CS 461 - ARTIFICIAL INTELLIGENCE

PROJECT FINAL PROPOSAL

Project Name: SeqVista

Group No: 03

Instructor: Özgür Salih Oğuz

TA: Arda Sarp Yenicesu

Group Members:

Zeynep Hanife Akgül - 22003356

Arshia Bakhshayesh - 22001468

Huzaifa Huzaifa - 22301342

Emre Karataş - 22001641

Faaiz Khan - 22001476

Project Name

SeqVista: Sequence to Sequence Model for Vision-Language Navigation

1. Project

1.1 Description

The project aims to develop a Sequence-Sequence (Seq) model for training an autonomous vision-language navigation (VLN) agent. VLN refers to the navigation of embodied agent in an unseen environment based on natural language instructions. The agent is required to understand the natural language instructions using the natural language processing and then use computer vision to identify potential landmarks mentioned in the instructions in order to initiate its navigation actions. As an example, the agent might be required to execute the instruction, "*Walk down the stairs to the bottom of the staircase. Continue down the next small flight of stairs towards the bathroom at the lower level"*.

1.1.1 Simulator:

For the implementation of the VLN agent, the Matterport 3D simulator would be used. The Matterport 3D simulator is a graph-based environment based on the Matterport 3D dataset [1], which consists of 90 building-scale environments. The Matterport 3D simulator would be used in conjunction with the Room-to-Room (R2R) dataset [2]. This dataset consists of a large number of diverse instruction-trajectory pairs associated with Matterport 3D environments. The action space in this simulator would consist of the following actions:

- Forward: This action moves the agent to the navigable node, which is closest to the center of the agent's field of view (FOV).
- Right, Left: These actions would change the camera heading by 30 degrees.
- Up, Down: These actions would change the camera elevation by 30 degrees.
- Stop: This special action is initiated for the termination of the navigation episode.

At each time step, the output of the Matterport 3D simulator consists of:

- First-person RGB observation of the environment.
- The set of next navigable nodes from each node.

1.1.2 Seq-to-Seq Model:

The VLN problem can be solved using a Sequence-Sequence model where the input is a sequence of encoded natural language instructions while the output is the sequence of navigation actions, as shown in Fig 1. The instructions are encoded using the encoder LSTM to compute the encoder context, and the output is obtained by applying the decoder LSTM over the context vector. The image features would be computed using RESNET-152, pre-trained on ImageNet architecture, which utilizes convolutional neural networks for the extraction of image features. The decoder LSTM would be using these image features at the output stage for giving out the navigation actions.

