Thierry-Séphine GOMA-LEGERNARD
Victor HOFFMANN

2023-2024

MAP670M - Causal Inference

CausalVAE:
Structured Causal Disentanglement in Variational Autoencoder

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, Jun Wang. 2023

GitHub Repository

`https://github.com/VictorHoffmann1/CausalVAE`

Contents

1 Background and motivation for the study

1.1 Limitation of the usual VAE frameworks used for disentangled representation learning

Generating synthetic counterfactual data is one of the key objectives behind Causal Inference. When it comes to generating unstructured data, researchers usually rely on generative models such as Variational Autoencoders (VAEs). VAEs [3] belong to a family of deep generative models. A VAE is an autoencoder that regulates its encoding distribution during training to ensure that its latent space possesses favourable properties for generating new data. Usually, its latent space distribution tends towards a centered gaussian distribution through the KL Divergence Loss. Then, a sample of the latent space distribution is used to reconstruct the input (achieved through the reconstruction loss).

$$\mathcal{L}_{\text{VAE}} = \underbrace{-\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]}_{\text{Reconstruction Loss}} + \underbrace{\text{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))}_{\text{KL Divergence Loss}} \tag{1}$$

where $q_\phi(\mathbf{z}|\mathbf{x})$ is the approximate posterior distribution (encoder) parameterized by ϕ, $p_\theta(\mathbf{x}|\mathbf{z})$ the likelihood function (decoder) parameterized by θ, $p(\mathbf{z})$ the prior distribution over the latent space (usually a multivariate Gaussian) and KL represents the Kullback-Leibler divergence.

[5] explores the use of VAEs for causal inference and disentanglement. Disentangled representation learning aims to discover a low-dimensional representation comprising multiple explanatory and generative factors of observational data, which can potentially enhance model performance and improve generalizability. A disentangled latent space is usually a space where the variables are independent and uncorrelated to each other. Such representation is encouraged by the KL Divergence Loss, since it tries to tend the latent space distribution to a multivariate Gaussian distribution.

Different approaches exist in order to reinforce a disentangled representation of the latent space. For example, [1] proposes to increase the weight of the KL Divergence Loss, reducing total correlation among latent variables (known as β-VAE). However, these approaches share the strong hypothesis that real-world observations are generated by countable independent factors. In real scenarios, however, factors with semantics are not necessarily independent. Instead, there might be an underlying causal structure that renders these factors dependent.

1.2 Contribution of the paper

The authors of [5] propose a new VAE-based framework named CausalVAE, which facilitates causal disentangled representation learning. This framework is designed to transform independent exogenous factors into causal endogenous ones, aligning with causally related concepts found in data.

The CausalVAE framework architecture incorporates a Structural Causal Model layer, referred to as the Mask Layer which is introduced within a Variational Autoencoder (VAE) framework. This layer enables the recovery of latent factors imbued with semantics and structure through a causal Directed Acyclic Graph (DAG). The model's process involves the passage of input signals through an encoder to obtain independent exogenous factors, followed by the Causal Layer for generating causal representations, which are then decoded to reconstruct the original input, thereby constituting the Causal Disentangled Representation Learning process. As opposed to a standard VAE, CausalVAE is weakly supervised. In addition to the input data, some information regarding state of the entities in the input need to be passed to the Structural Causal Model Layer so that the causality of the latent variables is well captured. Additionally, as demonstrated in [5], such weak supervision enables the model to avoid identifiability problems, where different latent samples can achieve the same representation.

To effectively train the model, the authors of [5] propose a novel loss function incorporating the usual VAE Loss alongside an acyclicity constraint imposed on the learned causal graph to ensure its adherence to a Directed Acyclic Graph structure. This hybrid loss function facilitates the optimization process and reinforces the integrity of the learned causal relationships within the model.

Comprehensive experiments conducted on both synthetic and real-world face images showcase the efficacy of the learned factors with causal semantics, displaying their utility in generating counterfactual images not present in the training data.

2 Causal VAE framework

2.1 Architecture and equations

We present in this section the architecture of the model.

