A Handbook for Deep Learning with their Piecemeal Intuitions from Causal Theory

Tinghao Xie
Zhejiang University*
vtu@zju.edu.cn

Abstract

Deep learning has been fully adopted in various applications nowadays. On the other hand, causality, a powerful weapon to describe the relationship between causes and effects, is gaining increasing attention. Recent works begin to adopt intuitions from causal theory in order to improve deep learning, and the results are optimistic. We first introduce causal theory basics, then classify these works as improving 1) out-of-distribution generalization; 2) generation; 3) robustness, interpretability, and fairness. In addition, we explicitly point out the causal intuitions in these works, describing what causal intuitions are embraced and how do they help improve deep learning. We hope this "handbook" may help both beginners and researchers understand the underlying causal principles and see the promising future of deep learning with causality.

1 Introduction

Causality, the relationship between an event (cause) contributing to the production of another event (effect), shows up everywhere in the world. As a well-known geographic common sense, the temperature is generally lower in cities with higher altitudes. And as we all know, a city is colder because of its higher altitude, but not the reverse (freezing the city with a refrigerator would not change the city's altitude). The underlying relationships between events revealed by causality help people understand how everything works out, but were not taken seriously before the late 20th century. Causality fascinates a variety of researchers since it is where physics, economics, biology, mathematics, and computer science intersect. In physics, causality helps explain how every physic mechanism works with one change giving rise to another; In economics, researchers study what factors lead to a certain event; Biology help explain how the human brain understand causes and effects; Mathematics formalize causal theory rigorously; Finally, computer science becomes the best playground for causality.

Meanwhile, machine learning (ML), an attractive branch of computer science, has been studied thoroughly, with deep learning as the state-of-the-art method. While the well-studied computational (statistical) learning theory provides guarantees for the training and generalization of models focusing on data's correlation, it doesn't help much on causality. As twins, causality and correlation are often confused. According to Simpson's paradox [1], any statistical correlation between two variables can be reversed when new factors are considered. A famous example was when UC Berkeley was investigated for its graduate admission sex bias. The overall admission data showed a bias for admitting more male applicants. But when observed by each department, the bias was reversed – female applicants were found more in favor than male applicants. Moreover, it's proven theoretically [2] that given any joint distribution $P_{X,Y}$ of two variables, there exists a causal model that X causes Y. According to all the evidence, it's clear that the correlation tells nothing about the causality. To describe causal relationships, causal models should be much more expressive than statistical models: 1) they allow *intervention* and *i.e.* answer questions like "What would Y be if I do X?"; 2) they could answer *counterfactual* analysis and answer questions like "What if I had acted differently, say X had not occurred?". Whilst, statistical models only answer questions on *association* levels, *e.g.* "What would Y be if I see X?".

A perspective of the relationships between ML and causality is shown in Figure 1. The generation of a cat image could be a complex effect caused by various natural factors (weather, species, *etc*) and mechanisms, eventually captured by a

[1] Completed as visiting student at the University of Oxford.

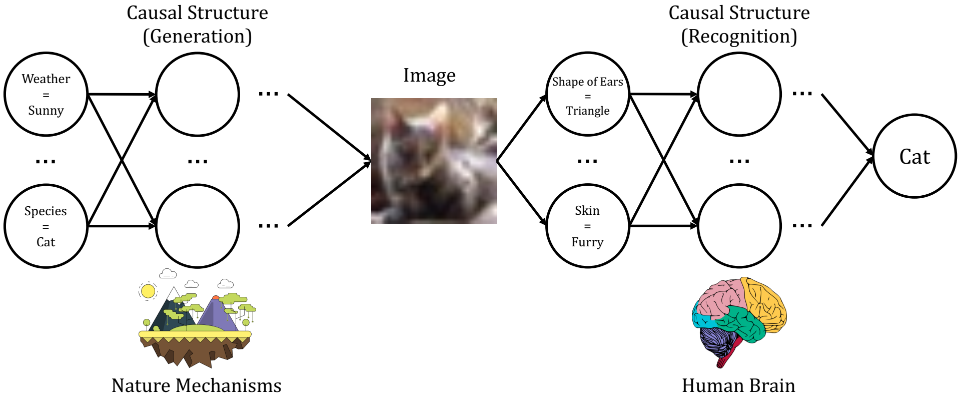


Figure 1: **ML and Causality.**

camera. On the other hand, how humans recognize the cat image is yet another causal process – we analyze different features of the image (shape of ears, types of the skin, *etc*), and at the end formulate the prediction "cat".

