

韦阳

✉ godweiyang@gmail.com · 📞 15221856016 · 🌐 <https://godweiyang.com>

教育经历

华东师范大学 2018.9 – 2021.6

硕士 计算机科学与技术 1/105

华东师范大学 2014.9 – 2018.6

本科 计算机科学与技术 1/110

研究方向

模型优化、机器翻译、AIGC、成分句法分析等

工作经历

字节跳动 2021.6 – 至今

AI Lab NLP 算法工程师

- **LightSeq 训练推理加速引擎**

项目地址：<https://github.com/bytedance/lightseq>

LightSeq 核心开发者之一，GitHub 获得 2700+ star。这是业界首个集（量化）训练、（量化）推理于一体的 Transformer 系列模型加速引擎，训练最高加速 3.5 倍，推理最高加速 14 倍，量化基本无损。

- **AI 绘画**

独立支持业务方的 AI 绘画需求，利用 stable diffusion，结合 textual inversion、lora、controlnet、超分等技术，实现了 AI 绘画从训练、推理到上线部署的完整流程。

- **大模型预训练**

为大模型预训练提供基础建设支持，利用 DeepSpeed、Megatron 等大模型并行技术，稳定训练 1.3b-30b 的模型。并利用训练后量化、head 裁剪等技术对大模型进行压缩，加速推理速度。

字节跳动 2020.5 – 2021.6

AI Lab NLP 算法工程师（实习）

- **模型压缩与量化**

研究 Transformer 压缩与量化方法，利用层间参数共享、词表分解等方法，结合模型量化，将 Transformer 模型参数量压缩至原来的 1/20，在机器翻译任务上效果几乎无损，并利用 TFLite+Swift 实现 iOS 端部署。

学术成果

LightSeq2: Accelerated Training for Transformer-based Models on GPUs

第二作者 SC 2022 CCF A

论文地址：<https://dl.acm.org/doi/abs/10.5555/3571885.3571935>

提出了 LightSeq2 训练加速引擎，支持 Transformer、BERT、GPT 和 ViT 等模型结构，支持 PyTorch 和 TensorFlow，相比 PyTorch 最高提速 3.5 倍。

LightSeq: A High Performance Inference Library for Transformers

第三作者 NAACL 2021 Industry Track CCF C

论文地址：<https://aclanthology.org/2021.naacl-industry.15>

提出了 LightSeq 推理加速引擎，支持 Transformer、BERT、GPT 和 ViT 等模型结构，相比 TensorFlow 最高提速 14 倍。

A Span-based Linearization for Constituent Trees

第一作者 ACL 2020 CCF A

论文地址：<https://aclanthology.org/2020.acl-main.299>

提出了一种成分句法树的序列化表示方法，将解码复杂度从 $O(n^3)$ 降低到了 $O(n \log n)$ ，解码速度从 30 句/秒提高到了 150 句/秒，并且效果无损。

获奖荣誉

| | |
|------------------------|------|
| 上海市优秀毕业生 | 2021 |
| 国家奖学金（硕士） | 2020 |
| 国家奖学金（本科） | 2015 |
| ACM-ICPC 全国邀请赛（西安站） 金牌 | 2017 |
| ACM-ICPC 亚洲区域赛（青岛站） 银牌 | 2016 |

编程技能

- 编程语言：熟悉 Python、C++、C、CUDA。
- 深度学习框架：熟悉 PyTorch、TensorFlow。

社交链接

- 技术博客：<https://godweiyang.com>
- GitHub：<https://github.com/godweiyang>
- 知乎（18000+ 关注）：<https://www.zhihu.com/people/godweiyang>
- 公众号（10000+ 关注）：算法码上来