Variational Autoencoder (VAE)

Like the plain Autoencoder that we have already encountered a *Variational Autoencoder (VAE)* is comprised of an Encoder and a Decoder

In both cases

- the Encoder produces a latent representation $\mathbf{z}^{(\mathbf{i})}$ of its input $\mathbf{x}^{(\mathbf{i})}$
- the Decoder attempts to reconstruct $\mathbf{x}^{(\mathbf{i})}$ from $\mathbf{z}^{(\mathbf{i})}$

However, the behavior of the Decoder is undefined when presented with a latent $\mathbf{z}$ that did not arise from a training example.

- We can only hope that the output is reasonable

As we saw, this has implications as to our ability to use the AE as a means of generating synthetic examples.

The Decoder part of a VAE is identical to that of the plain Autoencoder.

But the Encoder part of a VAE is different in an important way. Given input $\mathbf{x}$

- It creates a *distribution* for the the latent representation $\mathbf{z}$
- Rather than creating a unique $\mathbf{z}$

The Encoder part of a VAE, given input $\mathbf{x}$

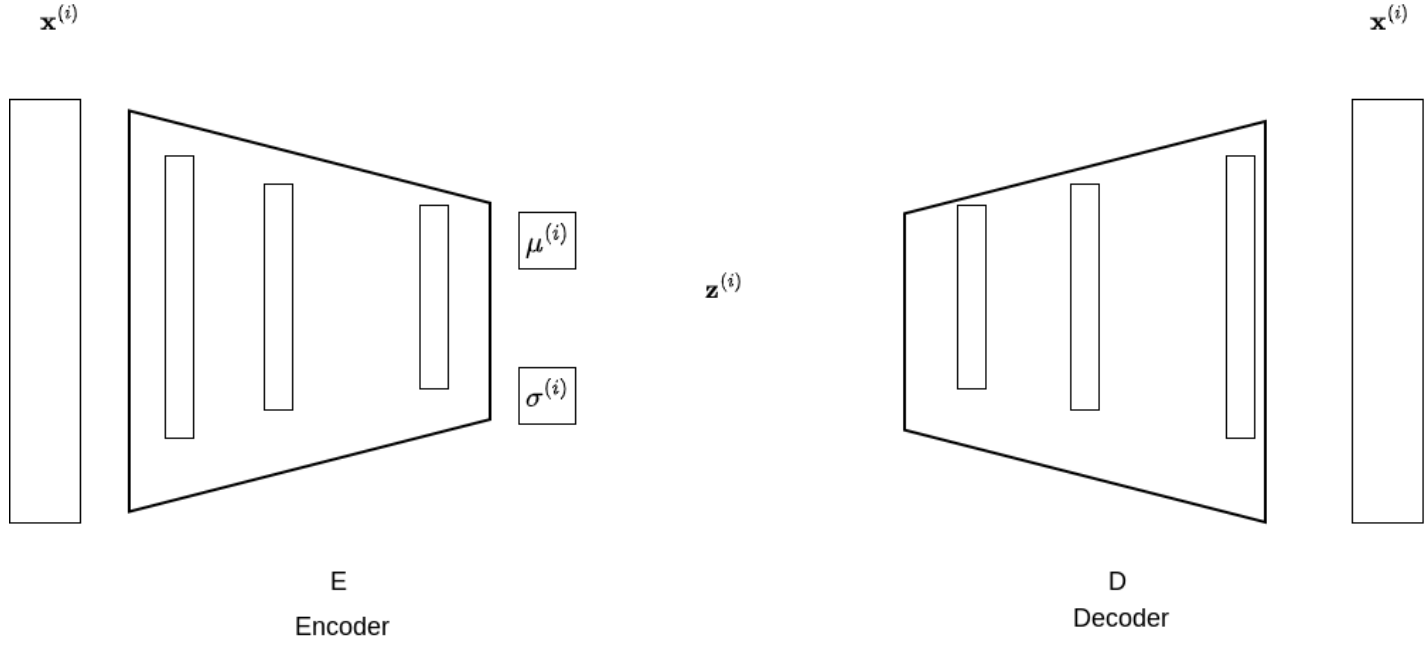
- Produces the parameters (e.g., $\mu^{(\mathbf{i})}$, $\sigma^{(\mathbf{i})}$) of a distributional form
- Draws a sample from the distribution as its output $\mathbf{z}^{(\mathbf{i})}$

Thus, the latent representation $\mathbf{z}$ of a given $\mathbf{x}$ is a probability distribution $q(\mathbf{z}|\mathbf{x})$.

This may address one of the concerns we raised with using a plain Autoencoder in a generative manner

- that a slight perturbation of the latent $\mathbf{z}^{(\mathbf{i})}$ obtained from input $\mathbf{x}^{(\mathbf{i})}$
- might have the Decoder produce $\tilde{\mathbf{x}}^{(\mathbf{i})}$ that is not similar to $\mathbf{x}^{(\mathbf{i})}$

Variational Autoencoder (VAE)



Note

$\mu^{(\mathbf{i})}$ and $\sigma^{(\mathbf{i})}$ are

- vectors
- computed values (and hence, functions of $\mathbf{x}^{(\mathbf{i})}$) and **not** parameters
- so training learns a *function* from $\mathbf{x}^{(\mathbf{i})}$ to $\mu^{(\mathbf{i})}$ and $\sigma^{(\mathbf{i})}$

This is not just straightforward engineering.

