Towards a Unified Framework for Learning and Reasoning

Han Zhao
Machine Learning Department
Carnegie Mellon University
han.zhao@cs.cmu.edu

Background: The recent decade has witnessed a phenomenal success in artificial intelligence. In particular, deep learning has gained an unprecedented impact across both research and industry communities by demonstrating better than human performance on various kinds of real-world competitions, e.g., the ImageNet recognition task [7], the Stanford question answering competition [9], the board game Go [12], etc. As a result, machine learning tools have been widely adopted to help decision making in various real-world scenarios, e.g., face recognition, machine translation, college admission, etc. While these empirical achievements are exciting, whether the learned model could deliver its promise in real-world scenarios crucially depends on the data used to train the model. However, often, it is computationally expensive, or sometimes infeasible, to collect labeled data under all the possible real-world scenarios. As a result, due to this distributional shift between the training and test data, the learned model may fail dramatically in practice. Perhaps more importantly, in high-stakes settings such as loan approvals, criminal justice and hiring process, if the data used to train the model contain historical bias, then the learned model, without bias mitigation, can only exacerbate the existing discrimination. Hence, the ability to learn representations that are invariant to the changes in the environment is crucial and can provide us the robustness against various noise and nuisance factors in real-world.

On the other hand, learning is only one of the two most fundamental cognitive abilities of intelligent agents. An intelligent agent needs to have the ability to learn from the experience, as well as the ability to reason from what has been learned. As foreseen by the Turing Award Laureate Prof. Leslie Valiant [13], one of the key challenges for AI in the coming decades is the development of integrated learning and reasoning mechanisms. However, classic symbolic reasoning cannot model the inherent uncertainty that ubiquitously exists, and it is not robust to noisy observations. Perhaps more fundamentally, inference and reasoning are computationally intractable in general. As a result, learning, which often takes inference/reasoning as a sub-procedure, is also hard.

Vision: I am interested in advancing the frontier of AI from the lens of representation learning and probabilistic reasoning. As shown in Figure 1, I believe an intelligent agent should consist of a knowledge base and an inference engine, and the agent interacts with the environment in loop. In each iteration, the agent receives input from the environment and uses its knowledge base to map the real-world observations or queries to its internal algebraic representations. The inference engine then carries out probabilistic reasoning to answer the query or choose an action to execute in the environment. My long-term research goal is to build a unified framework that provides a common semantics for learning and reasoning, by developing invariant representations that generalize across different environments as well as efficient inference engine that allows exact and tractable probabilistic reasoning.

Figure 1: A model of an intelligent agent, which consists of a knowledge base and an inference engine. Representation learning serves as a technique to help build our knowledge base that maps from real-world objects to their algebraic representations. The inference engine is powered by probabilistic reasoning that allows interactions between the agent and the environment.
Within the broad area of artificial intelligence and machine learning, my Ph.D. research primarily spans two themes: invariant representation learning and tractable probabilistic reasoning. Invariant representation learning serves as a bridge connecting abstract objects in real-world and their corresponding algebraic representations that are amenable for computation and allow generalization across different environments. Tractable probabilistic reasoning aims to provide an inference mechanism for exact and efficient reasoning under uncertainty. However, it is not well-understood what is the fundamental limit of invariant representations in terms of task utility, and it is well-known that even approximate probabilistic reasoning is computationally intractable [11] in the worst case.

Building on the fundamental concepts from information theory and theoretical computer science, my work aims to understand the inherent tradeoff between utility and invariance in learning the representations, and to develop efficient algorithms for learning tractable and exact probabilistic inference machines. The key contributions of my thesis research are as follows and summarized in Figure 2.

1. Analyzed and proved the fundamental limit of learning invariant representations in terms of task utility. With this result, we also identify and explain the inherent tradeoffs in learning domain-invariant representations for unsupervised domain adaptation [17], learning fair representations for algorithmic fairness [19], and learning representations for privacy-preservation under attribute-inference attacks [15].

2. Developed an algorithm on learning domain-invariant representations for unsupervised domain adaptation under multiple different source environments [23] using adversarial neural networks. To mitigate bias in automated decision making systems, my coauthors and I also proposed an algorithm to learn fair representations that can simultaneously guarantee accuracy parity and equalized odds [5] among different demographic groups. Analogously, we also developed an algorithm by learning invariant representations to filter out sensitive information and provided guarantees on inference error from malicious adversaries [15].

3. Established the first equivalence between Sum-Product networks [8], Bayesian networks with algebraic decision diagrams, and mixture models [21, 22]. Inspired by our theoretical results, we proposed efficient learning algorithms for Sum-Product networks in both offline [22] and online [10, 18], discrete [22] and continuous [6] settings, and from both frequentists’ and Bayesian principles [14].

