Efficient Deep Learning Methods

PD Dr. Haojin Yang

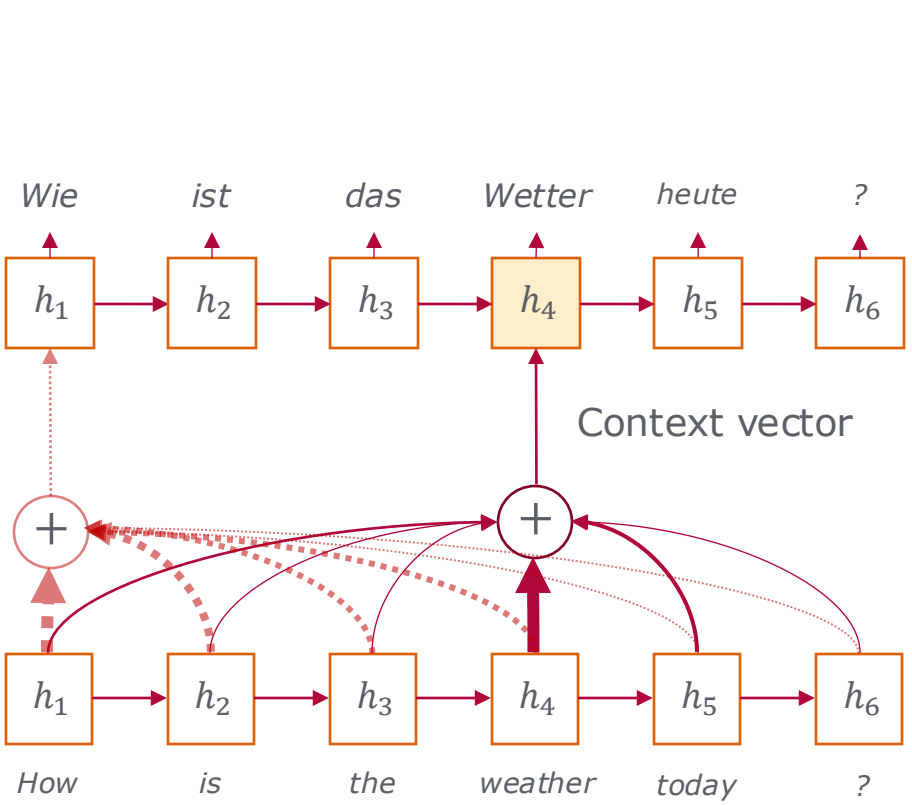
Multimedia and Machine Learning Group

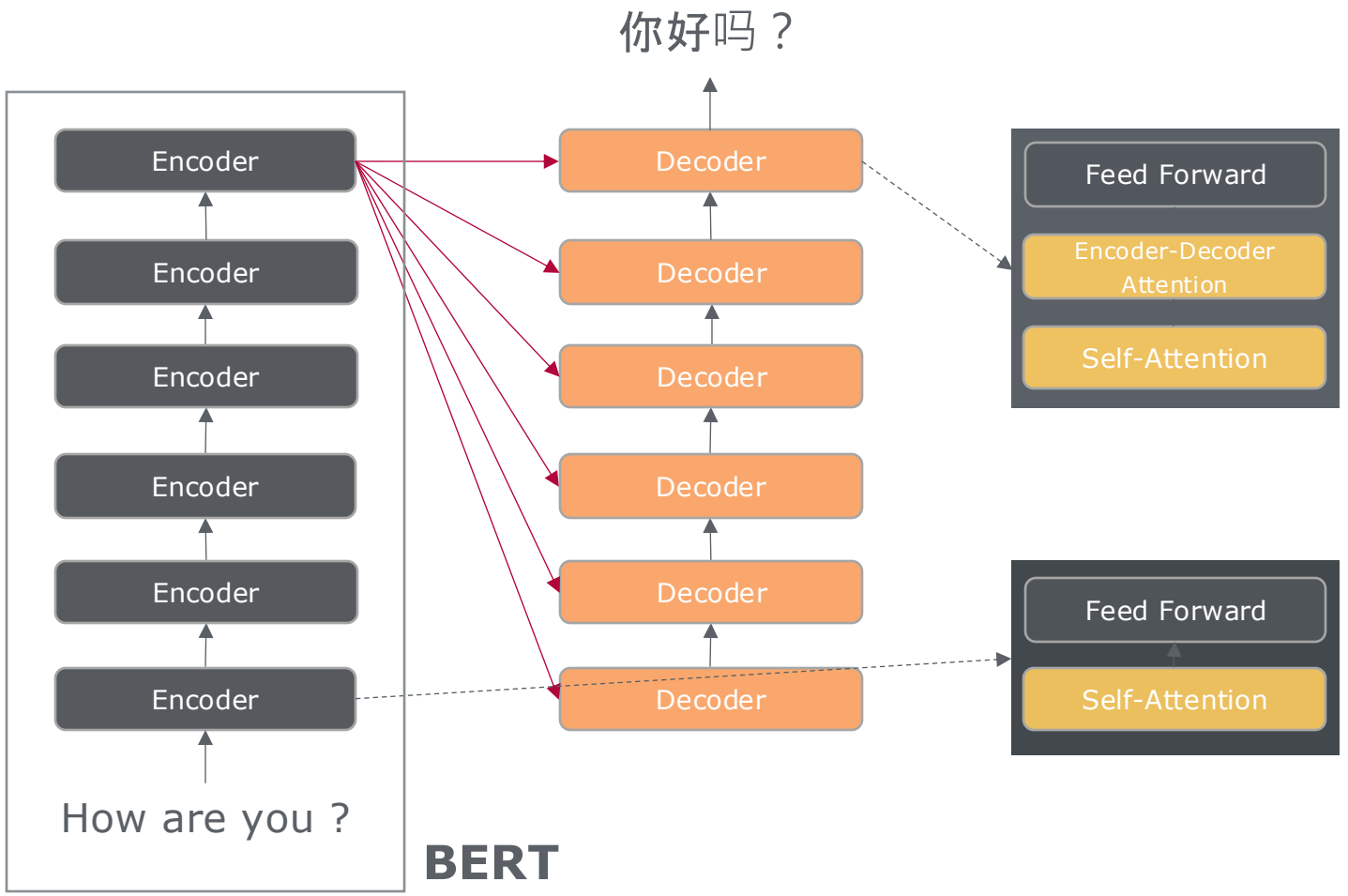
Hasso Plattner Institute

Design IT.
Create Knowledge.

www.hpi.de

Transformer, BERT and GPT





Vaswani, Ashish, et al. "Attention is all you need." NeurIPS 2017

Devlin, J et al. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT 2019*

Modern AI Revolution

2013

- AlexNet, ResNet, NASNet, ViT etc. (< 3GB)
- large-scale fully supervised dataset (ImageNet: 1.2 million images)
- training: dozens to hundreds of GPUs, days-weeks

