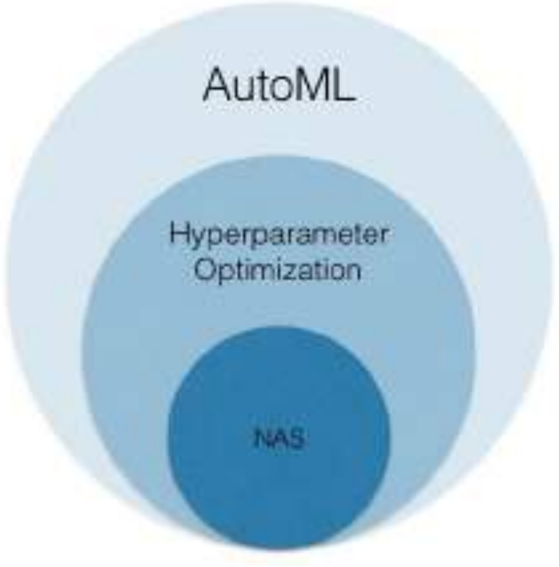
Computer Vision for Embedded Systems

Yung-Hsiang Lu
Purdue University
yunglu@purdue.edu

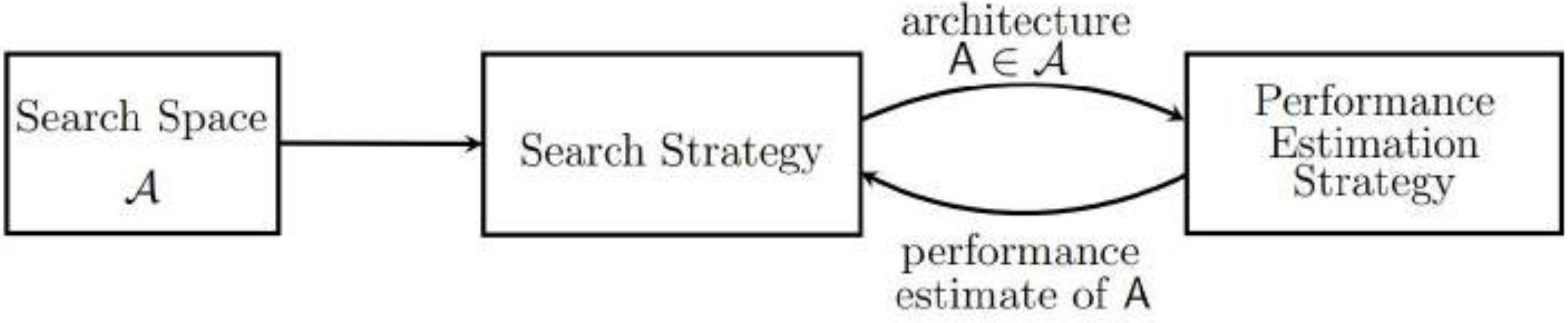
Why to search Neural Network Architectures?

- Many "hyper parameters" in neural networks
- Setting the right values is not easy
- Different neural networks are needed for
  - benchmarks (dataset)
  - performance objectives
  - constraints
  - hardware