In fact: the architecture of the VAE was obtained from the math rather than vice-versa !

We provide a brief overview of the mathematics.

The interested reader is referred to a highly recommended VAE tutorial (https://arxiv.org/pdf/1606.05908.pdf) for a detailed presentation.

Details

Notation summary

term	dimension	meaning
$\mathbf{x}$	n	Random variable for Input
$\tilde{\mathbf{x}}$	n	Output: reconstructed $\mathbf{x}$
$\mathbf{z}$	$n' << n$	Random variable for Latent representation
E	$\mathbb{R}^n \to \mathbb{R}^{n'}$	Encoder
		$E(\mathbf{x}) = \mathbf{z}$
D	$\mathbb{R}^{n'} \to \mathbb{R}^n$	Decoder
		$\tilde{\mathbf{x}} = D(\mathbf{z})$
		$\tilde{\mathbf{x}} = D(E(\mathbf{x}))$
		$\tilde{\mathbf{x}} \approx \mathbf{x}$
$p(\mathbf{x})$	prob. distribution	*prior* distribution of Inputs
		intractable. Only have empirical.
$q(\mathbf{z})$	prob. distribution	*prior* distribution of Latents
$q(\mathbf{z} \mid \mathbf{x})$	prob. distribution	*posterior* marginal distribution of Latents given Input
		intractable
$p(\mathbf{z} \mid \mathbf{x})$	prob. distribution	*posterior* marginal distribution of Latents given Input
		intractable
$q_\Phi(\mathbf{z} \mid \mathbf{x})$	Neural Network	NN to approximate $q(\mathbf{z} \mid \mathbf{x})$
		Encoder
$p_\Theta(\mathbf{x} \mid \mathbf{z})$	Neural Network	NN to approximate $p(\mathbf{x} \mid \mathbf{z})$
		Decoder

Let's pretend that we don't already know the architecture of a VAE

- that the latent $\mathbf{z}^{(\mathbf{i})}$ is generated as a probability function of $\mu^{(\mathbf{i})}$ and $\sigma^{(\mathbf{i})}$ given input $\mathbf{x}^{(\mathbf{i})}$.

Instead let

- $\mathbf{x}$ denote a random variable representing an Input
    - the random variable is from the (unknown) distribution $p(\mathbf{x})$
- $\mathbf{z}$ denote a random variable representing a Latent
    - the random variable is from the (unknown) distribution $q(\mathbf{z})$

Because $\mathbf{x}$ and $\mathbf{z}$ are related, there is also a joint distribution of $(\mathbf{x}, \mathbf{z})$ from which we can define the marginal distributions

- $q(\mathbf{z}|\mathbf{x})$: the marginal distribution of Latent, given an Input
- $p(\mathbf{x}|\mathbf{z})$: the marginal distribution of Input, given a Latent

But there's a problem !

- The distribution $p(\mathbf{x})$ is *intractable*
  - e.g., Who can say what the distribution of images is in the physical world ?
  - At best: we have an empirical distribution (our training dataset)

We will side-step the intractability issues by defining Neural Networks to learn an approximation.

- $q_{\Phi}(\mathbf{z}|\mathbf{x})$: Neural Network, with weights Φ to approximate $q(\mathbf{z} \mid \mathbf{x})$
  - The Encoder
- $p_{\Theta}(\mathbf{x}|\mathbf{z})$:Neural Network, with weights Θ, to approximate $p(\mathbf{x} \mid \mathbf{z})$
  - The Decoder

The mapping from Latent to reconstructed Input is not necessarily unique, thus we marginalize $\mathbf{x}$ over $\mathbf{z}$

$$p(\mathbf{x}) = \int_{\mathbf{z} \in q(\mathbf{z})} p(\mathbf{x}|\mathbf{z})\ q(\mathbf{z})$$

Some obvious concerns about the integral

- It may be very expensive to draw many samples of $\mathbf{z}$ from $q(\mathbf{z})$ for each training example $\mathbf{x^{(i)}}$
- There are many random choices of $\mathbf{z}$ from $q(\mathbf{z})$ which can't reconstruct $\mathbf{x^{(i)}}$
    - i.e., $p_\Theta(\mathbf{x^{(i)}}|\mathbf{z}') = 0$ for many $\mathbf{z}'$

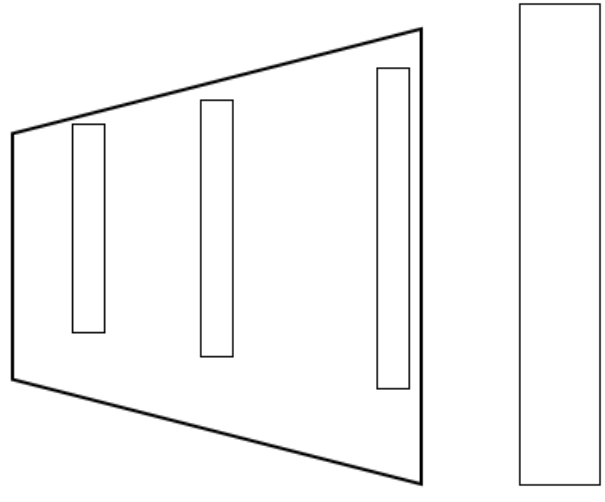
We can improve our sampling by considering only those choices of $\mathbf{z}$ that could generate $\mathbf{x}$ and re-write the objective as

$$p(\mathbf{x}) = \int_{\mathbf{z} \in q(\mathbf{z}|\mathbf{x})} p(\mathbf{x}|\mathbf{z})\ q(\mathbf{z})$$

