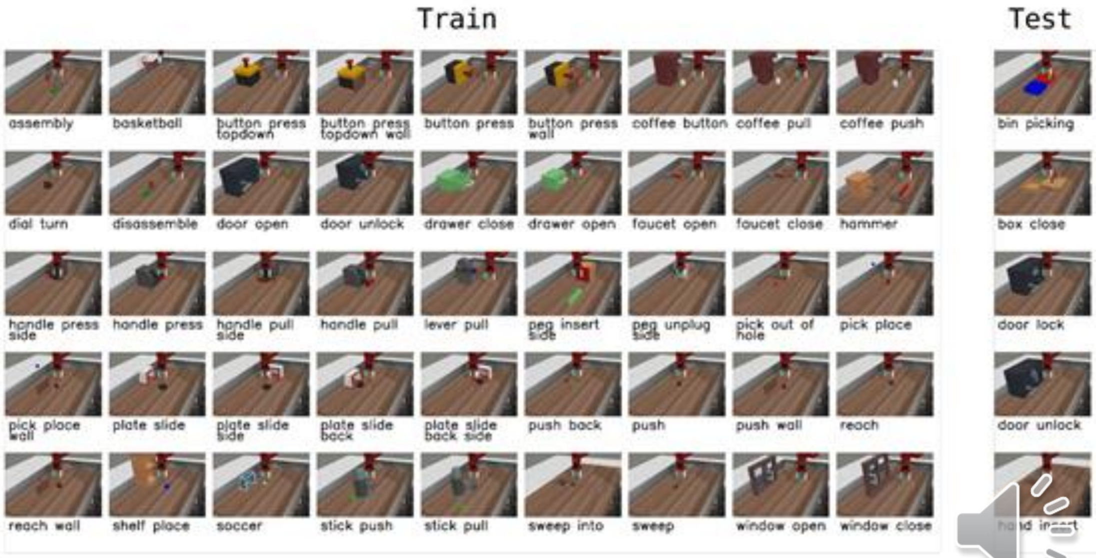
Meta-World: A Benchmark and Evaluation for Multi-Task and Meta Reinforcement Learning

Presented by Rowan Dempster



UNIVERSITY OF WATERLOO | FACULTY OF ENGINEERING

Intro and Agenda

- Meta-World is benchmark for multi-task and meta-RL algorithms based on robotic arm manipulation tasks. Meta-World mitigates issues present in existing evaluation methods such as narrow and dissimilar task distributions.
- Agenda:
  - Background: What is {multi-task, meta} RL? How are they benchmarked? 🤔
  - Current Solutions: Negative transfer Atari games, narrow parametric distributions 👎
  - Proposed Solution: Non-parametric distribution with positive transfer 👍
  - Empirical Evaluation: Existing {multi-task, meta} RL algorithms perform poorly 📉
  - Conclusions: New benchmark opens up opportunities for further work 📈

Background - Types of RL

Multi-Task RL

Learn a single, task conditioned policy $\pi(a|s, z)$ (where z is one-hot task ID) which maximizes expected cumulative rewards under task distribution $p(\mathcal{T})$: $\mathbb{E}_{\mathcal{T}\sim p(\mathcal{T})}[\mathbb{E}_{\pi}[\sum_{t=0}^{T} \gamma^t R_t(s_t, a_t)]]$

No separate test set of tasks, evaluation via average performance on training tasks

Meta-RL

Given a set of training tasks, learn a policy $\pi(a|s)$ that can "quickly" learn held-out test tasks

Requires the training task distribution to be sufficiently broad to share structure with held-out testing tasks

