EFFICIENTLY DISENTANGLE CAUSAL REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we propose a novel approach to efficiently learning disentangled representations with causal mechanisms, based on the difference of conditional probabilities in original and new distributions. We approximate the difference with model's generalization abilities so that it fits in standard machine learning framework and can be efficiently computed. In contrast to the state-of-the-art approach, which relies on learner's adaptation speed to new distribution, the proposed approach only requires evaluating the generalization ability of the model. We provide theoretical explanation for the advantage of the proposed method, and our experiments show that the proposed technique is 1.9–11.0$\times$ more sample efficient and 9.4–32.4$\times$ quicker than the previous method on various tasks. The source code is in supplementary material.

1 INTRODUCTION

Causal reasoning is a fundamental tool that has shown great impact in different disciplines (Rubin & Waterman, 2006; Ramsey et al., 2010; Rotmensch et al., 2017), and it has roots in work by David Hume in the eighteenth century (Hume, 2003) and in classical AI (Pearl, 2003). Causality has been mainly studied from a statistical perspective (Greenland et al., 1999; Pearl, 2009; 2018) with Judea Pearl's work on the causal calculus leading its statistical development.

More recently, there has been a growing interest to integrate statistical techniques into machine learning to leverage their benefits. Welling raises a particular question about how to disentangle correlation from causation in machine learning settings to take advantage of the sample efficiency and generalization abilities of causal reasoning (Welling, 2015). Although machine learning has achieved important results on a variety of tasks like computer vision and games over the past decade (e.g., Mnih et al. (2015); Silver et al. (2017); Szegedy et al. (2017); Hudson & Manning (2018)), current approaches can struggle to generalize when the test data distribution is much different from the training distribution (common in real applications). Further, these successful methods are typically "data hungry", requiring an abundance of labeled examples to perform well across data distributions.

In statistical settings, encoding the causal structure in models has been shown to have significant efficiency advantages. For example, in the context of speech recognition, Chorowski et al. (2014) and Bengio (2017) show that a model trained to learn $P(\text{acoustics})$ performs worse than a model trained to also learn a conditional distribution containing phonemes, $P(\text{acoustics}|\text{phonemes}) \cdot P(\text{phonemes})$. The model performs better when the causality—phonemes cause acoustics—is encoded in the model. Further, Schölkopf et al. (2012) show that encoding the causal structure leads to more robust estimation (and thus, the performance improves). Independent of any change in $P(\text{phonemes})$, the causal mechanism $P(\text{acoustics}|\text{phonemes})$ is not expected to change across tasks or datasets. These robustness and generalization characteristics lead to quicker training, because once the causal mechanism is learned, it requires minimal updates to transfer to new data.

In support of the advantages of encoding causal mechanisms, Bengio et al. (2020) recently introduced an approach to disentangling causal relationships in end-to-end machine learning by comparing the adaptation speeds of separate models that encode different causal structures. The core idea is explained with an atmospheric physics example of learning $P(A, B)$, where A (**Altitude**) causes B (**Temperature**) (Peters et al., 2016). The marginal probability of A can be changed depending on, for instance, the place (Switzerland to less mountainous country like Netherlands), but $P(B|A)$ as the underlining causality mechanism does not change. Therefore, the causality structure can be

inferred from the robustness of predictive models on out-of-distribution (Peters et al., 2016; 2017). The invariance of $P(B|A)$ across train and transfer distributions is the key assumption in their work, and we also follow it in this work.

$P(A, B)$ can be either modeled as $P(B|A) \cdot P(A)$ or $P(A|B) \cdot P(B)$, reflecting the possible causal directions. Since $P(B|A) \cdot P(A)$ reflects the right causal structure, it is more efficient to adapt to a change in $P(A)$ since parameters for $P(B|A)$ need not to be updated. On the other hand, the model that assumes the opposite causal direction, $P(A|B) \cdot P(B)$, must adapt both its marginal and conditional distribution parameters, leading to slower adaptation speed. The previous method proposed the state-of-the-art meta-learning approach to disentangling causal mechanisms using the adaptation speed as a meta-objective to optimize. In particular, the task is to determine whether the causal direction is $A \to B$ or $B \to A$.

In this work, we propose a more efficient approach to disentangling causal mechanisms than the state-of-the art one based on the same assumption. The baseline approach essentially captures the changes in $P(B|A)$ and $P(A|B)$, by examining how fast the joint probability models can adapt or learn after intervention. In contrast, our approach only evaluates the generalization ability of the model without the need of the adaptation process, therefore it is much quicker in disentangling the causal mechanisms. Our experiments show the proposed technique is 1.9–11.0$\times$ more sample efficient and 9.4–32.4$\times$ quicker than measuring adaptation speed on various tasks. We also argue that the proposed approach has less hyper parameters and it is straight-forward to implement the approach within the standard machine learning workflows.

Our contributions can be summarized as follows.

- We propose an efficient approach to disentangling representations for causal mechanisms by measuring generalization.
- We theoretically prove that the proposed estimators can identify causal direction and disentangles causal mechanisms.
- We empirically show that the proposed approach is significantly quicker and more sample efficient for various tasks. Sample efficiency is important when data size in transfer distribution is small.

2 Approach

To begin, we reflect on the tasks and the disentangling approach (as baseline) described by previous work (Bengio et al., 2020). The invariance of conditional distribution for the correct causal direction $P(B|A)$ is the key assumption in their work, and we also follow it in this work. We notice that their baseline approach compares the adaptation speed of models on a transfer data distribution, and hence requires significant time for adaptation. We propose an approach to learn causality mechanisms by directly measuring the changes in conditional probabilities before and after intervention for both $P(B|A)$ and $P(A|B)$. Further, we optimize the proposed approach to use generalization loss rather than a divergence metric—because loss can be directly measured in standard machine learning workflows—and we show that it is likely to correctly predict causal direction and disentangle causal mechanisms.

2.1 Causality Direction Prediction

As the first step towards learning disentangled representations for causal mechanisms, we start with the binary classification task. Given two discrete variables A and B, we want to determine whether A causes B, or vice-versa. The training (transfer) data contains samples (a, b) from training (transfer) distribution, P_1 (P_2). The baseline approach defines models that factor the joint distribution $P(A, B)$ into two causality directions $P_{A\to B}(A, B) = P_{A\to B}(B|A)P_{A\to B}(A)$ and $P_{B\to A} = P_{B\to A}(A|B)P_{B\to A}(B)$. It then compares their speed of adaptation to transfer distribution.

Intuitively, the factorization with correct causality direction should adapt more quickly to the transfer distribution. Suppose $A \to B$ is the ground-truth causal direction. For the correct factorization, they assume that conditional distribution $P_{A\to B}(B|A)$ does not change between the train and transfer distributions, so that only the marginal $P_{A\to B}(A)$ needs adaptation. In contrast, for the factorization

with incorrect causality direction, both the conditional $P_{B\to A}(A|B)$ and marginal distributions $P_{B\to A}(B)$ need adaptation. They analyze that updating a marginal distribution $P(A)_{A\to B}$ is likely to have lower sample complexity than the conditional distribution $P_{B\to A}(A|B)$ (only a part of $P_{B\to A}(A, B)$) because the later has more parameters. Therefore, the model with correct factorization will adapt more quickly, and causality direction can be predicted from adaptation speed.

