Carnegie Mellon University

Gated-Attention Architectures for Task-oriented Language Grounding

Devendra Singh Chaplot

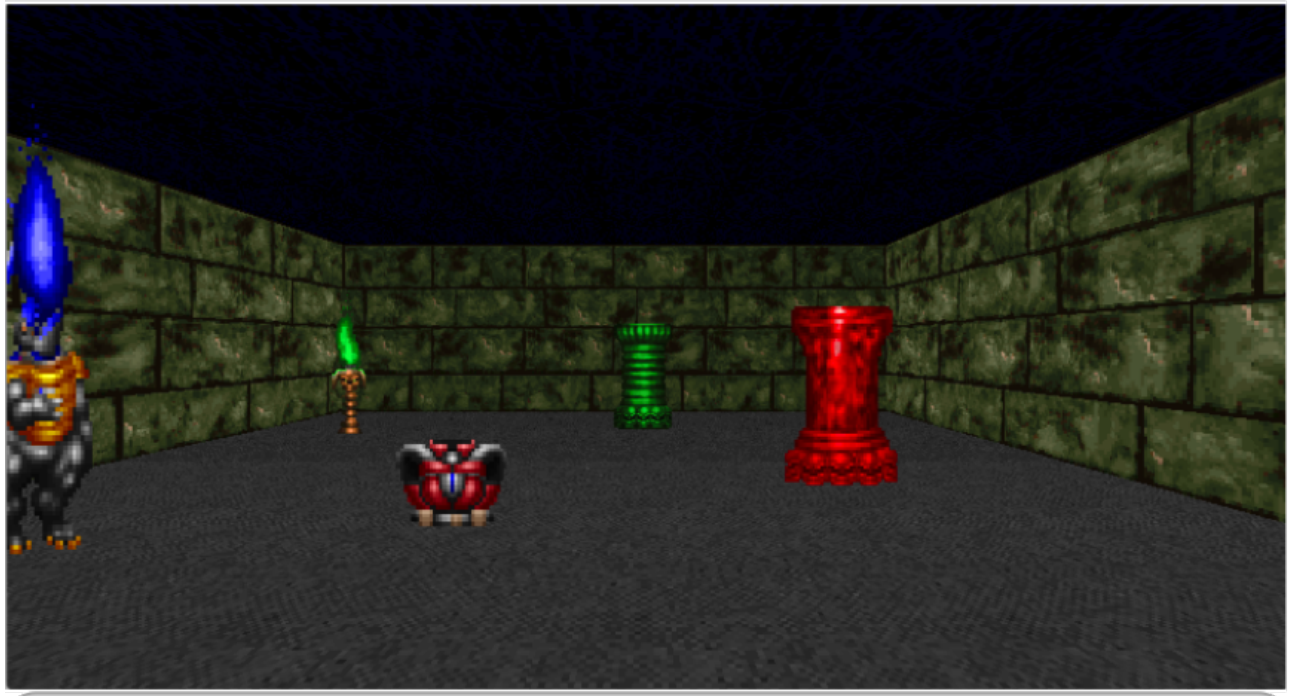
Kanthashree Mysore

Rama Kumar Pasumarthi

Dheeraj Rajagopal

Ruslan Salakhutdinov

Task-oriented language grounding



Go to the green torch

Train

Go to the short red torch
Go to the blue keycard
Go to the largest yellow object
Go to the green object

Test

Go to the tall green torch
Go to the red keycard
Go to the smallest blue object

Demo video

https://www.youtube.com/watch?v=JziCKsLrudE

Challenges

- *recognize* objects in raw pixel input,
- *explore* the environment, handle occlusion
- *ground* each concept of the instruction in visual elements or actions,
- *reason* about the pragmatics of language, and
- *navigate* to the correct object while avoiding incorrect ones.

Multitask Learning: Single model to tackle multiple instructions

Zero-Shot Learning: Generalize to unseen attribute-object pairs

Related work (1)

- Grounding Language in Robotics.
  - Guadarrama et al. 2014, Chao, Cakmak, and Thomaz 2011; Lemaignan et al. 2012, Chu et al. 2013, Kulick et al. 2013, Guadarrama et al. 2013, Bollini et al. 2013, Beetz et al. 2011 etc.
- Mapping Instructions to Action Sequences.
  - Chen and Mooney (2011) and Artzi and Zettlemoyer (2013): semantic parsing to map navigational instructions to a sequence of actions.
  - Mei, Bansal, and Walter (2015): neural mapping of instructions to sequence of actions

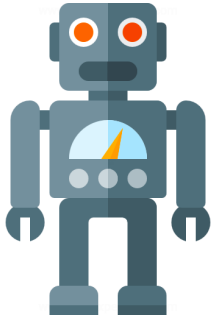
Related work (2)

- Deep reinforcement learning using visual data.
  - Deep Reinforcement learning approaches for playing FPS games (Lample and Chaplot 2016; Wu and Tian 2017; Dosovitskiy and Koltun 2017).
  - Zhu et al. (2016): target-driven visual navigation
  - Yu, Zhang, and Xu (2017): learning to navigate in a 2D maze-like environment and execute commands
  - Misra, Langford, and Artzi (2017): mapping raw visual observations and text input to actions in a 2D Blocks environment.
  - Oh et al. (2017): zero-shot task generalization in a 3D environment.

