Zero Shot Goal Based 2-D Object Navigation using CLIP

Tyler Gorman and Gowri Shankar Raju Kurapati

University of Colorado Boulder

Abstract. In today's space of deep learning, the number of parameters, input and the length of time that is required to train models are growing rapidly. Likewise, to establish top of the line results with reinforcement learning, we must run tens of thousands to millions of simulations. Although impressive results are being gained from these, at some point the models will need to be fine tuned to be more accurate using less data. At the same time, the true and inherent capabilities of a model are not being explored as much when compared to creating new models and fine tuning them. Lately, zero-shot models have exhibited outstanding performance in image classification of arbitrary objects (i.e., categorizing photos at inference using categories not explicitly encountered during training). We apply the success of the CLIP model to the embodied AI challenge of 2-D object navigation in this study. Though CLIP has been used as a part of the architecture in various 3-D Object Navigation Tasks (on Habitat and RoboTHOR) and it was able to achieve SOTA results in various downstream tasks like Object Goal Navigation, Room Rearrangement and Point Goal Navigation, no work has been done on using only CLIP in zero shot, with no training for object Navigation Tasks. We hypothesize that with no additional training and in zero shot, CLIP embeddings are powerful enough for Object Navigation Tasks. To this end, we explore the above hypothesis in a 2D environments like gSCAN and BabyAI.

Keywords: CLIP, zero shot, deep learning, object detection

1 Introduction

We believe that model's inherent capabilities are not currently explored fully and we lack full understanding of what models can and cannot do. Currently models are growing rapidly to keep up with the want of more performance, but the same emphasis is not put on understanding what the models are capable of. Our thought is that by probing models effectively and efficiently, we can gain new understanding and use cases out of older models. By doing so, we will limit the need for new discoveries of model architectures while still achieving the same or similar performance. For this study we are considering CLIP, which is a multi-modal model trained on 400 million image-caption pairs. This model aims to achieve zero-shot image classification as it "efficiently learns visual concepts

from natural language supervision" [1]. In this study, we are probing CLIP for the task of goal-based 2-D object navigation. Currently, there is ongoing research that is looking into making a model that will allow "humans to interactively train artificial agents to understand language instructions" [1] by using 2-D simulators. The idea behind this is to be able to provide prompts for an agent to learn and perform contextual based instructions based on objects, colors and tasks ([1] [2] [3]). Their reasoning for this is they believe that current deep learning methods are not strong enough in "learning a language with compositional properties"[1]. The findings show that imitation and reinforcement learning "methods scale and generalize poorly when it comes to learning tasks with a compositional structure"[1].

We believe that with efficient probing of CLIP we can draw out a few of these properties like navigation, shapes and colors in zero shot. We suspect this because of interesting results that have popped up through the use of CLIP embeddings to bolster agents in 3-D environments [4] [5] for tasks such as object goal navigation, room rearrangement and point goal navigation. Most 3-D environments with agents use RGB-D channels to have a full understanding of the environment. However, with recent studies, CLIP image embeddings have been shown to be used as an alternative with similar results using only RGB channels, without needing the depth dimension. When probed further the CLIP image embeddings were able to capture the necessary semantic information such as reach-ability, object presence on a grid, object presence in an image and free space. Because of these various uses and traits, we believe we should be able to navigate an agent to a goal destination without the cumbersome overhead caused by imitation and reinforcement learning. To this end, we leverage the embeddings of CLIP which we deem to be powerful enough to navigate in a 2D world guided by the goal text in Zero Shot with no additional training or pre-training.

2 Related Work

One related topic to this research is the ability to navigate within 3-D environments using CLIP. Some research that pursues such topics are **"CLIP on Wheels: Zero-Shot Object Navigation as Object Localization and Exploration"** [6] as well as **"Simple but Effective: CLIP Embeddings for Embodied AI"** [7]. Our work differs from these examples in 2 ways: Firstly, we are targeting a 2-D environment. Secondly, CLIP on Wheels focuses on object navigation by breaking it down into two steps of exploration and exploitation [6]. The CLIP image embeddings are leveraged only for the exploitation task, but not for the task of trajectory prediction (exploration) [6]. Also, both of them have pre-training steps, while we are using frozen CLIP embeddings with no training at all. Our research is aiming to use CLIP as an end to end solution of the object navigation task.

