Tutorial on Variational Autoencoders

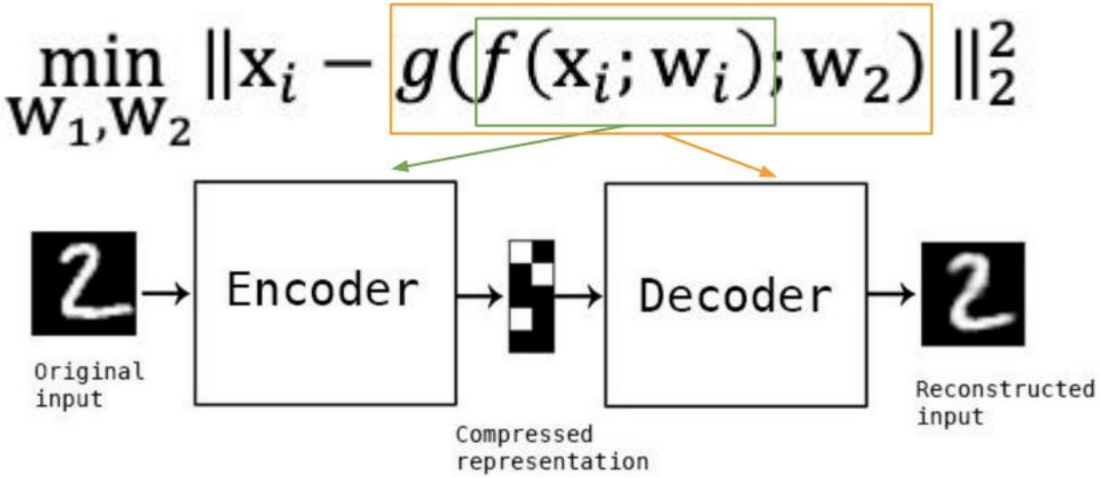
Group 2: Zhou Rong, Chen Siqi, Jiang Yutong, Tang Chenyu

Content Outline

- Introduction
- Generative Model Overview
- Variational Autoencoder
- Conditional Variational Autoencoders
- Examples

Introduction

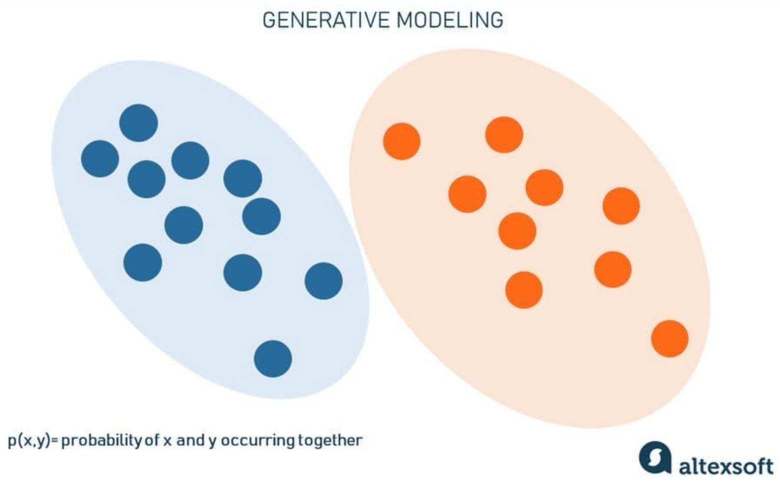
- Generative modeling is widely applied in machine learning area, and it deals with the distribution of P(x).
- People wants to get examples X distributed according to some unknown distribution Pg(x) and the goal is to learn the model P and make is similar with Pg(x)
- One of the popular framework is variational autoencoder.



Generative Model, Overview

Task: Generate new samples follows the same probabilistic distribution of a given a training dataset. eg. Given a dataset of images {X1,X2...} can we learn the distribution of X?

