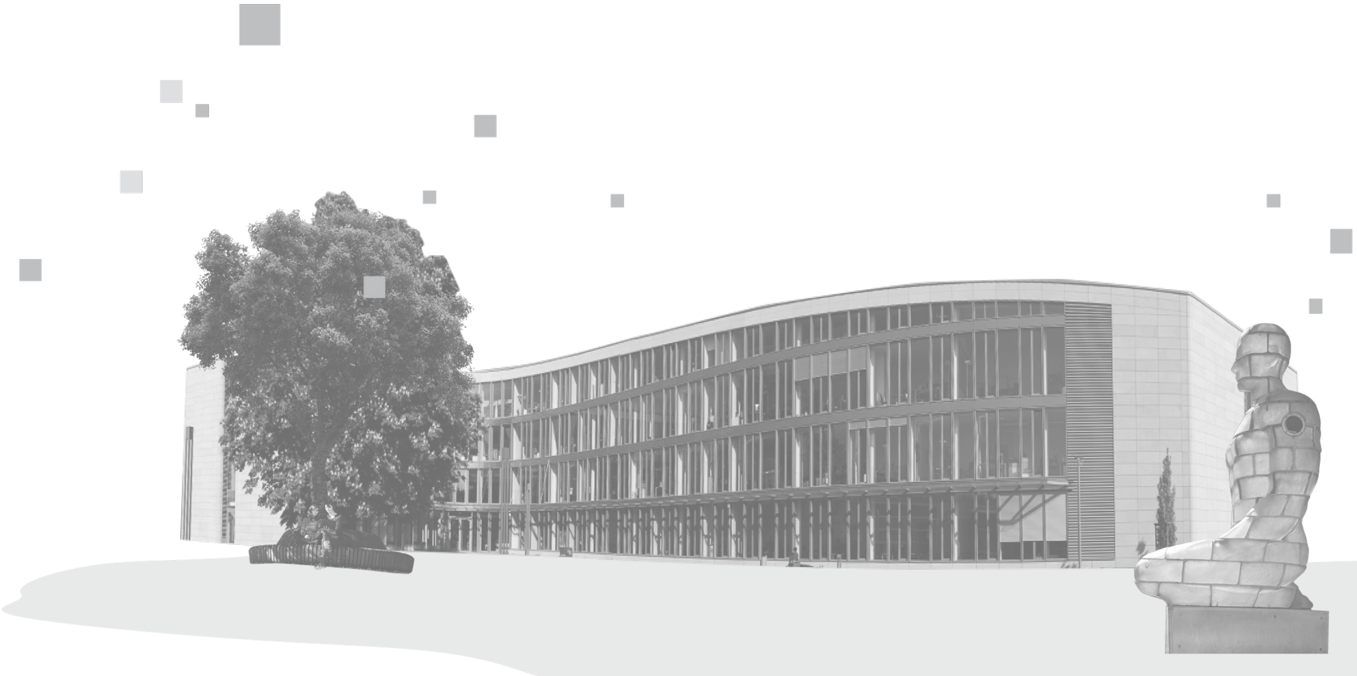
Dynamic Networks

PD Dr. Haojin Yang

Multimedia and Machine Learning Group

Hasso Plattner Institute

Hasso Plattner Institut
Digital Engineering · Universität Potsdam



Design IT.
Create Knowledge.

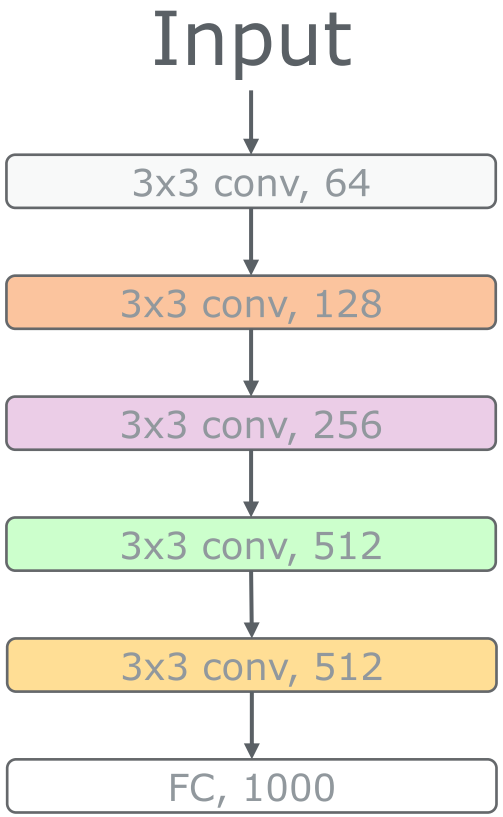
www.hpi.de

Dynamic Networks

Why Dynamic?