Another topic that we are focused on is using frozen CLIP to be a full end to end trajectory driver to navigate a 2-D environment. **"BabyAI: A Platform to Study the Sample Efficiency of Grounded Language Learning"** [1] and

"A Benchmark for Systematic Generalization in Grounded Language Understanding" [2] both focus on the task of navigating a 2-D environment using imitation learning and reinforcement learning for training. One similarity between our research and theirs is that we will be using a BabyAI like platform to conduct experiments.

3 Methods

For this project, we created a environment called "Grid2Photo" [8] which is similar to the BabyAI platform. More specifically, we created a 2-D grid world similar to MiniGrid [9], which is used in the BabyAI environment. This grid world, combined with the instructions inspired from BabyAI, will hopefully allow us to manipulate the environment using a prebuilt, frozen model. The model we have chosen for this task will be a CLIP model, given the recent findings of what these embeddings can be used for [6][7]. We intend to use Grid2Photo's output world for the image embeddings and the relevant task for the word embeddings.

In our Grid2Photo 2-D grid world where the grid is of size NxN, where N is a hyper-parameter, we are trying to navigate an object (i.e., a triangle or square) from its starting location to its goal destination which may be any shape or color. This task is annotated by a text prompt that will be specific in what it wants (i.e., "Move the triangle to the red circle"). From the initial configuration of the environment, we generate the images for the next possible states of the environment by considering only four moves (UP, LEFT, RIGHT, DOWN) that our object can make. We will encode these four images through CLIP's image encoder and find the similarity scores with the goal text embedding generated by passing the goal text to CLIP's text encoder (See Fig 1). We apply a softmax layer on the similarity scores to give us the probability function over the next possible states. We then sample out one state from the probability function and consider that image to be the next environment state. We repeat this until we reach the goal destination or max allowed number of actions (which is a hyper-parameter). We believe that by letting the model choose the direction to move, we can gain knowledge on whether CLIP is able to handle additional tasks outside of its predetermined purposes.

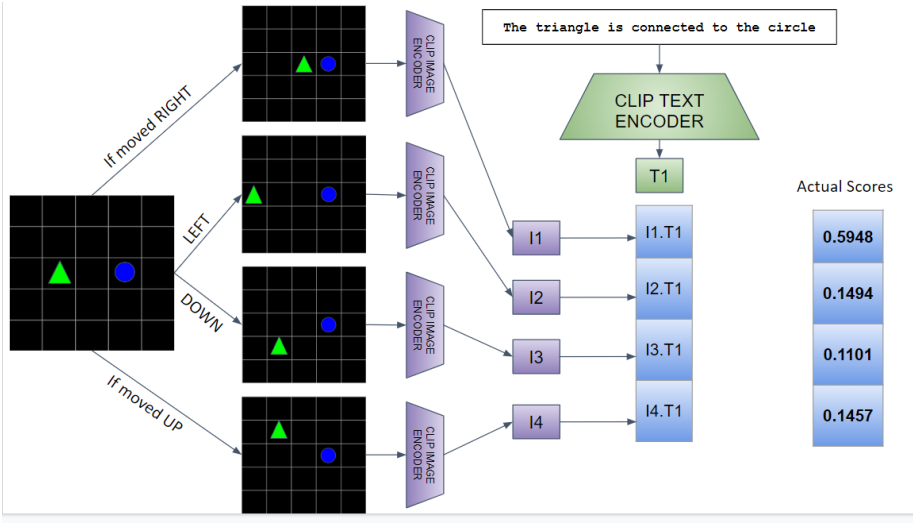

Fig. 1: Considering the second environment representation type of actually moving the object to the desired location. Four possible states are generated from the current state by moving the green triangle UP, DOWN, LEFT, and RIGHT. Similarity scores are calculated against the embeddings generated from the CLIP's encoders.

4 Experimental Design

For our experiments, we intend to use grids of various NxN sizes with a randomly placed object, and a randomly placed goal. The only restriction on the starting locations is that the starting object and the goal may not start on the same locations. By testing various sizes and locations, we can get a true understanding if CLIP is able to navigate relative positions on the 2-D grid world. By making the world simple without additional "distractions", we can analyze whether the model understands the task and provided image.

