# Dr. YINQIAO LI

TEL: +852 53132401 / +86 17740037095 | E-MAIL: li.yin.qiao.2012@hotmail.com

## EDUCATIONAL BACKGROUND

**Northeastern University** (Natural Language Processing Lab) **09/2018-07/2023** **Shenyang, China**

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

- **Advisor**: Prof. Jingbo Zhu
- **Co-advisor:** Prof. Tong Xiao
- **Thesis**: *Research on Neural Architecture Search for Language Modeling*

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

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

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

## WORK EXPERIENCE

**City University of Hong Kong** **08/2023-present** **Hong Kong SAR**

***Department: Department of Computer Science*** ***Postdoc Researcher***

- **PI**: Prof. Linqi Song
- **Research Directions**: Natural Language Processing and FinTech

## INTERNSHIP EXPERIENCE

**Tencent Holdings Ltd.** **06/2021-07/2022** **Beijing, China**

***Group: Platform & Content Group (PCG)*** ***Research Intern***

- **Mentor**: Qiaozhi He
- **Research**: *AuMLM: An Auto-regressive Training Approach to Masked Language Modeling*

## RESEARCH INTERESTS

**Natural Language Processing**

- Language Modeling
- Pre-training
- Neural Architecture Search for NLP tasks
- Machine Translation

## RESEARCH EXPERIENCE

**LoRAN: Improved Low-Rank Adaptation by a Non-Linear Transformation** **11/2023-06/2024**

- In this paper, we study parameter-efficient fine-tuning methods for large pre-trained models. Specifically, we improve LoRA approaches to alleviate the performance loss from the constrained adapter by introducing a non-linear transformation (call it LoRAN). For a better adaptation, we also design a new non-linear function to appropriately fit the accumulated weight updates.

- We test our method in multiple advanced large language models. Experimental results show that our LoRAN significantly outperforms a strong baseline on SAMSum and 20 Newsgroups tasks. Moreover, when a lower rank is applied, our approach even yields a 1.95-point improvement in the classification task.

**Learning Reliable Neural Networks with Distributed Architecture Representations** **10/2020-12/2022**

- Most NAS systems are unreliable due to the architecture gap brought by discrete representations of atomic architectures. In this work, we improve the performance and robustness of NAS var narrowing the gap between architecture representations. More specifically, we apply a general contraction mapping to model neural networks with distributed representations (call it ArchDAR).

- For a better search result, we present a joint learning approach to integrating distributed representations with advanced architecture search methods.
- We implement our ArchDAR in a differentiable architecture search model and test learned architectures on the language modeling task. On the PTB data, it outperforms a strong baseline significantly by 1.8 perplexity scores. Also, the search process with distributed representations is more stable which yields a faster structural convergence when it works with the DARTS model.

**AuMLM: An Auto-regressive Training Approach to Masked Language Modeling** **06/2021-07/2022**

- Traditional masked language modeling systems are forced to predict masked words simultaneously. In this work, we enable models to predict masked tokens in order. More specifically, we determine the order of mask learning according to the dependencies among masked words and sequentially produce the masked tokens in an auto-regressive manner (call it AuMLM).
- Moreover, we design a coverage-based masking approach to generating more efficient mask strategies for fast training.
- We apply our method to BERT and test it on the GLUE tasks. Experimental results show that it yields an average of 1.15 point improvement.

**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 learning 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 speed up the process of training neural models, we apply the 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 the all-reduce algorithm and the sampling-based gradient update strategy in the data transmission period.
- The experiments show the all-reduce algorithm and the sampling-based gradient update strategy can achieve 25% and 41% speedup on NVIDIA TITAN X. Moreover, we also discuss the applicability of the data parallel method and the influence of different hardware device connection modes on transmission speed.

## PUBLICATIONS

- Yinqiao Li, Linqi Song, Hanxu Hou. 2024. *LoRAN: Improved Low-Rank Adaptation by a Non-Linear Transformation*. In Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA. (Just Accepted)
- 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.
- Tong Xiao (my advisor), 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.

- Yinqiao Li, Runzhe Cao, Qiaozhi He, Tong Xiao, Jingbo Zhu. 2023. *Learning Reliable Neural Networks with Distributed Architecture Representations*. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP).
- 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.
- 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.
- 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.
- 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.
- Qi Chen, Oi Yee Kwong, Yinqiao Li, Tong Xiao, and Jingbo Zhu. 2022. *Coarse-to-Fine Output Predictions for Efficient Decoding in Neural Machine Translation*. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP).
- Quan Du, Nuo Xu, Yinqiao Li, Tong Xiao, Jingbo Zhu. 2021. *Topology-Sensitive Neural Architecture Search for Language Modeling*. IEEE Access.
- 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.
- 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.
- 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.
- Chi Hu, Chenglong Wang, Xiangnan Ma, Xia Meng, Yinqiao Li, Tong Xiao, Jingbo Zhu, Changliang Li. 2021. *RankNAS: Efficient Neural Architecture Search by Pairwise Ranking*. In Proc. of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), Punta Cana, Dominican Republic.
- 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.
- Junhao Ruan, Abudukeyumu Abudula, Xinyu Liu, Bei Li, Yinqiao Li, Chenglong Wang, Yuchun Fan, Yuan Ge, Tong Xiao, and Jingbo Zhu. 2024. *NDP: Next Distribution Prediction as a More Broad Target*. arXiv preprint arXiv:2408.17377. (Under Review)
- Yongyu Mu, Abudurexiti Reheman, Zhiquan Cao, Yuchun Fan, Bei Li, Yinqiao Li, Tong Xiao, Chunliang Zhang, and Jingbo Zhu. 2023. *Augmenting Large Language Model Translators via Translation Memories*. In Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada.
- 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.
- 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.

## EVALUATION TASKS

- The International Exhibition of Inventions Geneva (IEIG 2024) - Two Silver Awards 04/2024

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

## COMMUNITY ACTIVITIES

- Publicity Chair of AICIT 2024 07/2024
- Reviewer of WiNLP 2024 07/2024-09/2024
- Reviewer of ARR June 2024 06/2024-08/2024
- Reviewer of ARR April 2024 05/2024-06/2024
- Reviwer of AAAI 2023 08/2022-11/2022
- Reviwer of IALP 2022 08/2022-09/2022
- Reviwer of ICML 2022 03/2022-05/2022
- Reviwer of ACL 2022 02/2022-03/2022
- Reviwer of AAAI 2022 09/2021-11/2021
- Reviwer of ACL 2021 02/2021-04/2021
- Reviwer of AAAI 2021 09/2020-11/2020
- Reviwer of CCL 2019 06/2019-07/2019
- Member of Youth Working Committee of CIPS 09/2023-Present
- Tutorial representation on Workshop of the 18th China Conference on Machine Translation 08/2022
- Oral presentation on the PhD Debate of AI Time 03/2021
- 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 International Exhibition of Inventions Geneva (2024) 04/2024
- Poster presentation on Young Scholar Symposium on Natural Language Processing (2019) 05/2019
- 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 maintained the NiuTensor project and worked 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 (2020)
- CASC Scholarship (2020)
- Presidential Scholarship of Northeastern University (2016)

## PERSONAL INFORMATION

- Citizen of People's Republic of China.
- Live in Hong Kong SAR now.
- Born in November 1993, Anshan, Mainland China.
- Languages: Mandarin (native); English (read/write/listen/speak); Cantonese (listen/speak a little).
- Programming Languages: Python, Shell, C/C++.

---

*CV last modified on September 25, 2024.*