Workshop organizers make last-minute changes to their schedule. Download this document again to get the lastest changes, or use the ICLR mobile application.

Schedule Highlights

May 6, 2019

Room R01, **The 2nd Learning from Limited Labeled Data (LLD) Workshop: Representation Learning for Weak Supervision and Beyond** *Augenstein, Bach, Blaschko, Belilovsky, Oyallon, Platanios, Ratner, Re, Ren, Varma*

Room R02, **Deep Reinforcement Learning Meets Structured Prediction** *Liang, Lao, Ling, Marinho, Tian, Wang, Williams, Durand, Martins*

Room R02, **Deep Generative Models for Highly Structured Data** *Dieng, Kim, Reddy, Cho, Dyer, Blei, Blunsom*

Room R03, **Debugging Machine Learning Models** *Lakkaraju, Tan, Adebayo, Steinhardt, Sculley, Caruana*

Room R04, **Structure & Priors in Reinforcement Learning (SPiRL)** *Bacon, Deisenroth, Finn, Grant, Griffiths, Gupta, Heess, Littman, Oh*

Room R05, **AI for Social Good** *Luck, Sylvain, Sankaran, McGregor, Penn, Sylvain, Boucher, Cote, Toyama, Ghani, Bengio*

Room R06, **Safe Machine Learning: Specification, Robustness, and Assurance** *Chiappa, Krakovna, Garriga-Alonso, Trask, Uesato, Heinze-Deml, Jiang, Weller*

Room R07, **Representation Learning on Graphs and Manifolds** *Hamilton, Sala, Battaglia, Bruna, Kipf, Li, Pascanu, Romero, Veli■kovi■, Zitnik, Nickel, Gunel, Gu, Re*

Room R08, **Reproducibility in Machine Learning** *Ke, Lamb, Alias Parth Goyal, Courville, Bengio*

Room R09, **Task-Agnostic Reinforcement Learning (TARL)** *Hafner, Zhang, Touati, Pathak, Ebert, McAllister, Calandra, Bellemare, Hadsell, Pineau*

May 7, 2019

Room R01, **LatinX in AI and Black in AI Joint Workshop**

May 6, 2019

The 2nd Learning from Limited Labeled Data (LLD) Workshop: Representation Learning for Weak Supervision and Beyond

Isabelle Augenstein, Stephen Bach, Matthew Blaschko, Eugene Belilovsky, Edouard Oyallon, Anthony Platanios, Alex Ratner, Christopher Re, Xiang Ren, Paroma Varma

Room R01, Mon May 06, 09:45 AM

Modern representation learning techniques like deep neural networks have had a major impact on a wide range of tasks, achieving new state-of-the-art performances on benchmarks using little or no feature engineering. However, these gains are often difficult to translate into real-world settings because they usually require massive hand-labeled training sets. Collecting such training sets by hand is often infeasible due to the time and expense of labeling data; moreover, hand-labeled training sets are static and must be completely relabeled when real-world modeling goals change.

Increasingly popular approaches for addressing this labeled data scarcity include using weak supervision---higher-level approaches to labeling training data that are cheaper and/or more efficient, such as distant or heuristic supervision, constraints, or noisy labels; multi-task learning, to effectively pool limited supervision signal; data augmentation strategies to express class invariances; and introduction of other forms of structured prior knowledge. An overarching goal of such approaches is to use domain knowledge and data resources provided by subject matter experts, but to solicit it in higher-level, lower-fidelity, or more opportunistic ways.

In this workshop, we examine these increasingly popular and critical techniques in the context of representation learning. While approaches for representation learning in the large labeled sample setting have become increasingly standardized and powerful, the same is not the case in the limited labeled data and/or weakly supervised case. Developing new representation learning techniques that address these challenges is an exciting emerging direction for research [e.g., 1, 2]. Learned representations have been shown to lead to models robust to noisy inputs, and are an effective way of exploiting unlabeled data and transferring knowledge to new tasks where labeled data is sparse.

In this workshop, we aim to bring together researchers approaching these challenges from a variety of angles. Specifically this includes:
Learning representations to reweight and de-bias weak supervision
Representations to enforce structured prior knowledge (e.g. invariances, logic constraints).
Learning representations for higher-level supervision from subject matter experts
Representations for zero and few shot learning
Representation learning for multi-task learning in the limited labeled setting
Representation learning for data augmentation
Theoretical or empirically observed properties of representations in the above contexts

The second LLD workshop continues the conversation from the 2017 NIPS Workshop on Learning with Limited Labeled Data (http://lld-workshop.github.io). LLD 2017 received 65 submissions, of which 44 were accepted and was one of the largest workshops at NIPS 2017. Our goal is to once again bring together researchers interested in this growing field. With funding support, we are excited to again organize best paper awards for the most outstanding submitted papers. We also will have seven distinguished and diverse speakers from a range of machine learning perspectives, a panel on where the most promising directions for future research are, and a discussion session on developing new benchmarks and other evaluations for these techniques.

The LLD workshop organizers are also committed to fostering a strong sense of inclusion for all groups at this workshop, and to help this concretely, aside from $$1K for the paper awards, the remainder of the funding (both current and pending) will sponsor several travel awards specifically for traditionally underrepresented groups. We will also post a code of conduct emphasizing our commitment to inclusion, which we will expect all attendees to uphold.

[1] Norouzi et al. “Zero-Shot Learning by Convex Combination of Semantic Embeddings.” ICLR 2014.
[2] Liu et al. “Heterogeneous Supervision for Relation Extraction: A Representation Learning Approach.” EMNLP 2017.

Deep Reinforcement Learning Meets Structured Prediction

Chen Liang, Ni Lao, Wang Ling, Zita Marinho, Yuandong Tian, Lu Wang, Jason D Williams, Audrey Durand, Andre Martins

Room R02, Mon May 06, 09:45 AM

Website: https://sites.google.com/view/iclr2019-drlstructpred
ICLR page:
https://iclr.cc/Conferences/2019/Schedule?showEvent=630
Submission website:
https://openreview.net/group?id=ICLR.cc/2019/Workshop/drlStructPred

Important Dates
Submission open March 6
Submission deadline March 15 (11:59pm AOE)
Decisions April 6
Camera Ready April 28 (11:59pm AOE)
Workshop May 6

Deep reinforcement learning has achieved successes on numerous tasks such as computer games, the game of Go, robotics, etc. Structured prediction aims at modeling highly dependent variables, which applies to a wide range of domains such as natural language processing, computer vision, computational biology, etc. In many cases, structured prediction can be viewed as a sequential decision making process, so a natural question is can we leverage the advances in deep RL to improve structured prediction?