Experimental setting

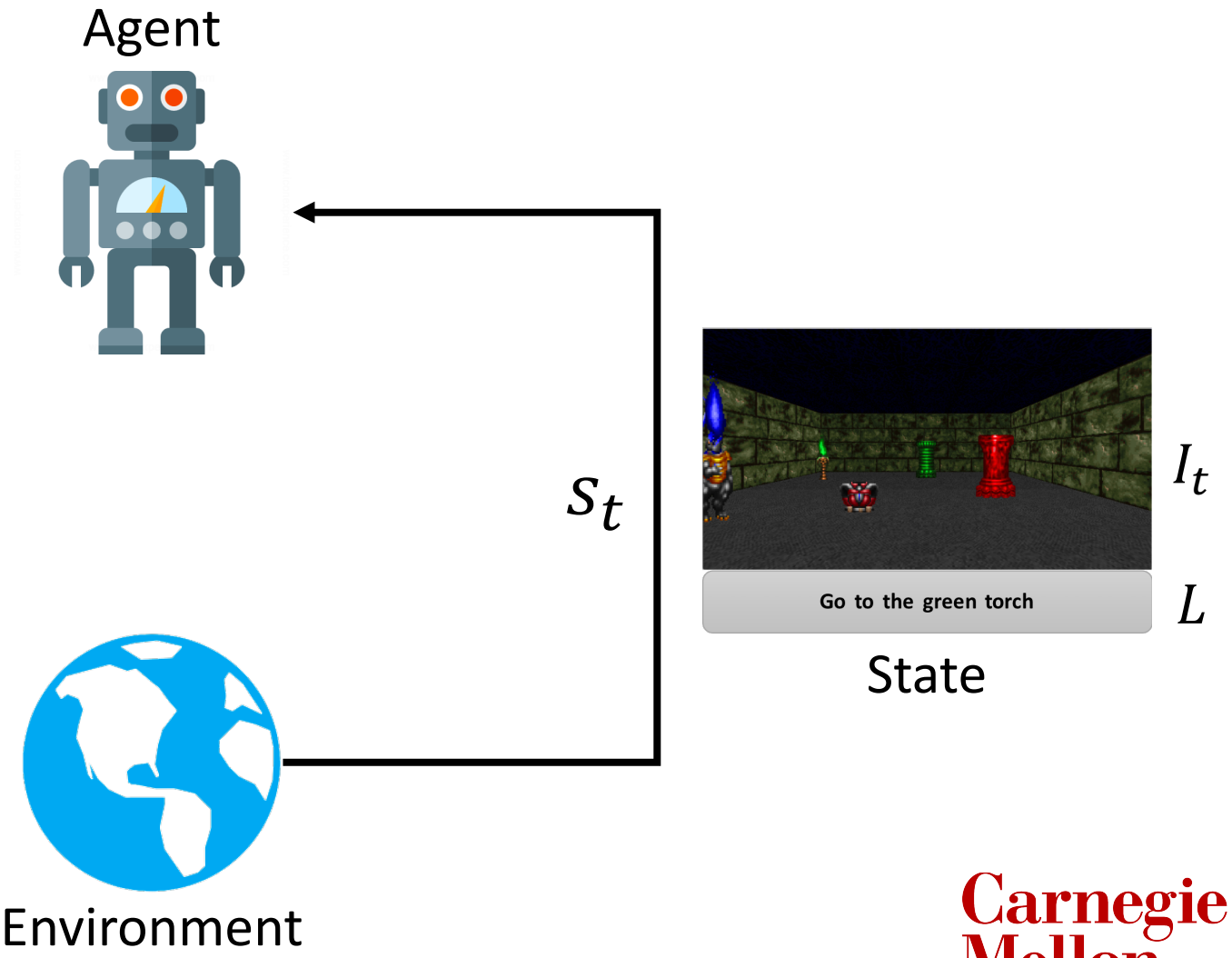
Experimental setting

Agent

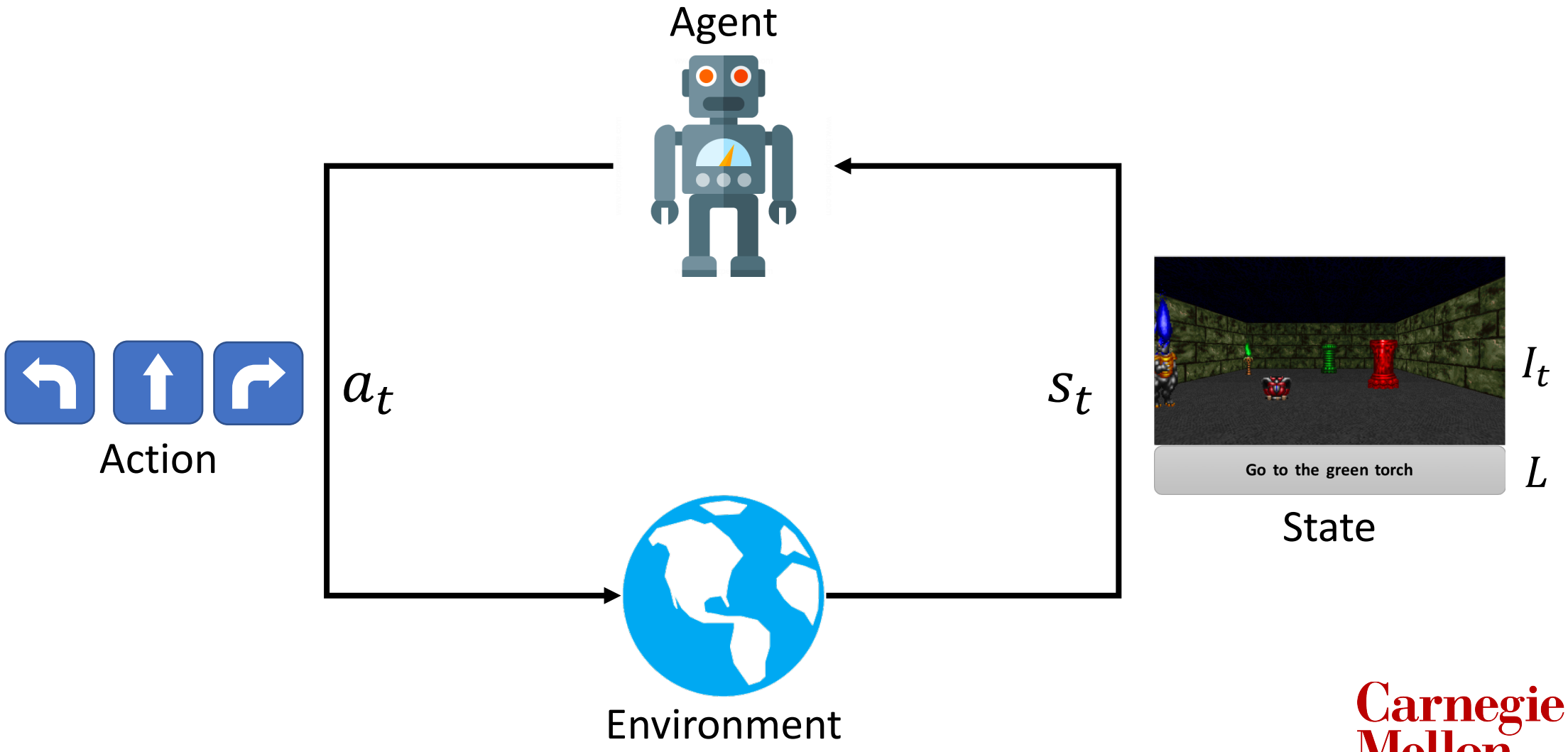


Environment

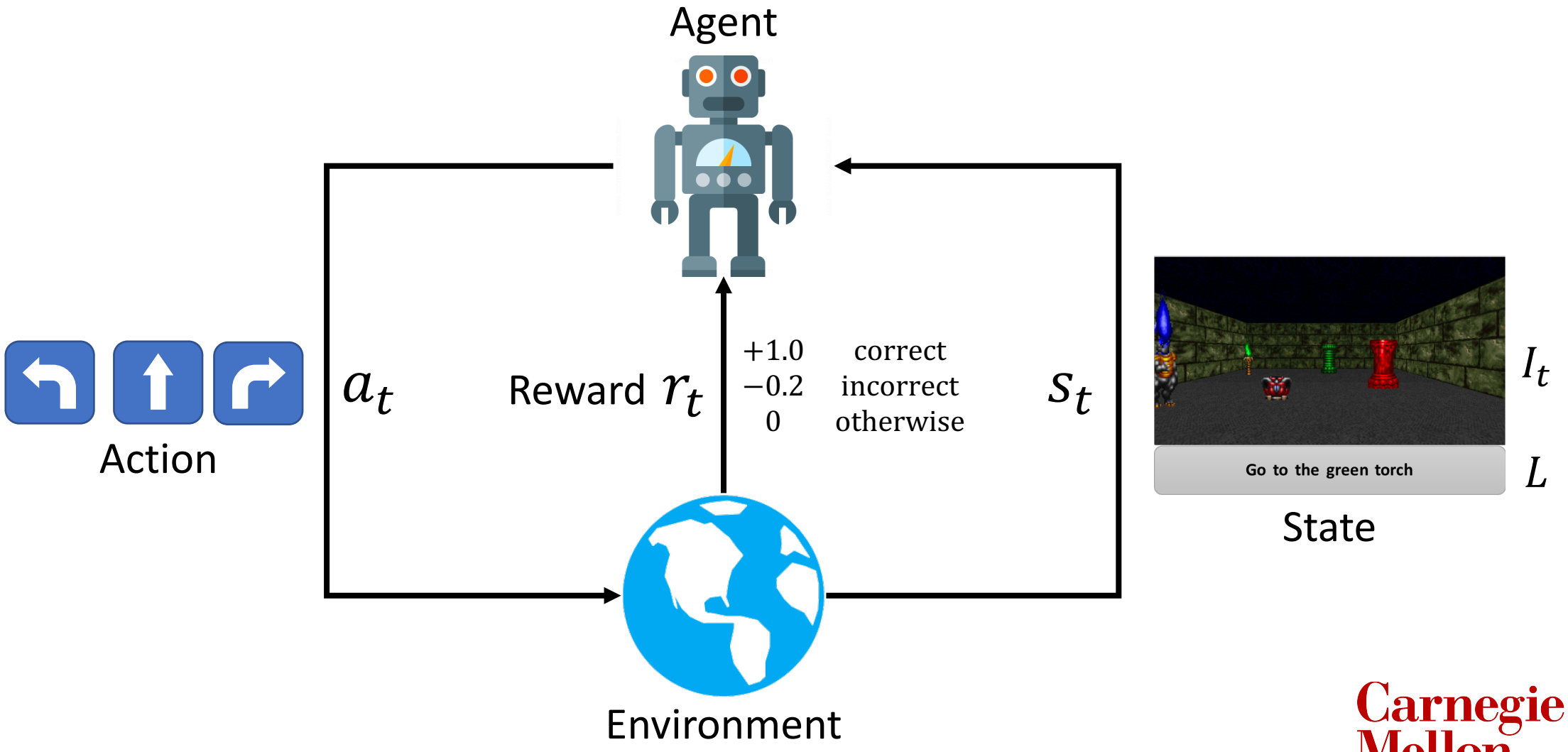
Experimental setting



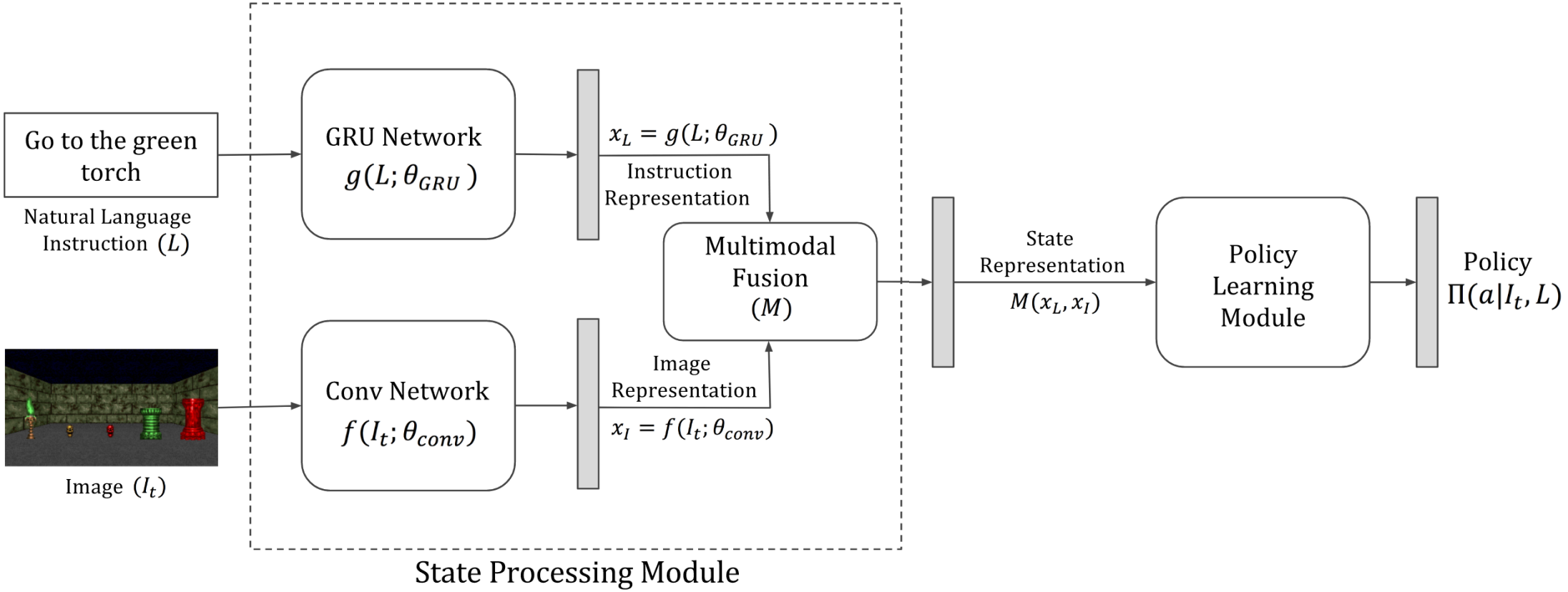
Experimental setting



Experimental setting



Network overview

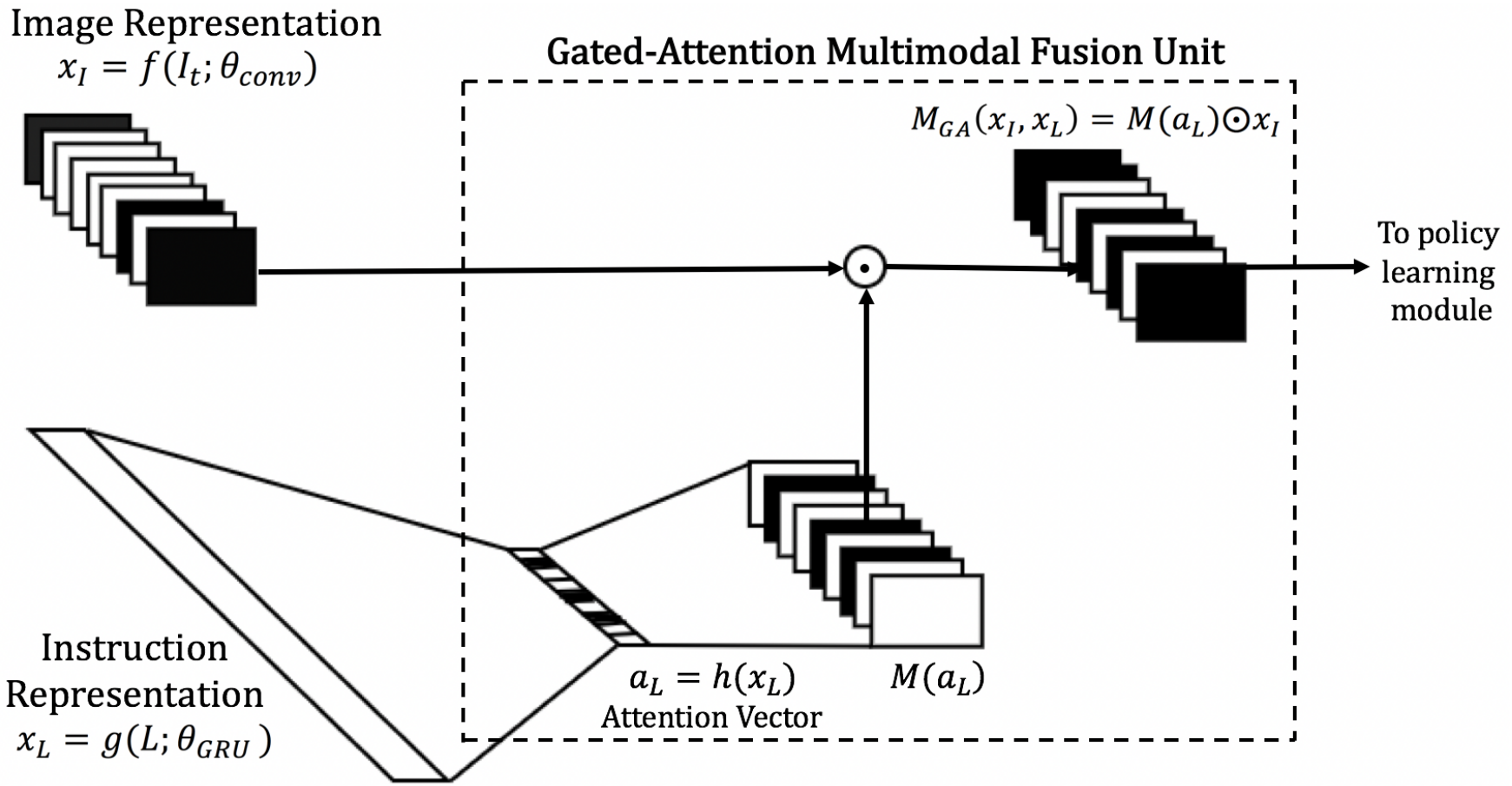


Multimodal Fusion

- Baseline Approach: Concatenation
- Proposed Approach: Gated-Attention

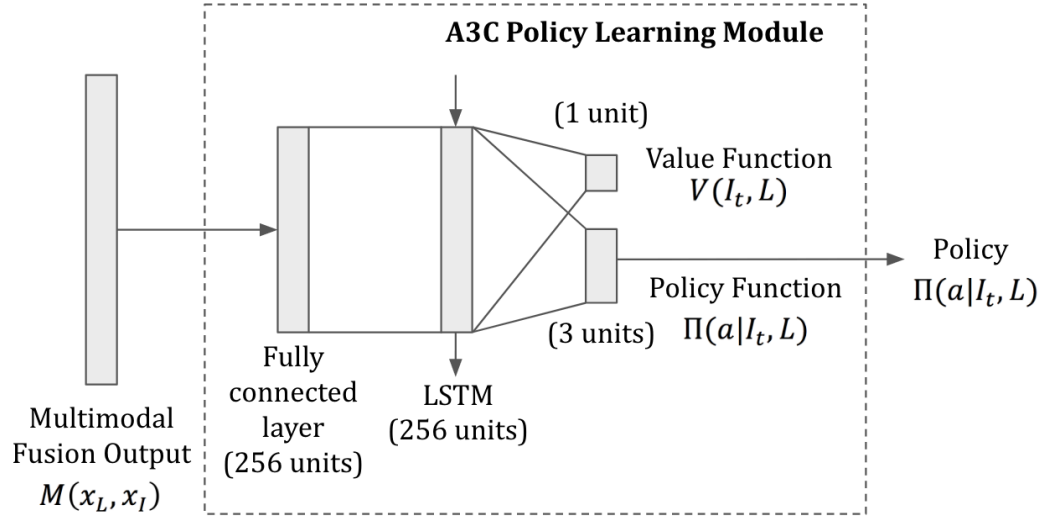
- Gated-Attention (Dhingra et al.)
  - attention weights for features maps, determines which filters to attend to
  - element-wise product (Gating)
  - creates instruction-specific convolutional filter representations

