Performance Prediction for Neural Architecture Search

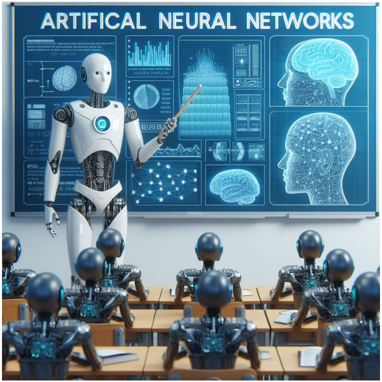
Gabriela Kadlecová, Petra Vidnerová,
Jovita Lukasik, Martin Pilát,
Mahmoud Safari, Roman Neruda, Frank Hutter

Hora Informaticae, March 19, 2024

Outline

Introduction – high level overview

- AutoML and Neural Architecture Search
- Search Spaces and objectives
- Search algorithms
- Speedup techniques



Performance prediction

- Analyzed search spaces
- Performance prediction
- Zero-cost Proxies

New predictor – graph properties

- Motivation, experimental results

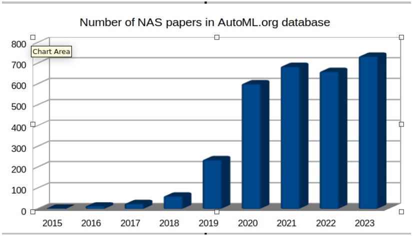
AutoML and Neural Architecture Search

Automated Machine Learning

The process of automating all steps in the machine learning pipeline, from data cleaning, to feature engineering and selection, to hyperparameter and architecture search.

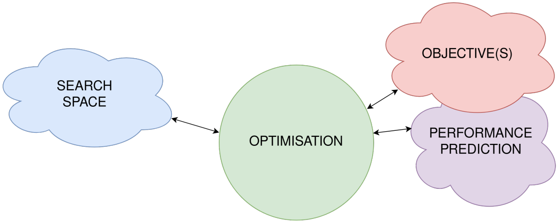
Neural Architecture Search (NAS)

Automating the design of neural network architecture. Given a problem, NAS looks for an optimal architecture.



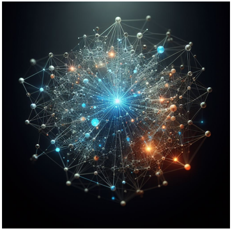
Neural Architecture Search (NAS)

- Optimization problem
- Minimize given objectives over the given search space
- Our focus - speed up the optimization process using performance prediction



Search Spaces

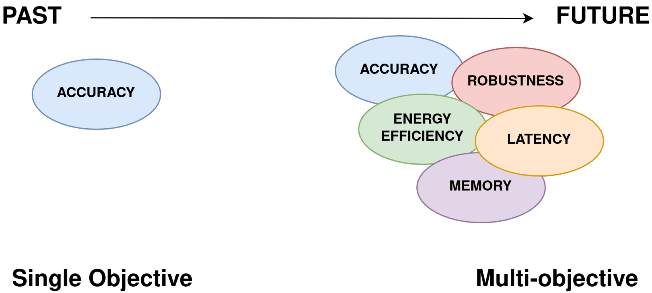
- Space of possible solutions (architectures)
- Trade-off between human bias and search efficiency



- Macro search spaces
  - Encode the entire architecture
  - Focus on macro-level hyperparameters
  - Slow to search
- Chain-structured search spaces
  - A sequential chain of operation layers
  - Easy to design and implement
  - Lower chance of discovering novel architecture
- Cell-based search spaces
  - Search for cells
  - Skeleton fixed
  - Popular, but have limits

Objectives

- Measure the quality of a solution



Optimisation

Black-box techniques

- Random search (baseline)
- Evolutionary and genetic algorithms
- Bayesian optimisation
- Reinforcement learning

One-shot techniques

- Training all at once using hypernet/supernet
- Differentiable architecture search

Speed-up Techniques

Parallelisation

- Easy, parallel objective evaluation
- Evolution with islands

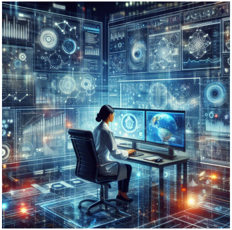
Performance prediction

- Regression of the objective
- Learning curve extrapolation
- Zero-cost proxies

Meta-learning

- Re-using information from previous experiments

Our work



Setting

- What benchmarks and datasets?
- Performance prediction details
- Zero-cost proxies