Broadly, all the results above enjoy sound theoretical guarantees and provide insights towards better understanding of learning invariant representations and building tractable inference machines. On the practical side, I have also devoted much effort to develop software tools for the proposed algorithms and publicly share them with the community. For instance, MDAN [23] (for domain adaptation with multiple sources) has been successfully used as a benchmark algorithm in various follow-up work on unsupervised domain adaptation for vision [24] and language [3].

---

[1] https://github.com/KeiraZhao/MDAN
1 Ph.D. Research on Invariant Representation Learning

**Domain Adaptation:** The success of machine learning has been partially attributed to rich labeled datasets and powerful computations. Unfortunately, collecting and annotating such large-scale training data is prohibitively expensive and time-consuming. To solve these limitations, different labeled datasets can be combined to build a larger one. However, due to the potential distributional shift between different datasets, models trained on the combined one still suffer from large generalization error on a target domain different from the training domain.

My work on domain adaptation focuses on understanding the limit of knowledge transfer from a labeled source environment to an unlabeled target environment by learning domain-invariant representations to bridge the gap [1, 20, 23]. The main idea of domain-invariant learning is simple and intuitive: we would like to learn representations that are invariant to the difference among different environments while still contain rich information for the desired task. Theoretically, my research sheds new light on this problem by proving an information-theoretic lower bound on the joint error of any domain adaptation algorithm that attempts to learn invariant representations: **there is a fundamental tradeoff between learning domain-invariant representations and achieving small joint error in both environments when when the difference between environments can be used to explain the target task** [17]. Specifically, our result implies that **any algorithm based on domain-invariant representations has to incur a large error on at least one of the environments.** This result characterizes the fundamental limit in terms of the joint utility when learning with domain-invariant representations.

**Algorithmic Fairness:** With the prevalence of machine learning applications in high-stakes domains, e.g., criminal judgement, medical testing, online advertising, etc., it is crucial to ensure that the automated decision making systems do not propagate existing bias or discrimination that might exist in historical data used to train these systems. Among many recent proposals for achieving different notions of algorithmic fairness, learning fair representations has received increasing attention due to recent advances in learning rich representations with deep neural networks. At a high level the underlying idea is that if representations from different groups are similar to each other, then any predictive model on top of them will certainly make decisions independent of group membership.

On the other hand, while it has long been empirically observed [2] that there is an underlying tradeoff between utility and fairness, theoretical understanding is lacking. In my recent work [19], I provided the first theoretical result that characterizes the inherent tradeoff between utility and fairness. More precisely, our result shows that any fair algorithm, in the sense of demographic parity, admits an information-theoretic lower bound on the joint error across different demographic groups. To escape such inherent tradeoff, we also propose an alternative algorithm [16] to learn conditionally group-invariant representations. The proposed algorithm constructs a classifier that is no worse than the optimal classifier in terms of demographic parity gap, and can achieve equalized false positive/negative rates and accuracy parity across different demographic groups simultaneously.

**Privacy-Preserving Learning:** With the growing demand for machine learning systems provided as services, a massive amount of data containing sensitive information, such as income level, age, etc., are generated and collected from local users. This poses a substantial privacy challenge and it has become an imperative object of study. In my recent work [15], my coauthors and I consider a practical scenario where the prediction vendor requests crowdsourced data for a target task, e.g., scientific modeling. The data owner agrees on the data usage for the target task while she does not want her other private information (e.g., age, race) to be leaked. The goal of privacy-preserving in this context is then to protect private attributes of the sanitized data released by data owner from potential attribute inference attacks of a malicious adversary. Again, by learning representations that are invariant to the sensitive attribute, we formulate the problem of utility maximization with privacy constraint as a minimax optimization problem that can be effectively and practically implemented. Our algorithm enjoys an information-theoretic guarantee on the inference error of the protected attribute under attacks from arbitrary adversaries. We also provide a theoretical analysis to formally characterize the inherent tradeoff between utility maximization and privacy-preservation.

