Unified Representations for Learning and Reasoning

Yitao Liang (yliang@cs.ucla.edu)

The recent decade has witnessed unprecedented success of deep learning in various high-profile challenges, e.g., the ImageNet recognition task, Atari and the board game Go, human-like text generation, and protein structure prediction, etc. Encouraged by those remarkable empirical achievements, both the research and industry communities cannot wait but further expand AI's impact in other scenarios. Einstein once famously said "the important thing is not to stop questioning". However, embracing the recent success, many have started paying less attention to what has exactly been learned and *what we can do with what has been learned*. Before the practice of "more data, more computation units" runs out of fuel, it is never too late to re-contemplate what is missing from current representations. My research focuses on investigating three such aspects, and proposing promising solutions to address them.

- Flexibility to reason on itself. Many popular machine learning models learn functions; e.g., classifiers are functions from features to designated classes. To increase the capabilities of classifiers, we often want to learn a world model (i.e., a distribution over features). Also known as generative models, world models permit sampling from its learned distribution. However, to better understand data and answer scientific questions, one may want to *reason* about instead of sampling from the learned distribution. For this, we propose tractable probabilistic representations; they can exactly answer queries that are hard to sample: e.g., probabilities of events, most likely events, expectations and higher-order moments conditioned on sub-population, etc. Surprisingly, by training these novel representations as deep computation graphs, tractable reasoning can be achieved without sacrificing model fit. Given their parameters are conditional probabilities factoring into the overall distribution, our proposed representations are also inherently *interpretable*.
- Support of reasoning with existing models. As previously mentioned, many existing models lack the probabilistic reasoning ability, yet have been widely adopted. Leveraging the tractability of our proposed representations and their inherent semantic linkage to the existing ones (e.g. logistic regression, decision trees, etc.), we enable reasoning of their predictions with respect to the feature distribution captured by our representations. This increases existing models' robustness in the presence of missing features. In fact, same as robustness, many valuable model properties can be approached through reasoning. For example, one can reason about what subsets of features retain the model's predictions in expectation, constructing *sufficient explanations*; *fairness can be verified* by reasoning about expected prediction difference across sensitive features.
- Knowledge incorporation. The value of data is well appreciated. However, some task-critical information could be outside pure data; e.g., constraints between features and labels, affordance of objects, etc. Same as data, this kind of knowledge is also ubiquitous and could significantly affect our decisions on certain tasks. We have developed effective *embeddings of knowledge* such that representations can leverage them during learning.

There is another fundamental theme behind these areas — *user-friendliness*. For AI to be welcomed by a larger crowd, we need to instill confidence in users, which can only be achieved through positive feedback between users and AI systems. My research can enable such feedback. Elaborating on the points above, for example users can ask the learned representations to back up their answers through reasoning about related tasks; users can teach representations with what they know; fairness can be more easily measured and expected through reasoning about expected results, etc. I possess a strong belief that all those great properties should not be given away when pursuing better predictive performance. We should not be asked to make a tradeoff between them. *My long-term research plan is to develop competitive unified representations with interpretable semantics for learning and reasoning* such that we can fully enjoy all the convenience of AI advancements.

Statistical Representations Capable of Symbolic Reasoning

Many perceive that robust reasoning and competitive learning performance sit on the two opposite ends of a representation's tradeoff spectrum. It appears that the two, though both are desirable, could not co-exist in one single representation. This dilemma would always be believed to be true unless someone demonstrates otherwise. One approach to shrink the distance between the two ends is to bring the performance of representations with reasoning power to the ballpark of leading deep learners.

Generative learning Probabilistic circuits are deep generative models that support tractable exact probabilistic reasoning, which means assorted reasoning tasks (e.g., marginals, most likely events, distribution multiplication, KL-divergence [13], etc.) can be answered in linear time with respect to the size of the representation. Those queries are typically computationally prohibiting for standard neural networks, as correctly computing the partition function to normalize probabilities would require enumerating exponentially many possible configurations of events. On the other hand, probabilistic circuits' tractability in reasoning is guaranteed by certain syntactic properties. We have built the first search-based structure learning algorithm for probabilistic circuits, with guarantees of monotone performance improvements [11]. Its final performance compares favorably to not just tractable but intractable representations as well.

Discriminative learning Standard neural network's performance lead in discriminative learning tasks (e.g., classification) on image domains is staggering. Tractable representations have had a hard time of catching up. Inspired by the counterpart relationship between naive Bayes and logistic regression, I have proposed the first discriminative learning algorithm on probabilistic circuits and proved the discriminative parameters are equivalent to log-odds of parameters associated with generative learning [14]. By associating features only with the representation's structure, I have successfully demonstrated that its parameter learning is convex optimization, and that a simple local search can induce strong structures from data. The resulting performance is on par with convolutional neural networks on standard benchmark datasets, with a clear advantage in the learned representation size and data efficiency. Some syntactic restrictions can be relaxed to readily incorporate convolutions and parameter-sharing, which further boosts our proposed circuit representation's classification performance [15].

