Heming Xia

https://hemingkx.github.io/

Email : hemingkx@gmail.com
Mobile : +86-188-0138-9565

Education

The Hong Kong Polytechnic University **Jan. 2024 –**
Ph.D. in Computer Science *Hong Kong, China*
Advisor: Prof. Wenjie Li

Peking University **Sep. 2020 – Jul. 2023**
Master in Software Engineering *Beijing, China*
Advisor: Prof. Zhifang Sui

Peking University **Sep. 2016 – Jul. 2020**
B.S. in Physics (Department of Astronomy) *Beijing, China*
Advisor: Asst. Prof. Lijing Shao

Papers

* indicates equal contribution.

- **Unlocking Efficiency in Large Language Model Inference: A Comprehensive Survey of Speculative Decoding**
  Heming Xia, Zhe Yang, Qingxiu Dong, Peiyi Wang, Yongqi Li, Tao Ge, Tianyu Liu, Wenjie Li, Zhifang Sui
  *Submitted to The 62nd Annual Meeting of the Association for Computational Linguistics. **ACL 2024 Submission**.*
- **ImageNetVC: Zero- and Few-Shot Visual Commonsense Evaluation on 1000 ImageNet Categories**
  Heming Xia*, Qingxiu Dong*, Lei Li, Jingjing Xu, Tianyu Liu, Ziwei Qin, Zhifang Sui
  *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. **EMNLP 2023 (Findings)**.*
- **Bi-Drop: Enhancing Fine-tuning Generalization via Synchronous sub-net Estimation and Optimization**
  Shoujie Tong*, **Heming Xia***, Damai Dai, Runxin Xu, Tianyu Liu, Binghuai Lin, Yunbo Cao, Zhifang Sui
  *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. **EMNLP 2023 (Findings)**.*
- **Speculative Decoding: Exploiting Speculative Execution for Accelerating Seq2seq Generation**
  Heming Xia*, Tao Ge*, Peiyi Wang, Si-Qing Chen, Furu Wei, Zhifang Sui
  *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. **EMNLP 2023 (Findings)**.*
- **Enhancing Continual Relation Extraction via Classifier Decomposition**
  Heming Xia, Peiyi Wang, Tianyu Liu, Binghuai Lin, Yunbo Cao, Zhifang Sui
  *The 61st Annual Meeting of the Association for Computational Linguistics. **ACL 2023 (Findings, Short Paper)**.*
- **Lossless Acceleration for Seq2seq Generation with Aggressive Decoding**
  Tao Ge, **Heming Xia***, Xin Sun*, Si-Qing Chen, Furu Wei
  Microsoft Research Technical Report.
- **Premise-based Multimodal Reasoning: Conditional Inference on Joint Textual and Visual Clues**
  Qingxiu Dong*, Ziwei Qin*, **Heming Xia**, Tian Feng, Shoujie Tong, Haoran Meng, Lin Xu, Zhongyu Wei, Weidong Zhan, Baobao Chang, Sujian Li, Tianyu Liu, Zhifang Sui
  *The 60th Annual Meeting of the Association for Computational Linguistics. **ACL 2022**.*
- **Improved deep learning techniques in gravitational-wave data analysis**
  Heming Xia, Lijing Shao, Junjie Zhao, Zhoujian Cao
  ***Physical Review D 103 (2021), 024040**.*

Research Experiences

A Comprehensive Survey and Unified Benchmark for Speculative Decoding **Oct. 2023 – Feb. 2024**
Advisor: Prof. Wenjie Li, Department of Computing, The Hong Kong Polytechnic University *Hong Kong, China*

- Made the first attempt to present a comprehensive survey of Speculative Decoding, including a formal definition and formulation of Speculative Decoding as well as a systematic categorization of leading techniques in current research.
- Developed Spec-Bench, an extensive benchmark designed for assessing Speculative Decoding methods across diverse scenarios, followed by a comparative evaluation of open-source methods under third-party testing conditions.

ImageNetVC: Zero- and Few-Shot Visual Commonsense Evaluation **Oct. 2022 – Jun. 2023**
Advisor: Prof. Zhifang Sui, Peking University, and Dr. Jingjing Xu, Shanghai AI Lab *Beijing, China*

- Constructed a unified visual commonsense benchmark, ImageNetVC, comprising 4K human-annotated, fine-grained question-answer pairs with sufficient sources of images, encompassing 1K ImageNet categories.