To leverage this observation, the baseline method defines a meta-learning objective. Let $\mathcal{L}_{A\to B}$ and $\mathcal{L}_{B\to A}$ be the log-likelihood of $P_{A\to B}(A, B)$ and $P_{B\to A}(A, B)$, respectively. Their baseline approach optimizes a regret, $\mathcal{R}$, to acquire an indicator variable $\gamma \in \mathbb{R}$. If $\gamma > 0$, the prediction is $A \to B$, otherwise $B \to A$.

$$\mathcal{R} = -\log[\sigma(\gamma)\mathcal{L}_{A\to B} + (1 - \sigma(\gamma))\mathcal{L}_{B\to A}] \tag{1}$$

2.2 Proposed Mechanism: Observe the Model's Conditional Distribution Divergence

Rather than relying on gradients of end-to-end predictions to identify causal direction, we propose to directly observe the divergence of each model's conditional distribution predictions under the intervention of the transfer dataset.

We consider the distribution of a single model's predictions on the training and transfer distributions. These distributions depend on both the marginal and conditional distributions learned by the models. It is expected that a model encoding the correct causal direction will have learned a correct conditional distribution, but the marginal distribution might shift from the training to transfer datasets. Hence, we would like to ignore changes in the model's predictions that are due to marginal distribution differences and focus directly on the conditional distribution.

Given access to the model's conditional distribution predictions, we could use the KL-divergence to directly measure the conditional distribution differences between the training and transfer datasets. A model that encodes the correct causal structure should show no change in its conditional distribution predictions, so the divergence of these predictions should be small. On the other hand, a model that encodes the incorrect causal structure is likely to show significantly large divergence. We tie these observations together in the following proposition:

Proposition 1. *Given two data distributions with the same directed causality between two random variables, A and B, the difference of the KL-divergences of their conditional distributions is an unbiased estimator of the correct causality direction between A and B.*

This approach shares the same intuition as the baseline approach. The model with correct causal direction should only witness minimal changes in the conditional distribution of its predictions between train and transfer distributions. Thus, the KL-divergence of the predictions should not change for models with correct causal structure. The proof for Proposition 1 is included in Appendix A.

2.3 Proposed Simplification: Generalization Loss Approximates the Divergence

Although we can use the KL-divergence of the conditional probabilities of two data distributions as an unbiased estimator of causality direction, it requires additional computation to model conditional distribution in transfer distribution. However, in many practical end-to-end learning settings, it is more efficient and straight-forward to acquire generalization loss, such as the cross-entropy, that can be used to approximate the conditional KL-divergence.

More specifically, we compare the generalization gaps between two causal models to approximate the causality direction. Let the generalization gap, $\mathcal{G}$, be the difference of a model's losses on the training and transfer datasets: $\mathcal{G}_{A\to B} = \mathcal{L}^{\text{transfer}}_{A\to B} - \mathcal{L}^{\text{train}}_{A\to B}$. Here, $\mathcal{L}_{\dot{A}\to B}$ is the loss on the specified set. Further, we define a directionality score as the difference of generalization gaps:

$$\mathcal{S}_{\mathcal{G}} = \mathcal{G}_{B\to A} - \mathcal{G}_{A\to B}$$

We show that, for an appropriately chosen loss, such as cross-entropy, $\mathcal{S}_{\mathcal{G}}$ is a biased but reasonable estimator of the correct causal direction. When $\mathcal{S}_{\mathcal{G}} > 0$, it is likely that $A \to B$ is the correct causal direction–the generalization gap for the incorrect causal model dominates the score. We formalize this notion in Proposition 2 in Appendix B:

Proposition 2. *Given two data distributions with the same directed causality between two random variables, A and B, the difference of their generalization gaps is a biased estimator of the correct causality direction between A and B.*

In practice for the tasks we test in next section, we find that although $\mathcal{S}_\mathcal{G}$ is biased, it always indicates the correct causal direction. An intuitive understanding is that this approach measures how well models trained on train distribution can predict in transfer distribution–their generalization ability. Algorithm 1 summarizes the process for identifying a correct causal model. In Appendix C, we describe that the generalization-based approach should converge more quickly than a gradient-based approach, especially in practical settings.

Algorithm 1 The proposed approach for causality direction prediction.

1: Train $f_{A\to B}$ on training data, and get train loss $\mathcal{L}^{\text{train}}_{A\to B}$.
2: Train $f_{B\to A}$ on training data, and get train loss $\mathcal{L}^{\text{train}}_{B\to A}$.
3: Get transfer loss $\mathcal{L}^{\text{transfer}}_{A\to B}$ with $f_{A\to B}$ on transfer data.
4: Get transfer loss $\mathcal{L}^{\text{transfer}}_{B\to A}$ with $f_{B\to A}$ on transfer data.
5: Get generalization loss $\mathcal{G}_{A\to B} = \mathcal{L}^{\text{transfer}}_{A\to B} - \mathcal{L}^{\text{train}}_{A\to B}$.
6: Get generalization loss $\mathcal{G}_{B\to A} = \mathcal{L}^{\text{transfer}}_{B\to A} - \mathcal{L}^{\text{train}}_{B\to A}$.
7: Compute $\mathcal{S}_\mathcal{G} = \mathcal{G}_{B\to A} - \mathcal{G}_{A\to B}$.
8: If $\mathcal{S}_\mathcal{G} > 0$ return $A \to B$, else return return $B \to A$.

The baseline approach and our proposed approach are based on the same intuition of unchanging conditional distributions for correct causality direction. However, the baseline approach must observe gradients during transfer distribution training, while the proposed approach looks directly at the changes in model outputs on the transfer data distribution. Informally, we note that these approaches are roughly equivalent by observing the following: For the correct causal direction,

$$\mathcal{G}_{\cdot\to\cdot} = \mathcal{L}^{\text{transfer}}_{\cdot\to\cdot} - \mathcal{L}^{\text{train}}_{\cdot\to\cdot} \qquad\qquad \nabla\mathcal{G}_{\cdot\to\cdot} = \nabla\mathcal{L}^{\text{transfer}}_{\cdot\to\cdot} - \nabla\mathcal{L}^{\text{train}}_{\cdot\to\cdot}$$

Since $\nabla\mathcal{L}^{\text{train}} = 0$, $\nabla\mathcal{G}_{\cdot\to\cdot} = \nabla\mathcal{L}^{\text{transfer}}_{\cdot\to\cdot}$. The previous work shows that $\nabla_{\theta_i}\mathcal{L}^{\text{transfer}}_{\cdot\to\cdot} = 0$ for model parameters, θ_i, representing correct causal structures (i.e., joint distribution), so $\nabla\mathcal{L}^{\text{transfer}}_{\cdot\to\cdot}$ should also be small. Similar to that, we expect $\mathcal{G}_{\cdot\to\cdot}$ to be small relative to the incorrect causal model.