Using the Decoder Neural Network $p_\Theta(\mathbf{x}|\mathbf{z})$ to approximate $p(\mathbf{x}|\mathbf{z})$ gives rise to the following architecture

$$\tilde{\mathbf{x}}^{(i)} = \int_{\mathbf{z}} p_\theta(\mathbf{x}|\mathbf{z}) q(\mathbf{z})$$

$\mathbf{z}^{(i)} \in q(\mathbf{z}|\mathbf{x})$



D
Decoder

We still can't train $p_\Theta(\mathbf{z}|\mathbf{x})$ because we don't know $q(\mathbf{z}|\mathbf{x})$.

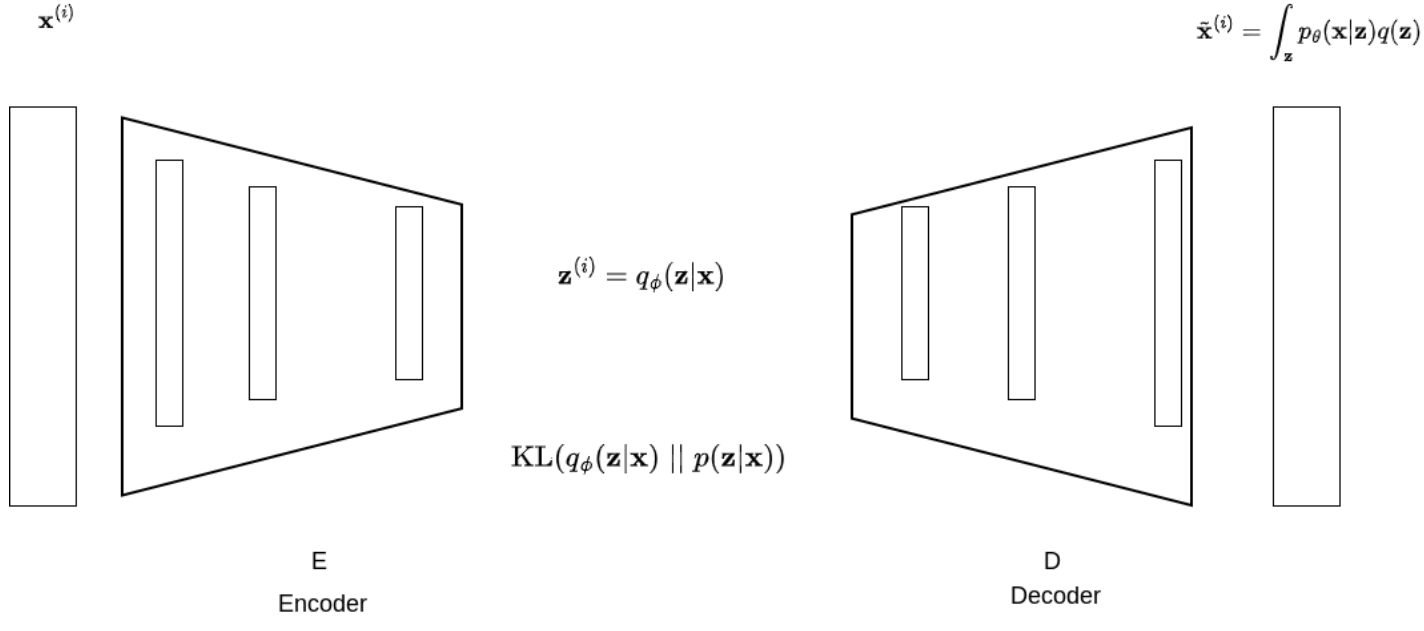
Let's use the Neural Network (Encoder) $q_\Phi(\mathbf{z}|\mathbf{x})$ to approximate $q(\mathbf{z}|\mathbf{x})$.

But we must constrain the Encoder to produce a distribution that approximates the true $q(\mathbf{z}|\mathbf{x})$.

We do so by including this as a constraint (part of the Loss function used in training) using KL divergence as a measure of dissimilarity of two distributions

$$\mathbf{KL}(q_\Phi(\mathbf{z}|\mathbf{x}) \;||\; q(\mathbf{z}|\mathbf{x}))$$

The resulting architecture:



We can train the Encoder/Decoder pair with the objective that the reconstructed $\tilde{\mathbf{x}}^{(\mathbf{i})}$ approximates the true $\mathbf{x}^{(\mathbf{i})}$ from the training set, across all examples i.

One one of stating this is as a Maximum Likelihood:

- Solve for the weights Φ, Θ
- That maximize the (log) Likelihood of the set of reconstructions $\tilde{\mathbf{X}}$ reproducing the training set $\mathbf{X}$

Since our practice is to minimize Loss (rather than maximize an objective function) we write our loss as (negative of log) likelihood

$$\mathcal{L} \quad = \quad -\log(p(\mathbf{X}))$$

Minimizing $\mathcal{L}$ is equivalent to maximizing likelihood.

