Cutting-edge Tutorial:
Machine Reasoning: Technology, Dilemma and Future

Nan Duan, Duyu Tang, Ming Zhou
Microsoft Research
{nanduan,dutang,mingzhou}@microsoft.com

1 Introduction

Machine reasoning research aims to build interpretable AI systems that can solve problems or draw conclusions from what they are told (i.e. facts and observations) and already know (i.e. models, common sense and knowledge) under certain constraints. Although its "formal" definitions vary in different publications (McCarthy, 1958; Pearl, 1988; Khardon and Roth, 1994; Bottou, 2011; Bengio, 2019), machine reasoning methods usually share some commonalities. First, such systems are based on different types of **knowledge**, such as logical rules, knowledge graphs, common sense, text evidence, etc. Second, such systems use different **inference algorithms** to manipulate available knowledge for problem-solving. Third, such systems have good **interpretability** to the predictions.

The developments of machine reasoning systems go through several stages. **Symbolic reasoning** methods represent knowledge using symbolic logic (e.g., propositional logic and first order logic) and perform inference using algorithms such as truth-table approach, inference rules approach, resolution, forward chaining and backward chaining. A major defect is that such methods cannot handle the uncertainty in data. **Probabilistic reasoning** methods combine probability and symbolic logic into a unified model. Such methods can deal with uncertainty, but suffer the combinatorial explosion when searching in a large discrete symbolic space. With the rapid developments of deep learning, neural reasoning methods attract much attention. **Neural-symbolic reasoning** methods represent knowledge symbols (such as entities, relationships, actions, logical functions and formulas) as vector or tensor representations, and allow the model to perform end-to-end learning effectively as all components are differentiable. **Neural-evidence reasoning** methods allow the model to communicate with the environment to acquire evidence for reasoning. As such models assume the reasoning layer is not required to be logical, both structured and unstructured data can be used as knowledge. Besides, as the interaction with the environment can be conducted multiple times, such approaches are good at solving sequential decision-making problems.

However, existing machine reasoning methods face with a **dilemma**: although they have many merits such as good abstraction, generalization and interpretability, their performance are still worse than black-box neural networks (such as pre-trained models) on most downstream tasks such as question answering, text classification, etc.

In this tutorial, we will review typical machine reasoning frameworks and talk about the dilemma between black-box neural networks with state-of-the-art performance and machine reasoning methods with better interpretability. We will also discuss possible research directions to escape this dilemma as the future work.

2 Description

We first review four machine reasoning frameworks.

Symbolic Reasoning This approach, also known as the Good, Old-Fashioned AI (GOFAI), was the dominant paradigm in the AI community before the late 1980s. By manipulating knowledge in the form of symbolic logic using inference algorithms, a symbolic reasoning system can solve deductive and inductive reasoning tasks. We will use deductive reasoning as an example to show how this task can be solved based on knowledge in the form of propositional logic and first-order logic, respectively. This part is also closely related to probabilistic reasoning and neural-symbolic reasoning.

Probabilistic Reasoning One drawback of symbolic reasoning is that it cannot handle data un-

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts, pages 1–6
Online, November 19 - 20, 2020. ©2020 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

certainty. To alleviate this problem, probabilistic reasoning is proposed, which integrates probabilistic models with symbolic knowledge in a unified framework. In such systems, probabilistic models handle the uncertainty issue while the symbolic logic represents types, relations, and the complex dependencies between them. We will use Bayesian Network (Pearl, 1988) and Markov Logic Network (Richardson and Domingos, 2006) as two representative models to show how probabilistic reasoning can solve typical reasoning tasks, such as diagnosis, prediction and maximum probable explanation.