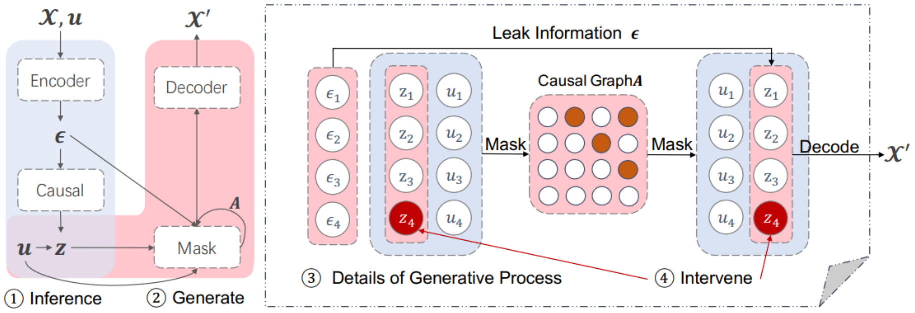


Figure 1: Overview of the CausalVAE model structure

2.1.1 Encoder and Decoder Layers

Let's consider n concepts of interest which are present in each observation of the dataset. Those concepts are causally structured by a Directed Acyclic Graph (DAG) with an adjacency matrix A.

Let's denote by $x \in \mathbb{R}^d$ the observed variables, $u \in \mathbb{R}^n$ the additional information (weak signals given with the input) where u_i represents the label of the ith concept of interest, $\epsilon \in \mathbb{R}^n$ the latent exogenous independent variables and $\mathbf{z} \in \mathbb{R}^n$ the latent endogenous variables with semantics where $\mathbf{z} = A^T\mathbf{z} + \epsilon = (I - A^T)^{-1}\epsilon$. For simplicity, we denote $C = (I - A^T)^{-1}$. Both $\mathbf{z}$ and ϵ are treated as latent variables.

Let $f(\mathbf{z})$ denote the decoder function, assumed to be invertible and $h(x, u)$ denote the encoder function. The model assumes that $\epsilon = h(x, u) + \zeta$ (encoding process) and that $x = f(z) + \xi$ (decoding process) where ξ and ζ are the vectors of independent noise with probability densities p_ξ and q_ζ. When ξ and ζ are infinitesimal, the encoder and decoder can be regarded as deterministic ones.

We consider the following conditional generative model parameterized by $\theta = (f, h, C, T, \lambda)$

$$p_\theta(x, \mathbf{z}, \epsilon | u) = p_\theta(x|\mathbf{z}, \epsilon, u) p_\theta(\epsilon, \mathbf{z}|u)$$

The generative and inference models are defined as follows:

$$p_\theta(x|\mathbf{z}, \epsilon, u) = p_\theta(x|\mathbf{z}) \equiv p_\xi(x - f(\mathbf{z})) \ , \ q_\phi(\mathbf{z}, \epsilon|x, u) = q(\mathbf{z}|\epsilon) q_\zeta(\epsilon - h(x, u))$$

We define the joint prior $p_\theta(\epsilon, \mathbf{z}|u)$ for latent variables $\mathbf{z}$ and ϵ as:$p_\theta(\epsilon, z|u) = p_\epsilon(\epsilon) p_\theta(z|u)$
where $p_\epsilon(\epsilon) = \mathcal{N}(0, I)$ and the prior of latent endogenous variables $p_\theta(z|u)$ is a factorized Gaussian distribution conditioning on the additional observation u

$$p_\theta(\mathbf{z}|u) = \prod_{i=1}^{n} p_\theta(\mathbf{z}_i|u_i), \quad p_\theta(\mathbf{z}_i|u_i) = \mathcal{N}(\lambda_1(u_i), \lambda_2^2(u_i)) \textit{ where } \lambda_1 \text{ and } \lambda_2 \text{ are arbitrary functions}$$

In [5], $\lambda_1(u) = u$ and $\lambda_2(u) \equiv 1$.
The distribution has two sufficient statistics, the mean and variance of $\mathbf{z}$, which are denoted by sufficient statistics $T(\mathbf{z}) = (\mu(\mathbf{z}), \sigma(\mathbf{z})) = (T_{1,1}(\mathbf{z}_1), \ldots, T_{n,2}(\mathbf{z}_n))$.