Recently, promising results have been shown applying RL to various structured prediction problems such as dialogue (Li et al, 2016; Williams et al, 2017; He et al, 2017), program synthesis (Bunel et al, 2018; Liang et al, 2018), semantic parsing (Liang et al, 2017), architecture search (Zoph & Le, 2017), chunking and parsing (Sharaf & Daumé III 2018), machine translation (Ranzato et al, 2015; Norouzi et al, 2016; Bahdanau et al, 2016), summarization (Paulus et al, 2017), image caption (Rennie et al, 2017), knowledge graph reasoning (Xiong et al, 2017), query rewriting (Nogueira et al, 2017; Buck et al, 2017) and information extraction (Narasimhan et al, 2016; Qin et al, 2018). However, there are also negative results where RL is not efficient enough comparing to alternative approaches (Guu et al, 2017; Bender et al, 2018; Xu et al, 2018). As a community it is very important to figure out the limit and future directions of RL in structured prediction.

This workshop will bring together experts in structured predictions and reinforcement learning. Specifically, it will provide an overview of existing approaches from various domains to distill generally applicable principles from their successes. We will also discuss the main challenges arising in this setting and outline potential directions for future progress. The target audience consists of researchers and practitioners in this areas. They include, but are not limited to, deep RL for:
dialogue
semantic parsing
program synthesis
architecture search
machine translation
summarization
image caption
knowledge graph reasoning
information extraction

Area: Reinforcement Learning, Applications

References

Sutton, Richard S., and Andrew G. Barto. (1998). Reinforcement learning: An introduction. Vol. 1. No. 1. Cambridge: MIT press.
Hal Daumé III, John Langford and Daniel Marcu. (2009). Search-based Structured Prediction. Machine Learning Journal.
Hal Daumé III. (2017). Structured prediction is *not* RL. Blogspot.
He, Di, et al. (2016). Dual learning for machine translation. NIPS.
Ranzato, Marc'Aurelio, et al. (2015). Sequence level training with recurrent neural networks. arXiv preprint arXiv:1511.06732.
Y. Efroni, G. Dalal, B. Scherrer, S. Mannor. (2019). How to Combine Tree-Search Methods in Reinforcement Learning, AAAI.
Bahdanau, Dzmitry, et al. (2016). An actor-critic algorithm for sequence prediction. arXiv preprint arXiv:1607.07086.
Bunel, Rudy, et al. (2018). Leveraging grammar and reinforcement learning for neural program synthesis. arXiv preprint arXiv:1805.04276.
Buck, Christian, et al. (2017) Ask the right questions: Active question reformulation with reinforcement learning. arXiv preprint arXiv:1705.07830..
Nogueira, Rodrigo, and Kyunghyun Cho. (2017). Task-oriented query reformulation with reinforcement learning. arXiv preprint arXiv:1704.04572.
Paulus Romain, Caiming Xiong, and Richard Socher. (2017).

A deep reinforced model for abstractive summarization. arXiv preprint arXiv:1705.04304.
Norouzi, Mohammad et al. (2016) Reward Augmented Maximum Likelihood for Neural Structured Prediction. NIPS.
Williams, Jason D., Kavosh Asadi, and Geoffrey Zweig. (2017). Hybrid Code Networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vol. 1.
Li, Jiwei, et al. (2016) Deep reinforcement learning for dialogue generation. arXiv preprint arXiv:1606.01541.
Kirthevasan Kandasamy, Yoram Bachrach, Ryota Tomioka, Daniel Tarlow, David Carter. (2017). Batch Policy Gradient Methods for Improving Neural Conversation Models. ICLR.
Narasimhan, K., Yala, A., & Barzilay, R. (2016). Improving Information Extraction by Acquiring External Evidence with Reinforcement Learning. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (pp. 2355-2365).
Rennie, Steven J., et al. (2017). Self-critical sequence training for image captioning. CVPR. Vol. 1. No. 2.
Michael Gygli, Mohammad Norouzi, Anelia Angelova. (2017). Deep Value Networks Learn to Evaluate and Iteratively Refine Structured Outputs. ICML.
Barret Zoph, Quoc V. Le. (2017). Neural Architecture Search with Reinforcement Learning. ICLR.
Chen Liang, Jonathan Berant, Quoc Le, Kenneth D. Forbus, Ni Lao. (2017). Neural Symbolic Machines: Learning Semantic Parsers on Freebase with Weak Supervision. ACL.
Kelvin Guu, Panupong Pasupat, Evan Zheran Liu, Percy Liang. (2017). From language to programs: Bridging reinforcement learning and maximum marginal likelihood. ACL.
Daniel A Abolafia, Mohammad Norouzi, Jonathan Shen, Rui Zhao, Quoc V. Le. (2018). Neural Program Synthesis with Priority Queue Training.
Chen Liang, Mohammad Norouzi, Jonathan Berant, Quoc Le, Ni Lao. (2018) Memory Augmented Policy Optimization for Program Synthesis with Generalization. NeurPS.
CJ Maddison*, D Lawson*, G Tucker*, N Heess, M Norouzi, A Doucet, A Mnih, YW Teh. (2017). Filtering Variational Objectives. NIPS.
Dieterich Lawson, Chung-Cheng Chiu, George Tucker, Colin Raffel, Kevin Swersky, Navdeep Jaitly. (2018). Learning hard alignments with variational inference. ICASSP.
Gabriel Bender, Pieter-Jan Kindermans, Barret Zoph, Vijay Vasudevan, Quoc Le. (2018). Understanding and simplifying one-shot architecture search. ICML.
Hoang M. Le, Nan Jiang, Alekh Agarwal, Miroslav Dudík, Yisong Yue, Hal Daumé III. (2018). Hierarchical Imitation and Reinforcement Learning. ICML.
Amr Sharaf, Hal Daumé III. (2017). Structured prediction via learning to search under bandit feedback. SP4NLP workshop.
Xiaojun Xu, Chang Liu, Dawn Song. (2018). Sqlnet: Generating structured queries from natural language without reinforcement learning.
W Xiong, T Hoang, WY Wang DeepPath. (2017). A Reinforcement Learning Method for Knowledge Graph Reasoning. EMNLP.
Pengda Qin, Weiran Xu, William Yang Wang. (2018). Robust Distant Supervision Relation Extraction via Deep Reinforcement Learning. ACL.
D. Bahdanau, P. Brakel, K. Xu, A. Goyal, R. Lowe, J. Pineau, A. Courville, Y. Bengio. (2017). An Actor-Critic Algorithm for Sequence Prediction. ICLR.