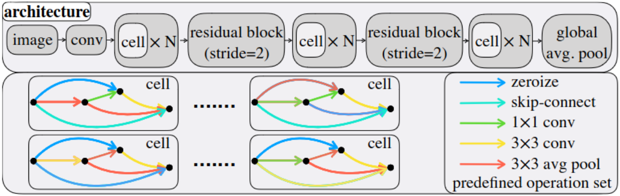
Goals

- Analyze zero-cost proxies as predictors
- Properties of the neural graph as a novel predictor
- Interpretability analysis of predictions
- Compare with predictors from related work

NAS Benchmarks

- Datasets of precomputed objectives on selected tasks
- Enables experiments and comparison of NAS algorithms, performance prediction algorithms
- Important for reproducible research

- NAS-Bench-101, NAS-Bench-201, NAS-Bench-301
- HW-NAS-Bench, TransNAS-Bench-101, robustness NB201

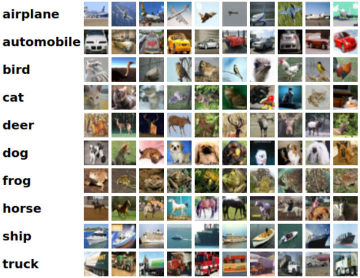


Source: NAS-Bench-201: Extending the scope of reproducible NAS. ICLR 2020

Image Classification Datasets

CIFAR10, CIFAR100

- Learning Multiple Layers of Features from Tiny Images, Alex Krizhevsky, 2009.
- 10/100 classes
- 60k images, 32x32 pixels



ImageNet-16-120

- A downsampled variant of Imagenet as an alternative to the Cifar dataset, Chrabaszcz et al, 2017
- 1000 classes

Limits of benchmarks

NAS-Bench-201

- Evaluated for different datasets and objectives
- Total of 15 625 candidates
- However, some of them are isomorphic
- Some have invalid branches
- The valid and unique set is quite small

NB101, NB301

- Larger, but evaluated only on CIFAR10
- Only one objective (accuracy)
- Cell-based – but models like LLMs are different

Performance Prediction

Predict objectives

- Imprecise prediction is enough (coarse to grain)
- Ranking is enough (who is the best)

Our goals

- Performance prediction of diverse objectives
- Accuracy, robustness, energy
- Exploring/combining zero cost proxies
- Proposal of new network encodings

Methodology

Regression

- Random forest regressor
- Predict accuracy or other metrics

Input data – network encodings

- Zero cost proxies
- One hot encoding (of chosen operations)
- Graph properties

Experiments

- Analyze predictions
- Compare different network encodings, predictors

Zero-cost proxies (ZCP)

- Fast to compute metrics that correlate with accuracy
- Zero-cost ... because we don't train the network at all!
- Some proxies depend on input data
- Other use artificial batches, e.g. a batch full of 1

How to compute ZCP?

- Sample one minibatch of data (or create an artificial batch)
- Pass it through the (untrained) network
- Compute a metric as a function of the forward pass and/or the gradient

ZCP in performance prediction

Main approaches

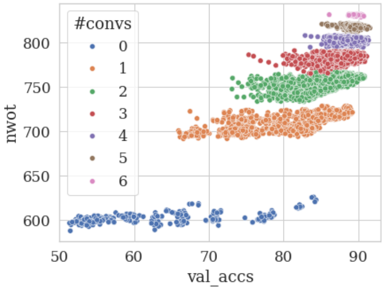
- Direct approximation of performance – choose nets with the highest score
- Warm-start search – initial generation are top-scored nets
- ZCP as net encoding – fit a regressor on multiple ZCP, predict performance

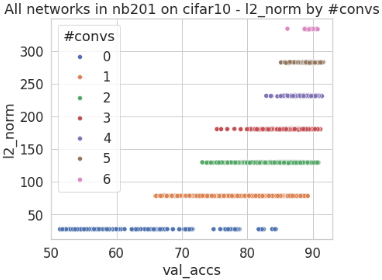
Examples of proxies

- flops, params – just simple metrics (no batch pass)
- synflow – from network pruning, product of network parameters
- nwot – activation of different ReLU regions (variance between batch examples)

