JUNTANG ZHUANG

Email: zhuangjt12@gmail.com ⋄ Website: https://juntang-zhuang.github.io ⋄ Phone: +1 475-224-8282 ⋄ [Google Scholar]

ABOUT ME

I am a Research Sceintist at OpenAI. I'm recognized as the primary contributor to DALL-E 3 and a core contributor to Embedding V3. I am the sole inventor of long-ctx algorithm enabling GPT-4 Turbo's 256k-context capability, notably surpassing the performance of Claude. My research interest lies in machine learning, optimization, large language models and numerical methods and theories.

WORK EXPERIENCE

OpenAI, Research Scientist April 2022 - now

- Working on next-generation GPT model architecture
- Sole inventor of long-context algorithm, delivered to GPT4-Turbo (256k capability, served 128k), outperformed Anthropic's Claude 2.1
- Lead multilingual capability for OpenAI Embedding v3, outperformed Cohere's embedding model
- Primary contributor to DALL·E 2.5 & 3 (contributors disclosed)
- Co-author of GPT-4 (contributors disclosed)

Google, Student Researcher June-Oct 2021

- Proposed GSAM, a generic method to improve the generalization of neural networks

OPEN-SOURCE PROJECTS

AdaBelief-optimizer (*>1k* stars on github, added to official repositories such as Deepmind optax, Tensorflow-Addons and Google Flax); ShelfNet; LadderNet; TorchDiffEqPack

AWARDS & SCHOLARSHIPS

- Henry Prentiss Becton Graduate Prize (1 out of Yale School of Engineering & Applied Science) 2022
- Best paper award, Machine Learning in Medical Imaging (MLMI) 2019
- Top-1 winner for CNI Transfer Learning Challenge, MICCAI 2019
- Graduate fellowship, Yale University 2016
- Award for excellent learning performance, Tsinghua University 2015
- Meritorious award for Mathematical Contest in Modeling (top 10% teams worldwide) 2015
- National encouragement award (for excellent learning performance), Tsinghua University 2014
- Sparks Program (Undergraduate High-tech Club) membership, Tsinghua University 2014

EDUCATION

Yale University
Ph.D. in Biomedical Engineering (Advisor: James S. Duncan) Sep 2016 - April 2022

Yale University
M.A. in Statistics, M.Phil in Biomedical Engineering Sep 2017 - May 2018

Tsinghua University
B.E. in Engineering Physics Sep 2012 - May 2016

SELECTED PUBLICATIONS

1. **J. Zhuang**, B. Gong, et al., Surrogate gap minimization improves sharpness-aware training *International Conference on Learning Representaitions* (ICLR 2022)[project page]
2. **J. Zhuang**, Y. Ding, et al., Momentum centering and asynchronous update for adaptive gradient methods *Conference on Neural Information Processing Systems* (NeurIPS 2021)[project page]
3. **J. Zhuang**, N. Dvornek, et al. MALI: a memory efficient and reverse accurate integrator for Neural ODEs, *International Conference on Learning Representations* (ICLR 2021)[project page]
4. **J. Zhuang**, N. Dvornek, et al. Multiple-shooting adjoint method for whole-brain dynamic causal modeling , *Information Processing in Medical Imaging* (IPMI 2021, oral presentation) [project page]
5. **J. Zhuang**, T. Tang, et al. AdaBelief Optimizer: adapting stepsizes by the belief in observed gradients, *Conference on Neural Information Processing Systems* (NeurIPS 2020, Spotlight) [project page]

6. **J. Zhuang**, N. C. Dvornek, et al. Adaptive Checkpoint Adjoint Method for Gradient Estimation in Neural ODE, *International Conference on Machine Learning* (ICML 2020) [project page]
7. **J. Zhuang**, J. Yang, et al., ShelfNet for fast semantic segmentation, *Workshop on Computer Vision for Road Scene Understanding and Autonomous Driving* (CVRSUAD 2019)

SELECTED RESEARCH EXPERIENCE

1. Optimization for deep learning

Surrogate gap minimization improves sharpness-aware training Jun - Oct 2021

- Proposed a generic method to improve the generalization of neural networks. Specifically, for the ImageNet top-1 accuracy the proposed method achieved **+11.3%** improvement over AdamW on Vision Transformer, and **+12%** improvement on MLP-Mixer.
- Paper accepted to ICLR 2022.[project page]

Momentum centering and asynchronous update for adaptive gradient methods Jan - May 2021

- Proposed ACprop, which is an adaptive optimizer combining momentum centering and asynchronous update. Theoretically, ACProp has the optimal convergence rate and weak convergence conditions. Validated ACProp in extensive empirical studies.
- Paper accepted to NeurIPS 2021. [project page]

AdaBelief optimizer: a fast, accurate and stable optimizer for deep learning Jan - June 2020

- Developed an optimizer for deep learning models. To our knowledge, it's the first to achieve three goals simultaneously: *fast training speed, good generalization performance, and stability of training.*
- Paper accepted as **Spotlight Presentation** by NeurIPS 2020. [project page]

2. Solvers for continuous-time neural networks

MALI: a memory efficient and reverse accurate integrator for Neural ODEs Sep - Nov 2020

- Proposed MALI, a new solver for Neural ODEs with numerical accuracy at a constant memory cost. MALI achieves new state-of-the-art (3.71 BPD on ImageNet64) for image generation with continuous models.
- Paper accepted by International Conference on Learning Representations (ICLR 2021) [project page]

Adaptive checkpoint adjoint method for gradient estimation in neural ODE Jun - Dec 2019

- Proposed and implemented a family of adaptive ODE solvers for accurate gradient estimation. Achieved both accuracy and computation efficiency. To our knowledge, our method is the first to enable neural-ODE to achieve comparable results to state-of-the-art discrete-layer models on benchmark classification tasks.
- Paper accepted by International Conference on Machine Learning (ICML 2020). [project page]

3. Prior-informed machine learning and biomedical applications

Evolutionary causal modeling of brain states from task-fMRI data Mar - Sep 2020

- Modeled the effective connectome of the brain, which is the directional influence between different regions of the brain. Developed a differential equation model to simulate the dynamical evolution of brain states.
- Paper accepted as **Oral Presentation** by IPMI 2021.

PROFESSIONAL ACTIVITY

Served as reviewer for MIDL, ICML, NeurIPS, ICLR, EMBC and MEDIA.

INVITED TALKS

- Computational Neuroscience Laboratory at Stanford University 2020
- SyncedTech (name in Chinese pinyin: JiQi ZhiXin) [video] 2020

SKILLS

C/C++, R, Python, MATLAB, PyTorch, Keras, Jax, Tensorflow, Triton