Deep Generative Models for Highly Structured Data

Adji Bousso Dieng, Yoon Kim, Siva Reddy, Kyunghyun Cho, Chris Dyer, David Blei, Phil Blunsom

Room R02, Mon May 06, 15:15 PM

Deep generative models are at the core of research in Artificial Intelligence. They have achieved remarkable performance in many domains including computer vision, speech recognition, and audio synthesis. In recent years, they have infiltrated other fields of science including the natural sciences, physics, chemistry and molecular biology, and medicine. Despite these successes, deep generative models still face many challenges when they are used to model highly structured data such as natural language, video, and generic graph-structured data such as molecules. This workshop aims to bring experts from different backgrounds and perspectives to discuss the applications of deep generative models to these data modalities.

Relevant topics to this workshop include but are not limited to:

--Generative models for graphs, text, video, and other structured modalities
--Unsupervised representation learning of high dimensional structured data
--Learning and inference algorithms for deep generative models
--Evaluation methods for deep generative models

--Applications and practical implementations of deep generative models
--Scalable algorithms to accelerate learning with deep generative models
--Visualization methods for deep generative models
--Empirical analysis comparing different architectures for a given data modality

Debugging Machine Learning Models

Hima Lakkaraju Lakkaraju, Sarah Tan, Julius Adebayo, Jacob Steinhardt, D. Sculley, Rich Caruana

Room R03, Mon May 06, 09:45 AM

Machine learning (ML) models are increasingly being employed to make highly consequential decisions pertaining to employment [Dastin, 2018], bail [Kleinberg et. al., 2017], parole [Dressel and Farid, 2018], and lending [Hurley et al., 2016]. While such models can learn from large amounts of data and are often very scalable, their applicability is limited by certain safety challenges. A key challenge is to be able to identify and correct systematic patterns of mistakes made by ML models before deploying them in the real world.

In order to address the aforementioned challenge, machine learning can potentially take cues from traditional software engineering literature, which has put significant emphasis on the development of rigorous tools for debugging and formal methods for program verification. While these methods are by no means complete or foolproof, there is ample evidence that they help in developing reliable and robust software [D'Silva et. al., 2008]. ML pipelines currently lack analogous infrastructure [Breck et. al. 2016] and it would be interesting to explore how to address this shortcoming. Furthermore, some recent research in machine learning has focussed on developing methods and tools for testing and verifying model violations to fairness, robustness, and security constraints [Cotter et. al. 2018, Dvijotham et. al. 2018, Kearns et. al. 2017, Odena et. al. 2018, Selsam et. al. 2017, Stock et. al. 2018, Tian et. al. 2017, Wicker et. al. 2017]. For example, interpretable models have been proposed to detect misclassifications and dataset biases [Koh and Liang, 2017; Kim et al., 2018; Lakkaraju et. al., 2017; Zhang et al., 2018]. The field of adversarial learning has proposed techniques which leverage the process of generation of adversarial examples (and defenses against them) to highlight vulnerabilities in ML models [Goodfellow et. al., 2014, Elsayed et. al., 2018]. Several of the aforementioned research topics have their own longstanding workshops. Yet, to the best of our knowledge, there has not been a single workshop that brings together researchers (spanning all the aforementioned topics) working on the common theme of debugging ML models.

The goal of this workshop is to bring together researchers and practitioners interested in research problems and questions pertaining to the debugging of machine learning models. For the first edition of this workshop, we intend to focus on research that approaches the problem of debugging ML models from the following perspectives:

• Interpretable and explainable ML
• Formal methods and program verification
• Visualization and human factors
• Security and adversarial examples in ML

By bringing together researchers and practitioners working in the aforementioned research areas, we hope to address several key questions pertaining to model debugging (some of which are highlighted below) and facilitate an insightful discussion about the strengths and weaknesses of existing approaches:

• How can interpretable models and techniques aid us in effectively debugging ML models?
• Are existing program verification frameworks readily applicable to ML models? If not, what are the gaps that exist and how do we bridge them?
• What kind of visualization techniques would be most effective in exposing vulnerabilities of ML models?
• What are some of the effective strategies for using human input and expertise for debugging ML models?
• How do we design adversarial attacks that highlight vulnerabilities in the functionality of ML models?
• How do we provide guarantees on the correctness of proposed debugging approaches? Can we take cues from statistical considerations such as multiple testing and uncertainty to ensure that debugging methodologies and tools actually detect 'true' errors?
• Given a ML model or system, how do we bound the probability of its failures?
• What can we learn from the failures of widely deployed ML systems? What can we say about debugging for different types of biases, including discrimination?
• What are standardized best practices for debugging large-scale ML systems? What are existing tools, software, and hardware, and how might they be improved?
• What are domain-specific nuances of debugging ML

models in healthcare, criminal justice, public policy, education, and other social good applications?

Call for Papers and More Info:
Please see the workshop website (https://debug-ml-iclr2019.github.io/).

Target Audience:
We anticipate this workshop to be of interest and utility to researchers in at least four different research areas that we have focused our workshop agenda on. Since there will be contributed posters and talks from students, we expect a good number of young researchers to attend. Additionally, we expect two components of our agenda -- the opinion piece and the panel -- to generate a lot of excitement and debate in the research community.