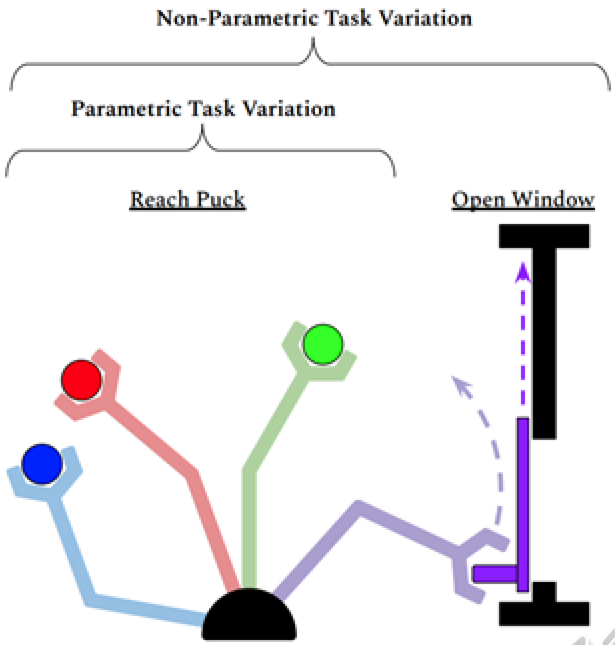
Background - Task Distributions

<u>Parametric Task Distributions</u>

- Variation in tasks is described by a continuous parameter, e.g. position, velocity, etc...
- E.g. the position of the puck in the *Reach Puck* task

<u>Non-Parametric Task Distributions</u>

- Variation in tasks cannot be described simply by continuous variables, the system is structurally unique, e.g. different joints, levers, etc...
- E.g. the "slide" motion required by the *Open Window* task is structurally different from the "grab" motion in *Reach Puck*



Current Benchmarking Approaches

Atari Games

- Significant differences in visual appearance, control schemes, etc…
- Challenging to leverage efficiency gains between games via learning shared structure

- In fact, proposed methods have observed large *negative* transfer learning between games (improved performance at one decreases performance in another)

Parametric Distribution

- Many meta-RL methods are evaluated using narrow parametric distributions, e.g. in legged robots holding-out certain running directions
- Far-cry from the “domain adaptation” promise of meta-RL that would yield real-world benefits

Proposed Solution

- Meta-World is a suite of 50 non-parametric robotic manipulation tasks
  - Advantage over Atari Suite: Shared structure of robotic manipulation tasks means positive transfer learning is possible
  - Advantage over Parametric Suite: Success in this large non-parametric domain does bring the community closer to solving general robotic intelligence that can quickly achieve never-before seen tasks in the real world

Proposed Solution

- What is shared structure? How are we sure that Meta-World has it?
  - Shared action space: 3D end-effector positions of Sawyer arm

Proposed Solution

- What is shared structure? How are we sure that Meta-World has it?
  - Shared action space: 3D end-effector positions of Sawyer arm
  - Shared workspace: Manipulation of variable object with a variable goal position, or manipulation of two variable objects with a fixed goal position. This allows for a uniformly 9-dimensional observation space across all tasks (3D positions of each variable object in workspace)

Proposed Solution

- What is shared structure? How are we sure that Meta-World has it?
  - Shared action space: 3D end-effector positions of Sawyer arm
  - Shared workspace: Manipulation of variable object with a variable goal position, or manipulation of two variable objects with a fixed goal position. This allows for a uniformly 9-dimensional observation space across all tasks (3D positions of each variable object in workspace)
  - Shared reward structure: Rewards are well shaped (each task in individually solvable), and exhibit similar structure and scale

Proposed Solution

- What is shared structure? How are we sure that Meta-World has it?
  - Shared action space: 3D end-effector positions of Sawyer arm
  - Shared workspace: Manipulation of variable object with a variable goal position, or manipulation of two variable objects with a fixed goal position. This allows for a uniformly 9-dimensional observation space across all tasks (3D positions of each variable object in workspace)
  - Shared reward structure: Rewards are well shaped (each task in individually solvable), and exhibit similar structure and scale

