Variational Autoencoders

Beta VAE, Ladder VAE, Causal VAE

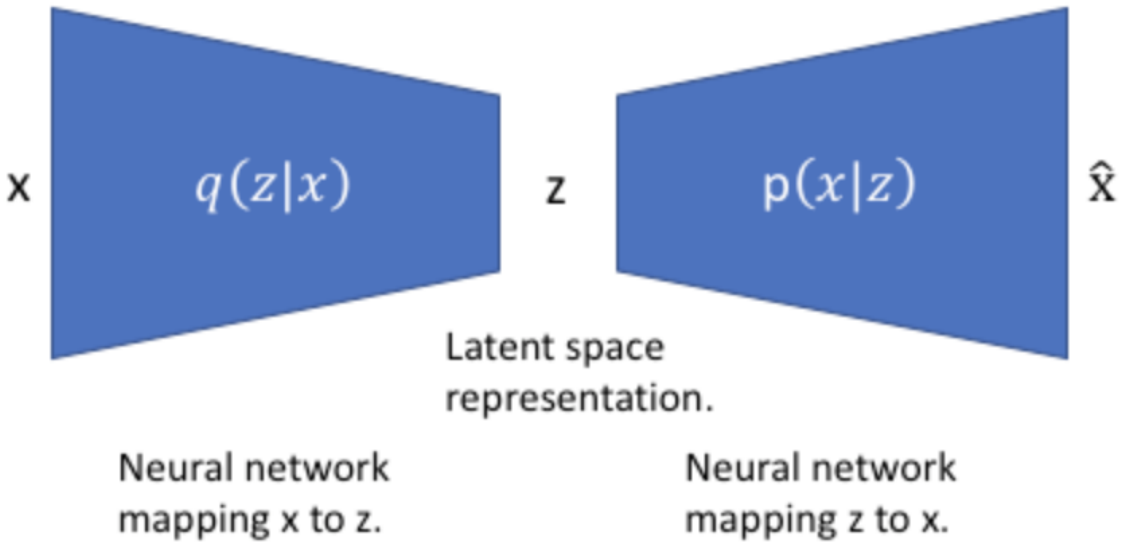
Presenter: Arshdeep Sekhon

https://qdata.github.io/deep2Read

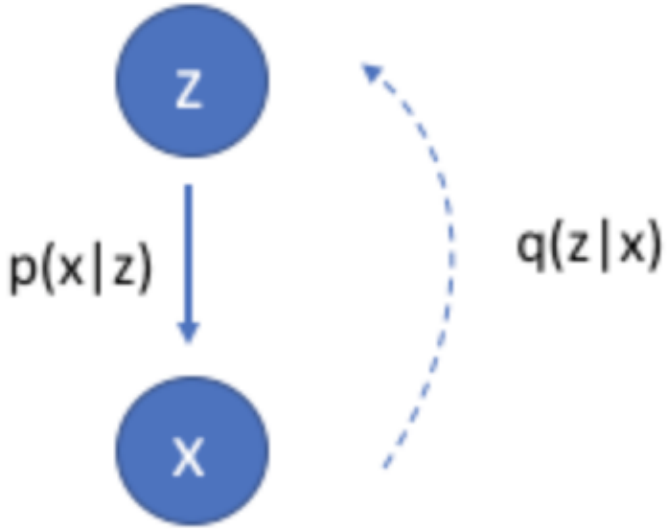
β−VAE: Learning Basic Visual Concepts with a Constrained Variational Framework

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, Alexander Lerchner

Variational Auto-Encoder



Variational Auto-Encoder



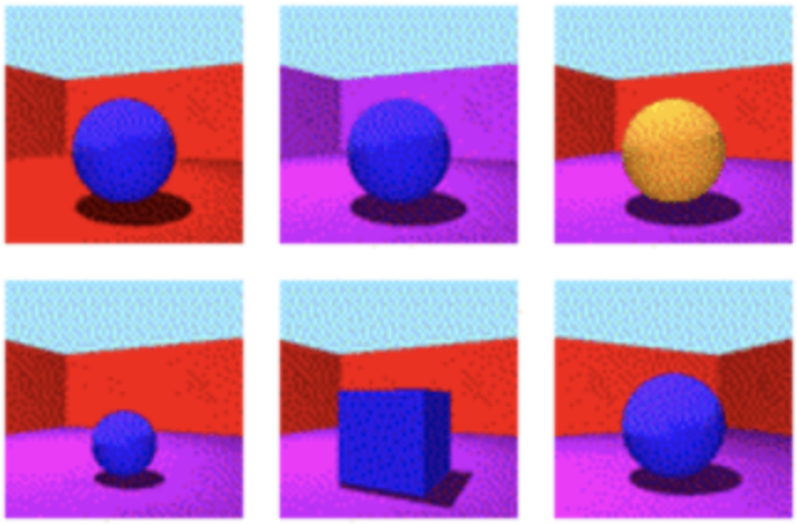
Maximize ELBO:

$$E_{q(z|x)} log(p(x|z)) - KL(q(z|x)||p(z)) \quad (1)$$

- Term 1: Reconstruction
- Term 2: prior

Objective

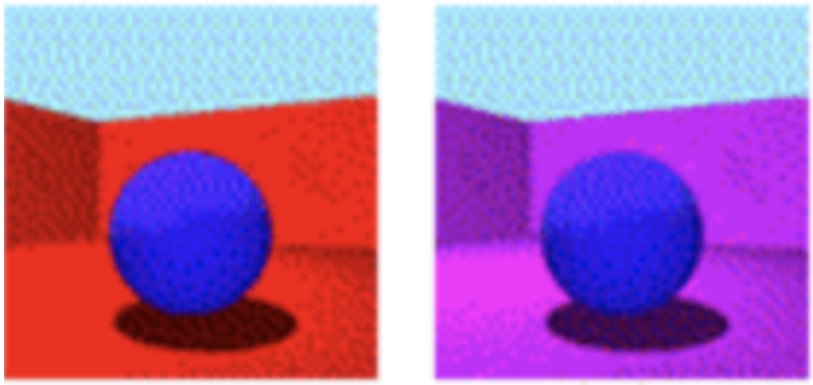
Learning independent factors of generation in an unsupervised manner



- In the above image, generated image depends on : color of walls, size, shape, color of object

Disentangled representations

- If we change one latent factor in a disentangled representation, it corresponds to changes in only one generative factor
- For example, in the above 3D image , change $\boldsymbol{z}$ corresponding to wall colors, image remains same, and only background changes.



