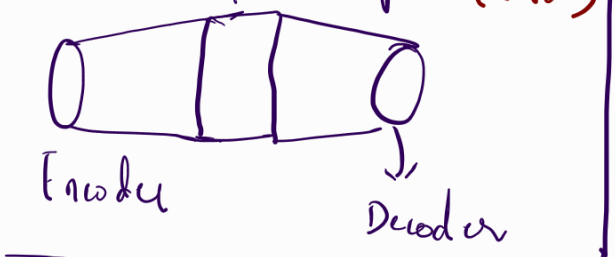
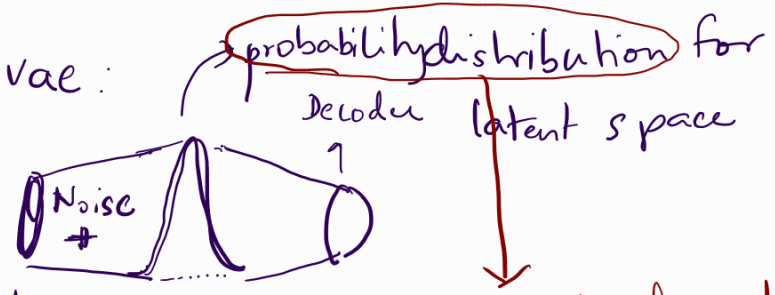
Normal Autoencoder



VAE:



main goal is to learn these distributions that can approximate the data.

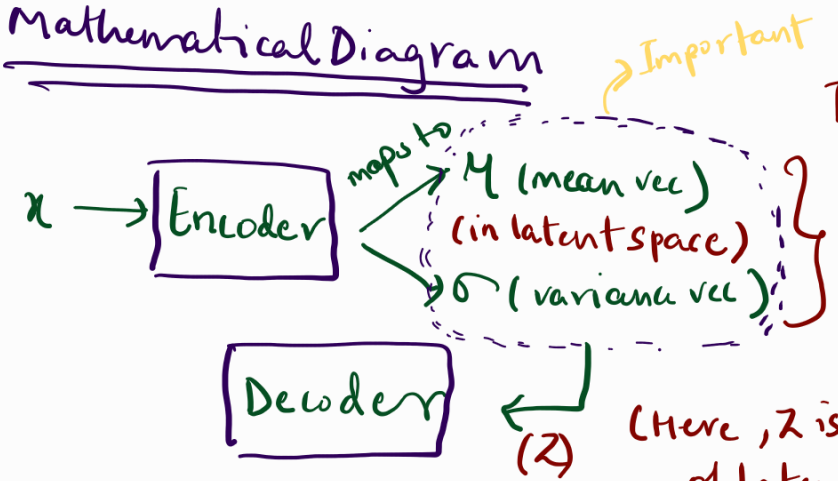
Mathematical Objective: Maximum Likelihood

intution: we want our data closely related to the actuals. Hence, if our distribution is close to what it should be → eureka

How do we achieve: By learning these params from our distributions.

Q. Why distribution approach is better than direct mapping into latent space and the providing output

Mathematical Diagram



How do we get the best sampling ? From deterministic to stochastic

From point value to distribution $N(\mu, \sigma)$

Why cannot we just use the mean to sample z ? (variance ?)

Answer: Random variation, rather than randomness causing deterministic nature to a certain degree.

Intuition: sample can have slight differences while still capturing the core features of original data.

Objective Function : (To estimate the a good distribution)

Reconstruction loss + KL divergence

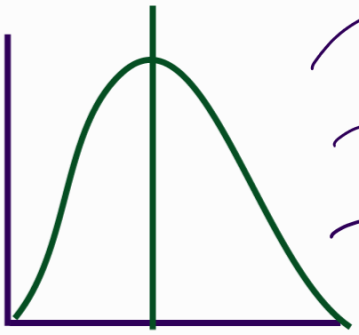
Motive : Decoded output is similar to actual data.

Motive : Makes the learned latent space distribution resemble a standard normal distribution → $N(0,1)$ (as closely as possible).

