# VLN

Literature Review week5

Towards Learning a Generic Agent for Vision-and-Language Navigation via Pre-training

# CONTENTS

# Solved Questions

The first to propose a pre-training and fine-tuning algorithm for VLN tasks

Trained in a self-supervised learning manner on a large number of image-text-action triples. The pre-training model provides a universal expression of visual environment and language instructions.

Be more effective in new tasks and expand to previously unseen environments.

# Direction

the first pre-trained models, grounding vision-language understanding with actions in a reinforcement learning setting.

The pre-trained model plays the role of providing generic image-text representations, and is applicable to most existing approaches to VLN, This paper isolate the encoder stage, and focus on pre-training a generic vision-language encoder for various navigation tasks.

Method used

Background: Partially Observable Markov Decision Process,POMDP

A vision-language encoder,
An action decoder

Pre-training Models

Pre-training Objectives

# Encoder

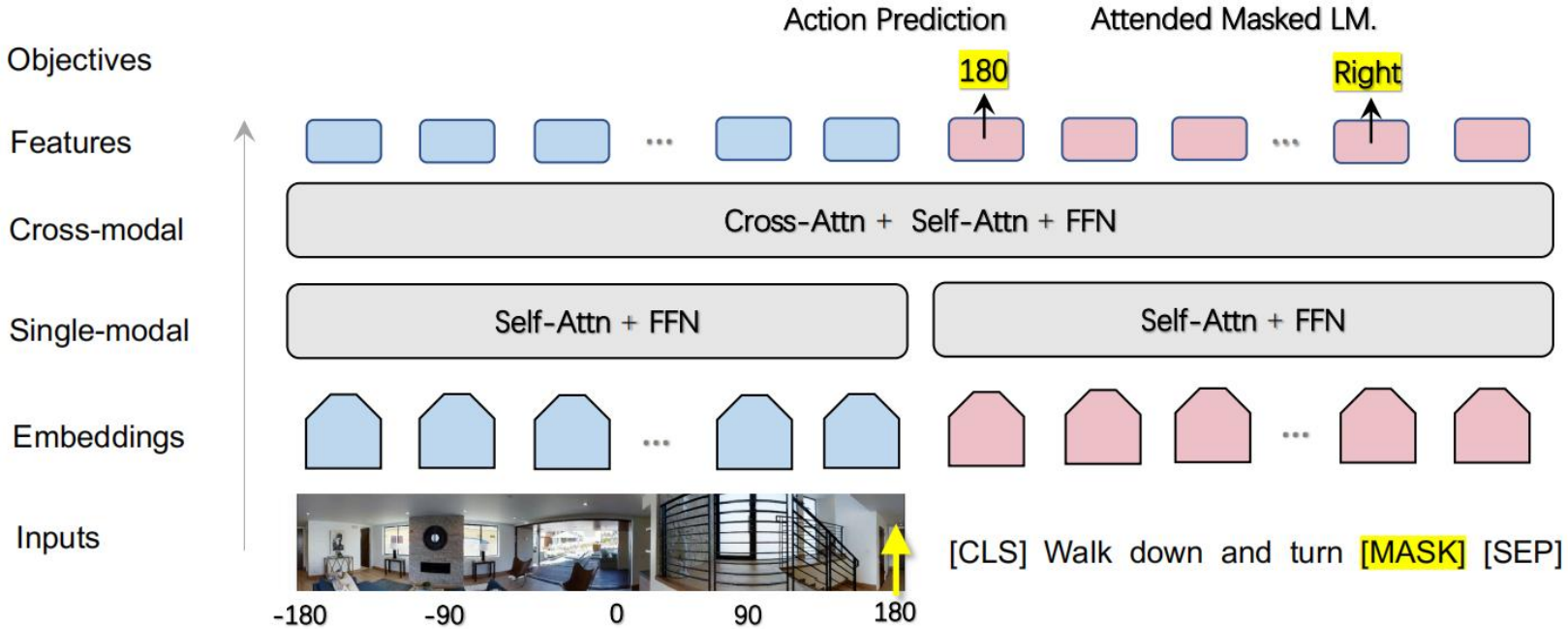
## Two Single-modal Encoders:

All of the keys, values and queries come from the output of the previous layer in the encoder. Each position in the encoder can attend to all positions that belong to its own modality in the previous layer.

# Encoder

## One Cross-modal Encoder:

This cross attention layer is followed by a self-attention layer and an FFN layer


Objectives
Action Prediction
Attended Masked LM.
180
Right
Features
Cross-modal
Cross-Attn + Self-Attn + FFN
Single-modal
Self-Attn + FFN
Self-Attn + FFN
Embeddings
Inputs
[CLS] Walk down and turn [MASK] [SEP]
-180
-90
0
90
180

# Pre-training Objectives

## Image-attended masked language modeling (MLM) :

Randomly mask out the input words with probability 15%, and replace the masked ones xi with special token [MASK].Masked word is recovered from surrounding words, but with additional image information to attend. It helps the learned word embeddings to be grounded in the context of visual states.

# Pre-training Objectives

## Action Prediction

The output on the special token [CLS] indicates the fused representation of both modalities. They apply an FC layer on top of the encoder output of [CLS] to predict the action. It scores how well the agent can make the correct decision conditioned on the current visual image and the instruction, without referring to the trajectory history

# Pre-training Dataset

construct the pre-training dataset based on the Matterport3D Simulator, a photo-realistic visual reinforcement learning (RL) simulation environment for the development of intelligent agents based on the Matterport3D dataset

1) The training datasets of R2R, which has 104K image-text-action triplets;

2) Also employed the Speaker model to synthesize 1,020K instructions for the shortest-path trajectories on the training environments.