Workflow of a static network

- Training
  - Pre-defined network architecture
  - Weight initialization
  - Optimization using training data
- Inference
  - **Fixed** network **architecture** and **weight**
  - Forward pass for arbitrary input samples

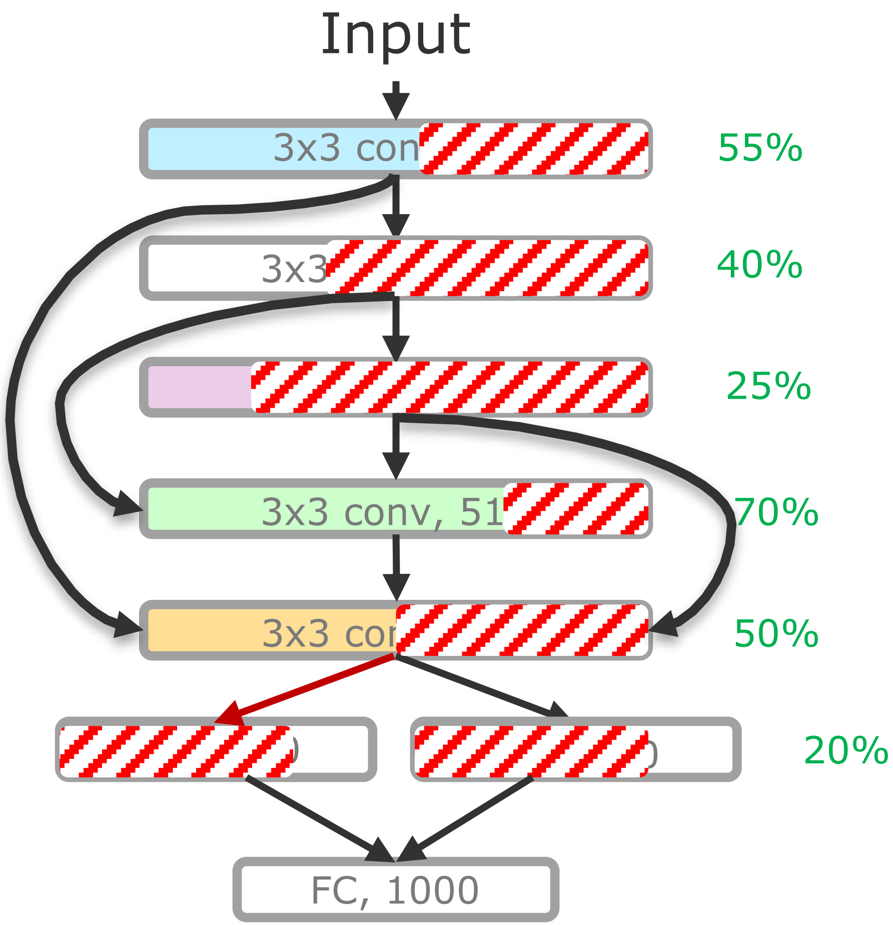


Dynamic Networks

Why Dynamic?