Keeping the time constraints in mind, we limited our experiments to use an environment where all points of interest take up space on the output image. No modifications of how they are represented location wise will be used. For the experiments, we will use four different shapes (square, rectangle, triangle, and circle) and six different colors (red, blue, green, yellow, grey, and purple) to generate our "movable object" (which from now on, we refer to as "agent") and goal object. We call an episode to be a success if the agent is able to reach the goal object in no less than the maximums number of steps (which is a hyperparameter). A unique pair of the agent and the goal including their locations on the grid is defined to be the configuration of the environment. We randomly generated 250 initial configurations of our environment for our experiments.

For our first experiment, we intend to find the effect of the grid size on the episode's success. We run the experiment on the 250 initial configurations on each

of the grid sizes of size 5, 8, 10 and 15 (25, 64, 100, 225 total spaces respectively). The maximum number of steps is capped at twice the grid size of the environment used in the episode (i.e., 10, 16, 20, 30). For our second experiment, we examine the effect of the text prompt used to guide the agent to the goal location. We use six different goal text objectives. which capture visual cues and the descriptors of the objects in the environment which include only shapes, only colors, and both the shapes and colors. This will directly reflect CLIP's capability of correlating surface-level image semantics with text semantics. See Fig(2) for reference.

The *<agent_description>* is *<visual_cue>* to the *<goal_description>*

Objective	Visual Cue	Description has SHAPE	Description has COLOR	Example
0	connected	YES	YES	The *green triangle* is *connected* to the *blue circle*
1	close	YES	YES	The *green triangle* is *close* to the *blue circle*
2	connected	YES	NO	The *triangle* is *connected* to the *circle*
3	close	YES	NO	The *triangle* is *close* to the *circle*
4	connected	NO	YES	The *green shape* is *connected* to the *blue shape*
5	close	NO	YES	The *green shape* is *close* to the *blue shape*

Fig. 2: This table shows the different types of text objectives used based on visual cues and descriptors of the agent and the goal.

Overall, for each of the 250 initial configurations of the environment, we run the episode for each of the four grid sizes and for each of the six-goal text objectives. Thus, in total, we run 6000 (250 * 4 * 6) episodes for both experiments combined.

To our knowledge, this is the first study that is probing CLIP to evaluate the navigation tasks in zero-shot with no training, in a 2D grid world environment. Thus, we do not have a pre-existing baseline to compare our approach with. For all our experiments, we assume that there is always a path to reach the goal object and a simple Breadth First Search would give us the path with 100% success for each episode. We strongly believe that this is the best-case baseline that we can compare our CLIP model against, as it allows us to analyze the contextual power of the input embeddings. Because the environments we are testing against have been used for imitation and reinforcement learning, our results will show CLIP's ability to navigate an environment in zero-shot without training.

5 Results

Prompt Engineering CLIP scores are very susceptible and hypersensitive to the text prompts used. After preliminary testing of CLIP with varied text prompts, we have identified that words that contain visual cues are better understood by the CLIP model. This is the primary reason why visual cues used in the text prompt design are "connected" and "close" as these words have a direct correlation with the image. Also, we have noticed that CLIP doesn't understand direction words such as left, right, up, and down. (See fig 4). We believe since CLIP is trained on 400M image caption pairs of diverse domains and is not exclusively trained on one domain, it is difficult for it to understand top/up, left, bottom/down, and right.

