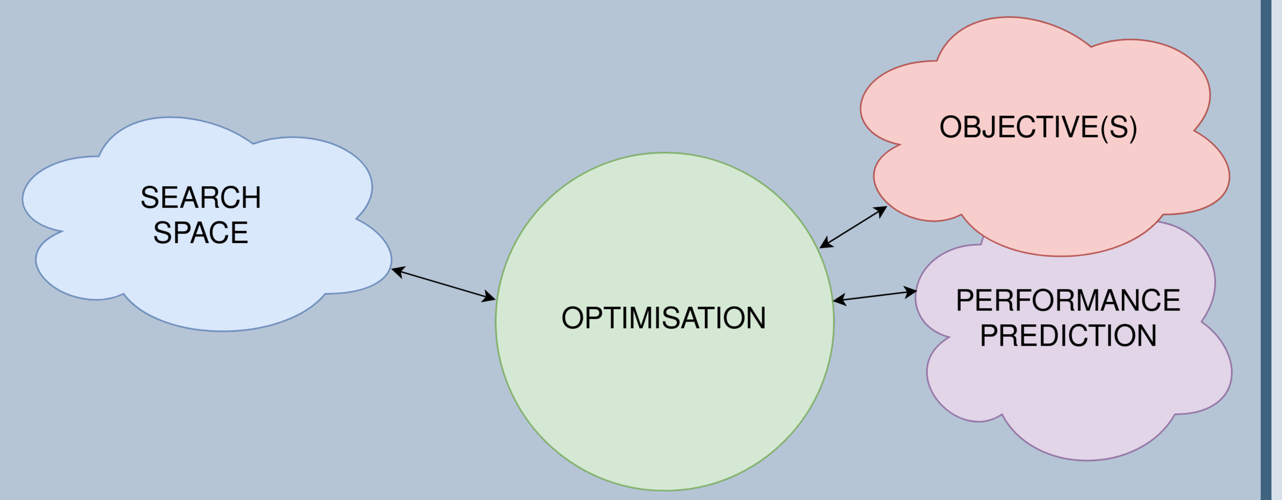
Performance Prediction for Neural Architecture Search

Roman Neruda, Gabriela Kadlecová, Petra Vidnerová
Institute of Computer Science, The Czech Academy of Sciences; Faculty of Mathematics and Physics, Charles University

