Variational Auto-Encoder

Yutao Chen

⧖ Mar 23 2026
⧗ Mar 23 2026

Contents

The variational auto-encoder (VAE) (Kingma & Welling, 2013) is:

1. a **generative model** that learns a parametrized distribution $p_\theta(x)$ to match an empirical data distribution $p_{\text{data}}(x)$, and
2. a **latent variable model** that explains observed data $x \sim p_{\text{data}}(x)$ by introducing latent variables $z \sim p_{\text{prior}}(z)$.

Specifically, a variational auto-encoder defines a marginal distribution $p_\theta(x)$ over observed data x as

$$p_\theta(x) = \int p_\theta(x, z)\, \mathrm{d}z,$$

where $p_\theta(x, z)$ is the joint distribution over x and z

$$p_\theta(x, z) = p_\theta(x|z) \cdot p_{\text{prior}}(z).$$

Evidence Lower Bound

For generative modeling, our objective is to match the empirical data distribution by minimizing the KL divergence from $p_{\text{data}}(x)$ to $p_\theta(x)$

$$\begin{aligned}\theta^* &= \arg\min_\theta \mathbb{D}_{\text{KL}}(p_{\text{data}}(x) \parallel p_\theta(x)) \\ &= \arg\max_\theta \mathbb{E}_{p_{\text{data}}(x)}[\log p_\theta(x)].\end{aligned}$$

That is, we want to find parameters θ^* that maximizes the *log-likelihood* $\log p_\theta(x)$ w.r.t. the empirical dataset $p_{\text{data}}(x)$.

However, optimizing $\log_\theta p(x)$ is often computationally intractable due to the integration over z, especially when z is continuous:

$$\log p_\theta(x) = \log \int p_\theta(x|z) p_{\text{prior}}(z) \, \mathrm{d}z.$$

Instead, a common alternative is to optimize a lower bound of $\log p_\theta(x)$, namely the *evidence lower bound* (ELBO):

$$\begin{aligned} \text{ELBO}(q(z)) &= \log p_\theta(x) - \mathbb{D}_{\text{KL}}(q(z) \parallel p_\theta(z|x)) \\ &= \mathbb{E}_{q(z)}[\log p_\theta(x, z) - \log q(z)] \\ &= \mathbb{E}_{q(z)}[\log p_\theta(x|z)] + \mathbb{D}_{\text{KL}}\big(q(z) \parallel p_{\text{prior}}(z)\big), \end{aligned}$$

where $q(z) \in \mathcal{Q}$ is a variational distribution from a variational family $\mathcal{Q}$, and $\text{ELBO}(q(z)) \leq \log p_\theta(x)$ with equality iff $q(z)$ perfectly matches the posterior $p_\theta(z|x)$:

$$q(z) = p_\theta(z|x) \propto p_\theta(x|z) \cdot p_{\text{prior}}(z).$$

Variational auto-encoders are typically parametrized using the encoder and decoder neural networks:

- **Encoder** $q_\phi(z|x) = q(z)$, also known as the *inference network*, maps observed data x to (a distribution over) latent variables z.
- **Decoder** $p_\theta(x|z)$, also known as the *generative network*, maps latent variables z back to (a distribution over) observed data x.

In summary, we optimize parameters $\{\phi, \theta\}$ of variational auto-encoders by maximizing the the evidence lower bound as follows:

$$\phi^*, \theta^* = \arg\max_{\phi,\theta} \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] + \mathbb{D}_{\text{KL}}\big(q_\phi(z|x) \parallel p_{\text{prior}}(z)\big)$$

Examples

Gaussian VAE

Gaussian VAEs assume that $p_{\text{prior}}(z), p_\theta(x|z), q_\phi(z|x)$ are all Gaussian:

- $p_{\text{prior}}(z)$ is typically a standard Gaussian prior $\mathcal{N}(0, I)$;

- $p_\theta(x|z)$ is the Gaussian likelihood $\mathcal{N}(\mu_\theta(z), I)$ with unknown mean $\mu_\theta(z)$ conditioned on z and known unit variance;
- $q_\phi(z|x)$ is the variational distribution $\mathcal{N}\Big(\mu_\phi(x), \text{diag}\Big(\sigma^2_\phi(x)\Big)\Big)$ with unknown mean and *diagonal* covariance conditioned on x.