Generally speaking, works involving both ML and causal theory are two-fold: 1) Using ML methods to learn causal effects and relations from data (*e.g.* learn the causal structure in Figure 1); 2) Improve ML with intuitions from causal theory (*e.g.* improving the existing generation and recognition ML methods based on the knowledge of underlying causal structures). There is no clear boundary between these two aspects, but only the focused subjects (*i.e.* to learn causality or to improve ML) are different. This paper focuses on the second type (and specifically deep learning), summarizing representative works integrating inspiring causal intuitions into existing ML frameworks. This line of works remains quite open and not systematic, and our goal is to put them into a bag and to link them with their causal intuitions. Specifically, we first introduce some basic assumptions and notations in causal theory (Section 2), then describe representative works making use of causal intuitions and link them with their corresponding causal concepts (Section 3).

2 Causal Theory

In this section, we briefly introduce causal models (2.1), a crucial assumption in causal theory (2.2), and a common lapse in learning scenarios involving causality (2.3), for better explaining the works' intuitions later in Section 3.

2.1 Causal Models

As mentioned earlier, causal models should be more expressive than statistical models. [3] defines causal models formally by introducing graphs and arrows, which are known as Structural Causal Models (SCM) (or Structural Equation Models (SEM)).

Definition 1. *Structural Causal Models (SCM)*

A structural causal model is a directed acyclic graph $G = \langle \mathcal{V}, \mathcal{E} \rangle$ and a set of structural equations $\mathcal{S}$. $\mathcal{V}$ is the set of nodes describing (exogenous and endogenous) variables, $\mathcal{E}$ is the set of directed edges describing the causal relationships from one variable to another, and $\mathcal{S}$ is the set of functions quantifying the causal effects in G.

For example, if Figure 2a is the graph of a SCM, then the corresponding structural equations $\mathcal{S}$ could be:

$$Z := N_Z, \tag{1}$$
$$X := f_X(Z, N_X), \tag{2}$$
$$Y := f_Y(X, Z, N_Y). \tag{3}$$

where N_X, N_Y, N_Z are exogenous noises (representing the unknown variables from the external environment) of the endogenous (observed) variables X, Y, Z, and f_X, f_Y are causal mechanisms.

2.2 Independent Mechanisms

Independent mechanisms are an essential assumption in causal theory. Intuitively, natural physical mechanisms are generally independent of each other. For example, even though heavy rain might promote grass growth in your lawn, the mechanism of rain and the mechanism of grass growing in your lawn have nothing to do with each other – knowing (or changing) the rain formulation process won't tell you (or alter) anything about the grass-growing mechanism w.r.t. the wetness in your lawn.

Back to our first example with only two variables, altitude (X) and temparature (Y), we may have a large dataset telling us their joint distribution $P(X, Y)$. The SCM of such a two-variable system could be modeled as Figure 2b, with its structural equations being $X := N_X, Y := f_Y(X, N_Y)$. Obviously, we may decompose the joint distribution $P(X, Y) = P(Y|X) \cdot P(X)$. Since we know it's X causes Y but not the reverse, we claim that $P(Y|X)$ is the causal mechanism from altitude to temperature – "how the altitude of a city determines its temperature". Here, the underlying independent mechanisms could be interpreted as: 1) altering $P(X)$ does not bother $P(Y|X)$, vice versa; 2) knowing $P(X)$ does not reveal any information about $P(Y|X)$, vice versa; 3) the noise variables, N_X and N_Y, are independent (or only weakly dependent). Notice that the independence of the two distributions $P(Y|X)$ and $P(X)$ is valid only when we decompose in the causal direction. If we decompose the joint distribution in another way, *i.e.* $P(X, Y) = P(X|Y) \cdot P(Y)$, we would find the $P(X|Y)$ and $P(Y)$ are dependent – *e.g.* changing $P(Y)$ leads to a completely different $P(X|Y)$.

