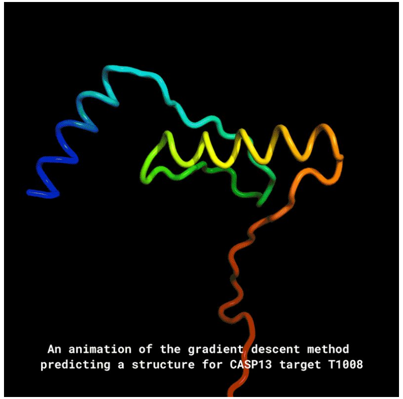
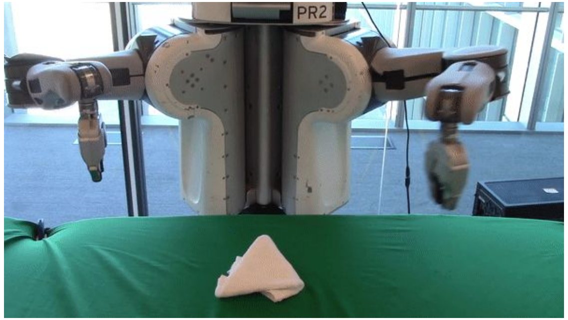
Evaluating the Robustness of Natural Language Reward Shaping Models to Spatial Relations

Antony Yun

The University of Texas at Austin
Computer Science

Successes of Reinforcement Learning





https://deepmind.com/blog/article/alphazero-shedding-new-light-grand-games-chess-shogi-and-go
https://bair.berkeley.edu/blog/2020/05/05/fabrics/
https://deepmind.com/blog/article/AlphaFold-Using-AI-for-scientific-discovery

My Work

- Construct a challenge dataset in the Meta-World reward shaping domain that contains spatially relational language
- Evaluate robustness of existing natural language reward shaping models

Outline

- Background on Deep Learning, Reinforcement Learning
- Natural language reward shaping
- Our Dataset
- Results

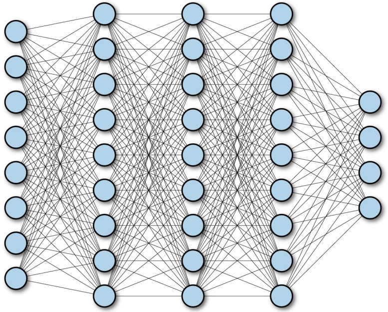
Background: Neural Networks

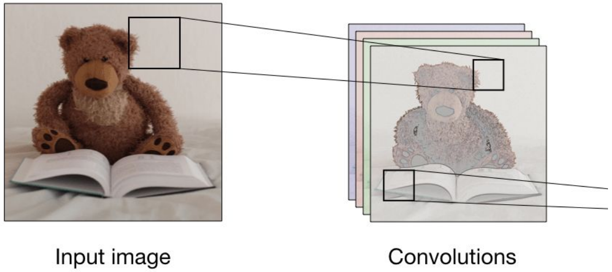
- Function approximators
- Trained with gradient descent

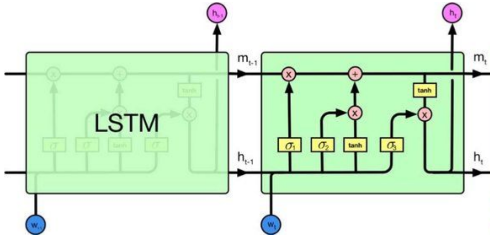
f() = [0.12, 0.05, ...]

https://github.com/caoscott/SReC

Background: Neural Networks



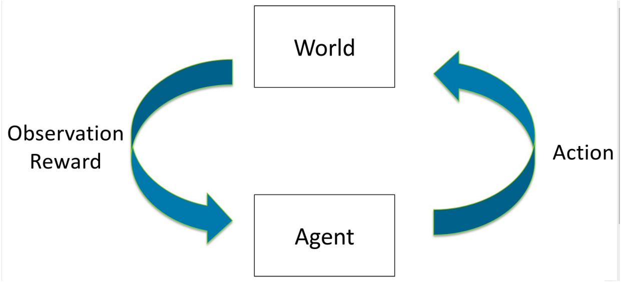


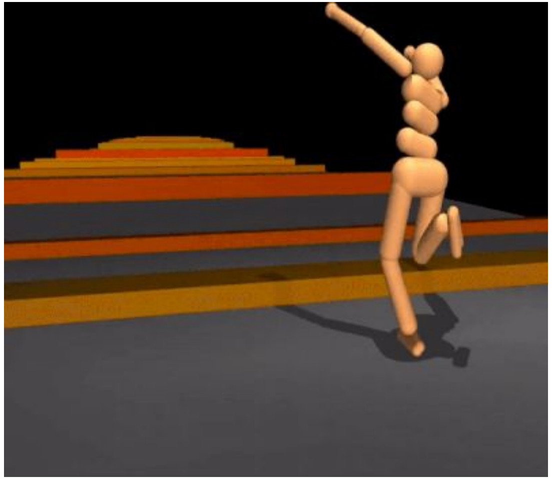


https://www.oreilly.com/library/view/tensorflow-for-deep/9781491980446/ch04.html
https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-convolutional-neural-networks
https://www.researchgate.net/figure/Illustration-of-LSTM-block-s-is-the-sigmoid-function-which-play-the-role-of-gates-during_fig2_322477802

