Gated-Attention Architectures for Task-oriented Language Grounding

Devendra Singh Chaplot
Kanthashree Mysore
Rama Kumar Pasumarthi
Dheeraj Rajagopal
Ruslan Salakhutdinov
Task-oriented language grounding

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.

- 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

Agent

Environment

State $S_t$

$I_t$

$L$

Go to the green torch
Experimental setting

Action $\alpha_t$ Environment $S_t$

Agent

$I_t$ State $L$

Go to the green torch
Experimental setting

Agent

Action  $\alpha_t$

Reward $r_t$

+1.0 correct

-0.2 incorrect

0 otherwise

State $S_t$

Environment

Reward

$I_t$

$L$

Go to the green torch

Action

Reward

State
Network overview

Go to the green torch
Natural Language Instruction (L)

Image (I_t)

State Processing Module

GRU Network $g(L; \theta_{GRU})$
$x_L = g(L; \theta_{GRU})$
Instruction Representation

Multimodal Fusion (M)
$M(x_L, x_I)$
State Representation

Policy Learning Module

Policy $\Pi(a|I_t, L)$

Conv Network $f(I_t; \theta_{conv})$
$x_I = f(I_t; \theta_{conv})$
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

Image Representation
\[ x_I = f(I_t; \theta_{\text{conv}}) \]

Instruction Representation
\[ x_L = g(L; \theta_{\text{GRU}}) \]

Gated-Attention Multimodal Fusion Unit
\[ M_{\text{GA}}(x_I, x_L) = M(a_L) \odot x_I \]

To policy learning module

a_L = h(x_L)

Attention Vector

M(a_L)
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

EASY

MEDIUM

[Images of game environments with different levels of difficulty]
Environment difficulty

EASY

MEDIUM

HARD
Results

![Graphs showing accuracy over time for different tasks and models.](image)

Gated-Attention Architectures for Task-oriented Language Grounding (Chaplot, Mysore Sathyendra, Pasumarthi, Rajagopal, Salakhutdinov)
Training Progress

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

Gated-Attention Architectures for Task-oriented Language Grounding (Chaplot, Mysore Sathyendra, Pasumarthi, Rajagopal, Salakhutdinov)
Attention map

Instruction Representation
\[ x_L = g(l; \theta_{GRU}) \]

Attention Vector
\[ a_L = h(x_L) \]

- blue
- armor
- pillar
- torch
- skullkey

Gated-Attention Architectures for Task-oriented Language Grounding (Chaplot, Mysore Sathyendra, Pasumarthi, Rajagopal, Salakhutdinov)

Carnegie Mellon University
Attention map

Gated-Attention Architectures for Task-oriented Language Grounding (Chaplot, Mysore Sathyendra, Pasumarthi, Rajagopal, Salakhutdinov)
Gated-Attention Architectures for Task-oriented Language Grounding (Chaplot, Mysore Sathyendra, Pasumarthi, Rajagopal, Salakhutdinov)
Attention map

Gated-Attention Architectures for Task-oriented Language Grounding (Chaplot, Mysore Sathyendra, Pasumarthi, Rajagopal, Salakhutdinov)
t-SNE Visualizations
t-SNE Visualizations

Red
Blue
Green
Yellow
Color unspecified
t-SNE Visualizations

- Red
- Blue
- Green
- Yellow
- Color unspecified

- Largest
- Smallest
- Tall
- Short
- Size unspecified
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

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

Thank you