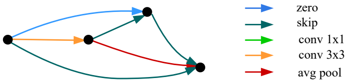
ZCP limitations

- For NB201, ZCP correlate surprisingly well with accuracy
- On some other searchspaces, the correlation is rather low
- We discovered the reason for the good correlation on NB201
- For proxies like nwot, l2_norm, the score directly depends on the number of convolutions in the network

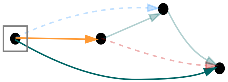




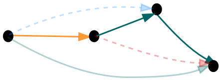
Properties of the neural graph



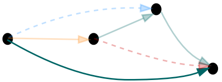
- Inspired by the finding, we look at properties of the network graph – paths, counts, ...
- Node degree (c3x3, skip) means input degree counting only conv3x3 and skip
- Similarly, max path computes the maximum path over allowed operations



Node degree (c3x3, skip): 2



Max path (c3x3, skip): 3



Max path (skip): 1

Properties of the neural graph

- Number of operations
- Min path from input over operations O
- Max path from input to over operations O
- Out degree of the input node counting only operations O
- In degree of the output node counting only operations O
- Mean in/out degree of intermediate nodes counting only operations O

Advantages

- Simple, interpretable, fast to compute

Disadvantages

- Highly correlated, dependent on search space

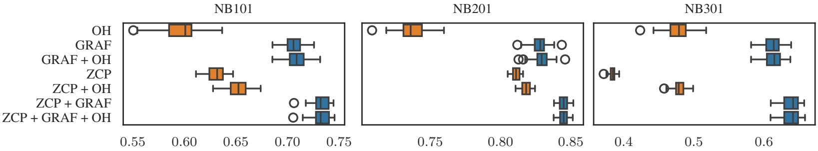
Accuracy prediction

Usage in performance prediction

- Gathering all properties, we use them as input data to a random forest regressor
- We compare with other network encodings – all ZCP, one-hot encoding (OH), and their combinations

Results

- Our results (GRAF) are better than ZCP and OH
- ZCP combined with network properties is the best

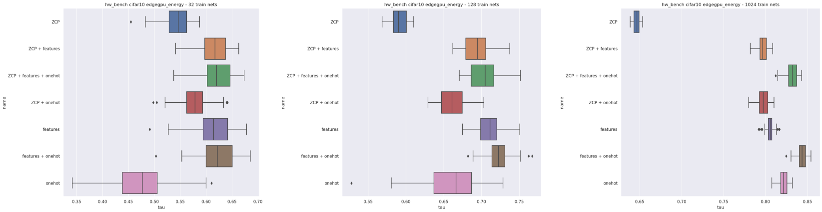


Interpretable prediction

- For network properties and ZCP, we compute Shapley coefficients (considers feature set importance)
- We look at the most important features
- Results – different features are important for diverse tasks
- nwot is important for CIFAR10, but not for autoencoder
- autoencoder needs skip connections – result from related work!

NB201 - `cifar-10`		TNB101-micro - `autoencoder`	
Feature name	Mean rank	Feature name	Mean rank
jacov	0.00	min path over skip	0.00
nwot	1.12	jacov	1.00
flops	3.62	fisher	2.00
synflow	4.08	min path over [skip,C3x3]	5.50
min path over [skip,C3x3,C1x1]	4.78	snip	5.58
params	5.04	min path over [skip,C1x1]	5.64
epe_nas	6.04	grad_norm	6.64
zen	6.36	zen	8.08
min path over [skip,C3x3]	11.08	grasp	9.34
min path over skip	11.88	l2_norm	9.74

