Causality for Machine Learning Summary

Jannik Gut
Department of Computer Science
ETH Zürich
Switzerland
jgut@student.ethz.ch

Zixin Shu
Department of Computer Science
ETH Zürich
Switzerland
zixshu@student.ethz.ch

Abstract

This summary shows the motivation behind including causality to machine learning and introduces basic concepts, like the structural causal model, disentanglement or half sibling entanglement. Problems when incorporating causality into models are explained, followed by the competing experts and causal autoencoder methods that solve these problems, as well as applications of half sibling regression in exoplanet detection by transit photometry and direct imaging methods.

1 Introduction

Machine learning is often mocked as a dark art that takes some input, learns some opaque distribution from that data and based on that almost magically can give an unreasonably good result. In a scenario like that, where the model is treated as a black box that just gives good results, problems can occur, as the learnt distributions might be based on (seemingly) non-sensible features of the input, that an intelligent learner would just ignore, like the time of day when a photo in the training set was taken for classifying tanks in the battlefield, or even worse, the learnt distributions are over non-ethical features, like race or sex when evaluating job candidates. One way to influence what distributions should be influential inside of the model is to use causality and learn causal representations, which give reasonable results, both in terms of robust accuracy against changes of unimportant features and in terms of transparency of decision making.

In this summary of a presentation, which is widely based on a paper[1], there are two main topics: learning causal representations (†) in a general fashion and the concrete application of denoising data(‡) with systematic noise. After this introduction, first some theoretical basics and notation are established in section 2, to understand the problems that occur in section 3, after that, the methods that are used to solve these problems in section 4 and the applications, where methods can be used in section 5. In the end there is a conclusion over the summary and an outlook on future challenges and goals in section 6. The subsections are marked with the symbol (†or ‡) of the topic they belong to.

2 Basics and notation

2.1 Structural causal model (SCM)†‡

Causation is an important property for all the following tasks in the paper. It is often modelled with a directed acyclical graph called structural causal model (SCM)[2], where the nodes are *observables* and the directed edges represent *direct causation*/influence on the target node. Looking at Figure 1 in the appendix one can see that *pig value* and *meat demand* directly influence the *pig price* payed in the end.
The set of variables that directly cause a variable X are the parents of it's node, denoted by *PA(X)* e.g. *pig health,* U_1*, pig weight* are the parents of *pig value*. Most observables have enlightening names, the exception are the U-terms, which stand for *unexplained* or noise random variable, e.g.

Lecture notes for the *Causal Representation Learning* course at ETH (fall semester 2020).

U_1 covers things like the spread of meat to fat and other not observed features. That noise-term is needed to better depict reality and also to guarantee that X_i can be fully deterministically explained by its parents, i.e. $X_i = f_i(PA(X_i))$, which is one *mechanism*.
The graph also infers a causal/topographical ordering π based on the direct dependencies. In Figure 1 *pig price* depends on *pig value*, which depends on *pig weight*, what means $\pi(pig\ price) > \pi(pig\ value) > \pi(pig\ weight)$. A *level* of this ordering is an ordering on sets of nodes, where each node in a set depends on/influences all nodes of an influencing/depending level. No two nodes in the same set can be strictly ordered. Taking an example from Figure 1, each level in the graph illustration can be a level in the topographical ordering, but these levels are not strict, as one could make a valid top-level of only *pig price* and another level below of only *cow price*. This ordering yields a *reduced form* over the noise variables.

2.2 (Dis)Entanglement †‡

As seen in the last subsection, the variable X_i only depends on its direct parents, this shows when trying to give the probability of all the observations, which factorises nicely thanks to the disentangled (causal) representation of the observations:

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i | PA(X_i))$$

where each $p(X_i|PA(X_i))$ is called **independent mechanism**, which entails that changing one mechanism on *i* does not change the mechanism on another variable *j*.
Opposed to that is an entangled factorisation, where X_i depends on other X_j and each change of $p(X_i|PA(X_i))$ often changes many of the dependent mechanisms $p(X_i|X_{i+1}, \ldots, X_n)$.

