Shuai Xie

+86 17788584418 | shuaixie@zju.edu.cn
GitHub: https://github.com/Shuai-Xie | Personal Page: https://shuai-xie.github.io
Google Scholar: https://scholar.google.com/citations?user=DzyIoEMAAAAJ&hl=zh-CN

Education

Zhejiang University — Hangzhou, Zhejiang
M.Sc. in Computer Science and Technology. Tutor: Professor Mingli Song. — 2018.9 – 2021.3

Central South University — Changsha, Hunan
B.Sc. in Computer Science and Technology. GPA: 3.63/4. CET6 510. — 2014.9 – 2018.6

Skills

Programming: Python, C, C++, CUDA, Java, SQL, Git, Linux, Docker, Kubernetes, Go
Frameworks: PyTorch, FasterTransformer, TensorRT, Triton, Flask, OpenCV, NumPy, Pandas, Matplotlib
Languages: Chinese, English

Work Experience

JD Explore Academy - Efficient LLM Inference — Beijing, China
Algorithm Engineer — 2023.3 – Present

- **Production-level LLM Inference Solution**: (1) Accelerated JD ChatRhino LLM with Int4 quantization to achieve an impressive 11× inference speedup and a remarkable 60% reduction in GPU memory usage compared to PyTorch, while maintaining over 95% accuracy with its Float16 counterpart. (2) Deployed the accelerated FasterTransformer (FT) model on A100 or V100 GPUs using Triton Inference Server, enabling support for streaming inference and dynamic batching. (3) Achieved competitive Time-To-First-Byte (TTFB) results when benchmarked against industry-leading language models like ChatGLM and ChatGPT. (4) Developed a semi-automated workflow to assess the model's quality in an agile manner, which generates QAs using carefully selected samples and presents them to experts in a user study fashion. This simulation of an online environment allows us to quickly evaluate the model performance and obtain timely user feedback. https://chatrhino.ttfb/
- **Post-Training Compression**: Implemented post-training weight-only quantization (PTQ) and semi-structured sparsity (PTS) techniques to compress the ChatRhino LLM. (1) PTQ accelerates the inference by reducing the weight movement time between HBM and SRAM. Benchmark results demonstrate that FT-Int4 achieves a 60% reduction in GPU memory usage and a 2× inference speedup compared to FT-Float16. (2) Additionally, as an orthogonal acceleration approach, PTS reduces computation demands by 50% using a 2:4 semi-structured sparsity format. Benchmark results indicate that FT-Float16 with sparsity achieves a 30% reduction in GPU memory usage and a 1.2× inference speedup. This method is gaining momentum, driven by the higher sparsity techniques and the rapid advancement of next-generation computational hardware such as H100 and Moffett S30.
- **Automatic Compression Framework**: Amalgamated the aforementioned acceleration features into an elegant and cohesive solution. This initiative seeks to develop an automatic compression framework capable of producing compressed models while adhering to stringent hardware constraints and maintaining nearly optimal model accuracy.

JD Explore Academy - OmniForce AutoML Platform — Beijing, China
Algorithm Engineer — 2021.3 – Present

- **On Human-Centered, Large Model Empowered and Cloud-Edge Collaborative AutoML Platform**: (1) Developed a large-scale cloud-native AutoML system on top of the Kubeflow and KServe machine learning toolkit, streamlining the AI development process into five essential steps: multi-objectives, data preparation, feature engineering, model training, and deployment. (2) Created two AutoML pipelines, named HPO and NoCode, which orchestrates the aforementioned five steps and seamlessly integrates various cloud-native workloads like PytorchJob and InferenceService to provide a ready-to-run blueprint solution. (3) This highly scalable design effectively harnesses diverse computational resources spanning the cloud, local infrastructure, and edge devices, while simultaneously collecting comprehensive metrics from both training and deployment phases to support multi-objective optimization. (4) Furthermore, we introduced informative visualizations akin to XAutoML to enhance human-friendly insights and incorporate humans at the center of optimization loop.
- **Application**: OmniForce developed multiple 2D and 3D annotation models for autonomous driving data and conducted a comprehensive search for a multi-task sensing model tailored to JD Logistics unmanned vehicles.
- **Publication**: OmniForce's design concept and application cases have been organized and refined in our paper, which is accessable on the arxiv preprint. https://arxiv.org/abs/2303.00501