Gated-Attention



Policy Learning

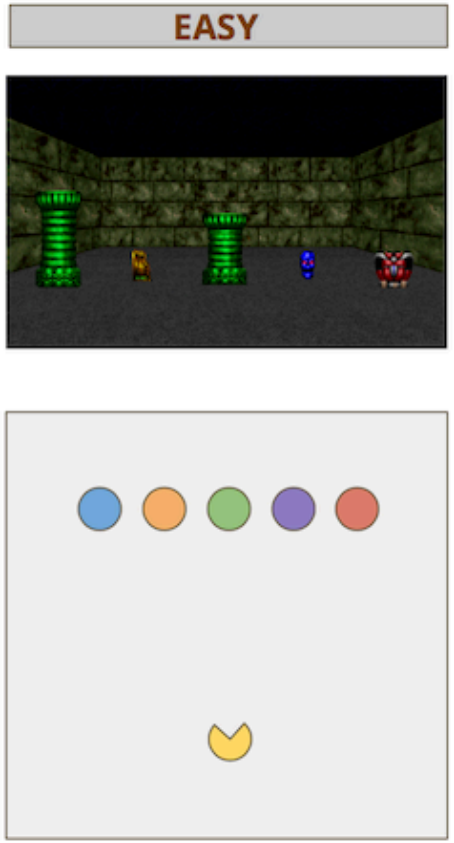
- Asynchronous Advantage Actor-Critic (A3C) (Mnih et al.)
  - uses a deep neural network to parametrize the policy and value functions and runs multiple parallel threads to update the network parameters.
  - use **entropy regularization** for improved exploration
  - use **Generalized Advantage Estimator** to reduce the variance of the policy gradient updates (Schulman et al.)



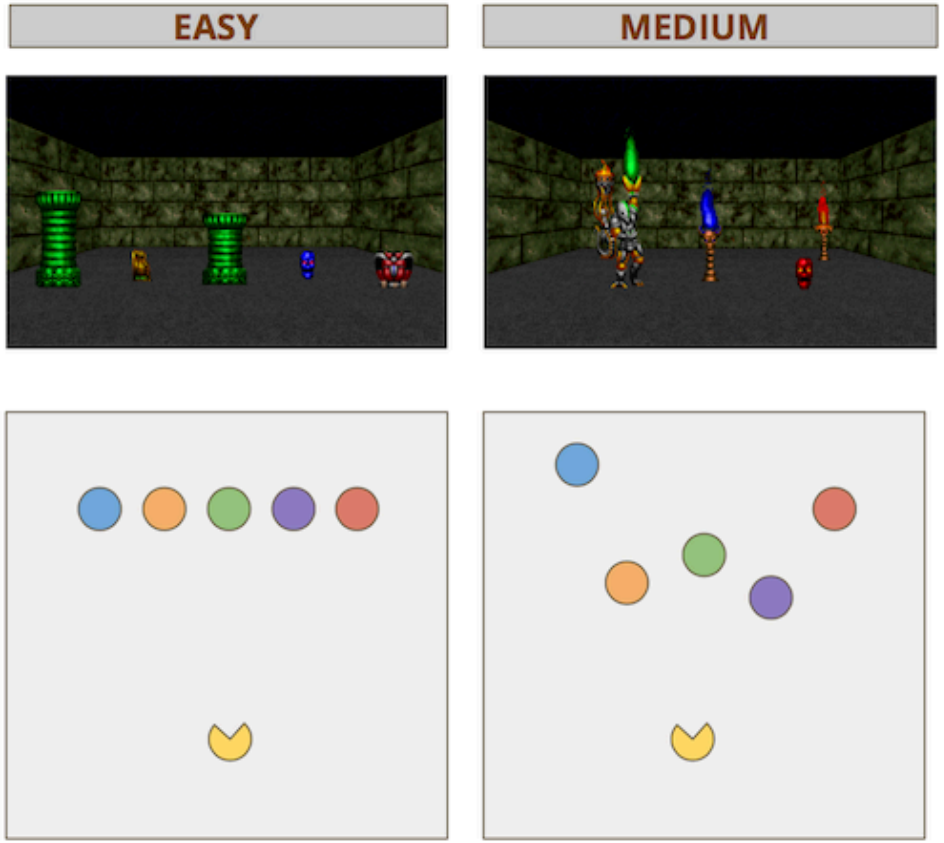
Environment

- 18 objects
- 5 types of objects
- Different colors and sizes
- Superlative instructions:
  - Largest, smallest
- Combinations
  - Tall green torch
  - Largest red object
- 70 instructions