2.4 Representation Learning

So far, we assume that the causal variables are already disentangled in the raw inputs. We are interested in extending it to situations where the variables are entangled in input, such as in pixels or sounds. Here, we want to learn representations that disentangle causal variables.

Following the previous work, we suppose the true causal variables (A, B) generate input observation (X, Y) with ground truth decoder, $\mathcal{D}$. Then, we want to train an encoder, $\mathcal{E}$, that converts input (X, Y) to a hidden representation, (U, V). Variables A, B, X, Y, U, V are all scalars, and the decoder and encoder are rotation matrices.

$$\begin{bmatrix} X \\ Y \end{bmatrix} = R(\theta_\mathcal{D}) \begin{bmatrix} A \\ B \end{bmatrix} \qquad\qquad \begin{bmatrix} U \\ V \end{bmatrix} = R(\theta_\mathcal{E}) \begin{bmatrix} X \\ Y \end{bmatrix}$$

Then we treat (U, V) as observed input, and build causal modules on them in a way similar to causality direction prediction problem (Section 2.1). If the encoder is learned correctly, we expect to obtain $(U, V) = (A, B)$ with $\theta_\mathcal{E} = -\theta_\mathcal{D}$, or $(U, V) = (B, A)$ with $\theta_\mathcal{E} = \theta_\mathcal{D}$. The baseline approach addresses this problem with loss function in Equation 1, but extending meta-parameters to encoder parameters in the meta-learning framework.

We propose extending our method to learn representation without using adaptation process. Suppose the conditional distributions $P(V|U)$ is modeled by $f_{U\to V}$, and $P(U|V)$ by $f_{V\to U}$. Also, the original losses are $\mathcal{L}_{U\to V}$ for $f_{U\to V}$ and $\mathcal{L}_{V\to U}$ for $f_{V\to U}$ on train data. We optimize the following objective.

$$\mathcal{L} = \mathcal{L}_{U\to V} + \mathcal{L}_{V\to U} + \lambda \min\{\mathcal{G}_{U\to V}, \mathcal{G}_{V\to U}\}$$

λ is a hyper parameter for interpolation. Note that we compute $\mathcal{L}_{U\to V}, \mathcal{L}_{V\to U}$ with train data, and $\mathcal{G}_{U\to V}, \mathcal{G}_{V\to U}$ with transfer data. When we use mini-batch, this means that there are two types of mini-batches for data from train and transfer datasets, respectively. Train data is used to learn $f_{U\to V}$ and $f_{V\to U}$, and transfer data is used to learn encoder $\theta_{\mathcal{E}}$. Please see Algorithm 2 for details.

Algorithm 2 The proposed approach for representation learning.

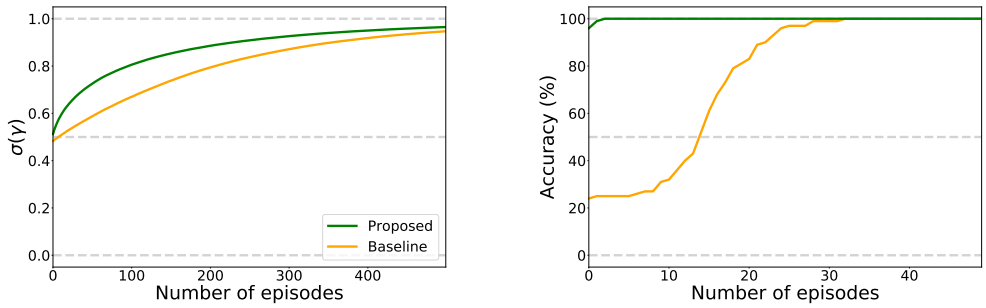
1: Initialize all parameters.
2: **for** each iteration **do**
3: Update predictor parameters with train samples, and loss $\mathcal{L}_{U\to V} + \mathcal{L}_{V\to U}$.
4: Update encoder parameters with transfer samples, and loss $\lambda \min\{\mathcal{G}_{U\to V}, \mathcal{G}_{V\to U}\}$.
5: **end for**

An intuitive understanding is that we want to learn a representation that works well in both causal directions in train distribution, and at least one causal direction in transfer data distribution. Therefore, the models are trained well in train distribution, and the representation recovers causal variables, because it works on both distributions in at least one causality direction.

3 Experiments

We run experiments to compare the baseline and proposed approaches on efficiency for causality direction prediction task, and representation learning task.

3.1 Causality Direction Prediction



(a) $\sigma(\gamma)$ (vertical) on transfer data along with number of episodes (horizontal).

(b) Accuracy (vertical) on transfer data with computation time in seconds (horizontal).

Figure 1: Experiments to evaluate efficiency in causality direction prediction. The model is discrete and both A and B are 10 dimensional variables ($N = 10$). The proposed approach is green. Baseline approach (Bengio et al., 2020) is orange.

We first compare the approaches in the same meta-learning setting. For the proposed approach, we replace log-likelihood of joint probabilities with the generalization loss. The overall loss for the proposed approach $\mathcal{R}'$ is similar to Equation (1) with $\mathcal{L}_{A\to B} = \mathcal{G}_{A\to B}$ and $\mathcal{L}_{B\to A} = \mathcal{G}_{B\to A}$.

$$\mathcal{R}' = -\log[\sigma(\gamma)\mathcal{G}_{A\to B} + (1-\sigma(\gamma))\mathcal{G}_{B\to A}]$$

The first case we consider is the discrete model with both A and B being 10 dimensional variables ($N = 10$). For other cases, please refer to Appendix E. The model architecture, same as in the baseline, has four modules: $P_{A\to B}(A), P_{A\to B}(B|A), P_{B\to A}(B), P_{B\to A}(A|B)$. The baseline uses all of them, and the proposed approach only uses the conditionals. We parameterize these probabilities via Softmax of unnormalized tabular quantities, and train the modules independently with maximum likelihood estimator. We then predict causality direction on transfer data with each approach.

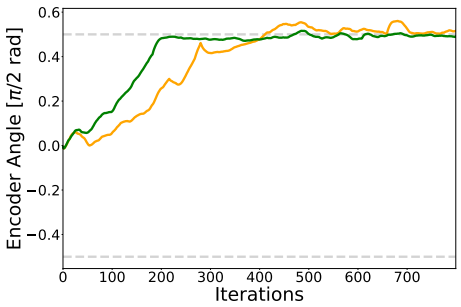
For each approach, we fix ground-truth causal direction as $A \to B$, and run 100 experiments with different random seeds. We plot the mean value of $\sigma(\gamma)$ in Figure 1a. The proposed method (green)

is above the baseline method (orange). This comparison shows that using generalization loss is more efficient than the log-likelihood of joint probabilities. Especially in the first several steps, the proposed approach moves toward the correct γ more quickly than the baseline approach. This quicker convergence is important when the transfer data is expensive. We also observe that for the baseline method, $\sigma(\gamma)$ is less than 0.5 for the first several steps, which might be because the factorization with incorrect causality direction has more parameters to be updated, leading to faster adaptation.

