Mitigating input-causing confounding in multimodal learning via the backdoor adjustment

Taro Makino[1] **Krzysztof J. Geras**[2,1] **Kyunghyun Cho**[1,3,4]

[1]NYU Center for Data Science
[2]NYU Grossman School of Medicine
[3]Genentech [4]CIFAR LMB
taro@nyu.edu

Abstract

We adopt a causal perspective to address why multimodal learning often performs worse than unimodal learning. We put forth a structural causal model (SCM) for which multimodal learning is preferable over unimodal learning. In this SCM, which we call the multimodal SCM, a latent variable causes the inputs, and the inputs cause the target. We refer to this latent variable as an input-causing confounder. By conditioning on all inputs, multimodal learning d-separates the input-causing confounder and the target, resulting in a causal model that is more robust than the statistical model learned by unimodal learning. We argue that multimodal learning fails in practice because our finite datasets appear to come from an alternative SCM, which we call the spurious SCM. In the spurious SCM, the input-causing confounder and target are conditionally dependent given the inputs. This means that multimodal learning no longer d-separates the input-causing confounder and the target, and fails to estimate a causal model. We use a latent variable model to model the input-causing confounder, and test whether its undesirable dependence with the target is present in the data. We then use the same model to remove this dependence and estimate a causal model, which corresponds to the backdoor adjustment. We use synthetic data experiments to validate our claims.

1 Introduction

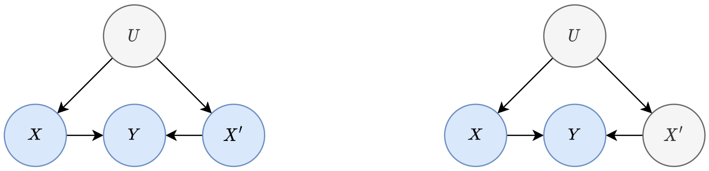
Multimodal learning refers to the joint modeling of data from different modalities such as images, speech, and text [Baltrusaitis et al., 2019]. It achieves promising results in areas such as joint vision and language modeling. For example, models trained to contrast images and their textual descriptions achieve strong zero-shot image classification performance on ImageNet [Deng et al., 2009, Radford et al., 2021, Jia et al., 2021, Pham et al., 2021]. Moreover, these models are highly robust to data distribution shifts that cripple the performance of their unimodal counterparts trained solely on images. These robustness results are significantly better than what has been achieved using only images, and also exhibits promising scaling behavior [Pham et al., 2021]. The models can also be fine-tuned in order to achieve state-of-the-art in-distribution classification performance while maintaining their robustness [Wortsman et al., 2021]. These results suggest that multimodal learning maybe a part of the solution to the longstanding problem of out-of-distribution generalization in deep learning.

Despite these successes, multimodal learning struggles in other areas such as visual question answering. In this setting, models are given an image and a question as inputs, and are asked to answer the question. Progress in this area is stalled by a degenerate behavior where models often answer a question while ignoring the image [Agrawal et al., 2016, 2018, Sheng et al., 2021]. More generally, it is frequently observed that multimodal learning performs worse than unimodal learning across a wide

NeurIPS 2022Workshop on Causal Machine Learning for Real-World Impact (CML4Impact 2022).

range of problems [Cadène et al., 2019, Gat et al., 2020, Wang et al., 2020, Wu et al., 2020]. Given these mixed results, there is an active research direction which aims to make multimodal learning more effective. In this work, we offer a causal perspective to complement the existing literature.

2 When is multimodal learning preferable over unimodal learning?



(a) Multimodal generating process

(b) Unimodal learning

Figure 1: (a) In the multimodal generating process, a latent variable $\mathbf{u}$ causes the inputs $\mathbf{x}$ and $\mathbf{x}'$, and the inputs cause the target $\mathbf{y}$. We refer to $\mathbf{u}$ as the input-causing confounder. (b) Unimodal learning only considers a single input $\mathbf{x}$ along with $\mathbf{y}$. Since the other modality $\mathbf{x}'$ is unobserved, there is an open backdoor path $\mathbf{x} \leftarrow \mathbf{u} \rightarrow \mathbf{x}' \rightarrow \mathbf{y}$ which makes $\mathbf{x}$ and $\mathbf{y}$ spuriously dependent. Multimodal learning is preferable over unimodal learning because it d-separates $\mathbf{u}$ and $\mathbf{y}$, resulting in a causal model.

