Review6: Mapping Navigation Instructions to Continuous Control Actions with Position-Visitation Prediction

Vivswan Shitole

CS539 Embodied AI

1 Summary

This paper introduces a new model for instruction-guided navigation with a continuous action space. Its a two-stage model where the first stage takes in provided instruction, the observation of the agent and the pose of the agent as input and outputs two position-visitation distributions predicting the positions that are likely to be visited by the agent during a correct instruction execution and the positions where the goal is likely to be present respectively. The second stage of the model controls the agent (a drone) to fly between the high probability positions (predicted in stage-1) to complete the task and reach the most likely goal location. The two stages are trained separately. The visitation prediction stage is trained with supervised learning using expert demonstrations with auxiliary losses. The plan execution stage is trained by mapping expert visitation distributions to actions using imitation learning. At test time the second stage uses predicted distributions from the first stage.

More concretely, in the first stage, the model takes the instruction u, the image observed by the agent I_t and the agent's pose P_t as input. Image I_t is passed through ResNet to extract features F_t^c that are projected onto the ground plane at elevation zero using a pinhole camera model to obtain a feature map F_t^w in the world co-ordinate frame. All the features $F_{\leq t}^w$ are integrated over time to obtain a semantic feature map S_t^w which maintains a learned high-level representation of every world location (x,y) that has been visible in any of the previous observed images. The instruction is encoded as an embedded vector u using an LSTM. A kernel K_g is computed using a learned linear transformation over the instruction embedding: $K_g = W_g u + b_g$. A grounding map R_t^w is computed using 1×1 convolution of semantic map S_t^w and kernel K_g. The concatenated matrix $[S_t^w R_t^w]$ is fed as input to LINGUNET (a language conditioned image-to-image encoder-decoder) to yield the two visitation distributions d_t^p and d_t^g. In the second stage, the obtained distributions d_t^p and d_t^g undergo an affine transformation to get aligned with agent's current egocentric frame as defined by its pose P_t and are center-cropped to two $K \times K$ regions. The cropped regions are flattened and concatenated to a vector of size $2K^2$ and passed through a multi-layer perceptron with leaky-ReLU activation function to yield the linear amd sngular velocities (v_t, w_t) or $STOP$ as the output action.

Experiments are conducted using the quad-copter simulation environment based on the Unreal Engine. Navigation instructions are obtained from the LANI corpus. Stopping distance of the agent from the goal is used as the evaluation metric. The experiments compare the introduced approach (PVN) to CHAPLOT and GSMN approaches. PVN demonstrates absolute task completion improvement of 16.85% over the second best GSMN system. Some ablation experiments are conducted as well. Removing the auxiliary losses or doing away with the goal-distribution prediction or the removing access to the instruction embedding lowered the performance significantly. Using perfect goal-visitation distribution or full-observability to the agent improved the performance drastically.

2 Strengths

The proposed model provides a significant improvement of 16.85% in performance over the previous best model for instruction-guided visual navigation in continuous action space. Visualizations of the position-visitation distributions can act as explanations for the agent's plan. Separately training the two stages of the model allows for sample efficient imitation learning. The figures provided in the paper (particularly figure 3) do a great job of summarizing the paper.

3 Weaknesses

Its not clear why and how are the position-visitation distributions discretized over positions in the otherwise continuous environment. The distributions are continuously updated with every new observation of the environment. Such a high planning frequency may be redundant and computationally expensive. The authors claim that the introduced approach can generalize to unseen tasks during test time, with the same visitation distributions working for a humanoid or a ground vehicle. This claim is never validated.

4 Reflections

The paper seems to introduce a yet another model for instruction-guided visual navigation in continuous action space. The proposed model is an innovation in the training pipeline but does not make a fundamental contribution to the concerned domain.

5 Most Interesting Thought

The most interesting idea from the paper for me are the position-visitation distributions which can be visualized and interpreted as the agent's plan.