EvalAI: Towards Better Evaluation of AI Agents

Deshraj Yadav[1] **Rishabh Jain**[1] **Harsh Agrawal**[1] **Prithvijit Chattopadhyay**[1]
Taranjeet Singh[2] **Akash Jain**[3] **Shiv Baran Singh**[4] **Stefan Lee**[1] **Dhruv Batra**[1]
[1]Georgia Institute of Technology [2]Paytm [3]Zomato [4]Cyware

Abstract

We introduce EvalAI, an open source platform for evaluating artificial intelligence algorithms (AI) at scale. EvalAI is built to provide a scalable solution to the research community to fulfill the critical need of evaluating ML models and AI agents in a dynamic environment against ground-truth annotations or by interacting with a human. This will help researchers, students, and data scientists to create, collaborate, and participate in AI challenges organized around the globe. By simplifying and standardizing the process of benchmarking these models, EvalAI seeks to lower the barrier to entry for participating in the global scientific effort to push the frontiers of ML and AI, thereby increasing the rate of measurable progress in this domain. Our code is available at https://github.com/Cloud-CV/EvalAI.

1 Introduction

Progress on several important problems in Computer Vision (CV) and Artificial Intelligence (AI) has been driven by the introduction of bold new tasks coupled with the curation of large, realistic datasets [4, 11, 17, 19, 22]. Not only do these tasks and datasets establish new problems and provide data necessary to analyze them, but more importantly they also establish reliable benchmarks where proposed solutions and hypothesis can be tested – an essential part of the scientific process. In recent years, the development of centralized evaluation platforms have lowered the barrier to compete and share results on these problems. As a result, a thriving community of researchers has grown around these tasks, thereby increasing the pace of progress and technical dissemination.

With the success of deep learning techniques on a wide variety of complex AI tasks such as grounded dialog generation [11] or generating aesthetically pleasing images [15] coupled with the widespread proliferation of AI-driven smart applications, there is an imminent to evaluate AI systems in the context of human collaborators. These tasks cannot be evaluated accurately using automatic metrics as performance on these metrics do not correlate well with human-judgment in practice[7]. Instead, to properly evaluate, they should be connected with a human workforce such as Amazon Mechanical Turk (AMT)[2] to mimic a setup which is closest to the one in which they may be eventually deployed.

Furthermore, the rise of reinforcement learning (RL) based problems in which an agent must interact with an environments introduces additional challenges for benchmarking. Unlike supervised learning, the performance in this setup cannot be measured by evaluating on a static test set. Evaluating these agents involves running the users code on a collection of unseen environments such that one can check if algorithms "overfit" on training environments.

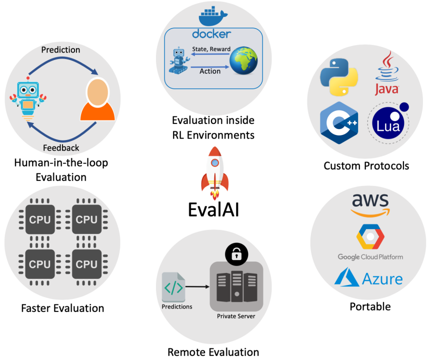

Figure 1. EvalAI is a platform to evaluate AI agents in dynamic environment with human-in-the-loop.

To address the aforementioned problems, we introduce a new evaluation platform called EvalAI that fullfills the critical need in the community for (1) human-in-the-loop evaluation of machine learning models and (2) the ability to run user's code in a dynamic environment instead of a static dataset enabling the evaluation of interactive agents.

2 Related work

In light of the requirements highlighted in the previous section, we compare EvalAI with existing platforms. We also provide a a head-to-head ocmparison in Table 1. Kaggle[16], CodaLab[8] and AICrowd[1] are some of the most popular platforms for hosting machine learning competitions but they have several limitations. Kaggle doesn't support custom evaluation metrics and multiple challenge phases – a common practice in popular challenges like COCO Caption Challenge, VQA etc. CodaLab provides an open-source alternative to Kaggle and fixes several of their limitations but doesn't support evaluating interactive agents in dynamic environments. EvalAI not only supports custom evaluation protocol but also allows evaluation of interactive agents in dynamic environments. In addition, we also support human-in-the loop evaluation of prediction based or code-upload based challenges, something AICrowd doesn't support. Similar to ParlAI [21], EvalAI integrates with Amazon Mechanical Turk (AMT) [2] for human based evaluation. However, unlike EvalAI, ParlAI is not a challenge hosting platform and only supports evaluation of dialog models, not for any AI task in general. OpenAI gym [5] and EvalAI have the same

