Yao Qiang

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

RESEARCH INTERESTS

- Natural Language Processing (NLP) & Large Language Model (LLM)
- Trustworthy AI: Fairness, Explainability, Robustness
- Machine Learning & Applications

WORK EXPERIENCE

Amazon, Applied Scientist Intern — May 2023 – Aug 2023

- Conducted comprehensive research on evaluating the capabilities of LLMs in generating structured hypotheses for Alexa API call tasks through extensive experiments with various LLMs, including ChatGPT, LLaMA, Flan-T5, etc.
- Applied prompt engineering techniques for supervised fine-tuning (SFT) of LLMs; 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.

RESEARCH EXPERIENCE

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

- **Natural Language Processing & Large Language Model**
  - Evaluated and improved the robustness of LLMs in generating structured hypotheses against perturbed prompts [7].
  - Proposed a novel hijacking attack algorithm to successfully induce LLMs to generate targeted malicious outputs under in-context learning (ICL) [9].
  - Introduced a novel information retrieval task and created a publicly available benchmark. Designed a feature-based neural architecture for entity ranking [10].
  - Proposed a certifiably robust defense technique improving the robustness of tiny RNN models against both character-level and word-level adversarial attacks by incorporating masked adversarial examples during the knowledge distillation process **IJCNN-2022** [3].
  - Proposed an aspect-based sentiment analysis model with self and position-aware attention mechanisms **IJCNN-2020** [4].
- **Trustworthy AI: Fairness, Explainability, Robustness**
  - Proposed a novel explanation technique via attentive class activation tokens to generate faithful and confident explanations of the Transformer's output **NeurIPS-2022** [1].
  - Proposed an interpretability-aware Vision Transformer (ViT) resulting in improved interpretability via a novel interpretability-aware objective [8].
  - Developed a novel DNN model explanation method leveraging the second-order gradients to a surface integration of the first-order gradients by applying the divergence theorem resulting in more faithful explanations **IJCAI-2023** [11].
  - Proposed a novel data augmentation strategy to improve DNN fairness via de-correlating the target variable from the sensitive attribute based on counterfactual causal inference **IJCAI-2022** [2].
  - Developed a novel fairness-through-blindness framework enforcing ViT to eliminate spurious features correlated with the sensitive attributes for bias mitigation [5].
  - Proposed a novel saliency-guided adversarial training scheme for learning generalizable features leading to impressive performance on OOD test sets **ICML-2022** Workshop [15].
- **Machine Learning & Applications**
  - Proposed a novel in-training representation alignment approach reducing mini-batch over-adaption **AAAI-2023** [11].
  - Proposed a proximal algorithm circumventing the need for large accuracy-dependent batch gradients and function evaluations to address the compositional optimization problems **ICML-2023** Workshop [14].
  - Designed a new focal transformer-based image segmentation architecture for CT images, efficiently extracting local features and global context **MICCAI-2023** [13].
  - Proposed a novel method for 3D multi-organ CT-based segmentation employing parameter-efficient adaptation techniques to facilitate the transformation of the SAM model's capabilities [16].

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.**, et al., "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". arXiv:2101.05624 [cs.LG].
[9] **Qiang, Y.**, Zhou, X. and Zhu, D. "Hijacking Large Language Models via Learning Adversarial In-Context Examples". under-review.
[10] **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.
[11] Li, X., Li, X., Pan, D., **Qiang, Y.** and Zhu, D. "Learning compact features via in-training representation alignment" AAAI 2023.
[12] Li, X., Pan, D., Li, C., **Qiang, Y.** and Zhu, D. "Negative Flux Aggregation to Estimate Feature Attributions" IJCAI 2023.
[13] 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
[14] 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.
[15] 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.
[16] 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].