Towards a Green AI

Evolutionary solutions for an ecologically-viable artificial intelligence

Nayat Sánchez-Pi and Luis Martí

Inria Chile Research Center, Santiago, Chile.
Instructors

Nayat Sánchez-Pi is currently the Director and CEO of the Inria Chile Research Center, created in 2012 by Inria, the French National Research Institute for Digital Sciences to facilitate scientific and industrial cooperation between France, Chile, and Latin America. Before that, she was a professor of Artificial Intelligence and Human-Computer Interaction at the Department of Informatics and Computer Science of the Institute of Mathematics and Statistics of the Rio de Janeiro State University. Prof. Sánchez-Pi’s research interests have broadened over the years and span topics that range from artificial intelligence, machine learning, and data mining to ambient intelligence, ubiquitous computing, and multi-agent systems. She received a degree in Computer Science in 2000 from the University of Havana and a Ph.D. degree in Computer Science in 2011 from the Universidad Carlos III de Madrid. She has led numerous research projects applying evolutionary computation, machine learning, and other artificial intelligence methods.

Luis Martí is currently the scientific director of Inria Chile, the Chilean Center of Inria, the French National Institute for Computational Sciences. Before that, he was a senior researcher of the TAU team at Inria Saclay since 2015. He was also an Adjunct Professor (tenured) at the Institute of Computing of the Universidade Federal Fluminense. Previous to that, Luis was a CNPq Young Talent of Science Fellow at the Applied Robotics and Intelligence Lab of the Department of Electrical Engineering of the Pontifícia Universidade Católica do Rio de Janeiro, Brazil. Luis did his Ph.D. at the Group of Applied Artificial Intelligence of the Department of Informatics of the Universidad Carlos III de Madrid, Madrid, Spain, and got his Computer Science degree from the University of Havana. He is mainly interested in artificial intelligence, and, in particular, machine learning, neural networks, evolutionary computation, optimization, machine learning, hybrid systems, and all that.

Inria in Chile

Since 2012

Inria
Computing (AI) for efficient green stuff

OPTIMIZATION
Plant and export layouts and robustness.

MODELING
Wind and wave modeling for solar, wind and tidal energies.

PREDICTION
Being able to predict the demand and production leads to efficient production.

Green energies and resilience

► Modeling and prediction of turbulent ocean flows.
  application: Power generation with tides.

► High resolution wind prediction
  application: Offshore wind generators.
  application: Impact of wind in solar generation.

► Ecological impact of industries.
► Understand catastrophes and their impact.
Ok, computing can help us reduce the environmental footprint...

...but, what about the impact of computing itself?

**BETTER MODELS**
Deep, reinforcement, transfer, active learning, etc.

**BETTER HARDWARE**
GPUs were developed for rendering graphics but are the best hardware for machine learning.

**BETTER DATASETS**
Important effort to create challenges datasets: vision, natural language, etc.
Progress in ML has come at the expense of exponential needs of computing power

Common carbon footprint benchmarks in lbs of CO2 equivalent

<table>
<thead>
<tr>
<th>Activity</th>
<th>CO2 Equivalent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roundtrip flight b/w NY and SF (1 passenger)</td>
<td>1,984</td>
</tr>
<tr>
<td>Human life (avg. 1 year)</td>
<td>11,023</td>
</tr>
<tr>
<td>American life (avg. 1 year)</td>
<td>36,156</td>
</tr>
<tr>
<td>US car including fuel (avg. 1 lifetime)</td>
<td>126,000</td>
</tr>
<tr>
<td>Transformer (213M parameters) w/ neural architecture search</td>
<td>626,155</td>
</tr>
</tbody>
</table>
The solution

Nature has provided a lot of inspiration for the AI/ML area:

↳ neural networks,
↳ learning theories,
↳ evolutionary computing, etc.

It is about time to return the favor!

HPC, AI, ML, simulation, modeling, etc.

