Paper Review - Assignment - Unit Test 1

1. Students are requested to select the research paper related to the course CS 365 Machine Learning and application in the field of respecitve discipline of the student strictly.
2. The paper should be selected from reputed conference or journal only. The journal is preferred.
3. Review the paper, analyze it thoroughly. Answer the following points.
4. One student from each department should coordinate the group of respective department students while selecting the paper to avoid the selection of same paper. It is student responsibility only student to avoid the duplication of paper for assignment.
5. If any students' papers found same, zero unit test assignment marks will be assgined and the second chance will not be given to the student.

* Required

Name of the student *

Bhagya Vinod Rana

Admission number of the student *

U19CS012

Deaprtment of the student *

Computer Science

Area of the paper selected *

Machine Learning, Artificial Intelligence and M

Title of the paper *

Emergent Tool Use from Multi-Agent Interactic

Is it conference paper? *

- (●) Yes
- ( ) No

Is it journal paper? *

- ( ) Yes
- (●) No

Attach a soft copy of the paper (file of paper uploaded should be name with Admission-number-Unit-Test-Assignment-One.pdf) *

PDF U19CS012_Unit_...

What is the problem addressed in the research paper? *

Problem Statement: Play Hide and Seek Game

Applications of the problem addressed in the research paper. *

Application of Multi-agent Reinforcement Learning (MARL)

1) Online Distributed Resource Allocation
Applying multi-agent learning on to come up with effective resource allocation in a network of computing.

2) Cellular Network Optimisation
Applying MARL in LTE networks, guide base stations to maximise mobile service quality.

3) Smart Grid Optimisation
Applying MARL to control power flow in an electrical power grid with optimum efficiency.

4) Smart Cross Light
Applying MARL to control traffic lights to minimise wait time for each car in a city, making them more adaptable based estimates of expected wait time.

5.) Multi-agent reinforcement learning studies how multiple agents interact in a common environment.

That is, when these agents interact with the environment and one another, can we observe them collaborate, coordinate, compete, or collectively learn to accomplish a particular task. It can be further broken down into three broad categories:
A.) Cooperative: All agents working towards a common goal
B.) Competitive: Agents competing with one another to accomplish a goal
C.) Some mix of the two: Think a 5v5 basketball game, where individuals on the same team are coordinating with one another, but the two teams are competing against one another.

Literature summary with key papers. For summary name the major algorithmic approaches reported in the literature for addressing the same problem. *

Through multi-agent competition, the simple objective of hide-and-seek, and standard reinforcement learning algorithms at scale, we find that agents create a self-supervised auto curriculum inducing multiple distinct rounds of emergent strategy, many of which require sophisticated tool use and coordination. We find clear evidence of six emergent phases in agent strategy in our environment, each of which creates a new pressure for the opposing team to adapt; for instance, agents learn to build multi-object shelters using moveable boxes which in turn leads to agents discovering that they can overcome obstacles using ramps. We further provide evidence that multi-agent competition may scale better with increasing environment complexity and leads to behavior that centers on far more human-relevant skills than other self-supervised reinforcement learning methods such as intrinsic motivation. Finally, we propose transfer and fine-tuning as a way to quantitatively evaluate targeted capabilities, and we compare hide-and-seek agents to both intrinsic motivation and random initialization baselines in a suite of domain-specific intelligence tests.

Demo Link: https://youtu.be/Lu56xVlZ40M

Logical discussion of the apporach for the problem addressed in the research paper. (Logical solution to the problem addressed) *

One of the side effects of learning by competition is that agents develop behaviors that are unexpected. In AI theory, this is known as agent autocurricula and represents a first row sit to observing how knowledge develops. Imagine that you are training an AI agent to master a specific game and, suddenly, the agent finds a strategy that has never been tested before. While the autocurricula phenomenon occurs in single-agent reinforcement learning systems, it is even more impressive when develops by competition which is what is known as multi-agent autocurriculum.

In a competitive multi-agent AI environment, the different agents compete against each other in order to evaluate specific strategies. When a new successful strategy or mutation emerges, it changes the implicit task distribution neighboring agents need to solve and creates a new pressure for adaptation. These evolutionary arms races create implicit autocurricula whereby competing agents continually create new tasks for each other. A key element of multi-agent autocurriculum is that the emergent behavior learned by the agents evolves organically and is not the result of pre-built incentive mechanisms. Not surprisingly, multi-agent autocurricula has been one of the most successful techniques when comes to training AI agents in multi-player games.

Technical solution description for the problem addressed in the research paper. *

The core algorithm: The agents are composed of two networks: a policy network to produce an action distribution and a critic network to predict the corresponding future returns. OpenAI researchers used Proximal Policy Optimization (PPO), the technique they have used in training Dota2 computer programs, to optimize the policy. The architecture is shown below.

Policy Architecture [https://i.ibb.co/Vt72FFC/multi-agent-policy-architecture.png]

The AI agents were trained millions of times in parallel. Training toward the final stage (surf defense) in the most complicated environment took three to four days on 16 GPUs and 4,000 CPUs.