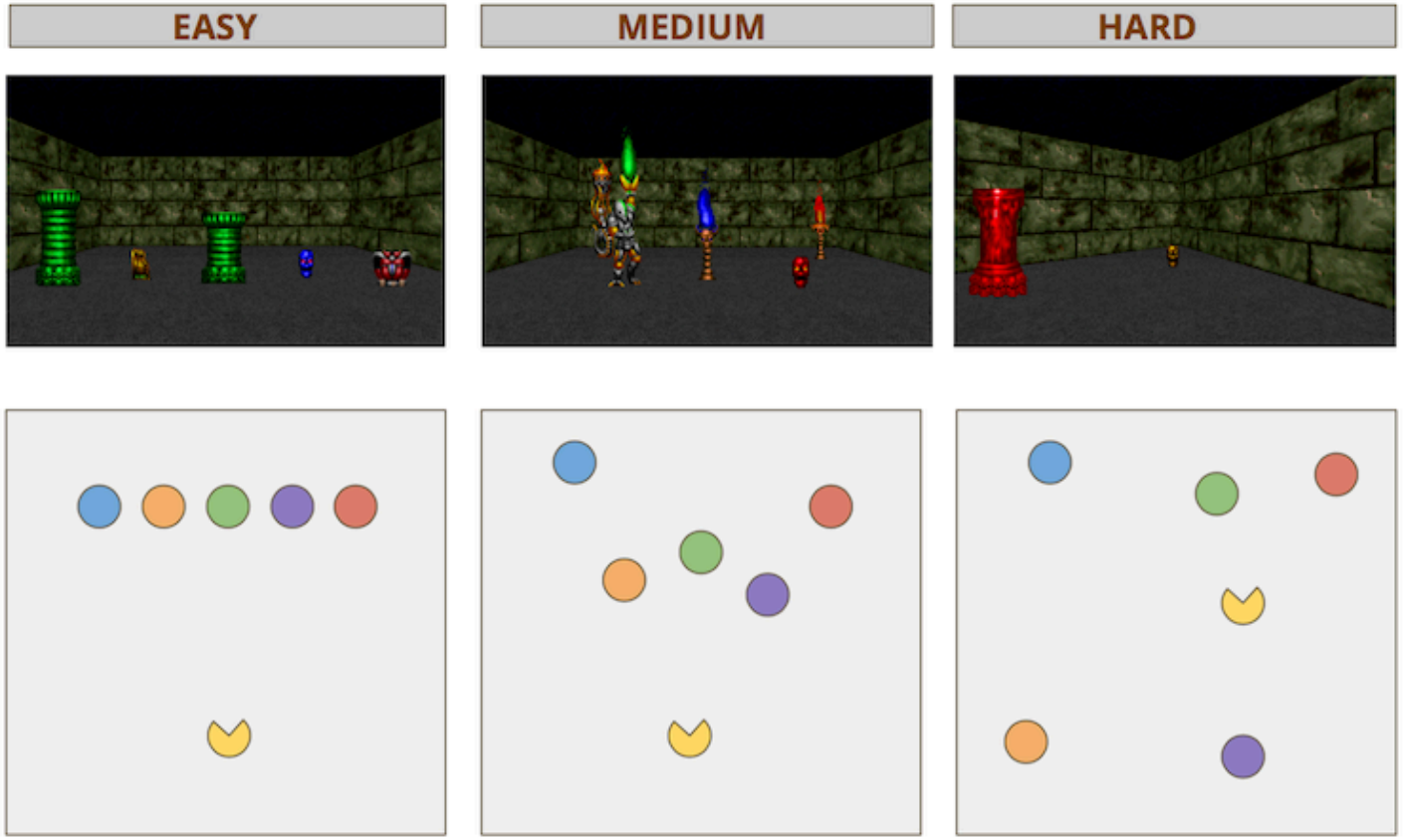
Environment difficulty



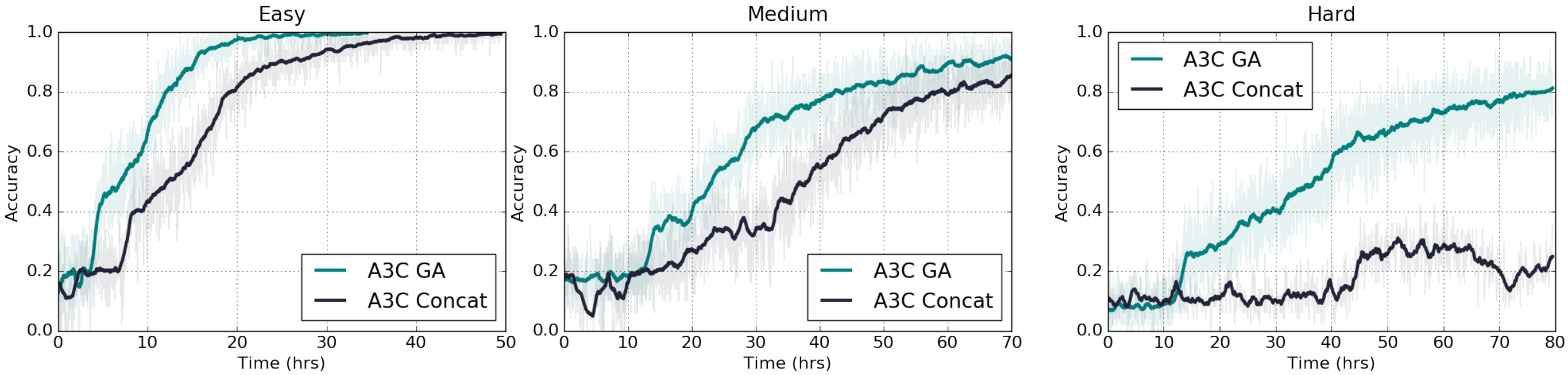
Environment difficulty



Environment difficulty

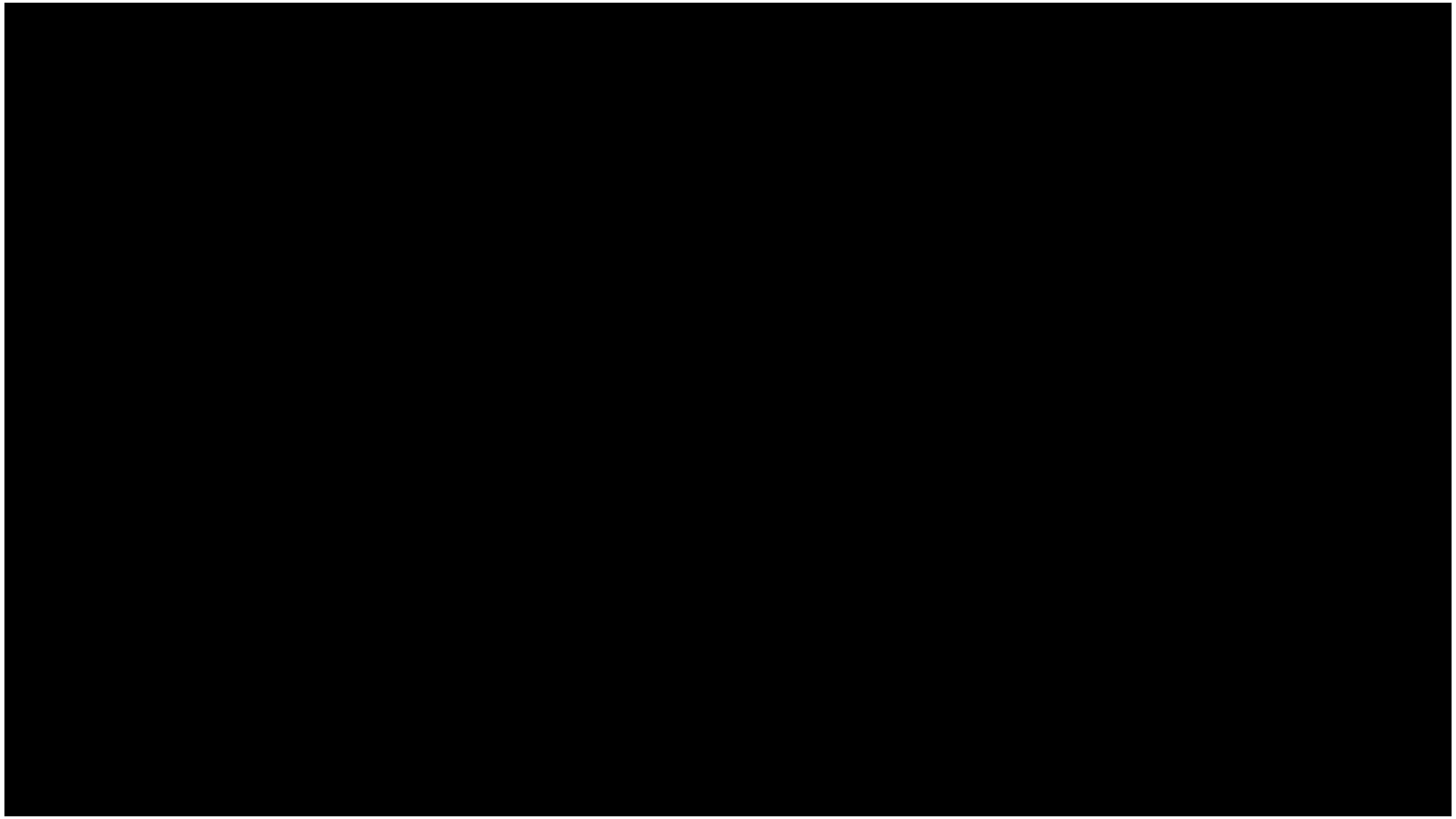


Results

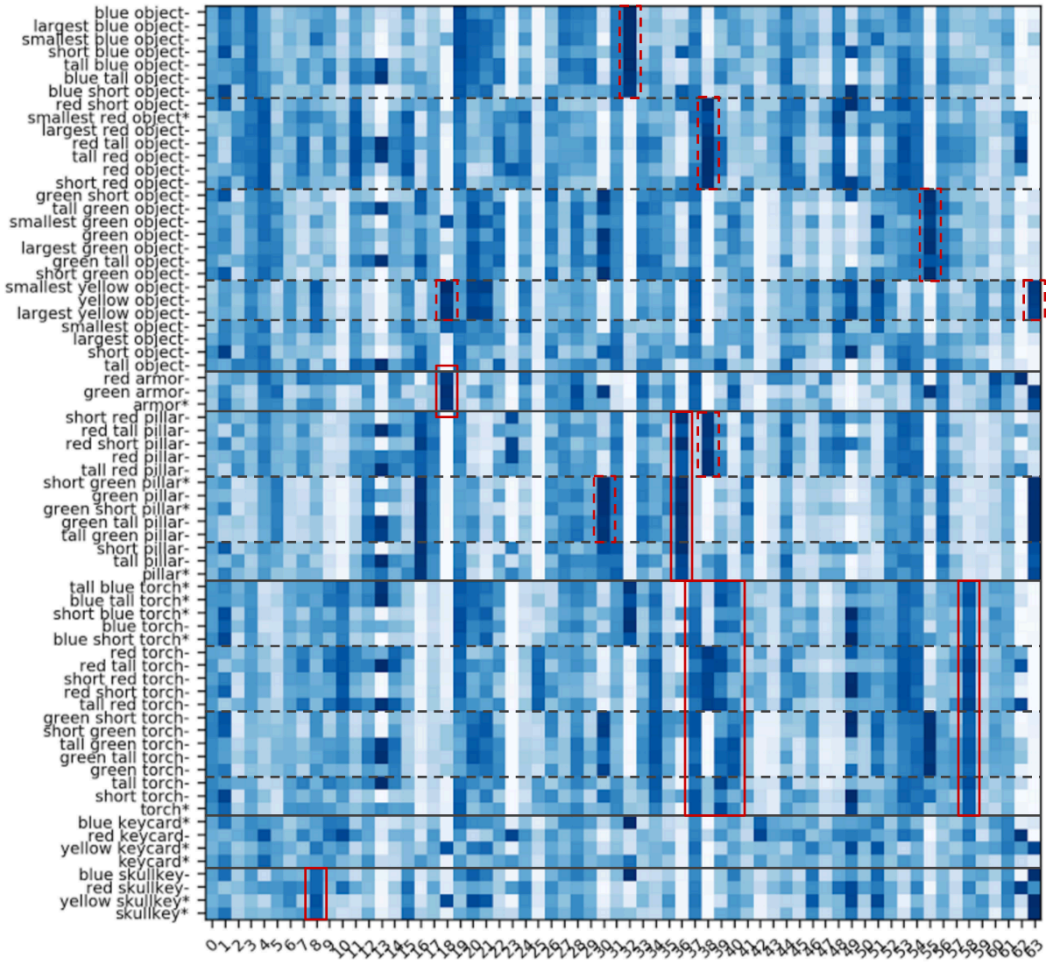

Easy
Medium
Hard
Accuracy
Time (hrs)
A3C GA
A3C Concat
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60
70
80