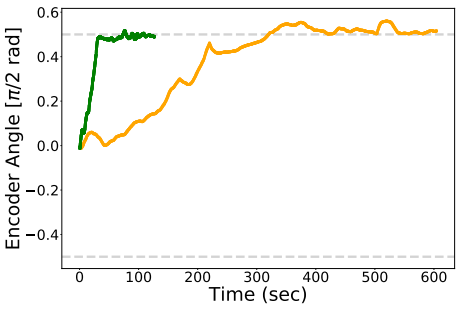
Since the proposed approach does not need an adaptation process, it actually does not need to update $\sigma(\gamma)$ with gradients. Instead, we can calculate a score, $\mathcal{S}_{\mathcal{G}} = \mathcal{G}_{B\to A} - \mathcal{G}_{A\to B}$, to detect causality direction. This score is not directly comparable with $\sigma(\gamma)$ in baseline approach, because they have different scales and semantics. To compare the approaches, we reformulate the causality direction prediction problem as a binary classification problem, and evaluate accuracy of the classification. For each approach, we run 100 experiments with the same setting as above. We then compute the accuracy as the number of experiments with successful prediction over the number of all experiments. For the baseline approach, we count an experiment as successful when $\gamma > 0$. For the proposed approach, we count when $\mathcal{S}_{\mathcal{G}} > 0$.

We plot accuracy along with the number of iterations to compare sample efficiency as shown in Figure 1b. The experimental result demonstrates that the proposed approach (green) achieves 100% accuracy after just 3 iterations with 5.1 ms, while the baseline approach (orange) takes 33 iterations to achieve 100% accuracy with 165.1 ms. Since our method only requires estimating a scale value of loss from transfer data, so that it is more sample efficient than the original method that requires learning / updating all parameters in a model during adaptation. The experiments show that our method is 11.0 times sample efficient and 32.4 times quicker compared with the baseline approach. Also, the baseline accuracy for the first several steps is less than 50%. This is consistent with the previous observation that the $\sigma(\gamma) < 0.5$ for the first several steps.

3.2 Representation Learning



(a) θ values ($\times\pi/2$) with number of iterations.



(b) θ values ($\times\pi/2$) with computation time in seconds.

Figure 2: Experiments to evaluate efficiency in representation learning with exactly the same setting. The curve of the proposed approach (green) is above that of baseline approach (orange), indicating the proposed method have both better sample and time complexity.

We evaluate representation learning in the same setting as in previous work. We set the correct causal direction as $A \to B$. With a sample (A, B), a fixed decoder $\mathcal{D} = R(\theta_{\mathcal{D}})$ converts it to an observation (X, Y), where R is a rotation matrix. We then use both approaches to learn encoders $\mathcal{E} = R(\theta_{\mathcal{E}})$ that map (X, Y) to representation (U, V). We set decoder parameter $\theta_{\mathcal{D}} = -\pi/4$, so that the expected encoder parameter should be $\pi/4$ or $-\pi/4$.

For both approaches, we use the same experiment setting and hyper parameters. In each iteration, the baseline approach uses transfer samples to adapt the model by learning gradient descent for several steps, and obtain regret to update encoder. On the other hand, we directly use loss instead of analyzing with statistical test, because this can be neurally integrated with neural network training for disentangled representation learning.

We plot the relation of θ values along with the number of iterations and time to compare sample and time efficiency. The result for sample efficiency is in Figure 2a. The proposed approach (green) reaches near $\pi/4$ after around 220 iterations. Baseline approach (orange) reaches near $\pi/4$ after around 410 iterations. The proposed approach is around 1.9 times sample efficient compared with the baseline approach. The result for time efficiency is in Figure 2b. The proposed approach (green) reaches near $\pi/4$ in around 34 seconds. Baseline approach (orange) reaches near $\pi/4$ in around 319 seconds. The proposed approach is 9.4 times quicker than the baseline approach.

The results show that proposed approach is both sample and time efficient for the representation learning task. Especially, the time efficiency shows the proposed method's advantage that it does not need multiple steps of adaptation in each iteration. It even does not need to compute gradient with back propagation or update model parameters.

4 Discussions

To further understand the proposed approach, we discuss alternative metrics for distribution differences, and the working conditions of the proposed approach and the baseline. We also extend experiments on real data and neural network models.

Alternative Metrics The proposed approach may even work when using other metrics that are easily accessed in standard machine learning workflows. Many loss functions share properties of distance metrics and smoothness conditions that make them useful for comparing the generalization ability of different models. Such metrics may be biased but can also be used to assess causality direction. We describe and experiment with metrics such as the KL-divergence loss and the gradient norm in Appendix F. Empirical tests show that both alternative metrics can identify causality direction in the prediction task. One may choose these alternative metrics depending on the workflow.

Robustness We compare the proposed and the baseline approach on the working conditions. One potential problem of the baseline approach is the dependency on marginal distributions. This not only requires more samples and computation, but when the marginal distribution for the correct factorization $P(A)$ is complicated, it may produce incorrect causality predictions (suppose $A \to B$ is the correct causality direction). We show an example for such a case in Appendix G. We design marginal distribution of cause variable $P(A)$ with a complicated hidden structure, which makes it harder to adapt after intervention than the marginal distribution of effect variable $P(B)$, and conditional distribution $P(A|B)$. We use a mixture model for $P(A)$ with K dimensional hidden variables C, and do not change models for other distributions. The result shows that the proposed approach is more robust than the baseline one in this case.

Real Data In addition to comparing the proposed approach with the previous work in the same setting to understand fundamental mechanism for causality learning, we also evaluate whether the propose method extends naturally to broader situations. We use real data of temperature and altitude (Mooij et al., 2016). We apply the proposed approach with both linear models and neural networks. We also add noise to the data and observe the changes on performance. The details can be found in Appendix H. The results show that the proposed approach works for the real data, and also applies for both linear models and neural networks. It resists some level of noise (standard deviation ≤ 1), but not when the noise is too large.

5 Related Work

Our work is closely related to causal reasoning, disentangled representation learning and its application to domain adaptation and transfer learning.

Causal Reasoning There is a rich literature on causal reasoning based on pre-conceived formal theory since Pearl's seminal work on do-calculus (Pearl, 1995; 2009). The major approach is structure learning in Bayesian networks. Discrete search and simulated annealing are reviewed in (Heckerman et al., 1995). Minimum Description Length (MDL) principles are commonly used to score and search (Lam & Bacchus, 1993; Friedman & Goldszmidt, 1998). Bayesian Information Criterion

(BIC) is also used to search when models have high relative posterior probability (Heckerman et al., 1995).

Prior work uses observation, but does not consider interventions, and focus on likelihood learning or hypothesis equivalence classes (Heckerman et al., 1995). Later, approaches have also been proposed to infer the causal direction only from observation (Peters et al., 2017), based on not generally robust assumptions on the underlying causal graph. The impact of interventions on probabilistic graphical models is also introduced in (Bareinboim & Pearl, 2016). Another early work (Peters et al., 2016) uses intervention to detect causal direction, but it does not address representation learning, which is an important problem in modern neural network, and the main topic of this paper. Further, our method should be more efficient than the fast approximation method (Peters et al., 2016), which needs to fit a model and takes many iterations. In contrast, we only run one forward pass on a fixed model.