[https://i.ibb.co/Mkgm4xQ/six-emergent-method.jpg]

Emergent Skill Progression From Multi-Agent Autocurricula. Through the reward signal of hide-and-seek (shown on the y-axis), agents go through 6 distinct stages of emergence.
(a) Seekers (red) learn to chase hiders, and hiders learn to crudely run away.
(b) Hiders (blue) learn basic tool use, using boxes and sometimes existing walls to construct forts.
(c) Seekers learn to use ramps to jump into the hiders' shelter.
(d) Hiders quickly learn to move ramps to the edge of the play area, far from where they will build their fort, and lock them in place.
(e) Seekers learn that they can jump from locked ramps to unlocked boxes and then surf the box to the hiders' shelter, which is possible because the environment allows agents to move together with the box regardless of whether they are on the ground or not.
(f) Hiders learn to lock all the unused boxes before constructing their fort.

Formulas and equations in support of the technical solution description with explaination should be handwritten on A4 white paper, signed and admission number should be written on top - left corner of the paper and should be uploaded as part of response to this point. *

PDF U19CS012_ML_P...

Simuation - data set selected, characteristic of the data set, or data set collected *

Data Set: mujoco-worldgen package [MuJoCo is a physics engine for detailed, efficient rigid body simulations with contacts.]

Agent policies are trained with
(1) Self-play [https://openai.com/blog/competitive-self-play/]
(2) Proximal Policy Optimization [https://openai.com/blog/openai-baselines-ppo/].

We use the same training infrastructure and algorithms used to train OpenAI Five [https://openai.com/blog/openai-five/] and Dactyl [https://openai.com/blog/learning-dexterity/].

However, in our environment each agent acts independently, using its own observations and hidden memory state. Agents use an entity-centric state-based representation of the world, which is permutation invariant with respect to objects and other agents.

Simulation - tools used - programming or ready to use software *

1.) MuJoCo is a physics engine for detailed, ef

Objective parameter used for analysis *

In the previous section, we qualitatively compare behaviors learned in hide-and-seek to those learned with intrinsic motivation. However, as environments increase in scale, so will the difficulty in qualitatively measuring progress. Tracking reward is an insufficient evaluation metric in multi-agent settings, as it can be ambiguous in indicating whether agents are improving evenly or have stagnated.

Metrics like ELO or Trueskill can more reliably measure whether performance is improving relative to previous policy versions or other policies in a population; however, these metrics still do not give insight into whether improved performance is caused by new adaptations or improving previously learned skills.

Finally, using environment-specific statistics such as object movement can also be ambiguous (for example, the choice to track absolute movement does not illuminate which direction agents moved), and designing sufficient metrics will become difficult and costly as environments scale.

We propose using a suite of domain-specific intelligence tests that target capabilities we believe agents may eventually acquire. Transfer performance in these settings can act as a quantitative measure of representation quality or skill, and we compare against pretraining with count-based exploration as well as a trained from scratch baseline.

Discussion on the simulation performed and results obtained. *

Experiment Results: Compared to previous algorithms such as intrinsic motivation, the hide-and-seek policy is much more human interpretable. Researchers also evaluated the multi-agent hide-and-seek method in object counting, lock and return, sequential lock, blueprint construction, and shelter construction intelligence tasks. The agents performed better than baseline models in three of the five tasks.

Increasing batch size speeds up time to convergence

[https://i.ibb.co/xMqrVL8/increase-batch-size.png]

Effect of Scale on Emergent Autocurricula.
Number of episodes (blue) and wall clock time (orange) required to achieve stage 4 (ramp defense) of the emergent skill progression presented in Figure 1. Batch size denotes number of chunks, each of which consists of 10 contiguous transitions (the truncation length for backpropagation through time).

[https://i.ibb.co/DQFPZs2/fine-tuning-results.png]

We plot the mean normalized performance and 90% confidence interval across 3 seeds smoothed with an exponential moving average, except for Blueprint Construction where we plot over 6 seeds due to higher training variance.
The Improvements are due to the learned feature representations, not from a reuse of skills. Therefore, we can conclude that skills are non-transferrable.

Conclusion (in your words not to be copied the same conclusion mentioned in the paper) *

Competition + Cooperation = Innovation

The OpenAI hide and seek experiments were absolutely fascinating and a clear demonstration of the potential of multi-agent competitive environments as a catalyzer for learning. Many of the OpenAI techniques can be extrapolated it to other AI scenarios in which learning by competition seems like a more viable alternative than supervised training.

As AI agents compete against each other in the environment explained before, they didn't only master hide and seek but they developed as many as six distinct strategies that were not part of the initial incentives.

(a) Running and Chasing
(b) Fort Building
(c) Ramp Use
(d) Ramp Defense
(e) Box Surfing
(f) Surf Defense

[IMAGE-six emergent methods]
We also concluded that Skills are non-Transferrable and the Improvements are due to the learned feature representations, not from a reuse of skills.

The fascinating thing about the emergent behavior developed by the hide and seek agents is that they evolved completely organically as part of the autocurriculum induced by the internal competition. In almost all cases, the performance of the emergent behaviors was superior than those learned by intrinsic motivations.

Your ideas to solve the same problem using any other techniques or approach learned in the subject with detail description. *

We also found that agents were very skilled at exploiting small inaccuracies in the design of the environment, such as seekers surfing on boxes without touching the ground, hiders running away from the environment while shielding themselves with boxes, or agents exploiting inaccuracies of the physics simulations to their advantage. Investigating methods to generate environments without these unwanted behaviors is another import direction of future research.

Additional Evidence:
https://openai.com/blog/openai-five-defeats-dota-2-world-champions/
https://deepmind.com/research/case-studies/alphago-the-story-so-far#our_approach

Submit Clear form