Modelling and estimating the cost of creating AI models

Many AI models require expensive training and tuning before they deliver production-ready accuracy. This training is not only time-consuming, but also resource-intensive: such models often train using multiple CPUs and GPUs, multiple runs to tweak parameters, and different architectures. However, we have a limited understanding of the cost of such models - be it in terms energy consumption, carbon emissions, or money. In this project we aim to define a cost model for such applications, and empirically validate it on one or more case-studies.

Methodology:

1. Define the design space of machine learning model design, training, and inference. Demonstrate how it is instantiated for several types of networks.
2. Propose a method to quantify the size of the design space and investigate pruning options.
3. Propose metrics and (software) measurement tools to determine the cost per design, and, combining with the size of the design space, evaluate the cost of exploring the full design space for a model.
4. Propose a *metric of efficiency*, assessing the success of the model versus the effort/cost taken to develop it.
5. Validate the approach with 1-2 case-studies, to showcase the flexibility of the proposed metric(s) and their interpretation.

Seed-papers:

D. de Goede et al - "The cost of Alpha-Zero" (ICPE'22)
D. Patterson at al - "The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink" (IEEE Computer Magazine)

Theory: 40%
Coding: 10%
Evaluation: 30%
Writing: 20%

Contact:

Supervisor (CAES group): Ana-Lucia Varbanescu, a.l.varbanescu@utwente.nl

Coordinator (CAES group): Ghayoor Gillani, s.ghayoor.gillani@utwente.nl, room ZI 5039