Why disentangled representations

- generalize better to unseen situations: useful in zero-shot or knowledge transfer
- boost AI performance: (Lake 2016)

$\beta-$VAE

- generative model to learn a disentangled $\boldsymbol{z}$ in an *unsupervised* manner
- no prior information regarding number of factors, or correspondence

Method

- Given dataset $D = \{\boldsymbol{X}, \boldsymbol{V}, \boldsymbol{W}\}$
- images $\boldsymbol{x} \in R^N$
- conditionally independent facotrs $\boldsymbol{v} \in R^K$
- conditionally dependent factors $\boldsymbol{w} \in R^H$
- these factors are independent $log(p(\boldsymbol{v}|\boldsymbol{x})) = \sum_k logp(\boldsymbol{v}_k|\boldsymbol{x})$
- assumption: $p(\boldsymbol{x}|\boldsymbol{v}, \boldsymbol{w}) = Sim(\boldsymbol{v}, \boldsymbol{w})$, where Sim is the true simulator that generates images given $\boldsymbol{v}, \boldsymbol{w}$

Goal

Learn a joint distribution of $p(\boldsymbol{x}, \boldsymbol{z})$ where $\boldsymbol{z} \in R^M, M \geq K$, such that

$$p(\boldsymbol{x}|\boldsymbol{z}) \sim p(\boldsymbol{x}|\boldsymbol{v}, \boldsymbol{w}) = Sim(\boldsymbol{v}, \boldsymbol{w}) \tag{2}$$

Maximize the marginal log likelihood of observed data $\boldsymbol{x}$:

$$max_\theta E_{p_\theta(\boldsymbol{z})}[p_\theta(\boldsymbol{x}|\boldsymbol{z})] \tag{3}$$

Posterior of the inferred latent factors $\boldsymbol{z}$:

$$q_\phi(\boldsymbol{z}|\boldsymbol{x}) \tag{4}$$

Goal is to ensure $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ corresponds to $\boldsymbol{v}$ in a disentagled manner.

$\beta-$VAE

To achieve disentanglement or statistical Independence, set the prior to be an isotropic Gaussian :

$$p(\boldsymbol{z}) = N(0, \boldsymbol{I}) \tag{5}$$

$$max_{\theta,\phi} E_{x\sim D} E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})} log\, p_\theta(\boldsymbol{x}|\boldsymbol{z}) \tag{6}$$

$$s.t. D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})) < \epsilon \tag{7}$$

Rewrite as Lagrangian under the KKT conditions:

$$max_{\theta,\phi} E_{x\sim D} E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})} log\, p_\theta(\boldsymbol{x}|\boldsymbol{z}) - \beta(D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})) - \epsilon) \tag{8}$$

$$max_{\theta,\phi} E_{x\sim D} E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})} log\, p_\theta(\boldsymbol{x}|\boldsymbol{z}) - \beta(D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z}))) \tag{9}$$

Trade off

- Term 1 encourages better representations for Reconstruction fidelity
- Term 2 or high beta values try to make the dimensions as independent of each other as possible

Likelihood is a poor metric to measure disentanglement

- Disentangled representations emerge when the right balance is found between reconstruction cost as regularisation and latent channel capacity restriction ($\beta > 1$).
- $\beta > 1$ can lead to poorer reconstructions due to the loss of high frequency details when passing through a constrained latent bottleneck.
- need of a new metric to measure disentanglement

New Disentanglement Meitrc

- disentangled must be interpretable: can generate images of *small*, *green* apples, and *large*, *green* apples by varying the *small* latent
- independence can be obtained using PCA or ICA, but not interpretable!
- cross correlation is not a good metric
- target is to measure both Independence and interpretability

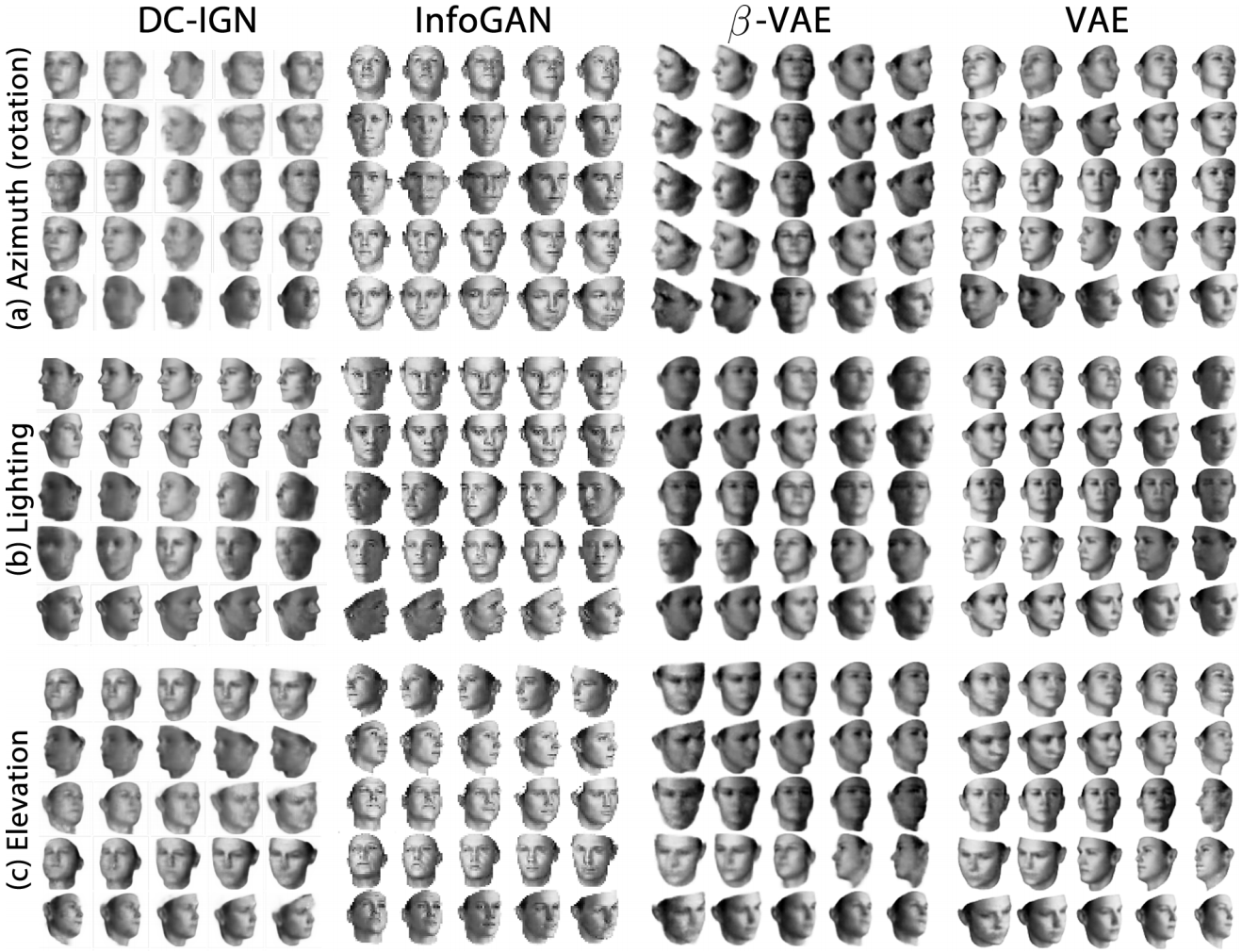
Disentanglement Meitrc

- we have labels of the generative factors $v \in V$ for some examples
- Choose an independent factor randomly $y \sim Unif[1, \ldots, K]$
- For a batch of L samples
  - sample two sets of latent representations $\boldsymbol{v}_{1,l}$ and $\boldsymbol{v}_{2,l}$ enforcing $[\boldsymbol{v}_{1,l}]_k = [\boldsymbol{v}_{2,l}]_k$ if $k = y$ the value of the factor is fixed
  - Get $\boldsymbol{x}_{1,l} \sim Sim(\boldsymbol{v}_{1,1})$ then infer $\boldsymbol{z}_{1,l} = \mu(\boldsymbol{x}_{1,l})$ using encoder $q(\boldsymbol{z}|\boldsymbol{x}) \sim N(\mu(\boldsymbol{x}), \sigma(\boldsymbol{x}))$
  - similarly for the second batch $\boldsymbol{v}_{2,l}$
  - $\boldsymbol{z}^b_{diff} = |\boldsymbol{z}_{1,l} - \boldsymbol{z}_{2,l}|$ the linear difference between the two latent
  - take average across the batch $\boldsymbol{z}^b_{diff} = \frac{1}{L}\sum_{l=1}^{L}(\boldsymbol{z}^l_{diff})$,and $p(y|\boldsymbol{z}^b_{diff})$, this score is disentanglement metric

Disentanglement Metric

- accuracy of this classifier over multiple batches is used as disentanglement metric score.
- $p(y|\boldsymbol{z}^{b}_{diff})$: linear classifier with low VC-dimension

Experiments



Experiments: Simulation Dataset

- 2D shapes
- Cartesian product of the shape and four independent generative factors defined in vector graphics: position X (32 values), position Y (32 values), scale (6 values) and rotation (40 values over the 2π range.