Background: Reinforcement Learning

- Learn a policy by interacting with the environment
- Optimize cumulative discounted reward





https://deepmind.com/blog/article/producing-flexible-behaviours-simulated-environments
http://web.stanford.edu/class/cs234/index.html

Background: Markov Decision Process (MDP)

$$M = \langle S, A, T, R, \gamma \rangle$$

- S = states
- A = actions
- T = transition function
- R = reward
- γ = discount factor

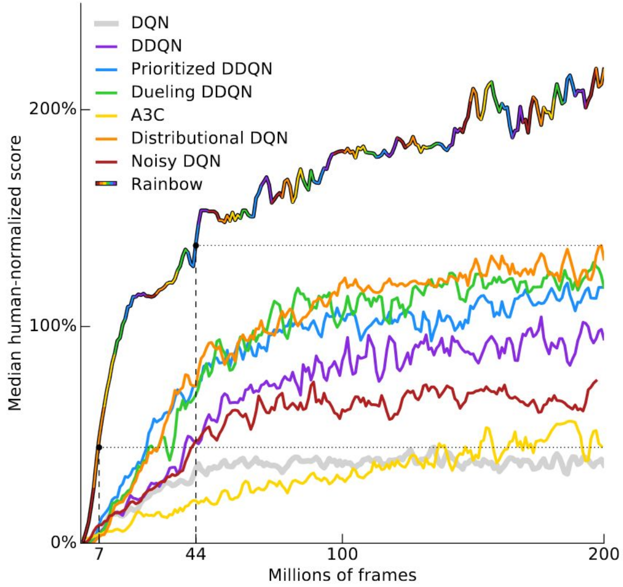
Background: Policy Based RL

- Parameterized policy
  $$\pi_\theta(s, a) = P[a \mid s, \theta]$$
- Want optimal policy that maximizes expected reward
- Learned by gradient descent on final reward
- We use Proximal Policy Optimization (PPO)

[Schulman et al, 2017]

Challenges with RL

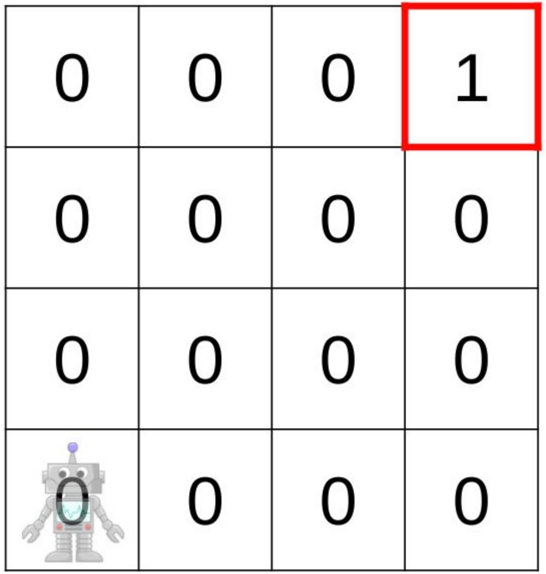
- Sample inefficient



https://www.alexirpan.com/2018/02/14/rl-hard.html

Challenges with RL

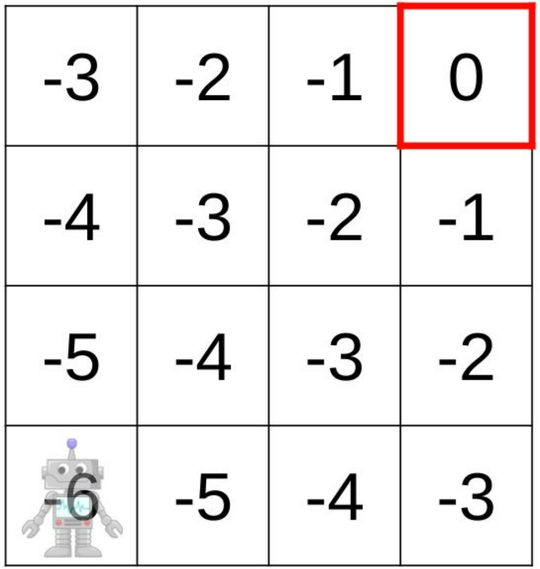
- Sample inefficient
- Good reward functions are hard to find
    - Sparse: easy to design



https://www.alexirpan.com/2018/02/14/rl-hard.html

Challenges with RL

- Sample inefficient
- Good reward functions are hard to find
    - Sparse: easy to design
    - Dense: easy to learn