- Learn joint distribution P(X, Y) = P(X|Y=y)P(Y=y)



Preliminaries: Latent Variable Models

- To effectively generate coherent data (like digits), models utilize a decision-making process through latent variables. A latent variable is an unseen (latent) decision that guides the generation process.
- The goal is to optimize the parameter vector θ to ensure that when sampling z from P(z), the function f(z; θ) produces outputs that closely resemble the real data points in the dataset, with a high probability.

(a) (b) (c)

Variational Autoencoders

- Problem Setting

The equation that we are trying to solve: $P(X) = \int P(X|z;\theta)P(z)dz.$

Two questions:

- how to define the latent variables z (i.e., decide what information they represent)?
- how to deal with the integral over z?

$$\mathcal{D}\left[Q(z)\|P(z|X)\right] = E_{z\sim Q}\left[\log Q(z) - \log P(z|X)\right].$$

Variational Autoencoders

2.1 Setting up the objective

1. give a computable formula of P(x)
2. calculate the gradient
3. use SDG to optimize

- If we can find a computable formula for P(X), and take the gradient of that formula, then we can optimize the model using SDG.

$$P(X) = \int P(X|z;\theta)P(z)dz.$$

- We are interested in z that are likely to produce X, which is P(z|X).

- However, P(z|X) is intractable, so we approximate it with Q(z|X).

Variational Autoencoders

2.1 Setting up the objective

$$\mathcal{D}\left[Q(z)\|P(z|X)\right] = E_{z\sim Q}\left[\log Q(z) - \log P(z|X)\right]. \quad (2)$$

for some arbitrary Q

$$\begin{aligned} D[Q(z)||P(z|X)] &= E_{z\sim Q}[\log Q(z) - \log P(z|X)] \\ &= E_{z\sim Q}\left[\log Q(z) - \log \frac{P(X|z)P(z)}{P(X)}\right] \\ &= E_{z\sim Q}[\log Q(z) - \log P(X|z) - \log P(z) + \log P(X)] \end{aligned}$$

$$\mathcal{D}\left[Q(z)\|P(z|X)\right] = E_{z\sim Q}\left[\log Q(z) - \log P(X|z) - \log P(z)\right] + \log P(X). \quad (3)$$

$$D[Q(z)||P(z)] = E_{z\sim Q}[\log Q(z) - \log P(z)]$$

$$\log P(X) - \mathcal{D}\left[Q(z)\|P(z|X)\right] = E_{z\sim Q}\left[\log P(X|z)\right] - \mathcal{D}\left[Q(z)\|P(z)\right]. \quad (4)$$

Variational Autoencoders

2.1 Setting up the objective

$$\log P(X) - \mathcal{D}\left[\boxed{Q(z)}\|P(z|X)\right] = E_{z\sim Q}\left[\log P(X|z)\right] - \mathcal{D}\left[\boxed{Q(z)}\|P(z)\right]. \quad (4)$$

Although Q(z) can be any distribution, we want to focus on Q(z|X)!

$$\log P(X) - \mathcal{D}\left[\boxed{Q(z|X)}\|P(z|X)\right] = E_{z\sim Q}\left[\log P(X|z)\right] - \mathcal{D}\left[\boxed{Q(z|X)}\|P(z)\right]. \quad (5)$$

$$\log P(X) - \mathcal{D}[Q(z \mid X)\|P(z \mid X)] = E_{z\sim Q}[\log P(X \mid z)] - \mathcal{D}[Q(z \mid X)\|P(z)]$$

Objective to maximize

Error term to minimize

ELBO: Evidence Lower Bound

$$\max \log P(X) \Leftrightarrow \min D[Q(z|X)\|P(z|X)] \Leftrightarrow \max \text{ELBO}$$

$$\max \text{ELBO} = \max E_{z\sim Q}[\log P(X \mid z)] - \mathcal{D}[Q(z \mid X)\|P(z)]$$