2 Ph.D. Research on Tractable Probabilistic Reasoning

**Sum-Product Networks:** Sum-Product networks (SPNs) are inference machines that admit exact probabilistic reasoning in linear time in the size of the network [8]. Because of its flexibility in modeling complex distributions and the tractability in performing exact inference, SPNs have been widely used in various application scenarios, e.g., image completion, language understanding, etc. The goal of my research work on SPNs is to theoretically study them as general inference machines, to design efficient structure and parameter learning algorithms for SPNs, and to apply SPNs as a powerful tool to various application domains that require the capability of reasoning. Together with my coauthors, I have built the first connection between SPNs, Bayesian networks and mixture models [21, 22]. We also provided a structural result showing that the representational power of SPN grows...
exponentially in its depth [22]. Inspired by our theoretical results, we proposed a series of efficient parameter learning algorithms for SPNs in both offline [22] and online [10, 18], discrete [22] and continuous [6] settings, and from both frequentists’ and Bayesian principles [14]. Two key techniques underpinning the design of our algorithms come from our theoretical insights: 1) an SPN can be equivalently reduced to a bipartite Bayesian network with one layer of latent variables pointing to one layer of observable variables; 2) the network polynomial computed by an SPN is a homogeneous polynomial in terms of the input boolean variables and a multi-linear function in terms of the network parameters. These observations allow us to efficiently evaluate a polynomial function with exponentially many terms via partial differentiation [3].

3 Future Research

My research contributes to two main themes in artificial intelligence: invariant representation learning and tractable probabilistic reasoning. Moving forward, I will continue working along these two themes towards the long-term goal of building a unified framework that provides a common semantics for learning and reasoning, and also branch out to explore applications related to algorithmic fairness and multilingual natural language understanding.

Information Analysis of Invariant Representation Learning: Invariant representation learning has abundant applications in domain adaptation, algorithmic fairness and privacy-preservation under attribute-inference attacks. Recently, the idea of learning language-invariant representations has also been actively explored in neural machine translation in order to enable knowledge transfer from high-resource language pairs to low-resource language pairs. Despite its broad applications, many fundamental questions remain open. Our work [15, 17, 19] has shown that utility has lower bound if exact invariance is attained. However, it is not clear what is the general form of tradeoff between utility and invariance. In particular, under a budget for approximate invariance, what is the maximum utility we can hope to achieve? This question calls for a characterization of the Pareto frontier between utility and variance. In my future research, I want to apply tools from information theory to provide a precise answer to the above question, and to use the theory of invariant representation learning to design Pareto-optimal algorithms in the above mentioned applications.

Efficient Structure Learning of Sum-Product Networks: SPNs distinguish themselves from other probabilistic graphical models, including both Bayesian Networks and Markov Networks, by the fact that reasoning can be performed exactly in linear time with respect to the size of the network. Similar to traditional graphical models, there are two main problems when learning SPNs: structure learning and parameter learning. In structure learning the goal is to infer the structure of SPNs directly from the data. As a direction for future work, I am highly interested in developing principled structure learning algorithms that can produce compact SPNs with directed acyclic structures. To date, most structure learning algorithms can only produce SPNs with tree structures, and they are based on various kinds of heuristics to generate SPNs from data without performance guarantees. In light of the existing limitations, it is favorable to come up with an algorithm that is able to produce SPNs with directed acyclic structures to fully exploit their representational power. I am also excited about extending the domain of SPNs from discrete/continuous data to more structured ones, e.g., string, graph, etc., and apply them to problems that require the capability of reasoning, including question answering, reading comprehension and statistical relation learning over graphs.

Unified Framework for Learning and Reasoning: I believe the holy grail of artificial intelligence is to build intelligent agents that have the ability to learn from the experience as well as to reason from what has been learned. In order to achieve this goal, we need to have a robust and probabilistic framework to unify learning and reasoning. Such framework is drastically different from the traditional one where symbolic representations are used to construct the knowledge base and first-order logic is used to build the inference engine. Instead, as shown in Figure 1, I propose to use invariant representation that maps real-world objects to their corresponding algebraic representations to serve as the foundation of knowledge base, and to use tractable probabilistic inference machine, e.g., Sum-Product networks, to act as the inference engine. Compared with the classic symbolic and logic-based framework, such new framework is inherently probabilistic and hence can handle the ubiquitous uncertainty. In particular, representations that are invariant to the change in the environment can provide us the robustness against various noise and nuisance factors in real-world, and the tractability of exact probabilistic inference machine can further allow us to efficiently deal with the uncertainty existing in real-world logic deduction.

Of course, the goal is challenging. First, in order to learn invariant representations, one needs to explicitly specify a set of nuisance factors that the representations should be invariant to. Due to the complexity of the real-world, such supervision is not always available or well-defined. Furthermore, when the invariant representations contain some internal structures, e.g., the hierarchical structure of sentence representations, it is not clear how to combine such structured data with existing tractable probabilistic inference machines. These problems are both fascinating and challenging, and I believe that being able to solve them could take us a significant step towards the goal of a unified framework for learning and reasoning.
References