Adding the KL divergence constraint to our Likelihood objective gives the loss function

$$\begin{aligned}\mathcal{L} &= -\log(p_\Theta(\mathbf{x})) + \mathbf{KL}(q_\Phi(\mathbf{z}|\mathbf{x}) \;||\; q(\mathbf{z}|\mathbf{x})) \\ &= \mathcal{L}_R + \mathcal{L}_D\end{aligned}$$

which now has two objectives

- Reconstruction loss $\mathcal{L}_R$: maximize the likelihood (by minimizing the negative likelihood)
- Divergence constraint $\mathcal{L}_D$: $q_\Phi(\mathbf{z}|\mathbf{x})$ must be close to $q(\mathbf{z}|\mathbf{x})$

$$\begin{aligned}\mathcal{L}_R &= -\log(p_\theta(\mathbf{x})) \\ \mathcal{L}_D &= \mathbf{KL}\left(q_\Phi(\mathbf{z}|\mathbf{x}) \;||\; q(\mathbf{z}|\mathbf{x})\right)\end{aligned}$$

We will show (in the next section: lots of algebra !) that the loss can be re-written as

$$\mathcal{L} = -\mathbf{E}_{\mathbf{z}\sim q_\Phi(\mathbf{z}|\mathbf{x})}\left(\log(p_\Theta(\mathbf{x}|\mathbf{z}))\right) + \mathbf{KL}(q_\Phi(\mathbf{z}|\mathbf{x}) \;||\; q(\mathbf{z}))$$

This is *almost* identical to our original express for $\mathcal{L}$ except

- Re-write
  $\log(p_\theta(\mathbf{x})) = \mathbf{E}_{\mathbf{z}\sim q_\Phi(\mathbf{z}|\mathbf{x})}\left(\log(p_\Theta(\mathbf{x}|\mathbf{z}))\right)$
- the KL term becomes
  $$\mathbf{KL}\left(q_\Phi(\mathbf{z}|\mathbf{x}) \;||\; q(\mathbf{z})\right)$$
  rather than the original
  $$\mathbf{KL}\left(q_\Phi(\mathbf{z}|\mathbf{x}) \;||\; q(\mathbf{z}|\mathbf{x})\right)$$

The purpose of re-writing: replace intractable $q(\mathbf{z}|\mathbf{x})$ with a tractable $q(\mathbf{z})$!

- So we can have a Loss function with which we can train !

TL;DR

- The VAE has a very interesting **two part** Loss Function
    - Reconstruction Loss, as in the Vanilla AE
    - Divergence Loss
- The Reconstruction Loss is not sufficient
    - Issues of intractability arise
    - The Divergence Loss skirts intractability
        - By constraining the Encoder to produce a tractable distribution

Advanced: Obtain $\mathcal{L}$ by rewriting

$\mathbf{KL}(q_\Phi(\mathbf{z}|\mathbf{x}) \,||\, q(\mathbf{z}|\mathbf{x}))$

Let's derive a simpler expression for $\mathcal{L}$ by manipulating $\mathbf{KL}(q_\Phi(\mathbf{z}|\mathbf{x}) \,||\, q(\mathbf{z}|\mathbf{x}))$:

$$
\begin{aligned}
\mathbf{KL}(q_\Phi(\mathbf{z}|\mathbf{x}) \,||\, q(\mathbf{z}|\mathbf{x})) &= \textstyle\sum_z q_\Phi(\mathbf{z}|\mathbf{x})(\log(q_\Phi(\mathbf{z}|\mathbf{x}) - \log(q(\mathbf{z}|\mathbf{x})) \\
&= \mathbf{E}_{z\sim q_\Phi(\mathbf{z}|\mathbf{x})}\left(\log(q_\Phi(\mathbf{z}|\mathbf{x}) - \log(q(\mathbf{z}|\mathbf{x}))\right) \\
&= \mathbf{E}_{z\sim q_\Phi(\mathbf{z}|\mathbf{x})}\big(\ \log(q_\Phi(\mathbf{z}|\mathbf{x})) \\
&\quad - (\ \log(p(\mathbf{x}|\mathbf{z})) + \log(q(\mathbf{z})) - \log(p(\mathbf{x}))\ )\ \ \big)
\end{aligned}
$$

$$
\begin{aligned}
&\mathbf{KL}(q_\Phi(\mathbf{z}|\mathbf{x}) \,||\, q(\mathbf{z}|\mathbf{x})) \\
&- \log(p(\mathbf{x})) &= \mathbf{E}_{z\sim q_\Phi(\mathbf{z}|\mathbf{x})}\big(\ \log(q_\Phi(\mathbf{z}|\mathbf{x})) - (\log(p(\mathbf{x}|\mathbf{z})) + \log(q(\mathbf{z}) \\
&&= \mathbf{E}_{z\sim q_\Phi(\mathbf{z}|\mathbf{x})}\big(\ -\log(p(\mathbf{x}|\mathbf{z})) + (\ \log(q_\Phi(\mathbf{z}|\mathbf{x})) - \log(q(\\
&&= -\mathbf{E}_{z\sim q_\Phi(\mathbf{z}|\mathbf{x})}\left(\log(p(\mathbf{x}|\mathbf{z}))\right) + \mathbf{KL}(q_\Phi(\mathbf{z}|\mathbf{x}) \,||\, q(\mathbf{z})) \\
&\mathcal{L} &= -\mathbf{E}_{z\sim q_\Phi(\mathbf{z}|\mathbf{x})}\left(\log(p(\mathbf{x}|\mathbf{z}))\right) + \mathbf{KL}(q_\Phi(\mathbf{z}|\mathbf{x}) \,||\, q(\mathbf{z}))
\end{aligned}
$$

The key step:

- Using Bayes Theorem to re-write
$$\log(q(\mathbf{z}|\mathbf{x}))$$
as
$$\log(p(\mathbf{x}|\mathbf{z})) + \log(q(\mathbf{z})) - \log(p(\mathbf{x}))$$
- This allows us do away with intractable conditional probability $q(\mathbf{z}|\mathbf{x})$
- In favor of unconditional probability $q(\mathbf{z})$

The LHS cannot be optimized via SGD (recall the tractability issue with $q(\mathbf{z}|\mathbf{x})$).