Workflow of a dynamic network

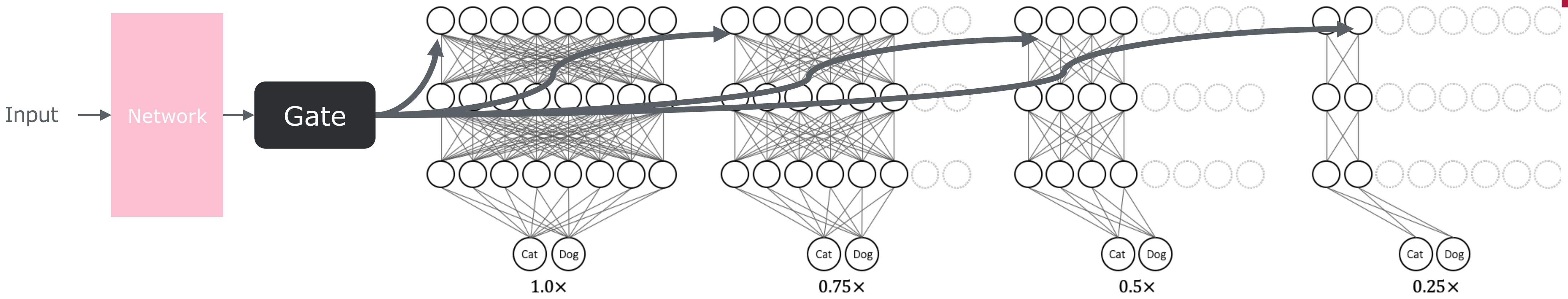
- Training
  - Learn a specific network structure for each sample
  - Joint training of neural network and decision-making mechanism
- Inference
  - Instance-wise dynamic network structure
  - The decision-making mechanism predicts the structure based on each input sample.
- Type:
  - Dynamic width
  - Dynamic depth
  - Dynamic path