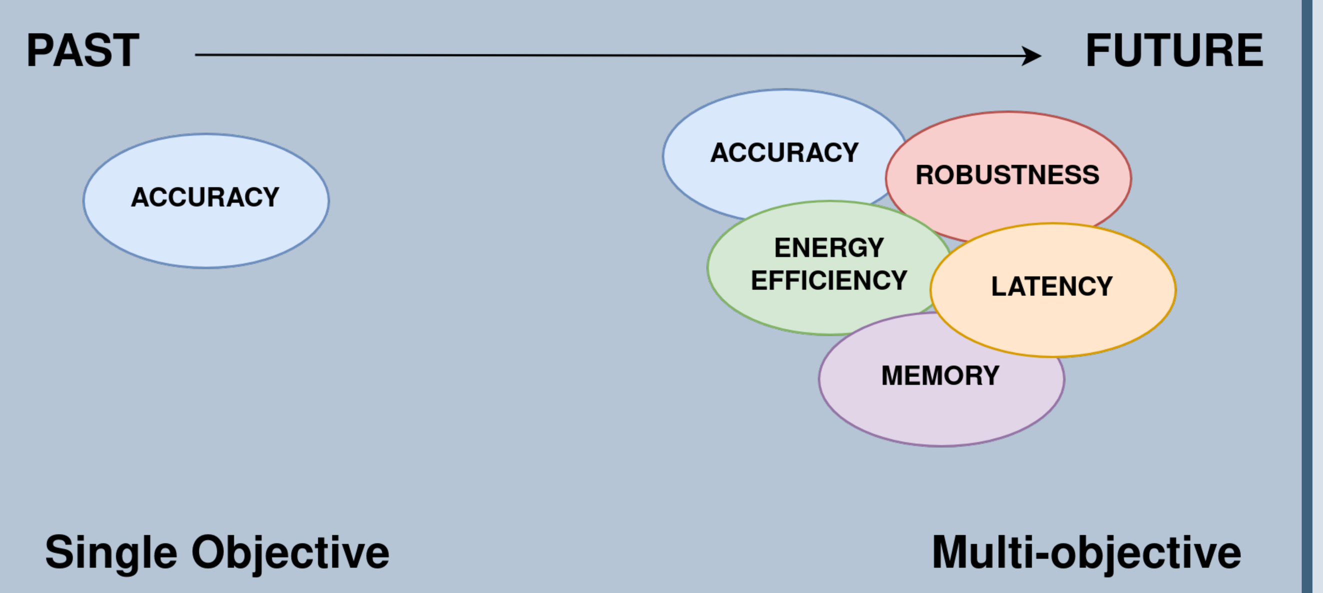
Neural Architecture Search

The performance of a neural network is highly influenced by its architecture. However, manual design and trial and error tuning is very resource-intensive. This is where Neural Architecture Search (NAS) [1] comes into play with automating the whole process.



Future is Multi-Objective

The optimal architecture must balance high accuracy with various constraints like robustness, energy efficiency, or limited memory capacity.
The NAS problem is a **multi-objective optimisation problem**.
Evaluation of objectives is the most time consuming part of the NAS process.



Performance Prediction

Methods to speed up the NAS process:
- parallel evaluation
- weight sharing - train large supernets
- early stopping and learning curve interpolation
- performance prediction and zero-cost proxies

Performance prediction techniques avoid expensive evaluations of objectives – accurate values are replaced by surrogate model predictions. Inaccurate predictions can be used during the optimisation process, especially in the exploratory part – although imprecise, the ranking of networks is often well-preserved.
Zero-cost proxies [2] are metrics that are computed using a single minibatch of data. They attracted considerable attention, as they correlate with the true performance without requiring any network training. But, their abilities are limited and need to be carefully studied.

