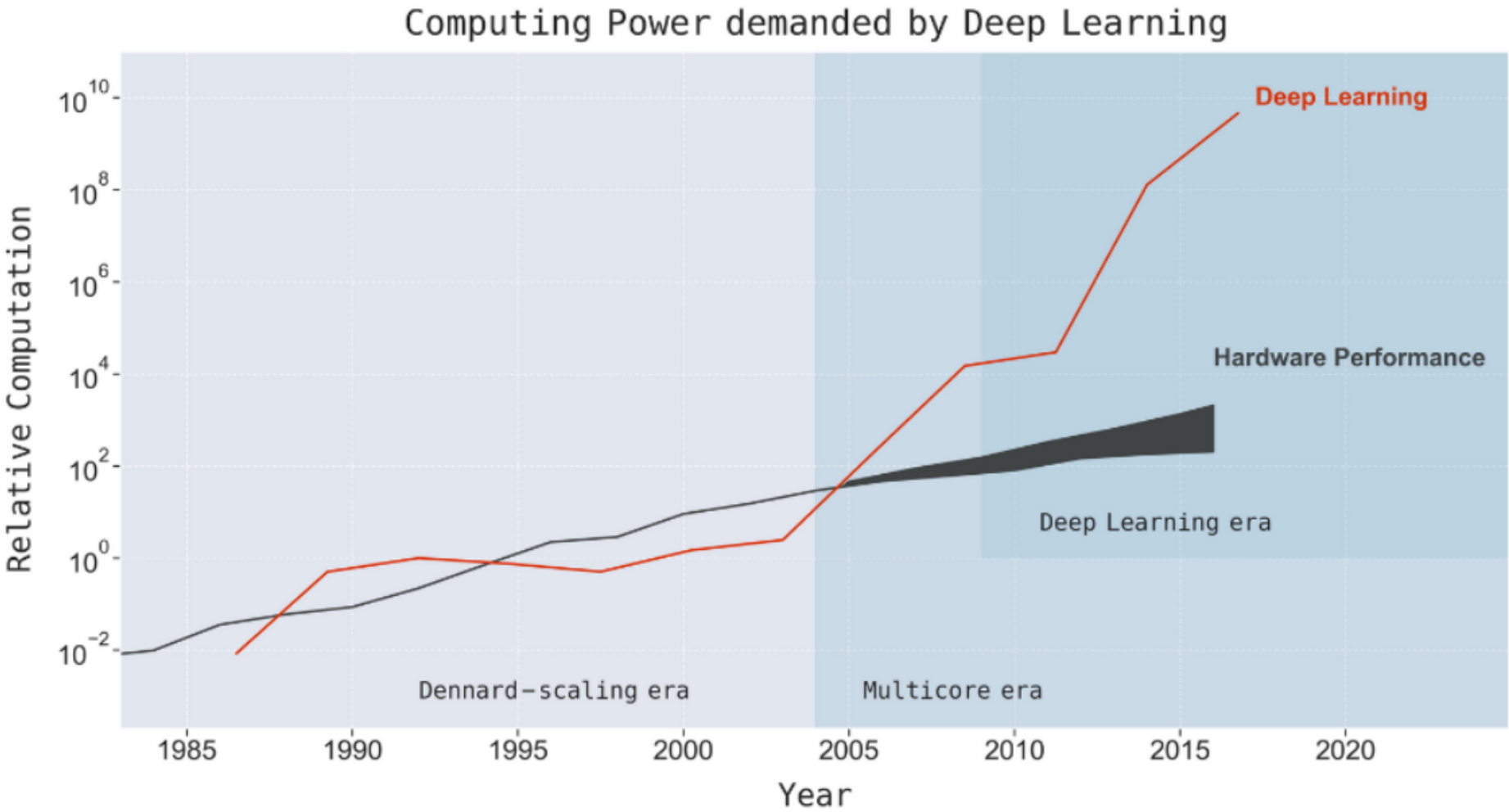
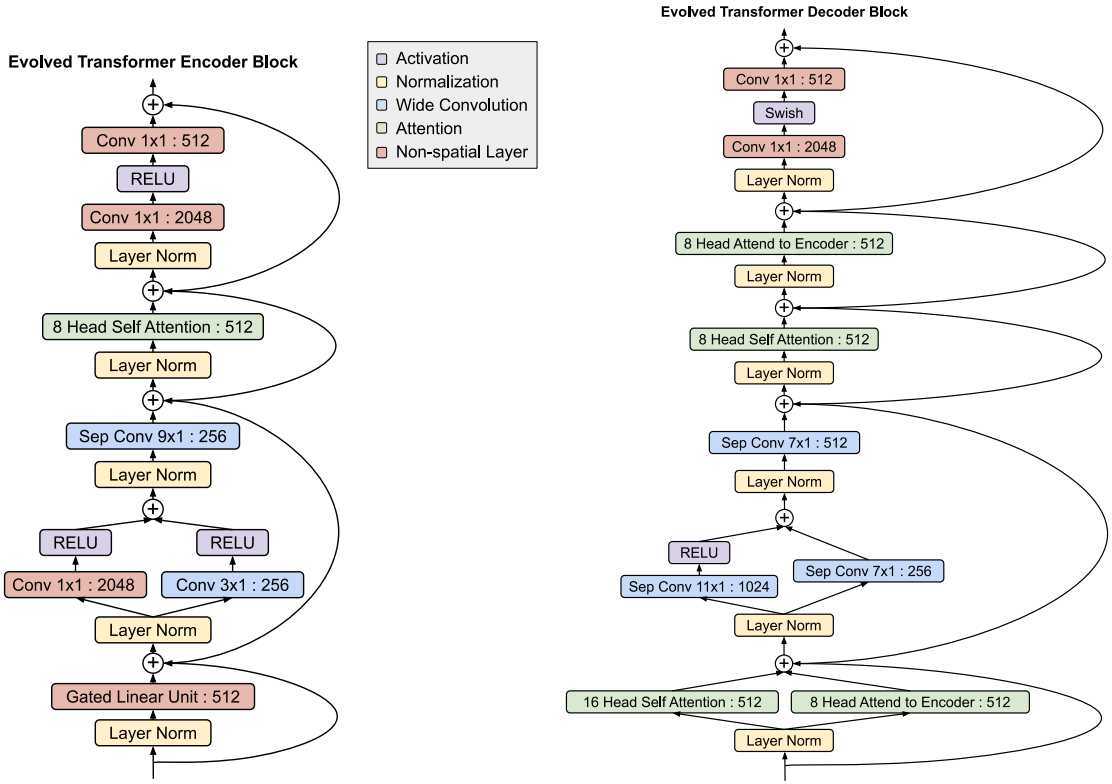


