Chengsong Zhang

continuerevolution@gmail.com | continue-revolution.github.io | continue-revolution | LinkedIn

Education

University of Illinois, Urbana-Champaign 08/2023 - 05/2025
Master of Science in Computer Science Urbana, IL, USA

University of Michigan, Ann Arbor 08/2021 - 08/2023
B.S.E. in Computer Science and Engineering Ann Arbor, MI, USA

Shanghai Jiao Tong University 09/2019 - 08/2023
B.S.E. in Electrical and Computer Engineering Shanghai, P. R. China

Stable Diffusion (SD) Open Source Projects

AnimateDiff for Stable Diffusion WebUI 07/2023 - 04/2024
Owner, ☆3.1k sd-webui-animatediff

- Insert AnimateDiff self attention motion module to SD UNet at runtime to force a batch of images to have temporal consistency while easy to restore UNet to image generation
- Interpolate prompt conditions to achieve smooth prompt condition transfer from one prompt to another.
- Re-write ControlNet main entry to process conditions in batch to tranfer styles from one video to another. It has proven production-ready performance within the domain of video 3D-to-2D transfer and style transfer, when several ControlNets are applied to SD with AnimateDiff.
- Optimize inference with flash attention and fp8 weights to improve speed and reduce VRAM by 3x. Native FP8 support let users run 1024x1024 high-res video-to-video transfer with only 18GB VRAM cost. Native LCM samplers let users generate reasonable videos within 8 steps.
- Contribute to sd-webui-controlnet and SD WebUI Forge, contributions include SparseCtrl and batch frame-by-frame control.
- Contribute to SD WebUI, contributions include LCM sampler and tweaked script hooks.

Segment Anything for Stable Diffusion WebUI 04/2023 - 01/2024
Owner, ☆3.4k sd-webui-segment-anything

- Automatically create bounding boxes and masks by clicking on images or entering text prompts in A1111 WebUI, both in single images and in batch, with the help of GroundingDINO (a powerful text-to-bounding-box model) and Segment Anything.
- It can automatically send masks to SD or ControlNet for inpainting.
- It can segment human or any other objects from source videos for
  - video style transfer with ControlNet and AnimateDiff
  - creating a better training dataset for LoRA or LyCORIS
- It can improve semantic segmentation and automatically send the semantic control map to ControlNet for retional-controlled image generation.

Research Projects – Systems for ML

Distributed Stable Diffusion with Regional Skip 08/2024 - Now
Co-leader advised by Fan Lai Ongoing

- Design and implement a system based on distrifusers that spawn N+1 processes where each control 1 diffusion process with 1 GPU. 1 process run warm-up steps to generate masks; N processes run main inference that diffuse latent representation against a certain region
- The system communicate with sub-processes to allocate tasks for each sub-process and decide when to skip steps in main inference for a particular region based on attention score, and automatically fill available GPUs with other available tasks to maximize throughput
- Collaborate with co-leader to redesign ML operators to communicate buffers effectively to help regional and skip-able convolution layers and self-attention layers

FedScale 05/2022 - 02/2023
Core contributor advised by Fan Lai and Mosharaf Chowdhury, ☆350 SymbioticLab/FedScale

- FedScale is a scalable and extensible open-source federated learning (FL) engine and benchmark.
- Design a distributed, hierarchical and serverless protocol to efficiently check-in clients and aggregate models
- Implement on-device training on various edge devices, such as clusters, PC and android. It supports a series of state-of-the-art execution frameworks, such as PyTorch, Alibaba MNN and TensorFlowLite.

Research Projects – Database Systems

PilotDB: Approximate Aggregation Query Rewrite Optimization 01/2024 - 04/2024
Core contributor advised by Daniel Kang In Submission

- Manually rewrite TPC-H and TPC-DS queries into table sampling and do experiments against duckdb to find potential bottleneck and speed-up opportunity.
- Compile runtime and profile data to find query runtime speed-up with table sampling when 5% error is guaranteed. Re-writing queries into table sampling can bring 2x - 100x speed-up for some queries with a guaranteed 5% error.
- Analyze and re-write duckdb source code to skip chunks at sequential scan and gain 2-20x extra speed-up. See code comparison here.

AIDB 08/2023 - 12/2023
Core contributor advised by Daniel Kang, ☆20 ddkang/aidb

- AIDB is a machine-learning analytics framework that can analyze unstructured data with machine learning in a structured way.
- Integrate cloud inference API and local inference from PyTorch (GroundingDINO object detection) and Detectron2 (document segmentation and OCR).
- Investigate and experiment vector databases (Faiss / ChromaDB / Weaviate) for querying embedings for approximate selection / aggregation.
- Design and implement Function-as-a-Service ML service, configuration schema and command line user interface. Implement several examples including NSFW detection and legal analysis.
- Improve querying speed via batching cached bound inference service.
- Design and implement downstream application query-your-video for AIDB that can automatically chain ffmpeg frame extraction, object detection, image classification, instance segmentation, image tagging via SQL queries, to select frames containing desired objects from videos.

Publications

[1] Tengjun Jin, Akash Mittal, Chenghao Mo, Jiahao Fang, Chengsong Zhang, Timothy Dai, and Daniel Kang. "AIDB: a Sparsely Materialized Database for Queries using Machine Learning". In: *DEEM*. 2024.

[2] Sanjay Sri Vallabh Singapuram, Chuheng Hu, Fan Lai, Chengsong Zhang, and Mosharaf Chowdhury. "Flamingo: A User-Centric System for Fast and Energy-Efficient DNN Training on Smartphones". In: *DistributedML*. 2023.

[3] Yuxuan Zhu, Tengjun Jin, Stefanos Baziotis, Chengsong Zhang, and Daniel Kang. "PilotDB: Database-Agnostic Online Approximate Query Processing with A Priori Error Guarantees". In: *In Submission*. 2024.

Skills

Languages: Python, C/C++, Go, CUDA, JavaScript, Java, OCaml, Rust
Framework: PyTorch, TensorFlow, LLVM, Flask, React.js

Courses

Compiler Construction: LLVM, lexer, parser, codegen, register allocator, LICM, SCCP
Distributed Systems: GoLang, sharded key/value service with paxos groups
Operating Systems: C++, thread library, virtual memory manager, network file systems
Deep Learning for Computer Vision: PyTorch, KNN, FC, CNN, BatchNorm, AutoGrad, Transformers, object detection (FCOS/Faster R-CNN), image captioning (CNN + RNN/LSTM/Attention), image generation (VAE/GAN), network visualization, style transfer
Web Systems: Python, JavaScript, Flack back-end API, React.js dynamic front-end, MapReduce, search engine
Misc.: applied parallel programming (CUDA), formal verification (Dafny), programming languages (OCaml/Rust), Java programming (Java), computer security

Teaching

CS598: Systems for Generative AI (Fall 2024), UIUC
CS510: Advanced Data Management (Spring 2024), UIUC
VE280: Programming and Elementary Data Structures (Summer 2021), SJTU
VV285: Honors Mathematics - Multi-Variable Integration (Summer 2021), SJTU
VV214: Linear Algebra (Spring 2021), SJTU
VV186: Honors Mathematics - Single-Variable Integration (Fall 2020), SJTU