2.1.2 Structural Causal Model Layer (SCM)

We precise that causal representation are ones structured by a causal graph. Aligned with the authors notation, we refer to Causal Discovery to describe the task of discovering the causal graph from pure observations.

Objective: Learn causal representations
Variables : $\epsilon \to z$
Equation: $z = A^T z + \epsilon$
$z = (I - A^T)^{-1} \epsilon$
where :
A is the parameter to be learned in this layer.
ϵ are Independent Gaussian exogenous factors
$\mathbf{z} \in \mathbb{R}^n$ is structured causal representation of n concepts that is generated by a DAG and thus A can be permuted into a strictly upper triangular matrix.

2.1.3 Mask Layer

Once the causal representation $\mathbf{z}$ is obtained, it passes through a Mask Layer to reconstruct itself. Parameters to do so are trained by minimizing the reconstruction error.

Let $\mathbf{z}_i$ denote the ith variable in the vector $\mathbf{z}$. The adjacency matrix associated with the causal graph is $A = [A_1| \ldots |A_n]$, where $A_i \in \mathbb{R}^n$ is the weight vector such that A_{ji} encodes the causal strength from $\mathbf{z}_j$ to $\mathbf{z}_i$.

While previous equation implies $\mathbf{z}_i = A_i^T \mathbf{z} + \epsilon_i$, the authors found that incorporating a mild nonlinear function g_i enhances stability. Thus, a set of mild non-linear and invertible functions is utilized. These are $[g_1, g_2, \ldots, g_n]$ that map parental variables to the child variable.

Consequently, we express the relationship as follows:

$$\mathbf{z}_i = g_i(A_i \circ \mathbf{z}; \eta_i) + \epsilon_i,$$

Objective: Learn adjacency matrix A and parameter η allowing to reconstruct the causal representation
Variables : $\mathbf{z} \to \mathbf{z}$
Equation: $\mathbf{z}_i = g_i(A_i \circ \mathbf{z}; \eta_i) + \epsilon_i$
where :
$\circ$ denotes element-wise multiplication
η_i is the parameter of $g_i(\cdot)$

This layer makes intervention or "do-operation" possible.
To intervene on $\mathbf{z}_i$, its value is fixed on the right side of Equation, corresponding to the i-th node of $\mathbf{z}$ in the first layer in Figure 1. Consequently, its effect propagates to all its children nodes, including itself, on the left side of Equation corresponding to certain nodes of $\mathbf{z}$ in the second layer.
[5] notes that intervening the cause will change the effect, whereas intervening the effect, on the other hand, does not change the cause because information can only flow into the next layer from the previous one in our model, which is aligned with the definition of causal effects.

2.2 Learning strategy

In this section, we outline the methodology for training the CausalVAE model to simultaneously learn the causal representation and the causal graph.

2.2.1 General reminders regarding ELBO

The Evidence Lower Bound (ELBO) is an expression used in probabilistic modeling for approximating complex probability distributions. It allows us to approximate the posterior distribution of latent variables given observed data, which is often intractable due to model complexity. Variational inference simplifies this by approximating the true posterior with a parameterized distribution.

The ELBO provides a lower bound on the log marginal likelihood, or evidence, which represents the probability of observing the data under the model. Mathematically, it is expressed as the expectation of the log joint distribution minus the expectation of the log variational distribution.

$$\text{ELBO} = \mathbb{E}_q[\log p(x,z)] - \mathbb{E}_q[\log q(z)]$$

Here, $p(x,z)$ is the joint distribution of observed data x and latent variables z, and $q(z)$ is the variational distribution approximating the true posterior. Maximizing the ELBO minimizes the KL divergence between the variational distribution and the true posterior, typically achieved iteratively through optimization techniques like gradient descent. In the context of a classic VAE, we can notice that its loss function shown in Equation (1) is exactly equivalent to maximizing the ELBO.

