much data there are for training, there always exist cases that the training data cannot cover. How to deal with the long tail problem poses a significant challenge to deep learning. By resorting to deep learning alone, this problem would be hard to solve.

Language data is by nature symbol data, which is different from vector data (real-valued vectors) that deep learning normally utilizes. Currently, symbol data in language are converted to vector data and then are input into neural networks, and the output from neural networks is further converted to symbol data. In fact, a large amount of knowledge for natural language processing is in the form of symbols, including linguistic knowledge (e.g. grammar), lexical knowledge (e.g. WordNet) and world knowledge (e.g. Wikipedia). Currently, deep learning methods have not yet made effective use of the knowledge. Symbol representations are easy to interpret and manipulate and, on the other hand, vector representations are robust to ambiguity and noise. How to combine symbol data and vector data and how to leverage the strengths of both data types remain an open question for natural language processing.

There are complex tasks in natural language processing, which may not be easily realized with deep learning alone. For example, multi-turn dialogue amounts to a very complicated process. It involves language understanding, language generation, dialogue management, knowledge base access and inference. Dialogue management can be formalized as a sequential decision process and reinforcement learning can play a critical role. Obviously, combination of deep learning and reinforcement learning could be potentially useful for the task, which is beyond deep learning itself.

In summary, there are still a number of open challenges with regard to deep learning for natural language processing. Deep learning, when combined with other technologies (reinforcement learning, inference, knowledge), may further push the frontier of the field.

FUNDING

This work is supported in part by the National Basic Research Program of China (973 Program, 2014CB340301).

Hang Li
Noah's Ark Lab, Huawei Technologies, Hong Kong, China
E-mail: hangli65@hotmail.com

REFERENCES

1. Blunsom P, Grefenstette E and Kalchbrenner N. A convolutional neural network for modelling sentences. In: *52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore, USA, 2014, 655–65.
2. Hu B, Lu Z and Li H. Convolutional Neural Network Architectures for Matching Natural Language Sentences. In: *Advances in Neural Information Processing Systems 27*. Montreal, Canada, 2014, 2042–50.
3. Ma L, Lu Z and Shang L *et al*. Multimodal Convolutional Neural Networks for Matching Image and Sentence. In: *IEEE International Conference on Computer Vision*. Santiago, Chile, 2015, 2623–31.
4. Cho K, Van Merriënboer B and Gulcehre C *et al*. Learning phrase representations using rnn encoder-decoder for statistical machine. In: *Conference on Empirical Methods in Natural Language Processing*. Doha, Qatar, 2014, 1724–34.
5. Bahdanau D, Cho K and Bengio Y. Neural machine translation by jointly learning to align and translate. In: *3rd International Conference on Learning Representations*. San Diego, USA, 2015.
6. Wu Y, Schuster M and Chen Z. CoRR, vol. abs/1609.08144, 2016.
7. Shang L, Lu Z and Li H. Neural Responding Machine for Short-Text Conversation. In: *53th Annual Meeting of Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. Beijing, China, 2015, 1577–86.
8. Chen D and Manning CD. A Fast and Accurate Dependency Parser using Neural Networks. In: *Conference on Empirical Methods in Natural Language Processing*. Doha, Qatar, 2014, 740–50.

National Science Review
5: 24–26, 2018
doi: 10.1093/nsr/nwx110
Advance access publication 8 September 2017

MULTIDISCIPLINARY

Special Topic: Machine Learning

Learning causality and causality-related learning: some recent progress

Kun Zhang[1,*], Bernhard Schölkopf[2], Peter Spirtes[1] and Clark Glymour[1]

INTRODUCTION

Causality is a fundamental notion in science, and plays an important role in explanation, prediction, decision making and control. Recently, with the rapid accumulation of huge volumes of data, it is even more desirable to abstract causal knowledge from data. Furthermore, such data are usually time series measured over a relatively long time period or aggregated data from multiple data sets collected in different environments or under different experimental conditions, leading to the issue of data heterogeneity. Causality also provides a way to understand and tackle data heterogeneity, while traditional machine learning

typically assumes that the given data follow a fixed distribution.

A traditional way to discover causal relations is to use interventions or randomized experiments, which are in many cases too expensive, too time-consuming, or even impossible. Therefore, revealing causal information by analyzing purely observational data, known as causal discovery, has drawn much attention [1]. Past decades have seen a series of cross-disciplinary advances in algorithms for identifying causal relations and effect sizes from observational data or mixed experimental and observational data. These developments promise to enable better use of appropriate 'big data'. They have already been applied in genomics, ecology, epidemiology, space physics, clinical medicine, neuroscience and many other domains, often with experimental or quasi-experimental validation of their predictions. Causal discovery will be a main focus of this perspective. In traditional causality research, algorithms for identification of causal effects, or inferences about the effects of interventions, when the causal relations are completely or partially known, address a different class of problems; see [2] and references therein. Moreover, causal models provide compact descriptions of the properties of data distributions, and it has recently been demonstrated that causal knowledge can facilitate various machine learning tasks, including semi-supervised learning and domain adaptation (or transfer learning).