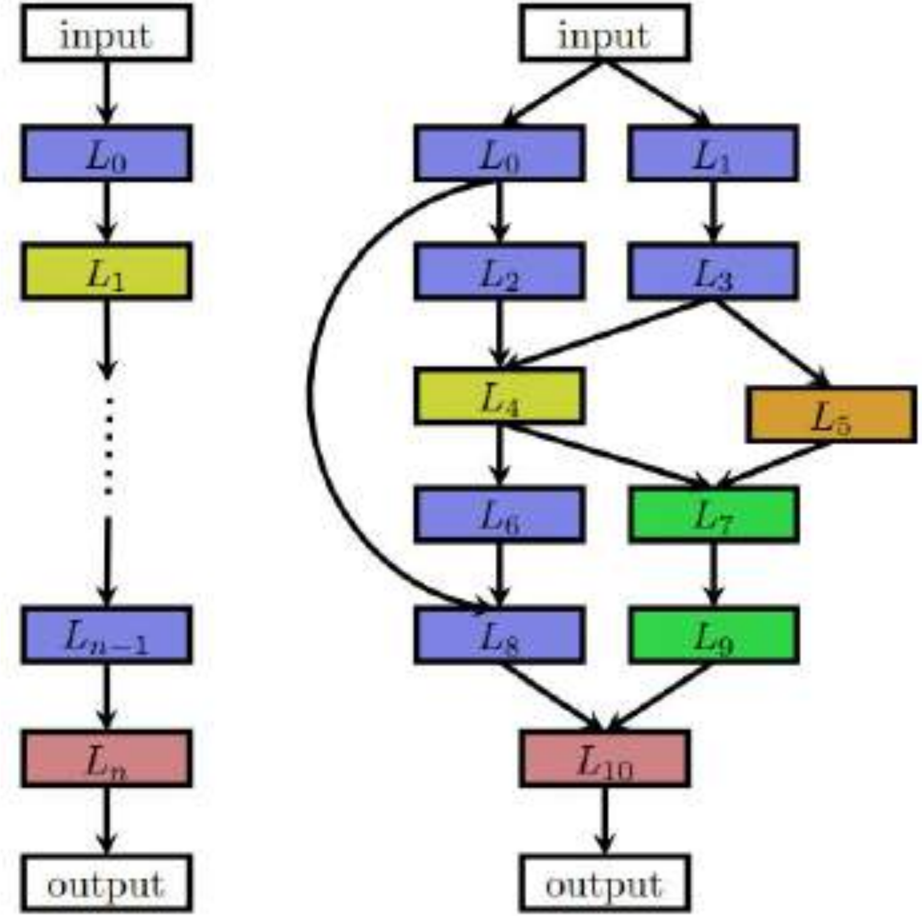
https://www.oreilly.com/content/what-is-neural-architecture-search/