In a causal model with multiple variables, it's assumed similarly, that the mechanism of each variable is independent of one another. [2] states the three aspects of independent mechanisms assumption:

1. The causal mechanisms are autonomous, modular, and invariant, while possible to perform intervention by controlling the mechanism inputs, and able to transfer to different domains.
2. The information contained in mechanisms are independent of each other.
3. The exogenous noises are independent (or weakly dependent) of each other.

This assumption could be quite general. For example, in Figure 1, both the nature generative mechanisms and the recognition mechanisms of human brains could be assumed to be independent.

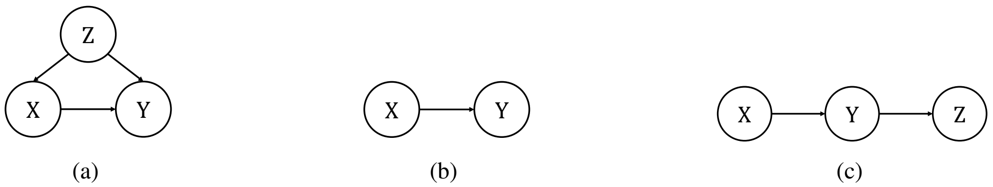
2.3 Spurious Correlation

Spurious correlations are a common lapse made in statistical learning. According to [3], spurious correlations are "correlations that do not imply causation." An example of cows and camels [4] best demonstrates this issue. Suppose there is a dataset of cows and camels, where most pictures of cows have green pastures as backgrounds, while most camels are accompanied by deserts. A simple deep neural network (DNN) trained on this dataset could falsely focus on the background to make predictions – it may simply answer "cows" when seeing green pastures and "camels" when seeing deserts. On the one hand, this obviously hit high scores on the train set; On the other hand, recognizing pastures or deserts is much easier than cows or camels. But when tested on images of cows standing on sandy beaches, the DNN classifier would always fail. Here, the DNN classifier falls into spurious correlations, since it's not the background causing the image label to be "cow" or "camel", but the creature does.

We may also formulate this example into a SCM as Figure 2c. Consider the case that a human is synthesizing the dataset: the human is given some pixels of a creature (a cow or a camel) and is then asked to add a background (*e.g.* from Internet resources) for the creature. It's plausible to assume that the human intends to believe a cow should be on a green pasture and a camel should be in a barren desert, though in the real world this is might not be true. The human also labels the creature as either "cow" or "camel". Then in Figure 2c, X is the image pixels of the creature (a cow or a camel), Y is the label of the image ("cow" or "camel") provided by the human, and Z is the background pixels picked by the human.

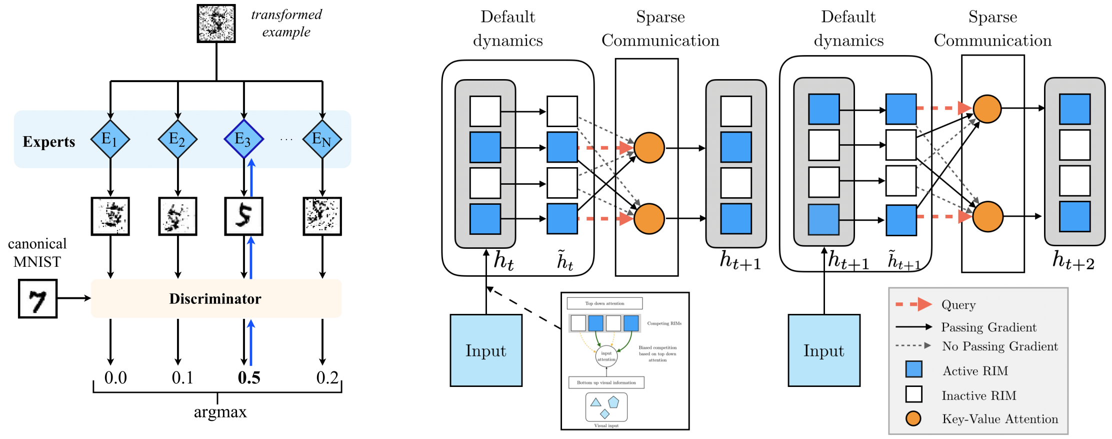
Then as stated, we can see the correlation learned by the classifier is from Z to Y (using background pixels to decide the image label), which is not the causal direction and therefore a spurious correlation. Although this spurious correlation causes no problem when the test set's distribution is the same as the train set (cows with pastures and camels with deserts), but would fail severely whenever the distribution shifts (*e.g.* another human insists cows should be at bullpens and camels should be at zoos), *i.e.* the causal mechanism $Y \rightarrow Z$ alters.