Recall that our objective is to maximize

$$\text{ELBO}\big(q_\phi(z|x)\big) = \underbrace{\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)]}_{\text{reconstruction}} + \underbrace{\mathbb{D}_{\text{KL}}\big(q_\phi(z|x) \parallel p_{\text{prior}}(z)\big)}_{\text{prior matching}}.$$

- The prior matching term is the KL divergence between two Gaussians, and can be computed in closed form analytically.
- The reconstruction term is approximated with Monte Carlo sampling

$$E_{q_\phi(z|x)}[\log p_\theta(x|z)] \approx \frac{1}{N}\sum_{n=1}^{N} \log p_\theta(x|z_n),$$

  where $z_n \sim q_\phi(z|x)$.

However, Monte Carlo sampling introduces a challenge where gradients of the reconstruction term can not be back-propagated to ϕ as sampling $z \sim q_\phi(z|x)$ is a stochastic operation with no well-defined derivatives. We resort to the *reparametrization trick* for Gaussian distributions:

$$z = \mu_\phi(x) + \sigma^2_\phi(x) \cdot \varepsilon, \text{where } \varepsilon \sim N(0, I).$$

This way, we can sample $z \sim \mathcal{N}\Big(\mu_\phi(x), \sigma^2_\phi(x)\Big)$ and yet delegates the stochasticity to $\varepsilon \sim \mathcal{N}(0, I)$, so that $\frac{\mathrm{d}z}{\mathrm{d}\phi}$ is well-defined.

Hierarchical VAE

Hierarchical VAEs introduce a chain of Markovian latent variable $z_{1:T} = \{z_1, ..., z_T\}$ such that

$$p_\theta(x, z_{1:T}) = p_\theta(x|z_T) \prod_{t=1}^{T-1} p_\theta(z_{t+1}|z_t) p(z_1),$$

and an encoder $q_\phi(z_{1:T}|x)$ with the same Markovian hierarchy

$$q_\phi(z_{1:T}|x) = q_\phi(z_T|x) \prod_{t=1}^{T-1} q_\phi(z_t|z_{t+1}).$$

The evidence lower bound of a hierarchical VAE can then be written as

$$\begin{aligned}
&\mathrm{ELBO}\big(q_\phi(z_{1:T}|x)\big) \\
&= \mathbb{E}_{q_\phi}\big[\log p_\theta(x, z_{1:T}) - \log q_\phi(z_{1:T}|x)\big] \\
&= \mathbb{E}_{q_\phi}\left[\log \frac{p_\theta(x|z_T)\prod_{t=1}^{T-1} p_\theta(z_{t+1}|z_t)p(z_1)}{q_\phi(z_T|x)\prod_{t=1}^{T-1} q_\phi(z_t|z_{t+1})}\right] \\
&= \mathbb{E}_{q_\phi}[\log p_\theta(x|z_T)] - && \text{(reconstruction)} \\
&\quad \mathbb{E}_{q_\phi}\big[\mathbb{D}_{\mathrm{KL}}\big(q_\phi(z_T|x) \parallel p_\theta(z_T|z_{T-1})\big)\big] - \\
&\quad \sum_{t=2}^{T-1} \mathbb{E}_{q_\phi}\big[\mathbb{D}_{\mathrm{KL}}\big(q_\phi(z_t|z_{t+1}) \parallel p_\theta(z_t|z_{t-1})\big)\big] - && \text{(consistency)} \\
&\quad \mathbb{E}_{q_\phi}\big[\mathbb{D}_{\mathrm{KL}}\big(q_\phi(z_1|z_2) \parallel p(z_1)\big)\big]. && \text{(prior matching)}
\end{aligned}$$

While this particular bound appears intimidating, we shall later reveal surprising and yet elegant connections between hierarchical VAE and denoising diffusion probabilistic models (DDPM) (Ho et al., 2020).

References

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems: Vol. 33. Advances in Neural Information Processing Systems.*

Kingma, D. P., & Welling, M. (2013,). Auto-Encoding Variational Bayes. *International Conference on Learning Representations.* https://openreview.net/forum?id=33X9fd2-9FyZd