Cuize(CZ) HAN

hancuize@gmail.com (608)-320-2673 linkedin.com/in/cuizehan Bay Area, CA

EXPERIENCE

Amazon Inc. Palo Alto, CA
Senior Applied Scientist *Aug 2022 - Present*

- **LLM Trust and Safety**
  - **Trust and Safety**
    Working as the science lead, design and implement the hydra model structure where we train multiple light-weighted heads that can evaluate the (partial) generated responses according to different sub-dimensions of human values such as harmlessness and helpfulness. Those heads are parallel to the LM head and trained while freeze the base model. Such design enables us to do online content moderation while streaming the response with minimum latency increased.
  - **Base Modeling**
    Collaborate with multiple teams building a billion-parameter-level large language model for shopping applications.
  - **Large-scale LM Training**
    Leverage model sharding through pytorch FSDP and model paralleism based on Nvidia Nemo Megatron to efficiently train billion parameters scale LLM with trust and safety heads.
  - **Model inference optimization**
    Implement the hydra model structure in Nvidia Faster Transformer package in C++ and CUDA for more efficient inference.

- **New Product Discovery**
  - Lead the science part of the new product exploration project that aims for solving the cold start problem in product search ranking.
  - Designed, implemented, experimented and launched an Empirical Bayes (EB) based online feature exploration solution in 8 locales that significantly improved new product discovery experience.
  - Based on the uncertainty quantification and UCB (upper confidence bound) principle, designed an efficient algorithm that balanced exploration and exploitation.
  - Extend the EB based online feature exploration framework to worldwide launch. Lead the collaborations with multiple teams.
  - Improve the product exploration framework through Bayesian deep learning.

Amazon Inc. Palo Alto, CA
Applied Scientist *Mar 2019 - Aug 2022*

- **Search ML Platform**
  - Help building an efficient search machine learning platform that support deep learning use cases including ranking and extreme label classification.
  - Design and implement APIs for the platform that help data processing, model training, evaluation and compression.
  - Develop and implement AGB (Accelerated Gradient Boosting) algorithm based on recent research that produces a more compact model that enables faster inferences.
  - Integrate and optimize the one-pass feature selection and model training process with HPO (Hyperparameter optimization) in Amazon SageMaker through Constrained Bayes Hyperparameter Tuning strategy.

Amazon Inc. Palo Alto, CA
Applied Scientist Intern *Aug 2018 - Dec 2018*

- **Fast Feature Selection**
  - Develop and write working C++ code for gradient boosting feature selection (GBFS) algorithm and its improvement and variant.
  - Develop and implement a fast GBFS algorithm for high dimensional data.
  - Develop the Multitask GBFS algorithm that can perform multitask learning in the GBDT setting.
  - Write shell scripts for doing extensive tuning on various parameters and testing on different search defect and relevance datasets.

EDUCATION

University of Wisconsin, Madison — Madison, WI
Ph.D. in Statistics; (GPA: 3.96/4.0) — *Sep. 2013 – Jan. 2019*

Shanghai Jiao Tong University — Shanghai, China
B.S. in Mathematics and Physics; (GPA: 3.9/4.0) — *Sep. 2009 – Jun. 2013*

RESEARCH PAPERS

- "Mitigating Exploitation Bias in Learning to Rank with an Uncertainty-aware Empirical Bayes Approach" Tao Yang, **Cuize Han**, Chen Luo, Parth Gupta, Jeff M Phillips, Qingyao Ai (submitted to CIKM 2023)
- "Measuring service-level learning effects in search via query-randomized experiments" Paul Musgrave, **Cuize Han**, Parth Gupta (SIGIR 2023)
- "Addressing Cold Start in Product Search via Empirical Bayes" **Cuize Han**, Pablo Castells, Parth Gupta, Xu Xu, Vamsi Salaka (CIKM 2022)
- "Imputation Balanced GAN: Sequence Classification with Data Augmentation" Grace Deng, **Cuize Han**,Tommaso Dreossi, Clarence Lee and David S. Matteson (SDM 2022)
- "Learning to Rank with Missing Data via Generative Adversarial Networks" Grace Deng, **Cuize Han** and David S. Matteson (Journal of Data Mining and Knowledge Discovery, 2022)
- "Scalable Feature Selection for (Multitask) Gradient Boosting Machines" **Cuize Han**, Nikhil Rao, Daria Sorokina and Karthik Subbian (AISTAT2020)
- "Information Based Complexity for High Dimensional Sparse Functions" **Cuize Han**, Ming Yuan (Journal of Complexity 2020)

AWARDS/ACKNOWLEDGEMENT

- JSM Best student paper (2021): Learning to Rank with Missing Data via Generative Adversarial Networks
- An Amazon and US patent (2019): Faster and Robust Feature Selection for Gradient Boosting Machines
- Determine a tight sample complexity lower bound for high dimensional sparse function approximation which posed as an open problem in information complexity for 8 years (2017)
- First place in Ph.D. Qualification Exam of Statistics Department in University of Wisconsin-Madison (2014)
- Outstanding student award in Zhi Yuan College of Shanghai Jiao Tong University (2013)
- S.T.Yau College Mathematics Contests Honorable Mention in 'Analysis and Differential Equations' and 'Applied and Computational Mathematics' (2012)
- A-class scholarship in Shanghai Jiao Tong University (2012)
- First prize in National High School Mathematics competition (2008)

PROGRAMMING SKILLS

- **Languages**: Python, SQL, Shell, C++, R, Matlab
- **Technologies**: AWS, PyTorch, Huggingface, Faster Transformers