Recently, meta-learning approach is adopted to draw causal inferences from purely observational data. Dasgupta et al. (2019) focuses on training an agent to learn causal reasoning by meta reinforcement learning; Bengio et al. (2020) focuses on predicting causal structure based on how fast a learner adapts to new distributions. In contrast, we propose an efficient approach for disentangling causal mechanisms, based on the difference of conditional probabilities in original and new distributions, assuming causal mechanisms hold in both distributions. More recently, generative models have been augmented with causal capabilities to learn to generate images and also to plan (Kocaoglu et al., 2018; Kurutach et al., 2018).

Disentangled Representation Learning This work discovers the underlying causal variables and their dependencies. This is related to disentangling variables (Bengio et al., 2013) and disentangled representation learning (Higgins et al., 2017). Some work point out that assumptions, including priors or biases should be necessary to discover the underlying explanatory variables (Bengio et al., 2013; Locatello et al., 2018). Another work (Locatello et al., 2018) reviews and evaluates disentangling and discusses different metrics.

A strong assumption of disentangling is that the underlying explanatory variables are marginally independent from each other. Many deep generative models (Goodfellow et al., 2016) and independent component analysis models (Hyvärinen et al., 2004; Higgins et al., 2017; Hyvarinen et al., 2018) are built on this assumption. However, in realistic cases, this assumption is often not likely to hold. Some recent work addresses compositionality learning (Li et al., 2019; 2020) without such assumption, but they do not address how to use causal mechanisms for representation learning.

Domain Adaptation and Transfer Learning Previous work finds a subset of features to best predict a variable of interest (Magliacane et al., 2018). However, the work is about feature selection, and our work is about causality direction and representation learning.

Other work also examine incorrect inference and turn the problem into a domain adaptation problem (Johansson et al., 2016). Counterfactual regression is proposed to estimate each effect from observation (Shalit et al., 2017). Another approach is to find a subset which makes the target independent from the variable selection (Rojas-Carulla et al., 2018). To do that, they assume the invariance of conditional distribution in train and target domains.

Also, competition may be introduced to recover a set of independent causal mechanisms, which leads to specialization (Parascandolo et al., 2017).

6 Conclusion

In this paper, we propose an efficient simplification to a recent technique to disentangle causal representations. Unlike the baseline approach which requires significant adaptation time and sophisticated comparisons of the models' adaptation speed, the proposed approach directly measures generalization ability based on the divergence of conditional probabilities from the original to transfer data distributions. We provide theoretical explanation for the advantage of the proposed method, and our experiments demonstrate that the technique is significantly more efficient than adaptation speed based approach in causality direction prediction and representation learning tasks. We hope this work in causal representation learning will be helpful for more advanced artificial intelligence.

REFERENCES

Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

Yoshua Bengio. Cifar workshop talk. CIFAR Workshop on Learning in Machines & Brains, 2017.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Nan Rosemary Ke, Sebastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A meta-transfer objective for learning to disentangle causal mechanisms. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=ryxWIgBFPS`.

Sergei Bernstein. The theory of probabilities, 1946.

Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. End-to-end continuous speech recognition using attention-based recurrent nn: First results. In *NeurIPS workshop*, 2014.

Ishita Dasgupta, Jane Wang, Silvia Chiappa, Jovana Mitrovic, Pedro Ortega, David Raposo, Edward Hughes, Peter Battaglia, Matthew Botvinick, and Zeb Kurth-Nelson. Causal reasoning from meta-reinforcement learning. *arXiv preprint arXiv:1901.08162*, 2019.

Nir Friedman and Moises Goldszmidt. Learning bayesian networks with local structure. In *Learning in graphical models*, pp. 421–459. Springer, 1998.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

Sander Greenland, Judea Pearl, James M Robins, et al. Causal diagrams for epidemiologic research. *Epidemiology*, 10:37–48, 1999.

David Heckerman, Dan Geiger, and David M Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243, 1995.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, volume 3, 2017.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. *arXiv preprint arXiv:1803.03067*, 2018.

David Hume. *A treatise of human nature*. Courier Corporation, 2003.

Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent component analysis*, volume 46. John Wiley & Sons, 2004.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard E Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. *arXiv preprint arXiv:1805.08651*, 2018.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pp. 3020–3029, 2016.

Murat Kocaoglu, Christopher Snyder, Alexandros G. Dimakis, and Sriram Vishwanath. CausalGAN: Learning causal implicit generative models with adversarial training. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=BJE-4xW0W`.

Harold W. Kuhn and Albert W. Tucker. Nonlinear programming. In *Berkeley Symposium*, pp. 481–492. Berkeley: University of California Press, 1951.

Thanard Kurutach, Aviv Tamar, Ge Yang, Stuart J Russell, and Pieter Abbeel. Learning plannable representations with causal infogan. In *Advances in Neural Information Processing Systems*, pp. 8733–8744, 2018.

Wai Lam and Fahiem Bacchus. Using causal information and local measures to learn bayesian networks. In *Proceedings of the Ninth international conference on Uncertainty in artificial intelligence*, pp. 243–250. Morgan Kaufmann Publishers Inc., 1993.

Yuanpeng Li, Liang Zhao, Jianyu Wang, and Joel Hestness. Compositional generalization for primitive substitutions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4284–4293, 2019.

Yuanpeng Li, Liang Zhao, Kenneth Church, and Mohamed Elhoseiny. Compositional continual language learning. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=rklnDgHtDS`.

Francesco Locatello, Stefan Bauer, Mario Lucic, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. *arXiv preprint arXiv:1811.12359*, 2018.

Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems*, pp. 10846–10856, 2018.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research*, 17(1):1103–1204, 2016.

Yurii Nesterov. *Introductory Lectures on Convex Optimization*. Springer US, 2004.

Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla, and Bernhard Schölkopf. Learning independent causal mechanisms. *arXiv preprint arXiv:1712.00961*, 2017.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

Judea Pearl. *CAUSALITY: MODELS, REASONING, AND INFERENCE*. Cambridge university press, 2003.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Judea Pearl. Does obesity shorten life? or is it the soda? on non-manipulable causes. *Journal of Causal Inference*, 6(2), 2018.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

Joseph D Ramsey, Stephen José Hanson, Catherine Hanson, Yaroslav O Halchenko, Russell A Poldrack, and Clark Glymour. Six problems for causal inference from fmri. *neuroimage*, 49(2): 1545–1558, 2010.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

Maya Rotmensch, Yoni Halpern, Abdulhakim Tlimat, Steven Horng, and David Sontag. Learning a health knowledge graph from electronic medical records. *Scientific reports*, 7(1):5994, 2017.

Donald B Rubin and Richard P Waterman. Estimating the causal effects of marketing interventions using propensity score methodology. *Statistical Science*, pp. 206–222, 2006.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*, 2012.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3076–3085. JMLR. org, 2017.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354, 2017.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Max Welling. Are ml and statistics complementary? In *IMS-ISBA Meeting on 'Data Science in the Next 50 Years*, 2015.