https://www.alexirpan.com/2018/02/14/rl-hard.html

Background: Reward Shaping

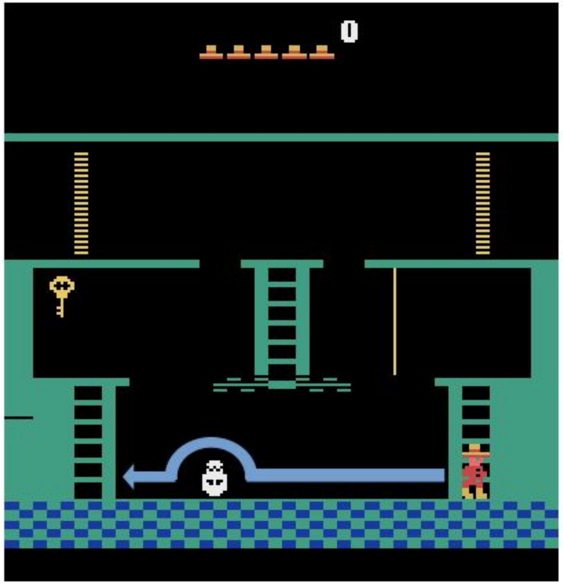
- Provide additional *potential* reward
- Does **not** change the optimal policy

$$R' = R + F$$
$$F(s, a, s') = \gamma\phi(s') - \phi(s)$$

[Ng et al, 1999]

Prior Work: LEARN

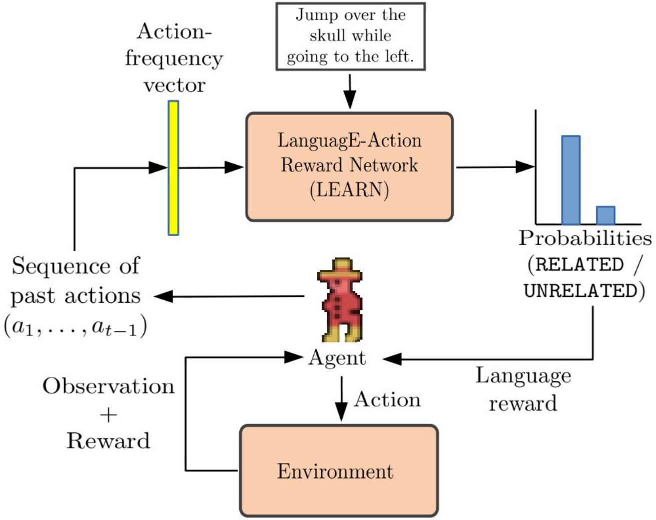
- Language-based shaping rewards for Montezuma's Revenge
- Non-experts can express intent
- 60% improvement over baseline

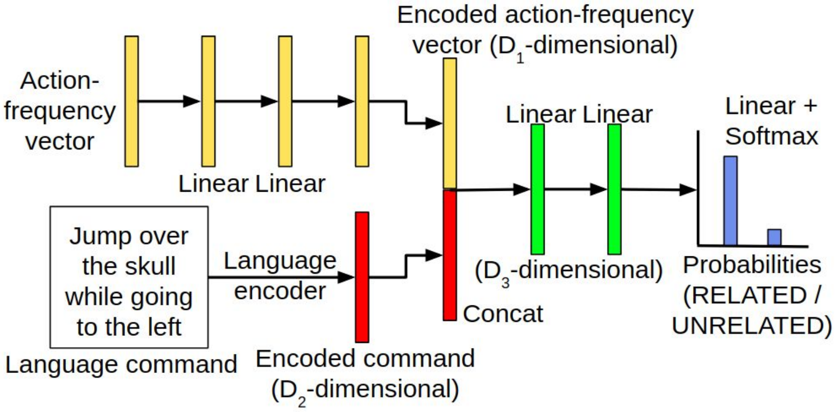


"Jump over the skull while going to the left"

[Goyal et al, 2019]

Prior Work: LEARN





[Goyal et al, 2019]

Meta-World

- Object manipulation domain involving grasping, placing, and pushing
- Continuous action space, multimodal data, complex goal states

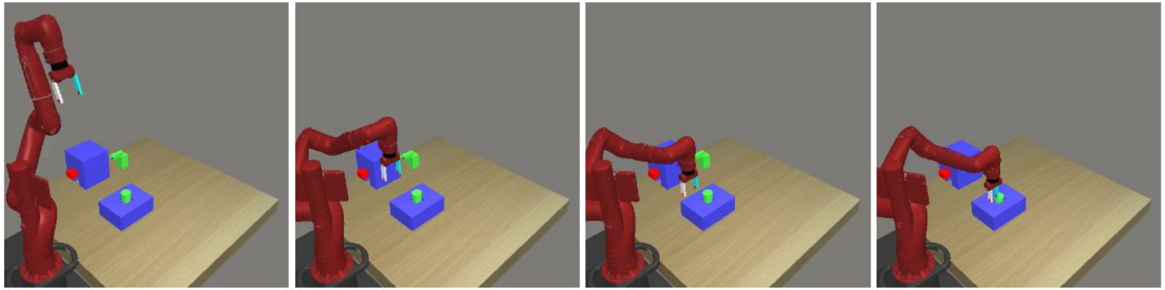


Figure 1. A simulated robot completing a task in the Meta-World domain

[Yu et al, 2019]

