Xiaoyue Xu

(+86) 18800187931 | xiaoyue.xu.me@gmail.com | xiaoyue2002.github.io

Education

Tsinghua University, *Undergraduate* — Beijing, China
B.Sc. in Computer Science and Technology — *2021 – Present*
B.Sc. in Department of Automation — *2020 – 2021*

▷ **Cumulative GPA: 3.81/4.00**

▷ Selected Courses of A & A+: Programming and Training, Introduction to Complex Analysis, Artificial Neural Networks, Media Computing, Database Special Topic Training

Publications

1. **Stress-Testing Long-Context Language Models with Lifelong ICL and Task Haystack** [pdf].
   Xiaoyue Xu*, Qinyuan Ye*, Xiang Ren
   NeurIPS Dataset&Benchmark Track, 2024

2. **Boosting Inference Efficiency: Unleashing the Power of Parameter-Shared Pre-trained Language Model.** [pdf]
   Weize Chen*, **Xiaoyue Xu***, Xu Han, Yankai Lin, Ruobing Xie, Zhiyuan Liu, Maosong Sun, Jie Zhou.
   Findings of EMNLP, 2023

3. **Learning Heterogeneous Mixture of Hash Experts for Highly Scalable Neural Radiance Fields.**
   Zhenxing Mi, **Xiaoyue Xu**, Dan Xu.
   Technical report.

** indicates equal contribution*

Research Experience

Summer Research Intern, INK Lab, USC. Advisor: Prof. Xiang Ren — *2024 – Present*

▷ **Stress-testing Long-context Language Models with Lifelong ICL**

- Proposed Lifelong ICL, a novel problem setting that challenges long-context language models to learn language tasks sequentially through in-context learning. Developed a evaluation suite called Task Haystack to diagnose and benchmark long-context LMs.
- Demonstrated that SOTA long-context LMs struggle in our setting, with GPT-4o failing 15% of the cases on average, and open-source models lagging behind by a large margin. Performed detailed controlled analyses, uncovering models' susceptibilities to recency bias, distractability, and inefficiencies in true context utilization.
- Accpeted by Neurips 2024 D&B Track.

Undergraduate Research Intern, THUNLP, THU. Advisor: Prof. Zhiyuan Liu — *2022 – 2023*

▷ **Efficient Inference for Parameter-sharing PLMs**

- Developed a straightforward technique to significantly improve inference efficiency in parameter-sharing PLMs by utilizing an ODE perspective. This approach allows for a reduction in hidden state update iterations by increasing the step size.
- Proposed a novel pre-training strategy, which further expedited the inference process of models with fully or partially shared parameters, retaining 99% performance at around 1.5x acceleration.
- Accpeted by EMNLP 2023 findings.

▷ **Addressing Long-tail Distribution Problem via Hypernetwork**

- Developed a hypernetwork-based solution to tackle long-tail data distribution challenges in NLP tasks such as relation extraction, effectively bridging the gap between low-resource and high-resource scenarios.

- Engineered a novel approach by modeling the training process as a stochastic differential equation (SDE) to simulate parameter trajectories, achieving optimal few-shot learning performance.

Visiting Research Intern, HKUST. Advisor: Prof. Dan Xu *2023*

▷ **Transferable Monocular Depth Estimation**

- Combined relative depth pre-training and metric depth fine-tuning to enhance model generalization across diverse environmental conditions. Experimented with incorporating prompting methods to improve zero-shot performance.

▷ **Heterogenerous MoE for Scalable Large Scale NeRF**

- Contributed to experiment design and writing of the research paper, which proposed a scalable and efficient large-scale NeRF framework by employing heterogeneous models with mixture of experts method.

Awards&Honors

Academic Excellence Scholarship, Dept. of CST, Tsinghua University (2021)

Skills

English Skills

▷ TOEFL (Best) 110/120 (Reading 30, Listening 30, Speaking 23, Writing 27).

▷ GRE Verbal 158/170, Quant 170/170, Analytical Writing 3.5/6.

Technical Skills

▷ Proficient in C/C++, Python (PyTorch), LaTeX, Linux.