We begin by specifying a structural causal model (SCM) under which multimodal learning has a clear advantage over unimodal learning. It is given by

$$\begin{aligned}
\mathbf{u} &:= f_{\mathbf{u}}(\epsilon_{\mathbf{u}}), \\
\mathbf{x} &:= f_{\mathbf{x}}(\mathbf{u}, \epsilon_{\mathbf{x}}), \\
\mathbf{x}' &:= f_{\mathbf{x}'}(\mathbf{u}, \epsilon_{\mathbf{x}'}), \\
\mathbf{y} &:= f_{\mathbf{y}}(\mathbf{x}, \mathbf{x}', \epsilon_{\mathbf{y}}),
\end{aligned} \tag{1}$$

where the ϵ's are noise variables. We refer to this SCM as the **multimodal SCM**, and its corresponding graph is shown in Fig. 1a. Since $\mathbf{u}$ causes the inputs, we refer to it as the input-causing confounder. This SCM exhibits unobserved confounding that unimodal learning is vulnerable to, and multimodal learning is invariant to.

Suppose we conduct unimodal learning by only modeling a single modality $\mathbf{x}$ along with the target $\mathbf{y}$. Since the other modality $\mathbf{x}'$ is unobserved, this corresponds to the graph in Fig. 1b. This opens the backdoor path $\mathbf{x} \leftarrow \mathbf{u} \rightarrow \mathbf{x}' \rightarrow \mathbf{y}$, making $\mathbf{x}$ and $\mathbf{y}$ spuriously dependent. A unimodal model would learn this spurious dependency, in addition to the causal relation $\mathbf{x} \rightarrow \mathbf{y}$. In other words, we have

$$p(\mathbf{y} \mid do(\mathbf{x})) \neq p(\mathbf{y} \mid \mathbf{x}). \tag{2}$$

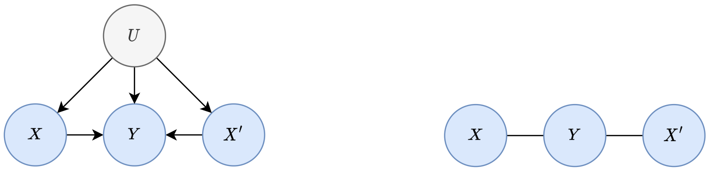
$p(\mathbf{y} \mid do(\mathbf{x}))$ is unidentifiable from observations alone, since neither $\mathbf{u}$ nor $\mathbf{y}$ are observed. This is a clear weakness for unimodal learning, since it can fail to generalize under test-time shifts in $p(\mathbf{u})$. In contrast, a multimodal model, that considers both modalities $\mathbf{x}$ and $\mathbf{x}'$, d-separates $\mathbf{u}$ and $\mathbf{y}$, rendering it invariant to shifts in $p(\mathbf{u})$. In other words, we have

$$p(\mathbf{y} \mid do(\mathbf{x}), do(\mathbf{x}')) = p(\mathbf{y} \mid \mathbf{x}, \mathbf{x}').$$

Therefore, under the multimodal SCM, multimodal learning is preferable over unimodal learning, since the former yields a causal model that is more robust to shifts in $p(\mathbf{u})$.

3 Explaining the failures of multimodal learning

We have just discussed the multimodal SCM, an SCM for which multimodal learning is more robust than unimodal learning to shifts in $p(\mathbf{u})$. How then, do we explain the frequent failures of multimodal learning? Our argument is based on the idea that a finite dataset sampled from an SCM can appear as though it came from an alternative SCM. In our case, although $\mathbf{u} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{x}, \mathbf{x}'$ holds in the original multimodal SCM, a finite dataset can appear as though there is a residual dependency, i.e.