Obviously, the correct causal correlation is the one from X to Y. Once the classifier learns this correct correlation, it could stay invariant and transfer to different domains, *i.e.* generalizing well on images of cows and camels with various backgrounds besides pastures and deserts. This capability is referred to as out-of-distribution (OOD) generalization.



(a) (b) (c)

Figure 2: **Examples of SCM.**



(a) Learning Independent Causal Mechanisms [5]

(b) Recurrent Independent Mechanisms [6]

Figure 3: **Improving OOD Generalization by Modular Design.**

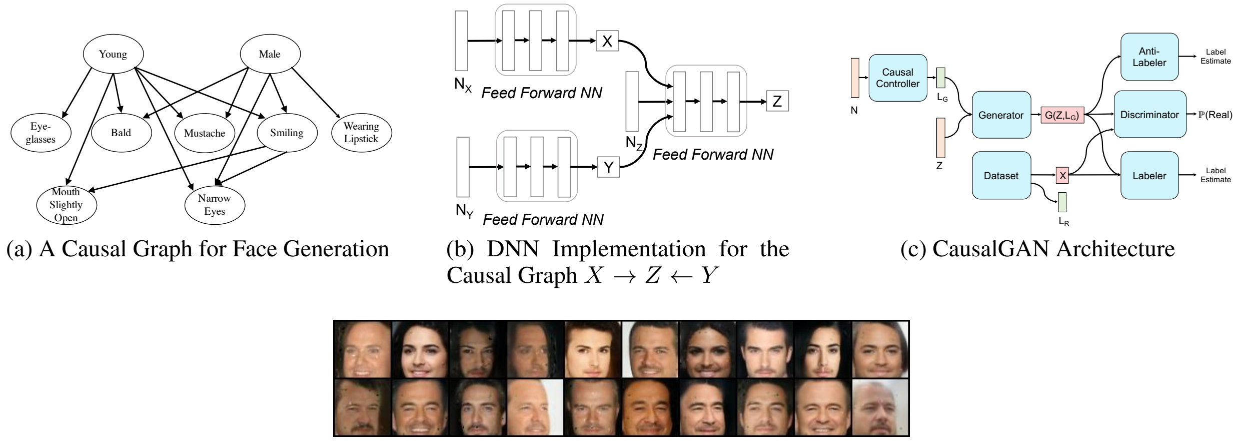
3 Deep Learning with Causal Intuitions

In this section, we introduce some representative works adopting causal intuitions into deep learning and thus improving the latter. We classify these works into three parts: 1) improving the out-of-distribution generalization ability of deep models (3.1); 2) improving the deep generation model (3.2); 3) improving the interpretability and fairness of deep models (3.3).

3.1 Out-Of-Distribution Generalization

To this end, we have mentioned the concept of spurious correlation and the ability of OOD generalization in Section 2.3. It turns a majority of work easing the issue of spurious correlation (and therefore improving OOD generalization of deep models) adopt crucial ideas from causal theory. These works mostly focus on machine learning recognition (the right part of Figure 1).

A work directly dealing with spurious correlation is **Invariant Risk Minimization (IRM)** [7]. The authors first point out the problem underlying conventional training with empirical risk minimization: the model could incorrectly learn an easier shortcut to make predictions (*e.g.* the background in the cows and camels example), and thus cannot stand any distribution shifts (*e.g.* changing the distribution of the background). Specifically, they explicitly assume there is a complete real-world environment (each representing a different domain or distribution) set $\mathcal{E}_{all}$, but the training set is only sampled from a limited environment subset $\mathcal{E}_{train} \subset \mathcal{E}_{all}$. Their goal is forcing deep models to learn the correct correlation from $\mathcal{E}_{train}$ but stands correct for $\mathcal{E}_{all}$. We may roughly interpret their goal as to learn a causal mechanism, which is independent of other mechanisms in the implicit causal structure. In other words, the learned mechanism should be *invariant* when the environment changes (*e.g.* when the cows are never standing on pastures anymore). Their original attempt to reach such a goal is to **minimize the maximum risk w.r.t. each environment**. Mathematically, $\min_{\mathbf{w}} \max_{e \in \mathcal{E}_{train}} R^e(f_{\mathbf{w}})$, where f and $\mathbf{w}$ are the deep model and its parameters, R^e is the risk function (expectation of loss) under environment e. They further instantiate such an intractable optimization problem into a practical one (omitted here). Their experiments on a synthetic dataset and colored MNIST further validate the generalization abilities of IRM across multiple environments.