2.2.2 Evidence Lower Bound of CausalVAE

The authors of [5] utilize variational Bayes to learn a tractable distribution $q_\phi(\epsilon, z|x, u)$ approximating the true posterior $p_\theta(\epsilon, z|x, u)$. The parameters θ and ϕ are optimized by maximizing the Evidence Lower Bound (ELBO) over a dataset X with empirical data distribution $q_X(x, u)$. The ELBO is defined as:

$$\text{ELBO} = \mathbb{E}_{q_X}\left[\mathbb{E}_{\epsilon,z\sim q_\phi}[\log p_\theta(x|z,\epsilon,u)] - D(q_\phi(\epsilon,z|x,u)||p_\theta(\epsilon,z|u))\right]$$

where $D(\cdot \parallel \cdot)$ denotes the KL divergence. Simplifying the variational posterior due to the one-to-one correspondence between ϵ and z, the ELBO further simplifies to:

$$\text{ELBO} = \mathbb{E}_{q_X}\left[\mathbb{E}_{q_\phi}(z|x,u)[\log p_\theta(x|z)] - D(q_\phi(z|x,u)||p_\theta(z|u)) - D(q_\phi(\epsilon|x,u)||p_\epsilon(\epsilon))\right]$$

2.2.3 Learning the Causal Structure of Latent Codes

In addition to the encoder and decoder, the CausalVAE model integrates a Causal Layer with a Directed Acyclic Graph (DAG) structure for learning.
To ensure the causal graph's identifiability, extra labels u are used to impose constraints. These constraints include a label constraint l_u and a latent code constraint l_m, ensuring the adherence to predefined limits κ_1 and κ_2, respectively, where $l_u = \mathbb{E}_{q_X}||u - \sigma(A^T u)||_2^2, l_m = \mathbb{E}_{\mathbf{z}\sim q_\phi} \sum_{i=1}^n ||\mathbf{z}_i - g_i(A_i \circ \mathbf{z}; \eta_i)||_2^2$.

Moreover, the causal adjacency matrix A must conform to a DAG structure, enforced by a differentiable constraint function $H(A)$, maintaining $H(A) = 0$, $H(A) \equiv \text{tr}((I + c \cdot A \circ A)^n)$
This translates the training objective into a constrained optimization problem, aiming to maximize the ELBO while satisfying $l_u \leq \kappa_1$, $l_m \leq \kappa_2$, and $H(A) = 0$.

Using the Lagrangian multiplier method, the new loss function is defined as

$$\mathcal{L}_{\text{CausalVAE}} = -\text{ELBO} + \alpha H(A) + \beta l_u + \gamma l_m$$

where α, β, and γ serve as regularization hyperparameters.

3 Results

3.1 Identifiability

The authors of the paper have shown that the model allows for identifiability, which is the ability to uniquely determine the parameters of the generative model from the observed data. Identifiability is guaranteed under several key assumptions and properties of the model:

- The observed data is generated according to the introduced conditional generative model parameterized by $\theta = (f, h, C, T, \lambda)$.
- The set $\{x \in X | \phi_\xi(x) = 0\}$ has measure zero, where ϕ_ξ is the characteristic function of the density p_ξ defined previously.
- The decoder function f is differentiable, and the Jacobian matrix of f is of full rank, ensuring invertibility in the mapping from latent variables to observed data.
- The sufficient statistics are such that $T_{i,s}(\mathbf{z}_i) \neq 0$ almost everywhere for all $1 \leq i \leq n$ and $1 \leq s \leq 2$, where $T_{i,s}(\mathbf{z}_i)$ is the s-th statistic of variable $\mathbf{z}_i$.
- Additional observations $u_i \neq 0$.

3.2 Achieved Performances

The Causal VAE framework has been assessed on data (*Flow and Pendulum*) and on image dataset (*CelebA*). In this section, we detail the achieved performances of the authors on *CelebA* data.