Interpretability Considering probabilistic circuits' inheritance from symbolic knowledge representation, it is fairly straightforward to interpret the learned representations. Without parameters, each node of a probabilistic circuit recursively encodes a logical sentence. In the context of generative learning, parameters are conditional probabilities of logical sentences; for discriminative learning, parameters are probabilities of the logical sentence being true with respect to the input data.

Open-source Efforts To facilitate collaboration and accelerate adoption of our proposed representations in other fields, my colleagues and I have built an open-source library that contains all our past research efforts [1]. The library is optimized to fully leverage modern parallel computing and GPU architecture. Furthermore, our proposed structure learning algorithms have been used by colleagues in the field of computer architecture to benchmark the throughput of custom processors for probabilistic reasoning kernels [9]. I also look forward to working with researchers and experts in system machine learning to further speed up learning of probabilistic circuits.

Representation Coordination for Statistical Reasoning

Many existing representations have asserted their prominence in certain scenarios. It is unrealistic to ask users to abandon the representations that have been learned. Thanks to their tractability, our proposed probabilistic circuits can serve as an intermediary to help certain existing discriminative representations to reason their prediction results with respect to a distribution.

Robustness to uncertainties Real-world data is full of noise, could even be partially missing. When faced with uncertainties, most existing methods impute possible values to fill in, which requires making strong assumptions of the data distribution; for example, being independent and identically distributed is typically assumed, yet this could be hard to satisfy in realistic scenarios. Fortunately, when we possess a pair of generative and discriminative representations with counterpart relations to one another, we could repudiate settling down on imputation and its associated assumptions. Rather we could pursue a more principled holistic solution that considers all possible situations of missingness. Essentially, we can attempt computing the expectation of the discriminative representation's prediction with respect to missing features following the probability distribution captured by the generative representation. My colleagues and I have proved the necessary conditions when such computation is efficient on our proposed circuit representation [3] and other representations (e.g., logistic regression [5] and decision tress [4], etc.). Empirical evaluations suggest that our methods retain existing models' performance with all features observed and outperform standard imputation techniques when features do go missing. We further move beyond standard supervised learning and demonstrate reasoning's effectiveness in stabilizing rollout error estimates in model-based reinforcement learning [8].

Sufficient explanation On a high level, the task of explaining a particular model can be thought of as quantifying which subset of the features contribute the most to the model's prediction. Many existing methods on this assume a black-box setting, and to evaluate the importance of each feature they need to compare the predictor across multiple perturbations of the given instance. Instead, as an extension of our efficient reasoning framework, we leverage the feature distribution to produce explanations and show reasoning helps make this process straightforward. We consider the smallest subset of features that are sufficient to get the same model prediction, in expectation with respect to the feature distribution [5]. Intuitively, this is the smallest set of support features that, in expectation, result in the same prediction result despite all the evidence to the contrary. In other words, we explain a model's prediction using the strongest evidence towards it.

Fairness verification Discrimination associated with the learned representations can be easily concealed when each prediction, classification or decision is considered individually and separately. However, by reasoning with respect to a complete distribution, whether a representation misbehaves can be principally measured and becomes obvious to users. In particular, we reason about the expected prediction conditioned on withholding a certain sensitive attribute and compare the expected prediction with the prediction using all features (including the sensitive attribute). A significant difference between the two points to the possibility that the learned representations have a tendency to discriminate against that sensitive attribute. A closer scrutiny on different sub-populations could be further conducted by conditioning on multiple attributes at the same time [3].

Neural Representations with Knowledge Embedding

Though the syntax and semantics of neural representations have pre-determined their incompatibility with many reasoning tasks, a good starting point to further empower them could be the ability to comfortably incorporate existing knowledge. This would be critical to extend neural representations' success to structured domains, where the structure is typically not captured by pure data.

Formulable knowledge Some domain structure can be well formulated as symbolic knowledge, such as one-to-one correspondence in recommendations, hierarchical relations in a tree, and validity as a path of several edges, etc. These can be well summarized as logical sentences and can be further compiled into logical circuits. Our proposed probabilistic circuits are naturally compatible with the compiled logical circuits, as they share many syntactic properties [11]. However, for neural representation, this can be problematic: symbolic knowledge is discrete by its nature whereas neural representations are continuous and smooth. To reconcile this fundamental conflict, we have proposed a semantic loss

that grants a probabilistic interpretation to logical circuits' evaluation, i.e., how likely the provided vector is to satisfy the encoded logical sentence considering each element in the vector as a Bernoulli distribution [10]. This creates a seamless linkage between the neural representations' outputs and the logical circuits' inputs. We further prove this function's uniqueness and monotonicity, setting up theoretical guarantees that it can indeed function as a loss function for deep learning.

Informulable knowledge Some complicated structure and affordance cannot be easily formulated. A notable example would be semantic segmentation in a visual input. High-quality semantic segmentation, though expensive to acquire, has proven to be extremely useful in many vision tasks. One can usually only expect to have it in limited amount or during limited time. To leverage this kind of side-task knowledge to improve autonomous driving control, my colleagues and I have proposed to squeeze it into an embedding, and then learn to mimick this embedding such that the representations continue to work well even when the side-task knowledge is unaccessible [16]. On a related note, what has done above is one example of coordinating representations in a two-task system. The most famous two-task situation would be the classic "exploration vs. exploitation", that is relevant to every control problem. My colleagues and I have proposed disentangled representations with mutual restraints to best fulfill respective objectives while mitigating biased updates to one another [7].

