CS539 - Embodied AI
Paper Review - Mapping Navigation Instructions to Continuous Control Actions with Position-Visitation Prediction

The paper introduces a new neural architecture for the task of mapping language instructions, image observations to continuous control for a quadcopter drone. It does this by decomposing the task into two stages, planning and execution. Planning phase is trained in a supervised learning manner, given instruction, observation, pose of the drone use feature extractor network, semantic segmentation network (LingUNet) to predict path (as visitation distribution map) and goal location (as goal distribution map). The 2nd half of execution, is trained via imitation learning to use these maps as input and produce suitable action to lead the drone to the desired goal location. Auxiliary losses have been added to ensure the network develops an understanding of objects, co relates it to the instruction sentence and is able to identify objects given the sentence.

Main contributions of the paper are its development of a language learnable model which can understand object references, co references and spatial, temporal relations and transfer that to control. The paper produces improvements of 16.85% on absolute task completion accuracy over current state of the art instruction following methods.
Another contribution is that of interpretable visualization of agent plans which supported by their experiments do show the ability of the model to plan, predict goal location and understand language semantics like on the left/right/around an object.

Some strength of the paper are that it is well written, the experiments and ablation studies do support their approach. The idea of comparison to baselines with weaker language understanding doing poor on the tasks was well supported, auxiliary losses were justified and made sense. Additionally some robustness/ generalization studies have also been done to prove that the model is indeed able to understand, plan and predict the goal locations well. Shows generalizability of their approach.

Some weakness were maybe, the paths taken were too simple, fairly straight, one good experiment could also be to have a convoluted path around different object types which could further prove better understanding. But the model does show promise. Some bits about how the model learns and predicts the global environment map is unclear. How would this method scale to larger maps.

The most interesting points would be their interpretable planning visualizations and their ability to show language understanding and mapping it to control.

For future work it would be interesting to see how providing it with more language instruction data and in a complex environment would scale things.