Variational Autoencoders

2.2 Optimizing the objective

- Once we can calculate the gradient, we can use SGD to optimize the parameters

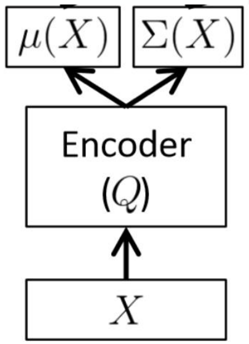
$$\log P(X) - \mathcal{D}[Q(z \mid X) \| P(z \mid X)] = E_{z \sim Q}[\log P(X \mid z)] - \mathcal{D}[Q(z \mid X) \| P(z)] \quad (5)$$

- The equation we want to optimize is ELBO:

$$\max \ \log P(X)$$

$$\Leftrightarrow \max \ \text{ELBO} = \max \ E_{z \sim Q}[\log P(X \mid z)] - \mathcal{D}[Q(z \mid X) \| P(z)]$$

$$P(z) \sim N(0, I), \quad Q(z|X) \sim N(\mu(X), \Sigma(X))$$



10

Variational Autoencoders

2.2 Optimizing the objective

$$\max \log P(X) \Leftrightarrow \max \text{ELBO} = \max \boxed{E_{z\sim Q}[\log P(X \mid z)]} - \boxed{\mathcal{D}[Q(z \mid X)||P(z)]}$$

① ②

- We average the gradient over arbitrarily many samples of X:

$$E_{X\sim D}[\log P(X) - \mathcal{D}[Q(z|X)\|P(z|X)]] = E_{X\sim D}[E_{z\sim Q}[\log P(X|z)] - \mathcal{D}[Q(z|X)\|P(z)]]. \tag{8}$$

$$\begin{aligned} &\nabla \mathrm{E}_{X\sim D}[\mathrm{E}_{z\sim Q}[\log P(X \mid z)] - \mathcal{D}[Q(z \mid X)||P(z)]] \\ &= \mathrm{E}_{X\sim D}[\nabla \mathrm{E}_{z\sim Q}[\log P(X \mid z)] - \nabla\mathcal{D}[Q(z \mid X)||P(z)]] \\ &= \mathrm{E}_{X\sim D}[\mathrm{E}_{z\sim Q}[\nabla \log P(X \mid z)] - \nabla\mathcal{D}[Q(z \mid X)||P(z)]]. \end{aligned}$$

- each time we only need to sample a single value of X and a single value of z from the distribution Q(z|X), and compute the gradient of:

$$\log P(X|z) - \mathcal{D}[Q(z|X)\|P(z)]. \tag{9}$$

Variational Autoencoders

2.2 Optimizing the objective

$$\log P(X|z) - \mathcal{D}\left[Q(z|X) \| P(z)\right]. \quad (9)$$

$$\max \ \log P(X) \Leftrightarrow \max \ \text{ELBO} = \max \ \underbrace{\boxed{E_{z \sim Q}[\log P(X \mid z)]}}_{①} - \underbrace{\boxed{\mathcal{D}[Q(z \mid X) \| P(z)]}}_{②}$$

- ② is a KL-divergence between two multivariate Gaussian distributions, which can be easily computed.

$$P(z) \sim N(0, I), \quad Q(z|X) \sim N(\mu(X), \Sigma(X))$$

$$\mathcal{D}[\mathcal{N}(\mu(X), \Sigma(X)) \| \mathcal{N}(0, I)] = \frac{1}{2}\left(\operatorname{tr}(\Sigma(X)) + (\mu(X))^\top (\mu(X)) - k - \log\det(\Sigma(X))\right). \quad (7)$$

Variational Autoencoders

2.2 Optimizing the objective

Analytical compute this

$$\text{ELBO} = \boxed{E_{z\sim Q}[\log P(X \mid z)]} - \boxed{\mathcal{D}[Q(z \mid X)||P(z)]}$$

① ②

$$\log P(X|z) - \mathcal{D}\left[Q(z|X)\|P(z)\right]. \qquad (9)$$

- But ① is a bit more tricky.