(a) A Causal Graph for Face Generation

(b) DNN Implementation for the Causal Graph $X \to Z \leftarrow Y$

(c) CausalGAN Architecture

(d) Generated Examples (Top: Intervened on Mustache = 1; Bottom: Conditioned on Mustache = 1)

Figure 4: **CausalGAN [12].**

Other research works improve the OOD generalization ability of ML indirectly, by **modular design**. [5] adopts independent mechanisms into the framework design. They first distort their input images by a set of unknown transformations (in their example, each input image is either noised or translated to a certain direction, and the exactly applied transformation is not recorded or labeled), then feed the input to several independent "experts" (modules) which attempt to undo the deliberate transformation. As shown in Figure 3a, their framework composes of N completely independent experts, which are modeled by convolutional neural networks (CNNs), followed by a discriminator (borrowed from Generative Adversarial Nets (GAN) [8]) to pick the most plausible expert to produce the final de-transformed output. In another sentence, the experts compete for each input and only one expert takes effect, eventually undoing the transformation. Here, each expert could be viewed as an independent causal mechanism. Changing one mechanism does not cause the change of another mechanism, so do the experts. Furthermore, their experiments show the framework's OOD generalization ability – even though they only train experts on MNIST [9], the experts generalize to the Omniglot letters dataset [10]. **Recurrent Independent Mechanisms (RIMs)** [6] applies the intuition of independent mechanisms into their recurrent architecture, see Figure 3b. Their framework sets up independent recurrent units (named RIMs), which likewise compete for the input at each time step. To be mentioned, these units compete according to the (soft-)attention mechanism [11]. Only a fixed number of RIMs with the highest attention scores are activated and allowed to forward the current input according to its internal dynamics, while other non-activated RIMs stay unchanged. Moreover, they introduce *sparse communications* among RIMs, determined by top-k attention. Their results demonstrate RIMs' strong specializing ability (only a few RIMs are effecting, not all) over temporal patterns and objects, along with the OOD generalization ability when the distribution shifts. Again, each modular RIM could be thought of as a causal mechanism, which specializes in dealing with a certain feature.

While [7] embraces the independence and invariance of causal mechanisms by reforming the optimization target, both [5] and [6] turn the idea of independent mechanisms into an inductive bias of modular design (to simulate the possibly modular structure of human brains while recognizing things). It's obvious that these works simultaneously adopt the first two aspects of independent mechanisms assumption in Section 2.2. We also point out that [6] actually takes the third aspect into their design – exogenous noises may be weakly dependent, and so are the RIMs (which are allowed to sparsely communicate with each other).

3.2 Generation

See the left part of Figure 1, another attractive perspective is to study the causal generative mechanisms in nature, *e.g. what's the causes of the picture that a cat is lying in the sun*. In deep learning practice, this may correspond to deep generative networks, *e.g.* Generative Adversarial Nets (GAN) [8] and Variational Autoencoders (VAEs) [13].

Following the famous conditional GAN [14], authors of [12] propose **CausalGAN** (Figure 4). The spirit of CausalGAN is to simulate a hypothesized causal generative model (*e.g.* Figure 4a) with a carefully designed neural network (as shown in Figure 4b), which corresponds to the structural causal model. This neural network is known as a *causal controller*. Firstly, the causal controller is trained to map a set of independent noises to (binary) labels L_G (*e.g.* Mustache). The