Dong Yin, Ashwin Pananjady, Max Lam, Dimitris Papailiopoulos, Kannan Ramchandran, and Peter L. Bartlett. Gradient diversity: a key ingredient for scalable distributed learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

A PROOF THAT KL-DIVERGENCE CAN BE USED TO IDENTIFY CAUSALITY DIRECTION

Here, we prove that we could use conditional KL-divergence to detect causality direction.

Proposition 1. *Given two data distributions with the same directed causality (e.g., $A \to B$) between two random variables, A and B, the difference of the KL-divergences of their conditional distributions (i.e., $P(B|A)$) is an unbiased estimator of the correct causality direction.*

Proof. First, let P_1 and P_2 be the distributions such that $P_1 \neq P_2$, but their conditional distributions are the same. Let $D_{\text{KL}}(\cdot\|\cdot)$ be conditional KL-divergence. $\text{CrossEntropy}_{A\to B}(P_i, P_j)$ is the conditional cross-entropy, $-\sum_{\forall(a,b)} P_i(a,b)\log P_j(b|a) = \mathbb{E}_{A,B\sim P_i}[-\log P_j(B|A)]$. Let $H_{A\to B}(\cdot)$ be conditional entropy.

By the Kraft–McMillan theorem, we have

$$\begin{aligned} D_{A\to B} &= D_{\text{KL}}(P_2(B|A)\|P_1(B|A)) \\ &= \mathbb{E}_{A,B\sim P_2}[\log P_2(B|A) - \log P_1(B|A)] \\ &= \mathbb{E}_{A,B\sim P_2}[-\log P_1(B|A)] - \mathbb{E}_{A,B\sim P_2}[-\log P_2(B|A)] \\ &= \text{CrossEntropy}_{A\to B}(P_2, P_1) - H_{A\to B}(P_2) \end{aligned}$$

Without loss of generality, suppose $A \to B$ is the correct causality direction. Since $P_1(B|A) = P_2(B|A)$, we have that $\text{CrossEntropy}_{A\to B}(P_2, P_1) = H_{A\to B}(P_2)$ and

$$\begin{aligned} D_{A\to B} &= \text{CrossEntropy}_{A\to B}(P_2, P_1) - H_{A\to B}(P_2) \\ &= H_{A\to B}(P_2) - H_{A\to B}(P_2) \\ &= 0 \end{aligned}$$

Finally, if we define a directionality score, $\mathcal{S}_{D_{\text{KL}}}$, as the difference of KL-divergence of the two distributions, we can identify the incorrect causal direction by its greater divergence. Namely,

$$\begin{aligned} \mathcal{S}_{D_{\text{KL}}} &= D_{B\to A} - D_{A\to B} \\ &= [\text{CrossEntropy}_{B\to A}(P_2, P_1) - H_{B\to A}(P_2)] - 0 \\ &> 0 \implies A \to B \end{aligned}$$

Here we use the property that entropy is less than cross entropy when the two distributions are not equal. □

B PROOF THAT GENERALIZATION LOSS CLOSELY APPROXIMATES DIVERGENCE

Here, we prove that we can replace conditional KL-divergence with a standard generalization loss calculation to approximate the causal directionality score, $\mathcal{S}_{D_{\text{KL}}}$, in Proposition 1. This approximation introduces bias into the score, but we expect this bias to be small relative to the difference of conditional KL-divergences of the incorrect causal model. Thus, it can be used as a good estimator of causality direction.

Proposition 2. *Given two different data distributions, P_1 and P_2, with the same directed causality between random variables, A and B, consider the generalization loss defined as the conditional cross-entropy of a given distribution:*

$$\begin{aligned} \mathcal{L}^{P_i}_{A\to B} &= \textit{CrossEntropy}_{A\to B}(P_i, P_1) \\ &= \mathbb{E}_{A,B\sim P_i}[-\log P_1(B|A)] \end{aligned}$$

And define the generalization gap as $\mathcal{G}_{\cdot\to\cdot} = \mathcal{L}^{P_2}_{\cdot\to\cdot} - \mathcal{L}^{P_1}_{\cdot\to\cdot}$. Finally, define the modified causality score, $\mathcal{S}_{\mathcal{G}} = \mathcal{G}_{B\to A} - \mathcal{G}_{A\to B}$. If $\mathcal{S}_{D_{KL}} = D_{B\to A} - D_{A\to B}$ as defined in Proposition 1, then $\mathcal{S}_{\mathcal{G}} = \mathcal{S}_{D_{KL}} - [\Delta H(B) - \Delta H(A)]$. $\mathcal{S}_{\mathcal{G}}$ is a biased estimator of which model has correct causality direction.

Proof. Let $H(\cdot,\cdot)$ be the joint entropy. The generalization gap is

$$\begin{aligned}
\mathcal{G}_{A\to B} &= \mathcal{L}^{P_2}_{A\to B} - \mathcal{L}^{P_1}_{A\to B} \\
&= \text{CrossEntropy}_{A\to B}(P_2, P_1) - H_{A\to B}(P_1) \\
&= [\text{CrossEntropy}_{A\to B}(P_2, P_1) - H_{A\to B}(P_2)] + H_{A\to B}(P_2) - H_{A\to B}(P_1) \\
&= D_{A\to B} + H_{A\to B}(P_2) - H_{A\to B}(P_1) \\
&= D_{A\to B} + H(P_2(B|A)) - H(P_1(B|A)) \\
&= D_{A\to B} + [H(P_2(A,B)) - H(P_2(A))] - [H(P_1(A,B)) - H(P_1(A))] \\
&= D_{A\to B} + [H(P_2(A,B)) - H(P_1(A,B))] - [H(P_2(A)) - H(P_1(A))] \\
&= D_{A\to B} + \Delta H(A,B) - \Delta H(A) \\
\mathcal{G}_{B\to A} &= D_{B\to A} + \Delta H(A,B) - \Delta H(B)
\end{aligned}$$

Here $\Delta H(A,B)$, $\Delta H(A)$ and $\Delta H(B)$ are the change in entropy from P_1 to P_2 for (A,B), A and B, respectively. Further, the modified causality score is the difference of these generalization gaps:

$$\begin{aligned}
\mathcal{S}_{\mathcal{G}} &= \mathcal{G}_{B\to A} - \mathcal{G}_{A\to B} \\
&= [D_{B\to A} + \Delta H(A,B) - \Delta H(B)] - [D_{A\to B} + \Delta H(A,B) - \Delta H(A)] \\
&= [D_{B\to A} - D_{A\to B}] - [\Delta H(B) - \Delta H(A)] \\
&= \mathcal{S}_{D_{\text{KL}}} - [\Delta H(B) - \Delta H(A)]
\end{aligned}$$

Without loss of generality, suppose $A \to B$ is the correct causal direction, then $D_{A\to B} = 0$.

$$\mathcal{S}_{\mathcal{G}} = D_{B\to A} - [\Delta H(B) - \Delta H(A)]$$

When $\mathcal{S}_{\mathcal{G}} > 0$ (and $[\Delta H(B) - \Delta H(A)]$ is small), the prediction of causality direction is correct. □

Informally, we expect the bias term, $[\Delta H(B) - \Delta H(A)]$, to be small relative to the divergence, $D_{B\to A}$, in real application settings. If $A \to B$, then A and B will have some (positive or negative) correlation–they will vary in a correlated way under intervention on A. Unless the conditional distribution, $P_1(B|A) = P_2(B|A)$, contains large portions with high entropy, it is likely that a change in A will cause a commensurate change in B. Thus, it is likely that $|\Delta H(A)| \approx |\Delta H(B)|$, so it is unlikely that $\Delta H(A)$ and $\Delta H(B)$ are both large and have opposite sign.