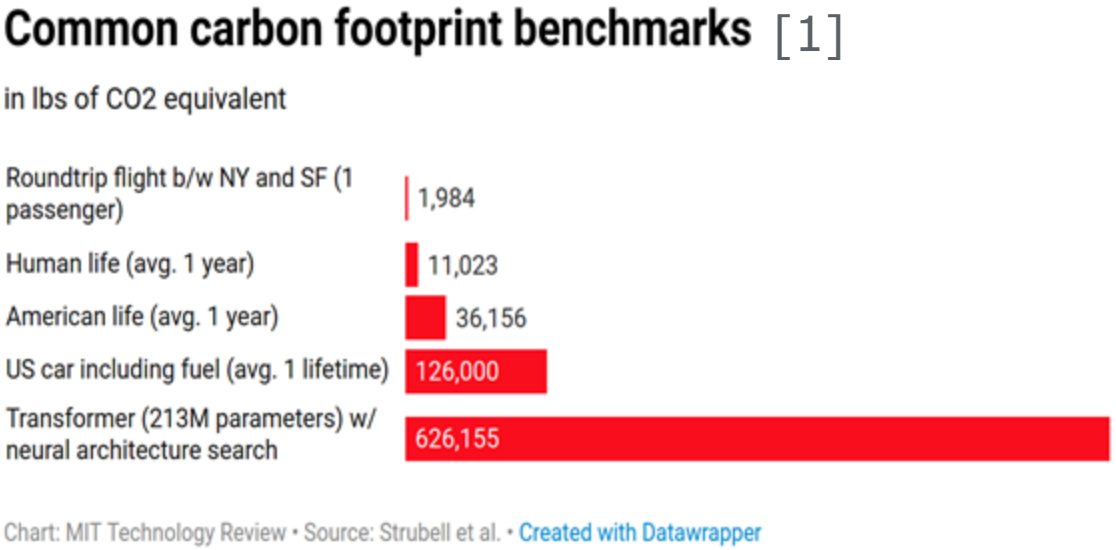
2025

- LLMs: GPT, LLaMA, DeepSeek (25-670GB)
- self-supervised learning using Internet (>3 trillion tokens)
- training: 4-20k GPUs, months

Example:

• GPT-1 (12 blocks,	125 million weights)	– June 2018	**1x**
• GPT-2 (48 blocks,	1.558 million weights)	– Februar 2019	**13x**
• GPT-3 (96 blocks,	175.000 million weights)	– June 2020	**1400x**
• GPT-4 (estimated	~1.000.000 million weights)	– March 2023	**8000x**

Deep Learning Models Spend Lots of Energy



GPT-3: 1287MW, 552 tons [2]

43 cars or 24 US families / year

GPT-4 is ~8x larger

[1] Strubell, Emma, Ananya Ganesh, and Andrew McCallum. "Energy and Policy Considerations for Deep Learning in NLP." In the 57th Annual Meeting of the Association for Computational Linguistics (ACL). Florence, Italy. July 2019

[2] David Patterson et al., "Carbon emissions and large neural network training", April 2021

Efficient AI Methods

- Knowledge distillation
  - Distills a large model (teacher) into a small model (student)
- Network pruning and dynamic network
  - Remove non-essential weights or dynamic width & depth
- Compact network designs
  - Use layer structures with less weights and operations
- Low-bit quantization
  - Quantizes 32-bit floating point params to a lower bit-width

Our selected publication 2020-2024:
[1] Supervised Knowledge May Hurt Novel Class Discovery Performance, Transaction on Machine Learning Research (TMLR) 2023
[2] SMKD: Selective Mutual Knowledge Distillation IJCNN 2023
[3] Not All Knowledge Is Created Equal: Mutual Distillation of Confident Knowledge, NeurIPS workshop 2022
[4] Flexible BERT with Width-and Depth-dynamic Inference, IJCNN 2023
[5] Boosting Bert Subnets with Neural Grafting, ICASSP 2023
[6] AsymmNet: Towards ultralight convolution neural networks using asymmetrical bottlenecks. CVPR 2021
[7] MeliusNet: Can Binary Neural Networks Achieve MobileNet-level Accuracy? WACV 2020
[8] Empirical Evaluation of Post-Training Quantization Methods for Language Tasks 2022
[9] A Study on Ultra Low-bit Compression of Generative Pretrained Transformers 2023
[10] Diode: Reinventing Binary Neural Networks Training with Sign Descent Optimization 2023
[11] BoolNet: Minimizing the Energy Consumption of Binary Neural Networks AAAI workshop 2024
[12] Towards Optimization-Friendly Binary Neural Network, TMLR 2024
[13] Enhancing Optimization Robustness in 1-bit Neural Networks through Stochastic Sign Descent. ECCV 2024