Wanlong Fang (方万隆)

Research Interest: multi-modal learning (e.g., temporal sentence grounding), large language model (e.g., multi-modal LLM), computer vision (e.g., object detection) and natural language processing (e.g., text classification)
Professional Activities: ACM MM 2024 Reviewer, EMNLP 2023 reviewer, etc.
Homepage: Link   Google Scholar: Link   Github: Link   Email: wanlongfang@gmail.com
Tel.: (86)-18003818127   Address: School of Software, Henan University, Kaifeng, Henan Province, China

Education and Research Experience

Huazhong University of Science and Technology (***Top 10 University in China***)
Research Assistant (Supervisor: Pan Zhou, 2021- now)

Henan University (***Double first-class university in China***)
B.Eng. in Software Engineering (2019-2023)   ***GPA***: *86.01/100 (**1**st in 392)*

Publications about Artificial Intelligence

[1] Fewer Steps, Better Performance: Efficient Cross-Modal Clip Trimming for Video Moment Retrieval Using Language, Accepted by AAAI Conference on Artificial Intelligence 2024 (**co-first author**)

[2] Annotations Are Not All You Need: A Cross-modal Knowledge Transfer Network for Unsupervised Temporal Sentence Grounding, Accepted by Findings of Empirical Methods in Natural Language Processing 2023 (**co-first author**)

[3] Towards Robust Temporal Activity Localization Learning with Noisy Labels, Accepted by in International Conference on Computational Linguistics 2024 (co-author)

[4] Any-shot Compressed-domain Temporal Sentence Grounding, Under Review in IEEE Transactions on Pattern Analysis and Machine Intelligence (**co-first author**)

[5] Multi-Query Temporal Sentence Grounding via Co-Tokenization Cross-Modal Multi-Stream Network, Under Review in Conference on Computer Vision and Pattern Recognition 2024 (**co-first author**)

Selected Awards & Honors

- National encouragement scholarship (award ratio: **3%**) in 2020
- Merit student (award ratio: **7%**) in 2020
- Outstanding award for innovation and entrepreneurship among college students (award ratio: **3%**) in 2022
- Merit student (award ratio: **7%**) in 2023
- Henan University Scholarship (award ratio: **8%**) in 2023
- Merit undergraduate graduates (award ratio: **7%**) in 2023

Implemented Open-Source Projects about Artificial Intelligence

- **A Multi-modal LLM Framework for Text-rich Visual Question Answering (Link)**
  This project develops a multi-modal large language model (LLM) framework for the significant yet challenging visual question answering (VQA) task. VQA is a task of generating natural language answers when a question in natural language is asked related to an image. Considering that images with text are prevalent in many real-world scenarios, it is essential for human visual perception to comprehend such textual content. Therefore, the motivation is to introduce the LLM technology to enhance the understanding and interpretation of text within images for the VQA task. The main method leverages both learned query embeddings and encoded patch embeddings to improve the understanding of text within images and enhance the capabilities of multi-modal language models. Specially, I enhance image understanding by combining learned query embeddings and image-encoded patch embeddings. The utilized model consists of a vision encoder, a Q-Former module, and a projection layer. In the pre-training stage, I pre-train the projection layer on image-text pairs from various datasets, which can align the visual encoder and the language model. In the fine-tuning stage, I first initialize the Q-Former module from InstructBLIP, and then fine-tune Q-Former and the projection layer. The vision tower encodes visual representations into patch embeddings, which are then sent to the Q-Former to extract refined query embeddings. The projection layer allows the language model to grasp rich visual knowledge. The combined visual embeddings and question text embeddings are fed to the language model for inference. Thus,

this method can effectively address the constraint of image information by a multi-modal LLM framework for understanding text-image visual perception in the challenging VQA task.

- **Fast Temporal Sentence Grounding (Link)**
This project implements an effective and efficient framework for the fast temporal sentence grounding (TSG) task. TSG aims to localize a target segment in an untrimmed video semantically according to a given sentence query. Conventional TSG methods follow the top-down or bottom-up strategy with a time-consuming framework, which is inefficient and inflexible for a large number of untrimmed videos in real-world applications. This project presents an end-to-end framework that models hours-long videos in a single network execution. Specially, the framework is structured in a coarse-to-fine manner, where context knowledge is extracted from non-overlapping video clips (anchors), followed by the supplementation of highly responsive anchors to the query for detailed content knowledge. Therefore, the introduced approach enhances efficiency and enables the capture of long-range temporal correlations in overlong videos for more precise video grounding.

- **An Effective and Efficient Approach for Human Pose Estimation (Link)**
The project focuses on the human pose estimation task, which identifies and classifies the poses of human body parts and joints in images or videos. The project presents an efficient single-stage approach based on YOLOv5 (You Only Look Once version 5), which excels at estimating the poses of multiple individuals within complex scenes. Specially, I model both keypoints and poses as distinct objects, and then meticulously analyze these objects within a dense anchor-based detection framework. Since the presented method leverages YOLOv5 for object detection on the video, the approach can accurately and efficiently detect the human keypoints in the given image/video. By the parameter-efficient tuning and data augmentation techniques, I can train efficiently the utilized model with YOLOv5 on PyTorch.

- **A Bert-based Method for Offensive Language Detection (Link)**
The Natural Language Processing project targets the challenging offensive language detection task that aims to detect whether a text is considered offensive or non-offensive. The key methodology involves utilizing BERT's contextual embeddings and pre-trained language understanding to capture the nuanced features associated with offensive language. The model is fine-tuned on labeled datasets containing instances of offensive language, enabling it to generalize and identify offensive patterns across various contexts.

- **DSAIPC: Data Structures and Algorithms Implemented in Python and C++ (Link)**
Data structures and algorithms are the cornerstone of computer science and software engineering. Their mastery is vital for developing efficient software and solving complex problems. This project serves as an organized and methodical showcase of my rigorous study and practical implementation of core data structures and algorithms. It encapsulates a wide spectrum of implementations, each tailored to resolve specific computational challenges with utmost efficiency and precision. Within these projects, you will find a comprehensive display of my exploration into the domain of data structures and algorithms, implemented meticulously in C++ and Python.

Skills

- **Data analysis:** Experienced in data manipulation, analysis, and visualization using pandas, NumPy, and Matplotlib, machine learning algorithms and data mining techniques.
- **Programming languages:** Python, Java, C++, MATLAB, SQL.
- **English:** GRE (V=165/170, Q=170/170, AW=3.5/6.0), TOEFL (105/120).
- **Programming software and tools:** Jupyter, Anaconda, Github; Proficient in IntelliJ IDEA, PyCharm and Clion.

Advantages

- **Clear plan:** During my Ph.D. period, I want to study the research area of artificial intelligence, such as **multi-modal learning** and **large language model**. Also, I will publish multiple top-tier papers as the first author. After graduation, I will look for a faculty job in a top university.
- **High self-motivation & stress resistance:** I like to participate in various competitions and achieve satisfactory performance. Besides, I can effectively relieve the stress by running and swimming.
- **Strong independent thinking ability:** When facing a complex research task, I can decompose it into several small tasks. By consulting relevant information, I can solve these small tasks one by one independently.
- **Satisfactory self-study ability:** For a new field of research, I am able to search for related works on my own. Based on these works, I can design and implement my proposed model independently.
- **Solid foundation in mathematics:** I received **excellent grades** in advanced mathematics, linear algebra and probability and mathematical statistics. Satisfactory performance shows my solid mathematical foundation.