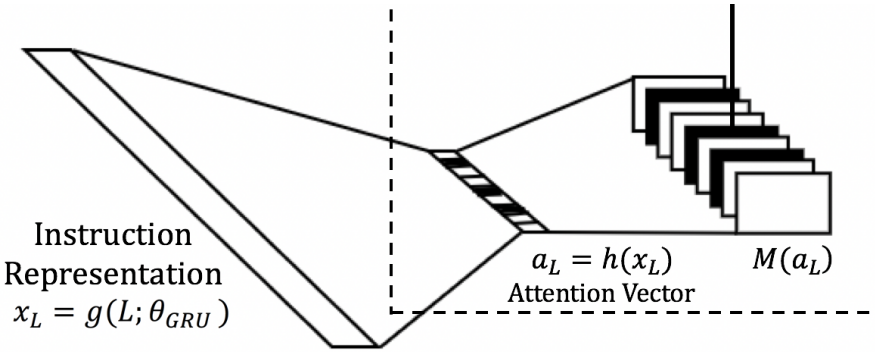
Training Progress

https://www.youtube.com/watch?v=o_G6was03N0

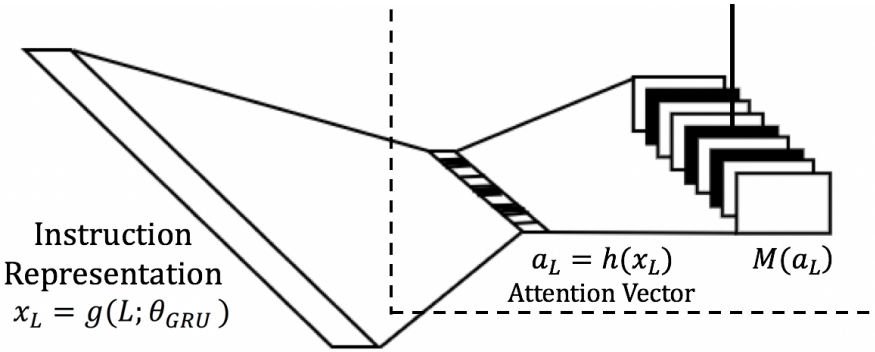


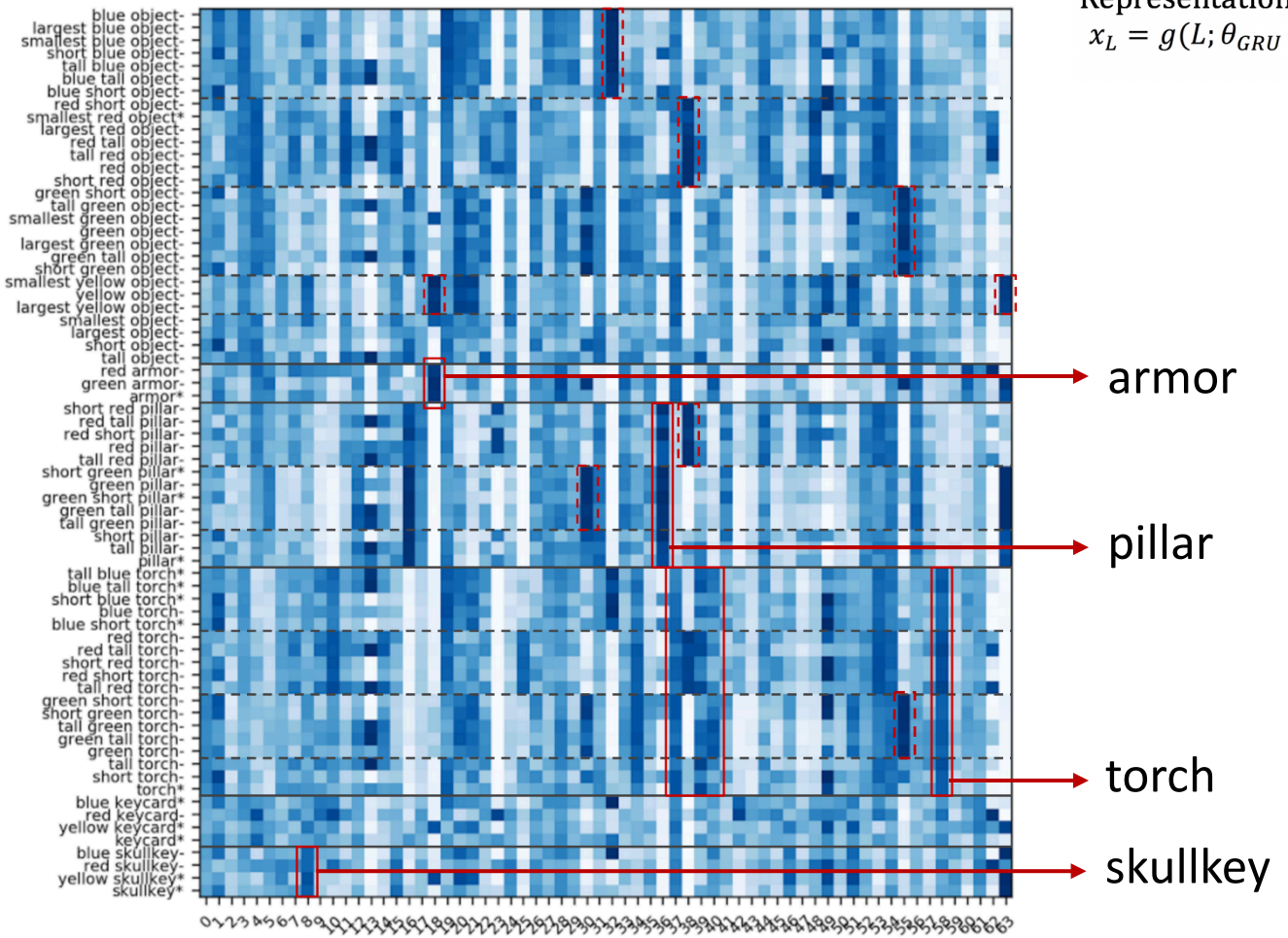
Attention map



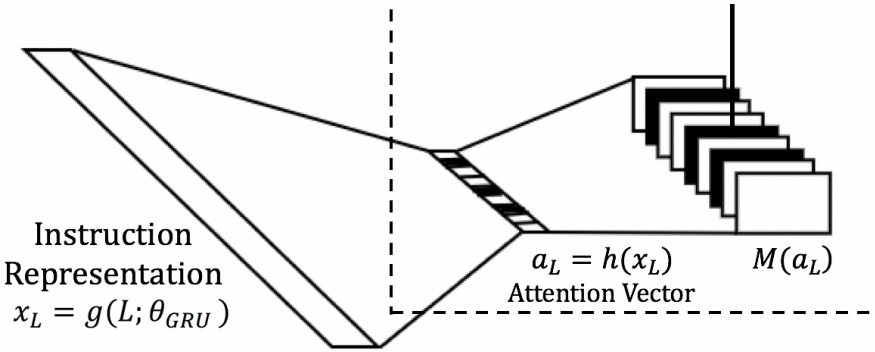


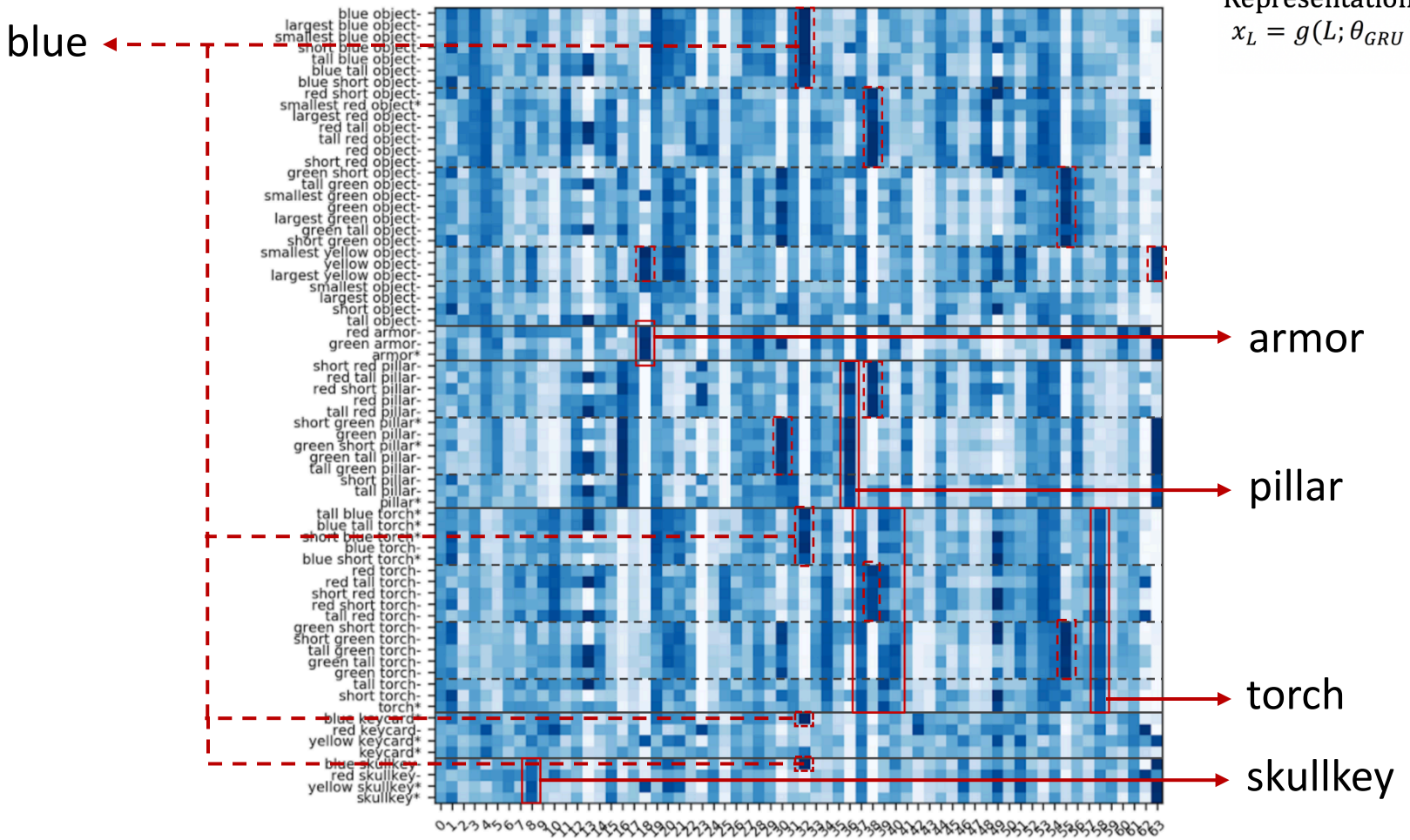
Attention map



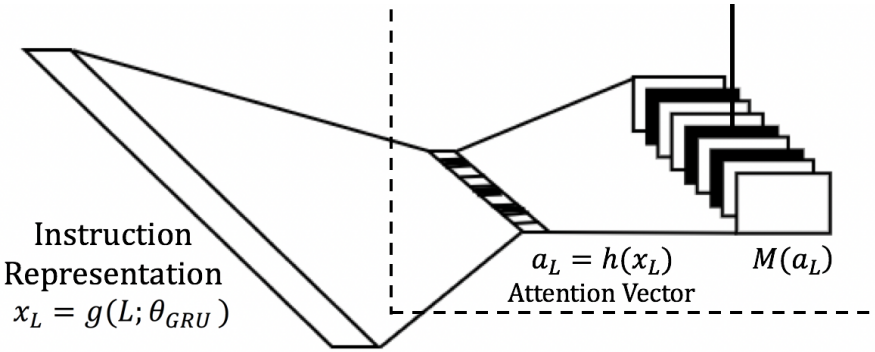


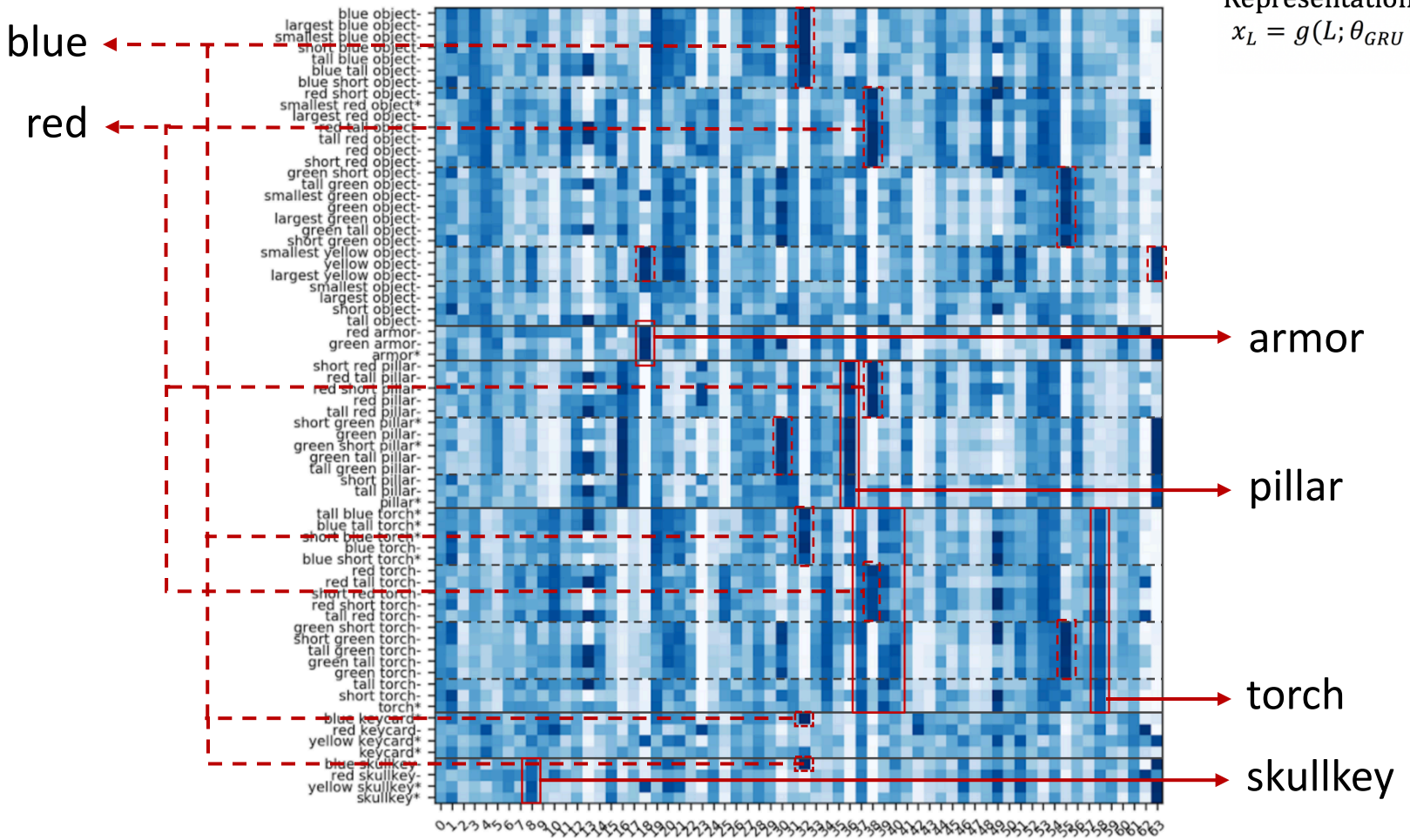
Attention map



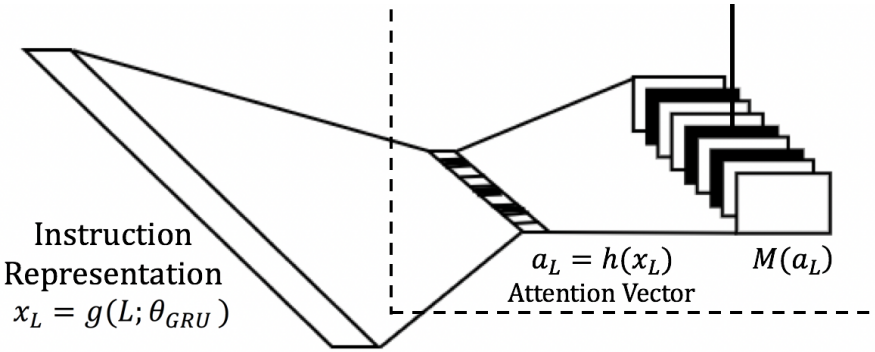


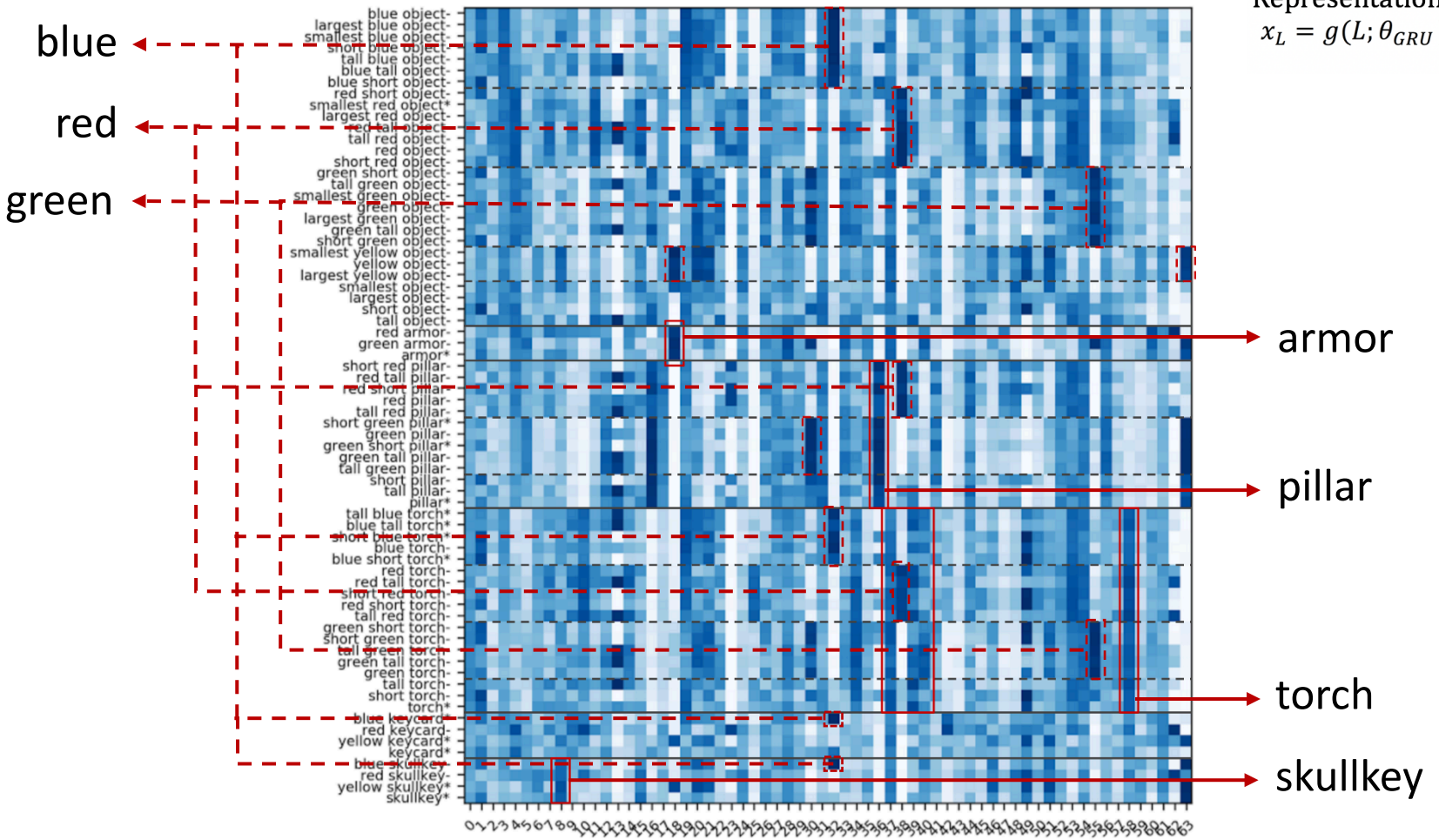
Attention map



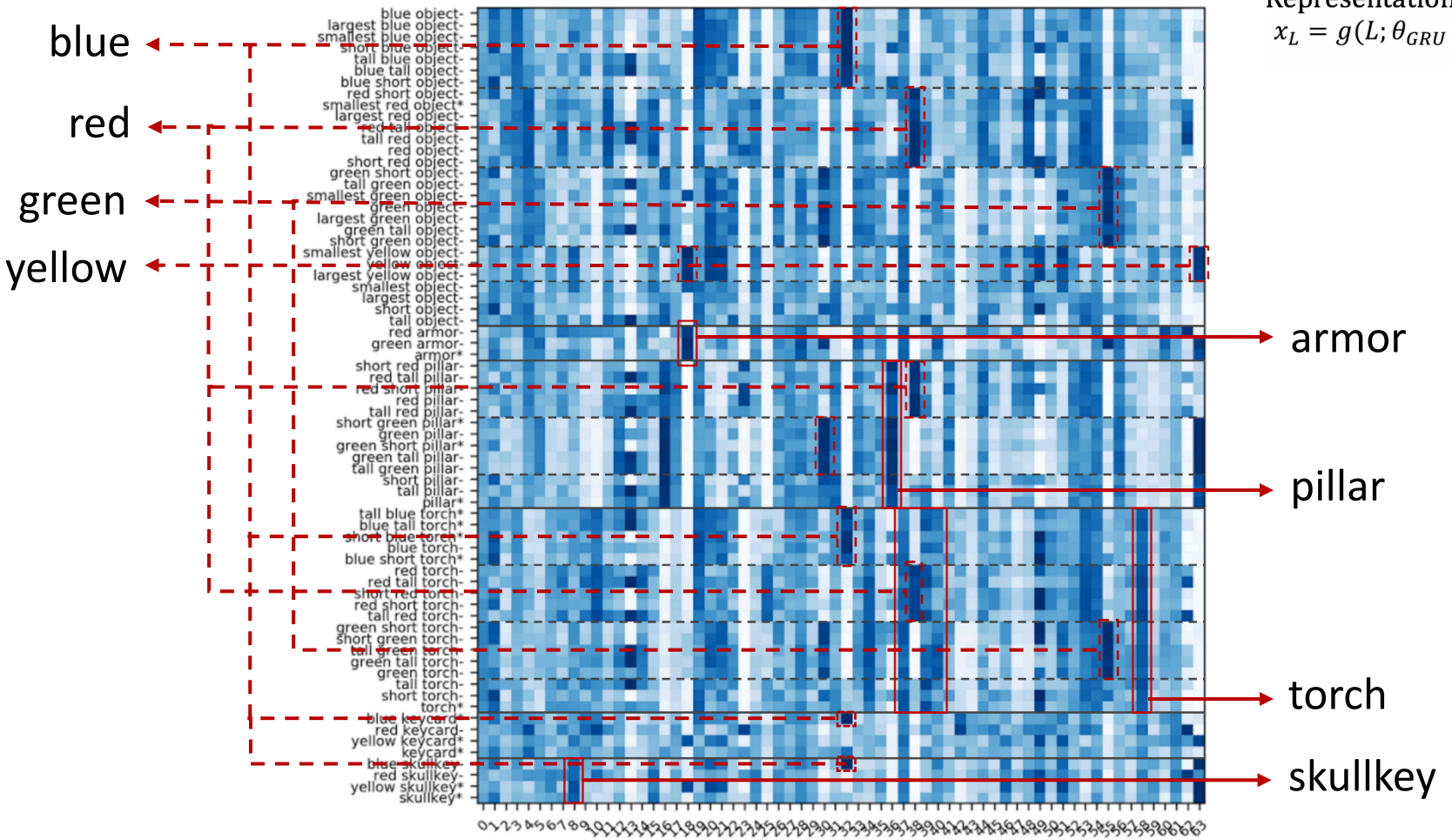


Attention map



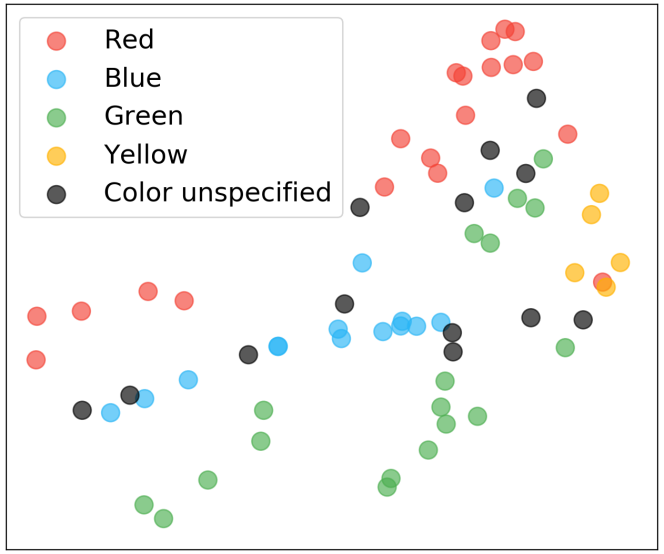


Attention map

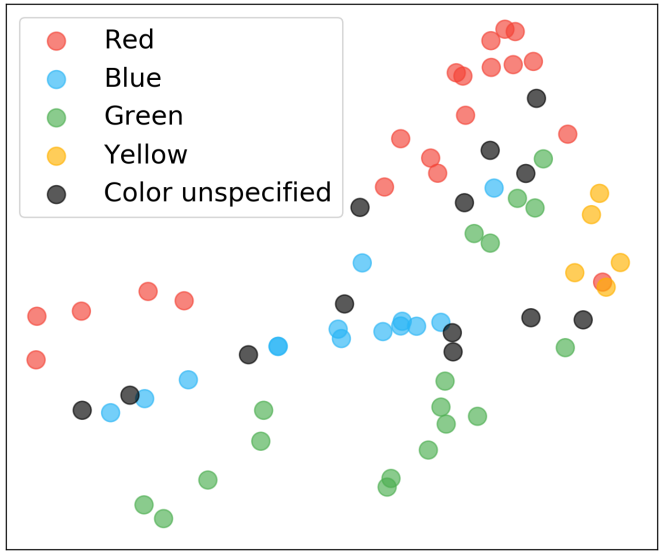


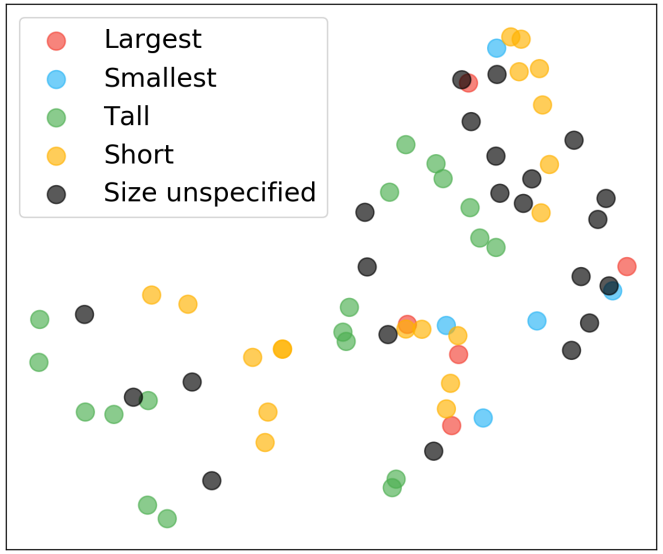
t-SNE Visualizations

t-SNE Visualizations