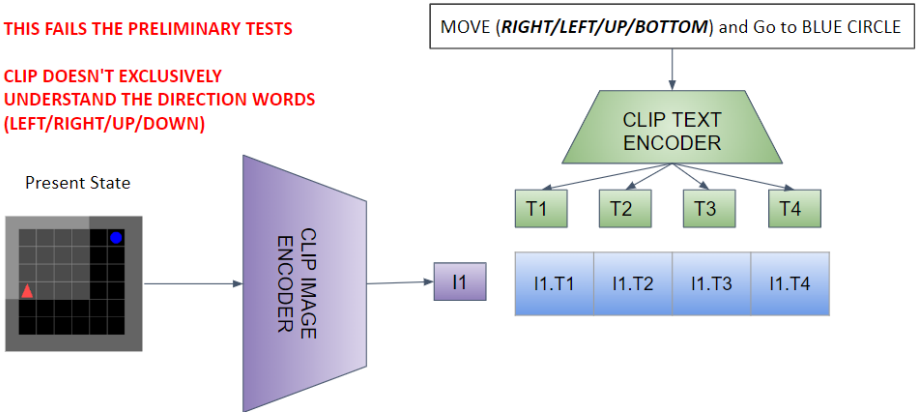

Fig. 3: This architecture fails the preliminary tests as the CLIP model doesn't understand the direction words when given in the text prompts. This is one of the reasons to consider using visual cues in the text prompts design.

Effect of grid sizes The per grid size and per goal text objective success rates are shown in Fig 4. The success rate is defined as the percentage of successful episodes in 250 initial configurations that we start the episode with. We observe that for each goal objective, the success rate increases as the grid size increases. This is an interesting result and also counter-intuitive. To explain this, we have dug deep in understanding why the episode fails and interestingly, we have found that the CLIP scores form a cyclic path all the time to max out the number of steps and thus the episode fails. This cyclic formation decreases as the grid size increases and thus explains the increase in success rate.

	success_rate					
goal_objective	0	1	2	3	4	5
grid_squares						
5	0.372	0.360	0.376	0.376	0.356	0.392
8	0.572	0.516	0.580	0.544	0.576	0.548
10	0.596	0.516	0.604	0.508	0.608	0.560
15	0.548	0.556	0.640	0.604	0.484	0.428

Fig. 4: This table shows the success rates for the 250 configurations for each goal text objective and for each grid size.

Shapes and Colors The success rates against goal text objectives of colors and shapes, just shapes, and just colors over various grid sizes are plotted in Fig 5. The goal text objectives 0, 2 and 4 use the visual cue "connected" and the objectives 1, 3 and 5 use the visual cue "close". We observe that for each grid size, the success rates of "connected" text objectives are greater than their "close" counterparts. This concludes that the visual cue "connected" is more powerful to correlate the distance in the corresponding image compared to the visual cue "close". Also, as the grid size increases, we observe that specifying the colors of the object in the text prompt negatively impacts the success rates. Specifying just the shapes of the objects in the prompts gives the highest success rates. According to [10], there exists an inherent reporting bias in the language model's perception of color. In the case of CLIP, the text encoder suffers from this bias and thus performs weaker compared to specified shapes.

Our observations on this environment have been only the agent and the goal object, meaning we are not adding other "noisy" objects to confuse the model. It would be an interesting direction to explore the behavior of CLIP when obstacles are placed randomly on the grid. The challenge would be to come up with an appropriate visual cue by carefully designing and experimenting with the text prompts to move past the obstacles. The only reason that we found for failure to reach the goal object is because of the agent taking a cyclic path. We can mitigate this by using CLIP scores for performing a heuristic search for the goal object in the grid. Although our baseline is whether or not our the agent was able to reach the goal within X actions, we believe that future work could use a search algorithm such as A* [11] to be a best case baseline (100%) and have worst case be the total number of squares in a world (0%). This would allow for a different understanding of the models accuracy in zero-shot tasks. Lastly, future work could not only implement additional tasks/actions more closely aligned

with BabyAI[1] for added complexity, but the images used for the input could be modified to probe specific traits of the tested model.