Schedule

09:50 AM	**Welcome**	
10:00 AM	**Research Invited Talk**	*Madry*
10:30 AM	**Similarity of Neural Network Representations Revisited**	*Kornblith*
10:40 AM	**Error terrain analysis for machine learning: Tool and visualizations**	
10:50 AM	**Coffee Break**	
11:10 AM	**Research Invited Talk**	*Bastani*
11:40 AM	**Debugging Machine Learning via Model Assertions**	*Kang*
11:50 AM	**Improving Jobseeker-Employer Match Models at Indeed Through Process, Visualization, and Exploration**	
12:00 PM	**Break**	
12:10 PM	**Research Invited Talk**	*Singh*
12:40 PM	**Practical Invited Talk**	*Raji*
01:00 PM	**NeuralVerification.jl: Algorithms for Verifying Deep Neural Networks**	*Arnon, Lazarus*
01:10 PM	**Lunch**	
03:20 PM	**Welcome Back**	
03:30 PM	**Research Invited Talk**	*Saria*
04:00 PM	**Practical Invited Talk**	*Moldovan*
04:20 PM	**Poster & Demo Session + Coffee Break**	
05:20 PM	**The Scientific Method in the Science of Machine Learning**	*Paganini*
05:30 PM	**Invited Opinion Piece**	*Rudin*
06:00 PM	**The Future of Debugging**	*Caruana, Rudin, Raji, Saria, Madry, Moldovan, Bastani, Singh*
06:25 PM	**Closing**	

Structure & Priors in Reinforcement Learning (SPiRL)

Pierre-Luc Bacon, Marc Deisenroth, Chelsea Finn, Erin Grant, Thomas L Griffiths, Abhishek Gupta, Nicolas Heess, Michael L. Littman, Junhyuk Oh

Room R04, Mon May 06, 09:45 AM

Abstract

Generalization and sample complexity remain unresolved problems in reinforcement learning (RL), limiting the applicability of these methods to real-world problem settings. A powerful solution to these challenges lies in the deliberate use of inductive bias, which has the potential to allow RL algorithms to acquire solutions from significantly fewer samples and with greater generalization performance [[Ponsen et al., 2009](https://link.springer.com/chapter/10.1007/978-3-642-11814-2_1 "M. Ponsen, M. E. Taylor, and K. Tuyls. Abstraction and generalization in reinforcement learning: A summary and framework. In International Workshop on Adaptive and Learning Agents, pages 1–32. Springer, 2009.")]. However, the question of what form this inductive bias should take in

the context of RL remains an open one. Should it be provided as a prior distribution for use in Bayesian inference [[Ghavamzadeh et al., 2015](https://arxiv.org/abs/1609.04436 "M. Ghavamzadeh, S. Mannor, J. Pineau, A. Tamar, et al. Bayesian reinforcement learning: A survey. Foundations and Trends in Machine Learning, 8(5-6):359–483, 2015.")], learned wholly from data in a multi-task or meta-learning setup [[Taylor and Stone, 2009](http://www.jmlr.org/papers/v10/taylor09a.html "M. E. Taylor and P. Stone. Transfer learning for reinforcement learning domains: A survey. JMLR, 10(1):1633–1685, 2009.")], specified as structural constraints (such as temporal abstraction [[Parr and Russell, 1998](https://papers.nips.cc/paper/1384-reinforcement-learning-with-hierarchies-of-machines "R. Parr and S. J. Russell. Reinforcement learning with hierarchies of machines. In NeurIPS, pages 1043–1049, 1998."), [Dietterich, 2000](https://arxiv.org/abs/cs/9905014 "T. G. Dietterich. Hierarchical reinforcement learning with the MAX-Q value function decomposition. JAIR, 13:227–303, 2000."), [Sutton et al., 1999](https://www.sciencedirect.com/science/article/pii/S0004370299000521 "R. S. Sutton, D. Precup, and S. Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. Artificial intelligence, 112(1-2):181–211, 1999.")] or hierarchy [[Singh, 1992](https://link.springer.com/article/10.1007/BF00992700 "S. P. Singh. Transfer of learning by composing solutions of elemental sequential tasks. Machine Learning, 8:323–339, 1992."), [Dayan and Hinton, 1992](https://papers.nips.cc/paper/714-feudal-reinforcement-learning "P. Dayan and G. E. Hinton. Feudal reinforcement learning. In NeurIPS,1992.")]), or some combination thereof?

The computational cost of recently successful applications of RL to complex domains such as gameplay [[Silver et al., 2016](https://www.nature.com/articles/nature16961 "D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587):484, 2016."), [Silver et al., 2017](https://www.nature.com/articles/nature24270 "D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al. Mastering the game of Go without human knowledge. Nature, 550(7676):354, 2017."), [OpenAI, 2018](https://blog.openai.com/openai-five/ "OpenAI. OpenAI Five, 2018.")] and robotics [[Levine et al., 2018](https://arxiv.org/abs/1603.02199 "S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. IJRR, 37(4-5):421–436, 2018."), [Kalashnikov et al., 2018](https://arxiv.org/abs/1806.10293 "D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, et al. QT-Opt: Scalable deep reinforcement learning for vision-based robotic manipulation. arXiv preprint arXiv:1806.10293, 2018.")] has led to renewed interest in answering this question, most notably in the specification and learning of structure [[Vezhnevets et al., 2017](https://arxiv.org/abs/1703.01161 "A. S. Vezhnevets, S. Osindero, T. Schaul, N. Heess, M. Jaderberg, D. Silver, and K. Kavukcuoglu. FeUdal networks for hierarchical reinforcement learning. In ICML, pages 3540–3549, 2017."), [Frans et al., 2018](https://arxiv.org/abs/1710.09767 "K. Frans, J. Ho, X. Chen, P. Abbeel, and J. Schulman. Meta-learning shared hierarchies. In ICLR, 2018."), [Andreas et al., 2017](https://arxiv.org/abs/1611.01796 "J. Andreas, D. Klein, and S. Levine. Modular multitask reinforcement learning with policy sketches. In ICML, 2017.")] and priors [[Duan et al., 2016](https://arxiv.org/abs/1611.02779 "Y. Duan, J. Schulman, X. Chen, P. L. Bartlett, I. Sutskever, and P. Abbeel. RL2: Fast reinforcement learning via slow reinforcement learning. arXiv preprint arXiv:1611.02779, 2016."), [Wang et al., 2016](https://arxiv.org/abs/1611.05763 "J. X. Wang, Z. Kurth-Nelson, D. Tirumala, H. Soyer, J. Z. Leibo, R. Munos, C. Blundell, D. Kumaran, and M. Botvinick. Learning to reinforcement learn. arXiv preprint arXiv:1611.05763, 2016."), [Finn et al., 2017](https://arxiv.org/abs/1703.03400 "C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In ICML, 2017.")]. In response to this trend, the ICLR 2019 workshop on "Structure & Priors in Reinforcement Learning" (SPiRL) aims to revitalize a multi-disciplinary approach to investigating the role of structure and priors as a way of specifying inductive bias in RL.