# Evaluation

## TL Trajectory Length

measures the average length of the navigation trajectory.

## SR Success Rate

is the percentage of the agent’s final location that is less than 3 meters away from the target location.

## NE Navigation Error

The mean of the shortest path distance in meters between the agent’s final location and the target location.

## SPL Success weighted by

A higher score represents more efficiency in navigation.

# Data analysis

| | Agent | Validation Seen TL ↓ | NE ↓ | SR ↑ | SPL ↑ | Validation Unseen TL ↓ | NE ↓ | SR ↑ | SPL ↑ | Test Unseen TL ↓ | NE ↓ | SR ↑ | SPL ↑ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Greedy, S | RANDOM | 9.58 | 9.45 | 16 | - | 9.77 | 9.23 | 16 | - | 9.93 | 9.77 | 13 | 12 |
| | SEQ2SEQ | 11.33 | 6.01 | 39 | - | 8.39 | 7.81 | 22 | - | 8.13 | 7.85 | 20 | 18 |
| | RPA | - | 5.56 | 43 | - | - | 7.65 | 25 | - | 9.15 | 7.53 | 25 | 23 |
| | SPEAKER-FOLLOWER | - | 3.36 | 66 | - | - | 6.62 | 35 | - | 14.82 | 6.62 | 35 | 28 |
| | SMNA | - | - | - | - | - | - | - | - | 18.04 | 5.67 | 48 | 35 |
| | RCM+SIL(TRAIN) | 10.65 | 3.53 | 67 | - | 11.46 | 6.09 | 43 | - | 11.97 | 6.12 | 43 | 38 |
| | REGRETFUL | - | 3.23 | 69 | 63 | - | 5.32 | 50 | 41 | 13.69 | 5.69 | 48 | 40 |
| | FAST | - | - | - | - | 21.17 | 4.97 | 56 | 43 | 22.08 | 5.14 | 54 | 41 |
| | ENVDROP | 11.00 | 3.99 | 62 | 59 | 10.70 | 5.22 | 52 | 48 | 11.66 | 5.23 | 51 | 47 |
| | PRESS | 10.57 | 4.39 | 58 | 55 | 10.36 | 5.28 | 49 | 45 | 10.77 | 5.49 | 49 | 45 |
| | PREVALENT (ours) | 10.32 | 3.67 | 69 | **65** | 10.19 | **4.71** | **58** | **53** | 10.51 | 5.30 | **54** | **51** |
| M | PRESS | 10.35 | 3.09 | 71 | 67 | 10.06 | 4.31 | 59 | 55 | 10.52 | 4.53 | 57 | 53 |
| | PREVALENT | **10.31** | 3.31 | 67 | 63 | **9.98** | **4.12** | **60** | **57** | **10.21** | **4.52** | **59** | **56** |
| | Human | - | - | - | - | - | - | - | - | 11.85 | 1.61 | 86 | 76 |

Comparison with the state-of-the-art methods on R2R. Blue indicates the best value in a given setting. S indicates the single-instruction setting, M indicates the multiple-instruction setting.

| Agent | Validation Unseen | | | Test Unseen | | |
|---|---|---|---|---|---|---|
| | **Oracle** | **Navigator** | **Mixed** | **Oracle** | **Navigator** | **Mixed** |
| RANDOM | 1.09 | 1.09 | 1.09 | 0.83 | 0.83 | 0.83 |
| SEQ2SEQ | 1.23 | 1.98 | 2.10 | 1.25 | 2.11 | 2.35 |
| PREVALENT (Ours) | **2.58** | **2.99** | **3.15** | **1.67** | **2.39** | **2.44** |
| SHORTEST PATH AGENT | 8.36 | 7.99 | 9.58 | 8.06 | 8.48 | 9.76 |

Table 2: Results on CVDN measured by Goal Progress. Blue indicates the best value in a given setting.

| | Agent | SEEN-ENV | | | | UNSEEN-ALL | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | SR ↑ | SPL ↑ | NE ↓ | #R ↓ | SR ↑ | SPL ↑ | NE ↓ | #R ↓ |
| Rule | RANDOM WALK | 0.54 | 0.33 | 15.38 | 0.0 | 0.46 | 0.23 | 15.34 | 0.0 |
| Rule | FORWARD 10 | 5.98 | 4.19 | 14.61 | 0.0 | 6.36 | 4.78 | 13.81 | 0.0 |
| | NO ASSISTANCE | 17.21 | 13.76 | 11.48 | 0.0 | 8.10 | 4.23 | 13.22 | 0.0 |
| | ANNA | 88.37 | 63.92 | 1.33 | 2.9 | 47.45 | 25.50 | 7.67 | 5.8 |
| | PREVALENT (Ours) | 83.82 | 59.38 | 1.47 | 3.4 | **52.91** | **28.72** | **5.29** | 6.6 |
| Skyline | SHORTEST | 100.00 | 100.00 | 0.00 | 0.0 | 100.00 | 100.00 | 0.00 | 0.0 |
| Skyline | Perfect assistance | 90.99 | 68.87 | 0.91 | 2.5 | 83.56 | 56.88 | 1.83 | 3.2 |

Table 3: Results on test splits of HANNA. The agent with "perfect assistance" uses the teacher navigation policy to make decisions when executing a subtask from the assistant. Blue indicates the best value.