CelebA Dataset

The CelebA dataset comprises images depicting the faces of celebrities, accompanied by annotations detailing various attributes. The focal input x encompasses attributes such as gender, smile, eyes open, and mouth open. The causal graph proposed by the authors exhibits the following relation:

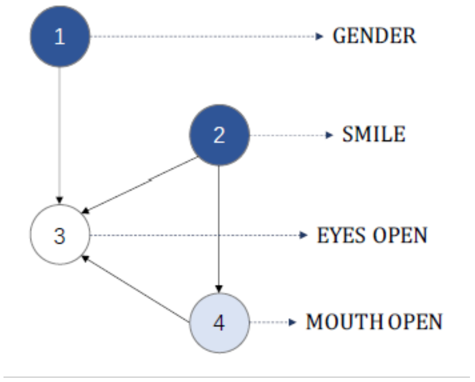
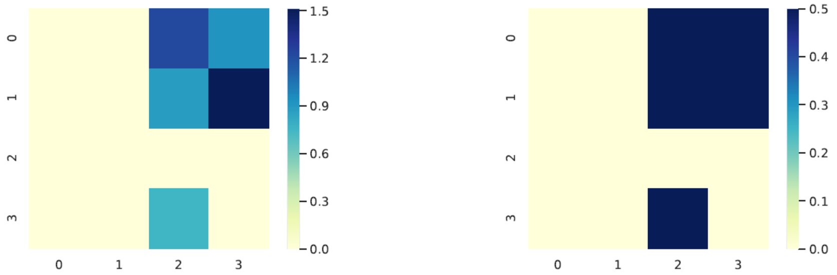


Figure 2: Causal graph of the CelebA dataset

The weak signals u consist of 5 attribute localisations (right eye, left eye, nose, right-side mouth, left-side mouth), along with 40 binary variables such as *High_Cheekbones*, *Arched_Eyebrows*, or *Male Mouth_Slightly_Open*, as well as indicators like *Wearing_Necklace* or *Wearing_Earrings*, for example.

Hereafter, we present the results obtained by [5] for the task of learning the Adjacency Matrix of the underlying Directed Acyclic Graph.



(c) After 5 epochs

(d) The true

Figure 3: Visualization of the learning process for the Adjacency Matrix

The results obtained through this framework are conclusive in terms of approximating the ground truth of the Adjacency Matrix.

4 Paper Implementation

To implement the paper, we drew inspiration from the GitHub repository of the authors, which is available here. We adapted their repository to fit our resources (memory, GPU, disk storage), the datasets used for training and evaluation, and the scores/plots we intended to display. Due to the lengthy process of obtaining results and the necessity of a GPU, we won't provide a notebook but instead a modified version of the original repository on GitHub. Our repository can be accessed here (or you can use the link from the cover page).

4.1 Training

We decided to train on two datasets mentioned in the paper, which were previously described: *Flow* and *Pendulum*. Since these datasets contain relatively simple images, we aimed to keep the model as simple as possible. Therefore, we designed an architecture relatively similar to the one described in the article, with the encoder consisting of 4 layers and the decoder consisting of 4 linear layers per concept (There are 4 concepts in *Flow*: Ball Size, Water Size, Hole, Water Flow, and 4 concepts in *Pendulum*: Light, Pendulum Angle, Shadow Length, Shadow Position). No convolutional layers were used. The number of epochs was set to 100, and the regularization hyperparameters α, β, and γ, as shown previously, were all set to 1. The optimizer used was Adam [2] with a learning rate of 0.001.