Beyond machine learning, other disciplines such as neuroscience and cognitive science have traditionally played, or have the potential to play, a role in identifying useful structure [[Botvinick et al., 2009](https://www.ncbi.nlm.nih.gov/pubmed/18926527 "M. M. Botvinick, Y. Niv, and A. C. Barto. Hierarchically organized behavior and its neural foundations: a reinforcement learning perspec tive. Cognition, 113(3):262–280, 2009."), [Boureau et al., 2015](https://www.ncbi.nlm.nih.gov/pubmed/26483151 "Y.-L. Boureau, P. Sokol-Hessner, and N. D. Daw. Deciding how to decide: self-control and meta-decision making. Trends in cognitive sciences, 19(11):700–710, 2015.")] and priors [[Trommershauser et al., 2008](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2678412/ "J. Trommershauser, L. T. Maloney, and M. S. Landy. Decision making, movement planning and statistical decision theory. Trends in cognitive sciences, 12(8):291–297, 2008."), [Gershman and Niv, 2015](https://www.ncbi.nlm.nih.gov/pubmed/25808176 "S. J. Gershman and Y. Niv. Novelty and inductive generalization in human reinforcement learning. Topics in cognitive science, 7(3):391–415, 2015."), [Dubey et al., 2018](https://arxiv.org/abs/1802.10217 "R. Dubey et al., Investigating Human Priors for Playing Video Games. In ICML, 2018.")] for use in RL. As such, we expect attendees to be from a broad variety of backgrounds (including RL and machine learning, Bayesian methods, cognitive science and neuroscience), which would be beneficial for the (re-)discovery of commonalities and under-explored research directions.

Schedule

09:45 AM	**Opening remarks**	
09:50 AM	**TBA**	*Abbeel*
10:20 AM	**Efficient off-policy meta-reinforcement learning via probabilistic context variables**	*Rakelly*
10:30 AM	**Poster Session #1**	*Ghasemi*
11:00 AM	**TBA**	*Botvinick*
11:30 AM	**TBA**	*Hofmann*
12:00 PM	**TBA**	*Kulkarni*
12:30 PM	**TBA**	*Lillicrap*
03:20 PM	**TBA**	*Narasimhan*
03:50 PM	**State marginal matching with mixtures of policies / Reinforcement learning with unknown reward functions**	*Eysenbach*
04:00 PM	**Poster Session #2**	
04:30 PM	**TBA**	
05:00 PM	**TBA**	*Wang*
05:30 PM	**TBA**	

AI for Social Good

Margaux Luck, Tristan Sylvain, Kris Sankaran, Sean McGregor, Jonnie Penn, Virgile Sylvain, Geneviève Boucher, Myriam Cote, Kentaro Toyama, Rayid Ghani, Yoshua Bengio

Room R05, Mon May 06, 09:45 AM

#AI for Social Good
##Important information
Contact information: [aisg2019.iclr.contact@gmail.com](mailto:aisg2019.iclr.contact@gmail.co

Submission deadline: ***EXTENDED*** to March 22nd 2019 11:59PM ET

[Workshop website](https://aiforsocialgood.github.io/iclr2019/index.htm)

[Submission website](https://cmt3.research.microsoft.com/ICLRAISGW2019/)

Poster Information:

* Poster Size - ** 36W x 48H inches or 90 x 122 cm**

* Poster Paper - **lightweight paper - not laminated**

##Abstract
Our workshop “AI for Social Good” will focus on applying artificial intelligence to solve problems important for society. The focus is on machine learning for the following areas: health, education, protecting democracy, urban planning, assistive technology for people with disabilities, agriculture, environmental sustainability, social welfare and justice and, sustainable development. We believe that these fields are those where AI can have its strongest impact on society by reducing human suffering and improving democratic institutions. This workshop builds on our [AI for Social Good](https://aiforsocialgood.github.io/2018/) workshop at NeurIPS 2018.

If managed correctly, the rapidly expanding field of AI has the potential to improve many aspects of our lives. However, two main problems arise when attempting to tackle social issues. First, there is often little incentive for researchers to tackle social problems as there are few conferences and

journals that explicitly deal with such issues. Second, it is also difficult for researchers seeking to have a social impact to find problems to address. The convening of this workshop addresses these problems by networking impactful researchers and providing a venue for presentation.

This workshop brings together machine learning researchers, social impact leaders, stakeholders, government and policy leaders, and philanthropists to present and discuss ideas and applications linked to social issues, similarly to the [AI Commons](http://www.aicommons.com) project. We are partnering with AI Commons so that accepted proposals are invited to submit their work there. Moreover, the workshop inspires the creation of new tools by the community to tackle critical problems. We also wish to promote the sharing of information and datasets that might prove relevant to researchers who share our goals.

We invite contributions relating to at least one of the previously mentioned domains. The models or approaches presented do not necessarily need to be of outstanding theoretical novelty, but should demonstrate potential for a strong social impact. We especially encourage work where machine learning and in particular representation learning could meaningfully amplify existing efforts for social good. We invite two types of submissions:

* **Short Papers Track (Up to four page papers + unlimited pages for citations)** for oral and/or poster presentation. The short papers should focus on past and current research work, showcasing actual results and demonstrating beneficial effects on society. We also accept short papers of recently published or submitted journal contributions to give authors the opportunity to present their work and obtain feedback from conference attendees.

