Hao Xu

Machine Learning R&D Engineer | PhD Applicant

+86 15151869028 haoxu96@qq.com github.com/haoxuhao

Shanghai, China

Summary

- Currently working as a machine learning R&D engineer at Pinduoduo with a focus on **GPU-related performance engineering and deep-learning based recommendation systems**.
- Abundant hands-on R&D experience in designing GPU-accelerated ML algorithms and building scalable ML pipelines for real-world recommendation systems that work with over 700 million users and billions of items.
- Strong academic background in machine learning and evolutionary algorithms with a CGPA of 90.5% and 83% for the bachelor's and master's degrees.
- Ample research experiences; Published two research papers (IEEE TEVC'22 and CCC'19) and hold one patent.
- Now looking for a **PhD position** in the area of **machine learning or evolutionary computation**.

Education

2021 **MSc in Control Science and Engineering**, **Southeast University**, Dissertation: Study of the Adaptive and Distributed Evolutionary Multi-Task Optimization Algorithm, CGPA 83%, Selected Courses: Data Structures, Algorithms, Intelligent Robotics, Computer Vision, Image Processing, Statistical Machine Learning, Neural Networks, Applied Machine Learning, Linear Systems

2018 **BEng in Automation**, **Southeast University**, Dissertation: Face Recognition in Real-Time Surveillance Video, GPA 90.5%, Distinguished Graduate, Selected Courses: Machine Vision, Pattern Recognition, Control Theory, C++, Probability Theory, Communication Networks, Microcomputer Systems

Selected Research Projects

Evolutionary Multi-task Optimization with Adaptive Knowledge Transfer

Research Paper

1) Proposed a framework to reduce the occurrence of negative knowledge transfer in evolutionary multitasking (EMT), where multiple related tasks are solved simultaneously by individual evolutionary solvers. 2) Design the algorithm to adapt knowledge transfer frequency, knowledge source selection, and knowledge transfer intensity synergistically to make the best use of knowledge transfer, where each solver can utilize the historical rewards of knowledge transfer to guide the transferring frequency and the attention on those tasks with higher benefits. 3) Implemented the proposed framework and conducted extensive experiments that involve problems with 2, 10, 50, 2000 tasks in comparison to several state-of-the-art EMT methods with certain adaptation strategies regarding knowledge transfer.

Evolutionary Computation | Evolutionary Multitasking | Adaptive Knowledge Transfer | Multitask Optimization | Algorithm

Vehicle Detection in Aerial Images

Presentation Slides

1) Trained the base model based on RetinaNet Xt101 to detect vehicles in aerial images. 2) Developed an algorithm to leverage the overlapped slide window and generate image patches for training and inference, which could maximally preserve the information of small objects while retaining the object size and dealing with large-scale images. 3) Optimized the model by introducing a hard example resampling phase, which involves small average car sizes, high car densities, or low foreground ratios. 4) Designed a post-processing method to deal with the overlapped detection and improved the model's accuracy.

Computer Vision | Machine Learning | RetinaNet | Object Detection | Image Processing | Algorithm Design

Computer Vision Based Coal/Gangue Recognition

Demo

1) Built a black box over the belt with a strictly controlled light condition for image capturing and preprocessed the image with high-band filtering to strengthen the imaginary features. 2) Leveraged the YOLOv2-tiny model with ImageNet pretrained weights fine-tuned on the coal mine images for detection and optimized the model to adapt to power-constrained embedded systems. 3) Leveraged OpenCV as the inference engine and achieved a throughput of 20FPS and an accuracy of over 90% in the production environment. 4) Designed the algorithm to collect the images with low confidence for re-labeling and re-training to maintain a consistent accuracy through the long run.

Computer Vision | Machine Learning | YOLO | ImageNet | OpenCV | Object Recognition | Algorithm Design

Crowd Counting with Segmentation Map Guidance

Research Paper

1) Propose a new deep CNN model to integrate a segmentation map to compensate for the false response under a complex environment. 2) Designed a new method to generate segmentation ground truth merely based on the density map instead of manual labeling. 3) Optimized the geometry-adaptive method for density generation employed in the highly congested area. 4) Conducted extensive experiments and the results show that the proposed method could effectively reduce the false response of extremely complex backgrounds.

Machine Learning | Crowd Computing | Crowd Segmentation | Deep Learning | CNN

Professional Experience

Present – Jul 2021 | **R&D Engineer - Recommendation Systems, Pinduoduo, Shanghai, China**

- Devise a GPU-based Cross-ID generation method, which utilizes the cache blocking method to alleviate the memory load latency in GPU and achieves a 5% performance improvement.
- Design a batch padding and fine-grained profiling scheme to get better GEMM parameters and optimize multi-layered perceptron (MLP) computation with FP16 precision (mixed with FP32). The proposed solution could decrease the computational cost of the online model serving system by over 20% (from 300 Nvidia A10 GPU to 200 Nvidia A10 GPU).
- Design an embedding update scheme for the embedding layer training in GPU, which could achieve a 30% speedup over the baseline scheme.

Machine Learning | Recommendation Systems | ML Acceleration | NN Compression | CUDA | Algorithm Design

Jun 2021 – Jun 2020 | **Software Engineer - Operations Research , Ant Group, Hangzhou, China**

- Developed and implemented a reusable software framework for multi-objective optimization. Implemented and verified two multi-objective optimization methods, namely the weight-sum method and the augmented e-constraint method.
- Verified the effectiveness of the framework on the optimization problem of fund transfer in Ant Group, which provides the decision maker the knee point on the Pareto front for design making with an improved 5% average global interest rate under the required risk rate.
- Utilized the multi-objective Bayesian method to determine the points to search on the Pareto front. This strategy minimized the number of expensive optimization process (with billions decision variables) performed on the less-critical Pareto front, which saved computational cost by roughly 30% while maintained satisfied accuracy.
- Developed algorithms to perform exploratory analysis on historical transactions to find statistical patterns. Trained CNN-based models to detect anomalies that may be related to money laundering.

Machine Learning | Multi-Objective Optimization | Financial AI | CNN | Algorithm Design

Publications

- **H. Xu**, A. K. Qin and S. Xia, "Evolutionary Multitask Optimization with Adaptive Knowledge Transfer," in IEEE Transactions on Evolutionary Computation, vol. 26, no. 2, pp. 290-303, April 2022, doi: 10.1109/TEVC.2021.3107435.
- **H. Xu**, C. Zheng, Y. Nie and S. Xia, "Crowd Counting with Segmentation Map Guidance," 2019 Chinese Control Conference (CCC), 2019, pp. 7716-7721.
- **H. Xu** and S. Xia "A Crowd Density Estimation Method Based on Foreground Segmentation Map," CN110276264A[P]. 2019.

Skills

Programming	C++, Python, CUDA, Java, LaTeX
ML Frameworks	PyTorch, TensorFlow, Keras
ML Models	DCNN, YOLO, KNN, SVM, LSTM, ResNet, RetinaNet, GoogleNet, DenseNet, NLP, AlexNet, LeNet
DevOps	git, SVN, Maven, Docker

Awards

- 2022 **SEU Distinguished Graduate Award** (Top 2%)
- 2021 **The Distinguished Master's Thesis Award, JAAI**
- 2019 **First Prize in 2019 Object Detection and Recognition Competition, CSAA** (Top 2%)
- 2016 **National Scholarship** (Top 5%)
- 2015 **Texas Instruments Scholarship**

Languages

- GRE 317 (Percentiles: 97/45)
- TOEFL 100