(a) Spurious SCM (b) Markov random field

Figure 2: (a) The spurious SCM contains the edge $\mathbf{u} \rightarrow \mathbf{y}$, which makes $\mathbf{u}$ and $\mathbf{y}$ conditionally dependent given the inputs. Multimodal learning no longer d-separates $\mathbf{u}$ and $\mathbf{y}$ in this setting, and fails to estimate a causal model. (b) When $\mathbf{u}$ and $\mathbf{y}$ are perfectly dependent given the inputs, $\mathbf{u}$ collapses into $\mathbf{y}$, and the graph becomes a Markov random field. In this case, multimodal learning becomes equivalent to unimodal learning.

$\mathbf{u} \not\perp\!\!\!\perp \mathbf{y} \mid \mathbf{x}, \mathbf{x}'$. This can make the data appear as if it came from an alternative SCM, where all variables are defined the same as in Eq. (1) except for $\mathbf{y}$, which is given by

$$\mathbf{y} \coloneqq f_{\mathbf{y}}(\mathbf{x}, \mathbf{x}', \mathbf{u}, \epsilon_{\mathbf{y}}), \tag{3}$$

This changes the graph by introducing an additional edge $\mathbf{u} \rightarrow \mathbf{y}$, as seen in Fig. 2a. We refer to this alternative SCM as the **spurious SCM**.

The edge $\mathbf{u} \rightarrow \mathbf{y}$ poses a problem for multimodal learning, since conditioning on both inputs no longer d-separates $\mathbf{u}$ and $\mathbf{y}$. That is,

$$p(\mathbf{y} \mid do(\mathbf{x}), do(\mathbf{x}')) \neq p(\mathbf{y} \mid \mathbf{x}, \mathbf{x}').$$

Under this setting, multimodal learning may not be more robust than unimodal learning to shifts in $p(\mathbf{u})$.

The influence of the edge $\mathbf{u} \rightarrow \mathbf{y}$ depends on the degree of dependence between $f(\mathbf{u})$ and $\mathbf{y}$ given the inputs. For the sake of explanation, let us rewrite Eq. 3 as

$$\mathbf{y} \coloneqq f_{\mathbf{y}}(\mathbf{x}, \mathbf{x}', f(\mathbf{u}), \epsilon_{\mathbf{y}}),$$

where f is a function that modulates how much information about $\mathbf{u}$ is passed to $\mathbf{y}$. On one extreme, if f is a constant function, then we recover the original multimodal SCM, and there are no issues with multimodal learning. On the other extreme, if $f(\mathbf{u}) = \mathbf{y}$, then $\mathbf{u}$ collapses into $\mathbf{y}$. Now, each edge between $\mathbf{y}$ and the inputs are bidirectional, turning the graph into a Markov random field (MRF). This is shown in Fig. 2b, and the joint distribution of the observed variables is given by

$$p(\mathbf{x}, \mathbf{x}', \mathbf{y}) \propto \exp(\phi_{\mathbf{x}}(\mathbf{x}) + \phi_{\mathbf{x}'}(\mathbf{x}') + \phi_{\mathbf{y}}(\mathbf{y}) + \phi_{\mathbf{x},\mathbf{y}}(\mathbf{x}, \mathbf{y}) + \phi_{\mathbf{x}',\mathbf{y}}(\mathbf{x}', \mathbf{y})),$$

where the ϕ's are potential functions. Prediction is given by

$$\operatorname*{argmax}_{\mathbf{y}} \phi_{\mathbf{y}}(\mathbf{y}) + \phi_{\mathbf{x},\mathbf{y}}(\mathbf{x}, \mathbf{y}) + \phi_{\mathbf{x}',\mathbf{y}}(\mathbf{x}', \mathbf{y}).$$

If we assume that $\phi_{\mathbf{y}}$ is constant, then this is equivalent to prediction with an ensemble of unimodal models.

This explains why multimodal learning frequently fails to perform better than unimodal learning. Even though $\mathbf{u} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{x}, \mathbf{x}'$ holds in the multimodal SCM, a finite dataset may easily appear to have come from the spurious SCM, where $\mathbf{u} \not\perp\!\!\!\perp \mathbf{y} \mid \mathbf{x}, \mathbf{x}'$ holds instead. The extent to which this is problematic depends on the strength of the relation $\mathbf{u} \not\perp\!\!\!\perp \mathbf{y} \mid \mathbf{x}, \mathbf{x}'$. In the most extreme case where $f(\mathbf{u}) = \mathbf{y}$, a multimodal model becomes equivalent to an ensemble of unimodal models.