* **Problem Introduction Track (Application form, up to five page responses + unlimited pages for citations)** which will present a specific solution that will be shared with stakeholders, scientists, and funders. The workshop will provide a suite of questions designed to: (1) estimate the feasibility and impact of the proposed solutions, and (2) estimate the importance of data in their implementation. The application responses should highlight ideas that have not yet been implemented in practice but can lead to real impact. The projects may be at varying degrees of development, from formulation as a data problem to structure for effective deployment. The workshop provides a supportive platform for developing these early-stage or hobby proposals into real projects. This process is designed to foster sharing different points of view ranging from the scientific assessment of feasibility, discussion of practical constraints that may be encountered, and attracting interest from philanthropists invited to the event. Accepted submissions may be promoted to the wider AI solutions community following the workshop via the [AI Commons](http://www.aicommons.com), with whom we are partnering to promote the longer-term development of projects.

Schedule

11:00 AM	**Poster session**	*Tufa*

Safe Machine Learning: Specification, Robustness, and Assurance

Silvia Chiappa, Victoria Krakovna, Adrià Garriga-Alonso, Andrew Trask, Jonathan Uesato, Christina Heinze-Deml, Ray Jiang, Adrian Weller

Room R06, Mon May 06, 09:45 AM

ABSTRACT
The ultimate goal of ML research should be to have a positive impact on society and the world. As the number of applications of ML increases, it becomes more important to address a variety of safety issues; both those that already arise with today's ML systems and those that may be exacerbated in the future with more advanced systems.

Current ML algorithms tend to be brittle and opaque, reflect undesired bias in the data and often optimize for objectives that are misaligned with human preferences. We can expect many of these issues to get worse as our systems become more advanced (e.g. finding more clever ways to optimize for a misspecified objective). This workshop aims to bring together researchers in diverse areas such as reinforcement learning, formal verification, value alignment, fairness, and security to further the field of safety in machine learning.

We will focus on three broad categories of ML safety problems: [specification, robustness and assurance](https://medium.com/@deepmindsafetyresearch/building-safe (Ortega et al, 2018). Specification is defining the purpose of the system, robustness is designing the system to withstand perturbations, and assurance is monitoring, understanding and controlling system activity before and during its operation. Research areas within each category include:

Specification
- **Reward Hacking**: Systems may behave in ways unintended by the designers, because of discrepancies between the specified reward and the true intended reward. How can we design systems that don't exploit these misspecifications, or figure out where they are? (Over 40 examples of specification gaming by AI systems can be found here: http://tinyurl.com/specification-gaming.)
- **Side effects**: How can we give artificial agents an incentive to avoid unnecessary disruptions to their environment while pursuing the given objective? Can we do this in a way that generalizes across environments and tasks and does not introduce bad incentives for the agent in the process?
- **Fairness**: ML is increasingly used in core societal domains such as health care, hiring, lending, and criminal risk assessment. How can we make sure that historical prejudices, cultural stereotypes, and existing demographic inequalities contained in the data, as well as sampling bias and collection issues, are not reflected in the systems?

Robustness
- **Adaptation**: How can machine learning systems detect and adapt to changes in their environment (e.g. low overlap between train and test distributions, poor initial model assumptions, or shifts in the underlying prediction function)? How should an autonomous agent act when confronting radically new contexts, or identify that the context is new in the first place?
- **Verification**: How can we scalably verify meaningful properties of ML systems? What role can and should verification play in ensuring robustness of ML systems?
- **Worst-case robustness**: How can we train systems which never perform extremely poorly, even in the worst case? Given a trained system, can we ensure it never fails catastrophically, or bound this probability?
- **Safe exploration**: Can we design reinforcement learning algorithms which never fail catastrophically, even at training time?

Assurance
- **Interpretability**: How can we robustly determine whether a system is working as intended (i.e. is well specified and robust) before large-scale deployment, even when we do not have a formal specification of what it should do?
- **Monitoring**: How can we monitor large-scale systems to identify whether they are performing well? What tools can help diagnose and fix the found issues?
- **Privacy**: How can we ensure that the trained systems do not reveal sensitive information about individuals contained in the training set?
- **Interruptibility**: An artificial agent may learn to avoid interruptions by the human supervisor if such interruptions lead to receiving less reward. How can we ensure the system behaves safely even under the possibility of shutdown?

Schedule

09:50 AM	**Opening remarks**	
10:00 AM	**Cynthia Rudin**	*Rudin*
10:30 AM	**Posters and Coffee Break 1**	
11:30 AM	**Dylan Hadfield-Menell**	
12:00 PM	**Misleading meta-objectives and hidden incentives for distributional shift**	*Krueger*
12:20 PM	**Panel: Exploring overlaps and interactions between AI safety research areas**	*Olsson, Lakkaraju*
01:10 PM	**Lunch break**	
03:20 PM	**Bridging Adversarial Robustness and Gradient Interpretability**	*Kim*
03:40 PM	**Uncovering Surprising Behaviors in Reinforcement Learning via Worst-Case Analysis**	*Ruderman*
04:00 PM	**Posters and Coffee Break 2**	
05:00 PM	**Ian Goodfellow**	*Goodfellow*
05:30 PM	**Panel: Research priorities in AI safety**	*Goodfellow, Shah*

Abstracts (3):

Abstract 5: **Misleading meta-objectives and hidden incentives for distributional shift in Safe Machine Learning: Specification, Robustness, and Assurance**, *Krueger* 12:00 PM

David Krueger, Tegan Maharaj, Shane Legg and Jan Leike.

Decisions made by machine learning systems have a tremendous influence on the world. Yet it is common for machine learning algorithms to assume that no such influence exists. An example is the use of the i.i.d. assumption in online learning for applications such as content recommendation, where the (choice of) content displayed can change users' perceptions and preferences, or even drive them away, causing a shift in the distribution of users. A large body of work in reinforcement learning and causal machine learning aims to account for distributional shift caused by deploying a learning system previously trained offline. Our goal is similar, but distinct: we point out that online training with meta-learning can create a hidden incentive for a learner to cause distributional shift. We design a simple environment to test for these hidden incentives (HIDS), demonstrate the potential for this phenomenon to cause unexpected or undesirable behavior, and propose and validate a mitigation strategy.