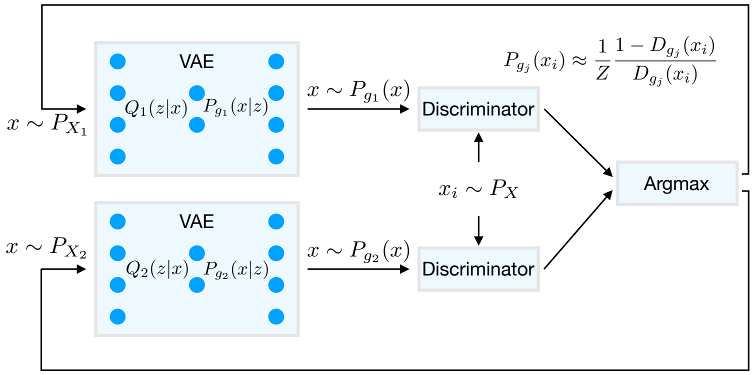


Figure 5: **Competitive Training of Mixtures of Independent Deep Generative Models [15].**

overall CausalGAN architecture is demonstrated in Figure 4c. The causal controller is then fixed, always providing generated labels L_G to the generator. The rest of the architecture is a *conditional generative network*, which could generate conditioned data from $P_G(X|L = L_G) \approx P_R(X|L = L_G)$. The training fashion for the conditional generative network is as conventions and thus omitted here. The generated results (Figure 4d) are impressive, since the labels L_G could be intervened, force *e.g.* Mustache $= 1$. "Intervention" is different from "condition", and the results in Figure 4d directly manifest this – by intervention on Mustache $= 1$ CausalGAN could generate images of females with mustaches, while condition on Mustache $= 1$ could only generate males with mustaches.

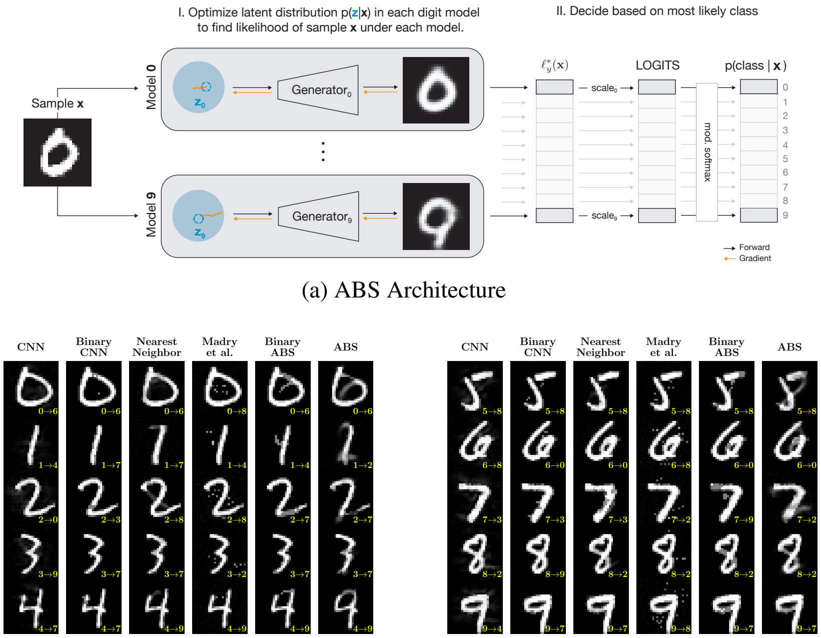
We reveal the causality in CausalGAN from two perspectives. Firstly, the causal controller module (a neural network with a specified structure) *explicitly* simulates an assumed causal graph (Figure 4a, 4b), and is capable of mapping independent exogenous noises N to labels L_G (*i.e.* generating the labels). Secondly, the conditional generative network takes L_G as input, *inherently* and *implicitly* simulating (a part of) the natural causal mechanisms, mapping labels L_G (and noises Z) to images. This architecture subsequently allows interventions on L_G (*e.g.* forcing Male $= 0$ and Mustache $= 1$), the observed endogenous variables, and generating imaginative but still plausible images.

Another work [15] makes the independent mechanisms assumption for data generation, *i.e.* $P_{model} = \sum_{j=1}^{k} \alpha_j P_{g_j}$. Similar to the works ([5], [6]) mentioned in 3.1, they set up several independent "expert" generative models, and let the models compete for training points in order to force them to learn different parts of the data distribution. Specifically, given K as a hyperparameter, indicating the assumed number of disentangled mechanisms for the entire data distribution, they set up K independent VAEs $g_j, j \in [K]$. Also, they set K discriminators $D_j, j \in [K]$, estimate $P_{g_j}(x)$ with D_j, and use $c_j = \mathbb{I}(j = \arg\max_i P_{g_i}(x))$ to ensure the specialization of each mechanism on a single mode. By their training pipeline (Figure 5), they managed to split complex data distributions into K modes, each handled by an independent generator neural network. Obviously again, they adopt the independent mechanisms as their essential intuition, by setting up K experts for generating different modes of the entire distribution (*e.g.* for the distribution in MNIST, they could set up 10 experts that generate the 10 different digits respectively).