Features	OpenML	Topcoder	Kaggle	AICrowd	ParlAI	CodaLab	EvalAI
AI Challenge Hosting	✗	✓	✓	✓	✗	✓	✓
Custom metrics	✗	✗	✗	✓	✓	✓	✓
Multiple phases/splits	✗	✗	✗	✓	✗	✓	✓
Open Source	✓	✗	✗	✓	✓	✓	✓
Remote Evaluation	✗	✗	✗	✗	✓	✓	✓
Human Evaluation	✗	✗	✗	✗	✓	✗	✓
Environments	✗	✗	✗	✓	✗	✗	✓

Table 1. Head-to-head comparison of capabilities between existing platforms and EvalAI

underlying philosophy of encouraging easy accessibility and reproducibilty of Reinforcement Learning (RL) agents but OpenAI gym is not a dedicated evaluation platform and lacks support for prediction based challenges, custom evaluation protocol, and human-in-the loop evaluation.

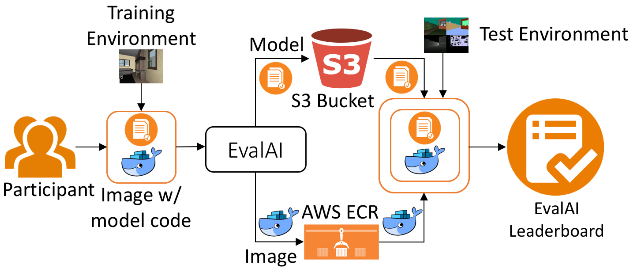

Figure 2. System architecture for code upload challenges

3 Key features

Evaluation inside RL environments: We have developed an evaluation framework to evaluate agents for tasks situated in active environments instead of static datasets (Figure 2). Participants upload Docker images with their pre-trained models using a command line interface. At the time of evaluation, the instantiated worker evaluates the user-submitted model against test-environment provided by the challenge organizer. Once the evaluation is complete, the results are sent over to the leaderboard using the message queue.

Human-in-the-loop evaluation: Automatic evaluation of tasks like image captioning [10, 18], visual dialog [11, 12] or image generation [15] is complicated by the huge set of possibly 'correct' responses and relatively sparse ground truth annotations. Given the interactive nature of tasks, it is clear that the most appropriate way to evaluate these kind of tasks is with a human in the loop, i.e. a Visual Turing Test [14]! Unfortunately, human-in-the-loop evaluation is still limited by financial and infrastructural challenges that must be overcome by each interested research group independently.

To address this, we have developed the infrastructure to pair AMT users in real-time with artificial agents (for instance, visual conversational agents [11]). We provide:

- **Custom HTML Templates**: Organizers can choose to provide their own HTML templates satisfying the unique requirements specific to their challenge.
- **Worker Pool**: We maintain a pool of good quality workers which have a history of high quality work and strong acceptance rate. Additionally, organizers can provide us with a list of whitelisted and blocked workers.

Year	# submissions	# participants	# challenges	# page views
2018	12,516	357	11	306,517
2019 (YTD)	23,357	1,069	25	642,383
Growth	**86%**	**186.5%**	**127%**	**109.6%**

Table 2. EvalAI growth statistics

- **Uninterrupted back-and-forth communication**: For tasks that require multiple rounds of human-AI communication, we do a lot of book-keeping to ensure that incompleted HITs are re-evaluated and turkers can reconnect with the same agent after temporary network failure.
- **Flexible schema**: We provide a flexible JSON based schema and APIs to fetch the results from the evaluation tasks once they are completed. These results are automatically updated on the leaderboard for each submission.

Private and Remote Evaluation: Certain large-scale challenges have special compute requirements for evaluation. For instance, challenges in medical domain such as FastMRI Image Reconstruction challenge [23] have sensitive data which cannot be shared with the evaluation platform. Some other AI challenges like CARLA Autonomous Driving challenge [13], and Animal-AI Olympics [9] need to run RL agents in a dynamic environment - requiring powerful clusters with GPUs. For these types of challenges, organizers can easily setup their own cluster of worker nodes to process participant submissions while we take care of hosting the challenge, handling user submissions and the maintaining the leaderboard. On submission, all related metadata is relayed to an external pool of workers through dedicated message queues - decoupling the worker nodes from the challenge front-end.