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

Reach Puck Reward

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

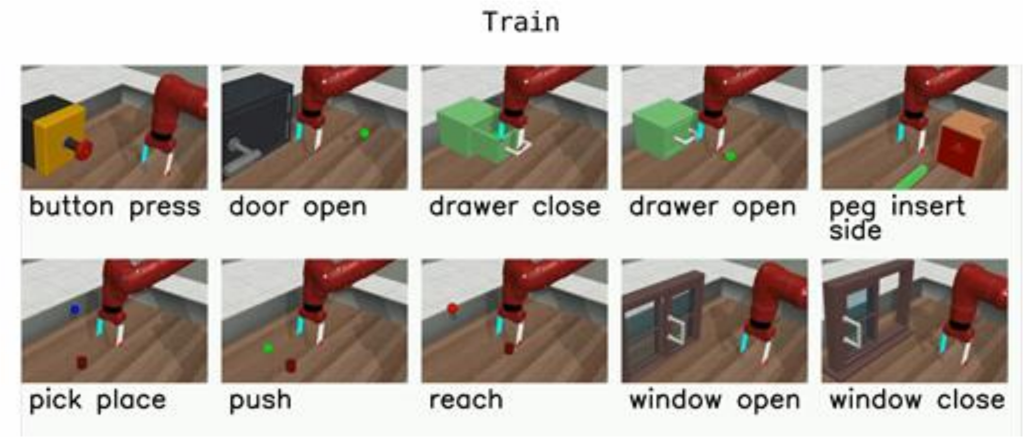
Open Window Reward

Proposed Solution

- Meta-World is a suite of 50 non-parametric robotic manipulation tasks
- Evaluation levels:
  - Meta-Learning 1 (ML1): Few-shot adaptation to goal variation within one task

Proposed Solution

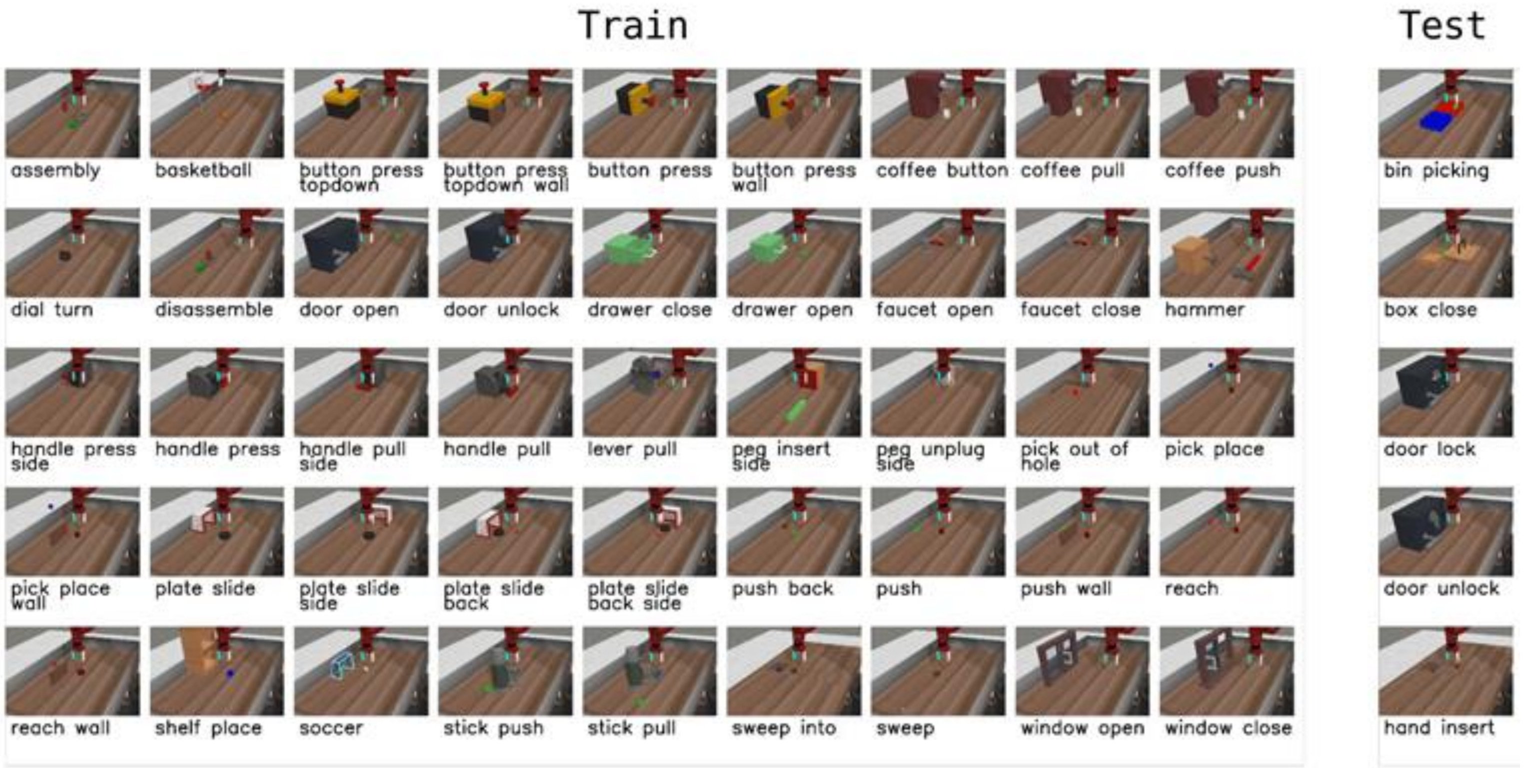
- Meta-World is a suite of 50 non-parametric robotic manipulation tasks
- Evaluation levels:
  - Meta-Learning 1 (ML1): Few-shot adaptation to goal variation within one task
  - Multi-Task 10, Multi-Task 50 (MT10, MT50): Learning one multi-task policy that generalizes to 10 and 50 training tasks