LEARNING CAUSAL RELATIONS

It is well known in statistics that 'causation implies correlation, but correlation does not imply causation'. Perhaps it is fairer to say that correlation does not directly imply causation—in fact, it has been shown that under various sets of assumptions, the underlying causal structure over a set of random variables can be recovered from their observed data, as least to some extent [1]. Since the 1990s, conditional independence relationships in the data have been exploited to recover the underlying causal structure. Typical (conditional independence) constraint-based algorithms include PC and fast causal inference (FCI) [1]. PC assumes that there is no confounder (unobserved direct common cause of two measured variables), and its discovered causal information is asymptotically correct. FCI gives asymptotically correct results even in the presence of confounders. Such approaches are widely applicable because they can handle various types of data distributions and causal relations, given reliable conditional independence testing methods. However, they do not necessarily provide complete causal information because they output (independence) equivalence classes, i.e. a set of causal structures satisfying the same conditional independences. The PC and FCI algorithms produce graphical representations of these equivalence classes. In cases without confounders, there also exist score-based algorithms that aim to find the causal structure by optimizing a properly defined score function. Among them, the greedy equivalence search (GES) [2] is a well-known two-phase procedure that directly searches over the space of equivalence classes. A parallelized modification (FGES) is able to search for causal relations in very high-dimensional data sets. Such algorithms have been implemented in the Tetrad package (http://www.phil.cmu.edu/tetrad/).

Recently it has been shown that algorithms based on properly defined functional causal models (FCMs) are able to distinguish between different directed acyclic graphs (DAGs) in the same equivalence class. This benefit is owing to additional assumptions on the data distribution apart from conditional independence relations. An FCM represents the effect variable Y as a function of the direct causes X and some noise term E, i.e. $Y = f(X, E)$, where E is independent of X. Thanks to the constrained functional classes, the causal direction between X and Y is identifiable because the independence condition between the noise and cause holds only for the true causal direction and is violated for the wrong direction (for details one may see [3]). Typical FCMs include the linear, non-Gaussian, acyclic model (LiNGAM) [4], in which $Y = aX + E$ with linear coefficient a, the nonlinear additive noise model (ANM) [5], in which $Y = f(X) + E$, and the post-nonlinear (PNL) causal model [6], which further considers possible nonlinear sensor or measurement distortion f_2 in the causal process: $Y = f_2(f_1(X) + E)$.

The identifiability of the causal direction is a crucial issue in functional causal discovery. The conditions for identifiability of causal directions for the PNL causal model entail those for LiNGAM and ANM, because they are special cases of the PNL causal model. Under a smoothness assumption on the involved functions and a positivity assumption on the densities of X and E, there are only five specific situations where the causal direction is not identifiable if data were generated according to the PNL causal model [6]. Accordingly, one way to estimate the causal structure from observed data based on the FCM is to first fit the model on given data and then test for independence between the estimated noise term and the hypothetical cause. So far functional causal discovery has been mainly concerned with cases without confounders or feedbacks, with several exceptions [7,8].

In practice, for reliable causal discovery one needs to address specific challenges that are often posed in the causal process or the sampling process to generate the observed data. Below are some particular issues that have recently been considered:

(i) **Deterministic case.** In a particular deterministic case where $Y = f(X)$ without noise, it is impossible to make use of the independence between noise and the cause to find the causal direction. However, one may exploit a certain type of independence between the transformation f and the distribution of the cause X to characterize the causal asymmetry and determine the causal direction [9].

(ii) **Nonstationary/heterogeneous data.** It is commonplace to encounter nonstationary or heterogeneous data, in which the underlying generating process changes over time or across data sets. Interestingly, if the

qualitative causal structure is fixed and the mechanisms or parameters associated with the causal structure may change across data sets or over time (the mechanisms may change such that some causal links in the structure vanish over some time periods or domains), causal discovery may benefit from distribution shift because causal modeling and distribution shift are heavily coupled. This, in particular, inspires a framework for causal mechanism change detection, causal skeleton estimation, causal direction identification and nonstationary driving force estimation [10].

(iii) **Measurement error.** Measurement error in the observed values of the variables can greatly change the output of various causal discovery methods. Given the ubiquity of measurement error caused by instruments or proxies used in the measuring process, this problem has received much attention, and sufficient conditions under which the causal model for the underlying measurement-error-free variables can be partially or completely identified in the presence of measurement error with unknown variance have been established [11]. This will hopefully inspire a set of causal discovery methods dealing with measurement error.

(iv) **Selection bias.** Selection bias is an important issue in statistical inference, which arises when the probability of including a data point in the sample depends on some attributes of the point. Selection bias, if not corrected, often distorts the results of statistical analysis and causal discovery and inference. In the presence of outcome-dependent selection bias, with FCM-based causal discovery it is possible to identify the correct causal direction and estimate the properties of the causal mechanism [12]. More general situations with selection bias remain to be studied.

(v) **Subsampled or temporally aggregated time series.** In many times series, data are subsampled or temporally aggregated due to the measuring device or sampling procedure, or for the purposes of efficient collection and storage. It has been shown that under suitable assumptions, the true causal relations are identifiable from both subsampled and temporally aggregated data; interested readers may refer to [13] and references therein.