Dense Rewards in Meta-World

$$
\begin{aligned}
R &= R_{\text{reach}} + R_{\text{grasp}} + R_{\text{place}} \\
&= \underbrace{-\|h - o\|_2}_{R_{\text{reach}}} + \underbrace{\mathbb{I}_{\|h-o\|_2<\epsilon} \cdot c_1 \cdot \min\{o_z, z_{\text{target}}\}}_{R_{\text{grasp}}} + \underbrace{\mathbb{I}_{|o_z - z_{\text{target}}|<\epsilon} \cdot c_2 \cdot \exp\{\|o - g\|_2^2/c_3\}}_{R_{\text{place}}}
\end{aligned}
$$

$$
\begin{aligned}
R &= R_{\text{reach}} + R_{\text{push}} \\
&= \underbrace{-\|h - o\|_2}_{R_{\text{reach}}} + \underbrace{\mathbb{I}_{\|h-o\|_2<\epsilon} \cdot c_2 \cdot \exp\{\|o - g\|_2^2/c_3\}}_{R_{\text{push}}}
\end{aligned}
$$

[Yu et al, 2019]

Dense Rewards in Meta-World

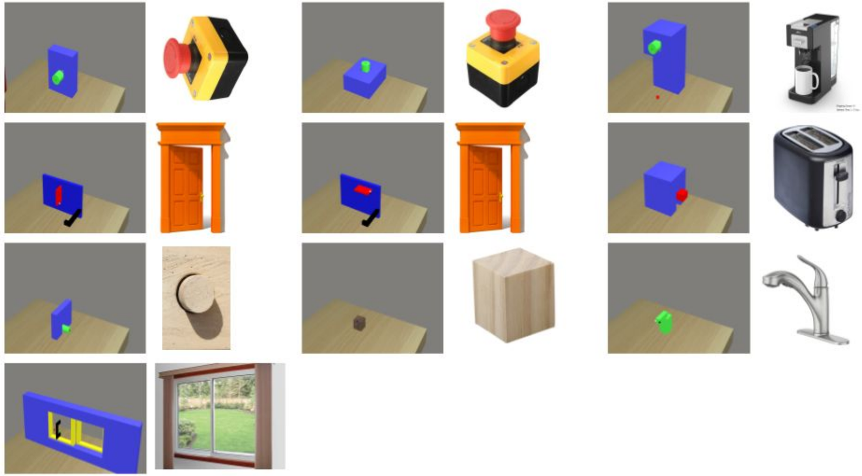
Task	Reward
turn on faucet	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
sweep	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
pick out of hole	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 100 \cdot \min\{o_z, z_{\text{target}}\} + \mathbb{I}_{\lvert o_z-z_{\text{target}}\rvert<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
turn off faucet	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
push with stick	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 100 \cdot \min\{o_z, z_{\text{target}}\} + \mathbb{I}_{\lvert o_z-z_{\text{target}}\rvert<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
get coffee	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
pull handle side	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
basketball	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 100 \cdot \min\{o_z, z_{\text{target}}\} + \mathbb{I}_{\lvert o_z-z_{\text{target}}\rvert<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
pull with stick	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 100 \cdot \min\{o_z, z_{\text{target}}\} + \mathbb{I}_{\lvert o_z-z_{\text{target}}\rvert<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
sweep into hole	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
disassemble nut	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 100 \cdot \min\{o_z, z_{\text{target}}\} + \mathbb{I}_{\lvert o_z-z_{\text{target}}\rvert<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
assemble nut	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 100 \cdot \min\{o_z, z_{\text{target}}\} + \mathbb{I}_{\lvert o_z-z_{\text{target}}\rvert<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
place onto shelf	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 100 \cdot \min\{o_z, z_{\text{target}}\} + \mathbb{I}_{\lvert o_z-z_{\text{target}}\rvert<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
push mug	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
press handle side	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
hammer	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 100 \cdot \min\{o_z, z_{\text{target}}\} + \mathbb{I}_{\lvert o_z-z_{\text{target}}\rvert<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
slide plate	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
slide plate side	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
press button wall	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
press handle	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
pull handle	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
soccer	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
retrieve plate side	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
retrieve plate	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
close drawer	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
reach	$1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
press button top wall	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
reach with wall	$1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
insert peg side	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 100 \cdot \min\{o_z, z_{\text{target}}\} + \mathbb{I}_{\lvert o_z-z_{\text{target}}\rvert<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
push	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
push with wall	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
pick&place w/ wall	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 100 \cdot \min\{o_z, z_{\text{target}}\} + \mathbb{I}_{\lvert o_z-z_{\text{target}}\rvert<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
press button	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
press button top	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
pick&place	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 100 \cdot \min\{o_z, z_{\text{target}}\} + \mathbb{I}_{\lvert o_z-z_{\text{target}}\rvert<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
pull	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
pull mug	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
unplug peg	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 100 \cdot \min\{o_z, z_{\text{target}}\} + \mathbb{I}_{\lvert o_z-z_{\text{target}}\rvert<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
turn dial	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
pull lever	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
close window	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
open window	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
open door	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
close door	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
open drawer	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
insert hand	$1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
close box	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 100 \cdot \min\{o_z, z_{\text{target}}\} + \mathbb{I}_{\lvert o_z-z_{\text{target}}\rvert<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
lock door	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
unlock door	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$
pick bin	$-\lVert h-o\rVert_2 + \mathbb{I}_{\lVert h-o\rVert_2<0.05} \cdot 100 \cdot \min\{o_z, z_{\text{target}}\} + \mathbb{I}_{\lvert o_z-z_{\text{target}}\rvert<0.05} \cdot 1000 \cdot \exp\{\lVert h-g\rVert_2^2/0.01\}$