But the RHS can be made tractable by

- Approximating $p(\mathbf{x}|\mathbf{z})$ with $p_\Theta(\mathbf{x}|\mathbf{z})$
- Assuming that the distributions $q(\mathbf{z})$ (and $p(\mathbf{x})$) as Normal

Choosing $q(\mathbf{z})$

So what distribution should we use for the prior $q(\mathbf{z})$?

- It should be differentiable, since we use Gradient Descent for optimization
- It should be tractable with a closed form (such as a normal)
- If we choose $q(\mathbf{z})$ as normal, we can require $q_\phi(\mathbf{z}|\mathbf{x})$ to be normal too
    - The KL divergence between two normals is an easy to compute function of their means and standard deviations.
    - See VAE tutorial (https://arxiv.org/pdf/1606.05908.pdf) Section 2.2

Re-parameterization trick

There is still one impediment to training.

It involves the random choice of $\mathbf{z} \sim q_\Phi(\mathbf{z}|\mathbf{x})$ in

$$\mathcal{L}_R = \mathbf{E}_{z \sim q_\Phi(\mathbf{z}|\mathbf{x})} \left(\log(p_\Theta(\mathbf{x}|\mathbf{z})) \right)$$

This is not a problem in the forward pass.

But in the backward pass we need to compute

$$\frac{\mathcal{L}_R}{\partial \Theta}$$

How do we back propagte through a random choice ?

The "reparameterization trick" redefines the random choice $\mathbf{z}$ as

$$\begin{aligned} \mathbf{z} &= \mu_\theta(\mathbf{x}) + \sigma_\theta(\mathbf{x}) * \epsilon \\ \epsilon &\sim N(0,1) \end{aligned}$$

That is

- Once we have defined $q(\mathbf{z})$ to be a Normal distribution
- We can re-write an element of the distribution
    - as the distribution's mean plus a random standardized Normal ϵ times the distribution's standard deviation

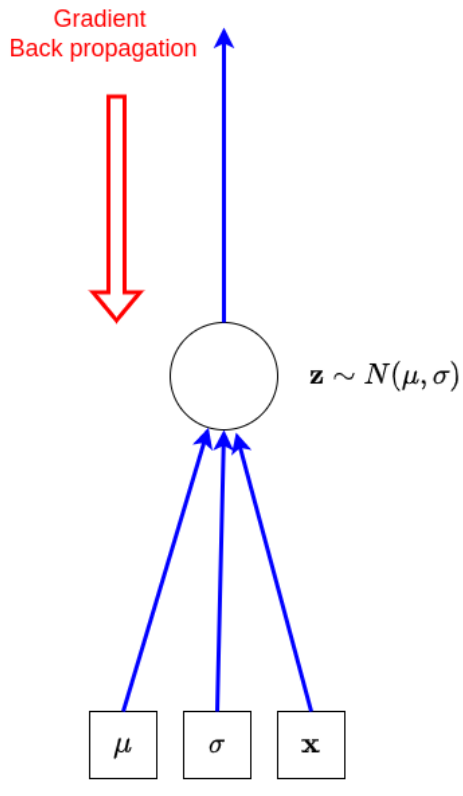
In this formulation, the random variable ϵ appears in a product term

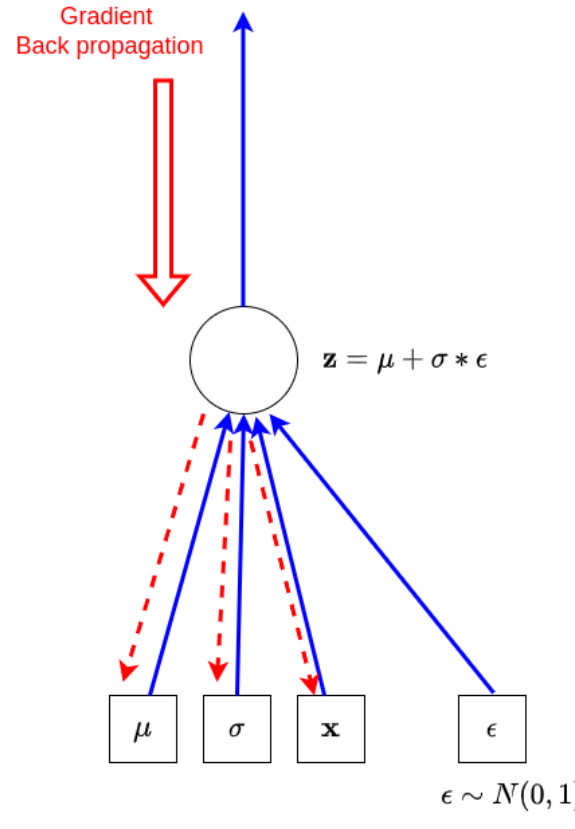
- We can differentiate the product with respect to Θ
- ϵ can be treated as a constant in $\frac{\partial \epsilon}{\partial \Theta}$

