Modern Approaches for Heuristic Algorithms in Network Architecture Searches (NAS) and tinyML

B.Sc. Daniel Duclos-Cavalcanti

Advisor: M.Sc. Alex Hoffman

Wissenschaftliches Seminar VLSI Entwurfsverfahren

SoSe 2022

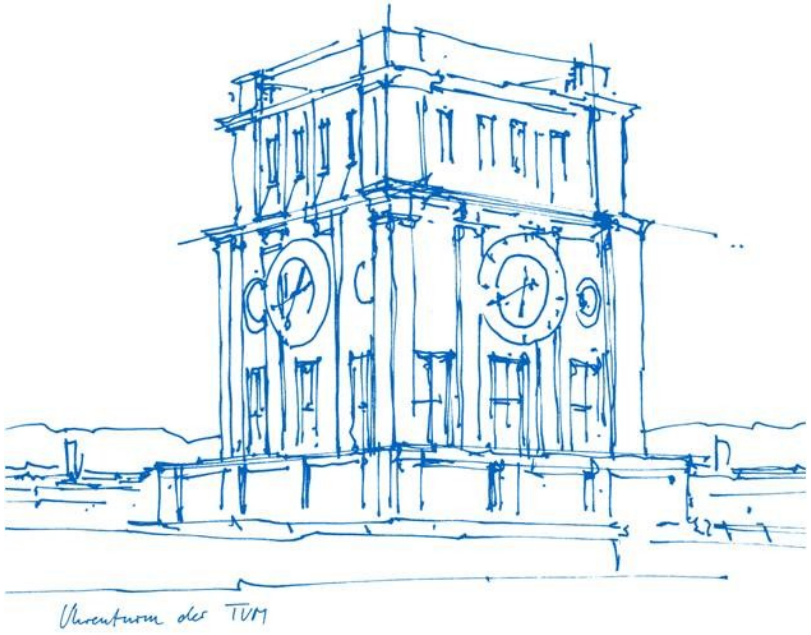


Table of Contents

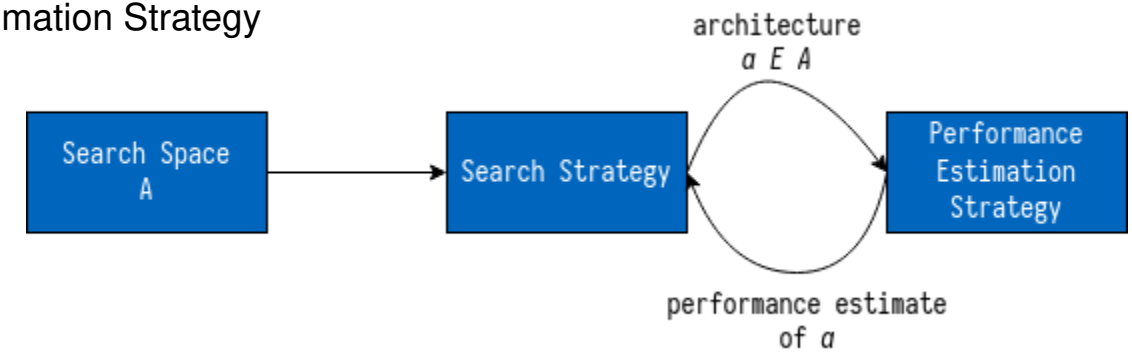
Introduction

- Deep Learning
  - Groundbreaking technology
  - Algorithms complexity increase
  - Architecture engineering and domain expertise

- Neural Architecture Search (NAS)
- TinyML and NAS

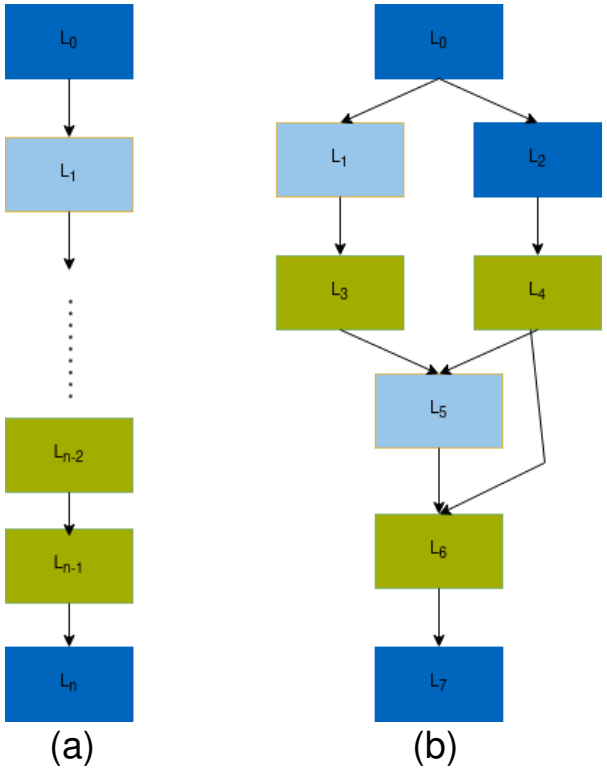
Background – Neural Architecture Search (NAS)