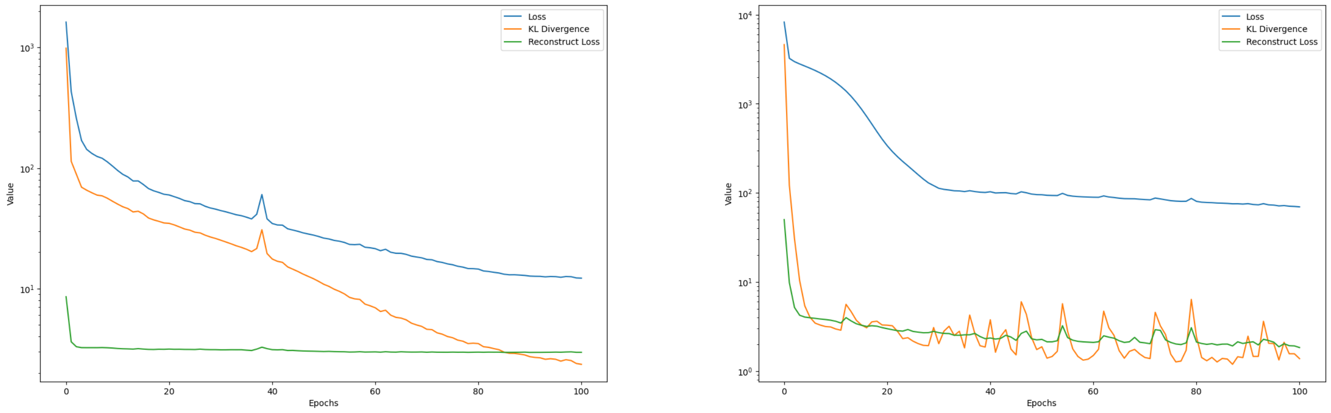
At the beginning of training, the reconstruction loss can dominate, causing the VAE to ignore the KL divergence and learn a trivial latent variable distribution. To prevent this, the KL divergence term is gradually "warmed-up" from 0 to its full contribution over a number of epochs. This is called Linear Deterministic Warm-Up [4]. We employed this approach in training, allowing the contribution of the KL divergence to transition from 0 to 1 over the first 50 epochs.

Figures 4a and 4b show the training losses (Total Loss, KL Divergence, and Reconstruction Loss) as functions of the number of epochs on the *Flow* and *Pendulum* datasets. As we can see, Causal-VAE manages to decrease the training loss on both datasets as the epochs progress. However, the behaviors of KL Divergence and Reconstruction Loss differ between *Flow* and *Pendulum*.

- In the case of *Flow*, the Reconstruction Loss decreases rapidly in the initial epochs and then stabilizes. This suggests that our model quickly learns to reconstruct the images well (see Figure 5) and subsequently fine-tunes the reconstruction. Conversely, the KL Divergence decreases sharply in the initial epochs and continues to decrease significantly thereafter. This indicates that our latent distribution before the Structural Causal Model Layer approaches a standard normal distribution more closely while maintaining the same reconstruction loss, thereby enhancing the ease of sampling and generating new data.

- Regarding *Pendulum*, both the Reconstruction Loss and KL Divergence decrease rapidly after only a few epochs and then converge to a minimum value. They also exhibit more noise compared to the *Flow* dataset, likely due to the higher variance in *Pendulum*. Additionally, we observe that between the 5th and 30th epochs, the Total Loss decreases much more than the Representation Loss and KL Divergence. This indicates that the additional terms in the loss: l_u, l_m, and $H(A)$ described in the equation... also decrease, suggesting that the Causal Layer with a DAG Structure is being learned.

The figures 5 and 6 show examples of the reconstruction of an image over the epochs by Causal-VAE on both datasets.



(a) Total Loss, KL-Divergence and Reconstruction Loss as a function of the number of epochs for *Flow*

(b) Total Loss, KL-Divergence and Reconstruction Loss as a function of the number of epochs for *Pendulum*.

Figure 4: History of the training for both *Pendulum* and *Flow* datasets.

- On *Flow*, we can clearly observe the progress of the reconstruction as the epochs progress. Initially, the water and the sphere are blurry, and there is no leakage. By epoch 40, the model begins to reproduce the leakage despite minimal water and a small hole. By the final epoch, the leakage is well-reconstructed, and the sphere matches the size of the original image.
- Similarly, for *Pendulum*, in the first epoch, the model struggles to reproduce the sun, the length, and position of the shadow. However, after 40 epochs, the model gradually improves its ability to reproduce the angle, light, and shadow accurately.