$\mathcal{L}_{VAE}$ = reconstruction loss + KL Divergence ($N(0,1)$)

Q) Why do we need the latent space distribution to approximate a standard normal distribution?

Standard normal Distribution (ND)



→ Symmetric about mean

→ $\mu = 0, \sigma = 1$

→ (All normal distributions are symmetrical but vice versa → not true)

Note1 : Normal distribution is key to central limit theorem.

Note2 : One has to study ND to interpret KL.

Intuition: we introduce stability and regularity into space.

⇒ while sampling points, one is less likely to encounter extreme or erratic values } Hence, coherent outputs.

⚠ Regularization proposed by KL, provides continuity in latent space, making it easier to transit from one feature to another

Why KL divergence? → measures difference between two probability distributions

in our case ① → Distribution learned by encoder (specific to each data point)

② → $N(0, 1)$

$D_{KL}(q(z|x) \,||\, N(0,1))$ → Learned distribution for z given x

A B

{intuition: to keep A close to B}

⇒ So that the learned representation stay close to actual data.

Note: If the distribution stays close to $N(0,1)$ no large shifts in latent spaces will be learnt

The Reparametrization trick: (Why, What, How)

Since the beginning of learning process we have been generating and approximating distribution

↓

we need gradients to back propogate | Due to random sampling

↓

Relu son → smooth gradients → continuous functions

makes it impossible

for the gradients to flow through encoder → Model doesn't learn properly

The trick: Reparameterize sampling step } This will make it differentiable.

(X) Instead of sampling z directly from $N(0,1)$

we break it into 2 parts.

First, sample a standard normal random variable ϵ from $N(0,1)$ → Then shift and scale ϵ using μ and σ

$$z = \mu + \sigma \cdot \epsilon$$

• Our latent vector would look like

Since, the randomness parameter is separated from μ and σ, the gradients can flow through μ and σ properly.

Generates randomness but is independent of μ, σ

Why would the above trick work?

→ Since μ and σ are now deterministicly used, we have a differentiable path for the latent variable z back to the encoders parameters.

→ Providing stochastic nature (ϵ) to VAE and also differentiable at the same time.

Reparemetrization ⟶ back propogation.

Training objective of the VAE: { Reconstruction loss + KL divergence.

(A) (B)

$$\mathcal{L}_{VAE} = E_{q(z|x)}[\log p(x|z)] - D_{KL}(q(z|x) \| p(z))$$

Q) Why might balancing these two terms be essential for training an effective VAE? (output sensible)

(A) Accurate reconstruction from latent space. Sufficient reconstruction accuracy. Capturing key features.

(B) Prevents latent space from overfitting. Helping the system to not copy and be realistic at the same time.

The balance allows realistic reconstructions while maintaining generalizable and structured latent space. Hence, allowing VAE to generate coherent new samples and interpolate smoothly between different points in latent space.

Until now, we have built these amazing latent spaces that provide the representational learning in vector format. But, does we really knows whats within that latent space, what core feature representations has the model learnt → Ans → INTERPRETABILITY

* Visualizing Latent Space Interpretability.

- **Latent space traversals** → systematically altering values in the latent space to understand how changes in latent variables affect the output.
  - Latent space: compressed, lower space dimensions
  - traversals: (Interpolating) between data points
    Eg: $3 \rightarrow 7(z)$ (z)

Examples: (a) Stroke thickness in MNIST (b) Presence of smile in CelebA

Note: The following traversals involve varying one specific dimension, keeping the remaining constant.

Mathematical Intuition (Important)

Latent space z is modeled as a probability distribution $N(\mu, \sigma)$

↓

During Latent Space Traversal

- Fix all dimensions except one (Eg: z_i)
- Linearly vary z_i across range of values (Eg: $-3, 3$)
- Pass the modified z through decoder $p(x|z)$ to generate

$\leftrightarrow$

$$z = \mu + \sigma \cdot \epsilon$$

Disentangled representation

a way of encoding data such that each dimension in the latent space corresponds to a distinct, interpretable factor of variation of data

Q) Why might the ability to vary only one dimension in the latent space be useful for understanding and controlling the behavior of generative model?