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i | X_{i+1}, \ldots, X_n)$$

Figure 1 in the appendix gives an example; *pig weight* does not grow, independent on how high the *pig value* was or the *pig price* the vendor payed. On the other hand, if *pig weight* is changed (*intervened on*), the *pig value* changes.
The sparse causal shift hypothesis[3] states that a shift in the input distribution always comes from a sparse change of mechanisms, which gives a good approximation of reality and allows to reason about the quality of the modelled factorisation/ SCM, as passive or active changes to the input distribution only gives a small change to independent mechanisms, but a large change to dependent mechanisms.

2.3 Half sibling entanglement‡

Much like the real world, half siblings in the SCM share (at least) one parent, but diverge in (at least) another parent and are independent of each other. An example for that are *pig price* and *cow price* which share *meat demand*, but have different *values*. In other words: predicting *cow price* from the pig specific observables (only) picks up information about the *meat demand*.
Some additional assumptions are needed to make usable computations[4]: first, the **additive noise model** is assumed, which models a observable Y with a *shared noise (N)* and a desired *unobserved value (Q)*: $Y = (Q) + f(N)$. Where all variables are random variables and f is measurable. *N* influences another observable *X*. For *N* and *Q* being independent of each other and other influences on *X*. In Figure 1 the *N* is the *meat demand*, that influences *Y*, the *pig price*, and its half sibling *X,cow price*, while *Q* is *pig value*.
With these assumptions, *Q* can be estimated ($\hat{Q}$) up to a constant in the L^2 norm if either the other influences on *X* go to zero or are a random vector whose components are jointly independent.

$$\mathbb{E}[(\hat{Q} - (Q - \mathbb{E}[Q]))^2] = \mathbb{E}[Var[f(N|X)]] \overset{assumption}{\approx} 0$$

In the special case that there are no other influences and the noise can be written as a function of *X* (**complete information**) , *Q* can be recovered up to a constant offset.

$$\hat{Q} = Q - \mathbb{E}[Q] = Y - \mathbb{E}[Y|X]$$

3 Problems

3.1 Towards causal representational learning †

SCMs can be used well to reason about different hypotheses and enable computations, but are an abstracted, symbolic model, which assumes that the causal model is already given, what unfortunately is not always the case, as the mechanisms are sometimes not known beforehand or hard to formulate in a way that the machine learning model could use it. The goal is to embed a SCM into a deep learning model.

3.2 Disentangle noise ‡

There are some problems, where the signal is only very small, comparably to systematic noise. One way to solve this problem is to use half sibling regression[4], where two samples of the same distribution share the same systematic noise, but one sample has a faint interesting signal, while the influence of signals is negligible in the other sample.

4 Methods

4.1 Competing experts †

The proposed model is split into two parts: a set of experts, that generate samples given an input and a discriminator as the second part, that does a desired task like classification. There are M different experts that modify one input to M different samples that the shared discriminator has to distinguish. If the discriminator can give a good enough output on some sample, the adjacent expert gets enforced/trained, while other experts are neglected for this round. The discriminator already is well trained at the beginning and does not receive additional training.
In the experiment, at each training round, an image with one disturbance is given. The experts then try to undo the disturbance to the best of what they have learnt so far and the discriminator tries to classify the expertly modified images to the correct digit. The result of the experiment was reassuring, as each expert learnt one different mechanism starting from the identity modification. But the number of different mechanisms was known and could be applied in a balanced fashion, one mechanism at a time.

4.2 Causal autoencoders †

Autoencoders[5] are an established method for various tasks in machine learning, but for causality, special attention has been given to dimension reduction. A high-dimensional input gets encoded by an encoder to a lower-dimensional latent space and decoded by a decoder back to the high-dimensional input space. The goal of the autoencoder is to generate the id transformation with the lower dimensional bottleneck.
The autoencoder can be used to map the input to the latent space, where each dimension is a mechanism of the underlying SCM, so that interventions can be measured more easily.

4.3 Training autoencoders †

There are multiple degrees of training causal autoencoders, where each uses more or fewer assumptions and is, therefore, more or less complicated, reliably computable and resource intensive.
In the easiest case, the latent dimension and decoder are already known and the only thing left is to train the encoder which is a more or less straight-forward neural network training task.
The most general way of training is **causal training**, where the latent space and decoder also have to be learnt. One way to train in this fashion is to again use the *sparse causal shift hypothesis* to enforce a latent space, where the actions/interventions on the input lead to only few changes in the latent features. Another way is **structural learning**, where the model tries to learn the reduced form of the SCM by giving one set of latent variables after another to the decoder[6] to emulate the order of the SCM.
The last way presented is to use **counterfactual training**, which uses a Generative Adversarial Network[7] and demands that the discriminator gives good results on the *counterfactuals* that the generative part (the autoencoder) has created[8].