Neural-Symbolic Reasoning Both symbolic reasoning and probabilistic reasoning support strong abstraction and generalization. Such systems have good interpretability but are fragile and inflexible duo to the finite and discrete symbolic representations. On the contrary, neural network models achieve state-of-the-art performance on various AI tasks, due to their good representation and learning capabilities. However, such models cannot capture compositionality and generalization in a systematic way. They cannot provide explicit decision-making evidence to explain their outputs as well, which make such systems look like a black box. So it is straightforward to integrate neural networks with symbolic reasoning, which is called neural-symbolic reasoning in this tutorial. In general, a neural-symbolic reasoning system (1) integrates existing reasoning technologies with symbolic knowledge based on neural networks and (2) implements inference as a chain of differentiable modules, where each module represents a program with a specific function. By doing these, such systems are usually more interpretable than black-box neural networks. We will review knowledge graph reasoning (Bordes et al., 2013; Wang et al., 2014; Lin et al., 2015; Wang et al., 2017; Glorot et al., 2013; Socher et al., 2013; Dong et al., 2014; Liu et al., 2016; Dettmers et al., 2018; Guo et al., 2019; Ren et al., 2020; Xiong et al., 2017; Dong et al., 2019; Rocktäschel and Riedel, 2017; Qu and Tang, 2019; K. Teru et al., 2020), neural semantic parsing (Dong and Lapata, 2016, 2018; Sun et al., 2018; Guo et al., 2018; Mao et al., 2019; Zhong et al., 2020), neural module network (Andreas et al., 2016; Hu et al., 2017; Gupta et al., 2020; Chen et al., 2020) and symbolic knowledge as constraints (Rocktaschel et al., 2015; Hu et al., 2016; Xu et al., 2018; Li and Srikumar, 2019; Wang et al., 2020) as four representative models.

Neural-Evidence Reasoning Previously mentioned three reasoning pipelines have the merits of utilizing abstractive logical or symbolic functions, which are interpretable to developers and users at concept level. The design of such symbolic functions in real applications are typically conducted by domain experts, thus these models cannot be easily extend to broader applications. Here, we review neural-evidence models that find external evidence and combine evidence with the input to make predictions. We group existing methods into three categories, including unstructured textual evidence retrieval models, structured fact evidence retrieval models, and iterative evidence retrieval models. Applications include open question answering (Chen and Yih, 2020), CommonsenseQA (Talmor et al., 2019), fact checking and verification (Thorne et al., 2018), inferential text generation (Rashkin et al., 2018; Sap et al., 2019), and multi-hop question answering (Yang et al., 2018).

We then talk about the dilemma between black-box neural networks with state-of-the-art performance and machine reasoning approaches with better interpretability.

Dilemma: Interpretability vs. Performance Despite the appealing properties of the previously mentioned machine reasoning approaches in terms of interpretability, the reality is that the leading systems on open benchmarks, evaluated by accuracy, are typically black-box models. We will discuss this dilemma of "interpretability versus performance" by showing the empirical success of pre-trained models on natural language understanding challenges, including Grade 8 New York Regents science exam (Clark et al., 2019), discrete reasoning over natural language (Dua et al., 2019), reasoning over rules in natural language (Clark et al., 2020), and logical reasoning (Yu et al., 2020). Afterwards, we will review model interpretation methods, including post-hoc ones and intrinsic ones. Post-hoc methods aim to interpret what an existing model learned without making changes to the original model. We will cover saliency maps (Simonyan et al., 2013), local interpretable model-agnostic explanations (LIME) (Ribeiro et al., 2016), testing with concept activation vectors (TCAV) (Kim et al., 2018), and visual explanation generation (Hendricks et al., 2016). Intrinsic methods are that inherently interpretable (to some extent). We will cover attention (Bahdanau et al., 2014), interpretable CNN (Zhang et al., 2018), and neural module network (Andreas et al., 2016).

We last summarize the content of this tutorial and discuss possible future directions.