CAUSALITY-RELATED MACHINE LEARNING

Learning under data heterogeneity has been becoming important because of the potential distribution shift in the data and the expense or neglect of labeling procedures. Typical learning problems in this category include semi-supervised learning, domain adaptation or transfer learning, and learning with positive and unlabeled examples. To solve such problems, one has to gain information about the underlying process behind the given data.

The distinction between causal and 'anticausal' learning was discussed in [14], together with a causal view of semi-supervised learning. In this learning setting, an important issue is to determine whether unlabeled data points are useful to improve the prediction model. It has been noticed that if the features are causes of the target (or label) with no confounder between them, then unlabeled data points are not helpful. In domain adaptation or transfer learning, it is essential to determine what knowledge to transfer from source domains to the target and how to do the transfer. Causal modeling has been shown to provide a nice tool to address this issue [14–16]. Causal diagrams have been used to establish conditions that allow transportation of results across domains [15]. Even when such conditions do not hold, it is still possible to leverage causal knowledge together with some technical conditions for domain adaptation [16]; the basic idea is that if there is no confounder between them, $P(\texttt{cause})$ and $P(\texttt{effect} \mid \texttt{cause})$ are reflections of true causal processes and change independently, allowing separate parameterization of the changes in a simple form.

Modern causality research has benefited a great deal from the advances in machine learning techniques, which provide an essential tool to extract information from data. On the other hand, causal information describes properties of the process behind the observed data, and is able to facilitate the solution of a number of learning problems involving distribution shift or concerning the relationship between different modules of the joint distribution. Open problems in this research area include the development of computationally efficient causal discovery methods that apply to more general situations and the characterization of what information of the underlying causal process is useful and the optimal ways to use it in various machine learning settings.

FUNDING

C.G. and K.Z. would like to acknowledge the support from the National Institutes of Health (NIH-1R01EB022858-01 FAIN-R01EB022858, NIH-1R01LM012087, and NIH-5U54HG008540-02 FAINU54HG008540). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Kun Zhang[1,*], Bernhard Schölkopf[2], Peter Spirtes[1] and Clark Glymour[1]
[1]Department of Philosophy, Carnegie Mellon University, USA
[2]Max Planck Institute for Intelligent Systems, Germany
***Corresponding author.** E-mail: kunz1@cmu.edu

REFERENCES

1. Spirtes P, Glymour C and Scheines R. *Causation, Prediction, and Search*, 2nd edn. Cambridge: MIT Press, 2001.
2. Chickering DM. *J Mach Learn Res* 2003; **3**: 507–54.
3. Spirtes P and Zhang K. *Appl Informat* 2016; **3**.
4. Shimizu S, Hoyer PO and Hyvärinen A *et al. J Mach Learn Res* 2006; **7**: 2003–30.
5. Peters J, Mooij JM and Janzing D *et al. J Mach Learn Res* 2014; **15**: 2009–53.
6. Zhang K and Hyvärinen A. On the identifiability of the post-nonlinear causal model. In: *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*. OR, USA: AUAI Press, 2009.

7. Lacerda G, Spirtes P and Ramsey J *et al.* Discovering cyclic causal models by independent components analysis. In: *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI2008).* OR, USA: AUAI Press, 2008.
8. Hoyer PO, Shimizu S and Kerminen AJ *et al. Int J Approx Reason* 2008; **49**: 362–78.
9. Janzing D, Mooij J and Zhang K *et al. Artif Intell*, 2012; **182-183,** 1–31.
10. Zhang K, Huang B and Zhang J *et al.* Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In: *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.
11. Zhang K, Gong M and Ramsey J *et al.* Causal discovery in the presence of measurement error: Identifiability conditions. In: *UAI 2017 Workshop on Causality: Learning, Inference, and Decision-Making*, 2017.
12. Zhang K, Zhang J and Huang B *et al.* On the identifiability and estimation of functional causal models in the presence of outcome-dependent selection. In: *Proceedings of the 32rd Conference on Uncertainty in Artificial Intelligence (UAI 2016).* OR, USA: AUAI Press, 2016.
13. Gong M, Zhang K and Schölkopf B *et al.* Causal discovery from temporally aggregated time series. In: *Proc. Conference on Uncertainty in Artificial Intelligence (UAI).* OR, USA: AUAI Press, 2017.
14. Schölkopf B, Janzing D and Peters J *et al.* On causal and anticausal learning. In: *Proc. 29th International Conference on Machine Learning (ICML 2012).* NY, USA: Omnipress, 2012.
15. Pearl J and Bareinboim E. Transportability of causal and statistical relations: A formal approach. In: *Proceedings of the 25th AAAI Conference on Artificial Intelligence*. San Francisco: AAAI Press, 2011, 247–54.
16. Zhang K, Schölkopf B and Muandet K *et al.* Domain adaptation under target and conditional shift. In: *Proceedings of the 30th International Conference on Machine Learning, JMLR: W&CP Vol. 28*, 2013.

National Science Review
5: 26–29, 2018
doi: 10.1093/nsr/nwx137
Advance access publication 17 November 2017