Zachery Ye

(+86) 131-****-6485 - yez****@mail2.sysu.edu.cn - yzq986.github.io

TECHNICAL SKILLS

Algorithms: Recommendation Systems Architecture, Model Optimization, Deep Learning
Programming Languages: C++, Python, Java
Machine Learning: TensorFlow, PyTorch
Data & Engineering: Spark, Hadoop, Docker, Kubernetes, AWS, Git

AWARDS & HONORS

- Gold Medal — ACM-ICPC Asia Regional Contest Finals 2017 (December 2017)
- Gold Medal (5th Place) — ACM-ICPC Asia Regional Contest Hong Kong 2017 (November 2017)
- Gold Medal — ACM-ICPC Asia Regional Contest Shenyang 2017 (October 2017)
- Gold Medal — ACM-ICPC Asia Regional Contest Qingdao 2016 (November 2016)
- Gold Medal (5th Place) — CCPC Changchun Regional Contest 2016 (September 2016)
- Silver Medal (18th Place) — CCPC Finals 2016 (December 2016)
- Bronze Medal — National Olympiad in Informatics (NOI) 2013 (August 2013)
- Meritorious Winner — Mathematical Contest in Modeling (MCM) 2017 (April 2017)

WORK EXPERIENCE

A Top-tier Exchange - Square Feed Stream - Algorithm Lead

Head of Square Recommendation Algorithms | Tech Lead April 2025 — Present

- As the algorithm lead for the Square Feed recommendation system, I am responsible end-to-end for algorithm optimization, engineering architecture, and team collaboration, driving a breakthrough in trading conversion for the Square feed. By establishing a complete trading attribution system, enabling cross-team collaboration, and optimizing full-stack system performance, I delivered an estimated annual incremental trading contribution of 4.11B USD.
- End-to-end Leadership in Recommendation System Architecture Upgrades & Business Growth
  - As Tech Lead, I led full-pipeline optimization of the recommendation system (v23 → v24 → v25), coordinating among algorithm, engineering, and business teams (10+ members), and establishing a cross-department collaboration mechanism.
  - Core annual achievements: component-driven order volume +30.95%, daily trading volume 126M USD (annual increment 4.11B USD), conversion efficiency +23.94%.
  - Built a complete performance evaluation system and AB testing framework to ensure all improvements are purely from model contributions, excluding product or operations influence.
- System Performance Optimization & Cost Control
  - Led offline training pipeline optimization: identified and solved memory explosion issues in offline feature construction, designed a two-phase deduplication strategy, reduced memory peak by 50%, lowered failure rate from 30% to <1%, and reduced processing time from 50 minutes to 30 minutes.
  - Led SageMaker online inference optimization: collaborated deeply with engineering to upgrade instance configurations from 300×4x large to 150×8x large, reducing resource cost by 50% while keeping inference stable and performant.
  - Annual cost optimization: offline training costs reduced from 6000 USD/day to 3000 USD/day, online inference cost -50%, total estimated yearly savings 2M+ USD.
- Organizational Collaboration & Technical System Building
  - Established deep collaboration with the trading team (Tat/Maya/Choco), connecting content recommendation with trading conversion and enabling the first accurate quantification of content-driven trading contribution.
  - Drove engineering team (Jeffrey/JerryM) to optimize feature storage (30% compression), inference performance, and deployment pipelines, raising team-wide iteration efficiency.
  - Built standardized SOPs for model launch, evaluation frameworks, and diagnostic workflows; documented methodologies to support knowledge transfer and team capability growth.
- Feature Engineering & Model Architecture Evolution

    - v23 model: Introduced trading objectives into the recommendation model for the first time; designed a complete trading attribution pipeline from scratch: mapped 20+ trading events, created 60s/1h backfill mechanisms, and integrated futures/spot/alpha scenarios into a unified trading base table.
    - Innovatively incorporated trade_click, place_order, main_trade and other trading targets into a multi-objective model, enabling end-to-end optimization of click → dwell → trade and improving attribution accuracy.
    - 60-second order volume +16.07%, orders per user +21.78%; in high-conversion scenarios, orders per user increased by 59%–76%.
    - v24 model: systemically enhanced the feature system, introducing Category Features v3 (Track/Product/Content tri-axis classification), enriching user behavior sequence features (click/like/share/follow/comment with category statistics), leading to +32.78% 60-second orders per user.
    - v25 model: enhanced historical trading-widget click sequence features (adding 8 new side-info dimensions including author/token/zone/keyword), integrating them as sequences in the model to better understand user trading intent, increasing daily orders by +1.08%.
    - Proactively identified and fixed multiple legacy feature bugs, establishing a complete data quality monitoring system to ensure fundamental model reliability.

