Variational Autoencoders

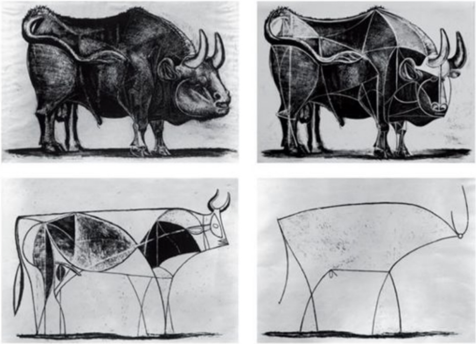
Presented by: Achint Kumar

Generative AI Reading Club

May 11, 2023

Desiderata

VAE: A framework for Latent Variable Modelling



The Bull by Picasso (1945)

- Six strokes capture the essence of the bull (latent variables)
- Six strokes is scaffolding for any bull (generative modelling)

Why should we care?

- Latent variable models: Working in latent space is simpler than working in data space.

  "Make things as simple as possible, but not simpler"

  Albert Einstein

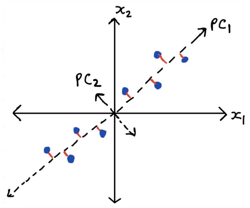
- Generative models: We can generate new data. Abstractly, it lets us represent and sample from high dimensional data distribution, p(x) efficiently.

  "What I cannot create, I do not understand"

  Richard Feynman

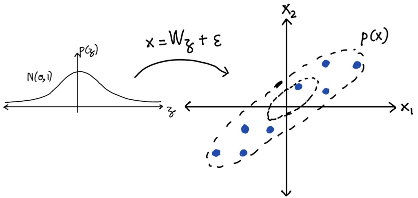
Latent Variable Modelling: Historical Perspective

Principal Component Analysis (Pearson, 1901. Hotelling, 1933)



- Maximize variance in projected space.
- Projected space has less noise, no redundancy
- Linear, not generative, needs covariance matrix diagonalization

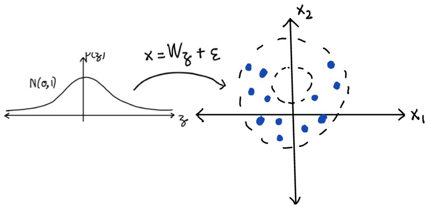
Probabilistic PCA (Tipping & Bishop, 1999)



- Maximize $\log p(x)$ wrt W
- Both latent variable and generative model
- Linear, needs covariance matrix diagonalization

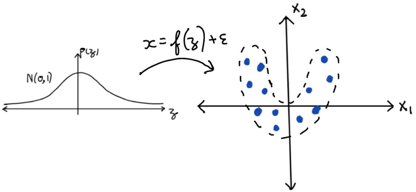
pPCA to Variational Autoencoders

Probabilistic PCA (Tipping & Bishop, 1999)



- Maximize $\log p(x)$
- Both latent variable and generative model
- Linear, needs covariance matrix diagonalization

Variational Autoencoders (Kingma & Welling, 2013)



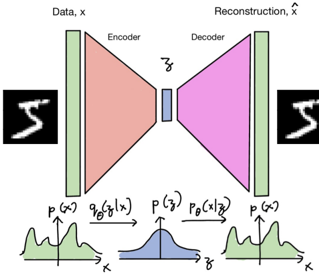
- Minimize variational free energy
- Non-linear, scalable and works for high dimensional data
- Fuzzy generation, non-interpretable latent space

Variational Autoencoders: Network Architecture

State-of-the-art framework for latent variable modelling. It consists of 2 neural networks:

1. Encoder: Data, $\mathbf{x}$ are input and its latent representation, $\mathbf{z}$ is output.
2. Decoder: Latent representation, $\mathbf{z}$ is input and reconstruction of data, $\hat{\mathbf{x}}$ are output.

To create probabilistic framework, we add noise to latent space.



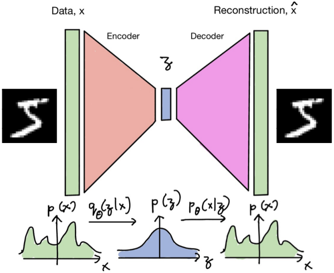
Variational Autoencoders: Training Objective

Want: $\max_{\theta} \log p_\theta(x) = \max_{\theta} \log \int p_\theta(x|z)p(z)\,dz$.