Proposed Solution

- Meta-World is a suite of 50 non-parametric robotic manipulation tasks

- Evaluation levels:
    - Meta-Learning 1 (ML1): Few-shot adaptation to goal variation within one task
    - Multi-Task 10, Multi-Task 50 (MT10, MT50): Learning one multi-task policy that generalizes to 10 and 50 training tasks
    - Meta-Learning 10, Meta-Learning 45 (ML10, ML45): Few-shot adaptation to new test tasks with 10 and 45 meta-training tasks


Train
Test
assembly
basketball
button press topdown
button press topdown wall
button press
button press wall
coffee button
coffee pull
coffee push
bin picking
dial turn
disassemble
door open
door unlock
drawer close
drawer open
faucet open
faucet close
hammer
box close
handle press side
handle press
handle pull side
handle pull
lever pull
peg insert side
peg unplug side
pick out of hole
pick place
door lock
pick place wall
plate slide
plate slide side
plate slide back
plate slide back side
push back
push
push wall
reach
door unlock
reach wall
shelf place
soccer
stick push
stick pull
sweep into
sweep
window open
window close
hand insert

Empirical Evaluation - Candidate Algorithms

Multi-Task Candidates

Multi-task proximal policy optimization (PPO)

Multi-task trust region policy optimization (TRPO)

Multi-task soft actor-critic (SAC)

Task embeddings (TE)

Meta-RL Candidates

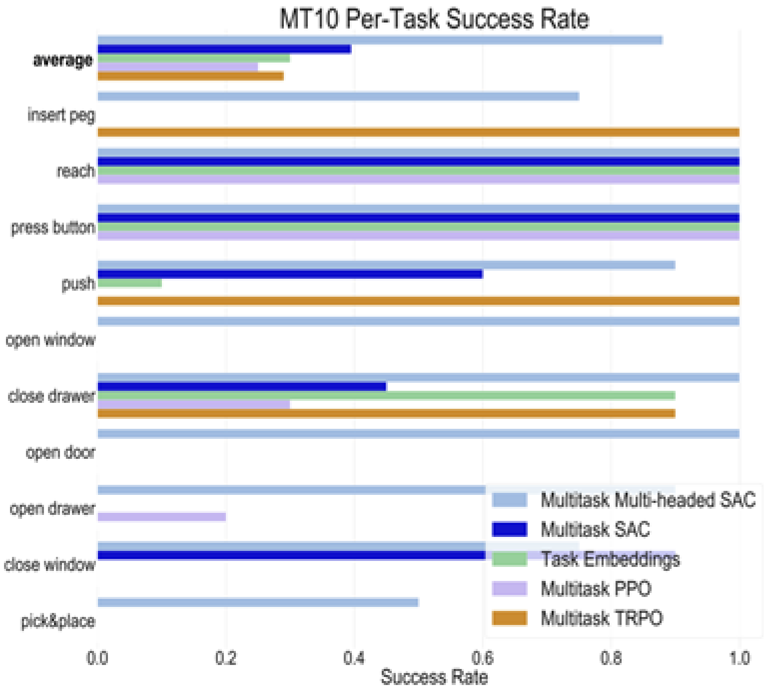
RL^2

Model-agnostic metalearning (MAML)

