Counterfactual Vision-and-Language Navigation via Adversarial Path Sampling

UC Santa Barbara

Tsu-Jui Fu
Xin Wang
Matthew Peterson
Scott Grafton
Miguel Eckstein
William Wang

UC Santa Barbara
Vision-and-Language Navigation (VLN)

- Achieve the goal based on the instruction in a room
  - learns to align the **linguistic semantic** and **visual understanding**

- Difficult to collect (instruction, path) pairs
  - the **data scarcity** makes learning the optimal match challenging
Data Augmentation with Speaker

- Expand the training set
  - a **speaker** to back-translate path into instruction
  - randomly sample paths as augmented data
  - however, the **help is limited** since the augmented path are arbitrary

Adversarial Path Sampling (APS)

- To make the sampled path more useful
  - APS learns to sample **challenging paths** that NAV cannot navigate easily
  - NAV tries to solve the paths from APS
Adversarial Path Sampling (APS)

- To make the sampled path more useful
  - APS learns to sample **challenging paths** that NAV cannot navigate easily
  - NAV tries to solve the paths from APS

**Adversarial Path Sampler (APS)**

**Adversarial Training**
Adversarial Path Sampling (APS)

- To make the sampled path more useful
  - APS learns to sample **challenging paths** that NAV cannot navigate easily
  - NAV tries to solve the paths from APS

![Adversarial Path Sampler (APS)](image)

![Adversarial Training](image)
Pre-Exploration with APS

- Under **unseen environments**, we can do **pre-exploration** to make NAV more robust
  - use APS to sample paths and optimize NAV for unseen adaption
  - then, NAV can run each instruction in a **single turn**
● Randomly sampled **stop improving** when using more than 60%
● **APS sampled** helps both seen and unseen environments
● **Pre-Exploration** further helps unseen environments
Walk out of the bathroom and straight across the hall. Walk down the steps and stop.

(a) without Pre-Exploration  
(b) with Pre-Exploration