For Vanilla VAE, we assume

$$p(z) \sim \mathcal{N}(0, \mathbb{I})$$
$$p(x|z) \sim \mathcal{N}(f(z), \mathbb{I})$$

Problem: Calculating $p_\theta(x)$ leads to an intractable integral!
Solution: Exploit connection to statistical physics.



Training Objective: Statistical Physics to rescue

In Statistical Physics, partition function (ML: $p_\theta(x)$) is related to free energy.

$$-\log[p_\theta(x)] \doteq \underbrace{\mathcal{F}(x)}_{\text{Free Energy}} = \underbrace{\langle E(x) \rangle}_{\text{Energy}} - T \underbrace{\langle S(x) \rangle}_{\text{Entropy}}$$

where

$$\langle E(x) \rangle = -\mathbb{E}_{p(z|x)}[\log(p(x,z))]$$
$$\langle S(x) \rangle = -\mathbb{E}_{p(z|x)}[\log(p(z|x))]$$

This relation gives another way of calculating, $p_\theta(x)$.
Problem: We don't know the posterior $p(z|x)$.
Solution: Variational Inference

Training Objective: Statistical Physics to rescue

In Statistical Physics, partition function (ML: $p_\theta(x)$) is related to free energy.

$$-\log[p_\theta(x)] \doteq \underbrace{\mathcal{F}(x)}_{\text{Free Energy}} = \underbrace{\langle E(x)\rangle}_{\text{Energy}} - T\underbrace{\langle S(x)\rangle}_{\text{Entropy}}$$

where

$$\langle E(x)\rangle = -\mathbb{E}_{p(z|x)}[\log(p(x,z))]$$
$$\langle S(x)\rangle = -\mathbb{E}_{p(z|x)}[\log(p(z|x))]$$

This relation gives another way of calculating, $p_\theta(x)$.
Problem: We don't know the posterior $p(z|x)$.
Solution: Variational Inference

Training Objective: Variational Inference

Approximate true posterior, $p_\phi(z|x)$ by Gaussian distribution, $q_\phi(z|x) \sim \mathcal{N}(\mu(x), \sigma^2(x))$.

Variational free energy upper bounds true free energy,

$$\underbrace{\mathcal{F}_{\text{true}}(x)}_{\text{Free Energy}} \leq \mathcal{F}_{\text{var}}(x) = \underbrace{E_{\text{var}}(x)}_{\text{Energy}} - T\underbrace{S_{\text{var}}(x)}_{\text{Entropy}}$$

After making substitutions, variational free energy becomes

$$\mathcal{F}_{\text{var}}(x) \propto \underbrace{||x - \hat{x}||_2^2}_{\text{reconstruction}} + \underbrace{D_{KL}(q_\phi(z|x)||p(z))}_{\text{regularizer}}.$$

VAEs minimize $\mathcal{F}_{\text{var}}(x)$.

Training Objective: Variational Inference

Approximate true posterior, $p_\phi(z|x)$ by Gaussian distribution, $q_\phi(z|x) \sim \mathcal{N}(\mu(x), \sigma^2(x))$.

Variational free energy upper bounds true free energy,

$$\underbrace{\mathcal{F}_{\text{true}}(x)}_{\text{Free Energy}} \leq \mathcal{F}_{\text{var}}(x) = \underbrace{E_{\text{var}}(x)}_{\text{Energy}} - T\underbrace{S_{\text{var}}(x)}_{\text{Entropy}}$$

After making substitutions, variational free energy becomes

$$\mathcal{F}_{\text{var}}(x) \propto \underbrace{||x - \hat{x}||_2^2}_{\text{reconstruction}} + \underbrace{D_{KL}(q_\phi(z|x)||p(z))}_{\text{regularizer}}.$$

VAEs minimize $\mathcal{F}_{\text{var}}(x)$.

Story so far: Executive Summary

How to infer underlying latent variables of high-dimensional data?

