Xiaochuan Yu (hillyu@live.com)

https://hillyu.linkpc.net

Ph.D. in Computer Science and Information Technology, with expertise in Large Language Models, Computer Vision, Deep Learning, Internet of Things (IoT), and Wearable Computing. 10+ years working experience in AI with extensive team management skills.

Career History

Wisers AI Lab, Wisers Information Limited **Feb 2020 – Jun 2024**

Lead Researcher, Manager Sep 2021 – Jun 2024

High-Impact Initiatives:

- Led the development of custom vertical LLMs (*GLMv3*-based architecture, fine-tuned by *LoRA*).
- Spearheaded the creation of a large language model (LLM) pipeline for core business applications such as sentiment analysis, information extraction and translation, leading to a notable enhancement in performance and substantial conservations in R&D expenditure.
- Performed in-depth research and assessment of cutting-edge technological advancements, with a focus on the practical utility, limitations and benefits of various LLMs (*Phi, Gemma, Llama, chatGLM, Qwen, Baichuan* etc.); conducted evaluations of RAG-based LLM frameworks (e.g., *Lang Chain, Llama Index, ChatGPT plugin*) for robust media insight analytics.
- Devised a multilingual social media analytical solution utilizing LLMs following agentic design principle, catalysing the corporation's business expansion from national markets to a global clientele.
- Instituted a standardized approach and prescribed best practices for LLM-focused workflows, including prompt engineering, continuous model finetuning, and the automation of model deployment.
- Enhanced model efficiency using cutting-edge acceleration tools (e.g., *vllm, fastllm, ollama/llama.cpp*), superseding the *fastapi+onnx* model deployment approach-yielding a 30% reduction in development expenditure and a quadruple increase in performance.
- Productized and refined OCR subtitle extraction processes, slashing associated costs by 75%.
- Optimized the efficiency of logo detection technology, resulting in an 80% cost reduction.

Professional Experience:

- Led a diverse team specializing in computer vision (CV) and natural language processing (NLP). Offered mentorship in harnessing individual proficiencies, task alignment and provided comprehensive support on pivotal technical challenges. Assured project success via hands-on technical direction, management, and risk mediation. Maintained alignment across teams through effective coordination.
- Tackled enterprise-wide technical and business issues, directly contributing to a 20% enhancement in sentiment analysis precision and a 30% improvement in average response speed (RPS) of API calls.
- Forecasted strategic business trajectories and advised senior leadership, enabling informed decision-making (such as the transition to LLMs over traditional deep learning models and the RAG pipeline adoption).
- Actively collaborated with stakeholders to identify the potential for application and commercialization of new technologies, showcasing capabilities through PoC (proof of concept) demonstrations, performance benchmarks and knowledge-sharing forums.
- Monitored technological trends and seamlessly integrated relevant innovations into business solutions; drafted and oversaw research initiatives and interpreted experiment results to guide strategic directions.

Senior Researcher (Leader of Computer Vision team) Feb 2020 – Sep 2021

High-Impact Initiatives:

- Product detection solution (for a prestigious French luxury brand) - securing a multi-year service agreement.
- Brand exposure solution for social listening (For a global pharmaceutical and medical aesthetics company) - winning a long-term contract for ongoing service.
- Table recognition solution for a renowned global insurance company, enhancing their data collection and analytics capabilities.
- Fashion element recognition PoC for a famous global sportswear company, to analyze fashion/outfitting trends dynamically from social media feeds.

Professional Experience:

- Directed foundational research in Computer Vision, encompassing a wide array of specialties including object detection, image retrieval, image classification, few-shot learning, text recognition, and integrated video understanding.
- Led and participated in computer vision competitions (**2nd place in RVC/ECCV 2020 competition**) and international conferences and workshops, significantly raising the profile and reputation of the organization within the tech community.
- Delivered comprehensive training, technical support, and consultancy, augmenting competencies within multiple business units and for a diverse client base.
- Integrated modern project management methodologies, such as Scrum and Kanban, into the research workflow, optimizing efficiency and output by utilizing tools like Jira and Trello.

- Strategically extrapolated insights from retrospective projects to identify technological opportunities and pre-emptively address business risks, catalysing R&D initiatives in areas like few-shot learning and image retrieval that expanded the potential for market penetration and fortified the product portfolio against future challenges.

Embedded System, Applied Science and Technology Research Institute (Astri) **Oct 2017 – Oct 2019**

Senior Engineer (Lead Eng. Sr. Professional)

Mainly responsible for system architecture design, algorithm design, implementation of back- end service, deployment and maintenance automation. Also actively participated in tech consultation, including enterprise specifics & requirement analysis, technology selection, solution design & prototype.

