Yao Qiang

+1 (313) 329-3094 | yaocsphd@gmail.com | Google Scholar | LinkedIn

RESEARCH EXPERIENCE

Trustworthy AI Lab at Wayne State University, Graduate Research Assistant — Aug 2019 – Present

- **Explainability**
  - Proposed a novel explanation technique via attentive class activation tokens named AttCAT, leveraging encoded features, their gradients, and their attention weights to generate a faithful and confident explanation for the Transformer's output. Published one top conference paper in **NeurIPS-2022** [1] as the first author.
  - Proposed an interpretability-aware variant of Vision Transformer (ViT) named IA-ViT, which consists of three major components: a feature extractor, a predictor, and an interpreter. By training the predictor and interpreter jointly with a novel interpretability-aware objective, IA-ViT achieves improved interpretability without a significant loss in predictive performance. Submitted one top conference paper in **AAAI-2024** [8] as the first author.
  - Developed a novel DNN model explanation method named NeFLAG. This approach converts a volume integration of the second-order gradients to a surface integration of the first-order gradients by applying the divergence theorem, resulting in more faithful explanations. Published one conference paper in **IJCAI-2023** [11].
- **Fairness**
  - Proposed counterfactual interpolation augmentation (CIA), a novel data augmentation strategy to improve DNN fairness via de-correlating the target variable from the sensitive attribute based on counterfactual causal inference. Designed counterfactual gradient integration leveraging the counterfactual interpolations from CIA to generate high-quality and fair explanations. Published one conference paper in **IJCAI-2022** [2] as the first author.
  - Developed a novel framework, named debiased self-attention (DSA), which is a fairness-through-blindness approach that enforces ViT to eliminate spurious features correlated with the sensitive attributes for bias mitigation. DSA leads to improved fairness guarantees over prior works on multiple prediction tasks without compromising target prediction performance. Submitted one top conference paper in **AAAI-2024** [5] as the first author.
- **Robustness**
  - Proposed a certifiably robust defense technique named GradMASK, which improves the robustness of tiny models against both character-level and word-level adversarial attacks, by incorporating masked adversarial examples during the knowledge distillation process. Published one conference paper in **IJCNN-2022** [3] as the first author.
  - Proposed a novel saliency-guided adversarial training scheme for learning generalizable features. This approach demonstrates impressive performance on OOD test sets compared to the baseline methods. Published one conference paper in **ICML-2022** Workshop [14].
- **Natural Language Processing**
  - Proposed an aspect-based sentiment analysis model with self and position-aware attention mechanisms. This model can extract both aspect level and overall sentiments from text and explain the prediction based on polarity among different aspects. Published one conference paper in **IJCNN-2020** [4] as the first author.
  - Introduced a novel information retrieval task called Conversational Entity Retrieval from a Knowledge Graph (CER-KG) and created QBLink-KG, a publicly available benchmark. Designed a feature-based neural architecture for entity ranking in CER-KG, considering various lexical and semantic matching signals within the knowledge graph and dialog history. Submitted one top conference paper in **EMNLP-2023** [9] as the first author.
- **Foundational Machine Learning**
  - Proposed a DNN training method named In-Training Representation Alignment. To reduce mini-batch over-adaption during the training, this method augments SGD with regularization by explicitly aligning feature distributions of two different mini-batches. Published one conference paper in **AAAI-2023** [10].
  - Proposed Prox-DRO, a proximal algorithm to solve compositional optimization problems that often arise in distributionally robust optimization formulations. Prox-DRO circumvents the need for large accuracy-dependent batch gradients and function evaluations, demonstrated to be feasible for most practical settings. Published one conference paper in **ICML-2023** Workshop [13].
- **Medical Imaging Segmentation**
  - Designed a new focal transformer-based image segmentation architecture for CT images, efficiently extracting local features and global context. Introduced an auxiliary boundary-aware regression task alongside the main segmentation task to improve clarity in boundaries. Published one conference paper in **MICCAI-2023** [12].
  - Proposed a novel method, AutoSAM Adapter, specifically for 3D multi-organ CT-based segmentation. Employed parameter-efficient adaptation techniques to facilitate the transformation of the SAM model's capabilities, eliminating the need for manually generated prompts. Submitted one top conference paper in **AAAI-2024** [15].

Mike Ilitch School of Business at WSU, Student Research Assistant — Sep 2018 – Aug 2019

- Proposed programming solutions using Python and R for strategic management research projects, i.e., virtual currency, firm diversity, and stock sentiment analysis.

WORK EXPERIENCE

Amazon, Applied Scientist Intern — May 2023 – Aug 2023

- Conducted comprehensive research on evaluating the capabilities of LLMs in generating structured hypotheses, specifically focused on intent classification (IC) and slot filling (SF) for Alexa API call tasks. Demonstrated a significant performance gap between LLMs and SOTA discriminative models (e.g., JointBERT), through extensive experiments with various LLMs, including ChatGPT, LLaMA, Flan-T5, etc.
- Applied prompt engineering techniques for supervised fine-tuning (SFT) of LLMs on IC-SF tasks; Successfully reduced the performance gap between LLMs and JointBERT from 22% (48%) to 1% (0.5%) in IC accuracy (SF F1-score).
- Evaluated the robustness of LLMs under various prompt perturbations, i.e., synonyms, oronyms, and paraphrases. Demonstrated that LLMs exhibit vulnerability to these perturbations, resulting in an average performance drop rate of 13.07% (22.20%) in IC accuracy (SF F1-score).
- Proposed two mitigation strategies, i.e., perturbation consistency learning and data augmentation, recovering up to 59% (69%) performance drop in IC (SF) task.
- Finished writing a research paper manuscript.

