

A brief introduction to deep generative models

Clayton W. Seitz

February 3, 2022

Outline

Generative Models

References

The logic of generative modeling

Say we have a set of variables $x = (x_1, x_2, \dots, x_n)$ which might have some statistical dependence

The variable x might be an amino acid sequence, gene expression data, microscopy image, etc.

- ▶ Often we are handed a batch of empirical samples $\{x_i\}_{i=1}^N$
- ▶ We want to know the generating distribution $p(x)$

In supervised **generative learning**, we try to explicitly learn the joint distribution $p(x) = \prod_{i=1}^{N-1} p(x_i | x_{i+1:N}) p(x_N)$, which is generally more difficult than discriminative learning.

Perks of generative modeling

- ▶ Fitting complete multivariate distributions $p(\mathbf{x})$ goes beyond correlation-based or clustering approaches
- ▶ Correlations cannot discover partial correlation in the context of other neighbors
- ▶ Fitting $p(\mathbf{x})$ permits sampling based inference

Why generative modeling is difficult

When describing a distribution over multiple variables, we may not know the proper normalization Z . That is,

$$p(x) = \frac{1}{Z} \tilde{p}(x)$$

This **very important** situation arises in several contexts:

1. In **Bayesian inference** where $p(x_1|x_2) = p(x_2|x_1)p(x_1)/p(x_2)$ is intractable due to $Z = p(x_2) = \int p(x_2|x_1)p(x_1)dx_1$. This integral can be very difficult or impossible to compute.
2. In models from statistical physics, e.g. the Ising model, we only know $\tilde{p}(x) = e^{-H(x)}$ where $H(x)$ is the Hamiltonian

Variational autoencoders (VAEs)

A variational solution to generative modeling

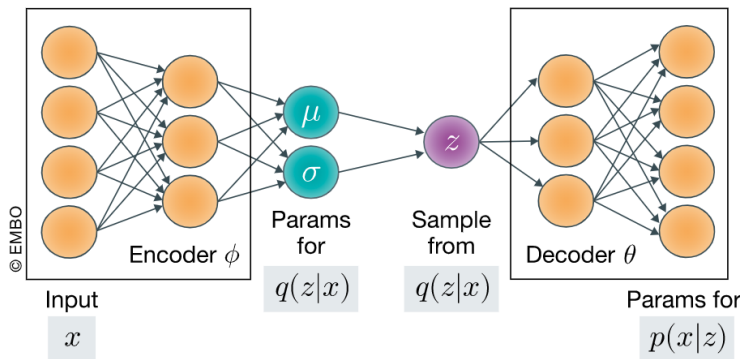


Figure 1: Variational autoencoder architecture Taken from Lopez 2020 in EMBO

Bayesian inference

The variable x has a latent representation or code z . We often say that z is the *causal source* of x . Ultimately, we would like to know the distribution $P_\phi(x)$

$$P_\phi(x) = \frac{P_\phi(x|z)P_\phi(z)}{Q_\psi(z|x)}$$

in order to find the model parameters that maximize the likelihood of the observed data:

$$\Phi^* = \underset{\Phi}{\operatorname{argmin}} -\log P_\Phi(x)$$

but we generally do not know $P_\psi(z|x)$ due to the intractable integral $Z = \int P_\phi(x|z)P_\phi(z)dz$ (see slide 5)

Computing the evidence

We can rewrite the evidence as

$$\begin{aligned}P_{\phi}(x) &= \int P_{\phi}(z)P_{\phi}(x|z)dz \\&= \int P_{\phi}(z)P_{\phi}(x|z)\frac{P_{\phi}(z|x)}{P_{\phi}(z|x)}dz \\&= \mathbb{E}_{z \sim P_{\phi}(z|x)}\frac{P_{\phi}(z)P_{\phi}(x|z)}{P_{\phi}(z|x)}\end{aligned}$$

where $P_{\phi}(z|x)$ is our model "encoder"

The evidence lower bound (ELBO)

$$\begin{aligned}\log P_{\phi}(x) &= \log \int_z P(x, z) dx \\&= \log \int_z P(x, z) \frac{Q(z|x)}{Q(z|x)} dz \\&= \log \mathbb{E}_{z \sim P_{\phi}(z|x)} \frac{P(x|z)P(z)}{Q(z|x)} \\&\geq \mathbb{E}_{z \sim P_{\phi}(z|x)} \log \frac{Q(x|z)}{P(z)} + \log P(x|z) \\-\log P_{\phi}(x) &\leq \mathbb{E}_{z \sim P_{\phi}(z|x)} \log \frac{Q(x|z)}{P(z)} - \log P(x|z)\end{aligned}$$