Table 3: A list of reward functions used for each of the Meta-World tasks.

[Yu et al, 2019]

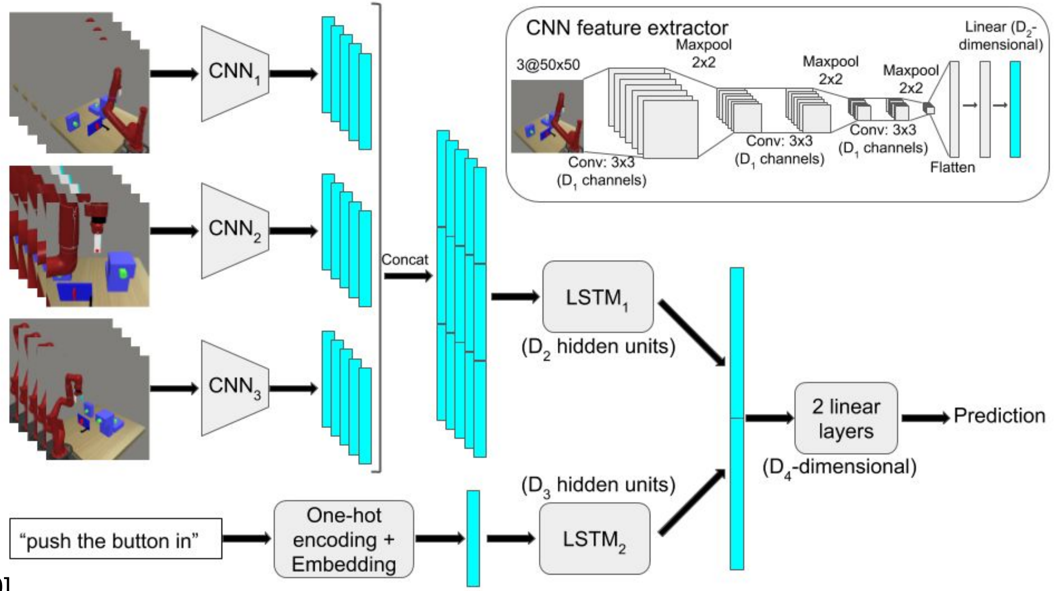
Pix2R Dataset

- 13 Meta-World tasks, 9 objects
- 100 scenarios per task
- Videos generated using PPO on dense rewards
- 520 human-annotated descriptions from Amazon Mechanical Turk
- Use video trajectories + descriptions to approximate dense reward



[Goyal et al, 2020]

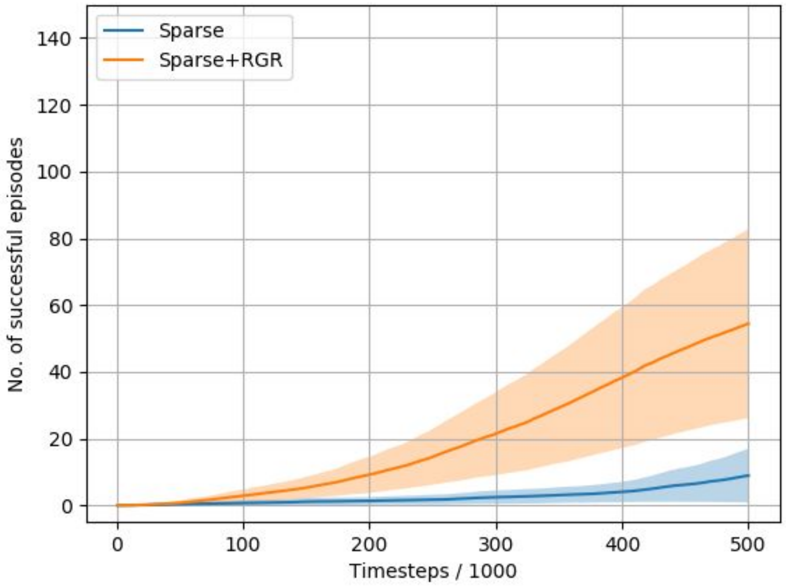
Pix2R Architecture



[Goyal et al, 2020]