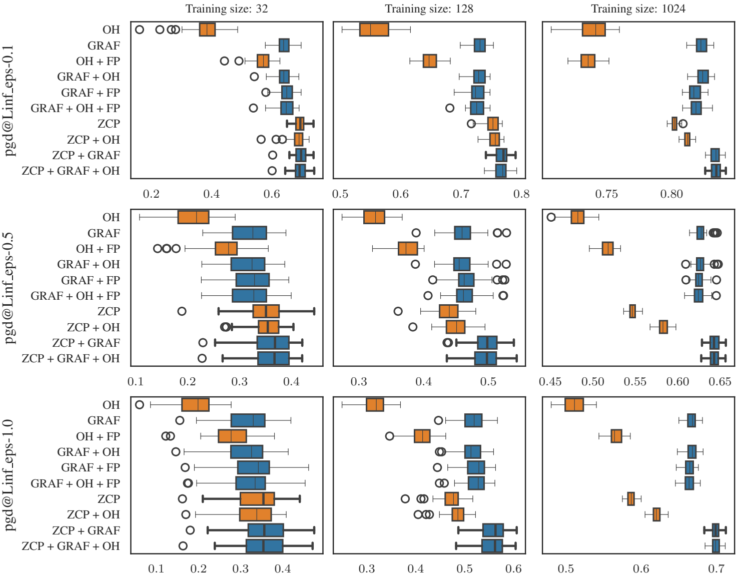
HW metrics prediction



Prediction of EDGEGPU energy, random forest.

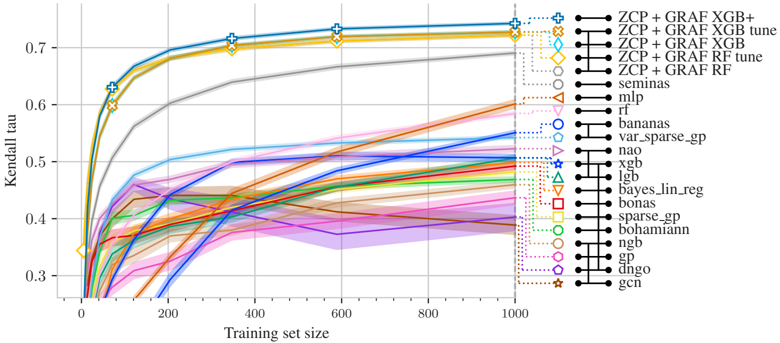
- HW metrics differ in difficulty of prediction
- `edgegpu energy` is one of difficult tasks

Robustness prediction



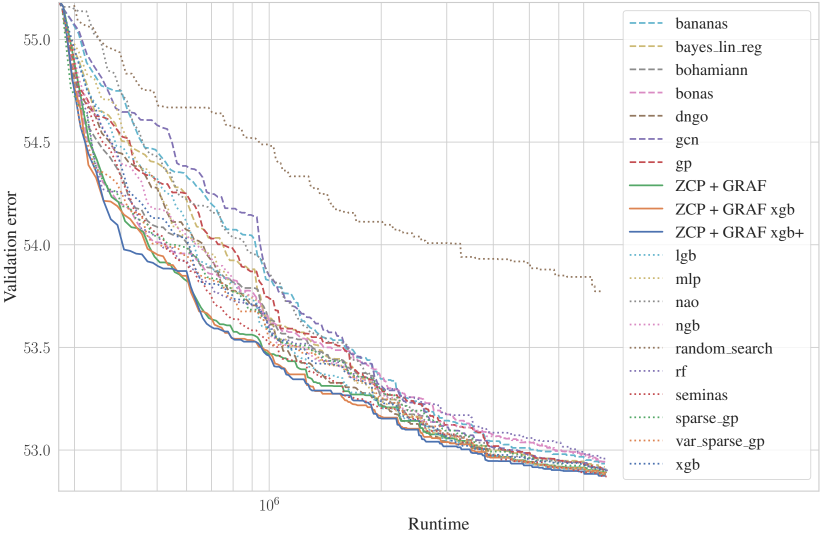
Comparison with existing predictors – NB101, CIFAR10

- Same experiment as in a predictor survey [1]
- Outperforms all available predictors
- Some predictors take much longer to train, e.g. graph neural networks!



[1] Colin White, Arber Zela, Binxin Ru, Yang Liu, & Frank Hutter (2021). How Powerful are Performance Predictors in Neural Architecture Search?. In Advances in Neural Information Processing Systems.

Usage in the NAS process – ImageNet16-120 search



Discussion and future work

Pros

- Better and faster than most predictors
- Great interpretability
- Works across tasks and search spaces
- Baseline for complex predictors

Cons

- Properties need to be used with ZCP for best performance
- Some graph neural networks with ZCP can be better

Future work

- Extension to transformer or LLM search spaces
- Study why ZCP are still needed

Thank you! Questions?

Images in the presentation generated by DALL-E3.