Abstract 8: **Bridging Adversarial Robustness and Gradient Interpretability in Safe Machine Learning: Specification, Robustness, and Assurance**, *Kim* 03:20 PM

Beomsu Kim, Junghoon Seo and Taegyun Jeon.

Adversarial training is a training scheme designed to counter adversarial attacks by augmenting the training dataset with adversarial examples. Surprisingly, several studies have observed that loss gradients from adversarially trained DNNs are visually more interpretable than those from standard DNNs. Although this phenomenon is interesting, there are only few works that have offered an explanation. In this paper, we attempted to bridge this gap between adversarial robustness and gradient interpretability. To this end, we identified that loss gradients from adversarially trained DNNs align better with human perception because adversarial training restricts gradients closer to the image manifold. We then demonstrated adversarial training causes loss gradients to be quantitatively meaningful. Finally, we showed that under the adversarial training framework, there exists an empirical trade-off between test accuracy and loss gradient interpretability and proposed two potential approaches to resolving this trade-off.

Abstract 9: **Uncovering Surprising Behaviors in Reinforcement Learning via Worst-Case Analysis in Safe Machine Learning: Specification, Robustness, and Assurance**, *Ruderman* 03:40 PM

Avraham Ruderman, Richard Everett, Bristy Sikder, Hubert Soyer, Charles Beattie, Jonathan Uesato, Ananya Kumar and Pushmeet Kohli

Reinforcement learning agents are typically trained and evaluated according to their performance averaged over some distribution of environment settings. But does the distribution over environment settings contain important biases, and do these lead to agents that fail in certain cases despite high average-case performance? In this work, we consider worst-case analysis of agents over environment settings in order to detect whether there are directions in which agents may have failed to generalize. Specifically, we consider a 3D first-person task where agents must navigate procedurally generated mazes, and where reinforcement learning agents have recently achieved human-level average-case performance. By optimizing over the structure of mazes, we find that agents can suffer from catastrophic failures, failing to find the goal even on surprisingly simple mazes, despite their impressive average-case performance. Additionally, we find that these failures transfer between different agents and even significantly different architectures. We believe our findings highlight an important role for worst-case analysis in identifying whether there are directions in which agents have failed to generalize. Our hope is that the ability to automatically identify failures of generalization will facilitate development of more general and robust agents.

Representation Learning on Graphs and Manifolds

Will Hamilton, Fred Sala, Peter Battaglia, Joan Bruna, Thomas Kipf, Yujia Li, Razvan Pascanu, Adriana Romero, Petar Veli■kovi■, Marinka Zitnik, Maximilian Nickel, Beliz Gunel, Albert Gu, Christopher Re

Room R07, Mon May 06, 09:45 AM

Many scientific fields study data with an underlying graph or manifold structure—such as social networks, sensor networks, biomedical knowledge graphs, and meshed surfaces in computer graphics. The need for new optimization methods and neural network architectures that can accommodate these relational and non-Euclidean structures is becoming increasingly clear. In parallel, there is a growing interest in how we can leverage insights from these domains to incorporate new kinds of relational and non-Euclidean inductive biases into deep learning.

Recent years have seen a surge in research on these problems—often under the umbrella terms of graph representation learning and geometric deep learning. For instance, new neural network architectures for graph-structured data (i.e., graph neural networks) have led to state-of-the-art results in numerous tasks—ranging from molecule classification to recommender systems—while advancements in embedding data in Riemannian manifolds (e.g., Poincaré embeddings, Hyperspherical-VAEs) and optimization on Riemannian manifolds (e.g., R-SGD, R-SVRG) have demonstrated how non-Euclidean geometries can provide powerful new kinds of inductive biases.

Perhaps the biggest testament to the increasing popularity of this area is the fact that five popular review papers have recently been published on the topic [1-5]—each attempting to unify different formulations of similar ideas across fields. This suggests that the topic has reached critical mass and requires a focused workshop to bring together researchers to identify impactful areas of interest, discuss how we can design new and better benchmarks, encourage discussion, and foster collaboration.

The workshop will consist of contributed talks, contributed posters, and invited talks on a wide variety of methods and problems in this area, including but not limited to:
- Deep learning on graphs and manifolds (e.g., graph neural networks)
- Riemannian optimization methods
- Interaction and relational networks
- Unsupervised geometric/graph embedding methods (e.g., hyperbolic embeddings)
- Generative models with manifold-valued latent variables
- Deep generative models of graphs
- Deep learning for chemical/drug design
- Deep learning on manifolds, point clouds, and for 3D vision
- Relational inductive biases (e.g., for reinforcement learning)
- Optimization challenges due to the inherent discreteness of graphs
- Theoretical analyses of graph-based and non-Euclidean machine learning approaches
- Benchmark datasets and evaluation methods

We welcome and encourage position papers under this workshop theme. We are also particularly interested in papers that introduce benchmark datasets, challenges, and competitions to further progress of the field, and we will discuss the challenge of designing such a benchmark in an interactive panel discussion.

[1] Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., & Vandergheynst, P. (2017). Geometric deep learning: going beyond euclidean data. IEEE Signal Processing Magazine, 34(4), 18-42.
[2] Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Representation learning on graphs: Methods and applications. IEEE Data Engineering Bulletin.
[3] Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., ... & Gulcehre, C. (2018). Relational inductive biases, deep learning, and graph networks. arXiv preprint arXiv:1806.01261.
[4] Goyal, P., & Ferrara, E. (2018). Graph embedding techniques, applications, and performance: A survey. Knowledge-Based Systems, 151, 78-94.
[5] Nickel, M., Murphy, K., Tresp, V., Gabrilovich, E. (2016). A review of relational machine learning for knowledge graphs. Proceedings of the IEEE. 104.1, 11-33.