One observation applicable to both datasets is that the colors of the reconstructed images appear slightly different from those of the original images. This discrepancy may be attributed to the size of the decoder, which is relatively small in terms of the number of parameters. Despite this, CausalVAE effectively captures the causal relationships between different objects in the image, which is the primary focus.

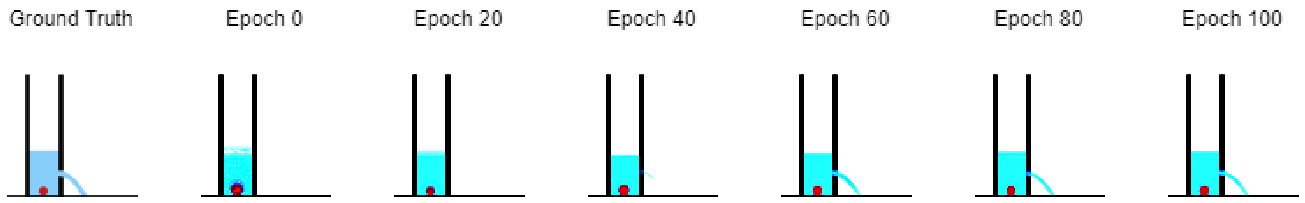


Figure 5: Example of the reconstruction of an image over the epochs by CausalVAE on *Flow*. As we can see, the model gradually improves its reconstruction quality over the epochs.

4.2 Inference

4.2.1 Reconstruction of the Image

After training the model, we evaluated its performance on both datasets. Figures 7 and 8 show eight different reconstruction tests on *Flow* and *Pendulum*.

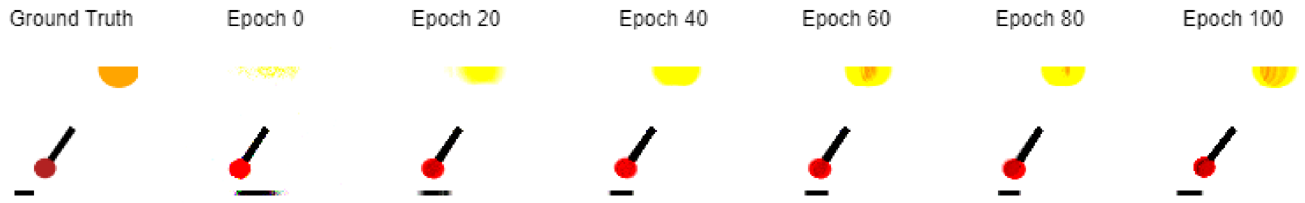


Figure 6: Example of the reconstruction of an image over the epochs by CausalVAE on *Pendulum*. As we can see, the model is gradually able to correctly reconstruct the light, pendulum and shadow so that the causality is respected (Light & Pendulum Angle → Position & Length of the Shadow).

- On *Flow*, the predictions appear very close to the ground truth images (despite the color issue mentioned earlier). The only noticeable difference may be the size of the hole causing the leakage: CausalVAE tends to predict smaller holes than in the original image.
- On *Pendulum*, the performance seems slightly weaker. Although the overall result is satisfactory, we can still observe that the model struggles to represent the sun as a disk, and the length of the shadows sometimes differs slightly from the originals.

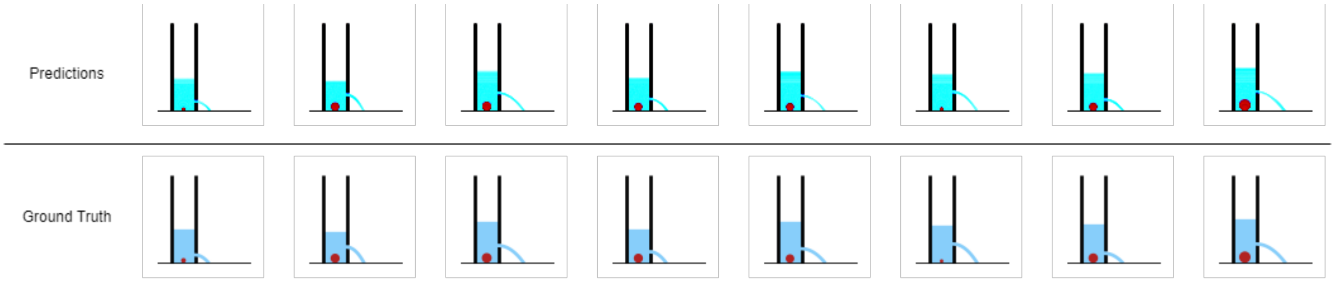


Figure 7: 8 different reconstruction tests on *Flow* dataset.