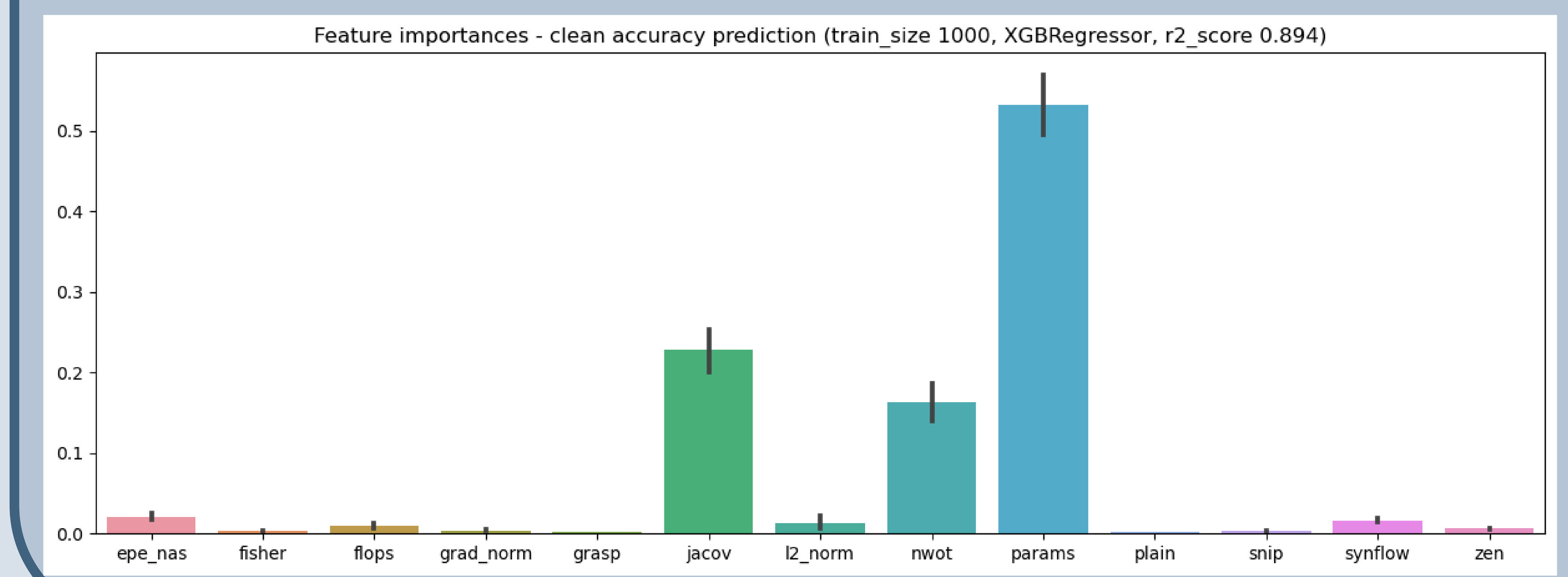
Future Challenges

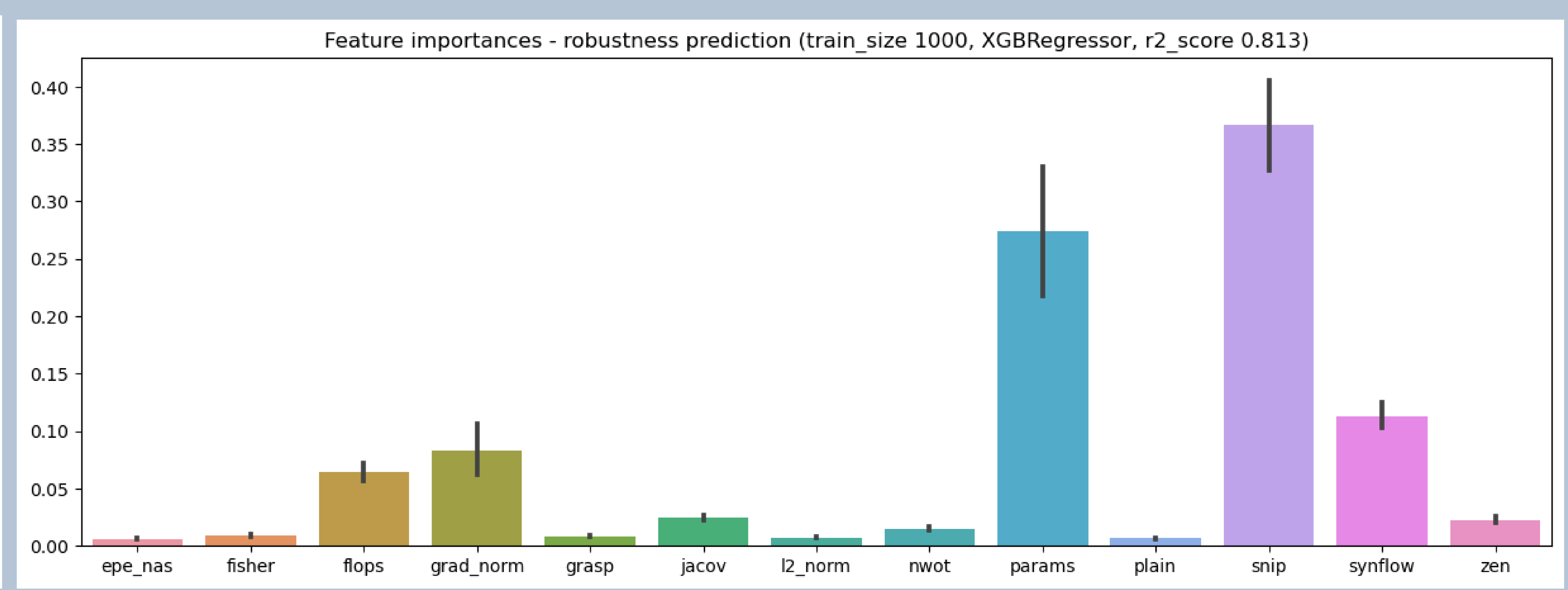
NAS and performance prediction have mostly focused on convolutional neural networks and predictive accuracy. However, there is high demand for large language model design and finetuning, and the discovery of scaling laws.
Objectives like fairness or energy efficiency are highly important when applying LLMs.