Reproducibility in Machine Learning

Nan Rosemary Ke, Alex Lamb, Anirudh Goyal Alias Parth Goyal, Aaron Courville, Yoshua Bengio

Room R08, Mon May 06, 09:45 AM

Papers from the Machine Learning community are supposed to be a valuable asset. They can help to inform and inspire future research. They can be a useful educational tool for students. They are the driving force of innovation and differentiation in the industry, so quick and accurate implementation is really critical. On the research side they can help us answer the most fundamental questions about our existence - what does it mean to learn and what does it mean to be human? Reproducibility, while not always possible in science (consider the study of a transient astrological phenomenon like a passing comet), is a powerful criteria for improving the quality of research. A result which is reproducible is more likely to be robust and meaningful and rules out many types of experimenter error (either fraud or accidental). There are many interesting open questions about how reproducibility issues intersect with the Machine Learning community:

-How can we tell if papers in the Machine Learning community are reproducible even in theory? If a paper is about recommending news sites before a particular election, and the results come from running the system online in production - it will be impossible to reproduce the published results because the state of the world is irreversibly changed from when the experiment was run.

-What does it mean for a paper to be reproducible in theory but not in practice? For example, if a paper requires tens of thousands of GPUs to reproduce or a large closed-off dataset, then it can only be reproduced in reality by a few large labs.

-For papers which are reproducible both in theory and in practice - how can we ensure that papers published in ICML would actually be able to replicate if such an experiment were attempted?
What is the best way of publishing the code of the papers so that it is easy for engineers to implement it? Just publishing ipython notebooks it is not sufficient and often hard to make it work in different platforms

-A lot of people tend to understand an algorithm by looking at code and not by following equations. How can we come up with a framework of publishing that includes them. Is pseudocode the best we can do?

-While scientific papers often do an importance analysis of the features, ML papers rarely do proper attribution on the importance of algorithmic components and hyperparameters. What is the best way to “unit-test” an algorithm and do attribution of the results to certain components and hyperparameters

-What does it mean for a paper to have successful or unsuccessful replications?

-Of the papers with attempted replications completed, how many have been published?

-What can be done to ensure that as many papers which are reproducible in theory fall into the last category?

-On the reproducibility issue, what can the Machine Learning community learn from other fields?

-Part of ensuring reproducibility of state-of-the-art is ensuring fair comparisons, proper experimental procedures, and proper evaluation methods and metrics. To this end, what are the proper guidelines for such aspects of machine learning problems? How do they differ among subsets of machine learning?

Our aim in the following workshop is to raise the profile of these questions in the community and to search for their answers. In doing so we aim for papers focusing on the following topics:

-Analysis of the current state of reproducibility in machine learning. Some examples of this include experimental-driven investigations as in [1,2,3]

-Investigations and proposals of proper experimental procedure and evaluation methodologies which ensure reproducible and fair comparisons in novel literature [4]

-Tools to help improve reproducibility

-Evidence-driven works investigating the importance of reproducibility in machine learning and science in general

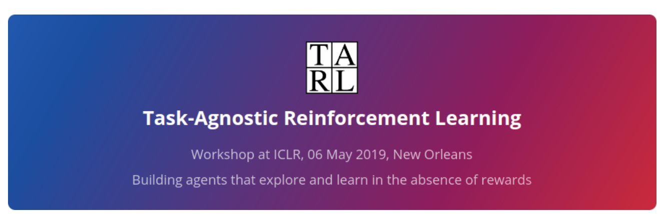
-Connections between the reproducibility situation in Machine Learning and other fields

-Rigorous replications, both failed and successful, of influential papers in the Machine Learning literature.

Task-Agnostic Reinforcement Learning (TARL)

Danijar Hafner, Amy Zhang, Ahmed Touati, Deepak Pathak, Frederik Ebert, Rowan McAllister, Roberto Calandra, Marc G Bellemare, Raia Hadsell, Joelle Pineau

Room R09, Mon May 06, 09:45 AM



Workshop website: https://tarl2019.github.io/
Start a submission:
https://cmt3.research.microsoft.com/TARL2019

Contact the organizers: taskagnosticrl@gmail.com

Summary

Many of the successes in deep learning build upon rich supervision. Reinforcement learning (RL) is no exception to this: algorithms for locomotion, manipulation, and game playing often rely on carefully crafted reward functions that guide the agent. But defining dense rewards becomes impractical for complex tasks. Moreover, attempts to do so frequently result in agents exploiting human error in the specification. To scale RL to the next level of difficulty, agents will have to learn autonomously in the absence of rewards.

We define task-agnostic reinforcement learning (TARL) as learning in an environment without rewards to later quickly solve down-steam tasks. Active research questions in TARL include designing objectives for intrinsic motivation and exploration, learning unsupervised task or goal spaces, global exploration, learning world models, and unsupervised skill discovery. The main goal of this workshop is to bring together researchers in RL and investigate novel directions to learning task-agnostic representations with the objective of advancing the field towards more scalable and effective solutions in RL.

We invite paper submissions in the following categories to present at the workshop:
- Unsupervised objectives for agents
- Curiosity and intrinsic motivation
- Few shot reinforcement learning
- Model-based planning and exploration
- Representation learning for planning
- Learning unsupervised goal spaces
- Automated curriculum generation
- Unsupervised skill discovery
- Evaluation of unsupervised agents

Submissions

Papers should be in anonymous ICLR style and up to 5 pages, with an unlimited number of pages for references and appendix. Accepted papers will be made available on the workshop website and selected submissions will be offered a 15 minute talk at the workshop. This does not constitute an archival publication and no formal workshop proceedings will be made available, meaning contributors are free to publish their work at journals or conferences.

Schedule

09:50 AM	**Martin Riedmiller**
10:20 AM	**Lightning Talks 1**
10:30 AM	**Poster Session 1**
11:00 AM	**Chelsea Finn**
11:30 AM	**Doina Precup**
12:00 PM	**Contributed Talk 1**
12:15 PM	**Contributed Talk 2**
12:30 PM	**Katja Hofmann**
03:20 PM	**Contributed Talk 3**
03:35 PM	**Contributed Talk 4**
03:50 PM	**Lightning Talks 2**
04:00 PM	**Poster Session 2**
04:30 PM	**Pierre-Yves Oudeyer**
05:30 PM	**Neil Bramley**
06:00 PM	**Panel Discussion**

May 7, 2019