The Encoder design is now to produce (trainable parameters) μ_Θ, σ_Θ

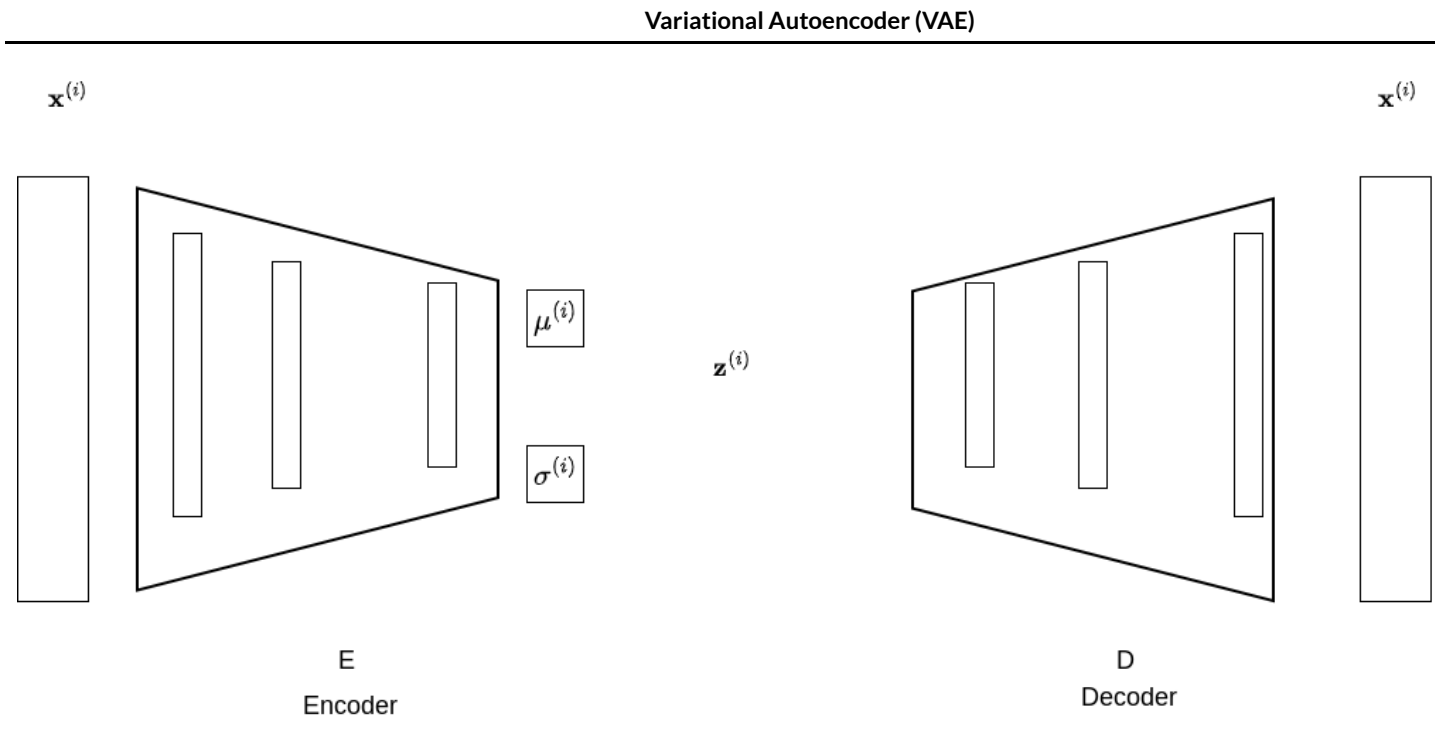
- And $\mathbf{z}$ indirectly

Reparameterization trick





This gets us to the final picture of the VAE:



To train a VAE:

