Intelligenza Artificiale: dalla ricerca fondazionale all’impatto su aziende e PA
Rome, Oct 20, 2023

Edge and Exascale Artificial Intelligence
Barbara Caputo
Politecnico di Torino
Mainstream AI is inefficient

Energy

“Training GPT-3 would cost over $4.6M. [...] With the increase in parameters, there’s an exponential increase in energy.” Lambda Labs

Hardware

“The computational demands of AI present an emerging problem for its implementation on different hardware platforms.”

Data

“Data labeling takes anywhere from 35 to 80% of project budgets.”

Environment

“Training a single AI model can emit as much carbon as five cars in their lifetimes.”

---

1Source: https://lambdalabs.com/blog/demystifying-gpt-3
2Source: https://www.nature.com/articles/s41928-018-0068-2
3Source: https://www.forbes.com/sites/cognitiveworld/2022/08/06/no-youre-not-alone-google-is-also-making-this-big-mistake-on-ai
Edge and Exascale AI

Tiny AI
- WP 7.1
- WP 7.3

- Higher energy efficiency
- Environment friendly
- Lower hardware requirement
- Faster training and inference

Parallel AI
- WP 7.2
- WP 7.3

Downstream tasks
- WP 7.4
- WP 7.5

Implication
- WP 7.6
Spotlight works
We study how to better build neural models

The Multiply and Max/Min (MAM) –based neuron

MAM-based neurons do not accumulate all the weighted inputs, but sum together only the maximum and minimum contributes.

Neural networks built with MAM learn to use only a small subset of interconnections during inference.

By using all the available memory on device, MAM achieves 33 dB reconstruction performance (where a standard DNN would achieve 18 dB)

STM32F767ZI
μC unit (512 kB RAM, 216 MHz Clock)

Case study: MAM for ECG autoencoder tail

Prono, Luciano; Bich, Philippe; Mangia, Mauro; Pareschi, Fabio; Rovatti, Riccardo; Setti, Gianluca [2023]. A Multiply-And-Max/min Neuron Paradigm for Aggressively Prunable Deep Neural Networks. TechRxiv. Preprint.

https://doi.org/10.36227/techrxiv.22561567.v1
We assess the reliability of existing architectures

Robustness w.r.t. external disturbance (e.g. neutron strikes)

Robustness w.r.t. permanent hardware faults through Evolutionary-based approach to generate ITLs

Robustness w.r.t. information representation (e.g. POSIT)

And study **how and what models are learning**

Miss-classification (in **red**) rate of **speech models** (intent classif.) also depend on subgroups – how to detect low-accuracy subgroups without supervision?

**Subgroup gain** as the difference in performance of two models $M_1, M_2$ on a specific subgroup $S$ for $f$

$$gain_f(S, M_1, M_2) = f(S, M_2) - f(S, M_1)$$

**Divergence** as a measure of anomalous behavior of a data subgroup $S$ w.r.t. overall dataset $D$ for a function $f$

$$\Delta(S) = f(S) - f(D)$$

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>Sup acc</th>
<th>Acc</th>
<th>Drop acc</th>
<th>Gain acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>age=20-35, gender=males, is-naive</td>
<td>0.81</td>
<td>74.78</td>
<td>1.18</td>
<td>4.7</td>
</tr>
<tr>
<td>is-naive, is-native</td>
<td>0.83</td>
<td>74.81</td>
<td>1.56</td>
<td>4.9</td>
</tr>
</tbody>
</table>

**Distribution of the gain in performance**

- Drop for **31.9%** of the subgroups
- Benefit for **63.7%** of the subgroups

**Gain in performance**

- **-20.97%**
- **+22.69%**
From edge to very large-scale applications

DATA

Probabilistic modeling/AI
prob. model: easy ~ 10 min. simple desktop
AI -> computationally heavy (gpus)

Achieved accuracy ~ 90 %

Phage Display / Selex

Deep mutational scanning

Reertoire sequencing data

Selected B-cells
Sequence reads
Antibody repertoire

In silico generation of new sequences

Model

Interest
Pharma [cancer therapies, monoclonal antibodies]
Rational drug design
Design of monoclonal Abs with specific target ****
Bioengineering
Molecular motors
Industrial production of artificially engineered bacteria

Total addressable market
~ 13 billion USD

Total addressable market
~ 6 billion USD

Fernandez de Cossio Diaz, Uguzzoni, Pagnani, "Unsupervised Inference of Protein Fitness Landscape from Deep Mutational Scanning Experiments" Molecular Biology and Evolution 38:1 318-328 (2021)
We care about applications:
Efficient 2D and 3D computer vision on edge and very large scale
We care about applications:
Physics-informed NNs for quadrotor dynamical modeling and control

Learn the full dynamics of a quadrotor enhancing trustworthiness and interpretability through Physics-informed neural networks.
We care about applications: Understanding human behavior from egocentric data

- Mobile models for first person action recognition
- Challenges:
  - Untrimmed videos
  - Robust to Multi-modal domain gap (time, space)
  - Model footprint for edge deployment

3rd place EPIC KITCHENS UDA challenge at CVPR2022 & CVPR2023
Not only “supervised”: RL for (soft) intelligent manipulators

(A) Policy robustness

BayesSim [1]

RF-DROPO (ours)

✔ Distance to target ~3mm

✗ Distance to target ~30mm

(B) Training efficiency (up to 8x faster)

(T) Task Push the object to a target location (green dot)

(A) Task Walk forward

(C) Env. Constraints Exploitation

(T) Task Lift up the object

✗ Wall exploitation

✔ Wall exploitation

The agent is not aware of the random location of the wall

AI and its societal impact

Predictive maintenance on district heating networks

Semantic segmentation of cultural heritage point clouds

Ethically-sensitive dataset labeling

Balancing the complexity and interpretability of AI-based energy management strategies
Spoke 7 in Numbers

- 30+ professors
- 8 Assistant Professors enrolled
- 4 PhD students enrolled
- 6 research assistants (all on board)
- 9+ journal papers with peer review
- 40+ conference papers with peer review
Dissemination activity

IROS 2023, Detroit, USA
ICIAP 2023, Udine, Italy
ECML-PKDD 2023, Turin, Italy
IEEE TPTR, Anaheim, USA
ECML-PKDD 2023, Turin, Italy
AUTOMATICA 2023, Catania, Italy
Spoke 7 in the world
Not only rocket science

• We do have **extensive experience** in academia-industry collaborations
• Many already existing collaborations
• Keen to use FAIR as flywheel to enhance current relationships and open new opportunities
Thanks

• Happy to chat in the networking session

• Come and visit us in Turin!

   giuseppe.averta@polito.it
   barbara.caputo@polito.it

This study was carried out within the FAIR - Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013). This manuscript reflects only the authors’ views and opinions, neither the European Union nor the European Commission can be considered responsible for them.