3.3 Robustness, Interpretability, and Fairness

In the era of deep learning, the robustness, interpretability, and fairness of deep models are gaining increasing attention. The three topics are usually brought up together since they are the direct concerns when a DNN is about to be deployed in the wild. But unlike Section 3.1 and 3.2, related works in this section are not usually borrowing just some tiny fractions from causal theory, but the entire significance of causality. Some views claim that causality is a man-made concept, and causal explanations are used to pass responsibilities. This means that introducing causality into ML might possibly help people understand models' behaviors (*interpretability*), therefore rectifying them to make fair decisions (*fairness*), and reducing the chance of them making errors due to some human imperceptible perturbations (*robustness*).

It's known that DNN classifiers are susceptible to *adversarial attacks* [16]. That is, though DNN classifiers could hopefully make a correct prediction on a clean input, but would make an error when the clean input is maliciously perturbed (by adding a human-invisible noise). On one hand, one can think of the adversarial perturbed samples are from a different distribution compared with the training distribution (*i.e.* adversarial attack is yet another special case of out-of-distribution problems). On the other hand, humans never make wrong predictions due to invisible noises, which means that current DNN classifiers are quite different from human cognition. Many defenses against adversarial attacks are proposed, *e.g.* adversarial training [17] and neural network verification [18][19]. But these works and their



(a) ABS Architecture

(b)

(c)

Figure 6: **Towards the first adversarially robust neural network model on MNIST [21].** Adversarial examples for the ABS models (the 6th column in 6b and 6c) are perceptually meaningful.

follow-ups are not making improvement satisfying enough against adversarial attacks – either the clean accuracy drops too much or the adversarial attack success rate stays non-trivial. It could be assumed ([20]) that if a model learns the classification task in the natural causal generative direction, adversarial attacks would be impossible or much harder to cast on the model. **Analysis by Synthesis (ABS)** [21] models the causal generative process class label → image in MNIST, obtain the generative distribution $P(x|y)$, and make the final prediction with $P(y|x) \propto P(x|y)P(y)$ (see Figure 6a). Their experiment results on MNIST comply with their causal intuition – adversarial attacks are more difficult, and the adversarial noises against ABS are larger, while having clear semantical meaning (See Figure 6b and 6c). Actually, the adversarial examples could be interpreted as counterfactuals, *i.e.* answering the question *what would the input looks like if the class label is 9, not 4?* The design of ABS clearly shares the benefits and advantages of causal models.

Interpretability and fairness play important roles in human-centric scenarios. For example, if a DNN is deployed to decide whether to grant loans to loan applicants, according to their attributes of age, annual income, *etc.* As the model beholder, we want to ensure that whenever the model denies an applicant, it's not because of the applicant's race or any other unfair and non-causal attributes. Hence, we demand to interpret the model and force it to act fairly. There are various traditional explanation and interpretability methods (*e.g.* saliency maps [22]), which are not mentioned here. We focus on causal interpretability and fairness.

Suppose a model takes the input X and outputs the prediction Y. There are three levels of interpretability according to [23]: 1) *statistical*, *e.g.* "How does seeing X make me believe Y"; 2) *interventional*, *e.g.* "What would the expectation of Y be if we force intervention on $X = 1$?"; 3) *counterfactual*, *e.g.* "What would Y be if we acted differently last time, say let $X = 1$?". [24] has already provided a comprehensive survey of the causal interpretability and fairness for ML, where the authors classify the existing works into four categories: 1) model-based interpretations; 2) counterfactual explanations generators; 3) fairness; 4) verifying causal relationships. Some of the awesome works include:

- [25] considers a DNN as a SCM, and builds a SCM as an abstraction for the DNN. They then quantitatively rank the filters of a convolution layer according to their counterfactual importance.