4 Testing whether $\mathbf{u} \rightarrow \mathbf{y}$ can be observed in our data

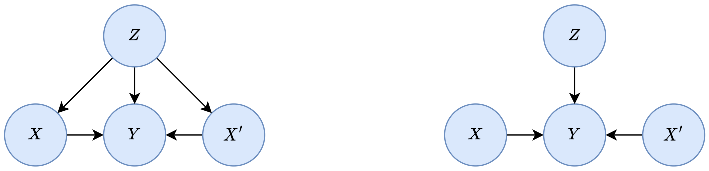
Given a finite multimodal dataset $\mathcal{D}$, we should ideally be able to test whether multimodal learning is worth pursuing over unimodal learning. That is, we want to know whether $\mathcal{D}$ is more consistent with the multimodal SCM, or the spurious SCM. This amounts to testing whether $\mathbf{u} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{x}, \mathbf{x}'$ holds. This cannot be done using traditional independence testing, since $\mathbf{u}$ is unobserved. Here, we present a way to test this condition using latent variable modeling.

We define a latent variable $\mathbf{z}$ to capture anything about $\mathbf{y}$ that cannot be explained using $\mathbf{x}$ and $\mathbf{x}'$. This information corresponds to the dependence between $\mathbf{u}$ and $\mathbf{y}$ given the inputs. We then train a conditional variational autoencoder (VAE) [Kingma and Welling, 2014] to maximize the lower bound of $\log p(\mathbf{y} \mid \mathbf{x}, \mathbf{x}')$. This model consists of an encoder $q(\mathbf{z} \mid \mathbf{x}, \mathbf{x}', \mathbf{y})$ and a decoder $p(\mathbf{y} \mid \mathbf{x}, \mathbf{x}', \mathbf{z})$, both parameterized by neural networks. The VAE is conditional because $\mathbf{x}$ and $\mathbf{x}'$ are always observed, and can be considered to be inverted, since a conditional VAE is typically conditioned on $\mathbf{y}$. The evidence lower bound (ELBO) to be maximized is the right-hand side of

$$\log p(\mathbf{y} \mid \mathbf{x}, \mathbf{x}') \geq \mathbb{E}_{q(\mathbf{z}\mid\mathbf{x},\mathbf{x}',\mathbf{y})}[\log p(\mathbf{y} \mid \mathbf{x}, \mathbf{x}', \mathbf{z})] - D_{KL}(q(\mathbf{z} \mid \mathbf{x}, \mathbf{x}', \mathbf{y}) \parallel p(\mathbf{z})).$$

Maximizing the ELBO can be thought of as using the encoder $q(\mathbf{z} \mid \mathbf{x}, \mathbf{x}', \mathbf{y})$ to encode information $\mathbf{z}$ that is useful for the decoder $p(\mathbf{y} \mid \mathbf{x}, \mathbf{x}', \mathbf{z})$ to later reconstruct $\mathbf{y}$. The KL divergence term in the ELBO measures how useful $\mathbf{z}$ is in modeling $\mathbf{y}$ given $\mathbf{x}$ and $\mathbf{x}'$. If there is no useful information available, this means that there is nothing for $\mathbf{z}$ to encode. Maximizing the ELBO would then push the KL divergence term to zero. This mechanism is the basis of our test for whether multimodal learning is preferable over unimodal learning. If the KL divergence term is close to zero, then it is likely that $\mathbf{z} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{x}, \mathbf{x}'$, and we can conclude that $\mathcal{D}$ is consistent with the multimodal SCM. Therefore, we would expect multimodal learning to be more robust than unimodal learning to shifts in $p(\mathbf{u})$.

5 Mitigating the influence of $\mathbf{u} \rightarrow \mathbf{y}$ in our data



(a) Original graph

(b) Mutilated graph

Figure 3: In order to intervene on $\mathbf{x}$ and $\mathbf{x}'$ in the original graph in (a), we remove all edges coming into $\mathbf{x}$ and $\mathbf{x}'$, resulting in the mutilated graph in (b). We can then compute $p_M(\mathbf{y} \mid \mathbf{x}, \mathbf{x}')$ in the mutilated graph to estimate $p(\mathbf{y} \mid do(\mathbf{x}), do(\mathbf{x}')$ under the original graph.