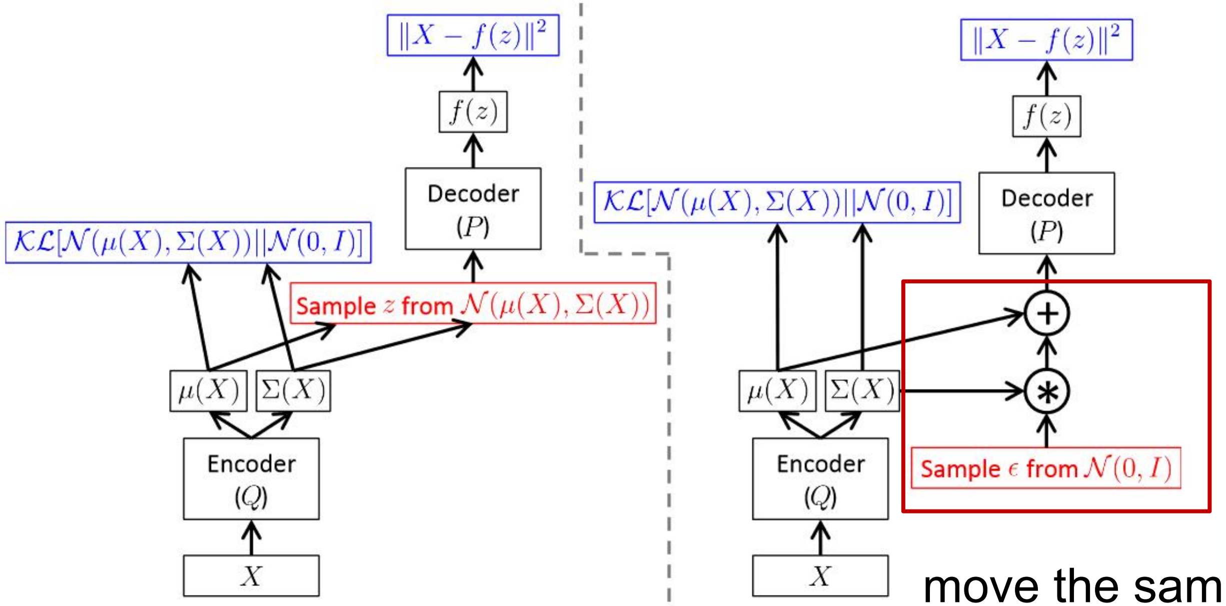
1. $E_{z\sim Q}\left[\log P(X|z)\right]$ depends not just on the parameters of P, but also on the parameters of Q. However, in $\nabla \log P(X \mid z)$ this dependency has disappeared!

2. we need to back-propagate through a layer that samples z from Q(z|X), which is a non-continuous operation and has no gradient.

Variational Autoencoders

2.2 Optimizing the objective

$$E_{X\sim D}\left[E_{\epsilon\sim\mathcal{N}(0,I)}[\log P(X|\boxed{z=\mu(X)+\Sigma^{1/2}(X)*\epsilon)}] - \mathcal{D}\left[Q(z|X)\|P(z)\right]\right]. \tag{10}$$



