Spoke 7 - Edge and Exascale AI

Giuseppe Averta, Barbara Caputo (PI)
Politecnico di Torino

Rome, Oct 20, 2023
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.”  
Source: https://lambdalabs.com/blog/demystifying-gpt-3

Hardware

“The computational demands of AI presents an emerging problem for its implementation on different hardware platform.”
Source: https://www.nature.com/articles/s41928-018-0068-2

Data

“Data labeling takes anywhere from 35 to 80% of project budgets.”
Source: https://www.forbes.com/sites/cognitiveworld/2022/08/06/no-youre-not-alone-google-is-also-making-this-big-mistake-on-ai

Environment

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

Tiny AI

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

Parallel AI

Foundation

WP 7.1

WP 7.3

Downstream tasks

WP 7.4

WP 7.5

WP 7.6

Implication

WP 7.2

WP 7.3
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.

The Multiply and Max/Min (MAM) neuron is defined as:

\[
y = \max_{i=1}^{N} x_i w_i + b
\]

where \( x_i w_i \) represents the product of an input \( x_i \) and a weight \( w_i \). The neuron sums together the maximum of these products and adds a bias \( b \).

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

**Case study: MAM for ECG autoencoder tail**

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)**

The FC layer is pruned by removing the parameters with the lowest magnitude.

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

$$\text{gain}_f(S, M_1, M_2) = f(S, M_2) - f(S, M_1)$$

**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</th>
<th>acc</th>
<th>$\Delta_{acc}$</th>
<th>$t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>{age=22-40, gender=none, loc=none, speakRate=high, tot_silence=high}</td>
<td>0.03</td>
<td>74.79</td>
<td>-18.38</td>
<td>4.7</td>
</tr>
<tr>
<td>{action=increa, gender=male, speakRate=high}</td>
<td>0.03</td>
<td>74.81</td>
<td>-18.36</td>
<td>4.9</td>
</tr>
</tbody>
</table>

We care about applications:
Efficient 2D and 3D computer vision on edge and very large scale
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

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

(C) Env. Constraints Exploitation

AI and its societal impact

Semantic segmentation of cultural heritage point clouds

Predictive maintenance on district heating networks

Balancing the complexity and interpretability of AI-based energy management strategies

Ethically-sensitive dataset labeling
Spoke 7 in Numbers

- 30+ professors
- 8 Assistant Professors enrolled (more coming soon)
- 4 PhD students enrolled (more coming soon)
- 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
- **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.