In order to mitigate the influence of $\mathbf{u} \rightarrow \mathbf{y}$, we must use $p(\mathbf{y} \mid do(\mathbf{x}), do(\mathbf{x}'))$ for prediction instead of $p(\mathbf{y} \mid \mathbf{x}, \mathbf{x}')$. To do so, we notice that $\mathbf{z}$ satisfies the backdoor criterion relative to $(\mathbf{x}, \mathbf{y})$, as well as to $(\mathbf{x}', \mathbf{y})$ [Pearl, 2009]. Therefore, $p(\mathbf{y} \mid do(\mathbf{x}), do(\mathbf{x}'))$ is identifiable via the backdoor adjustment. To do so, we remove the edges going into $\mathbf{x}$ and $\mathbf{x}'$ in the original graph in Fig. 3a, which results in the mutilated graph in Fig. 3b. $p(\mathbf{y} \mid do(\mathbf{x}), do(\mathbf{x}'))$ under the distribution P of the original graph is equivalent to $p_M(\mathbf{y} \mid \mathbf{x}, \mathbf{x}')$ under the distribution p_M of the mutilated graph. We therefore have

$$\begin{aligned} p(\mathbf{y} \mid do(\mathbf{x}), do(\mathbf{x}')) &= p_M(\mathbf{y} \mid \mathbf{x}, \mathbf{x}') \\ &= \int p_M(\mathbf{y}, \mathbf{z} \mid \mathbf{x}, \mathbf{x}') \mathrm{d}\mathbf{z} \\ &= \int p_M(\mathbf{y} \mid \mathbf{x}, \mathbf{x}', \mathbf{z}) p_M(\mathbf{z}) \mathrm{d}\mathbf{z} \\ &= \int p(\mathbf{y} \mid \mathbf{x}, \mathbf{x}', \mathbf{z}) p(\mathbf{z}) \mathrm{d}\mathbf{z} \\ &= \mathbb{E}_{p(\mathbf{z})}[p(\mathbf{y} \mid \mathbf{x}, \mathbf{x}', \mathbf{z})], \end{aligned} \tag{4}$$

where we used the fact that $p(\mathbf{z})$ and $p(\mathbf{y} \mid \mathbf{x}, \mathbf{x}', \mathbf{z})$ are the same across the original and mutilated graphs.

The expectation on the right-hand side of Eq. 4 may have high variance if we were to sample from $p(\mathbf{z})$. We therefore use importance sampling and take the expectation w.r.t. $p(\mathbf{z} \mid \mathbf{x}, \mathbf{x}')$, which we can estimate using our approximate posterior $q(\mathbf{z} \mid \mathbf{x}, \mathbf{x}', \mathbf{y})$ from our test in the previous section. We use our training set to minimize $D_{KL}(q(\mathbf{z} \mid \mathbf{x}, \mathbf{x}', \mathbf{y}) \parallel q(\mathbf{z} \mid \mathbf{x}, \mathbf{x}'))$, and then compute

$$p(\mathbf{y} \mid do(\mathbf{x}), do(\mathbf{x}')) \approx \frac{1}{M} \sum_{m=1}^{M} \frac{p(\mathbf{z}^{(m)})}{q(\mathbf{z}^{(m)} \mid \mathbf{x}, \mathbf{x}')} p(\mathbf{y} \mid \mathbf{x}, \mathbf{x}', \mathbf{z}^{(m)}),\ \mathbf{z}^{(m)} \sim q(\mathbf{z} \mid \mathbf{x}, \mathbf{x}').$$

6 Experimental setup

6.1 Dataset

We validate our claims using a synthetic data experiment, where the SCM is

$$\begin{aligned}
&U := \mathcal{N}(0,1), \\
&X := U + \mathcal{N}(0,1), \\
&X' := U^2 + \mathcal{N}(0,1), \\
&Y := X + X' + \beta U + \mathcal{N}(0,0.1).
\end{aligned} \tag{5}$$

Data sampled from this SCM appears to come from the spurious SCM, where $\beta \in [0,1]$ defines the degree of the strength of the dependence of U and Y given the inputs. Unless otherwise noted, we sample a dataset of one hundred thousand examples, and use 90% for the training set, 10% for the validation set, and 10% for the test set.