4 Impact

As shown in Table 2, EvalAI has already hosted 35+ challenges, with over 1400 participants from 84 countries who have created over 35000 submissions. Some of the large scale challenge that EvalAI hosted are CARLA Autonomous Driving Challenge [13], Animal-AI Olympics Competition [9], Vision and Language Navigation [3], Habitat Challenge [20], Visual Question Answering Challenge [4] and many more.

5 Conclusion

While traditional platforms were adequate for evaluation of tasks using automatic metrics, there is a critical need to support human-in-the-loop evaluation for more free-form multimodal tasks such as (visual) dialog and image generation. We develop, EvalAI, a large-scale evaluation platform to support the same. To this end, EvalAI supports pairing an AI agent with thousands of workers in an interactive dynamic environment so as to rate or evaluate the former over multiple rounds of interaction. By providing a scalable platform that supports such evaluations will eventually encourage the community to benchmark performance on tasks extensively, leading to better understanding of a model's performance both in isolation and in human-AI teams[6].

References

[1] AICrowd [n. d.]. AICrowd. Website - https://www.aicrowd.com/.

[2] AMT [n. d.]. Amazon Mechanical Turk (AMT). Website - https://www.mturk.com/.

[3] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian D. Reid, Stephen Gould, and Anton van den Hengel. 2017. Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2017), 3674–3683.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

[5] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. 2016. OpenAI Gym. *CoRR* abs/1606.01540 (2016).

[6] Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Mark Johnson. 2018. Do Explanations make VQA Models more Predictable to a Human?. In *EMNLP*.

[7] Prithvijit Chattopadhyay, Deshraj Yadav, Viraj Prabhu, Arjun Chandrasekaran, Abhishek Das, Stefan Lee, Dhruv Batra, and Devi Parikh. 2017. Evaluating Visual Conversational Agents via Cooperative Human-AI Games. In *Proceedings of the Fifth AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*.

[8] CodaLab [n. d.]. CodaLab. Website - https://competitions.codalab.org/.

[9] Matthew Crosby, Benjamin Beyret, and Marta Halina. 2019. The Animal-AI Olympics. *Nature Machine Intelligence* 1 (2019), 257.

[10] Bo Dai, Dahua Lin, Raquel Urtasun, and Sanja Fidler. 2017. Towards Diverse and Natural Image Descriptions via a Conditional GAN. *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), 2989–2998.

[11] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 1080–1089.

[12] Abhishek Das, Satwik Kottur, José M. F. Moura, Stefan Lee, and Dhruv Batra. 2017. Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning. *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), 2970–2979.

[13] Alexey Dosovitskiy, Germán Ros, Felipe Codevilla, Antonio López, and Vladlen Koltun. 2017. CARLA: An Open Urban Driving Simulator. In *CoRL*.

[14] Donald Geman, Stuart Geman, Neil Hallonquist, and Laurent Younes. 2015. Visual turing test for computer vision systems. *Proceedings of the National Academy of Sciences* (2015), 201422953.

[15] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems (NIPS)*.

[16] Kaggle [n. d.]. Kaggle. Website - https://kaggle.com/.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoff Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems (NIPS)*.

[18] Dianqi Li, Xiaodong He, Qiuyuan Huang, Ming-Ting Sun, and Lei Zhang. 2018. Generating Diverse and Accurate Visual Captions by Comparative Adversarial Learning. *arXiv preprint arXiv:1804.00861* (2018).

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[20] Manolis Savva*, Abhishek Kadian*, Oleksandr Maksymets*, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. 2019. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[21] Alexander H Miller, Will Feng, Adam Fisch, Jiasen Lu, Dhruv Batra, Antoine Bordes, Devi Parikh, and Jason Weston. 2017. Parlai: A dialog research software platform. *arXiv preprint arXiv:1705.06476* (2017).

[22] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100, 000+ Questions for Machine Comprehension of Text. In *EMNLP*.

[23] Jure Zbontar, Florian Knoll, Anuroop Sriram, Matthew J Muckley, Mary Bruno, Aaron Defazio, Marc Parente, Krzysztof J Geras, Joe Katsnelson, Hersh Chandarana, et al. 2018. fastMRI: An Open Dataset and Benchmarks for Accelerated MRI. *arXiv preprint arXiv:1811.08839* (2018).