Network Architecture Search



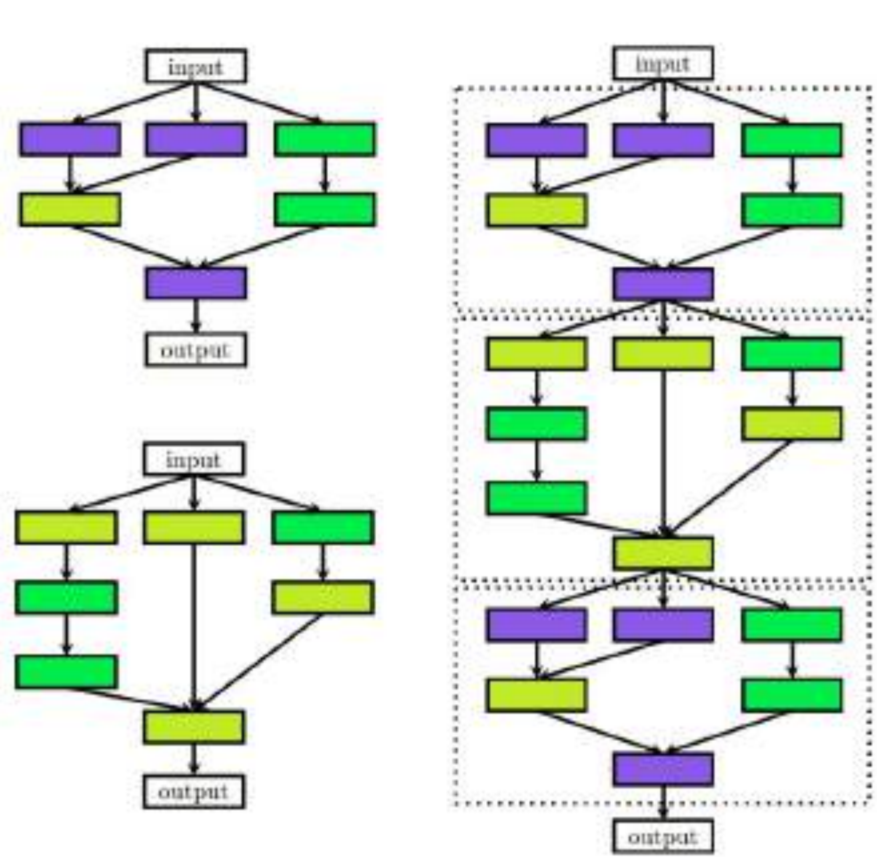
Neural Architecture Search: A Survey

Journal of Machine Learning Research (2019)


input
L_0
L_1
L_{n-1}
L_n
output
input
L_0
L_1
L_2
L_3
L_4
L_5
L_6
L_7
L_8
L_9
L_{10}
output

ce

Stack Architectures



Search Strategy

- random search
- Bayesian optimization
- evolutionary methods
- reinforcement learning
- gradient-based methods

Bayesian optimization

- Select a dataset and a task
- Define a search space
- Define an objective function f to evaluate the performance
- Define a distance function d between two architectures
- If one architecture is better than another, "move" toward the first architecture.

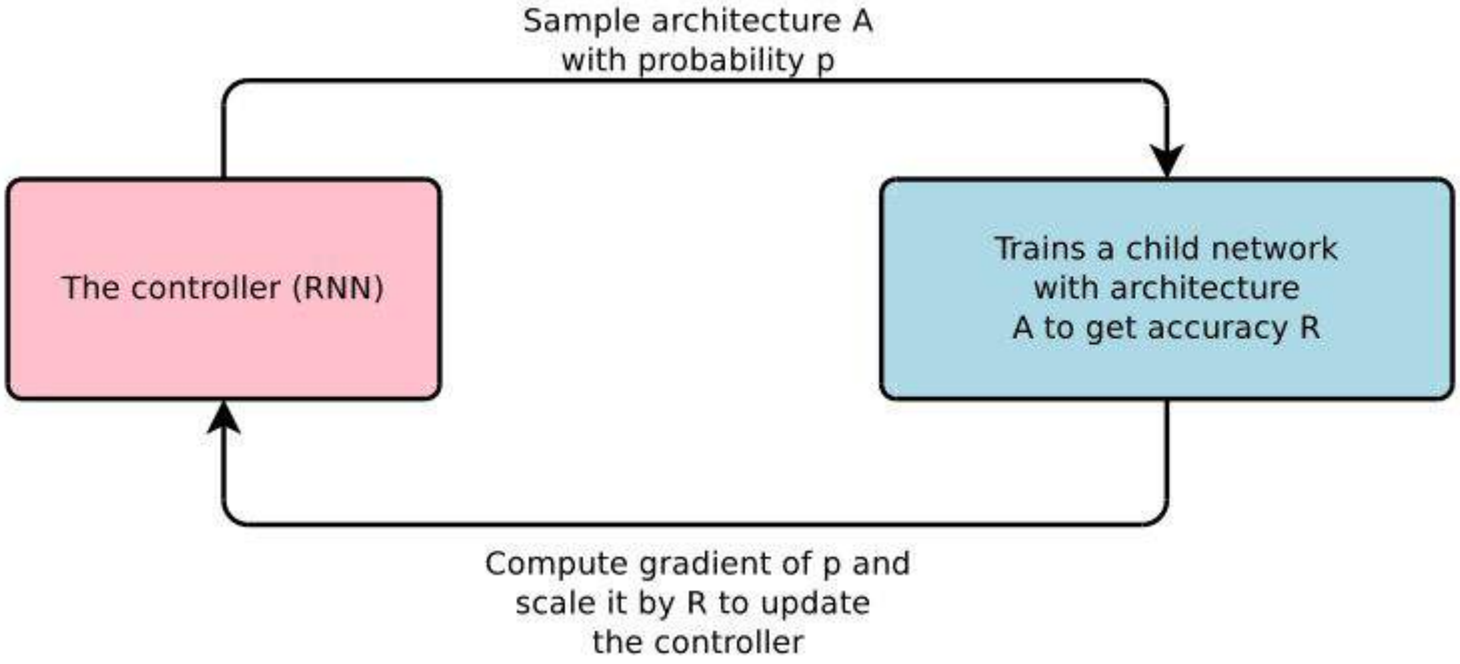
https://medium.com/abacus-ai/an-introduction-to-bayesian-optimization-for-neural-architecture-search-d324830ec781