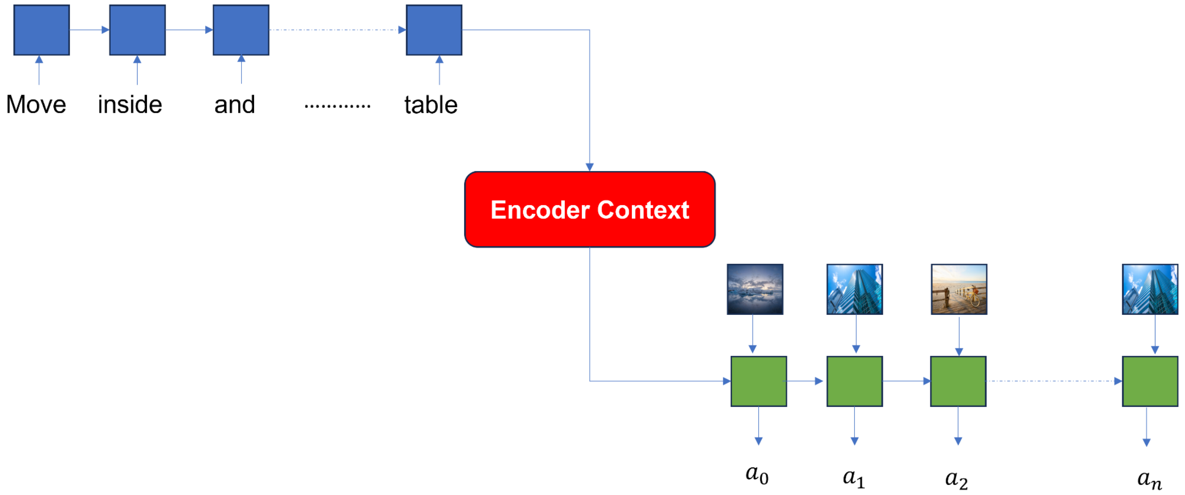

Figure 1: Seq-Seq Model for Vision-Language Navigation

1.1.3 Evaluation Metrics:

The evaluation metrics that would be used for the evaluation of our VLN agent are:

- ***Navigation Error (NE):*** It measures the distance between the actual goal location and the stopping location of the VLN agent.
- ***Success Rate (SR):*** The percentage of successful navigation episodes out of the total episodes that were carried out by the VLN agent is called the success rate. An episode is considered successful if the navigation error is *less than 3 meters.*
- ***Path Length (SR):*** The path length (PL) provides the total length of the path followed during the execution of navigational instruction.
- ***Success Weighted by Path Length (SPL):*** The success weighted by path length combines the success rate and path length. It penalizes the model if the total length of

the path followed by the agent is greater than the ground truth path demonstrated in natural language instructions.

1.2 Goals

The goals of this project are to:

- Successfully initialize the Matterport3D Simulator environment, which is a large-scale reinforcement learning environment based on real imagery,
- Successfully model the agent with a Recurrent Neural Network (RNN) policy using a Long Short-Term Memory (LSTM) based Sequence-to-Sequence (Seq2Seq) architecture for mapping natural language instructions to navigation actions,
- Establish proficiency with Matterport API and action space,
- Utilize the Matterport3D simulator to use the Room-to-Room (R2R) dataset for visually grounded natural language navigation in real buildings,
- Successfully develop Sequence to Sequence (Seq2Seq) model,
- Improve the agent's understanding of nuanced instructions, enabling it to handle a broader range of coarse language inputs,
- Evaluate the trained agent on VLN evaluation metrics mentioned in 1.1.3.

2. Literature

The paper addresses the challenge of Vision-and-Language Navigation (VLN) in robotics, aiming to enable robots to understand and execute natural language instructions in real-world environments. The authors introduced the Matterport3D Simulator, utilizing the Matterport3D panoramic RGB-D dataset, and created the Room-to-Room (R2R) dataset, consisting of open-vocabulary navigation instructions. They employed a sequence-to-sequence neural network model with LSTM units and an attention mechanism. The model processed sequential natural language instructions and images. During training, two approaches were explored: 'teacher-forcing,' where ground-truth actions were provided for training, and 'student-forcing,' where actions were sampled from the model's output probabilities. The model's action space included six actions for navigation. Image features were extracted using a pre-trained ResNet-152 CNN. An attention mechanism identified relevant parts of instructions for accurate action prediction. The study's implementation involved text pre-processing, specific LSTM configurations, embedding sizes, dropout techniques, and optimization methods. The results demonstrated the model's ability to effectively navigate unseen environments based on natural language instructions and real-world visual data, marking a significant advancement in VLN research [2].

3. Timeline

1. **Progress Report Preparation:** Understanding the use of Matterport API for this project, creation of Seq-Seq Model
2. **Final Report and Demo:** Trained & Evaluated VLN agent on R2R dataset using Seq-Seq Model

4. Work Distribution

- Zeynep Hanife Akgül: Preparation of Seq-Seq model
- Arshia Bakhshayesh: Understanding the Use of Matterport API
- Huzaifa Huzaifa: Preparation of Seq-Seq model
- Emre Karataş: Preparation of Seq-Seq model
- Faaiz Khan: Evaluation Setup for VLN agent

5. References

[1] Matterport3D: Learning from RGB-D data in indoor environments, https://niessner.github.io/Matterport/ (accessed Oct. 23, 2023).

[2] Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., & Hengel, A. V. (2017). Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments. ArXiv. /abs/1711.07280