Other Representations & Reinforcement Learning Research

It is also not hard to notice representations tend to get larger and more complicated to achieve better performance, which inadvertently increases the difficulty of decoding what specifically have been learned. However, for certain tasks, the specific learned representation values do not matter as much as the inductive bias induced by the representation. We have shown with careful designs, important inductive bias can be reproduced by simple representations (e.g. linear function approximation) [12]. This could be especially true for reinforcement learning and control problems where specific utility values of individual actions are less important than the relative advantages between actions. I am also generally passionate about all aspects of reinforcement learning. My colleagues and I are the first to theoretically establish that for parallel Monte Carlo tree search to have comparable performance with its sequential counterpart, its visit counts must be no smaller than the sum of complete and ongoing simulations [6]. We have also successfully applied reinforcement learning to crowdsourcing, leading to better-quality data labels compared to if treating the whole process as one-shot [2].

Future Outlook

The research agenda presented above naturally leads to assorted immediate and long-term important problems to investigate. To name a few: (i) extension of representations' reasoning ability to sequential settings; (ii) scale of knowledge embeddings to more structured domains (e.g., over graphs); (iii) pursuit of interdisciplinary applications (e.g. representations for biological data that is prone to corruption). The need for learning competitive representations with reasoning capacity will only increase, as AI moves towards more challenging and realistic use cases that require flexibility and robustness.

References

[1] Meihua Dang, Pasha Khosravi, **Yitao Liang**, Antonio Vergari, and Guy Van den Broeck. Juice: A julia package for logic and probabilistic circuits. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (Demo Track)*, 2021.

[2] Zehong Hu, **Yitao Liang**, Jie Zhang, Zhao Li, and Yang Liu. Inference aided reinforcement learning for incentive mechanism design in crowdsourcing. In *Advances in Neural Information Processing Systems 31*, 2018.

[3] Pasha Khosravi, YooJung Choi, **Yitao Liang**, Antonio Vergari, and Guy Van den Broeck. On tractable computation of expected predictions. In *Advances in Neural Information Processing Systems 32*, 2019.

[4] Pasha Khosravi, Antonio Vergari, YooJung Choi, **Yitao Liang**, and Guy Van den Broeck. Handling missing data in decision trees: A probabilistic approach. In *The Art of Learning with Missing Values Workshop at ICML (Artemiss)*, 2020.

[5] Pasha Khosravi, **Yitao Liang**, YooJung Choi, and Guy Van den Broeck. What to expect of classifiers? reasoning about logistic regression with missing features. In *The 28th International Joint Conference on Artificial Intelligence (IJCAI)*, 2019.

[6] Anji Liu, **Yitao Liang**, Ji Liu, Guy Van den Broeck, and Jianshu Chen. On effective parallelization of monte carlo tree search. In *Preprint*, 2020.

[7] Anji Liu, **Yitao Liang**, and Guy Van den Broeck. Off-policy deep reinforcement learning with analogous disentangled exploration. In *Proceedings of the 19th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 2020.

[8] Anji Liu, **Yitao Liang**, Guy Van den Broeck, and Erin Talvitie. Conditions for reliable error estimation in model-based reinforcement learning. In *Preprint*, 2020.

[9] Nimish Shah, Laura I. Galindez Olascoaga, Wannes Meert, and Marian Verhelst. Acceleration of probabilistic reasoning through custom processor architecture. In *2020 Design, Automation Test in Europe Conference Exhibition (DATE)*, 2020.

[10] Jingyi Xu, Zilu Zhang, Tal Friedman, **Yitao Liang**, and Guy Van den Broeck. A semantic loss function for deep learning with symbolic knowledge. In *The 35th International Conference on Machine Learning*, 2018.

[11] **Yitao Liang**, Jessa Bekker, and Guy Van den Broeck. Learning the structure of probabilistic sentential decision diagrams. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017.

[12] **Yitao Liang**, Marlos C. Machado, Erik Talvitie, and Michael Bowling. State of the art control of atari games using shallow reinforcement learning. In *The 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2016.

[13] **Yitao Liang** and Guy Van den Broeck. Towards compact interpretable models: Shrinking of learned probabilistic sentential decision diagrams. In *IJCAI 2017 Workshop on Explainable Artificial Intelligence*, 2017.

[14] **Yitao Liang** and Guy Van den Broeck. Learning logistic circuits. In *The Thirty-Third AAAI Conference on Artificial Intelligence*, 2019.

[15] **Yitao Liang** and Guy Van den Broeck. Learning convolutional logistic circuits. In *Preprint*, 2020.

[16] Albert Zhao, Tong He, **Yitao Liang**, Haibin Huang, Guy Van den Broeck, and Stefano Soatto. Sam: Squeeze-and-mimic networks for conditional visual driving policy learning. In *Conference on Robot Learning*, 2020.