Energy-conscious computing

- HPC, AI, ML, simulation, modeling, etc.
- Viable Eco World
- Computer Science
The solution: a multifaceted approach
Two distinctive scenarios that require different solutions

<table>
<thead>
<tr>
<th>Training</th>
<th>Inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>High demand of resources:</td>
<td>Current state of the art models (i.e. transformers) do not scale as-is.</td>
</tr>
<tr>
<td>- Big masses of data, and particularly,</td>
<td>- To take these results into edge/mobile computing and IoT scenarios:</td>
</tr>
<tr>
<td>- Power-hungry GPUs.</td>
<td>- i.e. phones, security cameras, etc.</td>
</tr>
<tr>
<td>Aggravated by the need of hyperparameter tuning.</td>
<td></td>
</tr>
<tr>
<td>- e.g. impossible to determine optimal network topology.</td>
<td></td>
</tr>
<tr>
<td>- Current “expensive” AutoML.</td>
<td></td>
</tr>
<tr>
<td>Limits access to technology to a few powerful companies.</td>
<td></td>
</tr>
<tr>
<td>- They are starting to monetize it directly: i.e. OpenAI’s GPT3.</td>
<td></td>
</tr>
<tr>
<td>- The good news: we do it once, use many times.</td>
<td></td>
</tr>
</tbody>
</table>

Better Hardware

It is unlikely that we get right of GPUs (or TPUs) at training time.

There are hardware alternatives at use time:

- Field Programmable Gate Arrays (FPGAs), Application-Specific Instruction-set Processors (ASIPs), etc.

We should also keep exploring the use of low-precision computing:

Reducing the quality (and therefore length) of the floating-point representation of numbers.
Everything starts with how we code, how we schedule, and organize resources.


HPC efficiency optimization and auto-tuning

Roofline model for evaluating AI applications

Observed characteristics DT:
- Balanced in memory and processing requirement
- The bottleneck is DRAM memory
- They have low arithmetic intensity and GFLOPs
Cloud computing for AI?

AI and ML pipelines are very computationally intensive but not continuously run:

**Training:** when we fit a model to the data available, very costly but run infrequently - or maybe once.

- Punctual use of expensive computing equipment.
- GPUs are both expensive to acquire and expensive to operate (high energy demands, cooling, etc.).

**Prediction:** using the model to make decisions, low computing requirements and used very frequently.

- High-availability low cost computing.

---

Not all cloud locations are the same

For example, different cloud locations and their CO₂ impact (indirectly, cost)

---

https://mlco2.github.io/
Self-scaling computing facilities make available a pool of shared resources.

Optimally schedule computing time.

Cloud computing allows to pick the location where programs will be run.

**Code is mobile!**

We can, for example, "track the sun" and ensure that the AI/ML processes use renewable sources.

"the choice of DNN, datacenter, and processor can reduce the carbon footprint up to ~100x - 1000x."

---

Smarter experimentation: Better AutoML

**Finding the right configuration of the hyperparameters** probably where more energy is consumed.

- This is an optimization problem => NP-Hard problem
- Neural architecture search, AutoML, AutoDL, etc.
- **Go beyond 'regular' hyperparameter search**
- However, better approaches like **evolutionary computing** is here to help!
  
  ...but they need populations of individuals, hence more energy.

  **This is a multi-objective optimization problem!**
High performance comes at high cost

Measuring AI’s footprint: proxies

- Carbon emissions generated.
- Electric energy consumed.
- Floating point operations.
- Elapsed time used for execution.
- Number of parameters (weights) and hyperparameters.

MLCommons Power Working Group:
Create power measurement techniques for various MLPerf benchmarks that enable reporting and comparing energy consumption, performance and power of benchmarks run on the submission systems.

https://mlcommons.org/en/groups/best-practices-power/
Self adaptation

To look for methods that adapt their complexity automatically to the complexity of the problem being solved.

Neural networks based on adaptive resonance theory (ART) and growing neural gas (GNG) have rules to adapt themselves to the complexity of the problem.

This self-adaptation is best profited when using cloud-based infrastructure.

Model reuse and transfer learning
Active learning and sample efficiency

The combination of mitigation factors

...(c/s)hould be posed as an optimization problem
Nature has provided a lot of inspiration for the AI/ML area:

- neural networks,
- learning theories,
- evolutionary computing,
- etc.

It is about time that we return the favor!

How the ocean helps to mitigate climate change?  How to protect this mitigation effect?

https://oceania.inria.cl/
Thank you! Obrigados! Merci! - ¡Gracias!

Questions?

Find more in greenai.inria.cl