Counterfactuals are samples generated based on real samples and applied *hypotheticals* to it, like ("Peppa had a value of X €, what *would have been* the price if Peppa had been heavier?) and can be created by clamping a subset of properties (e.g. Peppas health) that need to be enforced and compute the (remaining) latent features(**hybrid sampling**). One way of doing this is "drop-in". Clamping on noise encourages statistical independence (already by design in GANs, useful for autoencoders in general), but clamping on mechanisms encourages them to be independent according to the ICM principle[9].
In experiments it could be shown that causal autoencoder can distinguish and order mechanisms, even if they are multidimensional (e.g. colour) and help to discover the underlying mechanisms, which make the model more robust and better.

5 Applications

5.1 Transit photometry ‡

One case where the disentangled noise problem (subsection 3.2) occurs is when photographing transiting exoplanets in front of stars, as their signal is only very, very faint, but for the most part, noise is very faint too and either systematic (like from the electronics on the probe) or independently random. Since detectable exoplanets are seldom, they are hardly ever close enough to each other on the photo to interfere and are far away from each other to noticeably physically influence each other[10].
With these observations, systematic noise (N) can be mapped as a parent to an empty photograph(X) without other parents and one (Y) where an exoplanet(Q) is suspected, which is the only other parent of that photo. Once in the framework, different computations can be done, such as half sibling regression[11]. The reference patch was not chosen at random, but temporal and spatial data was taken into consideration, to get the most similar systematic noise.
The approach was used to find 36 tentative planets[10], which have been checked by the more local radial velocity method.[12] After this check 17 planets were confirmed for the first time[13], where K2-18b was national news worthy[14], as water was found on this planet.

5.2 Direct imaging ‡

Another method for finding planets is direct imaging, that takes a photograph of the sun and its surrounding, removes the sun from the picture and checks if there are any planets. What sounds super simple is highly complicated as suns tend to be bright, so bright that it heavily distorts the pixels close to the sun, where exoplanets could be found.
Instead of very small noise from the instruments of the probe, there is very big noise from the sun, but the photograph without a planet still has no additional non-negligible parent, so half sibling methods can be applied in the extensive pipeline, where de-rotation and PCA was applied among other methods.[15]
With this improved pipeline, the exoplanets appear more significantly and cleaner, which makes it easier to spot them and improve the direct imaging method.

6 Conclusion

Causal models can bring much needed structure into the field of machine learning, thanks to the mathematical concepts of causality now being mature enough, that they can be used in non-trivial models and, as shown by the exoplanet detection, can yield significant benefits for real-world problems. But the road still is long, answering open questions such as opening the space from the rather well-controlled test environments to a more diverse set of environments and tasks, how to tackle transfer learning and how to grasp a basic object in the input space can still improve causality for machine learning. By looking at the dates of the cited papers and initiatives, like the real robot challenge[16], it becomes apparent that research is fully ongoing and the appetite for new theoretical results is far from satisfied and other applications such as denoising gravitational waves are still waiting to be engineered, for that the digital revolution, that instead of energy like the industrial revolutions before, generates, converts and processes information to a new and hopefully better world, also thanks to causality.

Acknowledgments

These notes are based on a lecture given by Dr. Bernhard Schölkopf on 10th and 24th of Nov 2020 (title: Causal Representation Learning) as part of the Causal Representation Learning seminar course at ETH Zurich during the fall semester of 2020.

References

[1] B. Schölkopf. "Causality for Machine Learning". In: *CoRR*, abs/1911.10500 (2019). arXiv:1911.10500.

[2] J. Pearl. "Causality: Models, reasoning and inference cambridge university press". In: *Cambridge, MA, USA,* 9 (2000): 10–11.

[3] G. Parascandolo et al. "Learning Independent Causal Mechanisms". In: *CoRR*, abs/1712.00961 (2017). arXiv:1712.00961.