Summary This tutorial classifies machine reasoning methods into 4 categories based on their modeling mechanisms, including symbolic reasoning, probabilistic reasoning, neural-symbolic reasoning and neural-evidence reasoning. Symbolic reasoning can handle complex reasoning tasks by using logical rules. Probabilistic reasoning further alleviates the data uncertainty issue in symbolic reasoning systems by introducing probabilistic models. Neural-symbolic reasoning provides more robust representation and learning capabilities based on the latest deep learning technologies. Neural-evidence reasoning doesn't require the reasoning layer to be logical, so they can leverage both symbolic and non-symbolic evidence. All these methods have good applications in many real-world scenarios like expert system, medical diagnosis, knowledge base completion, question answering, search engine, fact checking, etc.

Of course, we also notice the dilemma of existing machine reasoning methods. We think this is only a short-term phenomenon. With the continue and rapid developments of different areas at the same time, such as knowledge base engineering, pre-training, interpretability modeling and neural-symbolic computing, we believe machine reasoning will definitely have a brighter future.

3 Outline

Opening (15 min.) will describe the motivation and outline of this tutorial and give our definition on machine reasoning.

Symbolic Reasoning (20 min.) will review typical methods based on propositional logic and first order logic, respectively.

Probabilistic Reasoning (20 min.) will review typical methods based on Bayesian Network and Markov Logic Network, respectively.

Neural-Symbolic Reasoning (40 min.) will review typical methods including knowledge graph reasoning, neural semantic parsing, neural module network and symbolic knowledge as constraints.

Neural-Evidence Reasoning (40 min.) will review text-base evidence retrieval models, fact-based evidence retrieval models, and interactive evidence retrieval models.

Dilemma: Interpretability vs. Performance (30 min.) will review post-hoc models and intrinsic models for interpretation, and discuss the dilemma of "interpretability versus performance".

Summary & Future Discussion (10 min.) will summarize the content of this tutorial and discuss possible future directions.

4 Prerequisites for the Attendees

We expect the attendees to be familiar with typical NLP tasks (such as question answering, semantic parsing, text generation, etc.), basic concepts of logic (such as propositional logic and first order logic) and knowledge graph, recent neural network architectures (such as convolutional neural network, recurrent neural network and Transformer) and pre-trained language models (such as GPT and BERT).

5 Small Reading List

- Domingos and Richardson (2004) - an introduction to Markov Logic as a unifying framework for statistical relational learning, which is closely related to probabilistic reasoning;
- Bottou (2011) - a nice introduction to machine reasoning;
- Besold et al. (2017) and Garcez et al. (2019) - two surveys on neural-symbolic reasoning;
- Storks et al. (2019) - a survey on benchmarks, knowledge resources, learning and inference approaches to natural language inference;
- Du et al. (2020) - a survey on interpretable machine learning techniques;
- Chen and Yih (2020) - a tutorial on open-domain question answering, in which many work can be categorized as neural-evidence reasoning;
- Sap et al. (2020) - a tutorial on commonsense reasoning for natural language processing.

6 Tutorial Abstract

Machine reasoning research aims to build interpretable AI systems that can solve problems or draw conclusions from what they are told (i.e. facts and observations) and already know (i.e. models, common sense and knowledge) under certain constraints. In this tutorial, we will (1) describe the motivation of this tutorial and give our definition on machine reasoning; (2) introduce typical machine reasoning frameworks, including symbolic reasoning, probabilistic reasoning, neural-symbolic reasoning and neural-evidence reasoning, and show their successful applications in real-world scenarios; (3) talk about the dilemma between black-box neural networks with state-of-the-art performance and machine reasoning approaches with better interpretability; (4) summarize the content of this tutorial and discuss possible future directions.

7 Presenters

Nan Duan is a Principal Researcher of the Natural Language Computing group at Microsoft Research Asia. His research focuses on question answering, semantic parsing, pre-trained models for learning joint representations of natural language and images/videos/codes/knowledge. His technologies have been widely used in Microsoft products like Bing, Ads, Chatbot, Azure, etc.

