CS539 - Embodied AI
Paper Review - Reinforced Cross-Modal Matching and Self-Supervised Imitation Learning for Vision-Language Navigation

The paper introduces a novel approach, Reinforced Cross-Modal Matching (RCM) for the task of vision language navigation (VLN) in indoor environments. Evaluation on benchmark datasets show that RCM model significantly outperforms pervious state of the art methods (a 10% increase in SPL) on the task of VLN. The paper also proposes a Self-Supervised Imitation Learning (SIL) method to improve generalizability of the learned policy to unseen environments, SIL has been shown to tremendously reduce the success rate performance gap between seen and unseen environments when compared to previous state of the art methods (from 30.7% to 11.7%).

Main contributions of the paper are the RCM framework, supported by extrinsic rewards, grounding textual instruction to local scene observation, learning trajectory history, focus of textual instruction to corelate instruction to observation better i.e Cross-Modal Matching. Matching Critic has been used to enforce overall agent trajectory matches human expert trajectories and a cycle reconstruction component has been added to interpret instruction from path and path from instruction to better learn their correlation, this acts as the intrinsic reward. SIL utilizes the trained Matching Critic in the unseen environment to pick the best trajectory from the model from its multiple predictions and further explore and improve its trajectory predictions in unseen environments.

Some strengths of the paper are that it is well written, the experiments and ablation studies do support their approach. Comparison to baselines with weaker language understanding models do show performance gains of 10% on SPL. Their claim on generalizability is an interesting one with high success rate gain from previous methods and less drop in performance from seen to unseen environments supports novelty in SIL idea.

Couldn't find many weaknesses in the paper.

The most interesting points would be potential for formulating more intrinsic rewards for their Self Supervised Imitation Learning (SIL) model which would enable directions for life long learning for the agent without supervision.