Tao YANG

Tech Lead, Senior Research Scientist | Tencent America

+1 (480) 347-8414 taoyang.ai@gmail.com linkedin.com/in/taoyang1989 taoyang1989

Research Interests

- Natural Language Processing
- Machine Learning
- Deep Learning
- Data Mining

Professional Experiences

Present - March 2018 — Tech Lead, Senior Research Scientist, Tencent America — Consistently Rated Top Performer

- Lead & manage research team with 10+ members, design & develop AI solutions for products;
- Built robust AI-Chatbot solutions for the Healthcare domain, *e.g.*, QA Chatbot, Triage Chatbot;
- R&D ML models for text feature extraction, semantic matching, triaging, diagnosis, *etc.*;
- Developed an elastic Image-OCR-IR pipeline to build patients' & doctors' user profiles;
- Established a scalable, high-throughput, modular medical NLP platform & Knowledge Graph system;
- Established CI pipelines for AI services with Docker and Kubernetes.

Chatbot | NLP | KG | Text Feature Extraction | Word/Sent/Para/Docu Embeddings | ANN Search | Semantic Matching | Q&A | NER | PoS-Tagging | Summarization | ML/DL models | Docker | Kubernetes

March 2018 - January 2017 — Senior Research Scientist, Baidu US Research

- Participated in Baidu Medical Brain projects;
- Investigated DL models for personalized medical examination recommendation and diagnosis;
- Built NLP & intention modules for understanding Medical Licensing Exam questions;
- R&D efficient approximate nearest neighbor search platform with locality-sensitive hashing.

NLP | KG | NER | PoS-Tagging | ML/DL models | Topic/Intention Models | ANN Search | LSH | HPC Optimizations

Education

Dec 2016 — PhD Computer Science (2016), MS in Computer Science (2013)

Aug 2011 — Arizona State University, School of Computing, Informatics, & Decision Systems Engineering

— *Supervised by Dr. Jieping Ye (PhD & MS) and Dr. Guoliang Xue (PhD)*

- PhD Dissertation: Structured Sparse Methods for Imaging Genetics
- MS Thesis: Machine Learning Methods for High-Dimensional Imbalanced Biomedical Data

2015-2016 — Visiting PhD Scholar, University of Michigan - Ann Arbor

2007-2011 — BEng Software Engineering, Beijing Jiaotong University, China

Selected Projects

AI Chatbots for Healthcare — 2019 - 2021

- AI Chatbot solutions for diverse application scenarios, from single-round to multiple-round conversations;
- **AI-enhanced TaskBots** – screening, triaging, pre-visit info collection Chatbots with AI models guided dynamic dialogue mechanism that support COVID-19 self-check, Tencent Medipedia & Tencent Doctor products;
- **Semantic FAQ/CQA Chatbots** – semantic feature extraction with multiple levels of embeddings; indexing & searching with advanced ANN approaches; fast & accurate semantic matching models; support COVID/Heart Failure/Diabetes QAs;
- Boost Chatbot QA with KG – knowledge generation & injection in MRC [SIGDIAL-2020]; augment QA with KG [AAAI-2020];
- Provided **millions** of online screening services in the early stage of the COVID-19 pandemic, supported **millions** of Chatbot & QA requests in the Heart Failure AI Nurse and Tencent medical products, outperform the SOTA methods.

Chatbot | TaskBot | FAQ/CQA Chatbot | Text Feature Extraction | ANN Search | Semantic Matching | Knowledge Graph | ML/DL models

Image-OCR-IR Pipeline for User Profiles Extraction — 2021

- Establish an Image-OCR-IR pipeline to extract valuable information from users' uploaded images;
- Improve OCR pipeline with a series of IP methods: angle/distortion/tilt correction, dewarping, dewatermark;
- Elastic IR pipelines to tackle different types of images: medical records, prescriptions, lab reports, *etc.*;
- Exploit NLP and KG to perform text correction [Pat.] & structurization; Precisions 95%+, Recalls 90%+.

