YINQIAO LI

TEL: +86 17740037095 | E-MAIL: li.yin.qiao.2012@hotmail.com

EDUCATIONAL BACKGROUND

Northeastern University (Natural Language Processing Lab) **09/2018-Present** **Shenyang, China**

Major: Computer Software and Theory ***Doctoral Degree Candidate***

- **Advisor**: Prof. Jingbo Zhu
- **Co-advisor**: Prof. Tong Xiao

Northeastern University (Natural Language Processing Lab) **09/2016-06/2018** **Shenyang, China**

Major: Computer Software and Theory ***Master Degree***

- **Advisor**: Prof. Tong Xiao
- **Thesis**: *Research of Decoding Acceleration Method for Neural Machine Translation Based on Self-attention Mechanism*

Northeastern University **09/2012-06/2016** **Shenyang, China**

Major: Internet of Things Engineering ***Bachelor Degree***

- **Grades:** GPA: 3.79/5; Rank: 3/29
- **Thesis**: *Design and Implementation of Reranking System Which Based on Neural Network Language Model*

RESEARCH INTERESTS

Natural Language Processing

- Machine Learning for NLP (My recent work has been focused on using Neural Architecture Search methods to find appropriate structures for NLP tasks)
- Language Modeling
- Machine Translation

RESEARCH EXPERIENCE

Neural Architecture Search with Node and Operation Embeddings **10/2020-Present**

- Current Neural Architecture Search methods do not model the relationships among nodes and operations in the neural structures. The neglect makes the search process extremely unstable. In this work, we are trying to apply the attention mechanism on the node and operation embeddings to catch the correlations of different parts in the model.
- We are going to try this method on language modeling tasks and apply the architecture to other NLP tasks such as machine translation (MT) and named entity recognition (NER).

Learning Architectures from an Extended Search Space for Language Modeling **09/2019-03/2020**

- Neural architecture search has advanced significantly in recent years but most NAS systems restrict search to learning architectures of a recurrent or convolutional cell. In this work, we extend the search space of NAS. In particular, we present a general approach to learn both intra-cell and inter-cell architectures (call it ESS).
- For a better search result, we design a joint learning method to perform intra-cell and inter-cell NAS simultaneously. We implement our model in a differentiable architecture search system.
- For recurrent neural language modeling, ESS outperforms a strong baseline significantly on the PTB and WikiText data, with a new state-of-the-art on PTB. Moreover, the learned architectures show good transferability to other systems. E.g., they improve state-of-the-art systems on the CoNLL and WNUT NER tasks and CoNLL chunking task, indicating a promising line of research on large-scale pre-learned architectures.

Sharing Attention Weights for Fast Transformer **02/2018-04/2019**

- Recently, the Transformer machine translation system has shown strong results by stacking attention layers on both the source and target-language sides. But the inference of this model is slow due to the heavy use of dot-product attention in auto-regressive decoding. In this work we speed up Transformer via a fast and lightweight attention model. More specifically, we share attention weights in adjacent layers and enable the efficient re-use of hidden states in a vertical manner. Moreover, the sharing policy can be jointly learned with the MT model.
- We test our approach on ten WMT and NIST OpenMT tasks. Experimental results show that it yields an average of 1.3X speed-up (with almost no decrease in BLEU) on top of a state-of-the-art implementation that has already adopted a cache for

fast inference. Also, our approach obtains a 1.8X speed-up when it works with the AAN model. This is even 16 times faster than the baseline with no use of the attention cache.

Analysis of Data Parallel Methods in Training Neural Language Models via Multiple GPUs 05/2017-10/2017

- In order to speedup the process of training neural models, we apply data parallel method to train them on multiple devices (such as GPUs). However, due to the frequent data transmission between multiple devices, we cannot get the ideal acceleration effect easily. In this work, we analyse the performance of all-reduce and sampling-based gradient update strategy in the data transmission period.
- The experiments show all-reduce and sampling-based gradient update strategy can achieve 25% and 41% speedup on NVIDIA TITAN X. Moreover, we also discuss on the the applicability of data parallel method and the influence of different hardware device connection modes on transmission speed.