6.2 Model

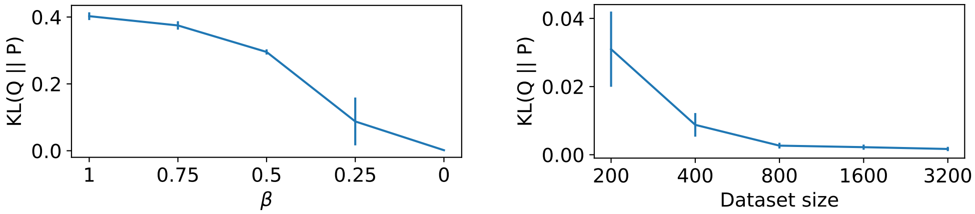
We experiment with a VAE with encoder $q(\mathbf{z} \mid x, x', y)$ and decoder $p(y \mid x, x', \mathbf{z})$. The encoder consists of two multi-layer perceptrons (MLPs), which turn (X, X', Y) into the mean and covariance of $\mathbf{z}$. The decoder also consists of two MLPs, which turn $(X, X', \mathbf{z})$ into the mean and covariance of Y. The latent variable $\mathbf{z}$ is a vector with length 100.

6.3 Training methodology

We optimize our models using AdamW [Loshchilov and Hutter, 2019] with a learning rate of 1×10^{-3} and a weight decay of 1×10^{-5}. Training is done for 200 epochs with a batch size of 100, and we early stop once the validation loss fails to improve for 20 epochs. We use NumPy [Harris et al., 2020], PyTorch [Paszke et al., 2019], and PyTorch Lightning [Falcon et al., 2019], where the licensing information can be found in the respective papers. We ran our experiments on a single NVIDIA RTX8000 GPU on our high performance computing system.

7 Results

First, we demonstrate that our test from Section 4 can detect the influence of the edge $U \to Y$. The results are shown in Fig. 4. Fig. 4a shows that if we vary β in Eq. 5 between one and zero, the KL divergence term in the ELBO is significantly larger than zero for $\beta = 1$, and gradually decreases to zero. This is in line with our expectations. As we decrease β, the influence of the edge $U \to Y$ lessens, and X and X' are sufficient for modeling Y. In Fig. 4b, we demonstrate that a similar effect can be observed when we set $\beta = 0$, and vary the dataset size. This confirms our suspicions - even though we generated data from the multimodal SCM where the edge $U \to Y$ is not present, our finite datasets can appear to come from the spurious SCM where the edge does exist.



(a) Varying β

(b) Varying the dataset size

Figure 4: (a) We vary β from one to zero to show that our test can detect the presence of the edge $U \to Y$. (b) A similar result can be seen when we set $\beta = 0$, and vary the size of the dataset. This confirms our suspicions that a finite dataset sampled from the multimodal SCM can appear to come from the spurious SCM.

In our next set of experiments, we apply our method from Section 5 to mitigate the influence of $U \to Y$. For this, we generate a training set from the spurious SCM by setting $\beta = 1$, and compare $\log p(y \mid do(x), do(x'))$ to $\log p(y \mid x, x')$ across a range of test sets, where we vary β from one to zero. This simulates a test-time shift in $p(u)$. Our results in Fig. 5 show that $p(y \mid do(x), do(x')$ is more robust than $p(y \mid x, x')$ to the shift in $p(u)$, which validates our method.

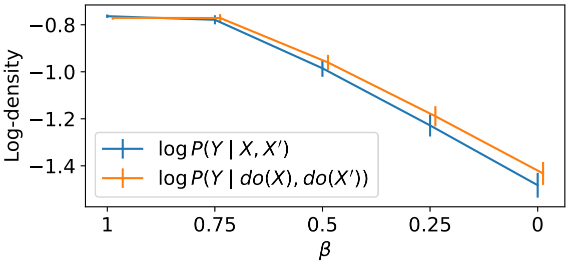


Figure 5: After training on data where $\beta = 1$, we evaluate across test sets where β ranges from zero to one, which represents a shift in $p(u)$. The causal model $p(y \mid do(x), do(x')$ is more robust than the statistical model $p(y \mid x, x')$ to this distribution shift.

