Applicant: Qianqi Yan   Intended Program: Ph.D. in Computer Science

Statement of Purpose

When I look back at my undergraduate years, I see how my experiences lead me to the doorway of pursuing a research career in **natural language processing (NLP), especially language grounding in embodied AI**. I believe language grounding and language-vision multimodality are essentials in building intelligent embodied AI agents to perform real-world multimodal tasks. My project ‘Grounding Language in Robotic Perception and Affordance” got presented at the 2022 Microsoft Research Summit. Another work “Hierarchical Semantic Segmentation” has utilized language to boost visual intelligence by 4% compared to state-of-the-art works, and I am currently preparing the conference submission of the project “Language-Aided Object Detection”. The experience in these projects gave me the vision and ability to push forward the frontier NLP topics through their high correlation with language grounding and language-vision tasks. It also prepares me for joining UW's Ph.D. program in Computer Science and Engineering through which I will bring innovation and vitality to the CSE community.

My previous experience in building robots motivated me to look deeper into machine intelligence. During high school, I built a biomimetic robot fish that received first place in the teenage innovation contest in Shanghai. After entering Shanghai Jiao Tong University (SJTU) as an ECE student, I was funded to design a service robot to be applied in hospital wards. With the completion of the projects, I kept thinking about potential optimization in efficiency and performance that could be achieved if the machines could be embraced with more intelligence - if they could **gradually learn based on experience or common sense**, if they could **communicate with humans about the world in natural language**, or if they could even **utilize language to help with visual intelligence**. I felt the necessity for systematic learning and applied for a dual degree in computer science and a math minor at the University of Michigan (UM). At UM, I dedicated all my credits to related courses and research.

Research in Grounding Language in Robotic Perception and Affordance. With the intuition that humans, or even animals, have dedicated spatial memory like a map as well as common sense knowledge to navigate themselves and manipulate tasks in familiar environments, I worked on a project incorporating memory and utilizing common sense knowledge encoded in large language models (LLM) to improve household agents' long-term performance under the supervision of Prof. Joyce Chai.

I broke down this long-standing yet challenging goal into boosting performance in navigation and manipulation, respectively. The challenging part comes with the latter. When pre-trained LLMs are queried to generate actionable plans based on the goal state and feedback from the environment using their common sense knowledge, they tend to give responses that lack progressive relationships and are even repetitive. Using the generative language model as a black box in a zero-shot manner, we cannot ground specific task-planning knowledge into the model itself. To address this dilemma, I proposed a switch to few-shot learning by introducing a few examples in the prompt for the model to learn the progressive pattern. This improvement solves most but the repetition issue, even with larger models such as GPT-3. After identifying the algorithm's assumptions and investigating relevant papers, I further found that the original method of ranking by string probability could be problematic due to surface form competition. I proposed a refined scoring function based on domain conditional PMI. *This work got presented* *at the 2022 Microsoft Research Summit.* This project motivated me to dive deeper into common sense encrypted in LLMs and gave me a glance at the frustration and accomplishment a researcher may experience all the failed attempts, exploring, doubting, and finally proving the goal.

Research in Hierarchical Semantic Segmentation. After completing the grad course *deep learning in computer vision* with A+, I became familiar with the underlying structure of state-of-the-art (SOTA) vision models. I observed that for existing segmentation models, the labels only get assigned at the last layer, indicating their inability to derive hierarchical semantics. My assumption was verified after evaluating the SOTA models on several datasets targeting partial-whole relationships. I was advised by Prof. Stella Yu to work on a project on matching the hierarchy of image segmentation and language entity at each level of granularity to create a visual model which conducts hierarchical image-level recognition and semantic segmentation.

In order to grasp the partial-whole relationship in language, I extract 50k images from the conceptual caption 12M dataset that contains hierarchical information in their captions using the CoreNLP dependency parser. And to match the language entity with the hierarchy of image segmentation at each level, I introduced an extra contrastive loss and fine-tuned the existing state-of-the-art GroupViT model based on the improved loss. After evaluating the DensePose dataset, which targets human part segmentation only, *the mean intersection over union (MIoU) was lifted by an average of 4% across all labels.* This project gives me insight into how language can play a role in boosting visual intelligence and motivates me to keep believing and exploiting my ideas' potential.

Research in Language-Aided Object Detection. Following the mindset of the previous two projects, I would like to know whether common sense in pre-trained LLM could be leveraged to improve state-of-the-art object detectors in a zero-shot manner. I co-led this project and was supervised by Prof. Joyce Chai to introduce a method for language models to infer possible refinement of labels assigned by object detectors, given the spatial relation between bounding boxes as well as the description of the scene.

After deliberately designing the algorithm and launching the pipeline confidently, the performance on the PASCAL VOC and COCO datasets did not improve. To understand why our initial approach did not work, I optimized the pipeline by ablating different modules, trying with prompts in various formats, and enabling more hyperparameter searches. With this analysis, I found the root cause to be the falsely ignored instances per class when reading tensor outputs from the detector. After refinement, we see an increase of 2% in mean Average Precision (mAP) for SOTA object detectors. *This work is under submission to ACL 2023.* This experience taught me that the fun lies in developing good hypotheses and well-conceived experiments to validate them.

Future Work. The speed I advanced myself in graduate-type works convinces me of the benefit of pursuing a Ph.D. degree. We are on the verge of a breakthrough change in embodied AI research, and I want to extend my previous experiences and advance them to a higher level at UW. My passion for building intelligent embodied AI agents to perform real-world multimodal tasks has driven me to this day. In the future, I want to continue exploring the intersection between robotics, language, and vision in order to investigate how language and visual intelligence could intertwine to boost embodied agents' performance. I would want to work with Prof. Yejin Choi, who did fantastic work in physical commonsense reasoning and language grounding with vision. I am also interested in working with Prof. Noah Smith, whose work emphasizes language prompting and understanding of deep models. If I am privileged to be accepted, I will keep pushing on the future exploration of embodied AI with research depth and a keen awareness of my social responsibilities.