AutoEncoder

- neural network that learns a compressed(compact) representation of data
- architecture
  - encoder: mapping input to latent space
  - decoder: reconstruct the input from the latent code
- objective
  - minimize reconstruction error
- deterministic latent representation
  - no explicit probabilistic modeling → not a generative model
    - not define a valid generative distribution
  - random sampling in latent space is unreliable
    - sampling random z does not guarantee meaningful outputs

---

Variational AutoEncoder (VAE)

- impose a probabilistic structure on the latent space → learn latent distirbution directly
  - encoder
    - $q_\phi(z \mid x) = \mathcal{N}(\mu(x), \sigma^2(x))$
    - approximate posterior: distribution of latent z given by x
    - output statistics of distribution
  - decoder
    - $p_\theta(x \mid z)$
    - likelihood: generative distibution of x given by latent z
- **Reparametrization Trick**
  - $z = \mu + \sigma \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$
  - separates randomness from parameters → enabling backpropagation

- objective
  - maximizing marginal likelihood of data $\log p(x)$ → intractable

latent variable model에서 MLE에 의해 $\log p(x)$ 최대화 → $p(x) = \int p(x,z)\,dz$ posterior $p(z|x)$ → intractable
variational inference : 잘 알려진 쉬운 분포 $q_\phi(z|x)$로 $p(z|x)$에 근사 $\log p(x) = \text{ELBO} + D_{KL}(q_\phi(z|x) \| p(z|x))$
ELBO 최대화 → $D_{KL}(q_\phi(z|x)\|p(z|x))\downarrow$ & $\log p(x)\uparrow$

ELBO (Evidence Lower BOund) in VAE

$$\log p(x) \;\geq\; \mathbb{E}_{q(z|x)}[\log p(x \mid z)] - \mathrm{KL}(q(z \mid x) \,\|\, p(z))$$

$\text{ELBO} = \mathbb{E}_{q(z|x)}[\log p(x,z) - q_\phi(z|x)]$
$= \mathbb{E}_{q(z|x)}[\log p(x|z) + \log p(z) - \log q_\phi(z|x)]$
$= \mathbb{E}_{q(z|x)}[\log p(x|z)] - D_{KL}(q_\phi(z|x)\|p(z))$
reconstruction / regularization

- instead of maximizing data log likelihood → maximizing ELBO ↑
- **Reconstruction term**
  - encourages accurate reconstruction → keep latent informative
- **KL regularization term**
  - pushes posterior toward prior (e.g., Gaussian distribution) → smooth ↑
    - preventing posterior from dirac delta distribution
- drawbacks
  - blurry samples & posterior collapse
    - due to Gaussian likelihood and averaging
  - continuous latent space → not ideal for discrete structure

Vector-Quantized Variational AutoEncoder (VQ-VAE)

- replace continous latent z with discrete codebook entries
  - encoder → outputs a vector
  - vector → quantized to nearest codebook vector
- objective
  - $\mathcal{L} = \|x - \hat{x}\|^2 + \|\mathrm{sg}[z_e] - e\|^2 + \beta\|z_e - \mathrm{sg}[e]\|^2$
  - reconstruction loss + codebook loss + commitment loss
- captures discrete structure

Generative Model

- push-forward mapping
  - learn a mapping that pushes a simple distribution into a complex data distribution
- two viewpoints:
  - density-based (likelihood): model $p_\theta(x)$ directly (e.g., flows)
  - implicit (sampling): model a generator that can sample but may not give tractable likelihood (e.g., GANs)

Evaluation Metrics

- low-level: perceptual metric
- high-level: compare real and generated sample distributions in a feature space
  - metrics often rely on pretrained networks (Inception) as a feature extractor
- caveats
  - not always aligned with human preference
  - sensitive to the feature extractor and dataset domain shift

Peak Signal-to-Noise Ratio (PSNR)

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2$$

$$\mathrm{PSNR} = 10\log_{10}\left(\frac{\mathrm{MAX}^2}{\mathrm{MSE}}\right)$$

- pixel-level reconstruction fidelity
  - based on MSE (mean squared error)
  - MAX: maximum possible pixel value (e.g., 255 or 1)
- higher is better (MSE lower)
- limitation
  - sensitive to small pixel shifts and blur
  - can favor over-smoothed results that look less sharp to humans

Structural Similarity Index (SSIM)

$$\mathrm{SSIM}(x, \hat{x}) = l(x, \hat{x}) \cdot c(x, \hat{x}) \cdot s(x, \hat{x})$$

- similarity of local structure rather than exact pixel match
  - compares luminance, contrast, and structure
- high-level form
  - typically computed on local windows and averaged
  - often correlates with perceived quality better than PSNR for some tasks
- higher is better
- limitation
  - still not a full perceptual metric
  - may not reflect semantic realism in generative outputs

Fréchet Inception Distance (FID)

$$\mathrm{FID} = |\mu_r - \mu_g|_2^2 + \mathrm{Tr}!\left(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}\right)$$

- distance between real and generated distributions in Inception feature space
  - approximates each feature distribution as a Gaussian
- extract features (often from Inception pool3)
  - compute mean (quality) and covariance (diversity)
  - real: (μ_r, Σ_r)
  - generated: (μ_g, Σ_g)
- lower is better
- known failure modes
  - biased for small sample sizes (needs enough samples)

    - can be gamed if the feature extractor is not appropriate for the domain

Inception Score

$$\text{IS} = \exp\left(\mathbb{E}_x[\text{KL}(p(y|x), |, p(y))]\right)$$

$$p(y) = \mathbb{E}_x[p(y|x)]$$

- uses a pretrained classifier's outputs $p(y|x)$
    - confident predictions for each image (low entropy per image)
        - quality proxy: $p(y|x)$ is sharp
    - diverse images across the set (high entropy marginal)
        - diversity proxy: $p(y)$ is broad
- higher is better
- limitations
    - does not compare to the real data distribution directly
    - can be high even when samples do not match the target dataset distribution
    - depends strongly on the classifier label space and domain

Learned Perceptual Image Patch Similarity (LPIPS)

$$\text{LPIPS}(x, \hat{x}) = \sum_l w_l |\hat{\phi}_l(x) - \hat{\phi}_l(\hat{x})|_2^2$$

- perceptual distance using deep features from a pretrained network
    - compares images in feature space rather than pixel space
- definition
    - extract multi-layer features $\phi_l(x)$ and $\phi_l(\hat{x})$
    - compute weighted feature differences across layers
    - $\hat{\phi}$: channel-normalized features
- interpretation
    - lower is better (more perceptually similar)