8 Related work

Our work is strongly influenced by Wang and Blei [2019], who used a latent variable model to estimate a surrogate confounder, and perform the backdoor adjustment to estimate a causal model. Their approach, called the deconfounder, assumes the same causal graph as the spurious SCM in our work. The difference between our work is that they use a latent variable $\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{x}, \mathbf{x}')$ to model the dependence between the inputs. In contrast, our latent variable $\mathbf{z} \sim q(\mathbf{z} \mid \mathbf{x}, \mathbf{x}', \mathbf{y})$ models the dependence of $\mathbf{u}$ and $\mathbf{y}$ given the inputs.

Our work also belongs to a recent line of work which estimates intervention distributions to improve out-of-distribution generalization in the setting of high-dimensional unstructured data [Makino et al., 2022]. While we consider multimodal learning as opposed to multitask learning, our work is conceptually similar to Makino et al. [2022], since we also use the notion of unobserved confounding to explain weaknesses in machine learning, and use ideas from causality to improve them.

9 Conclusion

We addressed the issue of multimodal learning failing to perform better than unimodal learning. We put forth the multimodal SCM, under which multimodal learning estimates a causal model, and is more robust than unimodal learning to shifts in $p(\mathbf{u})$. We argue that multimodal learning can fail in practice because our finite datasets can appear consistent with an alternative SCM, which we call the spurious SCM. The spurious SCM contains the edge $\mathbf{u} \to \mathbf{y}$, which makes $\mathbf{u}$ and $\mathbf{y}$ conditionally dependent given the inputs. Under the spurious SCM, multimodal learning no longer estimates a causal model. Since the problem is with the edge $\mathbf{u} \to \mathbf{y}$, we devise a test that uses latent variable modeling to detect the influence of $\mathbf{u} \to \mathbf{y}$ in a given dataset. The same latent variable model can then be used to perform the backdoor adjustment and estimate a causal model. A current limitation of our work is that we used synthetic data experiments to validate our claims. We plan to move to a more realistic setting and experiment with more complex multimodal datasets. Looking forward, we are excited by the opportunities to use causality to explain and resolve difficult issues in machine learning.

References

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. In *EMNLP*, 2016.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *CVPR*, 2018.

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):423–443, 2019.

Rémi Cadène, Corentin Dancette, Hédi Ben-Younes, Matthieu Cord, and Devi Parikh. Rubi: Reducing unimodal biases in visual question answering. In *NeurIPS*, 2019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

William Falcon et al. Pytorch lightning. *GitHub. Note: https://github.com/PyTorchLightning/pytorch-lightning*, 3, 2019.

Itai Gat, Idan Schwartz, Alexander G. Schwing, and Tamir Hazan. Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies. In *NeurIPS*, 2020.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Taro Makino, Krzysztof J. Geras, and Kyunghyun Cho. Generative multitask learning mitigates target-causing confounding. In *NeurIPS*, 2022.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

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

Hieu Pham, Zihang Dai, Golnaz Ghiasi, Hanxiao Liu, Adams Wei Yu, Minh-Thang Luong, Mingxing Tan, and Quoc V. Le. Combined scaling for zero-shot transfer learning. *arXiv*, abs/2111.10050, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Sasha Sheng, Amanpreet Singh, Vedanuj Goswami, Jose Alberto Lopez Magana, Wojciech Galuba, Devi Parikh, and Douwe Kiela. Human-adversarial visual question answering. In *NeurIPS*, 2021.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *CVPR*, 2020.

Yixin Wang and David M Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596, 2019.

Mitchell Wortsman, Gabriel Ilharco, Mike Li, Jong Wook Kim, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. *arXiv*, abs/2109.01903, 2021.

Nan Wu, Stanislaw Jastrzebski, Jungkyu Park, Linda Moy, Kyunghyun Cho, and Krzysztof J. Geras. Improving the ability of deep neural networks to use information from multiple views in breast cancer screening. In *MIDL*, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]