Research Experience

Visual Intelligence and Pattern Analysis Lab, Zhejiang University Hangzhou, Zhengjiang
Graduate Researcher 2018.9 – 2021.3

- **DEAL (Data Sampling)**: https://github.com/Shuai-Xie/DEAL
  Proposed a Difficulty-awarE Active Learning method for semantic segmentation. (1) Devised a plug-in segmentation branch to learn semantic difficulty maps directly from the binary error masks, along with two acquisition functions to select the most informative samples. (2) This approach attained competitive performance on semantic segmentation benchmarks, concurrently offering intuitive guidance for annotators. (3) The DEAL techniques have been incorporated into the semi-automated data annotation module of AI+X Platform.
- **SegHZ (Remote Sensing)**: https://github.com/Shuai-Xie/NBBs
  Engaged in a remote sensing segmentation project with the primary objective of streamlining the annotation process for satellite maps using semantic segmentation models. (1) Due to security considerations, our original dataset was limited to a small number of labels without RGB images, so we crawled the remote sensing tile maps from publicly accessible satellite mapping platforms, such as BaiduMap. (2) Employed an automated alignment method based on neutral best buddies pair matching techniques to establish a satellite map semantic segmentation dataset specifically for Hangzhou. (3) Enhanced the model's performance through a series of optimization techniques, including the utilization of the HRNet backbone, pyramid spatial pooling, positional attention, label smoothing, ensemble consistency loss, and test augmentation. (4) The SegHZ models have been successfully deployed in practical applications.
- **Wali-Turtlebot (EdgeAI)**: https://github.com/Shuai-Xie/Wali-turtlebot
  Developed a self-driving Turtlebot that leverages semantic segmentation results from RGBD data for path planning. (1) Established two operational modes: client-server mode, where computational tasks are centralized on the server-side, while the client is limited to data collection; edge-computing mode, which delegates computation to edge devices for increased efficiency and real-time processing. (2) Developed a high-frequency ROS communication system between the Turtlebot and server nodes, and implemented a custom RGBD message format to transfer stereo vision or kinect RGBD data. (3) Distillated a lightweight BiseNet model with structural knowledge distillation, resulting in enhanced performance when utilizing a ResNet18 backbone. (4) Deployed the lightweight model with optimized TensorRT engine format on an NVIDIA Jetson TX2 for real-time autonomous navigation.

Publications

1. **Xie S**, Feng Z, Chen Y, et al. Deal: Difficulty-aware active learning for semantic segmentation[C]. Proceedings of the Asian Conference on Computer Vision. 2020.
2. Yang J*, **Xie S***, Li Z, et al. Soft-Prompting with Graph-of-Thought for Multi-modal Representation Learning[C]. Proceedings of the 31th International Conference on Computational Linguistics. 2024. (Under Review, * is equal contribution).
3. Xue C, Liu W, **Xie S**, et al. OmniForce: On Human-Centered, Large Model Empowered and Cloud-Edge Collaborative AutoML System[J]. arXiv preprint arXiv:2303.00501, 2023.
4. Yang J, Li Z, Li C, **Xie S** et al. Generalizing to Unseen Domains via Patch Mix[J]. Multimedia Systems. 2023.
5. Yang J, Li Z, Li C, **Xie S** et al. DSDRNet: Disentangling Representation and Reconstruct Network for Domain Generalization[J]. Multimedia Tools and Applications. 2023. (Under Review).

Awards & Achievements

National Scholarship, 2015

National Encouragement Scholarship, 2016 & 2017

Honorable Mention of Mathematical Contest In Modeling, 2016

Silver Award of National Undergraduate Entrepreneurship Competition, 2017

Silver Award of Undergraduate Innovation and Entrepreneurship Competition in Sichuan Province, 2017

Outstanding Graduate of Hunan Province, 2018

Excellent Graduate of Zhejiang University, 2021

Hobbies

Fitness, Reading, Badminton, Basketball, Short Video Creation