TikTok - Global E-commerce - Algorithm Expert

Global E-commerce Recommendation Team Member — April 2024 – April 2025

- Ranking system optimization, iterating on feature signals, model structures, samples and ranking objectives for both coarse and fine ranking
- Coarse Ranking Listwise Data Flow Distillation
    - Achieved online-offline score alignment by supplementing fine ranking scoring samples into coarse ranking data flow
    - Optimized distillation strategy using implicit task heads, reversed heavy structures
    - Achieved significant improvements in US region: Shop GMV/capita +1.4957%, UV_CTR +0.4300%
- Coarse Ranking Fusion Formula Iteration, VT Pipeline Reconstruction
    - Optimized configuration management for coarse ranking fusion formula, simplified model adjustment process
    - Systematic parameter optimization strategy
    - Significant growth in US region: UV_CTR +0.3650%, GMV +0.9536%
- Fine Ranking Sub-scenarios, Small Models Replacing Large Models
    - Successfully replaced original large scenario models with Lite models by combining global large models and scenario-specific small models
    - Small models trained with scenario-specific sample flow, reducing resource consumption and accelerating training process
    - Maintained GMV and CTCVR while reducing training PS consumption by 52.73% and online PS consumption by 79.9%

Shopee SG - Senior Algorithm Engineer

Daily Discovery Team Member — October 2021 – April 2024

- Design and optimize ranking models for Shopee homepage recommendation system
- First Version Multi-objective Cascade Coarse Ranking Model Deployment Across All Regions
    - Shared training and inference framework with fine ranking, accelerated online inference by caching item-side embeddings
    - Adopted three-tower + top MLP fusion architecture to obtain cross-information while ensuring inference efficiency
    - Achieved ~46ms coarse ranking latency with 3000 item inputs, significant improvements across all regions
    - Multiple metrics improved: ID +2.33% click +1.59%, BR +2.3% order, VN +3.92% order
- Dual Coarse Ranking Joint Model Research and Implementation
    - Utilized in-batch random negative samples and adopted CLIP's bidirectional CE loss with adaptive temperature coefficient

    - Achieved +5.23% order improvement offline, +2.3% click in ID and +3.36% order in SG regions online
- Coarse Ranking Multi-objective ESMM Structure Upgrade
    - Introduced multi-level objectives, flexible adjustment of online scoring sequence weights, optimized item ranking and implicit objective training
    - Achieved 3.35% order/user improvement in ID region
- Fine Ranking JRC Structure Upgrade
    - Replaced pointwise training with listwise training, adopted RCR loss and CE+GE loss from JRC paper
    - Achieved 4% click/user improvement in ID region
- Shop Display Page Fine Ranking Model Feature Selection
    - Introduced slot multiplier optimization for feature selection, removed 59 low-weight multipliers
    - Reduced storage capacity by 50%, improved training speed by 50%
    - Improved order metrics by +1.27% while maintaining global click rate

WeChat Channels - Algorithm Engineer
Algorithm Recommendation Team Member June 2020 – October 2021

- Research and development of video recommendation algorithms and backend data support for WeChat Channels
- Lookalike Model Architecture Development
    - Built lookalike pipeline in hot feed recommendations, implemented user-to-user crowd expansion
    - Generated historical data through MQ buffer processing by aggregating impressions for the same item, handling ~330M daily impressions
    - Optimized online inference performance using cache for intra-round caching and bdemem for inter-round caching
- Framework Prediction Process Migration to Metis 1.0 Architecture
    - Completed model architecture migration, decoupling model inference embedding logic from original process
    - Created new RPC interface for computing inference embeddings for all requests
- Cold Start Recommendation Boost Formula Modification
    - Replaced original cold start pipeline with PID control-based fusion boost formula
    - Fine-tuned cold start volume control considering watch time, completion rate, likes and other features
- Array Type Feature Support for Recommendation Architecture
    - Added support for array type features in video feed recommendation pipeline
    - Implemented list feature support across all components in the data flow model
    - Completed integration between online data flow and model training

Google - Software Development Engineer
Software Development Engineer July 2019 – February 2020

- Development, implementation, testing, and deployment of advertising revenue prediction project
- Advertising Revenue Prediction Project
    - Designed algorithmic framework and implemented revenue prediction based on Google advertising business data
    - Implemented revenue predictions for daily, monthly, quarterly, and annual periods
    - Identified peak revenue periods annually, completed comparative testing and deployment

BIGO - Algorithm Engineer - Internship
Algorithm Team Member February 2019 – June 2019

- Completed audio fast clustering algorithm patent and implemented all code details
- Audio Fast Clustering Algorithm Patent
    - Designed hash algorithm for comparing large volumes of short video audio within time constraints, tailored for BIGO's Like app audio characteristics
    - Designed corresponding feature clustering algorithm, optimized time complexity based on business requirements
    - Adapted algorithm for large-scale data comparison needs in business scenarios

The Hong Kong Polytechnic University - Research Assistant
Research Assistant (Exchange Student) September 2018 – January 2019

- Served as a Research Assistant during exchange program, focusing on blockchain applications in supply chain management
- Blockchain-Driven Trusted Supply Chain System Research
  - Investigated the potential of blockchain technology in enhancing supply chain trust and transparency
  - Designed and implemented a blockchain-based supply chain tracking prototype

EDUCATION

Sun Yat-sen University 2015 – 2019
Computer Science and Technology GPA: 3.8/4.0

EXTRACURRICULAR ACTIVITIES

- Technical Director in SYSU ACM/ICPC Club — 2016 – 2017
- Organized algorithm lectures, programming contests for beginners, and school competitions