- [26] uses a generative model (specifically GAN) to generate counterfactual visual image explanations for a classification model. That is, *what would the given image of a woman (not wearing glasses) look like, if the prediction is eyeglass = True?*

- [27] designs a framework where an encoder is trained to learn low-dimensional latent representations (*e.g.* α for "shape" and β for "color") of an input image, and a generator is trained to map the latent variables to an image. Their decoder could disentangle an image into causal variables, and their encoder could simulate the causal generation mechanisms. A huge benefit of their framework is that the causal variables, *i.e.* latent representations, could be manually intervened and the generator could generate extremely descriptive explanation images.
- [28] proposes a criterion w.r.t. fairness – they define that a decision is fair towards an individual if it is the same in (a) the actual world and (b) a counterfactual world where the individual belonged to a different demographic group. They further develop a framework for modeling fairness using tools from causal inference.
- . . .

4 Conclusion

In this paper, we summarize some causal intuitions adopted in deep learning. Specifically, we first introduce the basics of causal theory, then bring up some representative works improving ML (deep learning) by adopting intuitions from the causal theory. We classify these works as improving 1) out-of-distribution generalization; 2) generation; 3) robustness, interpretability, and fairness. Also, we explicitly point out the causal intuitions (corresponding to Section 2) in these works, describing what causal intuitions are embraced and how do they help improve deep learning. From Section 3, we see causality as a great adjuvant. It's not hard to see – following causality to improve popular deep learning techniques was, is, and will be a bright and promising path.

References

[1] Colin R Blyth. On simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, 67(338):364–366, 1972.

[2] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[3] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[4] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018.

[5] Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla, and Bernhard Schölkopf. Learning independent causal mechanisms. In *International Conference on Machine Learning*, pages 4036–4044. PMLR, 2018.

[6] Anirudh Goyal, Alex Lamb, Jordan Hoffmann, Shagun Sodhani, Sergey Levine, Yoshua Bengio, and Bernhard Schölkopf. Recurrent independent mechanisms. *arXiv preprint arXiv:1909.10893*, 2019.

[7] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[9] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[10] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

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

[12] Murat Kocaoglu, Christopher Snyder, Alexandros G Dimakis, and Sriram Vishwanath. Causalgan: Learning causal implicit generative models with adversarial training. *arXiv preprint arXiv:1709.02023*, 2017.

[13] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[14] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[15] Francesco Locatello, Damien Vincent, Ilya Tolstikhin, Gunnar Rätsch, Sylvain Gelly, and Bernhard Schölkopf. Competitive training of mixtures of independent deep generative models. *arXiv preprint arXiv:1804.11130*, 2018.

[16] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[17] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

[18] Radosiaw R Zakrzewski. Verification of a trained neural network accuracy. In *IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No. 01CH37222)*, volume 3, pages 1657–1662. IEEE, 2001.

[19] Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. In *International Conference on Computer Aided Verification*, pages 97–117. Springer, 2017.

[20] Niki Kilbertus, Giambattista Parascandolo, and Bernhard Schölkopf. Generalization in anti-causal learning. *arXiv preprint arXiv:1812.00524*, 2018.

[21] Lukas Schott, Jonas Rauber, Matthias Bethge, and Wieland Brendel. Towards the first adversarially robust neural network model on mnist. *arXiv preprint arXiv:1805.09190*, 2018.

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

[23] Judea Pearl. Theoretical impediments to machine learning with seven sparks from the causal revolution. *arXiv preprint arXiv:1801.04016*, 2018.

[24] Raha Moraffah, Mansooreh Karami, Ruocheng Guo, Adrienne Raglin, and Huan Liu. Causal interpretability for machine learning-problems, methods and evaluation. *ACM SIGKDD Explorations Newsletter*, 22(1):18–33, 2020.

[25] Tanmayee Narendra, Anush Sankaran, Deepak Vijaykeerthy, and Senthil Mani. Explaining deep learning models using causal inference. *arXiv preprint arXiv:1811.04376*, 2018.

[26] Shusen Liu, Bhavya Kailkhura, Donald Loveland, and Yong Han. Generative counterfactual introspection for explainable deep learning. In *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1–5. IEEE, 2019.

[27] Matthew O'Shaughnessy, Gregory Canal, Marissa Connor, Mark Davenport, and Christopher Rozell. Generative causal explanations of black-box classifiers. *arXiv preprint arXiv:2006.13913*, 2020.

[28] Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *arXiv preprint arXiv:1703.06856*, 2017.