September 2019

Devendra Singh Chaplot — Research Statement

Webpage: https://www.cs.cmu.edu/~dchaplot

My primary scientific interest lies in the field of machine learning, computer vision, and robotics. Advances in machine learning in the last decade have led to 'digital intelligence', i.e. machine learning models capable of learning from vast amounts of labeled data to perform several digital tasks such as speech recognition, face recognition, machine translation and so on. My long-term research goal is to design algorithms capable of 'physical intelligence', i.e. building intelligent embodied autonomous agents capable of learning to perform complex tasks in the physical world involving perception, natural language understanding, reasoning, planning, and sequential decision making. To achieve this goal, my current research focuses on training embodied navigation agents in 3D environments capable of learning to localize, building semantic maps, path planning, navigating to semantic goals, following language instructions, and answering questions. While classical robotics and language navigation systems are brittle and fail to generalize to unseen scenarios, I aim to build learning-based embodied navigation models which are robust to dynamic environment changes and capable of generalizing to unseen environments and understanding novel language queries.

In order to navigate in 3D environments and perform complex tasks, embodied agents require both spatial understanding as well as semantic understanding of the scene. Spatial understanding involves recognizing obstacles and traversable space from raw RGB images (perception), estimating egomotion (pose estimation), remembering previously seen obstacles as the agent moves (mapping), exploring the environment efficiently and planning a path to the goal under uncertainty (path planning).

While spatial understanding gives an agent basic obstacle avoidance and geometric navigation capabilities, semantic understanding is essential for performing complex tasks such as finding semantic goals (specific objects, rooms, exit and so on), following natural language instructions and answering questions. Semantic understanding not only involves recognizing objects, regions and their properties from raw RGB observations but also understanding visual cues (like Exit signs) and learning common-sense (beds are more likely to be found in bedrooms). It also involves grounding words in visual objects and their properties (what does 'green' look like?), and complex contextual reasoning (such as relational reasoning: 'left of', 'not green', or pragmatics: 'largest'). Semantic understanding requires the agent to build a semantic map and perform complex reasoning on this map to follow natural language instructions and answer questions.

In addition to spatial and semantic understanding, another aspect which makes embodied intelligence challenging but also powerful at the same time is activeness or the ability to choose actions. Unlike traditional machine learning which learns from a static dataset passively, embodied agents have the ability to choose actions which affect their future observations. They can learn to decide task-dependent actions to be more efficient and effective. This makes training embodied navigation agents challenging, as they not only need to learn rich and meaningful spatial and semantic representations but also learn a task-dependent policy on top of these representations.

My research has provided several key advances towards building intelligent embodied navigation agents capable of spatial and semantic understanding. Specifically, I have worked on training autonomous agents capable of active localization [6], active mapping [4], pose estimation [12], path planning [11], visual navigation [10, 2], following natural language instructions [3] and answering questions [5]. These embodied navigation agents do not assume any prior perceptual or linguistic knowledge and learn end-to-end using deep reinforcement learning from raw-pixel based first-person view of the environment and language queries. They are not only capable of mapping, localizing and navigating in unseen environments but also able to tackle unseen instructions and questions and transfer the knowledge of grounded concepts across different tasks. Many of these models are the state-of-the-art for the respective tasks and have won the **CVPR 2019 Habitat Navigation Challenge** [9, 4] and the **Visual Doom AI Competition** 2017 [14, 10, 2] and received **Best Paper** [12] and **Best Demo** [2] awards. Several works have also been highlighted in media articles such as MIT TechReview [8], Techcrunch [7], Popular Science [1], and Engadget [13]. My near-term research goal is to combine the capabilities of the above models to build an autonomous agent capable of performing complex navigational tasks in the real-world.

References

[1] Atherton, K. D. 2016. Trained A.I. beats humans in doom deathmatches. https://www.popsci.com/trained-ai-bets-humans-in-doom-deathmatches.

[2] Chaplot, D. S., and Lample, G. 2017. Arnold: An autonomous agent to play FPS games. In *AAAI*, 5085–5086.

[3] Chaplot, D. S.; Sathyendra, K. M.; Pasumarthi, R. K.; Rajagopal, D.; and Salakhutdinov, R. 2018. Gated-attention architectures for task-oriented language grounding. In *AAAI*.

[4] Chaplot, D. S.; Gupta, S.; Gupta, A.; and Salakhutdinov, R. 2019a. Modular Visual Navigation using Active Neural Mapping. *under review*.

[5] Chaplot, D. S.; Lee, L.; Salakhutdinov, R.; Parikh, D.; and Dhruv, B. 2019b. Embodied Multimodal Multitask Learning. *under review*.

[6] Chaplot, D. S.; Parisotto, E.; and Salakhutdinov, R. 2018. Active neural localization. In *ICLR*.

[7] Coldewey, D. 2016. Scientists teach machines to hunt and kill humans — in doom deathmatch mode. http://tcrn.ch/2cWvt10.

[8] Gent, E. 2017. Machines are developing language skills inside virtual worlds. https://www.technologyreview.com/s/608380/machines-are-developing-language-skills-inside-virtual-worlds/.

[9] Habitat Challenge 2019. https://aihabitat.org/challenge/.

[10] Lample, G., and Chaplot, D. S. 2017. Playing FPS games with deep reinforcement learning. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[11] Lee, L.; Parisotto, E.; Chaplot, D. S.; Xing, E.; and Salakhutdinov, R. 2018. Gated Path Planning Networks. In *Proceedings of the 35th International Conference on Machine Learning*.

[12] Parisotto, E.; Chaplot, D. S.; Zhang, J.; and Salakhutdinov, R. 2018. Global pose estimation with an attention-based recurrent network. In *CVPR Deep Learning for Visual SLAM workshop*.

[13] Souppouris, A. 2016. Facebook and Intel reign supreme in 'Doom' AI deathmatch. https://www.engadget.com/2016/09/22/facebook-and-intel-reign-supreme-in-doom-ai-deathmatch/.

[14] Visual Doom AI Cmpetition. http://vizdoom.cs.put.edu.pl/competitions/vdaic-2017-cig/results.