Evolutionary Algorithm

1. Evaluate the fitness of each individual in the population
2. Select the fittest individuals for reproduction (Parents)
3. Breed new individuals through crossover and mutation operations to give birth to offspring
4. Replace the least-fit individuals of the population with new individuals

https://en.wikipedia.org/wiki/Evolutionary_algorithm

Reinforcement Learning for Architecture Search



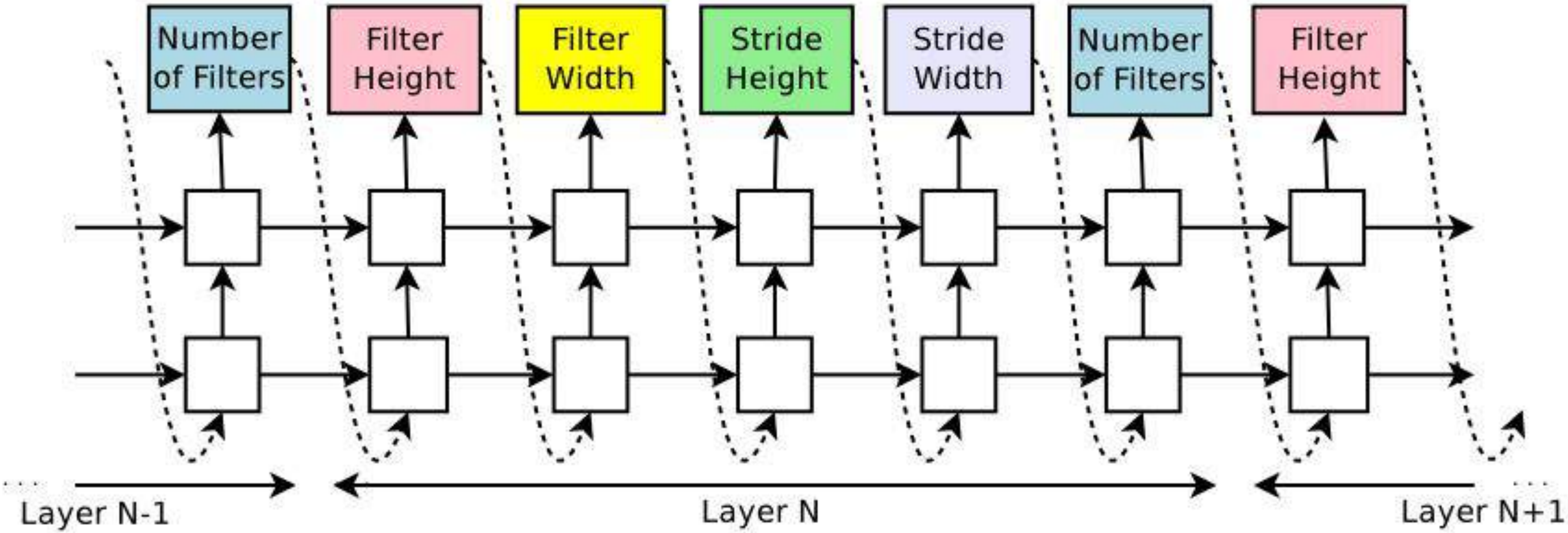
Sample architecture A
with probability p

