Xiang Li

Tianjin University
Master's Degree in Electronic Information Engineering
Address: Beijing, China

+86-18810589962
lixiang_eren@tju.edu.cn
GitHub Profile
Google Scholar

Education

•College of Intelligence and Computing, Tianjin University *Project 985*
Master's Degree in Electronic Information Engineering 2020-2023

•College of Chemical Engineering, Beijing University of Chemical Technology *Project 211*
Bachelor's Degree in Chemical Engineering 2016-2020

Research Experiences

•Disentangled and Robust Representation Learning for Bragging Classification in Social Media *2023*

***Xiang Li**, Yucheng Zhou.*

Accepted by IEEE International Conference on Acoustics, Speech, and Signal Processing(CCF-B)

- **Backgrounds & Motivations:**
  As online communication on social media becomes more prevalent and important in human life, bragging (or self-promoting) categorization has become an important topic in computational (socio)linguistics. It has been widely used in academia and industry, for example, to help linguists delve into the context and types of bragging, and to support social scientists in studying the relationship between bragging and other characteristics (e.g., gender, age, economic status, occupation), to enhance online users' self-presentation strategies. However, existing bragging classification datasets suffer from a serious data imbalance issue.
- **Methods & Results:**
  We propose a novel bragging classification method with disentangle-based representation augmentation and domain-aware adversarial strategy. The experimental results show that our method achieves a new SOTA performance, and the macro-F1 score increases by 4.27% compared with the previous SOTA.

•Impromptu Cybercrime Euphemism Detection *2023*

***Xiang Li**, Yucheng Zhou.*

- **Backgrounds & Motivations:**
  As criminals use euphemisms to evade regulation and scrutiny when communicating online, detecting euphemisms for cybercrime is of great significance to combat cybercrime and purify cyberspace. Impromptu dark web euphemisms refer to new euphemisms created during criminals' communication, which are highly time-sensitive and exist in very small numbers in the corpus. Existing methods cannot detect them.
- **Methods & Results:**
  We propose a detection framework consisting of coarse and fine-grained classification models. Moreover, we propose context augmentation and muti-turn iterative training for the fine-grained classification model and collect a development set via ChatGPT to alleviate the overfitting problem in model training. Experimental results show that the approach achieves a significant 76-fold improvement compared to previous SoTA methods.

•Jailbreaking GPT-4V via Self-Adversarial Attacks with System Prompts *2023*

*Yuanwei Wu, **Xiang Li**, Yixin Liu, Pan Zhou, Lichao Sun.*

Submitted to ACL2024

- **Backgrounds & Motivations:**
  Jailbreak is an attack method that breaks through the security limitations of a large language model and makes it output harmful content. It is helpful to study the jailbreak method to improve the security of large language models. Most of the existing methods are white-box attacks, and the model weight needs to be obtained to carry out the attack. But the commercial large models that are widely used by the public are often black boxes, such as GPT-4V.
- **Methods & Results:**
  We discover a system prompt leakage vulnerability in GPT-4V. Through carefully designed dialogue, we successfully steal the internal system prompts of GPT-4V. Based on the acquired system prompts, we propose a novel MLLM jailbreaking attack method termed SASP(Self-Adversarial Attack via System Prompt). Furthermore, in pursuit of better performance, we also add human modification based on GPT-4's analysis, which further improves the attack success rate to 98.7%. Also, We evaluated the effect of modifying system prompts to defend against jailbreaking attacks. Results show that appropriately designed system prompts can significantly reduce jailbreak success rates.

•**Visual In-Context Learning for Large Vision-Language Models** *2024*

*Yucheng Zhou, **Xiang Li**, Qianning Wang, Jianbing Shen*

Submitted to ACL2024

– **Backgrounds & Motivations:**
In-Context Learning is a method that allows the model to produce the correct output by simply changing the input prompt without updating the model parameters.Due to the large number of parameters of large models and the large demand for computing power for parameter updates, it is of great value to study in-context learning for the application of large models. However, many ICL methods that are applied to LLMs are greatly reduced or even ineffective in LVLM.

– **Methods & Results:**
Firstly, we analyzed the cause of the problem. This is because 1) the information interaction of visual language modalities occurs at a deeper level of LVLM; 2) The eigenvectors of vision and language are in different positions in space. And then, based on these two observations, we design a retrieval and cross-modal ICL methods, which greatly improve the accuracy

Industrial Experiences

•**GRDI CO., LTD.** *Feb 2022 - May 2022*

Intern

– Backend Development: Python, Flask, MySQL, MongoDB.
I maintain a Python backend built on top of Flask, do some CURD, and write some functional interfaces based on python.

– Knowledge Graph Development: Python, NEO4J
I maintain a threat intelligence knowledge graph and do some data processing.

•**BEIJING ABT NETWORKS CO., LTD.** *Aug 2020 - Apr 2022*

Intern

– Crawler Development: Scrapy
Write a distributed crawler based on Scrapy for scraping open-source threat intelligence, including indicators of compromise and unstructured text.

– Information Extractor Development: PyTorch, Bi-LSTM
I built and trained a NER model based on BiLSTM to extract useful information from unstructured intelligence text.

•**INSTRUCT AI CO., LTD.** *December 2023 - March 2024*

Intern

– LLM Agent Development: Langchain
I write prompts for roleplay models, develop langchains, and do some data processing. I review academic papers in the field of role-playing large models and present my findings and insights at company meetings.

Certificates & Award

•**IELTS Exam:** Overall 7.0, Listening 8.0, Reading 8.0, Speaking 6.0, Writing 6.5 *May 2022*

•**NumerAi Competition:** Gold Medal(top1%)*1 Silver Medal*3(top5%), Bronze Medal*3(top15%) *Jun 2023*