Experiments

Model	Disentanglement metric score
Ground truth	*100%*
Raw pixels	$45.75 \pm 0.8\%$
PCA	$84.9 \pm 0.4\%$
ICA	$42.03 \pm 10.6\%$
DC-IGN	$\mathbf{99.3 \pm 0.1\%}$
InfoGAN	$73.5 \pm 0.9\%$
VAE untrained	$44.14 \pm 2.5\%$
VAE	$61.58 \pm 0.5\%$
β-VAE	$\mathbf{99.23 \pm 0.1\%}$

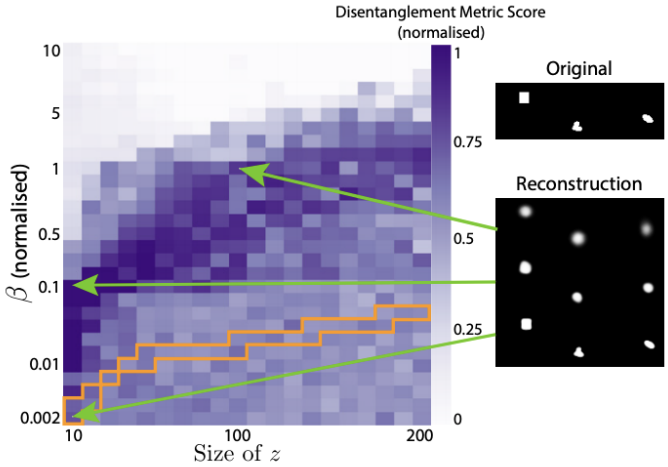
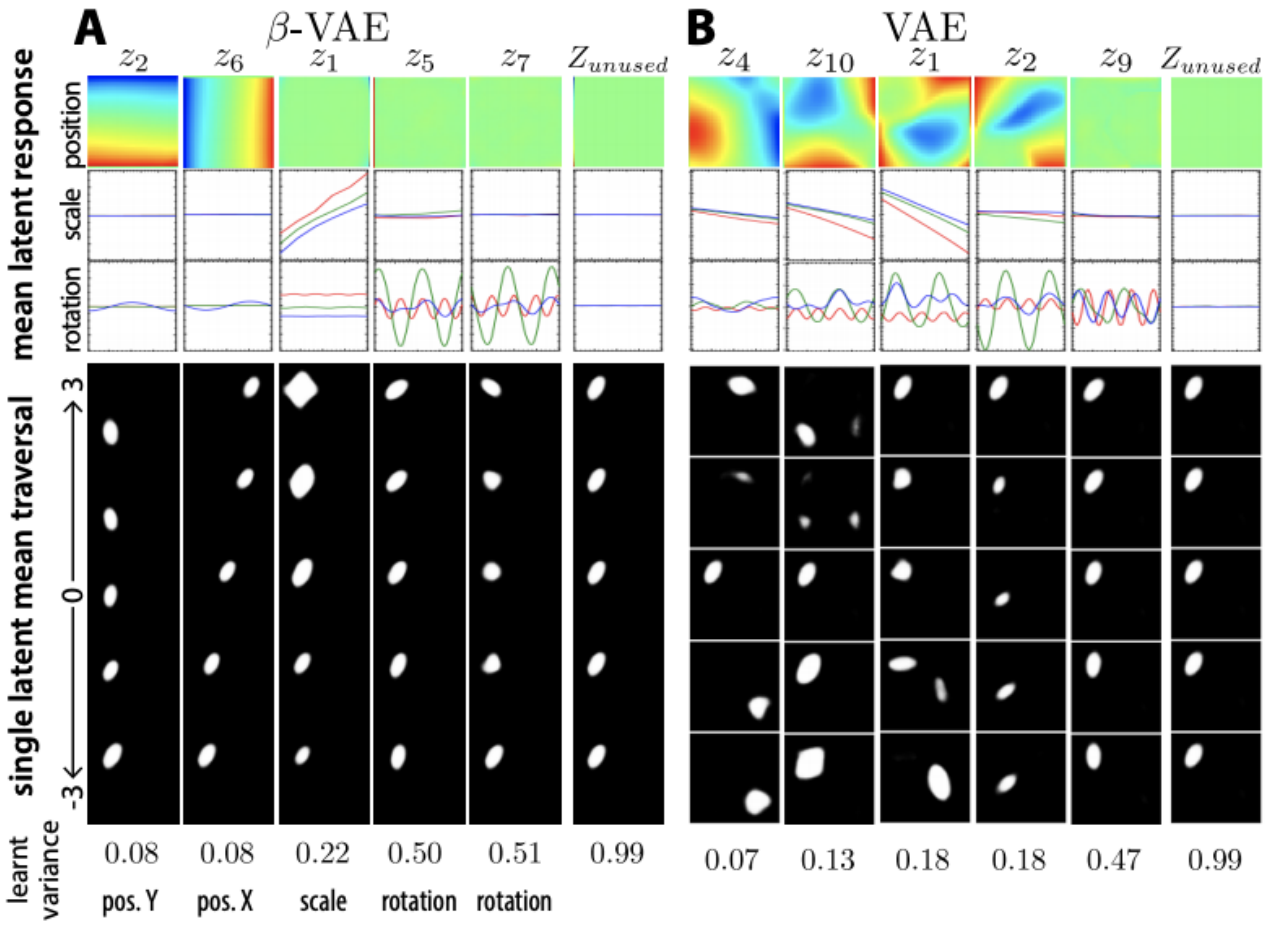
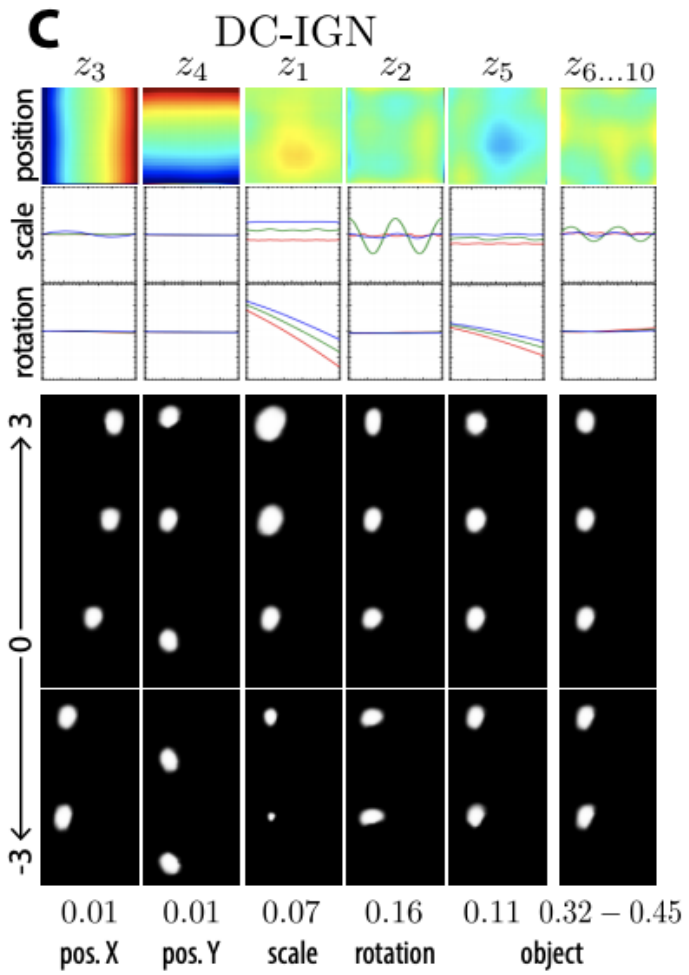