PUBLICATIONS

- Yinqiao Li, Chi Hu, Yuhao Zhang, Nuo Xu, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, Changliang Li. 2020. *Learning Architectures from an Extended Search Space for Language Modeling*. In Proc. of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), Seattle, USA.
- Chi Hu, Bei Li, Yinqiao Li, Ye Lin, Yanyang Li, Chenglong Wang, Tong Xiao, Jingbo Zhu. 2020. *The NiuTrans System for WNGT 2020 Efficiency Task*. In Proc. of the Fourth Workshop on Neural Generation and Translation (WNGT), Seattle, USA.
- Yuhao Zhang, Ziyang Wang, Runzhe Cao, Binghao Wei, Weiqiao Shan, Shuhan Zhou, Abudurexiti Reheman, Tao Zhou, Xin Zeng, Laohu Wang, Xiaoqian Liu, Xunjuan Zhou, Yongyu Mu, Jingnan Zhang, Yinqiao Li, Bei Li, Tong Xiao and Jingbo Zhu. 2020. *The NiuTrans Machine Translation Systems for WMT20*. In Proc. of the Fifth Conference on Machine Translation (WMT), online.
- Tong Xiao, Yinqiao Li, Jingbo Zhu, Zhengtao Yu and Tongran Liu. 2019. *Sharing Attention Weights for Fast Transformer*. In Proc. of the 28th International Joint Conference on Artificial Intelligence (IJCAI), Macao, China.
- Bei Li, Yinqiao Li, Chen Xu, Ye Lin, Jiqiang Liu, Hui Liu, Ziyang Wang, Yuhao Zhang, Nuo Xu, Zeyang Wang, Kai Feng, Hexuan Chen, Tengbo Liu, Yanyang Li, Qiang Wang, Tong Xiao and Jingbo Zhu. 2019. *The NiuTrans Machine Translation Systems for WMT19*. In Proc. of the Fourth Conference on Machine Translation (WMT), Florence, Italy.
- Yuhao Zhang, Nuo Xu, Yinqiao Li, Tong Xiao and Jingbo Zhu. 2019. *Research on inference acceleration method of Neural Machine Translation system based on coarse2fine*. In Proc. of the 15th China Conference on Machine Translation (CCMT), Nanchang, China.
- Nuo Xu, Yinqiao Li, Chen Xu, Yanyang Li, Bei Li, Tong Xiao and Jingbo Zhu. 2019. *Analysis of Back-translation Methods for Low-Resource Neural Machine Translation*. In Proc. of the 8th CCF International Conference on Natural Language Processing and Chinese Computing (NLPCC), Dunhuang, China.
- Yanyang Li, Tong Xiao, Yinqiao Li, Qiang Wang, Changming Xu and Jingbo Zhu. 2018. *A Simple and Effective Approach to Coverage-Aware Neural Machine Translation*. In Proc. of the fifty-sixth Annual Meeting of the Association for Computational Linguistics (ACL), Melbourne, Australia.
- Qiang Wang, Fuxue Li, Tong Xiao, Yanyang Li, Yinqiao Li and Jingbo Zhu. 2018. *Multi-layer Representation Fusion for Neural Machine Translation*. In Proc. of the 27th International Conference on Computational Linguistics (COLING), Santa Fe, New Mexico, USA.
- Qiang Wang, Bei Li, Jiqiang Liu, Bojian Jiang, Zheyang Zhang, Yinqiao Li, Ye Lin, Tong Xiao, Jingbo Zhu. 2018. *The NiuTrans Machine Translation System for WMT18*. In Proc. of the Third Conference on Machine Translation (WMT), Belgium, Brussels.
- Qiang Wang, Bei Li, Jiqiang Liu, Bojian Jiang, Zheyang Zhang, Yinqiao Li, Ye Lin, Tong Xiao and Jingbo Zhu. 2018. *Towards Building a Strong Transformer Neural Machine Translation System*. In Proc. of the 14th China Workshop on Machine Translation (CWMT), Wuyishan, China.
- Yufan Jiang, Bei Li, Ye Lin, Yinqiao Li, Tong Xiao and Jingbo Zhu. 2018. *Learning Neuron Connections for Language Models*. In Proc. of the 14th China Workshop on Machine Translation (CWMT), Wuyishan, China.
- Yinqiao Li, Ambyer Han, Le Bo, Tong Xiao, Jingbo Zhu, Li Zhang. 2017. *Analysis of Data Parallel Methods in Training Neural Language Models via Multiple GPUs*. In Proc. of the 13th China Workshop on Machine Translation (CWMT), Dalian, China.

CV last modified on Dec 11, 2020.

EVALUATION TASKS

- WMT19 Kazakh-English MT track - 1st place (auto-evaluation) 01/2019-04/2019
- WMT19 English-Kazakh MT track - 1st place (auto-evaluation) 01/2019-04/2019

COMMUNITY ACTIVITIES

- PC members of AAAI 2020 09/2020-11/2020
- PC members of CCL 2019 06/2019-07/2019
- Student member of Youth Working Committee of CIPS 10/2019-Present
- Oral presentation on the Doctoral Forum of the 15th China Conference on Machine Translation 10/2020
- Oral presentation on the 58th Annual Meeting of the Association for Computational Linguistics 07/2020
- Oral presentation on the 28th International Joint Conference on Artificial Intelligence 08/2019
- Poster presentation on the 13th China Workshop on Machine Translation 09/2017

SOFTWARE

NiuTensor Open-source Toolkit

- NiuTensor is an open-source toolkit developed by a joint team from NLP Lab. at Northeastern University and the NiuTrans Team. It provides tensor utilities to create and train neural networks.
- I maintain the NiuTensor project now and work on designing and developing the memory pool, data structures of tensor.
- https://github.com/NiuTrans/NiuTensor

HONORS AND AWARDS

- The First Class Scholarship for Outstanding Students of Northeastern University (2016, 2017,2018, 2019, 2020)
- Outstanding Graduate Students of Shenyang Province (2020)
- CASC Scholarship (2020)
- Presidential Scholarship of Northeastern University (2016)

PERSONAL INFORMATION

- Citizen of P. R. China.
- Born November 1993, Anshan, China.
- Languages: Mandarin (native); English (read/write/speak).
- Programming Languages: C/C++, Python, Shell.

CV last modified on Dec 11, 2020.