Probabilistic embeddings for actor-critic RL (PEARL)

Empirical Evaluation - MT10 / MT50 Results

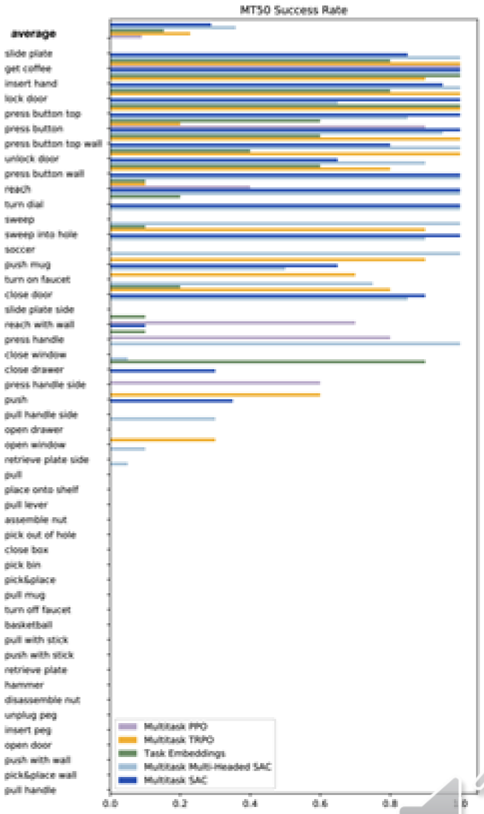
- MT10: Multi-task (10) learning, each with parametric distribution
  - Multi-headed SAC achieves 85% average success rate across the 10 tasks, other candidates ~30%



Empirical Evaluation - MT10 / MT50 Results

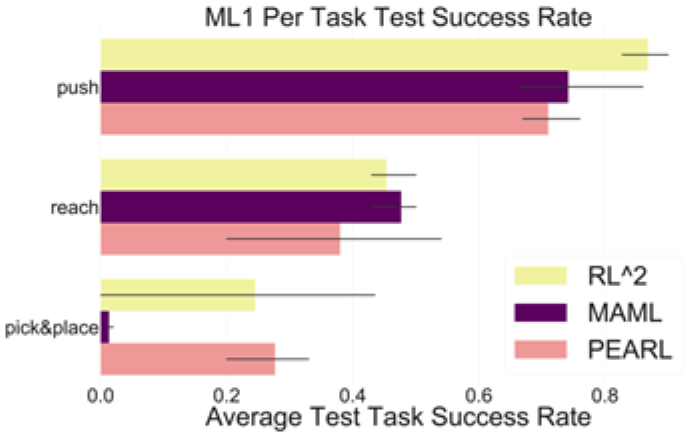
- MT10: Multi-task (10) learning, each with parametric distribution
  - Multi-headed SAC achieves 85% average success rate across the 10 tasks, other candidates ~30%

- MT50: Multi-task (50) learning, each with parametric distribution
  - Multi-headed SAC performance dropped to 40%, others < 30%



