Learning to Map Natural Language Instructions to Physical Quadcopter Control using Simulated Flight

Vals Blukis, Yannick Terme, Eyvind Niklasson, Ross A. Knepper, Yoav Artzi
https://github.com/clic-lab/drif

**Task:** Follow natural language navigation instructions on a physical quadcopter, assuming access only to first-person RGB images and pose estimates. **Challenges:** Language understanding, grounding, perception, spatial reasoning, exploration and control.

### Two-Stage Model (Position Visitation Network v2)

Stage 1 outputs a pair of 2D probability distributions over environment locations. Stage 2 generates actions to:
1. Visit high-probability positions
2. Stop at a likely goal location
3. Explore the environment

#### Stage 1: Visitation Distribution Prediction
- **Stop position distribution** predicts where the goal is, or if the goal is not observed yet.
- **Position visitation distribution** predicts which observed positions in the environment the agent should visit, or if it should visit positions that are not yet observed.

#### Stage 2: Action Generation
- **Deterministic Observability Mask**
- **Mask Generator**
- **Control Network**

### Joint Sim-to-Real Training with SuReAL

#### Adversarial Loss for Joint Sim and Real Training
- **Discriminator**
- **Wasserstein Loss**
- **Additional Components for Domain-Adversarial Training**

#### SuReAL – Supervised and Reinforcement Asynchronous Learning
- **Process A: Supervised Learning**
- **Process B: Reinforcement Learning**
- **SuReAL** simultaneously train Stage 1 with supervised learning and Stage 2 with reinforcement learning. Continuously add on-policy rollouts from RL to the data used to train Stage 1.

#### Advantages of SuReAL:
- More sample efficient than end-to-end reinforcement learning
- More robust to errors than end-to-end behavior cloning (supervised learning)
- Unlike behavior cloning, does not require action oracle, but requires only a visitation distribution oracle

### Intrinsic Reward for Language-Directed Exploration with Partial Observability
- **Exploration Reward**
- **Trajectory Reward**
- **Stopping Reward**
- **Step Reward**

#### Automated Evaluation
- **Goal Success Rate**
- **Trajectory Earth-Mover’s Distance**

#### Human Evaluation (M Thur 5-point Likert-scale scores of agent behavior)
- **Goal score:** How well the agent reached the correct goal. 5/5 points 40% of the time.
- **Path score:** How well the agent followed the correct path. 5/5 points 38% of the time.

---

**After the blue bale take a right towards the small white bush, before the white bush take a right and head towards the right side of the banana.**

---

**绡**

---

**绡**

---

**绡**