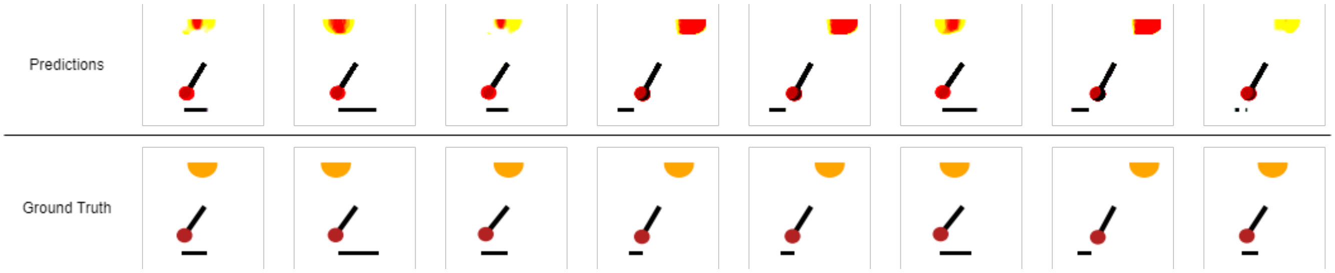


Figure 8: 8 different reconstruction tests on *Pendulum* dataset.

4.2.2 Measuring Causal Disentanglement

As mentioned in 2.1.1, the Causal VAE is trained in a weakly supervised manner, meaning that some information about the concepts in each image is provided. For example, in *Pendulum*, the position of the sun, the angle of the pendulum, and the position and length of the shadow will be given to the model in addition to the image. This means that we can compare how close the latent representation $\mathbf{z}$ is to the labels. In other words, we can measure if the latent representation corresponds to the concepts of the image and their causal links. We assessed this information using the Maximum Information Coefficient (MIC) and the Total Information Coefficient (TIC) between the latent representation $\mathbf{z}$ and the labels u.

MIC and TIC are both measures of statistical dependence between two variables, making them suitable for evaluating the relevance of latent variables in a Causal VAE.

MIC is a measure that captures a wide range of associations both functional and non-functional, and it assigns a score between 0 (no association) and 1 (perfect association). It's useful for detecting any kind of relationship, not just linear ones. This makes it a good choice for evaluating the relevance of latent variables, as the relationship between these variables and the labels may not be linear.

TIC, on the other hand, is a measure of total information between the joint distribution of two random vectors and the product of their marginal distributions. It's a multivariate extension of mutual information and can capture both linear and non-linear dependencies. This makes it a good choice for evaluating the relevance of latent variables in a Causal VAE, as it can capture complex, multivariate relationships.

Table 1 shows the MIC and TIC we obtained on both datasets. Our results are slightly lower than those in the article, which might be due to a different model architecture, a higher number of epochs, or a better learning rate. As we can see, the model is able to reconstruct the causality of the concepts quite well on *Pendulum*, whereas it struggles a bit more on *Flow*.

Dataset	**MIC**	**TIC**
Flow	69.1 ± 2.1	52.9 ± 1.3
Pendulum	91.3 ± 1.7	79.7 ± 2.3

Table 1: The Maximum Information Coefficient (MIC) and Total Information Coefficient (TIC) between learned representation $\mathbf{z}$ and the label u.

References

[1] Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae, 2018.

[2] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[3] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.

[4] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders, 2016.

[5] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Structured causal disentanglement in variational autoencoder. 2023.