Figure 6: Disentanglement metric classification accuracy for 2D shapes dataset. **Left:** Accuracy for different models and training regimes **Right:** Positive correlation is present between the size of $\mathbf{z}$ and the optimal *normalised* values of β for disentangled factor learning for a fixed β-VAE architecture. β values are normalised by latent $\mathbf{z}$ size m and input $\mathbf{x}$ size n. Note that β values are not uniformly sampled. Orange approximately corresponds to *unnormalised* $\beta = 1$. Good reconstructions are associated with entangled representations (lower disentanglement scores). Disentangled representations (high disentanglement scores) often result in blurry reconstructions.

Experiments



Experiments



Conclusion

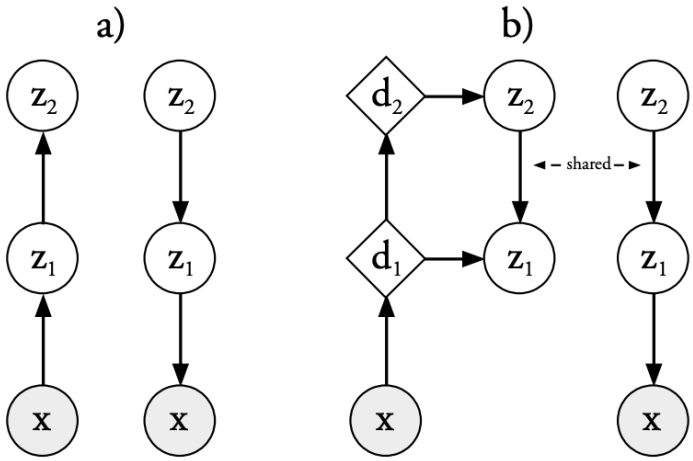
- reformulated the standard VAE framework as a constrained optimisation problem with strong latent capacity constraint and independence prior pressures.
- covers a wider range of factor values and is disentangled more cleanly than other benchmarks, all in a completely unsupervised manner

Ladder Variational Autoencoders

Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, Ole Winther

Motivation

- hierarchies of conditional stochastic variables
- structured inference model using the same top-down dependency structure both in the inference and generative models.



Ladder VAE

- VAEs and LVAEs simultaneously train a generative model $p_\theta(\boldsymbol{x}, \boldsymbol{z}) = p_\theta(\boldsymbol{x}|\boldsymbol{z})p_\theta(\boldsymbol{z})$ for data $\boldsymbol{x}$ using latent variables $\boldsymbol{z}$
- inference model $q_\phi(\boldsymbol{z}|\boldsymbol{x})$
- variational lower bound to the likelihood $p_\theta(\boldsymbol{x}) = \int p_\theta(\boldsymbol{x}, \boldsymbol{z})d\boldsymbol{z}$.

VAE: generative model

$$p_\theta(\boldsymbol{z}) = p_\theta(\boldsymbol{z}_L) \prod_{i=1}^{L-1} p_\theta(\boldsymbol{z}_i|\boldsymbol{z}_{i+1}) \tag{10}$$

$$p_\theta(\boldsymbol{z}_i|\boldsymbol{z}_{i+1}) = \mathcal{N}\left(\boldsymbol{z}_i|\boldsymbol{\mu}_{p,i}(\boldsymbol{z}_{i+1}), \boldsymbol{\sigma}^2_{p,i}(\boldsymbol{z}_{i+1})\right), \quad p_\theta(\boldsymbol{z}_L) = \mathcal{N}\left(\boldsymbol{z}_L|0, \boldsymbol{I}\right) \tag{11}$$

$$p_\theta(\boldsymbol{x}|\boldsymbol{z}_1) = \mathcal{N}\left(\boldsymbol{x}|\boldsymbol{\mu}_{p,0}(\boldsymbol{z}_1), \boldsymbol{\sigma}^2_{p,0}(\boldsymbol{z}_1)\right) \text{ or } P_\theta(\boldsymbol{x}|\boldsymbol{z}_1) = \mathcal{B}\left(\boldsymbol{x}|\boldsymbol{\mu}_{p,0}(\boldsymbol{z}_1)\right) \tag{12}$$

- p indicates generative model parameters, q indicates inference model parameters
- observation models is matching either continuous-valued (Gaussian $\mathcal{N}$) or binary-valued (Bernoulli $\mathcal{B}$) data
- The hierarchical specification allows the lower layers of the latent variables to be highly correlated

VAE inference model: bottom-up

$$q_\phi(\boldsymbol{z}|\boldsymbol{x}) = q_\phi(\boldsymbol{z}_1|\boldsymbol{x}) \prod_{i=2}^{L} q_\phi(\boldsymbol{z}_i|\boldsymbol{z}_{i-1}) \tag{13}$$

$$q_\phi(\boldsymbol{z}_1|\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{z}_1|\boldsymbol{\mu}_{q,1}(\boldsymbol{x}), \boldsymbol{\sigma}^2_{q,1}(\boldsymbol{x})\right) \tag{14}$$

$$q_\phi(\boldsymbol{z}_i|\boldsymbol{z}_{i-1}) = \mathcal{N}\left(\boldsymbol{z}_i|\boldsymbol{\mu}_{q,i}(\boldsymbol{z}_{i-1}), \boldsymbol{\sigma}^2_{q,i}(\boldsymbol{z}_{i-1})\right), \; i = 2 \dots L. \tag{15}$$

$$\boldsymbol{d}(\boldsymbol{y}) = \texttt{MLP}(\boldsymbol{y}) \tag{16}$$

$$\mu(\boldsymbol{y}) = \texttt{Linear}(\boldsymbol{d}(\boldsymbol{y})) \tag{17}$$

$$\sigma^2(\boldsymbol{y}) = \texttt{Softplus}(\texttt{Linear}(\boldsymbol{d}(\boldsymbol{y}))) \,, \tag{18}$$

