Literature Review

Author Name

November 2021

1 Some Terminology

Partially Observable Markov Decision Process (POMDP): S, A, P_s, r.

where S is the visual state space, A is a discrete action space, P_s is the unknown environment distribution from which we draw the next state, and r R is the reward function.

RGB image s_t and then takes an action a_t. The agent interacts with the environment sequentially and generates a trajectory of length T. The navigation is successfully completed if the trajectory terminates at the intended target location.

Training data set DE = , x consists of pairs of the instruction x together with its corresponding expert trajectory .

The agent then learns to navigate by performing maximum likelihood estimation (MLE) of the policy , based on the individual sequences.

2 Solved Questions

-

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

(2) 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.

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

3 Direction

While a number of methods have been proposed to improve language understanding, common to all existing work is that the agent learns to understand each instruction from scratch or in isolation, without collectively leveraging prior vision-grounded domain knowledge.

Although various approaches have been suggested to enhance language comprehension, a shared characteristic among all prior research is that the agent acquires an understanding of each instruction individually, without effectively utilizing pre-existing vision-grounded domain knowledge as a collective resource.

In order to better tackle the inherent ambiguity in instructions, we suggest a novel approach: pre-training an encoder to align language instructions and visual states, creating shared representations. At each time step, the model independently processes image-text-action triplets, training to predict masked word tokens and subsequent actions. This establishes a VLN pre-training framework within a self-learning paradigm. By removing language understandings lacking consensus from visual states, we can simplify the complexity of VLN learning.

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

4 Dataset

The task constructs 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 in [9] to synthesize 1,020K instructions for the shortest-path trajectories on the training environments.

5 Encoder

The backbone network has three principal modules: two single-modal encoders (one for each modality), followed by a cross-modal encoder.

5.1 Single-modal Encoder:

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.

5.2 Cross-modal Encoder

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

6 Pre-training Objectives

two main tasks to pre-train our model: 1. Image-attended masked language modeling (MLM) 2. action prediction (AP).

6.1 Attended Masked Language Modeling

We randomly mask out the input words with a probability of 15% and replace the masked ones xi with a 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.

6.2 Action Prediction

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.

7 Adapt new environment

7.1 R2R

In R2R, the goal is to navigate from a starting position to a target position with a minimal trajectory length, where the target is explicitly informed via language instruction. To use the pre-trained model for fine-tuning in R2R, the attended contextualized word embeddings are fed into an LSTM encoder-decoder framework, as in [9, 16]. In prior work, random initialization is used in [9], and BERT is used in [16]. In contrast, our word embeddings are pre-trained from scratch with VLN data and tasks.