Variational Autoencoders (VAEs) provide a framework for generative and latent variable modelling.
Training VAEs involves 3 steps:

1. Assume a form for prior distribution, $p(z)$ and variational posterior distribution, $q_\phi(z|x)$
2. Calculate variational free energy, $\mathcal{F}_{\mathrm{var}}(x) = \text{reconstruction} + \text{regularizer}$
3. Minimize $\mathcal{F}_{\mathrm{var}}(x)$ by backpropagation

β-VAE (Irina Higgins, *et. al.*, 2017)

$$\mathcal{F}_{\text{var}}(x) \propto \underbrace{||x - \hat{x}||_2^2}_{\text{reconstruction}} + \beta \underbrace{D_{KL}(q_\phi(z|x)||p(z))}_{\text{regularizer}}, \beta > 1$$

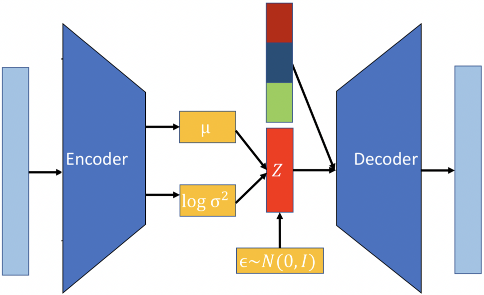
β-VAE produces disentangled representation

For β-VAE($\beta = 250$), for VAE($\beta = 1$). Single latent feature traversed from $[-3, 3]$ while others are kept fixed

CVAE (Kihyuk Sohn, *et. al*, 2015)

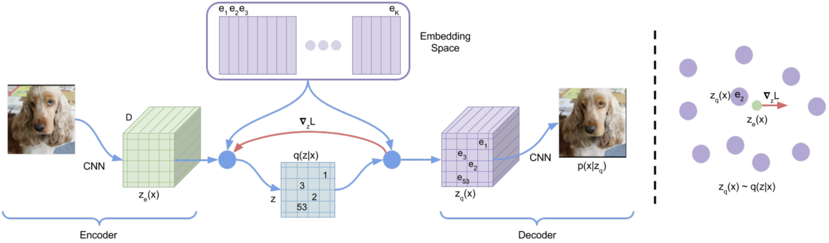
$[\text{Latent}] \longrightarrow [\text{Latent}, \text{Condition}]$ where Condition could be:

- Class label (eg: one-hot encoding of MNIST label)
- F(X) (learning all solutions of many-to-one functions).



VQ-VAE (Aaron van den Oord, *et. al.*, 2017)

Provides discrete latent representation

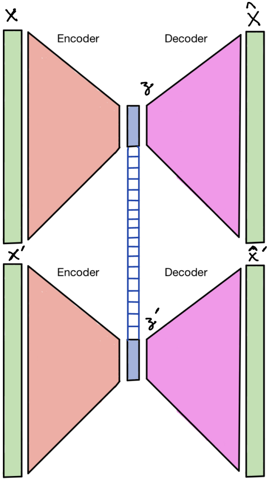


$$\mathcal{F}_{\text{var}}(x) \propto \underbrace{||x - \hat{x}||_2^2}_{\text{reconstruction}} + \underbrace{||sg[z_e(x)] - e||_2^2}_{\text{embedding loss}} + \beta \underbrace{||z_e(x) - sg[e]||_2^2}_{\text{commitment loss}}$$

Since, prior is uniform distribution and posterior is one-hot categorical distribution, KL term is a constant and can be ignored.

Multi-modal VAE

1. Multi-modal VAE is a variant of vanilla VAE in which multiple datasets can be jointly input to the network.
2. It can model nonlinear correlations between modalities. It is much more expressive than CCA.



Constructing multi-modal VAE

For any multi-modal VAE variational free energy is given by,

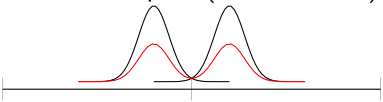
$$\mathcal{F}_{var}(x, x') = \underbrace{||x - \hat{x}||_2^2 + ||x' - \hat{x'}||_2^2}_{reconstruction} + \underbrace{D_{KL}(q_\phi(z, z'|x, x')||p(z, z'))}_{regularizer}$$

We need to assume a form for prior, $p(z, z')$ and posterior, $q_\phi(z, z'|x, x')$.