Empirical Evaluation - ML1 / ML10 / ML45 Results

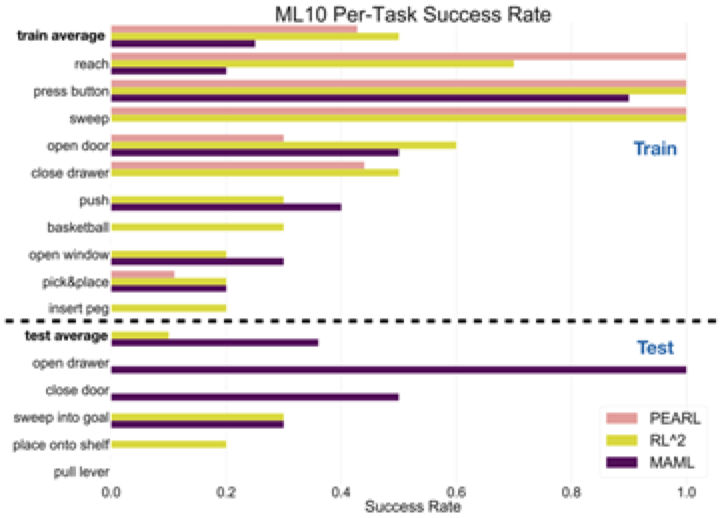
- ML1: Single-task (1) meta-learning with a only parametric distribution
  - Room for improvement even in parametric-only setting which these algorithms were originally designed to succeed in



Empirical Evaluation - ML1 / ML10 / ML45 Results

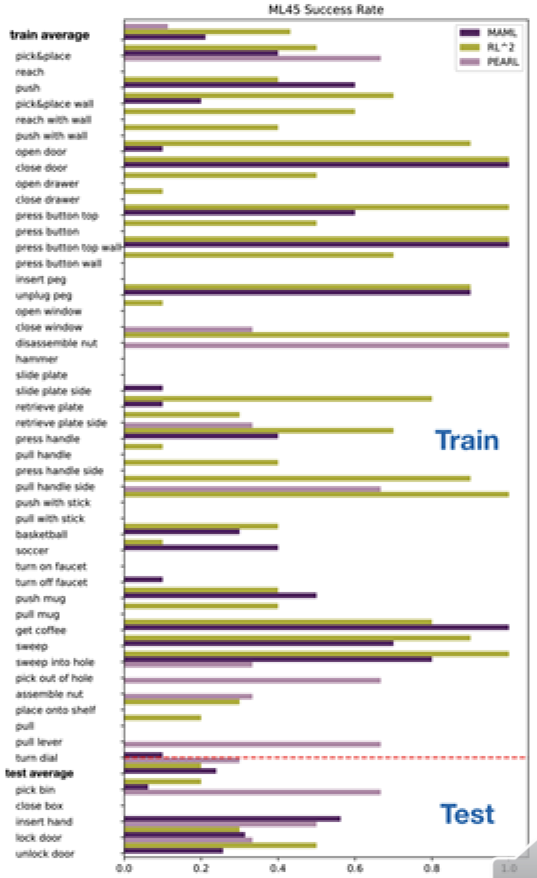
- ML1: Single-task (1) meta-learning with a only parametric distribution
  - Room for improvement even in parametric-only setting which these algorithms were originally designed to succeed in

- ML10: 10 meta-training tasks (with parametric distribution), 5 held-out tasks
  - MAML and RL^2 achieve 40% and 10% average success rate on hold-out, PEARL unable to generalize



Empirical Evaluation - ML10 / ML45 Results

- ML1: Single-task (1) meta-learning with a only parametric distribution
  - Room for improvement even in parametric-only setting which these algorithms were originally designed to succeed in
- ML10: 10 meta-training tasks (with parametric distribution), 5 held-out tasks
  - MAML and RL^2 achieve 40% and 10% average success rate on hold-out, PEARL unable to generalize
- ML45: 45 meta-training tasks (with parametric distribution), 5 held-out tasks
  - PEARL now generalizes best, 30% success rate, whereas MAML and RL^2 drop to 20%



Conclusions - Summary of Contributions

- Meta-World presents an advancement in the multi-task and meta-RL community's ability to benchmark algorithms in a shared structure setting that encourages positive transfer learning and is applicable to real world generalization requirements

- Current multi-task and meta-RL algorithms struggle with the larger scale MT50 and ML45 evaluation protocols, and thus the Meta-World benchmark provides oppurtunities for future algorithm development

Conclusions - Future Work

- **Algorithmic**: Existing meta-RL algorithms struggle in highly diverse (non-parametric) meta-training settings. Techniques to train meta-RL algorithms on broader task distributions are needed to enable methods to generalize effectively to meta-testing tasks.

- **Benchmark**: Current 3D pose observation space is not realistic. Changing to a partially observable setting (images of the workspace) would better match requirements of real world workspaces.