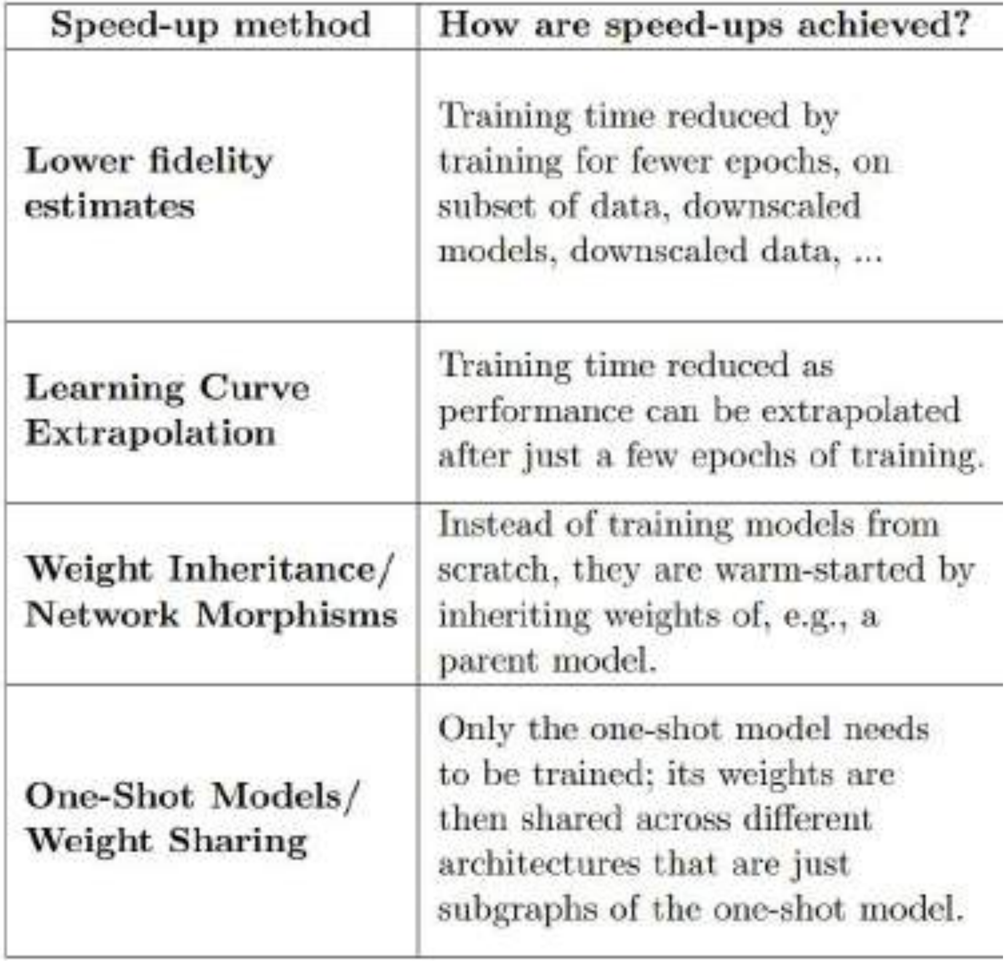
The controller (RNN)

Trains a child network
with architecture
A to get accuracy R

Compute gradient of p and
scale it by R to update
the controller

Neural Architecture Search with Reinforcement Learning, Barret Zoph, Quoc V. Le, ICLR 2017


Number of Filters
Filter Height
Filter Width
Stride Height
Stride Width
Number of Filters
Filter Height
Layer N-1
Layer N
Layer N+1



Speed-up method	How are speed-ups achieved?
Lower fidelity estimates	Training time reduced by training for fewer epochs, on subset of data, downscaled models, downscaled data, ...
Learning Curve Extrapolation	Training time reduced as performance can be extrapolated after just a few epochs of training.
Weight Inheritance/ Network Morphisms	Instead of training models from scratch, they are warm-started by inheriting weights of, e.g., a parent model.
One-Shot Models/ Weight Sharing	Only the one-shot model needs to be trained; its weights are then shared across different architectures that are just subgraphs of the one-shot model.