Formulating Posterior Distribution

There are two ways to define the multimodal posterior, $q_\phi(z_1, z_2|x_1, x_2)$ in terms of unimodal posterior, $q_\phi(z_1, z_2|x_1)$ and $q_\phi(z_1, z_2|x_2)$:

Mixture of Experts (MMVAE, 2018)



Black: Unimodal posterior
Red: Multimodal posterior

- Posterior is written as sum:

$$q_\phi(z_1, z_2|x_1, x_2) = \sum_i \alpha_i q_{\phi_i}(z_1, z_2|x_i)$$

- Like a healthy relationship (each expert can make decision). Does what either likes.

Product of Experts (MVAE, 2017)

Black: Unimodal posterior
Red: Multimodal posterior

- Posterior is written as product:

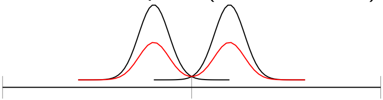
$$q_\phi(z_1, z_2|x_1, x_2) = p(z_1, z_2) \prod_{i=1}^{2} q_{\phi_i}(z_1, z_2|x_i)$$

- Like UN Security Council (each expert has veto power). Does what neither likes.

Formulating Posterior Distribution

There are two ways to define the multimodal posterior, $q_\phi(z_1, z_2|x_1, x_2)$ in terms of unimodal posterior, $q_\phi(z_1, z_2|x_1)$ and $q_\phi(z_1, z_2|x_2)$:

Mixture of Experts (MMVAE, 2018)



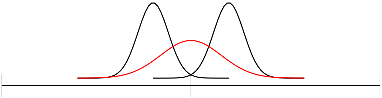
Black: Unimodal posterior
Red: Multimodal posterior

- Posterior is written as sum:

$$q_\phi(z_1, z_2|x_1, x_2) = \sum_i \alpha_i q_{\phi_i}(z_1, z_2|x_i)$$

- Like a healthy relationship (each expert can make decision). Does what either likes.

Product of Experts (MVAE, 2017)



Black: Unimodal posterior
Red: Multimodal posterior

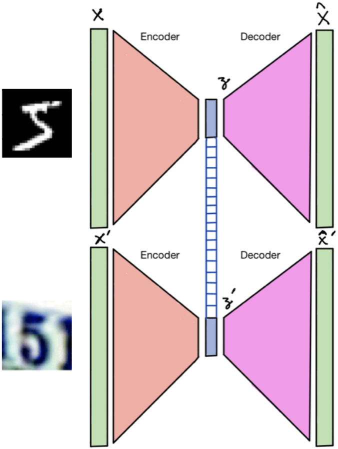
- Posterior is written as product:

$$q_\phi(z_1, z_2|x_1, x_2) = p(z_1, z_2) \prod_{i=1}^{2} q_{\phi_i}(z_1, z_2|x_i)$$

- Like UN Security Council (each expert has veto power). Does what neither likes.

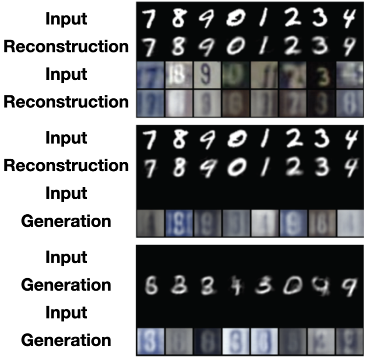
MNIST-SVHN Experiment

1. Input to POISE-VAE:
    - Modality 1: MNIST
    - Modality 2: SVHN
2. Methodology: Same digit class from the two modality is input to the VAE
3. Motivation: Can multimodal VAE learn to generate consistent digits in absence of one or both modalities?



MNIST-SVHN Results

- Top: Sample reconstructions.
- Center: Cross generation (one modality absent).
- Bottom: Joint generation (both modalities absent).



Strength of VAEs

- Generative Model: Allows us to generated new data
- Density Model: Finds approximate likelihood
- Latent Variable model: Forms compressed representation

Weakness of VAEs

- Weak Generative Model: Generated data is blurry
- Weak Density Model: Assuming prior and posterior is gaussian is too limiting
- Weak Latent Variables: Disentanglement works for only toy datasets

Basically, VAEs are a jack of all trades but master of none.

Research Frontier

- Combining VAE with other unsupervised deep learning models for better encoders and decoders
- Hierarchical latent representations
- More expressive priors and posteriors
- VAEs for non-image data: Using VAEs to generate molecular graph, video compression