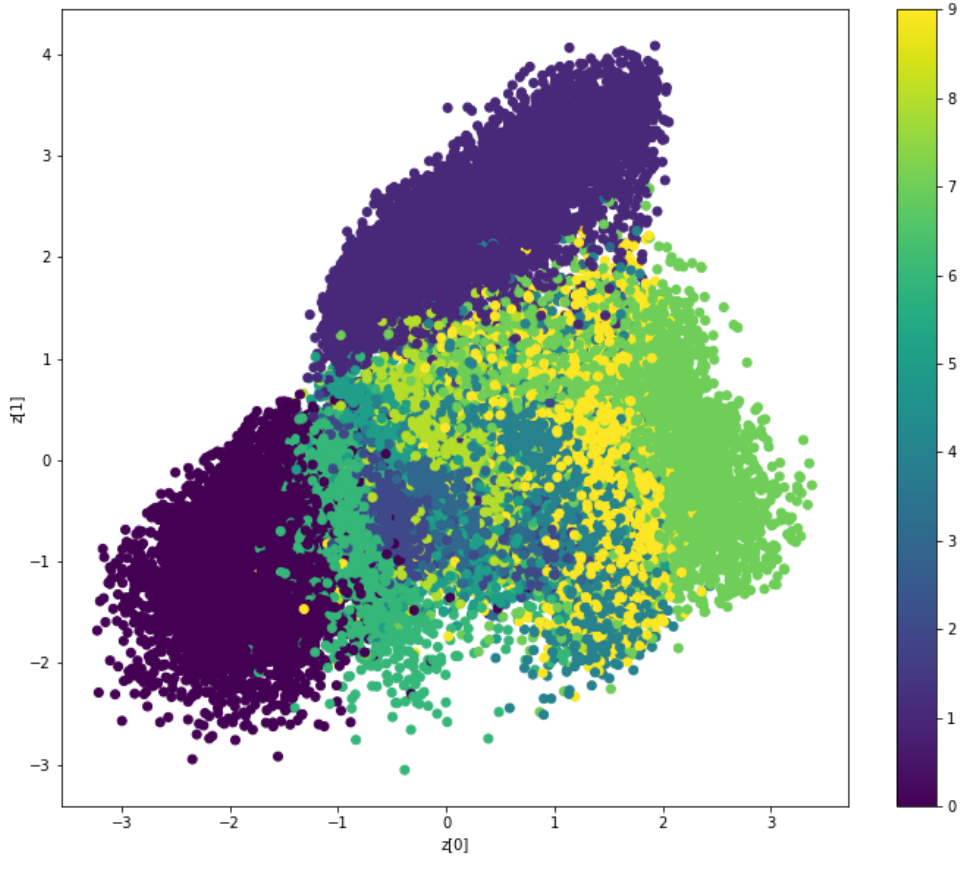
- pass input $\mathbf{x}^{(\mathbf{i})}$ through the Encoder, producing $\mu^{(\mathbf{i})}$, $\sigma^{(\mathbf{i})}$
    - use $\mu^{(\mathbf{i})}$, $\sigma^{(\mathbf{i})}$ to sample a latent representation $\mathbf{z}^{(\mathbf{i})}$ from the distribution
- pass the sampled $\mathbf{z}^{(\mathbf{i})}$ through the decoder, producing $D(\mathbf{z}^{(\mathbf{i})})$
- measure the reconstruction error $\mathbf{x}^{(\mathbf{i})} - D(\mathbf{z}^{(\mathbf{i})})$, just as in a plain AE
- back propagate the error, updating all weights and μ, σ

Each time that we encounter the same training example (e.g., in different epochs), we select another random element from the distribution.

Thus the VAE learns to represent the same example from multiple latents.

We can examine how the latent representations produced by the VAE form clusters:

MNIST examples: clustering of latent $\mathbf{z}$



In comparing the clusterings produced by the VAE against our previous example of a plain Autoencoder be aware

- The two models are displaying results on different data: MNIST digits verus Fashion MNIST
- The architecture of the Encoder and Decoder are different in the two models
    - The plain Autoencoder used extrememly simple architectures
        - Could the more complex architecture of the VAE Encoder/Decoder be the cause of tighter clustering ?

Certainly room for experimentation !

Using a VAE to produce synthetic examples

To give you an idea of the generative nature of the VAE, consider

- Creating latent vectors $\mathbf{z}$ from scratch
    - **not** as the output of the Encoder
- Varying these latent vectors systematically and examining the output created by the Decoder

Note that the outputs

- are **not** instances of any examples
- There was no guarantee that a random $\mathbf{z}$ would produce something that looked like a digit !

We may even be able to interpret the elements of $\mathbf{z}$

- $\mathbf{z}_0$: control slant ?
    - See the bottom row of 0's
- $\mathbf{z}_1$: control "verticality" ?
    - See right-most column

ELBo (Evidence-based Lower Bound)

By re-writing the Loss, we removed the intractable term $q(\mathbf{z}|\mathbf{x})$

It turns out that even this may not be necessary.

For the truly interested reader:

- The derivation uses a method known as *Variational Inference*. See this blog (https://mbernste.github.io/posts/variational_inference/) for a summary.
- One can show that loss $\mathcal{L}$ is equal to -1 times the *ELBo* (Evidence Based Lower Bound)

So if one knows how to maximize the ELBo (https://mbernste.github.io/posts/elbo/), one can minimize the loss.

Loss function: discussion

Let's examine the role of $\mathcal{L}_R$ and $\mathcal{L}_D$ in the loss function $\mathcal{L}$.

- What would happen if we dropped $\mathcal{L}_D$?
    - We would wind up with a deterministic $\mathbf{z}$ and collapse to a vanilla VAE
- What would happen if we dropped $\mathcal{L}_R$?
    - the encoding approximation $q_\Phi(\mathbf{z}|\mathbf{x})$ would be close to the empirical $p(\mathbf{z}|\mathbf{x})$ *in distribution*
    - but two variables with the same distribution are not necessarily the same ?
        - e.g., get a distribution p by flipping a coin
            - let distribution q be a relabelling of p by changing Heads to Tails and vice-versa
            - p and q are equal in distribution but clearly different !

Conditional VAE

The VAE would seem to offer a solution to the problem of creating synthetic data.

But there is a problem

- an *unlabeled* example is created
- we have no way of knowing the label
- nor do we have a way of *controlling* the output so as to be an example with a specified label

We can modify the VAE so as to *conditionally* generate an example with a specified label.