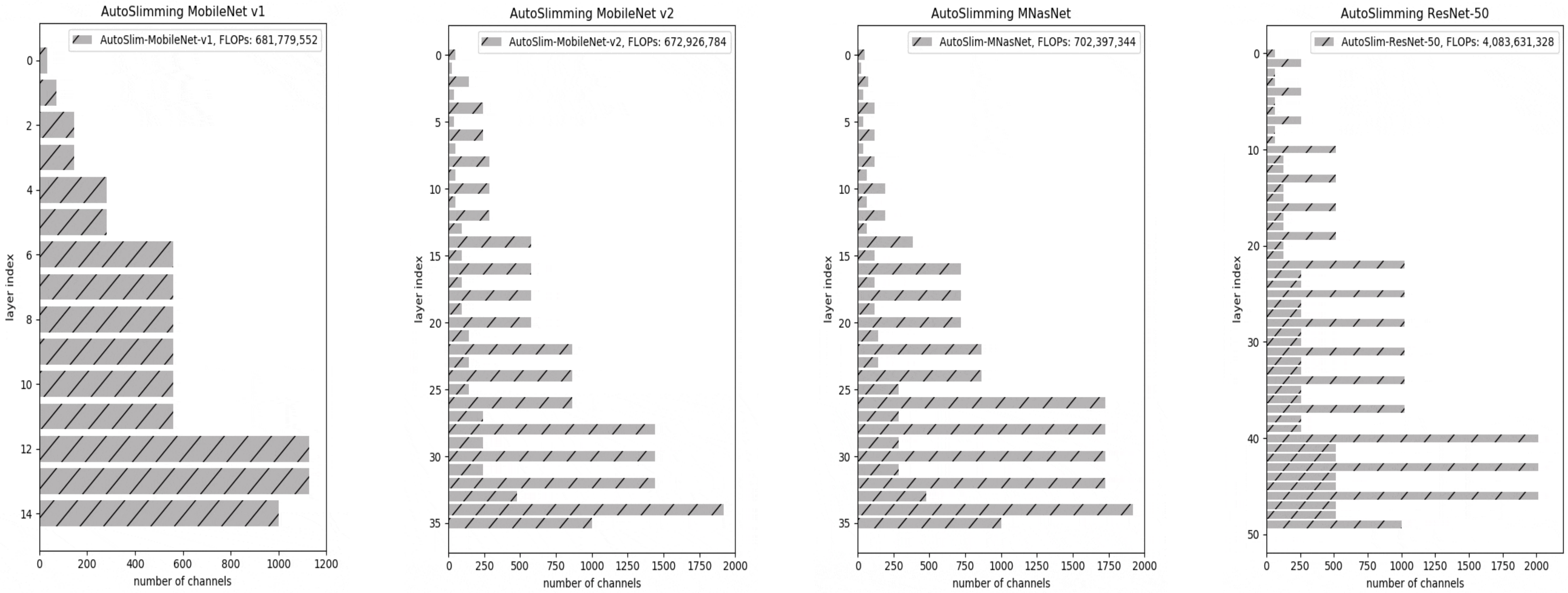


Advantages

- Efficiency: Strategically allocate computations on demand at test time
- Representation power: Enlarged parameter space and improved representation power
- Adaptiveness: Desired trade-off between accuracy and efficiency on the fly
- Compatibility: Compatible with most advanced techniques in deep learning
- Generality: Can be applied seamlessly to a wide range of applications
- Interpretability: It is believed that the brains process information in a dynamic way

Dynamic width

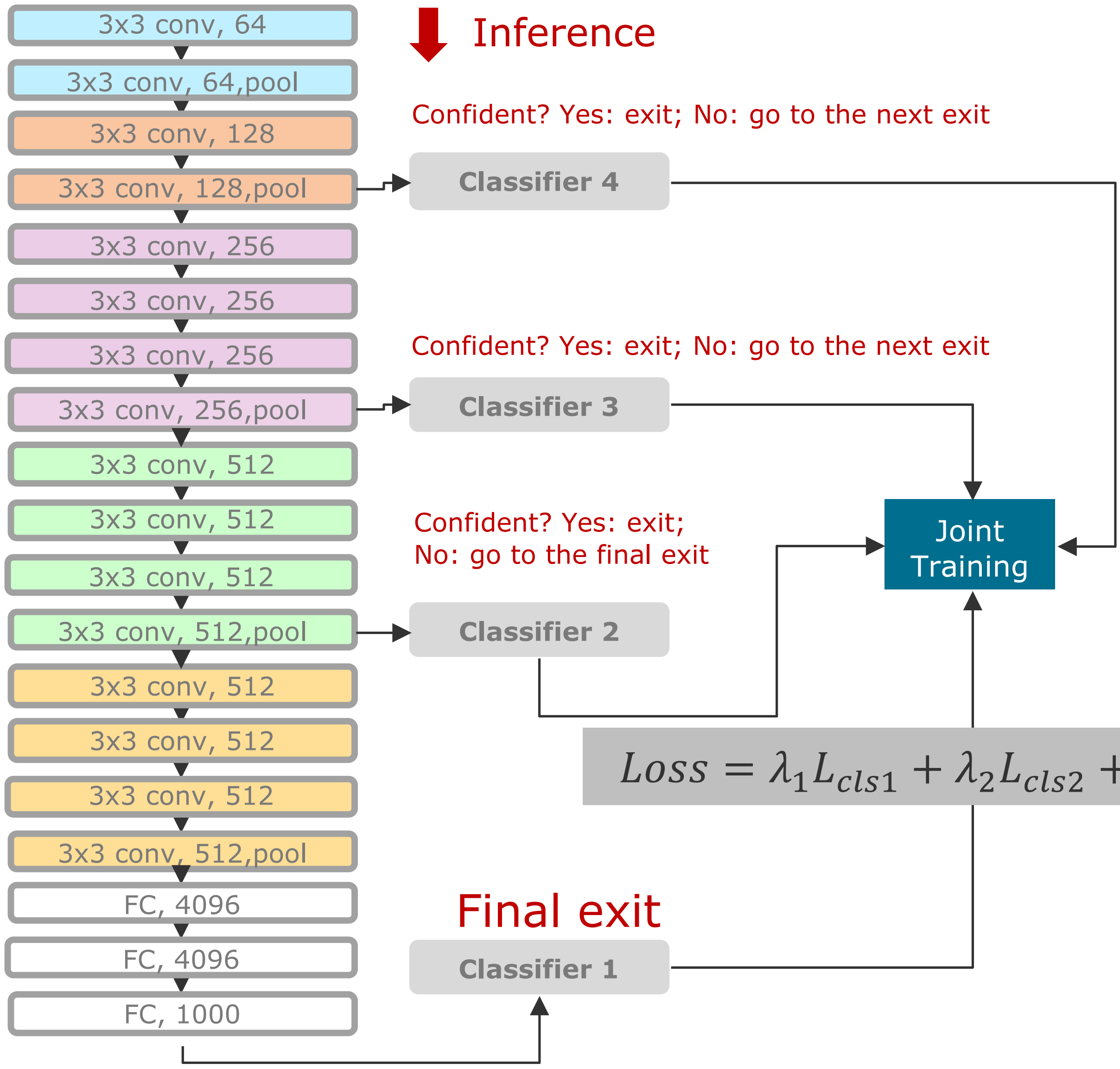




Yu, Jiahui, and Thomas Huang. "Autoslim: Towards one-shot architecture search for channel numbers." arXiv preprint arXiv:1903.11728, 2019

