Yang Wei (韦阳)

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

Education

East China Normal University 2018 – Present

M.S. Computer Science and Technology *Score: 91.5/100*

East China Normal University 2014 – 2018

B.S. Computer Science and Technology *Rank: 1/110*

Research

I mainly research *machine translation*, *model compression and acceleration*, *constituent parsing* and *dependency parsing*.

Experience

ByteDance AI Lab NLP Research Intern 2020.05 – Present

Mainly research compression and acceleration of machine translation models, using model compression techniques such as parameter sharing, embedding decomposition, quantization and knowledge distillation to reduce the size of the model and finally apply it to mobile devices. Finally, the model size is compressed by 20 times and the performance is basically lossless.

Publications

A Span-based Linearization for Constituent Trees **ACL 2020**

***Yang Wei**, Yuanbin Wu and Man Lan*

We propose a novel linearization of a constituent tree, together with a new locally normalized model. Compared with global models, our model is fast and parallelizable. Different from previous local models, our linearization method is tied on the spans directly and considers more local features when performing span prediction, which is more interpretable and effective.

LightSeq: A High Performance Inference Library for Transformers NAACL 2021 Industry Track

*Xiaohui Wang, Ying Xiong, **Yang Wei**, Mingxuan Wang, Lei Li*

We propose LightSeq, a highly efficient inference library for models in the Transformer family. LightSeq includes a series of GPU optimization techniques to both streamlining the computation of Transformer layers and reducing memory footprint. LightSeq supports models trained using PyTorch and Tensorflow. Experimental results on standard machine translation benchmarks show that LightSeq achieves up to 14x speedup compared with TensorFlow and 1.4x speedup compared with FasterTransformer.

Awards

National Scholarship (M.S.) 2020
Gold Medal ACM-ICPC Invitational Shaanxi Site 2017
Special Scholarship 2017
Silver Medal ACM-ICPC Asia Regional Programming Contest Qingdao Site 2016
First Prize Scholarship 2016
Bronze Medal ACM-ICPC Asia Regional Programming Contest Shanghai Site 2015
National Scholarship 2015

Programming Skills

- Programming Language: *Python, C++, C, CUDA, Java, SQL* e.g.
- Neural Network Toolkit: *TensorFlow, PyTorch, DyNet*.

Social Links

- Blog: *https://godweiyang.com*
- GitHub: *https://github.com/godweiyang*
- Zhihu: *https://www.zhihu.com/people/godweiyang*, 13000+ like, 7900+ fans。
- WeChat Official Accounts: *GodNLP*, 4000+ fans。