- NAS methods categorized by three dimensions:
  - Search Space
  - Search Strategy
  - Performance Estimation Strategy



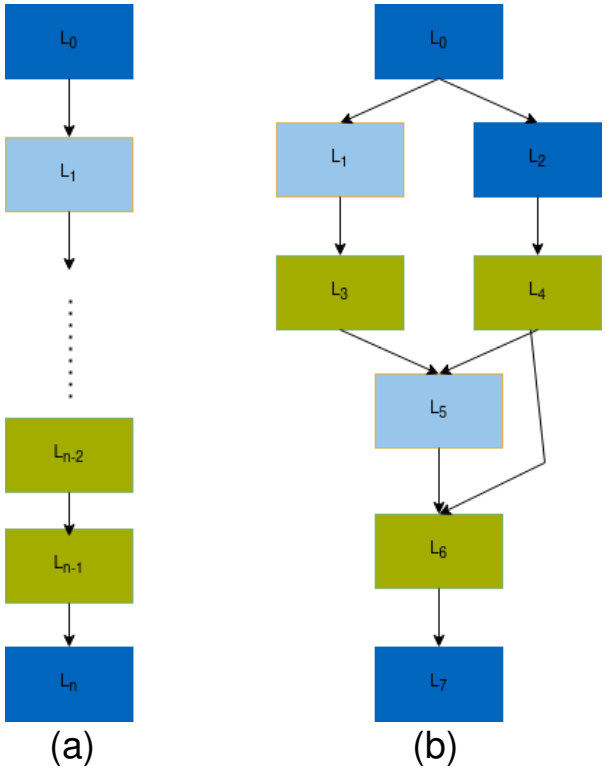
NAS – Search Space

- Optimization problem
- Exploration x Exploitation trade-of
- Chain-Structured NN (a) x Multi-Branch NN (b)



NAS – Search Space

- Optimization problem
- Exploration x Exploitation trade-of
- Cells (a) x Blocks (b)



NAS – Search Strategy

- Reinforcement Learning (RL) - Based
  - Agent's action – generation of a neural architecture
  - Action space – search space
  - Agent's reward – performance estimation

- Evolutionary Computation (EC) - Based
- Gradient Descent-Based

NAS – Performance Estimation Strategy

- Bias vs Speed trade-off

Speed-up Method	**How?**
Lower Fidelity Estimates	Shorter training time, training on subset of the data, training on downscaled data or with downscaled models.
Learning Curve Extrapolation	Performance is extrapolated after just a small number of epochs and then decided upon directly. Hyperparameters to predict promising architectures after partial learning. Extrapolate partial learning curves to predict and eliminate sub-optimal architectures.
Weight Inheritance	Weights from previously trained models passed down to new ones.
Weight Sharing or One-Shot Models	All architectures as subgraphs of a supergraph (one-shot model). Weights are shared between specific architectures. Only one-shot is trained.

State-of-the-Art

- Reinforcement Learning (RL) - Based
  - Costly, hundreds even thousands of GPU days
- Gradiant Descent - Based
  - Faster and efficient
  -
- Reinforcement Learning (RL) - Based
  - Costly, hundreds even thousands of GPU days

Grundlage der Masterfolien

Als Grundlage dient der Corporate Design Style Guide der TUM.
Die Präsentationsvorlage ist auf gute Lesbarkeit und klare Darstellung von Informationen optimiert.

Grundlage der Masterfolien

Als Grundlage dient der Corporate Design Style Guide der TUM.
Die Präsentationsvorlage ist auf gute Lesbarkeit und klare Darstellung von Informationen optimiert.

Grundlage der Masterfolien

Als Grundlage dient der Corporate Design Style Guide der TUM.
Die Präsentationsvorlage ist auf gute Lesbarkeit und klare Darstellung von Informationen optimiert.

Grundlage der Masterfolien

Als Grundlage dient der Corporate Design Style Guide der TUM.
Die Präsentationsvorlage ist auf gute Lesbarkeit und klare Darstellung von Informationen optimiert.

Conclusions

- Conclusion 1
  - foo
  - bar
- Conclusion 2
- Conclusion 3

Questions

Farben

Als erstes soll mit schwarz und weiß gearbeitet werden.
Für Aufwändigere Darstellungen sind Farben mit Bedacht und in möglichst geringem Umfang einzusetzen.
In diesem Folienmaster ist die Farbpalette festgelegt.

Zuerst mit den Primärfarben arbeiten.

Für z.B. komplexe Diagramme stehen noch Sekundärfarben zur Verfügung.

Gering im Einsatz sind die Akzentfarben.