- Conditional VAE (Cond_VAE_Generative.ipynb)

Experiments with Variational Autoencoders

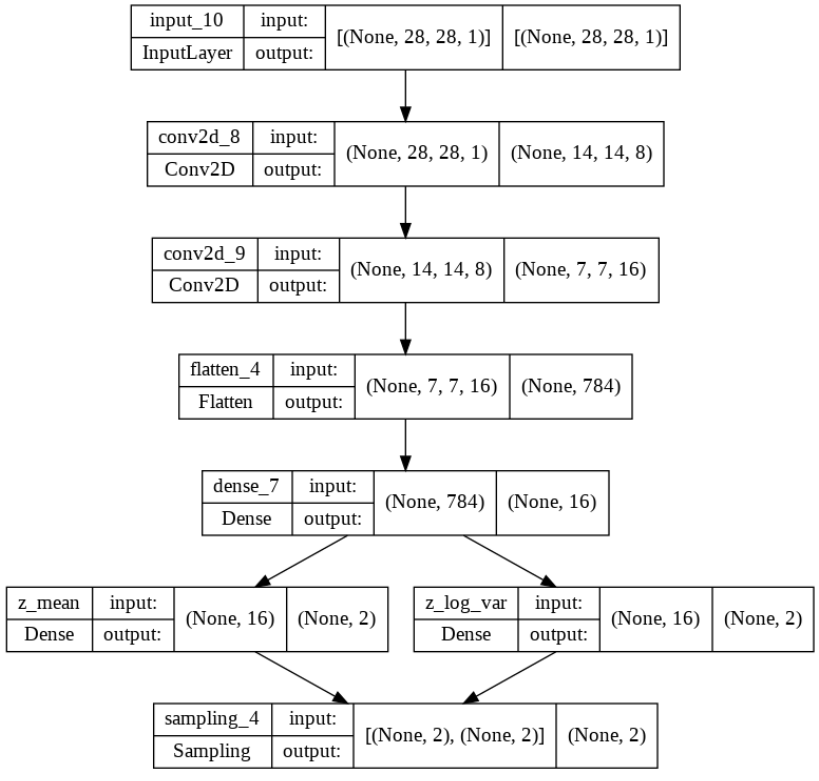
We can learn much more about the properties and use of a VAE through examples

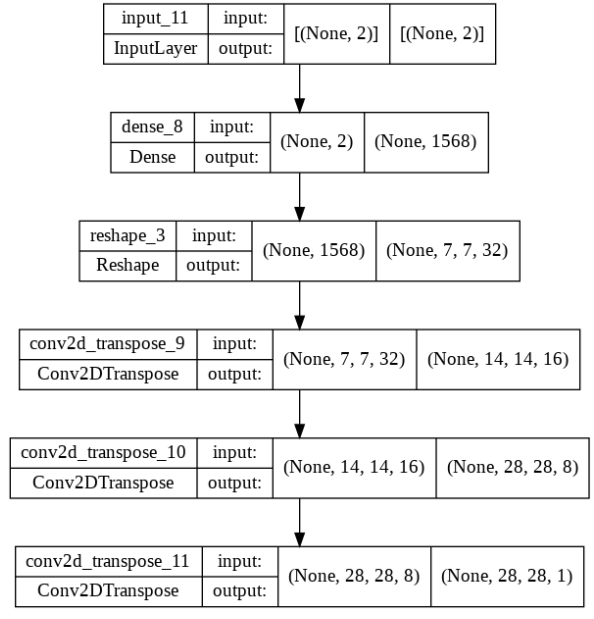
Let's explore this notebook (VAE_code.ipynb)

- illustrates Latent representation, Denoising, Anomaly Detection
- (secondary objective: study the code)

More complex architecture for the Encoder and Decoder than our plain VAE

- Illustrative, not a necessity
- Encoder and Decoder don't need to be symmetric


input_10 | input: | [(None, 28, 28, 1)] | [(None, 28, 28, 1)]
InputLayer | output:
conv2d_8 | input: | (None, 28, 28, 1) | (None, 14, 14, 8)
Conv2D | output:
conv2d_9 | input: | (None, 14, 14, 8) | (None, 7, 7, 16)
Conv2D | output:
flatten_4 | input: | (None, 7, 7, 16) | (None, 784)
Flatten | output:
dense_7 | input: | (None, 784) | (None, 16)
Dense | output:
z_mean | input: | (None, 16) | (None, 2)
Dense | output:
z_log_var | input: | (None, 16) | (None, 2)
Dense | output:
sampling_4 | input: | [(None, 2), (None, 2)] | (None, 2)
Sampling | output:


input_11 | input: | [(None, 2)] | [(None, 2)]
InputLayer | output:
dense_8 | input: | (None, 2) | (None, 1568)
Dense | output:
reshape_3 | input: | (None, 1568) | (None, 7, 7, 32)
Reshape | output:
conv2d_transpose_9 | input: | (None, 7, 7, 32) | (None, 14, 14, 16)
Conv2DTranspose | output:
conv2d_transpose_10 | input: | (None, 14, 14, 16) | (None, 28, 28, 8)
Conv2DTranspose | output:
conv2d_transpose_11 | input: | (None, 28, 28, 8) | (None, 28, 28, 1)
Conv2DTranspose | output:

Encoder

- Note the two branches to nodes `z_mean` and `z_log_var`
    - The output of their common parent is used to generate two separate values (μ and σ)
    - μ and σ are both vectors of length 2
        - Thus, the sampled $\mathbf{z}$ is also of length 2
        - In our grid illustration of generating synthetic examples, we vary each of the 2 components of $\mathbf{z}$
    - Latent is *much* shorter than in the plain VAE
        - does the random nature of sampling facilitate shorter representation ?