reparameterization trick

$$z=\mu(X)+\Sigma^{1/2}(X)*\epsilon.$$
$$\epsilon\sim\mathcal{N}(0,I).$$

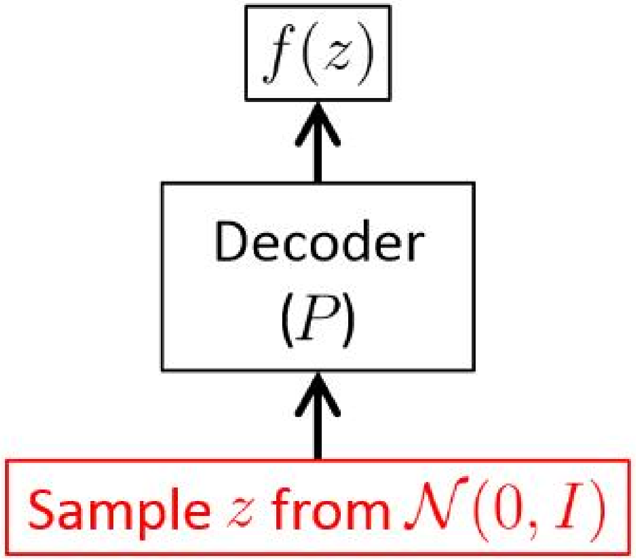
move the sampling to an input layer

Without reparameterization trick

Variational Autoencoders

2.3 Testing the learned model

- During the testing, VAE samples z from the prior distribution $z \sim \mathcal{N}(0, I)$
- Then generates new samples only through the decoder, requiring only random sampling and forward propagation.



Variational Autoencoders

2.4 Interpreting the objective

$$\log P(X) - \mathcal{D}[Q(z \mid X) \| P(z \mid X)] = E_{z \sim Q}[\log P(X \mid z)] - \mathcal{D}[Q(z \mid X) \| P(z)]$$

prior object

current interest

- $P(X)$ converges (in distribution) to the true distribution if and only if $D[Q(z|X) \| P(z|X)]$ goes to zero.
- Given **sufficiently high capacity neural networks**, there are many f functions that result in our model generating any given output distribution. Hence, all we need is one function f which both maximizes $\log P(X)$ and **results in $P(z|X)$ being Gaussian for all X**.

Variational Autoencoders

2.4 Interpreting the objective

$$-\log P(X) + \mathcal{D}[Q(z \mid X)||P(z \mid X)] = -E_{z\sim Q}[\log P(X \mid z)] + \mathcal{D}[Q(z \mid X)||P(z)]$$

- $\boldsymbol{-\log P(X)}$ can be seen as the total number of bits required to construct a given X under our model using an ideal encoding.
- $\boldsymbol{D[Q(z|X)||P(z)]}$ is the expected information that′ s required to convert an uninformative sample from $P(z)$ into a sample from $Q(z|X)$.
- $\boldsymbol{P(X|z)}$ measures the amount of information required to reconstruct X from z under an ideal encoding.
- $\boldsymbol{D[Q(z|X)||P(z|X)]}$ is the penalty we pay for Q being a sub-optimal encoding.