OCR | Image Correction | Image Classification | IR | NLP | KG | Text Error Inspection & Correction (Glyphic & Phonetic)

Medical NLP Platform and KG — 2018 - 2020

- NER with Bi-LSTM & multi-level attentions [ACL-2019];
- Multi-aspect entity relations understanding [Pat.];
- Specific parsers for prescriptions, lab reports, *etc.*;
- Embedding→modeling: topics, intentions [Pat.], *etc.*

Other AI Model Research — 2018 - 2021

- Disease diagnosis & triaging models [SIGIR-2021, Pat.];
- GNN-based text summarization [peer reviewing];
- Text anomaly detection [peer reviewing].

Publications

Under Review

1. [Mentor] Jing, Baoyu, Zeyu You, **Tao Yang**, Wei Fan, and Hanghang Tong. *Multi-GRAS: Multiplex Graph Neural Networks for Extractive Text Summarization*. Submitted to EMNLP 2021, Under Review.
2. [Mentor] You, Zeyu, Yichu Zhou, **Tao Yang**, and Wei Fan. *Anomaly-Injected Deep Support Vector Data Description for Text Outlier Detection*. Submitted to EMNLP 2021, Under Review.

Conference Proceedings

1. [Mentor] Liu, Zheng, Xiaohan Li, Zeyu You, **Tao Yang**, Wei Fan, and Philip Yu. "Medical Triage Chatbot Diagnosis Improvement via Multi-relational Hyperbolic Graph Neural Network". In: *Proceedings of the 44rd International ACM SIGIR Conference on Research and Development in Information Retrieval (***SIGIR-2021***)*. 2021.
2. [Mentor] Liu, Ye, **Tao Yang**, Zeyu You, Wei Fan, and Philip S. Yu. "Commonsense Evidence Generation and Injection in Reading Comprehension". In: *Proceedings of the 21st Annual Meeting of the Special Interest Group on Discourse and Dialogue (***SIGDIAL-2020***)*. 2020. arXiv: 2005.05240.
3. [Mentor] Shen, Sheng, Yaliang Li, Nan Du, Xian Wu, Yusheng Xie, Shen Ge, **Tao Yang**, Kai Wang, Xingzheng Liang, and Wei Fan. "On the Generation of Medical Question-Answer Pairs". In: *Proceedings of the AAAI Conference on Artificial Intelligence (***AAAI-2020***)*. 2020.
4. [Mentor] Xia, Congying, Chenwei Zhang, **Tao Yang**, Yaliang Li, Nan Du, Xian Wu, Wei Fan, Fenglong Ma, and Philip Yu. "Multi-Grained Named Entity Recognition". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (***ACL-2019***)*. May 2019, pp. 1430–1440.
5. li, Yan, **Tao Yang**, Jiayu Zhou, and Jieping Ye. "Multi-Task Learning based Survival Analysis for Predicting Alzheimer's Disease Progression with Multi-Source Block-wise Missing Data". In: *Proceedings of the 2018 SIAM International Conference on Data Mining (***ICDM-2018***)*. May 2018, pp. 288–296. ISBN: 978-1-61197-532-1. DOI: 10.1137/1.9781611975321.33.
6. **Yang**, **Tao**, Paul Thompson, Sihai Zhao, and Jieping Ye. "Identifying Genetic Risk Factors via Sparse Group Lasso with Group Graph Structure". In: 2017. arXiv: 1709.03645.
7. Li, Qingyang, **Tao Yang**, Liang Zhan, Derrek Paul Hibar, Neda Jahanshad, Yalin Wang, Jieping Ye, Paul M Thompson, and Jie Wang. "Large-scale collaborative imaging genetics studies of risk genetic factors for Alzheimer's Disease across multiple institutions". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (***MICCAI-2016***)*. Springer. 2016, pp. 335–343.
8. Liu, Chaochun, Huan Sun, Nan Du, Shulong Tan, Hongliang Fei, Wei Fan, **Tao Yang**, Hao Wu, Yaliang Li, and Chenwei Zhang. "Augmented LSTM Framework to Construct Medical Self-Diagnosis Android". In: *2016 IEEE 16th International Conference on Data Mining (***ICDM-2016***)*. 2016, pp. 251–260. DOI: 10.1109/ICDM.2016.0036.