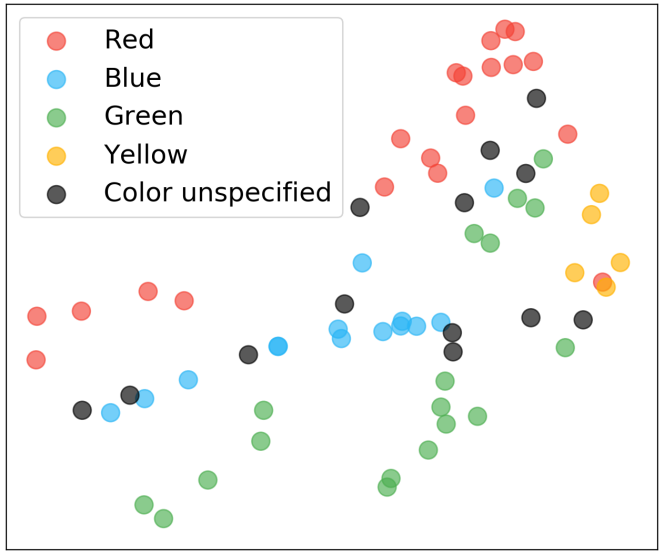


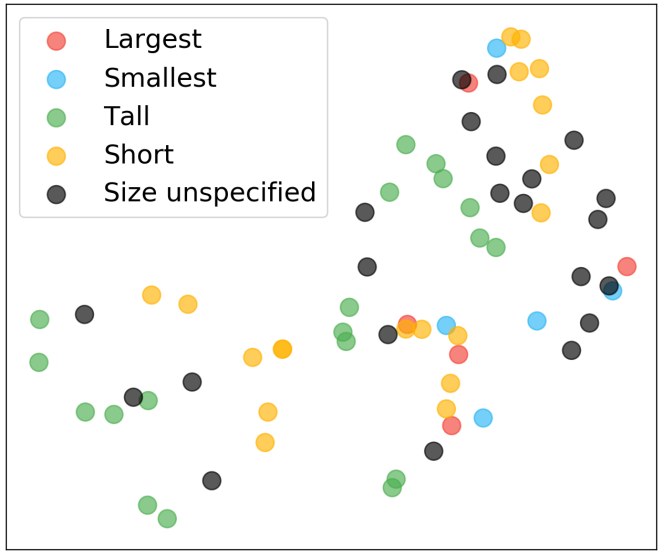
t-SNE Visualizations

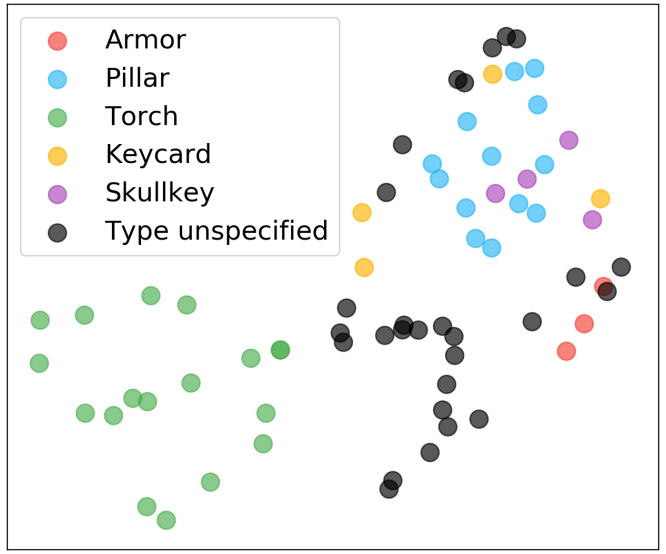




t-SNE Visualizations



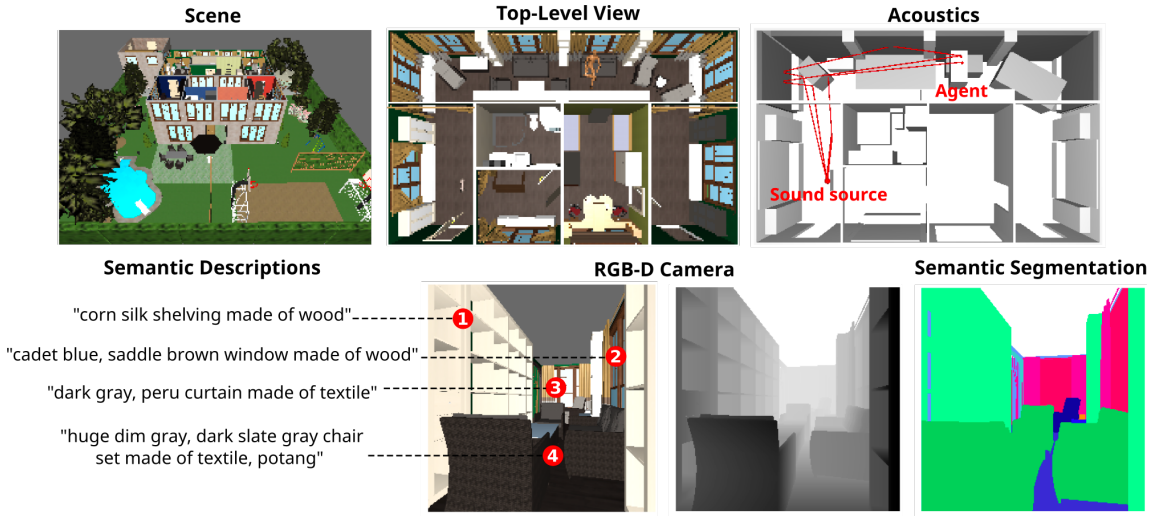




Recent work of language grounding

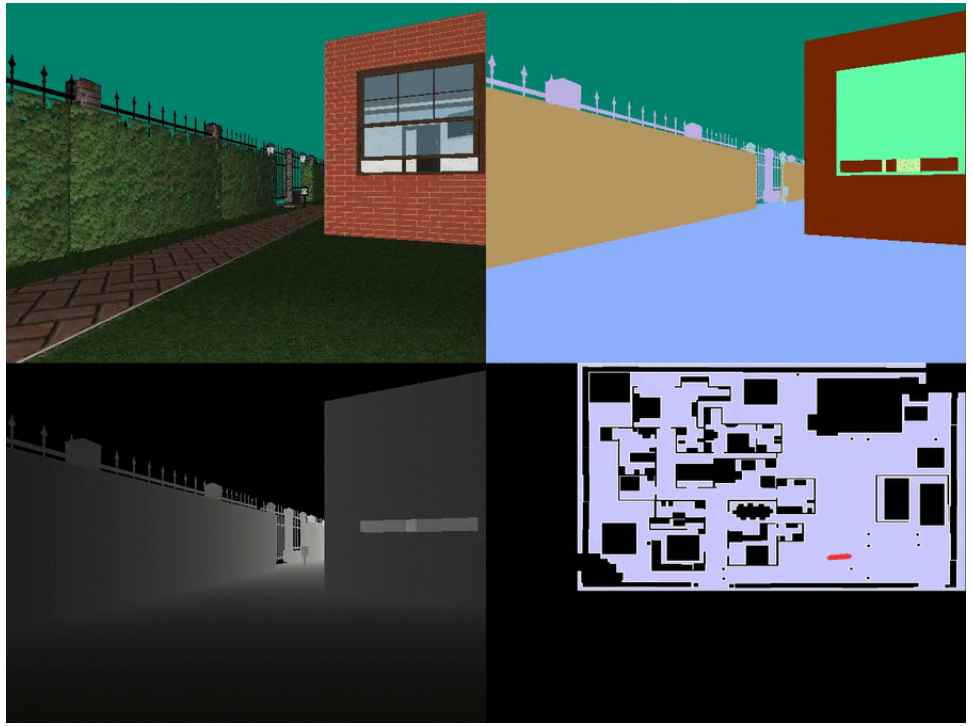
Recent work of language grounding

- Environments
  - Home-platform [MILA, Brodeur et al. 2017]



Recent work of language grounding

- Environments
  - Home-platform [MILA, Brodeur et al. 2017]
  - House3D [FAIR, Wu et al. 2017]



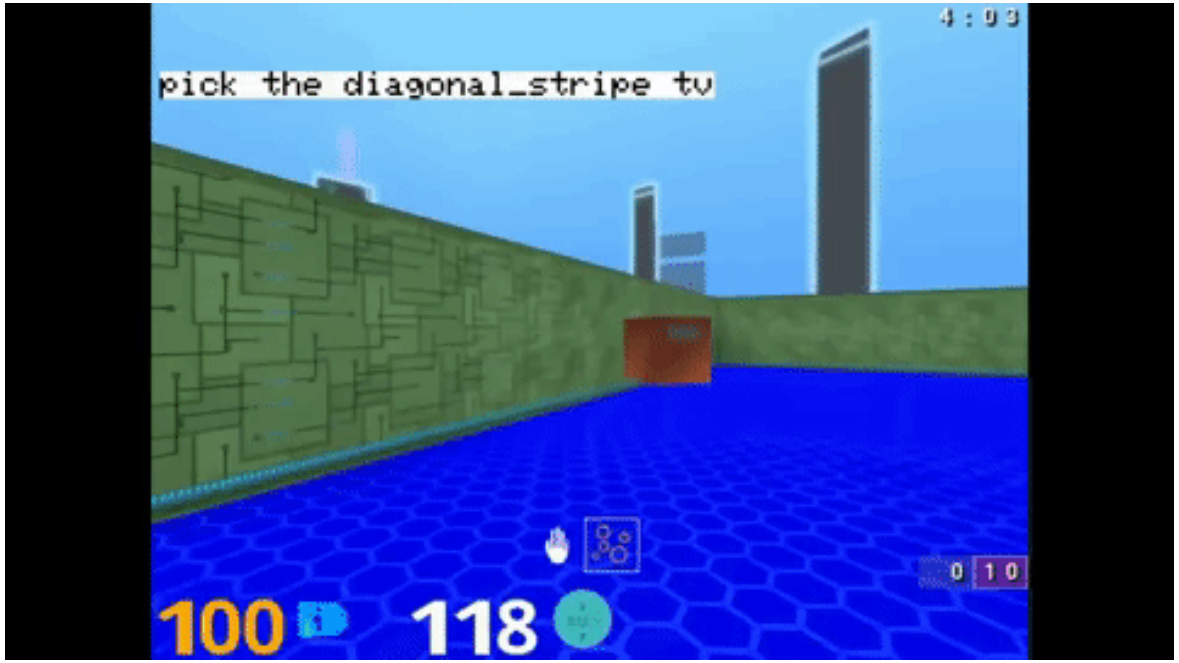
Recent work of language grounding

- Environments
  - Home-platform [MILA, Brodeur et al. 2017]
  - House3D [FAIR, Wu et al. 2017]
  - MINOS [Intel/Princeton, Savva et al. 2017]

