Yang Wei (韦阳)

✉ godweiyang@gmail.com · 📱 (+86)15221856016 · 🔗 https://godweiyang.com

Education

East China Normal University 2018.9 – 2021.6

M.S. Computer Science 1/105

East China Normal University 2014.9 – 2018.6

B.S. Computer Science 1/110

Research Direction

machine translation, model optimization, constituent parsing

Professional Experience

ByteDance 2021.6 – Present

AI Lab NLP algorithm engineer

One of the main developers of LightSeq. It is the first acceleration engine of Transformer-based models in the industry integrating training and inference, which won 2100+ stars in the GitHub. LightSeq supports mainstream models and training libraries, supports quantized inference, and provides rich training, exporting and inference examples. The maximal speedup of training is 3.5×, and the maximal speedup of inference is 14×. Quantized inference is further accelerated by 1.6× without performance loss. It is mainly used in Volctrans model training and deployment, and is widely used in internal and external businesses.

Link: `https://github.com/bytedance/lightseq`

ByteDance 2020.5 – 2021.6

AI Lab NLP algorithm engineer (intern)

Research Transformer compression and quantization methods. With cross-layer parameter sharing, vocabulary decomposition and quantization technologies, the Transformer model parameters are reduced to 1/20 of the original, and the performance on machine translation tasks is almost lossless.

Paper

LightSeq2: Accelerated Training for Transformer-based Models on GPUs

2nd auther SC 2022

This paper proposes LightSeq2 training acceleration engine, which supports Transformer, BERT, GPT, ViT, etc. LightSeq2 supports both PyTorch and TensorFlow, and gains maximal speedup of 3.5× compared to PyTorch.

Link: `https://arxiv.org/abs/2110.05722`

LightSeq: A High Performance Inference Library for Transformers

3rd auther NAACL 2021 Industry Track

This paper proposes LightSeq inference acceleration engine, which supports Transformer, BERT, GPT, ViT, etc. Compared with TensorFlow, the maximal speedup is 14×.

Link: `https://aclanthology.org/2021.naacl-industry.15`

A Span-based Linearization for Constituent Trees

1st auther ACL 2020

This paper proposes a linearization method of constituent trees, which reduces the decoding complexity from $O(n^3)$ to $O(n \log n)$. The decoding speed increases from 30 sentences/second to 150 sentences/second without performance loss.

Link: `https://aclanthology.org/2020.acl-main.299`

Award

Outstanding Graduates of Shanghai	2021
National Scholarships (M.S.)	2020
National Scholarships (B.S.)	2015
ACM-ICPC Invitational Shaanxi Site Gold Medal	2017
ACM-ICPC Asia Regional Programming Contest Qingdao Site Silver Medal	2016

Skill

- Programming language: Python, C++, C, CUDA.
- Deep learning framework: PyTorch, TensorFlow.

Social Link

- Blog: `https://godweiyang.com`
- GitHub: `https://github.com/godweiyang`
- Zhihu (14000+ followers): `https://www.zhihu.com/people/godweiyang`
- Wechat official account (9000+ followers): GodNLP