Variational Autoencoders

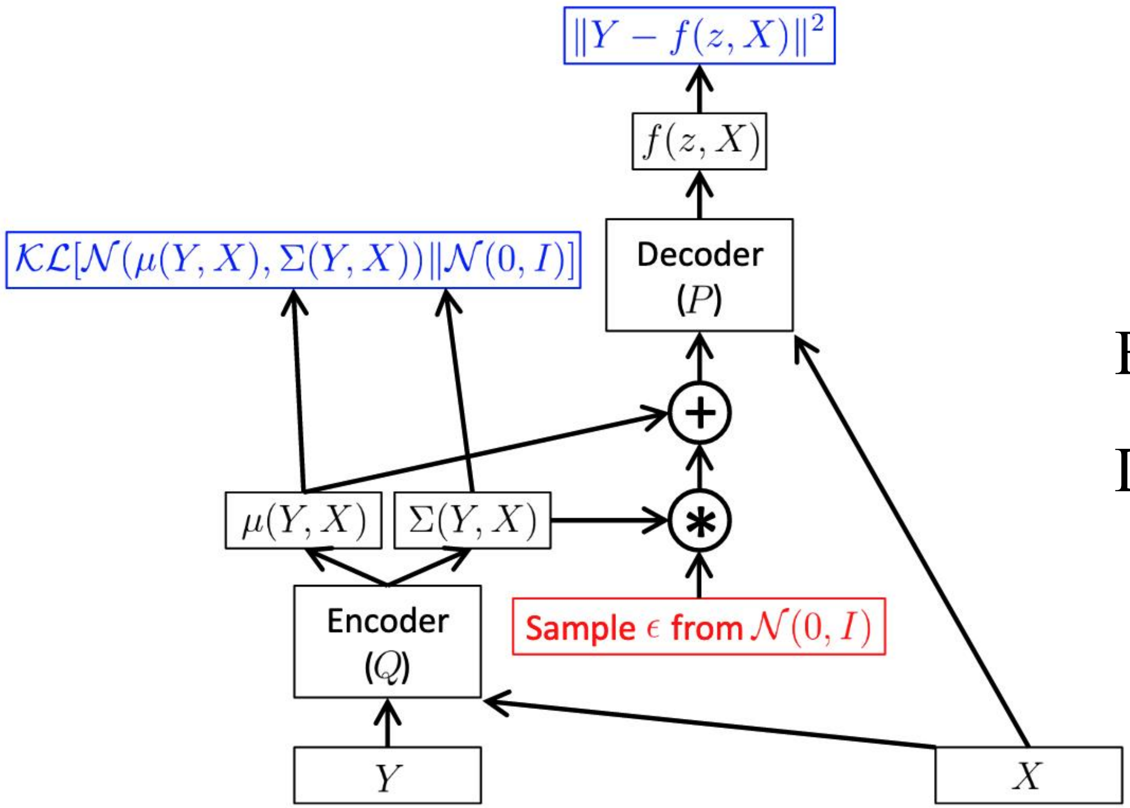
2.4 Interpreting the objective

$$\log P(X) - \mathcal{D}[Q(z \mid X) \| P(z \mid X)] = E_{z \sim Q}[\log P(X \mid z)] - \boxed{\mathcal{D}[Q(z \mid X) \| P(z)]}$$

“regularization” term

standard minimizer: $\|\phi(\psi(X)) - X\|^2 + \boxed{\lambda \|\psi(X)\|_0}$

Conditional Variational Autoencoders



Encoder: $Q(z|Y, X)$

Decoder: $P(Y|z, X)$

How CONDITION works in VAE

Conditional Variational Autoencoders

Optimization objective

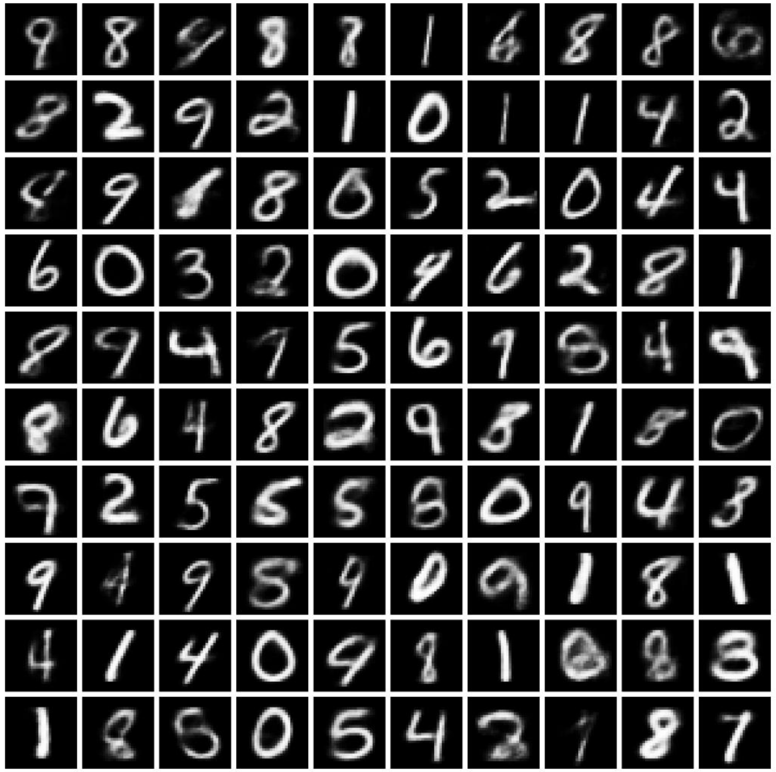
Maximize conditional ELOB:

