Reinforced Cross-Modal Matching and Self-Supervised and Imitiation Learning for Vision-Language Navigation: Paper Review

The authors leverages multi-reward reinforcement learning with self-supervised imitation learning to address 3 core challenges in Visual-Language Navigation (VLN) tasks : cross-modal grounding, sparse rewards, and generalization to unseen environments. Firstly, the multi-reward learning framework called the "Reinforced Cross-Modal Matching (RCM)" consists of Cross Modal reasoning navigator, which learns the trajectory history, focus on textual instructions and local visual attention for matching the instructions and trajectories, while also evaluating if the actions taken for the path matches the previous instructions. RCM is trained with intrinsic rewards from matching critic, which measures the alignment between the language instructions and the navigator's trajectory; as well from extrinsic sparse rewards from the environment. Secondly, another framework called "Self Supervised Imitation learning (SIL)" is used mainly for the exploration of unseen environments by imitating past successful decisions, to make it generalizable to unseen environment, with no labelled data or ground truth.

The approach has been tested on R2R dataset for the VLN tasks where the task is to train the agent in environments and test on seen and unseen validation tests. The methods were evaluated with 5 evaluation metrics (PL, NE, OSR, SR, SPL). The results demonstrate that RCM significantly outperforms the SOTA Methods, across all metrics. Moreover, using SIL to imitate the best behaviors of the RCM agent on training set, achieves the best result on SPL (38%). Further, SIL significantly improves RCM on unseen environment with no information of target location. The paper also discusses the importance of each of the components of the architecture (intrinsic reward, extrinsic reward, cross modelling navigator) with ablation study.

The strengths of the paper lies in its approach and experimental study to validate the proposed approach. The problem has been well motivated by how humans visually navigate to follow language instructions, based on extrinsic feedback from the environment, as well as intrinsic motivations in a self supervised way in unseen environment, based on previous best decisions made on different environments.The cross-model reasoning model combining both the language and visual modalities to similarly learn the local visual attention along with text and trajectory history is quite impressive. Further, exhaustic experiments have been done to compare with the SOTA methods and with unseen environments, to demonstrate the effectiveness of the approach. The ablation study done to demonstrate the importance of intrinsic reward, extrinsic reward and cross modelling navigator of the architecture is one of the key strengths of this paper.

However, there are few weakness in the paper as following. Firstly, in the cross-model reasoning navigator architecture design, it is not quite clear as to why the action predictor is fed with history information again, when that information is already encoded in the c_{text}. The intuition behind this architecture design could have been explained in more detail. Further, in the self supervised learning (SIL), it is not clear how the navigator produces initial several trajectories, among which the best trajectory is chosen as a ground truth. Also, I feel that this way of choosing the ground truth might make the learning biased towards this ground truth trajectory which makes it follow a moving target. Further, it is not explained why the Navigation error (NE) and Oracle Success rate (OSR) is lower in RCM as compared to RCM + SIL (train). Further, the performance (plots) with respect to the training time with mean and variance has not been demonstrated. There must be comparison for computation time and memory requirements with SOTA methods.

Overall, the idea of using multi-reward learning with self-supervised learning by leveraging both language and vision modalities to solve the VLN task is an interesting approach. The approach clearly stands out of all the existing SOTA methods and has set a good benchmark for similar future works for VLN tasks. I would be curious to apply the same approach to more complex environments/ simulators in order to analyse where and why it fails.