Duyu Tang is a Senior Researcher of the Natural Language Computing group at Microsoft Research Asia, working on natural language processing. Duyu's research has been advancing the state of art of robust, interpretable and trustworthy NLP systems, while making direct technical contributions to production. Over the years, Duyu worked on a wide range of NLP problems, from sentiment analysis, question answering, conversational semantic parsing, knowledge-driven machine reasoning, fact checking and fake news detection, to AI for software engineering. He has served as area chair for EMNLP 2020.

Ming Zhou Dr. Ming Zhou is Research Manager of the Natural Language Computing Group at Microsoft Research Asia and leads numerous research projects including next generation search engines, neural machine translation, machine reading comprehension, question-answering, chatbots, computer poetry, knowledge graph and recommendation systems. He has published over 200 papers at top conferences and journals. He has served as area chairs of ACL, EMNLP and many other conferences. He was ACL president in 2019.

References

Jacob Andreas, Jacob Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *CVPR*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Yoshua Bengio. 2019. The consciousness prior. In *arXiv*.

Tarek R. Besold, Artur S. d'Avila Garcez, Sebastian Bader, Howard Bowman, Pedro M. Domingos, Pascal Hitzler, Kai-Uwe Kühnberger, Luís C. Lamb, Daniel Lowd, Priscila Machado Vieira Lima, Leo de Penning, Gadi Pinkas, Hoifung Poon, and Gerson Zaverucha. 2017. Neural-symbolic learning and reasoning: A survey and interpretation. *CoRR*.

Antoine Bordes, Nicolas Usunier, and Alberto Garcia-Duran. 2013. Translating embeddings for modeling multi-relational data. In *NeurIPS*.

Léon Bottou. 2011. From machine learning to machine reasoning. In *arXiv*.

Danqi Chen and Wen-tau Yih. 2020. Open-domain question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 34–37.

Xinyun Chen, Chen Liang, Adams Wei Yu, Denny Zhou, D. Song, and Quoc V. Le. 2020. Neural symbolic reader: Scalable integration of distributed and symbolic representations for reading comprehension. In *ICLR*.

Peter Clark, Oren Etzioni, Daniel Khashabi, Tushar Khot, Bhavana Dalvi Mishra, Kyle Richardson, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord, Niket Tandon, et al. 2019. From'f'to'a'on the ny regents science exams: An overview of the aristo project. *arXiv preprint arXiv:1909.01958*.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. *arXiv preprint arXiv:2002.05867*.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2d knowledge graph embeddings. In *AAAI*.

Pedro Domingos and Matthew Richardson. 2004. Markov logic: A unifying framework for statistical relational learning.

Honghua Dong, Jiayuan Mao, Tian Lin, Chong Wang, Lihong Li, and Denny Zhou. 2019. Neural logic machines. In *ICLR*.

Li Dong and Mirella Lapata. 2016. Language to logical form with neural attention. In *ACL*.

Li Dong and Mirella Lapata. 2018. Coarse-to-fine decoding for neural semantic parsing. In *ACL*.

Xin Luna Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *KDD*.

Mengnan Du, Ninghao Liu, and Xia Hu. 2020. Techniques for interpretable machine learning. *Communications of the ACM*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*.

Artur d'Avila Garcez, Marco Gori, Luis C. Lamb, Luciano Serafini, Michael Spranger, and Son N. Tran. 2019. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *arXiv*.

Xavier Glorot, Antoine Bordes, Jason Weston, and Yoshua Bengio. 2013. A semantic matching energy function for learning with multi-relational data. In *arXiv*.

Daya Guo, Duyu Tang, Nan Duan, Ming Zhou, and Jian Yin. 2018. Dialog-to-action: Conversational question answering over a large-scale knowledge base. In *NeurIPS*.

Lingbing Guo, Zequn Sun, and Wei Hu. 2019. Learning to exploit long-term relational dependencies in knowledge graphs. In *ICML*.