$E_{z \sim Q(\cdot|Y,X)}[\log P(Y|z, X)] - D(Q(z|Y, X) \| P(z|X)),$

where $P(z|X) \sim N(0, I)$ still holds because z is sampled independently of X

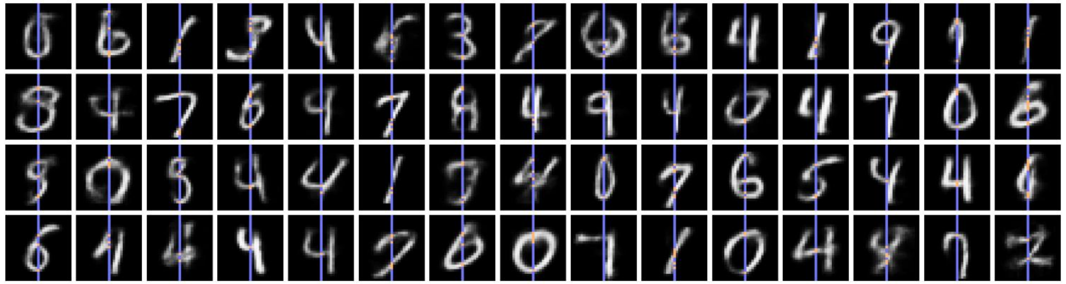
$\Rightarrow Loss = \|Y - f(z, X)\|^2 + aD(N(\mu(Y, X), \Sigma(Y, X)) \| N(0, I))$

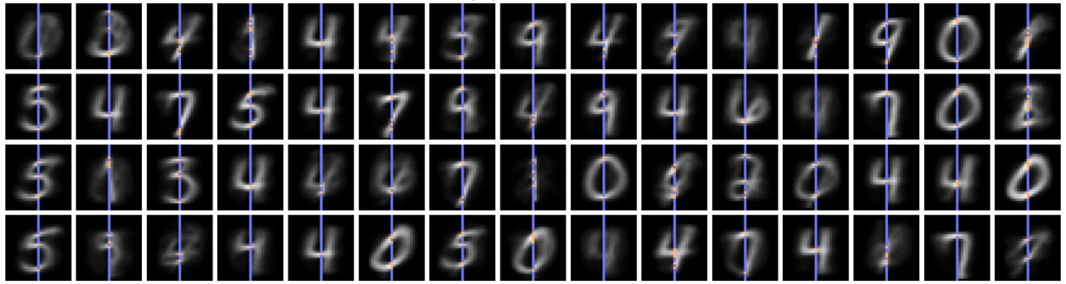
Examples – MINST VAE

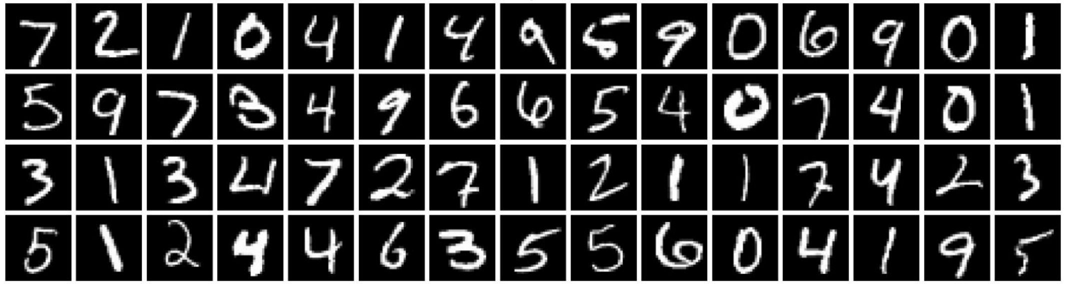


MINST AutoEncoder from Caffe with VAE structure

Examples – MINST CVAE


(a) CVAE


(b) Regressor


(c) Ground Truth

Sampled pixels from one column as condition

Thank You