ViAD : A Novel Strategic Robotic Navigation Methodology

For decades, visualization and comprehension of the real world environment has been a bottleneck in building robust navigation robotics systems. It has been quite popular to utilize LiDAR/ultrasonic cameras in order to understand 3-D space. This method of visualization leverages the obstacle coordinates in order to compute its distance from the obstacle, however it acts as a limitation to the real-time contextual scene understanding. Clutter, new lighting or any real-time changes in the environment usually confuses the system. When it came to planning, it was quite popular to use Nav2 which followed a very rigid planning pattern and a rule based approach. For example: “if an obstacle is detected, turn right, else go straight”. Its inability to adapt to dynamic tasks like “avoid moving objects” made it inflexible. Along with this, it lacked natural language control. In recent times, LLMs have been utilized to overcome this issue. When it came to low-level execution like directing wheel speed and frequency, it lacked memory based learning nor did it have any way to leverage offline datasets for policy improvement. Commands could be only understood by either feeding coordinates or numeric terms and had no semantic grounding for user commands. In order to overcome these issues, we propose ViAD - Visualize, Analyse and Decide methodology. It consists of leveraging Vision Transformers to get the raw visual perception of the environment, an LLM for comprehension and strategic planning, and Decision Transformers to implement low level execution leveraging memory-based learning. Each component is crucial to make our robot more adaptable, understandable and strategically navigable.

When it came to limited scope of visualization, we replaced ultrasonic sensors with a Depth Camera which helped us implement Vision Transformers for semantic segmentation. Given the large-scale adaptation of the transformer architecture in the industry, Vision Transformers have proved to be very efficient as they take less training time and have better raw contextual understanding due to the attention mechanism, replacing both Ultrasonic inputs or old school CNNs. Further, the embeddings produced by vision transformers are given as input to the LLM. In research, LLMs have been utilized in ways to interpret human instructions as its input and turn them into actionable sub goals in order for it to be comprehensible for robots. The sub-goals are usually a set of pre-defined high-level instructions including choices like “go_forward” or “turn_right”. These abstract commands along with the ViT learned embeddings are fed into LLM as inputs, helping in adaptation to changing context in the environment. The LLM acts as a strategist which combines perception and instructions in order to make the right high-level decision. It does not directly handle motor speed, wheel or any low-level execution.

Coming to low-level execution, reinforcement learning has been traditionally used in robotics which relies on trial-error based online learning which poses a risk when implemented on hardware due to its sample inefficiency. Another method that has been used widely is behaviour cloning. It imitates expert behaviour, however it poses risks in case the robot drifts off the expert path. In order to overcome these limitations, we leveraged decision transformers which utilizes the strength of reinforcement learning and behaviour cloning. It learns from offline datasets like trajectories by an expert and uses goal-oriented behaviour like reinforcement learning. It utilizes the memory of past trajectories. With all these factors combined, it predicts safe and smooth motions by recalling the past and still aiming for future success. Here, Vision Transformers have been used for visualizing, LLMs for planning and Decision Transformers for muscle memory and execution.