Notable projects:

- Drafted, proposed, and applied ITF funded project (approved): ARD/246CL Context Recognition from Multi-Stream of Audio and Video. 2019/2020.
- Smart Mattress System for Bed Occupancy Detection in Elderly Centre. In production, deployed in elderly centre of Yan Chai hospital, also exhibited at Gerontech Exhibition).
- Human Activity Recognition. Responsible for motion data analytics, specifically, the prediction of basic human activities with classical (SVM), and deep learning approach (LSTM, GRU), using motion sensor (3 axis accelerometer) readings.
- Movement Classification of Human Exercises with Deep Learning Approach.
- Deep learning models (LSTM, GRU) that provide real-time predictions of movement classes based on raw motion sensor readings.
- Intelligent Companion for Elderly. Responsible for design, development and validation of Exercise Detection Engine (detecting 5 movement types of human exercises) using deep learning methods.
- Chinese Handwriting Recognition Engine. part of milestone deliverable for a platform project ARD233CP iDMS (Intelligent Data Management System).
- One-shot Facial Recognition - Face identification API with minimum labelled data, using MTCNN face detection, Facenet embedding, DBSCAN, clustering, and nearest neighbour classification. Developed a barebone web demo for face identification.

Surelaw & Xunlei Fengniao Finance **Jul 2016 – Jun 2017**

Data Scientist /Tech Consultant/Project Manager

- Analysed the core issues (data-related) with the given business objective – underwriting and data-driven decision making in peer lending. Orchestrated a set of data-driven solution to support above objective.
- Designed and developed an automated underwriting data product that facilitates the planned business.
- Initiated the project with a budget, team and resource guideline.
- Built a small-scale dev-team. Managed the dev-team for the implementation of an event- driven peer lending platform with WeChat frontend. Ensured various deadlines for deployment and release were met.
- Led the development by overseeing the development process and facilitating the communication between executives and the dev-team.
- Facilitated in drafting a strategic and operational plan for the Big Data analytics and AI transformation.
- Helped the company selecting most suitable Big Data solution via series of due diligence, simulation, analysis and validation.

Institute of Textile and Clothing, The Hong Kong Polytechnic University **Sep 2013 – Dec 2015**

Research Associate

- Posture Correction Girdle for Adolescents with Early Scoliosis
- An innovative body-mapping tank top equipped with biofeedback system for adolescents with early scoliosis
- Successfully drafted and applied patents: US20160220174A1, CN105748037B

Department of Computing, The Hong Kong Polytechnic University **Sep 2008 – Aug 2013**

Research Administrative Assistant, Tutor, Research Assistant

- Thesis: A Distributed Publish-Subscribe Architecture for XML-Based Event Dissemination
- Tangible Social Networking: A Novel Approach to Teaching Computing
- Taught various subjects and lab sessions in computer science (master level and above).

Publications and Awards

Honors and awards:

- Apr 2022: 9th place, ICME Tianchi few-shot logo detection 2022.
- Oct 2021: Silver medal, Google Landmark Retrieval 2021 (in conjunction with ICCV 2021).
- Aug 2020: 2nd place on Google Open image/Robust Vision Challenge 2020 (in conjunction with ECCV 2020).
- Apr 2015: Gold Medal - The 43rd International Exhibition of Inventions Geneva, Switzerland.

- Apr 2015: A Diploma for the high scientific and technological level of the invention, Scientific Community of Romania.

Publications:

- Ka Ho Tong, Ka Wai Cheung and Xiaochuan Yu. ICME 2022 Few-shot LOGO detection top 9 solution, arXiv, June 2022.
- Jiaqi Fan, Junxin Huang, Xiaochuan Yu, and Chao He, Data Subset Selection for Object Detection, Pre-registration workshop paper, NeurIPS, 2020.
- Junxin HUANG, Jiaqi FAN, Xiaochuan YU, and Chao HE, WiseDet 2nd Place Solution for RVC2020, Workshop Paper, ECCV 2020.
- Xiaochuan Yu and Alvin Toong Shoon Chan. Hope: A fault-tolerant distributed Pub/Sub architecture for large-scale dynamic network environment. In proceedings of the 12th IEEE International Conference on Ubiquitous Computing and Communications, IUCC ’13, July 2013.
- Xiaochuan Yu and Alvin Toong Shoon Chan. A hypercubic overlay using bloom-filter based addressing for a non-dedicated distributed tag-based pub/sub system. In proceedings of the 11th IEEE International Symposium on Parallel and Distributed Processing with Applications, ISPA ’13, July 2013.
- Xiaochuan Yu and Alvin Toong Shoon Chan. A hypercubic event-dissemination overlay using structure-aware addressing for distributed xml-based pub/sub system. In Proceedings of the 2012 IEEE 14th International Conference on High Performance Computing and Communication & 2012 IEEE 9th International Conference on Embedded Software and Systems, HPCC ’12, June 2012, pages 179–186.
- Xiaochuan Yu and Alvin Toong Shoon Chan. A time/space efficient xml filtering system for mobile environment. In Proceedings of the 2011 IEEE 12th International Conference on Mobile Data Management - Volume 01, MDM ’11, June 2011, pages 184–193.

Patent:

- Body-sensing Tank Top with Biofeedback System for Patients with Scoliosis, US20160220174 A1 also published as CN1057480.

Interests:

- Large Language Models, Computer Vision, Deep Learning, Internet of Things (IoT), Wearable Computing.

Education

Sep 2008 ~ Oct 2013 – The Hong Kong Polytechnic University, Hong Kong

- PhD in Computer Science and Information Technology

Sep 2006 ~ Jun 2008 – Beijing Jiaotong University, Beijing, China

- M.S.E in Telecommunication and Information System

Sep 2002 ~ Jul 2006 – Beijing Jiaotong University, Beijing, China

- B.S.E in Telecommunication