Pix2R Results

- Adding shaping reward speeds up policy learning sparse rewards
- Sparse + Shaping rewards perform comparably to Dense rewards



[Goyal et al, 2020]

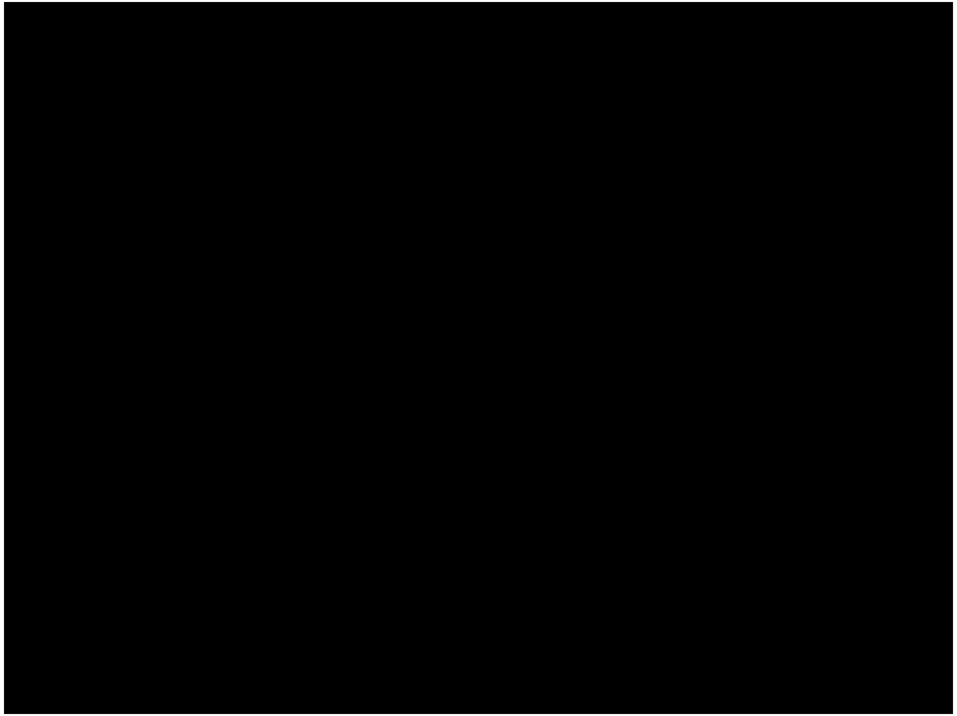
Extending Pix2R Dataset

- Each scenario has only one instance of each object
- Descriptions use simplistic language

- ***Goal:*** construct a dataset containing relational language
- Probe whether model is learning multimodal semantic relationships or just identification
- Motivate development of more robust models

Relational Data

- "Turn on the coffee machine on the left"
- "Press the coffee maker furthest from the button"



Video Generation

- Target object + duplicate object + distractors
- Train PPO with dense reward until success
- 6 tasks (button_top, button_side, coffee_button, handle_press_top, door_lock, door_unlock)
- 5 scenarios per task
- 30 total scenarios

Collecting Natural Language Descriptions

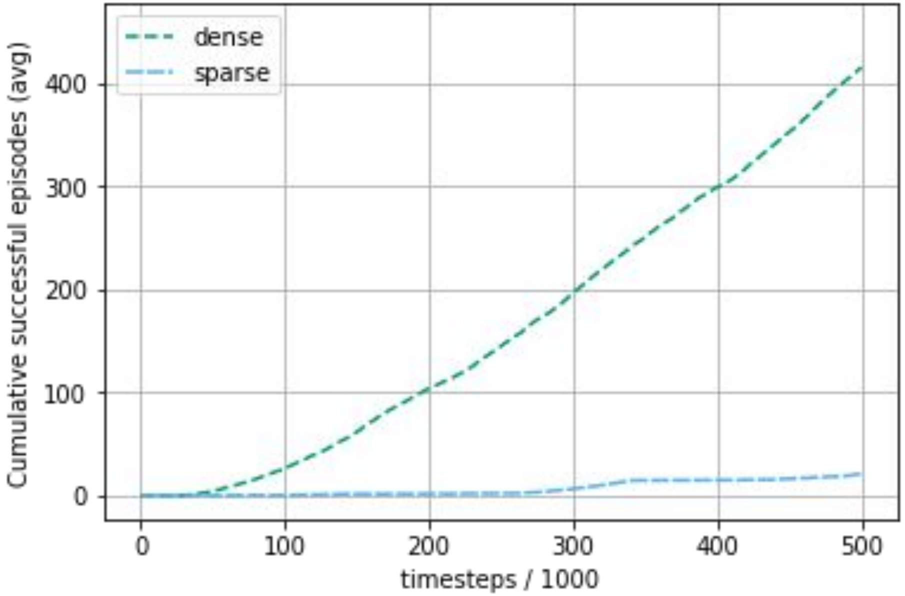
- Amazon Mechanical Turk
- *‘Please ensure that the instruction you provide uniquely identifies the correct object, for example, by describing it with respect to other objects around it.’*
- At least 3 descriptions per scenario (131 total)
- Manually create negative examples