Xi'an Microelectronics Technology Institute, Hardware Design Engineers — Aug 2010 – Dec 2017

- Conducted computer product and software design and test for manufacturing. Designed printed circuit boards (PCBs) and performed product simulations to optimize layouts and validate designs for maximum efficiency. Analyzed PCBs for signal integrity and power integrity, ensuring high-quality signal transmission and stable power distribution.

EDUCATION

Wayne State University, Detroit, Michigan, USA

- Doctor of Philosophy in Computer Science — Sep 2019 – Expected Spring 2024
  - Cumulative GPA: 3.95 / 4.0
  - Awards/Honors: Michael E. Conrad Award, Outstanding Graduate Teaching Assistant

Wayne State University, Detroit, Michigan, USA

- Master of Science in Computer Science — Sep 2018 – Dec 2019
  - Cumulative GPA: 3.95 / 4.0
  - Awards/Honors: Graduate School Master's Scholarship Award

Xidian University, Xi'an, China

- Bachelor of Science in Computer Science — Sep 2006 – Jul 2010

SKILLS

Tools: SQL, Matlab, Git, Scikit-learn, Amazon Web Services (AWS)
Programming: Python, R, C++, Java, PyTorch, TensorFlow, PySpark, HuggingFace
Machine Learning & AI: Large Language Models, Robustness, Explainability, Fairness, Computer Vision, Natural Language Processing, Unsupervised/self-supervised Learning, Medical Imaging
Languages: Mandarin (native), English (business fluent)

PUBLICATIONS AND PRE-PRINTS

[1] **Qiang, Y.**, Pan, D., Li, C., Li, X., Jang, R. and Zhu, D. “Attcat: Explaining transformers via attentive class activation tokens”. NeurIPS 2022.

[2] **Qiang, Y.**, Li, C., Brocanelli, M. and Zhu, D. “Counterfactual interpolation augmentation (CIA): A unified approach to enhance fairness and explainability of DNN”. IJCAI 2022.

[3] **Qiang, Y.**, Kumar, S.T.S., Brocanelli, M. and Zhu, D. “Tiny rnn model with certified robustness for text classification”. IJCNN 2022.

[4] **Qiang, Y.**, Li, X. and Zhu, D.“Toward tag-free aspect based sentiment analysis: A multiple attention network approach”. IJCNN 2020.

[5] **Qiang, Y.**, Li, C., Khanduri, P. and Zhu, D. “Fairness-aware Vision Transformer via Debiased Self-Attention”. arXiv:2301.13803 [cs.LG].

[6] **Qiang, Y.**, Kumar, S.T.S., Brocanelli, M. and Zhu, D. “Adversarially Robust and Explainable Model Compression with On-Device Personalization for Text Classification”. arXiv:2101.05624 [cs.LG].

[7] **Qiang, Y.**, Nandi, S., and Galstyan, A. “Prompt Perturbation Consistency Learning (PPCL) for Robust Language Models”. Under-review.

[8] **Qiang, Y.**, Li, C., Khanduri, P. and Zhu, D. “Interpretability-Aware Vision Transformer”. Under-review.

[9] **Qiang, Y.**, Kotov, A., Nikolaev, F., Zamiri, M., and Zhu, D. “Benchmark and Neural Architecture for Conversational Entity Retrieval from a Knowledge Graph”. Under-review.

[10] Li, X., Li, X., Pan, D., **Qiang, Y.** and Zhu, D. “Learning compact features via in-training representation alignment”. AAAI 2023.

[11] Li, X., Pan, D., Li, C., **Qiang, Y.** and Zhu, D. “Negative Flux Aggregation to Estimate Feature Attributions”. IJCAI 2023.

[12] Li, C., **Qiang, Y.**, Bagher-Ebadian, H., Goddla, V., Chetty, I.J. and Zhu, D. “FocalUNETR: A Focal Transformer for Boundary-aware Prostate Segmentation using CT Images”. MICCAI 2023

[13] Khanduri, P., Li, C., Sultan, R.I., **Qiang, Y.**, Kliewer, J. and Zhu, D. “Proximal Compositional Optimization for Distributionally Robust Learning”. ICML 2023 New Frontiers in Adversarial Machine Learning Workshop.

[14] Li, X., **Qiang, Y.**, Li, C., Liu, S. and Zhu, D. “Saliency guided adversarial training for learning generalizable features with applications to medical imaging classification system”. ICML 2022 New Frontiers in Adversarial Machine Learning Workshop.

[15] Li, C., Khanduri, P., **Qiang, Y.**, Sultan, R.I., Chetty, I. and Zhu, D. “Auto-Prompting SAM for Mobile Friendly 3D Medical Image Segmentation”. arXiv:2308.14936 [cs.LG].