Xinyu Zhao

xinyuzhao@stu.pku.edu.cn

Education

Peking University   **Sep. 2021 – Jun. 2024**
MS in Computer Technology - GPA: 3.52/4   *Beijing, China*

Huangzhong University of Science and Technology   **Sep. 2017 – Jun. 2021**
BA in Translation and Interpreting - GPA: 3.99/4   *Wuhan, China*

Huangzhong University of Science and Technology   **Sep. 2018 – Jun. 2020**
Minor in Computer Science and Technology   *Wuhan, China*

Publications

Sparse MoE with Language-Guided Routing for Multilingual Machine Translation

ICLR 2024   *First Author*

- Current Sparse Mixture-of-Experts (SMoE) solutions neglect the language features in multilingual machine translation. Therefore, we propose `Lingual-SMoE` with language-guided routing based on language hierarchy and difficulty.
- **Language hierarchy:** incorporate linguistic knowledge in routing to allocate experts into different group.
- **Language difficulty:** adjust the number of activated experts for each target language in an automatic manner, according to their data resource difficulties, mitigating the potential over-/under-fitting issues.
- Sufficient experiments are conducted with various network architectures, validating the superior performance of our proposals. For instance, `Lingual-SMoE` outperforms its dense counterpart by over 5% BLEU scores on OPUS-100.

Towards Building Reliable Language Models with Sparse Mixture-of-Experts

ICML 2024   *Co-First Author*

- MoE models sometimes underperform when transferred to new domains. Therefore,
- We design **MoE-RBench**, examines whether a MoE model matches with similar dense networks from multiple reliability dimensions, including safety and hallucination, adversarial robustness, and OOD Robustness.
- Our empirical observations suggest that with appropriate hyperparameters, training recipes, and inference techniques, we can build the MoE model more reliably than the dense LLM.
- Our study also reveals that MoE models are on par with dense models and further benefit from existing instruction tuning and inference techniques.

Enhancing Dynamic Image Advertising with Vision-Language Pretraining

SIGIR 2023   *Co-first Author*

- Query-image pairing in multimodal sponsored search consists of image retrieval and relevance modeling. Previous separate optimizing method leads to suboptimal result, we unify them by Vision-Language Pretraining.
- **Base model:** train a CLIP-based model with 1B multi-domain image-text pairs to learn general representation, outperforming SOTA in Recall@10 by 6.9% on business datasets.
- **Relevance model:** further train the base model on advertising data, achieving a 4.62% increase in AUC compared to separately trained model.
- **Retrieval model:** unify retrieval and relevance modeling with multitask training, improve Recall@10 and Relscore@10 by 19% and 1% compared to the base model.
- Our approach has been implemented in Baidu Search Ads, obtaining 1.04% and 1.87% increase in CPM and CTR.

Research Experiences

Internship: Baidu Search Ads | *Multimodal Deep Learning*   **Nov. 2022 – Jun. 2023**

- Scaling the vision-language model to 1B parameter is computationally expensive, we tackle the problem from training and data aspects.
- **Training efficiency:** partition optimizer states and gradients using ZeRO with Deepspeed, and randomly drop image patches during training.
- **Data quality:** generate additional texts with image captioning model, then use our two-tower model to filter low-quality image-text pairs.

Awards

Outstanding Graduate of Huangzhong University of Science and Technology   **Jun. 2021**

National Scholarship   **Oct. 2019**