Nitish Gupta, Kevin Lin, Dan Roth, Sameer Singh, and Matt Gardner. 2020. Neural module networks for reasoning over text. In *ICLR*.

Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. 2016. Generating visual explanations. In *European Conference on Computer Vision*, pages 3–19. Springer.

Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. 2017. Learning to reason: End-to-end module networks for visual question answering. In *ICCV*.

Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard Hovy, and Eric Xing. 2016. Harnessing deep neural networks with logic rules. In *ACL*.

Komal K. Teru, Etienne Denis, and William L. Hamilton. 2020. Inductive relation prediction by subgraph reasoning. In *ICML*.

Roni Khardon and Dan Roth. 1994. Learning to reason. In *AAAI*.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR.

Tao Li and Vivek Srikumar. 2019. Augmenting neural networks with first-order logic. In *ACL*.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *AAAI*.

Quan Liu, Hui Jiang, Andrew Evdokimov, Zhen-Hua Ling, Xiaodan Zhu, Si Wei, and Yu Hu. 2016. Probabilistic reasoning via deep learning: Neural association models. In *arXiv*.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. 2019. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *ICLR*.

John McCarthy. 1958. Program with common sense.

Judea Pearl. 1988. Probabilistic reasoning in intelligent systems: Networks of plausible inference. In *Morgan Kaufmann Publishers Inc.*

Meng Qu and Jian Tang. 2019. Probabilistic logic neural networks for reasoning. In *NeurIPS*.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A Smith, and Yejin Choi. 2018. Event2mind: Commonsense inference on events, intents, and reactions. *arXiv preprint arXiv:1805.06939*.

Hongyu Ren, Weihua Hu, and Jure Leskovec. 2020. Query2box: Reasoning over knowledge graphs in vector space using box embeddings. In *ICLR*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Matthew Richardson and Pedro Domingos. 2006. Markov logic networks. volume 62.

Tim Rocktaschel, Sameer Singh, and Sebastian Riedel. 2015. Injecting logical background knowledge into embeddings for relation extraction. In *NAACL*.

Tim Rocktäschel and Sebastian Riedel. 2017. End-to-end differentiable proving. In *NeurIPS*.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *AAAI*.

Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. 2020. Commonsense reasoning for natural language processing. In *ACL*.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Richard Socher, Danqi Chen, Christopher Manning, and Andrew Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. In *NeurIPS*.

Shane Storks, Qiaozi Gao, and Joyce Y. Chai. 2019. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *arXiv*.

Yibo Sun, Duyu Tang, Nan Duan, Jianshu Ji, Guihong Cao, Xiaocheng Feng, Bing Qin, Ting Liu, and Ming Zhou. 2018. Semantic parsing with syntax- and table-aware sql generation. In *ACL*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *NAACL*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *NAACL*.

Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. 2017. Knowledge graph embedding: A survey of approaches and applications. In *IEEE-TKDE*.

Ruize Wang, Duyu Tang, Nan Duan, Wanjun Zhong, Zhongyu Wei, Xuanjing Huang, Daxin Jiang, and Ming Zhou. 2020. Leveraging declarative knowledge in text and first-order logic for fine-grained propaganda detection. In *EMNLP*.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *AAAI*.

Wenhan Xiong, Thien Hoang, and William Yang Wang. 2017. Deeppath: A reinforcement learning method for knowledge graph reasoning. In *EMNLP*.

Jingyi Xu, Zilu Zhang, Tal Friedman, Yitao Liang, and Guy Van den Broeck. 2018. A semantic loss function for deep learning with symbolic knowledge. In *ICML*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *EMNLP*.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. *arXiv preprint arXiv:2002.04326*.

Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. 2018. Interpretable convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8827–8836.

Victor Zhong, Mike Lewis, Sida I. Wang, and Luke Zettlemoyer. 2020. Grounded adaptation for zero-shot executable semantic parsing. In *EMNLP*.