Evaluation

- Can Pix2R encode relations between objects?
- Evaluate on test split of new data
- 6 scenarios, 3 descriptions, 5 runs each → 90 runs

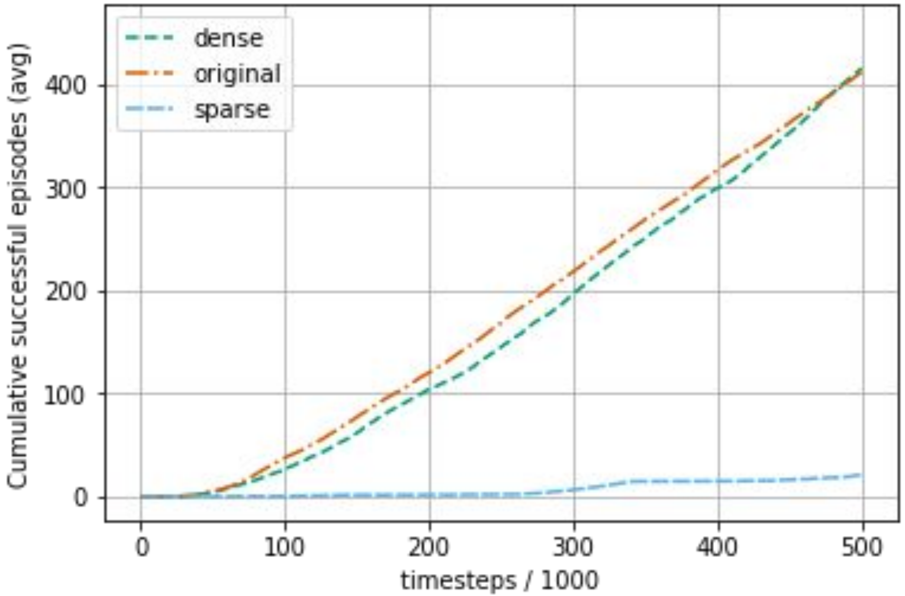
Baselines and Models

- Sparse: PPO with binary reward
- Dense: PPO with expert Meta-World reward



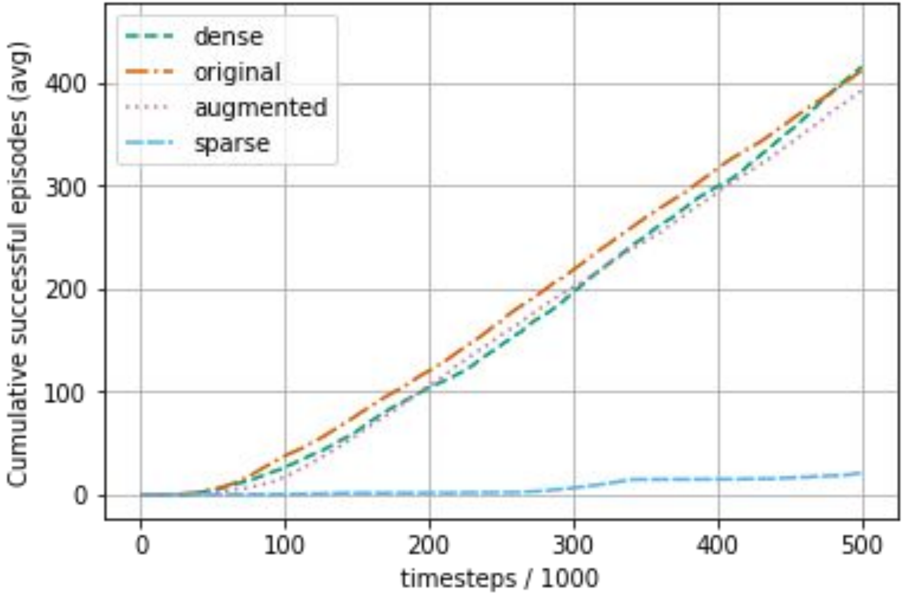
Baselines and Models

- Sparse: PPO with binary reward
- Dense: PPO with expert Meta-World reward
- Original: PPO shaped by Pix2R trained on original dataset