Further, in cases where $\Delta H(A)$ is large, but $\Delta H(B)$ is small, it is frequently an indicator that the selected P_i are "lucky" (uncommon) or the causality between A and B is weak. Consider the example in which $H(P_1(A))$ is small, but $H(P_1(B))$ is large. This case can occur if $H(P_1(B|A))$ is small, but it is unlikely, because the (low-entropy) conditional distribution must impose the disorder on $P_1(B)$. It is easy to see that the reverse situation ($H(P_1(B|A))$ is large) implies that the causality between A and B is weak, such that causality direction prediction may be difficult anyway.

Finally, if we want to relax the assumption that either $A \to B$ or $B \to A$ (e.g., a null-hypothesis that A and B are not causally related), we could introduce a significance test to decide whether we can accept the selection of causal direction from either method. Such a significance test would need to be mindful of the magnitude of any biases introduced in the approach.

C Generalization Loss is Likely to be More Practical

Here, we explain how generalization loss is likely to be more practical than gradient-based approaches for identifying a model with correct causal direction. Both the baseline and proposed approaches rely on convergence of estimators toward a ground truth model, but gradient-based approaches require much stronger assumptions to guarantee convergence, and they are more susceptible to practical challenges like tuning optimizer hyperparameters.

First, both gradient- and generalization-based approaches have the same asymptotic convergence characteristics, with the prediction error proportional to $\frac{1}{\sqrt{n}}$, where n is the number of steps or samples. Many theorems state that gradient-based approaches can have asymptotic convergence to within an ε error based on number of gradient descent steps, k: $\varepsilon = O(\frac{1}{\sqrt{k}})$. These proofs rely on many assumptions that cannot be guaranteed in real-world datasets, such as convexity, regularity,

or smoothness of error surface, using optimal settings of optimization parameters for best-case step sizes, etc. (Kuhn & Tucker, 1951; Nesterov, 2004; Yin et al., 2018). These bounds can also tend to be loose for real data distributions.

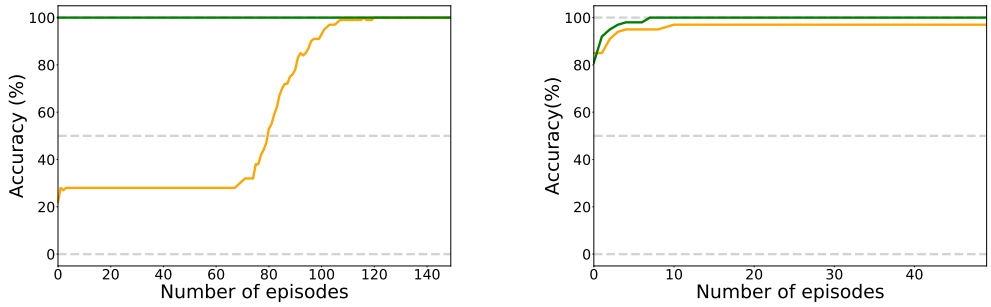
On the other hand, non-parametric estimators, such as the sample average generalization loss proposed in this work, do not have a model, so their convergence rates can depend only on the structure of the dataset (e.g., first or second moments). For instance, early estimator convergence results show that after summarizing n samples, the error in an estimator can be within ε of the true value: $\varepsilon = O(\frac{1}{\sqrt{n}})$ (Bernstein, 1946; Hoeffding, 1963). The constants in these bounds are just the moments that characterize the dataset (e.g., mean, variance). Like the gradient-based convergence rates, these non-parametric convergence rates can also be loose for real data distributions.

Thus, in practical settings, it is likely that a generalization-based approach (non-parametric estimators) will converge more quickly on causality identification in real-world datasets. The proposed approach does not require further model optimization (fine-tuning), so it is not susceptible to poorly chosen hyperparameters. Further, it is unlikely to encounter complex error surface topology. Thus, in practice, it is likely to converge more quickly than gradient-based causality identification techniques.

D Representation Learning

We run representation learning experiments for both approaches on a GeForce GTX TITAN Z with PyTorch. We use the original implementation for the baseline approach, and extend it for the proposed approach. The hyperparameters are the same as the original setting (Bengio et al., 2020).

E More Experiments on Causality Direction Predictions



(a) $\sigma(\gamma)$ (vertical) on transfer data along with number of episodes (horizontal).

(b) Accuracy (vertical) on transfer data with computation time in seconds (horizontal).

Figure 3: Experiments to evaluate efficiency in causality direction prediction. The model is discrete and both A and B are 100 dimensional variables ($N = 100$).

We would like to extend causality direction prediction experiments to high dimensional variable and continuous variables. We still follow the experiment settings in previous work (Bengio et al., 2020), and compare the proposed and the baseline approaches. We run all the experiments of causality direction prediction with Intel Core i9 9920X 12-Core Processors.

For high dimensional variable, we run experiments with $N = 100$. We use the same setting as the experiments with $N = 10$, but only changed the N. The result for sample efficiency is in Figure 3a. Similar to the case of $N = 10$, the proposed approach (green) is significantly more efficient than the baseline approach (orange) in both sample.

For continuous variable, we used the same setting as the previous work (Bengio et al., 2020). For the proposed method, we only use one sample for each episode. The plot shows that the proposed approach (green) is more efficient than the baseline approach (orange) in sample number (Figure 3b).

F ALTERNATIVE METRICS

The proposed approach may even work when using other metrics that are easily accessed in standard machine learning workflows. Many loss functions share properties of distance metrics and smoothness conditions that make them useful for comparing the generalization ability of different models. Such metrics may be biased but can also be used to assess causality direction. We describe and experiment with metrics such as the KL-divergence loss and the gradient norm. Empirical tests show that both alternative metrics can identify causality direction in the prediction task. One may choose these alternative metrics depending on one's workflow.