9. **Yang**, **Tao**, Jun Liu, Pinghua Gong, Ruiwen Zhang, Xiaotong Shen, and Jieping Ye. "Absolute Fused Lasso and Its Application to Genome-Wide Association Studies". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (***KDD-2016***)*. ACM. Aug. 2016, pp. 1955–1964. DOI: 10.1145/2939672.2939827.
10. Nie, Zhi, **Tao Yang**, Yashu Liu, Qingyang Li, Vaibhav A Narayan, Gayle Wittenberg, and Jieping Ye. "Melancholic depression prediction by identifying representative features in metabolic and microarray profiles with missing values". In: *Pacific Symposium on Biocomputing (***PSB-2015***)*. 2015, pp. 455–466. ISBN: 2335-6936; 2335-6936.
11. **Yang**, **Tao**, Jie Wang, Qian Sun, Derrek P Hibar, Neda Jahanshad, Li Liu, Yalin Wang, Liang Zhan, Paul M Thompson, and Jieping Ye. "Detecting Genetic Risk Factors for Alzheimer's Disease in Whole Genome Sequence Data via Lasso Screening". In: *Proceedings. IEEE International Symposium on Biomedical Imaging (***ISBI-2015***)*. Apr. 2015, pp. 985–989. ISBN: 1945-7928; 1945-8452. DOI: 10.1109/ISBI.2015.7164036.
12. **Yang**, **Tao**, Xinlin Zhao, Binbin Lin, Tao Zeng, Shuiwang Ji, and Jieping Ye. "Automated gene expression pattern annotation in the mouse brain". In: *Pacific Symposium on Biocomputing (***PSB-2015***)*. Vol. 20. 2015, pp. 144–155. ISBN: 2335-6936; 2335-6936.
13. Xiang, Shuo, **Tao Yang**, and Jieping Ye. "Simultaneous Feature and Feature Group Selection Through Hard Thresholding". In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (***KDD-2014***)*. KDD '14. New York, New York, USA: ACM, 2014, pp. 532–541. ISBN: 978-1-4503-2956-9. DOI: 10.1145/2623330.2623662.
14. Manji, Husseini K, Lynn Yieh, Jieping Ye, Yashu Liu, **Tao Yang**, Michael Farnum, Xiang Yao, Willem Talloen, Thomas Steckler, Pim Drinkenberg, et al. "Biomarkers Differentiating Major Depressive Disorder Subtypes". In: *52nd Annual Meeting of* **ANCP** *(American College of Neuropsychopharmacology)*. Vol. 38. 2013, S262–S263.

Journal Articles

1. Lin, Ming, Pinghua Gong, **Tao Yang**, Jieping Ye, Roger L Albin, and Hiroko H Dodge. "Big Data Analytical Approaches to the NACC Dataset: Aiding Preclinical Trial Enrichment". In: **Alzheimer Disease and Associated Disorders** 32.1 (2018), pp. 18–27. DOI: 10.1097/WAD.0000000000000228.

2. Liu, Li, Yung Chang, **Tao Yang**, David P Noren, Byron Long, Steven Kornblau, Amina Qutub, and Jieping Ye. "Evolution-informed modeling improves outcome prediction for cancers". In: **Evolutionary Applications** 10.1 (2017), pp. 68–76. DOI: 10.1111/eva.12417.
3. Liu, Yashu, Lynn Yieh, **Tao Yang**, Wilhelmus Drinkenburg, Pieter Peeters, Thomas Steckler, Vaibhav A. Narayan, Gayle Wittenberg, and Jieping Ye. "Metabolomic biosignature differentiates melancholic depressive patients from healthy controls". In: **BMC Genomics** 17.1 (2016), p. 669. DOI: 10.1186/s12864-016-2953-2.
4. Noren, David P. et al. "A Crowdsourcing Approach to Developing and Assessing Prediction Algorithms for AML Prognosis". In: **PLOS Computational Biology** 12.6 (June 2016), pp. 1–16. DOI: 10.1371/journal.pcbi.1004890.