Recent work of language grounding

- Environments
  - Home-platform [MILA, Brodeur et al. 2017]
  - House3D [FAIR, Wu et al. 2017]
  - MINOS [Intel/Princeton, Savva et al. 2017]

- Grounded Language Learning [Deepmind, Hermann et al. 2017]



Recent work of language grounding

- Environments
  - Home-platform [MILA, Brodeur et al. 2017]
  - House3D [FAIR, Wu et al. 2017]
  - MINOS [Intel/Princeton, Savva et al. 2017]

- Grounded Language Learning [Deepmind, Hermann et al. 2017]
- Embodied QA [FAIR, Das et al. 2017]

Contributions

- End-to-end trainable architecture that handles raw pixel-based input for task-oriented language grounding in a 3D environment and assumes no prior linguistic or perceptual knowledge.
- Model effective at multi-task as well as zero-shot learning.
- Novel Gated-Attention mechanism for multimodal fusion of representations of verbal and visual modalities.
- New environment for task-oriented language grounding with a rich set of actions, objects and their attributes. The environment provides a first-person view of the world state, and allows for simulating complex scenarios for tasks such as navigation

Gated-Attention Architectures for Task-oriented Language Grounding

Devendra Singh Chaplot, Kanathashree Mysore Satyendra, Rama Kumar Pasumarthi, Dheeraj Rajagopal, Ruslan Salakhutdinov

Code + Environment: https://github.com/devendrachaplot/DeepRL-Grounding

Website: https://sites.google.com/view/gated-attention/home

Thank you