By altering a single dimension in latent space we get a clear view of how specific features are represented and controlled. This helps us understand the structure of the latent space and whether the VAE has learned disentangled representations where different features are encoded independently.

- **Disentanglement Score (DS)**: The DS corresponds to how well each latent dimension corresponds to a single interpretable feature. (altering one dimension shouldnt affect the other features)

(compact and meaning full compression)

why do we measure this score?

These shouldnt overlap ↗

Eg: Dim1 : controls presence of smile
Dim2 : controls presence of glasses.

(2 metrics to calculate DS)

Beta β -VAE Disentanglement Metric

- Extension of VAE, that promotes disentangled representations by modifying the obj function.
- Hyperparam : β : is used to control the tradeoff between reconstruction accuracy and latent space regularization
- $\mathcal{L}_\beta = A - \beta \cdot B$
  - β > 1 increases regularization encouraging disentanglement by making latent space more independent
  - β < 1 Prioritizes reconstruction accuracy
- Promotes ↙
- Increased Regularization
  more independent and sparse rep
- Penalty and Complexity
  Forces latent space to encode only essentials → avoid overlapping

Mutual Information Gap (MIG)

Measures the difference in mutual information (MI) between the most latent dim and second most latent dim

$$MIG = \frac{1}{K} \sum_{k=1}^{K} \frac{MI(z_i ; v_k) - MI(z_j ; v_k)}{H(v_k)}$$

$z_i \Rightarrow$ most informative latent

$z_j \Rightarrow$ second most informative

$v_k \Rightarrow$ a known feature

* High MIG indicates that a particular single latent dim is primarily responsible for a feature, while others contribute little.

- High Score ⇒ Each dimension corresponds to a distinct dimension
- Low Score ⇒ Latent space is entangled making it harder to interpret individual features.

Robustness and Generalization: examining how the latent space and reconstructed outputs behave when the input data is perturbed {intentionally modifying or altering}

↳ Helps us understand ⓐ Robustness ⓑ Generalization.

- **Robustness**: stable the learned latent representations are when small changes like noise or adversarial perturbations are added to input.
- **Generalization**: meaningfull outputs by VAE.

Mathematical intuition: → ⓐ Perturbation Analysis
→ ⓑ Reconstruction stability

ⓐ • Add Gaussian noise η to the input data x: $\tilde{x} = x + \eta$, where $\eta \sim N(0, \sigma^2)$

• Encode both $x, \tilde{x}$ to obtain ⟶ $z, \tilde{z} \Rightarrow$ MSE $= \|z - \tilde{z}\|^2$

Stability.

ⓑ Decode z and $\tilde{z}$ to obtain ⟶ $\hat{x}, \tilde{\hat{x}} \Rightarrow$ MSE $= \|\hat{x} - \tilde{\hat{x}}\|$

Reconstruction Stability

Generative Quality Evaluation with metrics: Quality, Diversity

- **Frechet inception distance**: compares statistical properties like mean & covariance of features from real & generated extracted from a pretrained model (Eg Inception V3)

$$FID = \|\mu_r - \mu_g\|^2 + Tr(\Sigma_r + \Sigma_g - 2\sqrt{\Sigma_r \Sigma_g})$$

$\mu_r, \Sigma_r, \mu_g, \Sigma_g$: mean and covariance of real & generated outputs.

Inception Score: Measures diversity and quality of generated images based on the predicted class probabilities using pretrained classifiers.

$$IS = \exp(E_{x \sim p_g}[D_{KL}(p(y|x) \| p(y))])$$

$p(y|x) \Rightarrow$ predicted label distribution

$P(y)$ marginal label distribution

SSIM : Structural Similarity Index Measure.

↳ Considers → Luminance, contrast, structure as an quality assesment.

→ used for smoothness measuring

↓

continuous & meaningful transitions.

INTRODUCTION TO CHAOS THEORY AND ITS IMPLEMENTATION IN VAE

Chaotic Systems :→ deterministic systems that exhibit seemingly random behavior

Key characteristics : ⓐ Sensitive dependence to initial conditions ⓑ Deterministic Dynamics ⓒ Fractal and self similarity.