Figure source[1]

[1] Neil C Thompson et al. "The computational limits of deep learning". In: *arXiv preprint arXiv:2007.05558* (2020).

Example: Evolved Transformer[2] (NLP)



- Very large search space: $7.3 \cdot 10^{115}$ models
- Evolutionary algorithm to search for good models
- They discard "bad" models quickly, saving computation
- **All of this to make it feasible, but...**

[2] David So, Quoc Le, and Chen Liang. "The evolved transformer". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 5877–5886.

Example: Evolved Transformer (NLP)

It still too costly

Some performance information:

- They performed two experiments
  1. To validate the evolutionary algorithm search **(446M training steps)**
  2. To perform the "main" model search **(979M training steps)**
- **300k training steps** $\rightarrow$ **10 hours** with 1 Google TPU V.2 chip **(specialized hardware)**
  - **14866 TPU hours** for the first experiment, and **32633 TPU hours** for the second
  - **47499 TPU hours** in total

Putting in more understandable numbers...

Example: Evolved Transformer (NLP)

- TDP of a Google TPU v2 is 280 watts, and the PUE of the data center (`europe-west4`[3]) is 1.11[4]
- `europe-west4` costs 4.95 USD/TPU hour for on-demand (non preemptible) reservations[5]
- `europe-west4` server electricity is 60% carbon free, and it has 410gCO2/kWh of carbon intensity for the remaining 40%[6].
- **235000 USD paper**
- **2421 KgCO2 in emissions for the paper**
- **Four round trips Paris/NY in CO2 emissions**[7]

[3] `https://cloud.google.com/tpu/docs/types-zones#europe`

[4] David Patterson et al. "Carbon emissions and large neural network training". In: *arXiv preprint arXiv:2104.10350* (2021).

[5] `https://cloud.google.com/tpu`

[6] `https://cloud.google.com/sustainability/region-carbon`

[7] A round trip Paris New York is around 606 KgCO2 according to google flights.

Research Gaps

- **Gap 1: AI development is becoming computationally expensive or impossible**
  - The literature under explores resource efficiency in training
  - Ultra expensive methods are less accessible and lead to technology monopolization
    - We can share pretrained models and code (e.g., Facebook[8]), but do we have the means to change and retrain the models to suit our needs/desires?
- **Gap 2: Conducting reproducible experiments with AI methods is difficult**
  - For the same reasons as gap 1
  - 235 thousand dollars to reproduce the experiments of the Evolved transformer paper
    - Good luck if you want to make some variations

[8] https://ai.facebook.com/blog/democratizing-access-to-large-scale-language-models-with-opt-175b/

How can we help in this problem?

- **"Do better with less"**: Develop efficient AI models with less (or within a budget) resources (computing nodes, time, energy, etc.)
- **One step towards this goal:** Develop frameworks for *in silico* (simulation) experiments for AI research and development
- With simulation we can:
  - **Know in advance resource metrics** (e.g., execution time, data throughput, power consumption) that may guide decision-making, especially in neural architecture search
  - **We can search for AI models with less resources**, by simulating the training instead of actually training the models during architecture search

Challenges/Risks

What seems to be (easily) feasible

- **We can transfer HPC simulation for AI**
  - Everything is distributed computing in the end
  - We can instrument AI applications to extract and model the performance behavior

What seems to be challenging

- **How neural network architecture configurations correlate to resource usage?**
  - Can we easily model this correlation?
  - Is the configuration search space too large?

What may be very hard/impossible

- **How to estimate final model accuracy without training?**
  - Accurate estimates may be impossible to give
  - Extrapolation methods? Literature sourcing?

Thank you!

Towards resource aware AI development: AI simulation

Danilo Carastan dos Santos[1]

[2]Université Grenoble Alpes, Grenoble INP, Inria, LIG, France
email:danilo.carastan-dos-santos@inria.fr

May 12, 2022