Books and Chapters

1. Wang, Jie, **Tao Yang**, Pual Thompson, and Jieping Ye. "Chapter 5 - Sparse models for imaging genetics". In: **Machine Learning and Medical Imaging**. Ed. by Guorong Wu, Dinggang Shen, and Mert R. Sabuncu. Academic Press, 2016, pp. 129 –151. ISBN: 978-0-12-804076-8. DOI: 10.1016/B978-0-12-804076-8.00005-0.

Dissertation & Thesis

1. **Yang**, **Tao**. "Structured Sparse Methods for Imaging Genetics". Arizona State University, 2017.
2. — "Machine Learning Methods for High-Dimensional Imbalanced Biomedical Data". Arizona State University, 2013.

Patents

1. "Method and Apparatus for Medical Data Auto Collection Segmentation and Analysis Platform". Pat. US 10,943,673. 2021.
2. "Training Framework for Multi-Group Electrocardiography (MG-ECG) Analysis". Pat. US Patent App. 16/556,491. 2021.
3. "Understanding A Query Intention for Medical Artificial Intelligence Systems Using Semi-supervised Deep Learning". Pat. US Patent App. 16/560,440. 2021.
4. "Explainable Artificial Intelligence Framework for Electrocardiography Analysis". Pat. US Patent App. 16/253,942. 2020.
5. "Machine Learning Model Full Life Cycle Management Framework". Pat. US 11,030,086. 2020.
6. "Method and Apparatus for Natural Language Processing of Medical Text in Chinese". Pat. US Patent App. 16/395,439. 2020.
7. "Method for Determining Disease Symptom Relations Using Acceptance and Rejection of Random Samples". Pat. US Patent App. 16/277,430. 2020.
8. "Proximity Information Retrieval Boost Method for Medical Knowledge Question Answering Systems". Pat. US Patent App. 16/421,554. 2020.
9. "System and Method for Coronary Calcium Deposits Detection and Labeling". Pat. US 11,030,743. 2020.
10. "A Framework for Chinese Text Error Identification and Correction". Pat. Submitted.
11. "Efficient and Compact Text Matching System for Sentence Pairs". Pat. Submitted.
12. "Sp02 Miniprogram: AI Sp02 Measurement App". Pat. Submitted.

SERVICES

Journal Reviewer

- EURASIP Journal on Neurocomputing (2015 - 2020)
- EURASIP Journal on Pattern Recognition (2017, 2020)
- ACM Transactions on Knowledge Discovery from Data (2018 - 2019)
- IEEE Transactions on Knowledge and Data Engineering (2018 - 2019)
- The Annals of Applied Statistics (2018)
- EURASIP Journal on Advances in Signal Processing (2017 - 2018)
- EURASIP Journal on Bioinformatics and Systems Biology (2016)
- EURASIP Journal on Computational Statistics and Data Analysis (2016)
- International Journal of Bioinformatics Research and Applications (2015)

Conference PC & Reviewer

- [PC] AAAI Conference on Artificial Intelligence (AAAI 2020)
- The 13th international AAAI Conference on Web and Social Media (ICWSM-2019)
- [PC] Natural Language Processing and Chinese Computing (2017)
- Advances in Social Networks Analysis and Mining (ASONAM 2017)
- AAAI Conference on Artificial Intelligence (AAAI 2017)
- IEEE International Conference on Healthcare Informatics (ICHI 2015)
- SIAM International Conference on Data Mining (SDM 2014)