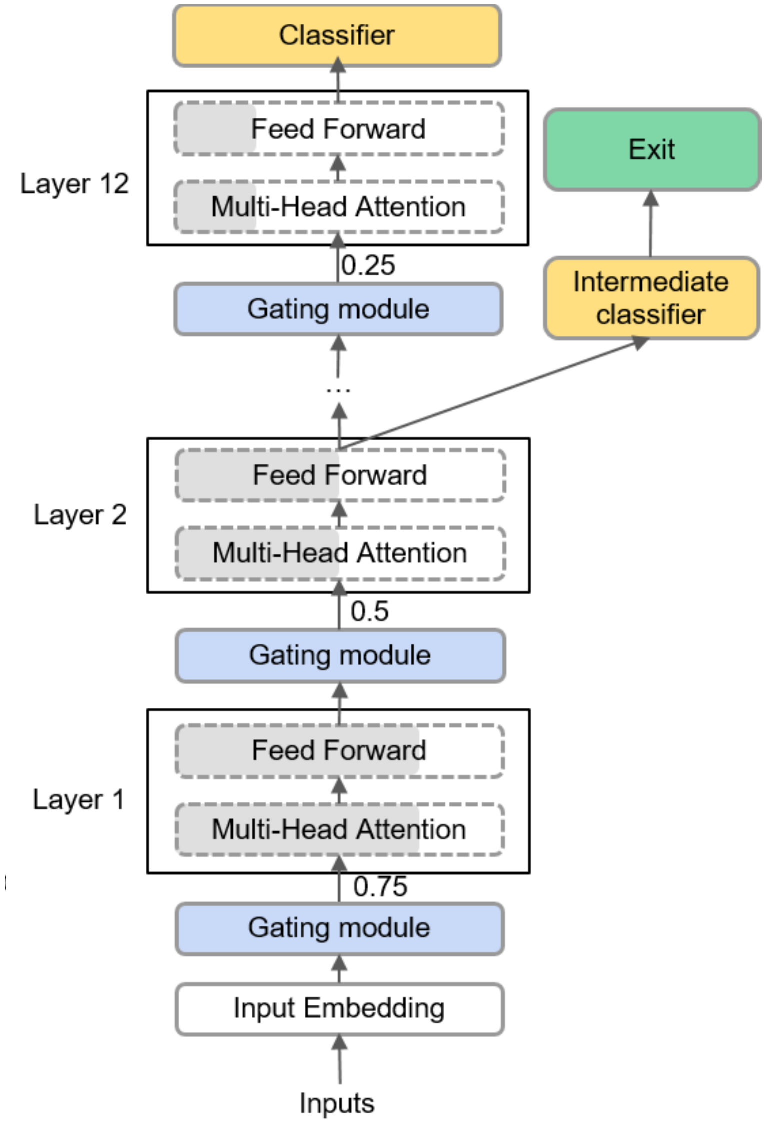
Dynamic depth

- Early Exits



$$Loss = \lambda_1 L_{cls1} + \lambda_2 L_{cls2} + \lambda_3 L_{cls3} + \lambda_4 L_{cls4}$$

Case study: Dynamic BERT



Classifier

Layer 12

Feed Forward

Multi-Head Attention

0.25

Gating module

...