- Conducted a comprehensive visual commonsense evaluation of existing large language models (LLMs) as well as their visually-augmented variants, across various model families, model scales and etc.
- Highlighted several experimental findings, such as LLaMA outperforms all model families in visual commonsense understanding and in-context learning plays a vital role for LLMs to understand visual commonsense tasks.

Enhancing Continual Relation Extraction via Classifier Decomposition **Jun. 2022 – Oct. 2022**
Advisor: Dr. Tianyu Liu, Liu Group, Tencent Cloud AI *Beijing, China*

- Analyzed two typical biases when models first learn new relations in Continual Relation Extraction (CRE): *classifier bias* and *representation bias*, which causes the previous knowledge that the model learned to be shaded.
- Proposed a classifier decomposition framework that splits the last FFN layer into separated previous and current classifiers, so as to maintain previous knowledge and encourage the model to learn more robust representations.

Speculative Decoding: Lossless Speedup of Seq2seq Generation **Nov. 2021 – Jun. 2022**
Advisor: Dr. Tao Ge, Natural Language Computing Group, Microsoft Research Asia *Beijing, China*

- Proposed Speculative Decoding (SpecDec), a general decoding paradigm for efficient sequence-to-sequence generation, which utilizes a high-efficiency drafter to speculate multiple future decoding steps of an autoregressive model.
- Proposed a specially designed non-autoregressive drafter for highly accurate speculation as well as an advanced verification strategy, which aims to make better use of the speculation results and guarantee high-quality outputs.
- Conducted extensive experiments on multiple seq2seq tasks and across various model architectures, showing that SpecDec achieves around 3x-5x speedup over autoregressive counterparts with comparable performances.

Premise-based Multimodal Reasoning on Joint Textual and Visual Clues **Mar. 2021 – Nov. 2021**
Advisor: Prof. Zhifang Sui, MOE Lab, Peking University *Beijing, China*

- Designed a benchmark termed Premise-based Multi-modal Reasoning (PMR), containing 15K manually annotated questions with images, where a textual premise is the background presumption on each source image.
- Benchmarked various state-of-the-art pretrained multi-modal inference models on PMR and conducted comprehensive experimental analyses to showcase the utility of the dataset.

Improved Deep Learning Techniques in Gravitational-wave Data Analysis **Mar. 2020 – Nov. 2020**
Advisor: Asst. Prof. Lijing Shao, Kavli Institute for Astronomy and Astrophysics, Peking University *Beijing, China*

- Introduced multiple deep-learning optimization techniques based on the task of gravitational-wave (GW) detection of binary black holes, such as batch normalization and dropout to CNN models.
- Investigated the generalization ability of CNN models on different parameter ranges of GW signals and pointed out that CNN models are robust to the variation of the parameter range of the GW waveform.

Open-Source Projects

- **Spec-Bench for Speculative Decoding (Python, PyTorch)**: Developed a comprehensive benchmark and unified evaluation platform for assessing leading Speculative Decoding methods across diverse application scenarios.
- **Seq2Seq Inference Acceleration with Speculative Decoding (Python, Fairseq)**: Released all the codes and checkpoints utilized in Speculative Decoding, which achieves 3x-5x inference speedup with only 300MiB of extra memory cost.
- **Deep Learning Toolkits for Gravitational-wave Analysis (Python, PyTorch)**: Developed a deep learning toolkit for gravitational-wave (GW) data analysis, which supports GW data generation, visualization and classification.

Services and Membership

- Reviewer: NeurIPS 2022, AACL 2022, AACL 2023, ARR (Feb, Apr-2024)
- Member: Department of Liaison, the Student Union of Peking University, 2018 - 2021

Technical Skills

Languages: Python, Latex, C/C++, Java, Shell, MATLAB, HTML/CSS
Developer Tools: VS Code, PyCharm, Git, Docker, Linux, Vim, Eclipse
Libraries/Frameworks: PyTorch, Transformers, Fairseq, TensorFlow, PyTorch-Lightning, spaCy, NumPy, WordPress

Honors and Awards

- Model Student of Social Work, Peking University 2022
- Merit Student, Peking University 2021
- Scholarship of National Astronomical Observatory, Chinese Academy of Sciences 2019
- Outstanding Volunteer Award of Volunteer and Cultural Program in Thailand, The Green Lion 2017
- Merit Student, Henan Province, China 2016