For KL-divergence, we define the metric as the follows, and show an example result in Figure 4a.

$$\mathcal{S}_{D_{\text{KL}}} = D_{\text{KL}}[P_2(B|A)\|P_1(B|A)] - D_{\text{KL}}[P_2(A|B)\|P_1(A|B)]$$

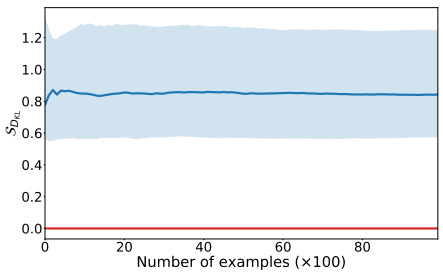
For gradient L_2 norm, we define the metric as the follows, and we show an example result in Figure 4b.

$$\mathcal{S}_{L_2} = L_2\Big(\frac{\partial \mathcal{L}^{\text{transfer}}_{A\to B}}{\partial \theta_{A\to B}}\Big) - L_2\Big(\frac{\partial \mathcal{L}^{\text{transfer}}_{B\to A}}{\partial \theta_{B\to A}}\Big)$$

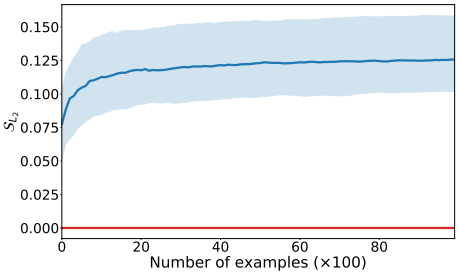
Note that the model or optimization algorithm does not have special requirement on gradient, such as gradient clipping.

These results indicate that other loss functions can behave similarly to the conditional cross-entropy when identifying causality direction. The initial samples already show a large gap between the correct and incorrect causal models, and after just a few tens of examples, the difference is close to convergence.

We expect that the same intuition applies for these other loss metrics. In particular, these loss metrics both satisfy distance metric properties of non-negativity and the triangle inequality. Further, they have useful continuity and smoothness characteristics that suggests any biases they may have should behave like conditional cross-entropy. Our empirical results confirm these intuitions.



(a) Difference of KL-divergence $\mathcal{S}_{D_{\text{KL}}}$ along with the number of transfer samples in the proposed approach.



(b) Difference of L_2 gradient norm $\mathcal{S}_{L_2}$ along with the number of transfer samples in the proposed approach.

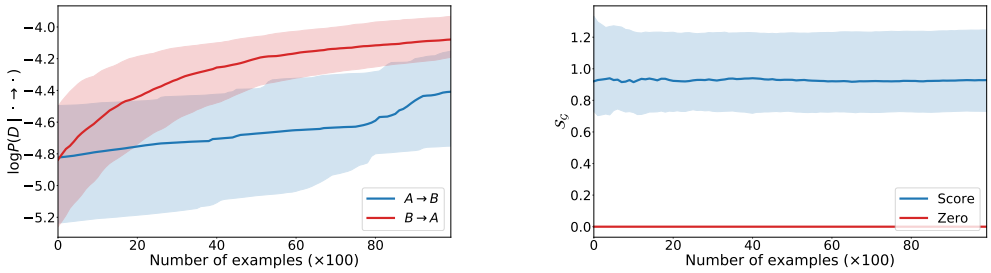
Figure 4: Experiments for other metrics The model is discrete and $N = 10$. Curve (blue) is median over 100 runs, with 25-75% quantiles intervals, and it is significantly above zero (red), indicating these metrics are good indicators for causality learning.

G ROBUSTNESS

To compare the robustness of the proposed approach with the baseline, we design $P(A)$ with a complicated hidden structure, so that it is harder to adapt after intervention than $P(B)$ and $P(A|B)$. For $P(A)$, we use a mixture model with K dimensional hidden variables C, and keep other distributions unchanged.

$$P(A;\theta,\phi) = \sum_{k=1}^{K} P(A|C_k;\phi)P(C_k;\theta)$$

where $\theta \in \mathbb{R}^K, \phi \in \mathbb{R}^{K \times N}$. We set $N = 10, M = 10, K = 200$. For the baseline approach, we plot the improvement of log likelihood during adaptation along with the number of samples ($\times 100$) in Figure 5a. The model with correct causality direction (blue) adapts slower than that of incorrect direction (red), which means the fundamental intuition of baseline approach does not apply in this situation. For the proposed approach, we plot the difference of generalization losses $\mathcal{S}_{\mathcal{G}} = \mathcal{G}_{B \to A} - \mathcal{G}_{A \to B}$ along with the number of samples ($\times 100$) in Figure 5b. The difference (blue) is significantly larger than zero (red). This is natural because the proposed approach does not use $P(A)$, so that it is not influenced by whether $P(A)$ has complicated hidden structure. This experiment shows that the proposed approach is more robust than the baseline one in this case.



(a) Incorrect result from baseline approach. Log likelihood during adaptation along with the number of samples ($\times 100$). The correct causal model (blue) adapts more slowly than that of the incorrect model (red).

(b) Result of the proposed approach. The difference of generalization losses $\mathcal{S}_{\mathcal{G}} = \mathcal{G}_{B \to A} - \mathcal{G}_{A \to B}$ along with the number of samples ($\times 100$). The difference (blue) is significantly larger than zero (red).

Figure 5: Experiments to evaluate robustness. $A \to B$ is the correct causality direction. Here, $N = 10, M = 10, K = 200$. Curves are median over 100 runs, with 25-75% quantiles intervals. The result shows that the models with correct causalities are slow to adapt in baseline approach (a), which means the baseline approach does not work, but the proposed approach (b) works.

H Real Data Experiments

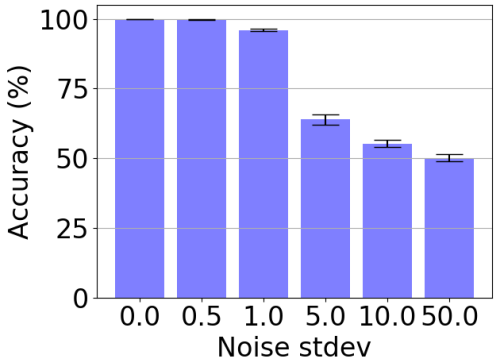


Figure 6: Accuracy when adding noise.

We use altitude and temperature dataset (Mooij et al., 2016), where altitude (A) causes temperature (B). We generate train and transfer datasets in the following way. We first divide the data to two sets D_1 and D_2, according to their A values. samples in D_1 has larger or equal A values than those in D_2. We respectively shuffle and separate D_1 and D_2 to two subsets with ratio of 9:1, generating $D_1^1, D_1^2, D_2^1, D_2^2$. We than merge sets to generate train data ($D_1^1 \cup D_2^2$) and transfer data ($D_1^2 \cup D_2^1$).

We use both linear regression model with L_2 loss and neural network model with two layers. The hidden layer has 100 nodes with ReLU activation. We use scikit-learn (Pedregosa et al., 2011) for

implementation. We run experiments 1,000 times with different random seeds, and the predicted causal directions are always correct.

For different noise levels, we use Gaussian noise with different standard deviations. In the altitude-temperature experiment with the linear model, the success rate start to drop to $99.8 \pm 0.1\%$ at 0.5, and becomes $96.0 \pm 0.4\%$ at 1.0, $63.9 \pm 1.8\%$ at 5.0, $55.3 \pm 1.2\%$ at 10.0, and $50.1 \pm 1.3\%$ at 50.0 (Figure 6). This means the method is robust to some noise (around 1 or less), but does not work when the noise is too large.