SIPENG ZHENG

Linkedin: spzheng@baai.ac.cn
Email: spzheng@baai.ac.cn
Cellphone: +86 15905058181

RESEARCH INTEREST

My name is Sipeng Zheng. Currently I work as a researcher in Multimodal Interaction Research Group, Beijing Academy of Artificial Intelligence(BAAI). I obtained my PhD and bachelor degree from Renmin University of China (RUC), advised by Prof. Qin Jin. My research interest focuses on human behavior understanding and learning algorithms for embodied agents. Relevant tasks what I'm interested in includes visual relation detection, scene graph generation, action recognition as well as multi-modal learning and multi-modal pre-training for the above tasks. In addition to these tradition vision and language tasks, recently, I'm exploring how to apply multi-modal pre-training for embodied agents to autonomously acquire the skills to cooperate and compete various tasks in our daily life.

EDUCATION

Information Institute of Renmin University of China *Sep. 2018 - Jun. 2023*
Ph.D in Computer Science Advisor: Qin Jin

Information Institute of Renmin University of China *Sep. 2014 - Jun. 2018*
B.S in Computer Science Advisor: Qin Jin

EXPERIENCE

Researcher at Beijing Academy of Artificial Intelligence *July. 2023 - Present*
Multimodal Interaction Research Group

Research Internship at Beijing Academy of Artificial Intelligence *Feb. 2023 - May.2023*
Multimodal Interaction Research Group

Research Internship at MSRA *Feb. 2022 - Oct.2022*
Multimedia Searching and Mining Group

Research Internship at Beijing Academy of Artificial Intelligence *Sep. 2021 - Jun. 2022*
Multi-modal and Multi-lingual Large-scale Pre-train

Research Internship at BianLiFeng *Apr. 2017 - Oct. 2017*

RESEARCH PROJECTS

Multi-modal Pre-training for Motor Control *Apr. 2022 - Present*

- By regarding motor control as sequence prediction, I study how multi-modal representations pre-trained on diverse ego-centric videos can greatly improve downstream robotic manipulation tasks (intern project).

Temporal Sentence Grounding of Long-form Videos *Jan. 2021 - Present*

- Grounding the location of a sentence in an untrimmed video can be actually regarded as a matching problem between video clips and sentences. However, the speech-to-noise ratio between visual and language modalities naturally does not match (language's is much higher than video's generally). Considering such mismatch, we propose an adaptive multi-scale local attention mechanisms, which is under review in ICCV 2023 (co-author).

Multi-modal Video Relation Detection *Nov. 2021 - Apr. 2022*

- I build a robust relation system with a temporal-aware relational predictor and improve the tracking module with better efficiency in multi-person scenarios, which is ranked 2nd in **ACM MM** Video Relation Understanding Grand Challenge. Then I propose the first end-to-end transformer-based model for video relation detection, which is distinct from previous multi-stage pipeline and is published in **CVPR 2022 Oral** (1st author).

Multi-modal Image Relation Understanding *Jun. 2018 - Present*

- **Image Relation Detection**. I propose a multi-level attention method to fuse multiple modalities (i.e. visual appearances, spatial locations, textual semantics) for relation prediction, which is published in **ACM MM 2019** (1st author).
- **Human-Object Interaction (HOI)** The HOI task only focuses on the relation between human and objects. I propose a graph-based method to effectively utilize human skeletons for fine-grained human-object relation prediction, which is published in **ICME 2020**. I also notice that the HOI task has long been plagued by the lack of data and supervision due to the large number of possible combinations of verbs and objects in real life. Models trained on limited data can only predict HOIs within a set of fixed categories. We therefore propose an open-category HOI pre-training model with specially designed proxy tasks that can well generalize to novel interaction categories. This work will be published in **CVPR 2023** (1st author).

Few-shot Action Recognition *Dec. 2020 - Aug. 2021*

- I propose a hierarchical matching approach to project videos into a metric space and classify videos via nearest neighboring, which supports comprehensive similarity measure at global, temporal and spatial levels. To avoid the lack of supervisions on fine-grained spatio-temporal matching, we propose a hierarchical contrastive learning algorithm. This work is published in **ECCV 2022** (1st author).

PUBLICATION

1. **Sipeng Zheng**, Boshen Xu, and Qin Jin. Open-category human-object interaction pre-training via language modeling framework. In *CVPR*, 2023
2. **Sipeng Zheng**, Shizhe Chen, and Qin Jin. Few-shot action recognition with hierarchical matching and contrastive learning. In *ECCV*, 2022
3. **Sipeng Zheng**, Shizhe Chen, and Qin Jin. Vrdformer: End-to-end video visual relation detection with transformers. In *CVPR*, pages 18836–18846, 2022
4. **Sipeng Zheng**, Shizhe Chen, and Qin Jin. Skeleton-based interactive graph network for human object interaction detection. In *ICME*, pages 1–6, 2020
5. **Sipeng Zheng**, Shizhe Chen, and Qin Jin. Visual relation detection with multi-level attention. In *ACM MM*, pages 121–129, 2019
6. **Sipeng Zheng**, Xiangyu Chen, Shizhe Chen, and Qin Jin. Relation understanding in videos. In *ACM MM*, page 2662–2666, 2019
7. **Sipeng Zheng**, Qi Zhang, and Qin Jin. Exploring anchor-based detection for ego4d natural language query. In *arxiv*, 2022
8. Bei Liu, **Sipeng Zheng**, Jianlong Fu, and Wen-Huang Cheng. Anchor-based detection for natural language localization in ego-centric videos. In *IEEC*, 2022

AWARDS

- ⋆ National Scholarship for Ph.D Students. 2022
- ⋆ Ranked 3th in Facebook CVPR 2022 Ego4D Natural Language Query Challenge. 2022
- ⋆ Ranked 4th, NIST TRECVID 2021 Ad-hoc Video Search (AVS) Challenge. (20+ teams) 2021
- ⋆ Ranked 1st in CVPR 2021 ActivityNet Entities Object Localization Challenge. 2021

- ⋆ Ranked 4th in CVPR 2021 HOMAGE Scene-graph Generation Challenge. 2021
- ⋆ Ranked 2th in ACM MM 2020 Video Relationship Understanding Grand Challenge. 2020
- ⋆ Ranked 2nd and won the Best Method Prize in ACM MM 2019 Video Relationship Understanding Grand Challenge. 2019
- ⋆ First Class Scholarship for Ph.D Students. 2018-2021
- ⋆ First Prize in National University Mathematical Modeling Competition. 2015

SERVICE

- ⋆ Conference Reviewer for CVPR, ICCV, ECCV, ACM MM, ICME, etc.
- ⋆ Teaching Assistant Introduction to the Theory of Computation 2020
- ⋆ Teaching Assistant Probability and Statistics 2019
- ⋆ Teaching Assistant Spoken Language Processing 2018

PARTICIPATED GRANTS

- ⋆ National Key Research and Development Plan No. 2016YFB1001202 (*Computational Principles of Human-Computer Interaction*)
- ⋆ Beijing Natural Science Foundation No. 4192028 (*Language Understanding and Interaction Based on Auditory Information*)