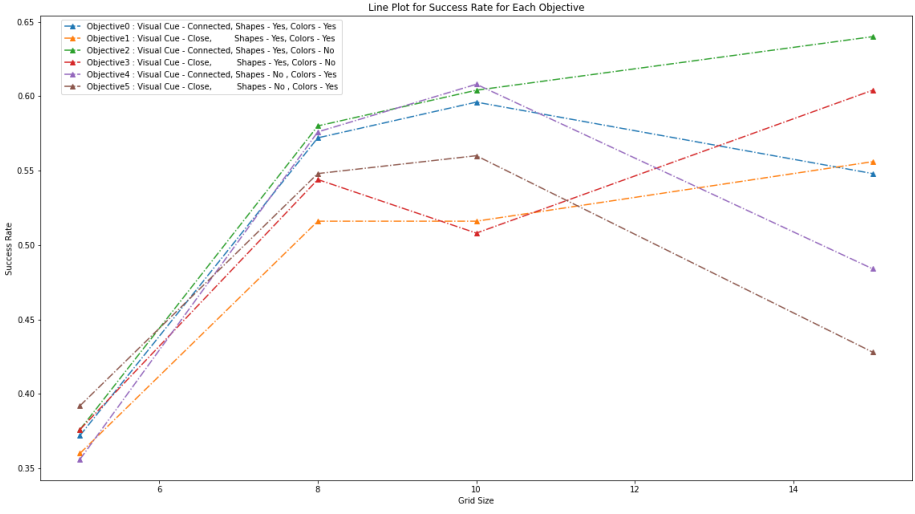

Fig. 5: The line plot illustrates the success rate for each objective as the grid size increases.

6 Conclusion

We think that the intrinsic capabilities of models are now underutilized and that our knowledge of what models can and cannot achieve is incomplete. To this end, we demonstrated that CLIP, a vision language model can be used for 2D Navigation in a 2D grid world in zero-shot and with no additional training/ fine-tuning. We also created a platform called "Grid2Photo" which can generate the images required based on the configuration of the environment and perform actions on the environment such as navigating the agent in the four directions.

Finally, we believe that with the direction that AI is taking in society, that we should comment on possible ethical implications of our research. For this research specifically, we believe that should a model be shown that it can do something with an extremely negative impact, the individuals who have found this discovery should be careful about releasing the information publicly. Many models are available publicly, and should these already very accessible models be shown to have a nefarious ability, we could potentially open a can of worms onto the community. Unlike some of the newer generative models, these older models would have more widespread freedom due to them being out in the community and under less scrutiny. That all said, we believe that probing models will provide many more positive uses than negative ones. It should be fair to

assume that whatever a model's negative traits are, they should be the main concern when working with it, due to them being inherited.

References

1. Chevalier-Boisvert, M., Bahdanau, D., Lahlou, S., Willems, L., Saharia, C., Nguyen, T.H., Bengio, Y.: Babyai: First steps towards grounded language learning with a human in the loop. CoRR **abs/1810.08272** (2018)
2. Ruis, L., Andreas, J., Baroni, M., Bouchacourt, D., Lake, B.M.: A benchmark for systematic generalization in grounded language understanding. CoRR **abs/2003.05161** (2020)
3. Lake, B., Baroni, M.: Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In: International conference on machine learning, PMLR (2018) 2873–2882
4. Deitke, M., Han, W., Herrasti, A., Kembhavi, A., Kolve, E., Mottaghi, R., Salvador, J., Schwenk, D., VanderBilt, E., Wallingford, M., Weihs, L., Yatskar, M., Farhadi, A.: Robothor: An open simulation-to-real embodied AI platform. CoRR **abs/2004.06799** (2020)
5. Manolis Savva*, Abhishek Kadian*, Oleksandr Maksymets*, Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., Parikh, D., Batra, D.: Habitat: A Platform for Embodied AI Research. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). (2019)
6. Gadre, S.Y., Wortsman, M., Ilharco, G., Schmidt, L., Song, S.: Clip on wheels: Zero-shot object navigation as object localization and exploration. arXiv preprint arXiv:2203.10421 (2022)
7. Khandelwal, A., Weihs, L., Mottaghi, R., Kembhavi, A.: Simple but effective: CLIP embeddings for embodied AI. CoRR **abs/2111.09888** (2021)
8. Gorman, T., Raju Kurapati, G.S.: Grid2Photo
9. Chevalier-Boisvert, M., Willems, L., Pal, S.: Minimalistic gridworld environment for gymnasium (2018)
10. Paik, C., Aroca-Ouellette, S., Roncone, A., Kann, K.: The world of an octopus: How reporting bias influences a language model's perception of color. arXiv preprint arXiv:2110.08182 (2021)
11. Hart, P.E., Nilsson, N.J., Raphael, B.: A formal basis for the heuristic determination of minimum cost paths. IEEE Transactions on Systems Science and Cybernetics **4**(2) (1968) 100–107