Ladder VAE

$$\boldsymbol{d}_n = \texttt{MLP}(\boldsymbol{d}_{n-1}) \tag{19}$$

$$\hat{\mu}_{q,i} = \texttt{Linear}(\boldsymbol{d}_i), i = 1 \dots L \tag{20}$$

$$\hat{\sigma}^2_{q,i} = \texttt{Softplus}(\texttt{Linear}(\boldsymbol{d}_i)), i = 1 \dots L \tag{21}$$

where $\boldsymbol{d}_0 = \boldsymbol{x}$. Recursive downward pass:

$$q_\phi(\boldsymbol{z}|\boldsymbol{x}) = q_\phi(\boldsymbol{z}_L|\boldsymbol{x}) \prod_{i=1}^{L-1} q_\phi(\boldsymbol{z}_i|\boldsymbol{z}_{i+1}) \tag{22}$$

$$\sigma_{q,i} = \frac{1}{\hat{\sigma}^{-2}_{q,i} + \sigma^{-2}_{p,i}} \tag{23}$$

$$\mu_{q,i} = \frac{\hat{\mu}_{q,i}\hat{\sigma}^{-2}_{q,i} + \mu_{p,i}\sigma^{-2}_{p,i}}{\hat{\sigma}^{-2}_{q,i} + \sigma^{-2}_{p,i}} \tag{24}$$

$$q_\phi(\boldsymbol{z}_i|\cdot) = \mathcal{N}\left(\boldsymbol{z}_i|\boldsymbol{\mu}_{q,i}, \boldsymbol{\sigma}^2_{q,i}\right), \tag{25}$$

where $\mu_{q,L} = \hat{\mu}_{q,L}$ and $\sigma^2_{q,L} = \hat{\sigma}^2_{q,L}$.

Ladder VAE

- precision-weighted combination of
  - $\hat{\mu}_q$ and $\hat{\sigma}_q^2$ carrying bottom-up information and
  - μ_p and σ_p^2 from the generative distribution carrying *top-down* prior information.
- $\hat{\mu}_q$ and $\hat{\sigma}_q^2$ as the approximate gaussian likelihood that is combined with a gaussian prior μ_p and σ_p^2 from the generative distribution.
- Together these form the approximate posterior distribution $q_\theta(\boldsymbol{z}|\boldsymbol{z}, \boldsymbol{x})$ using the same top-down dependency structure both in the inference and generative model.

Warm up from deterministic to VAE

- The variational regularization term causes some of the latent units to become inactive during training
- the approximate posterior for unit k, $q(z_{i,k}|\dots)$ is regularized towards its own prior $p(z_{i,k}|\dots)$, a phenomenon also recognized in the VAE setting
- presumably trapped in a local minima or saddle point at $KL(q_{i,k}|p_{i,k}) \approx 0$, with the optimization algorithm unable to re-activate them.

Warm up from deterministic to VAE

initializing training using the reconstruction error only (corresponding to training a standard deterministic auto-encoder), and then gradually introducing the variational regularization term:

$$\mathcal{L}(\theta, \phi; \boldsymbol{x})_T = -\beta KL(q_\phi(z|x)||p_\theta(\boldsymbol{z})) + E_{q_\phi(z|x)}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\right], \quad (26)$$

where β is increased linearly from 0 to 1 during the first N_t epochs of training.

Experiments

- Datasets: MNIST, Omniglot, NORB
- $L = 5$ with sizes $64, 32, 16, 8, 4$
- MNIST: bernoulli with sigmoid output layer

MNIST

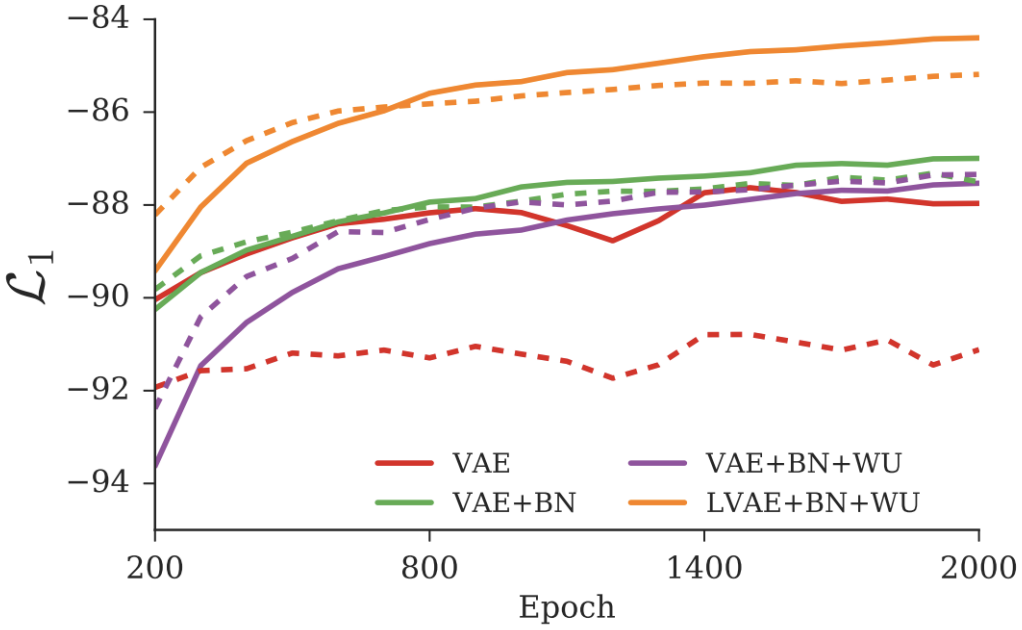


Figure 2: MNIST train (*full lines*) and test (*dashed lines*) set log-likelihood using one importance sample during training. The LVAE improves performance significantly over the regular VAE.

Experiments: MNIST

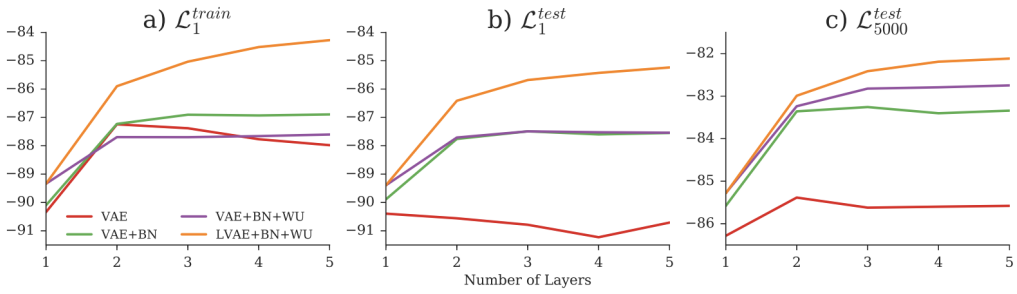
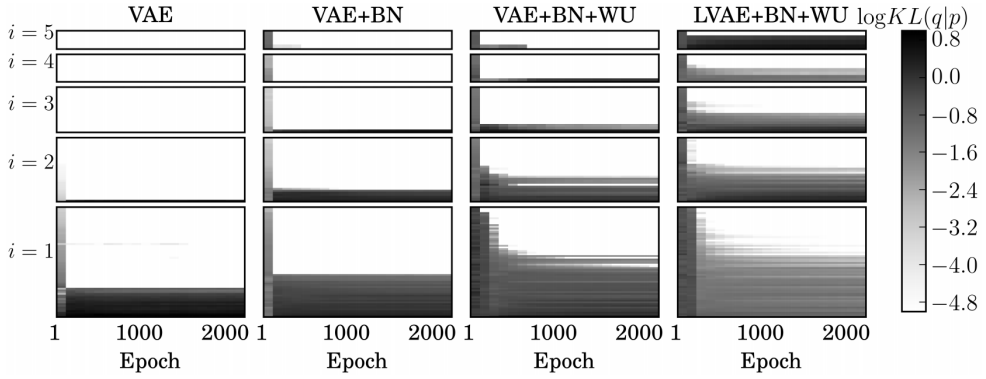


Figure 3: MNIST log-likelihood values for VAEs and the LVAE model with different number of latent layers

Experiments: MNIST



log $KL(q|p)$ for each latent unit is shown at different training epochs. Low KL (white) corresponds to an inactive unit.