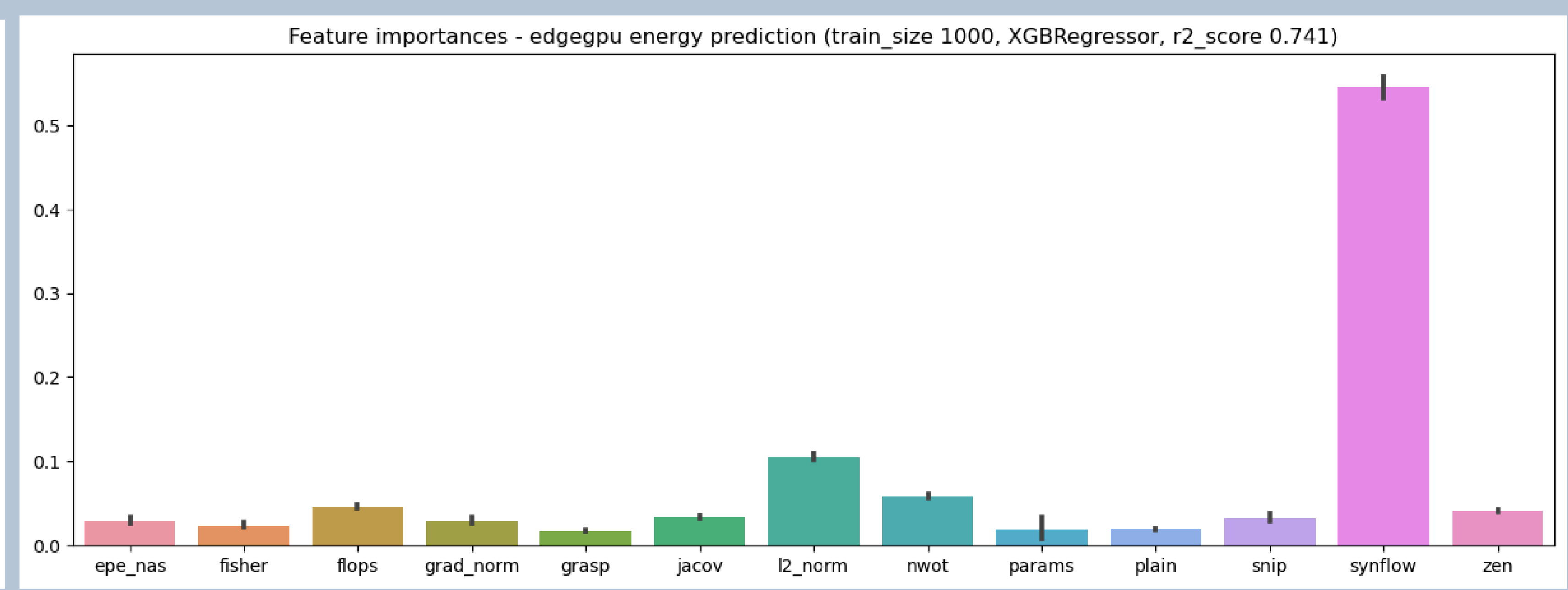
Experiments

Performance prediction by XGBoost Regressor based on zero cost proxies.

	Clean accuracy	Robust accuracy [4]	Edgegpu energy [3]
Kendall's tau	0.815	0.751	0.665







References

[1] White, et al. Neural Architecture Search: Insights from 1000 Papers. (2023)
[2] Krishnakumar, et al. NAS-Bench-Suite-Zero: Accelerating Research on Zero Cost Proxies. NeurIPS Datasets and Benchmarks Track (2022)
[3] Li, et al. HW-NAS-Bench:Hardware-Aware Neural Architecture Search Benchmark. ICLR (2021)
[4] Lukasik, et al. An Evaluation of Zero-Cost Proxies - from Neural Architecture Performance to Model Robustness. GCPR (2023)