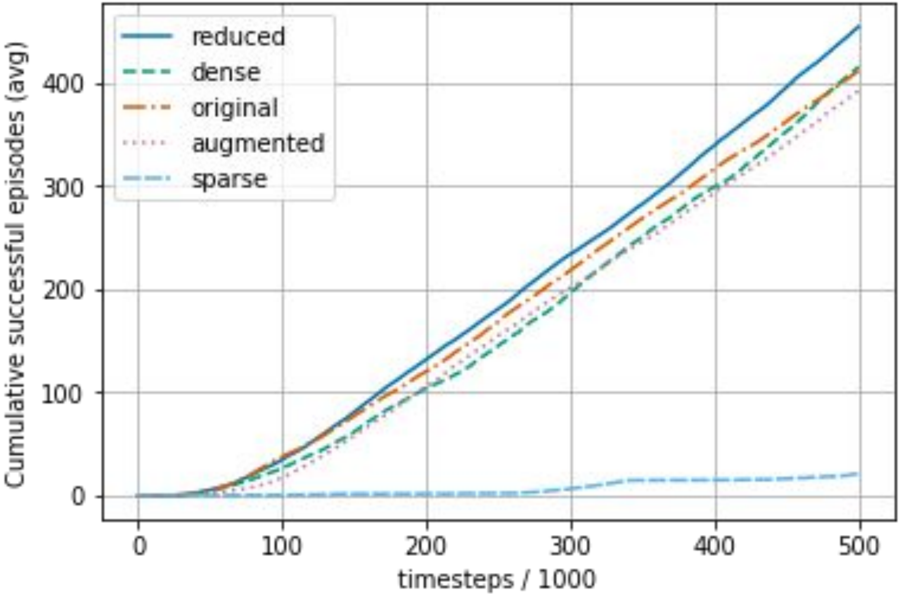
Baselines and Models

- Sparse: PPO with binary reward
- Dense: PPO with expert Meta-World reward
- Original: PPO shaped by Pix2R trained on original dataset
- Augmented: PPO shaped by Pix2R trained on combined dataset



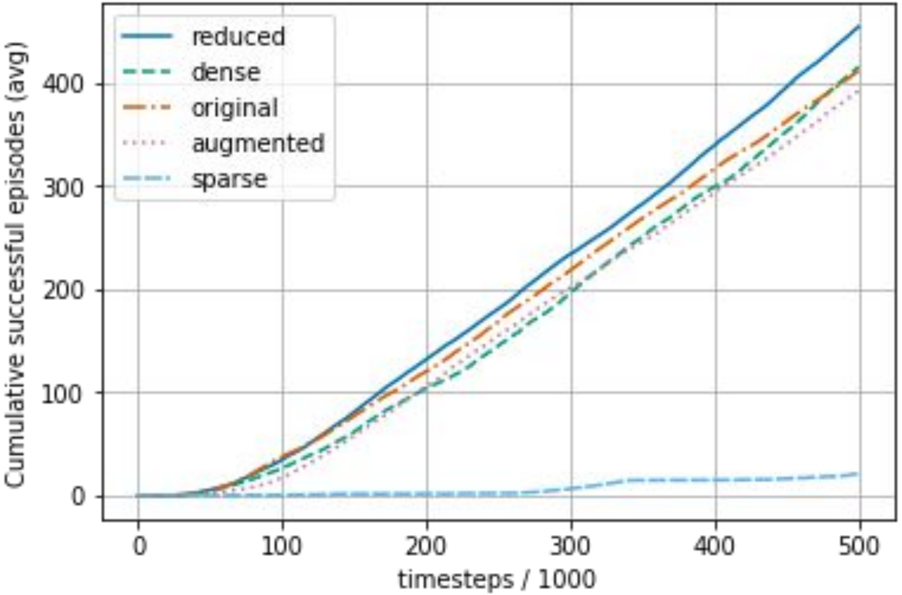
Baselines and Models

- Sparse: PPO with binary reward
- Dense: PPO with expert Meta-World reward
- Original: PPO shaped by Pix2R trained on original dataset
- Augmented: PPO shaped by Pix2R trained on combined dataset
- Reduced: PPO shaped by Pix2R trained on original dataset, excluding relational descriptions



Results

- All agents perform comparably, except sparse
- Reduced even performs slightly better
- Scenarios could be too simple
- Inconclusive, further experimentation needed



Conclusion

- Pix2R is robust to our specific challenge dataset
- No immediately obvious shortcomings
- Room for further probing through challenge datasets

Future Work

- Improving our existing challenge dataset
  - Refine environment generation to create more challenging scenarios
  - Multi-stage AMT pipeline for higher quality annotations
- Other challenge datasets
  - Can construct targeted, "adversarial" examples for any ML task

Acknowledgements

Dr. Ray Mooney

Prasoon Goyal