Experiments: MNIST

	$\leq \log p((x))$
VAE 1-layer + NF [17]	-85.10
IWAE, 2-layer + IW=1 [2]	-85.33
IWAE, 2-layer + IW=50 [2]	-82.90
VAE, 2-layer + VGP [20]	-81.90
LVAE, 5-layer	-82.12
LVAE, 5-layer + finetuning	-81.84
LVAE, 5-layer + finetuning + IW=10	-81.74

Experiments: MNIST

	VAE	VAE +BN	VAE +BN +WU	LVAE +BN +WU
OMNIGLOT				
64	-111.21	-105.62	-104.51	$-$
64-32	-110.58	-105.51	-102.61	-102.63
64-32-16	-111.26	-106.09	-102.52	-102.18
64-32-16-8	-111.58	-105.66	-102.66	-102.21
64-32-16-8-4	-110.46	-105.45	-102.48	**-102.11**
NORB				
64	2741	3198	3338	$-$
64-32	2792	3224	3483	3272
64-32-16	2786	3235	3492	**3519**
64-32-16-8	2689	3201	3482	3449
64-32-16-8-4	2654	3198	3422	3455

Experiments: MNIST

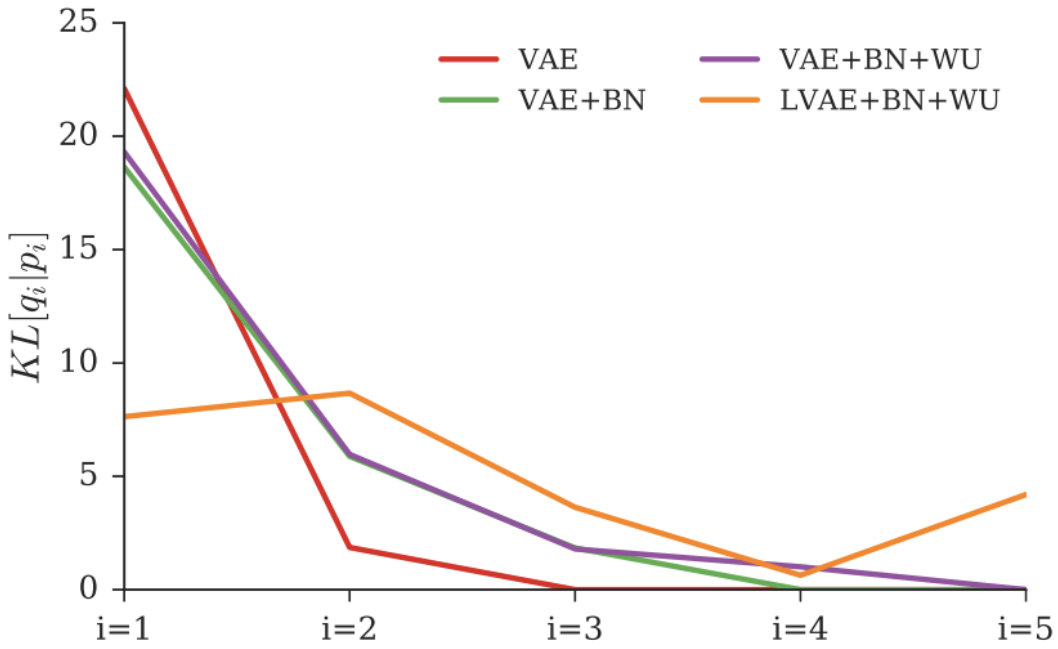


Figure 5: Layer-wise $KL[q|p]$ divergence going from the lowest to the highest layers. In the VAE models the KL divergence is highest in the lowest layers whereas it is more distributed in the LVAE model

Experiments: MNIST

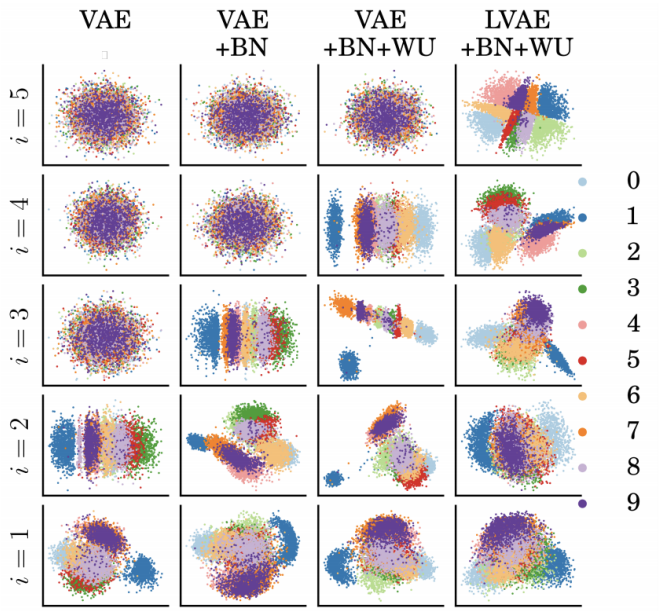


Figure 6: PCA-plots of samples from $q(z_i|z_{i-1})$ for 5-layer VAE and LVAE models trained on MNIST. Color-coded according to true class label

PCA plots of samples from $q(z_i|z_{i-1})$ for 5-layer VAE and LVAE models trained on MNIST

Conclusion

- new inference model for VAEs combining a bottom-up data-dependent approximate likelihood term with a prior information from the generative distribution
- learns a deeper and qualitatively different latent representation of data
- this parameterization makes the optimization easier since the inference is simply correcting the generative distribution instead of fitting the two models separately.