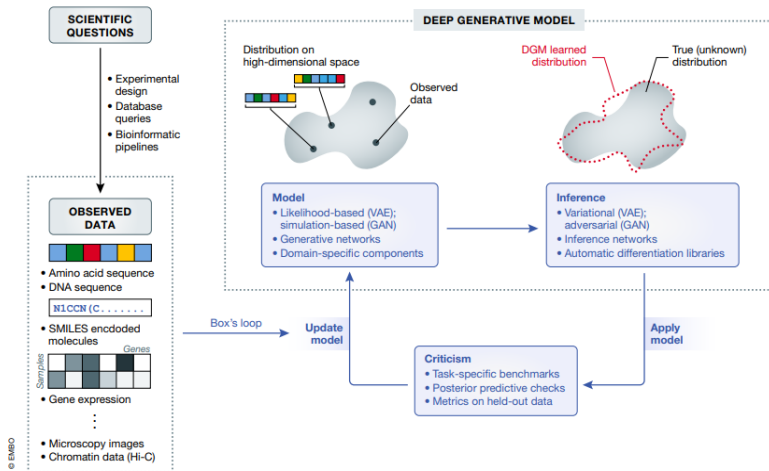
The ELBO objective

$$\begin{aligned}\Phi^* &= \mathcal{L}(\Phi) \\ &= \operatorname{argmin}_{\Phi} \mathbb{E}_{\mathbf{x} \sim \text{Pop}, \mathbf{z} \sim P_{\Phi}(\mathbf{z}|\mathbf{x})} \log \frac{Q_{\Psi}(\mathbf{z}|\mathbf{x})}{P(\mathbf{z})} - \log P(\mathbf{x}|\mathbf{z})\end{aligned}$$

The ELBO can be rewritten in terms of a KL-divergence and population entropy. We often think of Φ as “position” and the loss \mathcal{L} as an “energy”

Stochastic gradient descent

Applying deep generative models to biological data



A recent VAE for transcriptomics data: scV1

Recent improvements of interpretability

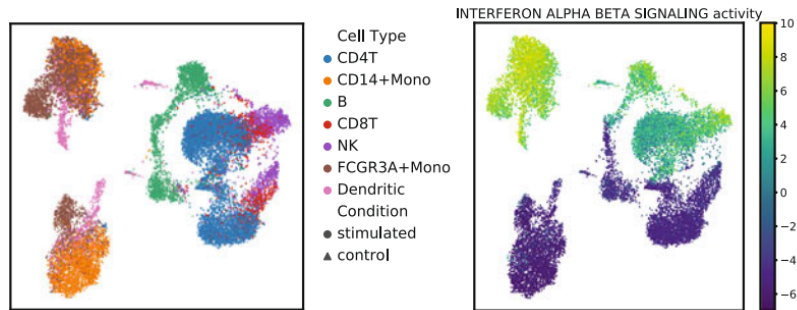


Figure 2: Phenotype segregation using a VAE on single-cell transcriptomics data. Taken from Seninge et al.

Previous studies lack interpretability

Previous studies lack spatial information

Multiplexed RNA imaging

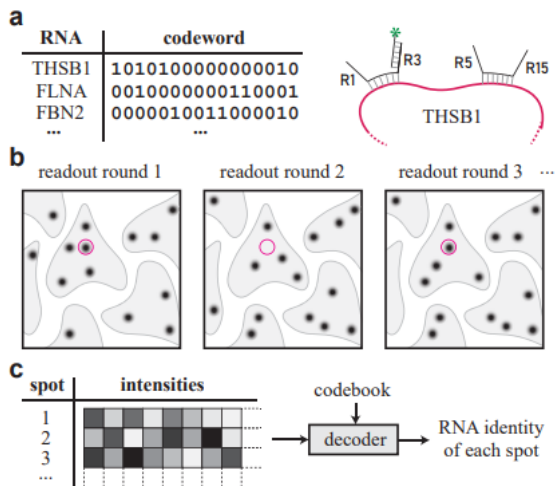


Figure 3:

Minimum Hamming Distance (MHD) codes

Very basic codes that generate codewords $x \in \mathcal{C}$ based on Hamming distance:

$$D = \sum_{n=1}^L \mathbb{I}(x_n = \hat{x}_n)$$

2^L possible code words (with $2^L - 1$ usable ones)

Binary symmetric channel (BSC)

References I