Layer 2

Feed Forward

Multi-Head Attention

0.5

Gating module

Layer 1

Feed Forward

Multi-Head Attention

0.75

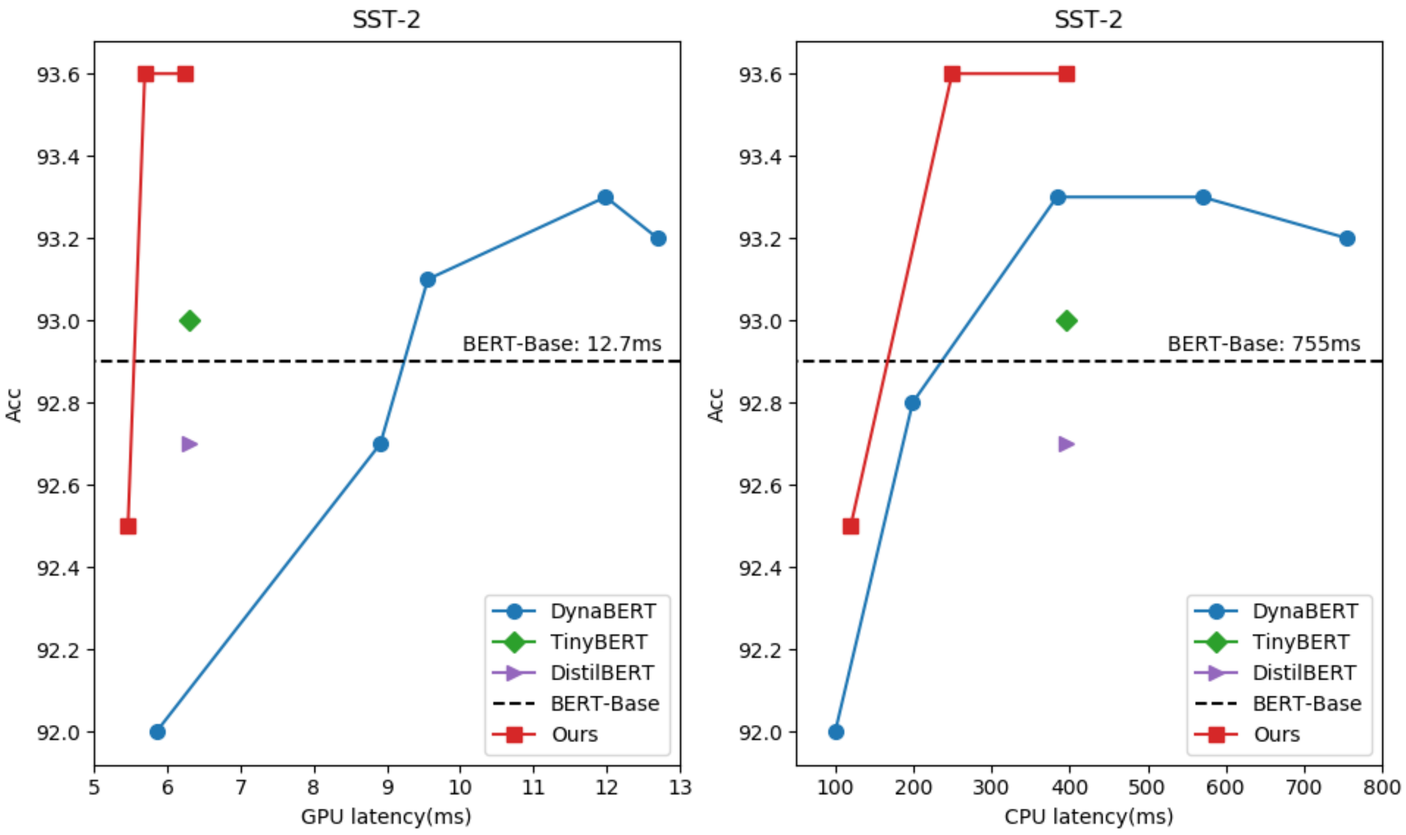
Gating module

Input Embedding

Inputs

Exit

Intermediate classifier



Hu, T., Meinel, C., & Yang, H. (2024). A flexible bert model enabling width-and depth-dynamic inference. Computer Speech & Language, 87, 101646.

Summary

- Differences between static and dynamic networks
- Advantages of dynamic networks
- Case study
    - Dynamic width
    - Dynamic depth
- Dynamic BERT example
- Upcoming: Challenges and solutions for efficient LLM deployment