CausalVAE: Disentangled Representation Learning via Neural Structural Causal Models

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

Motivation

- learning disentangled representations
- assumption: the data is indeed generated by countable independent factors
- this paper: the independent factors are causally related

Figure 1: A swinging pendulum: an illustrative example

pendulum angle and light position are the causes of (l, x) of shadow

Advantages of causal disentanglement

- above are not independent, indepndence based disentanglement cant extract these factors
- will still disentangle light and pnedulum?
- generating counterfactual data– do operation
- example: do($shadow = 0$)

Structural Causal model

- a causal model is an ordered triple $\langle \epsilon, X, F \rangle$,
- where ϵ : exogenous variables whose values are determined by factors outside the model;
- X : set of endogenous variables whose values are determined by factors within the model;
- F : structural equations that express the value of each endogenous variable as a function of the values of the other variables in X and ϵ
- $x_i = f_i(x_{pa_i}, \epsilon_i)$

Causal VAE

- Encoder $\boldsymbol{x} \rightarrow \boldsymbol{z}$
- SCM Layer
- Decoder $\boldsymbol{z} \rightarrow \boldsymbol{x}$

Causal VAE: Structural Causal Model

- Step 1: identify the exogeneous factors ϵ
- step 2: A 'Causal Structure Layer' that relates the exogeneous factors
- *the causal structure is learnt, not prespecified*

Method: Causal VAE: Structural Causal Model Layer

- Consider a Linear Structural Causal Model
- disentangled factors correspond to $\boldsymbol{u}$
- and the causal graph among them corresponds to $\boldsymbol{A}$
- $\boldsymbol{z} = \boldsymbol{A}^T \boldsymbol{z} + \boldsymbol{\epsilon} = (I - \boldsymbol{A}^T)^{-1} \boldsymbol{\epsilon}$
- $\boldsymbol{z} \in \boldsymbol{R}^n$ corresponding to n concepts
- $\boldsymbol{\epsilon} = N(0, I)$
- For example, chain graph $\boldsymbol{z}_1 \rightarrow \boldsymbol{z}_2 \rightarrow \boldsymbol{z}_3$
- $$\begin{bmatrix} \boldsymbol{z}_1 \\ \boldsymbol{z}_2 \\ \boldsymbol{z}_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{z}_1 \\ \boldsymbol{z}_2 \\ \boldsymbol{z}_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix}$$

Causal VAE: Step 1: $\boldsymbol{z}$ latent representations

- $$\boldsymbol{z} = \boldsymbol{A}^T \boldsymbol{z} + \boldsymbol{\epsilon} = (I - \boldsymbol{A}^T)^{-1} \boldsymbol{\epsilon} \tag{27}$$

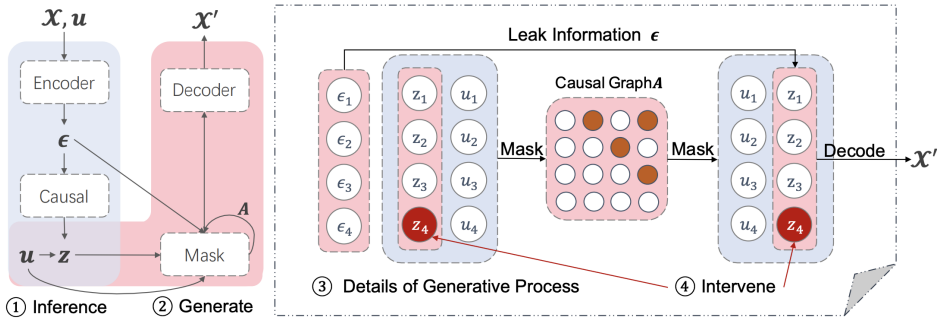
  $$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I}), \tag{28}$$

- ?? Only Markov Equivalence Graph if the variables are continuous, and the equations are linear with Gaussian (normal, or bell-shaped) errors.

Causal VAE: Step 2: Structural Causal Model Layer

- $\boldsymbol{z}_i = g_i(A_i \odot z_i; \eta_i) + \epsilon_i$
- Mask layer that mimics generating children from parents

Causal VAE



Figure

Generative Model

- $\boldsymbol{x} \in \boldsymbol{R}^d$, known concepts $\boldsymbol{u} \in R^n$, $\boldsymbol{\epsilon} \in R^n$
- $\boldsymbol{z} = (I - \boldsymbol{A}^T)^{-1}\boldsymbol{\epsilon} = \boldsymbol{C}\boldsymbol{\epsilon}$
- Conventional VAE: $p_\theta(\boldsymbol{x}, \boldsymbol{z}) = p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$
- Conditional VAE: $p_\theta(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{u}) = p_\theta(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{u})p(\boldsymbol{z}|\boldsymbol{u})$
- Causal VAE: $p_\theta(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\epsilon}|\boldsymbol{u}) = p_\theta(\boldsymbol{x}|\boldsymbol{\epsilon}, \boldsymbol{z}, \boldsymbol{u})p(\boldsymbol{z}, \boldsymbol{\epsilon}|\boldsymbol{u})$

Generative Model

- Decoder: $\mathrm{f}(\mathrm{z})$
- Encoder: $\mathrm{h}(\mathrm{x}, \mathrm{u})$

$$p_{\theta}(\mathrm{x}|\mathrm{z}, \epsilon, \mathrm{u}) = p_{\theta}(\mathrm{x}|\mathrm{z}) \equiv p_{\xi}(\mathrm{x} - \mathrm{f}(\mathrm{z})) \tag{29}$$

$$q_{\phi}(\mathrm{z}, \epsilon|\mathrm{x}, \mathrm{u}) \equiv q(\mathrm{z}|\epsilon) q_{\zeta}(\epsilon - \mathrm{h}(\mathrm{x}, \mathrm{u})), \tag{30}$$

$$\mathrm{x} = \mathrm{f}(\mathrm{z}) + \xi, \quad \epsilon = \mathrm{h}(\mathrm{x}, \mathrm{u}) + \zeta, \tag{31}$$

- 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.

Generative Model

-
- the joint prior $p_{\boldsymbol{\theta}}(\boldsymbol{\epsilon}, \mathrm{z}|\mathrm{u})$ for latent variables z and $\boldsymbol{\epsilon}$ as

$$p_{\boldsymbol{\theta}}(\boldsymbol{\epsilon}, \mathrm{z}|\mathrm{u}) = p_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}) p_{\boldsymbol{\theta}}(\mathrm{z}|\mathrm{u}), \tag{32}$$

- $p_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}) = \mathcal{N}(0, \mathrm{I})$
- the prior of latent endogenous variables $p_{\boldsymbol{\theta}}(\mathrm{z}|\mathrm{u})$ is a factorized Gaussian distribution conditioning on the additional observation u, i.e.

$$p_{\boldsymbol{\theta}}(\mathrm{z}|\mathrm{u}) = \mathcal{N}(\lambda_1(u_i), \lambda_2^2(u_i)), \tag{33}$$

- λ_1 and λ_2 are an arbitrary functions. let $\lambda_1(\mathrm{u}) = \mathrm{u}$ and $\lambda_2(\boldsymbol{u}) \equiv 1$.

Training the Generative Model

$$\mathbb{E}_{q_{\mathcal{X}}}[\log p_{\boldsymbol{\theta}}(\mathsf{x}|\mathsf{u})] \geq \mathsf{ELBO} = \mathbb{E}_{q_{\mathcal{X}}}[\mathbb{E}_{\boldsymbol{\epsilon},\mathsf{z}\sim q_{\boldsymbol{\phi}}}[\log p_{\boldsymbol{\theta}}(\mathsf{x}|\mathsf{z},\boldsymbol{\epsilon},\mathsf{u})] - \mathcal{D}(q_{\boldsymbol{\phi}}(\boldsymbol{\epsilon},\mathsf{z}|\mathsf{x},\mathsf{u})||p_{\boldsymbol{\theta}}(\boldsymbol{\epsilon},\mathsf{z}|\mathsf{u}))], \tag{34}$$

$$q_{\boldsymbol{\phi}}(\boldsymbol{\epsilon},\mathsf{z}|\mathsf{x},\mathsf{u}) = q_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}|\mathsf{x},\mathsf{u})\delta(\boldsymbol{z} = \mathsf{C}\boldsymbol{\epsilon}) = q_{\boldsymbol{\phi}}(\mathsf{z}|\mathsf{x},\mathsf{u})\delta(\boldsymbol{\epsilon} = \mathsf{C}^{-1}\boldsymbol{z}), \tag{35}$$

$$\mathsf{ELBO} = \mathbb{E}_{q_{\mathcal{X}}}[\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u})}[\log p_{\boldsymbol{\theta}}(\mathsf{x}|\mathsf{z})] - \tag{36}$$

$$\mathcal{D}(q_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}|\mathsf{x},\mathsf{u})||p_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})) - \mathcal{D}(q_{\boldsymbol{\phi}}(\mathsf{z}|\mathsf{x},\mathsf{u})||p_{\boldsymbol{\theta}}(\mathsf{z}|\mathsf{u}))]. \tag{37}$$