[4] B. Schölkopf et al. "Modeling confounding by half-sibling regression". In: *Proc. Natl. Acad. Sci. USA*, 113(27) (2016): 7391–7398. DOI: 10.1073/pnas.1511656113.

[5] D. Rumelhart et al. "Learning internal representations by error propagation". In: 1986.

[6] F. Leeb et al. "Structural autoencoders improve representations for generation and transfer". In: *arXiv preprint arXiv:2006.07796* (2020).

[7] I. J. Goodfellow et al. *Generative Adversarial Networks*. 2014. arXiv:1406.2661.

[8] M. Besserve et al. "Counterfactuals uncover the modular structure of deep generative models". In: *CoRR*, abs/1812.03253 (2018). arXiv:1812.03253.

[9] M. Besserve et al. "Group invariance principles for causal generative models". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2018, pages 557–565.

[10] D. Foreman-Mackey et al. "A systematic search for transiting planets in the K2 data". In: *The Astrophysical Journal*, 806(2) (2015): 215.

[11] B. Schölkopf et al. "Removing systematic errors for exoplanet search via latent causes". In: *International Conference on Machine Learning*. 2015, pages 2218–2226.

[12] *Doppler spectroscopy - Wikipedia*. https://en.wikipedia.org/wiki/Doppler_spectroscopy. Accessed: 2020-12-12.

[13] B. T. Montet et al. "Stellar and planetary properties of K2 Campaign 1 candidates and validation of 17 planets, including a planet receiving Earth-like insolation". In: *The Astrophysical Journal*, 809(1) (2015): 25.

[14] *Water found on a potentially life-friendly alien planet*. https://www.nationalgeographic.com/science/2019/09/first-water-found-in-habitable-exoplanets-atmosphere-hubble-kepler-k2-18b. Accessed: 2020-12-07.

[15] T. D. Gebhard et al. "Physically constrained causal noise models for high-contrast imaging of exoplanets". In: *arXiv preprint arXiv:2010.05591* (2020).

[16] *Real Robot Challenge*. https://real-robot-challenge.com/. Accessed: 2020-12-07.

Appendix

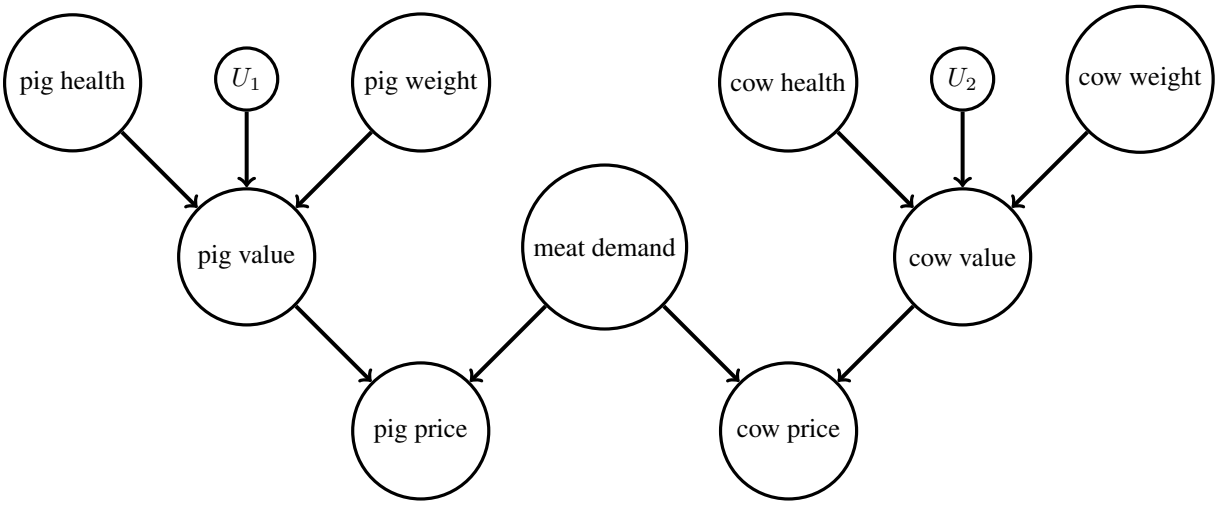

Figure 1: Structural causal model of animal prices