Additional Constraints

$$I_u = \mathbb{E}_{q_{\mathcal{X}}} \|\mathsf{u} - \sigma(\mathsf{A}^T \mathsf{u})\|_2^2 \leq \kappa_1, \tag{38}$$

$$I_m = \mathbb{E}_{\mathsf{z} \sim q_\phi} \sum_{i=1}^{n} \|z_i - g_i(\mathsf{A}_i \circ \mathsf{z}; \boldsymbol{\eta}_i)\|^2 \leq \kappa_2, \tag{39}$$

DAG constraint:

$$H(\mathsf{A}) \equiv tr((\mathsf{I} + \mathsf{A} \circ \mathsf{A})^n) - n = 0. \tag{40}$$

Training Loss Function:

$$\mathcal{L} = -\text{ELBO} + \alpha H(\mathsf{A}) + \beta I_u + \gamma I_m, \tag{41}$$

Experiments

- Two datasets: Synthetic (Pendulum) and CelebA
- Pendulum:
  - 3 entities (pendulum, light, shadow)
  - 4 concepts ((pendulum angle, light angle) $\rightarrow$ (shadow location, shadow length)).
- CelebA
  - 4 causally related concepts (gender, smile, eyes open,mouth open), where gender and smile cause eyes open, and smile causes mouth open.
- Evauation Criteria:
  - MIC: Maximal Information Criterion (MIC)
  - Total Information Criterion (TIC)

Results

	CausalVAE		DC-IGN		β-VAE		CausalVAE-unsup		LadderVAE	
Metrics(%)	MIC	TIC	MIC	TIC	MIC	TIC	MIC	TIC	MIC	TIC
Pendulum	**96.3**$\pm$3.6	**89.0**$\pm$2.9	61.8$\pm$8.7	48.1$\pm$7.3	22.6$\pm$4.6	12.5$\pm$2.2	21.2$\pm$1.4	12.0$\pm$1.0	22.4$\pm$3.1	12.8$\pm$1.2
CelebA	**83.7**$\pm$6.2	**71.6**$\pm$7.2	78.8$\pm$10.9	66.1$\pm$12.1	22.5$\pm$1.2	9.92$\pm$1.2	27.2$\pm$5.3	14.6$\pm$4.2	23.5$\pm$3.0	10.3$\pm$1.6

Intervention Experiments

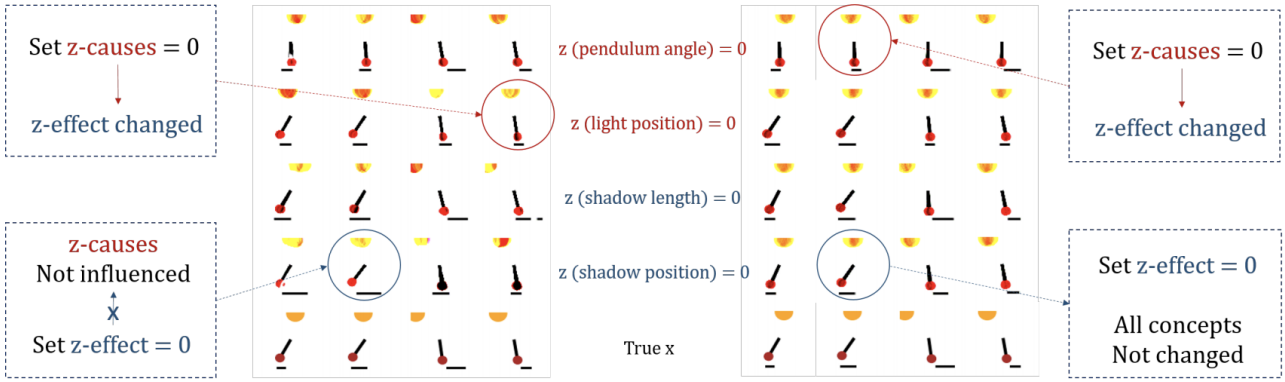


Figure 3: The results of Intervention experiments on the pendulum dataset. Each row shows the result of controlling the PENDULUM ANGLE, LIGHT ANGLE, SHADOW LENGTH, and SHADOW LOCATION respectively. The bottom row is the original input image.

Intervention Experiments

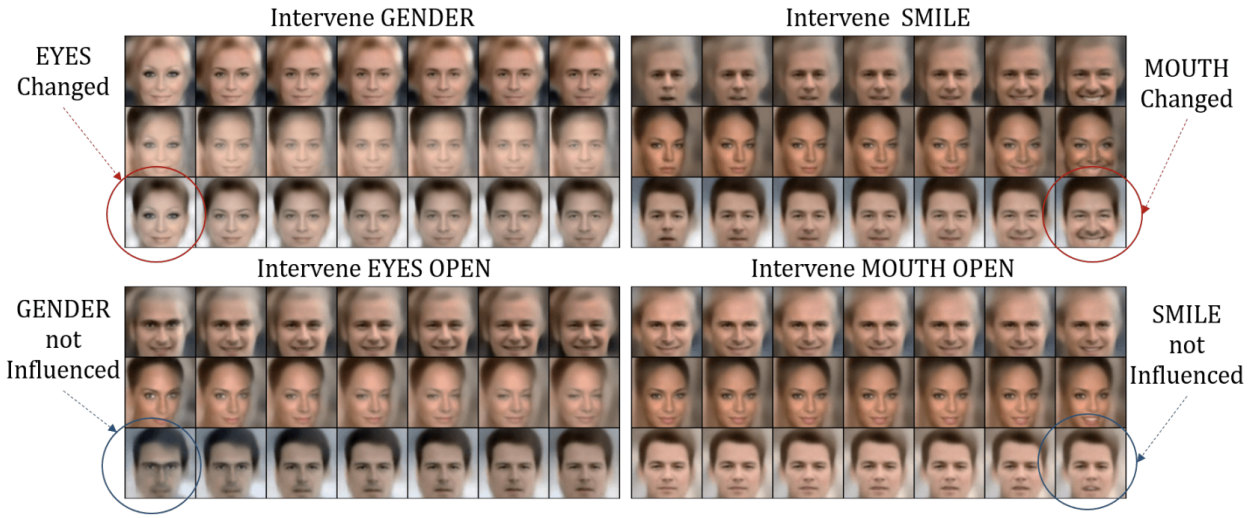


Figure 4: Results of CausalVAE model on CelebA. The controlled factors are GENDER, SMILE, EYES OPEN and MOUTH OPEN respectively.

Conclusion

- learning disentangled representations of causally related concepts in data
- allows intervention to generate counterfactual outputs as expected according to our understanding of the causal system.