# A very short intro on Variational Autoencoders (VAEs)

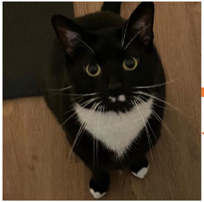
Student Seminar

October 27, 2023

GEORGE MASON UNIVERSITY

# Intro

## What makes an image a 'cat'? What defines 'cat-ness'?



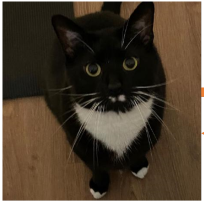
**[256, 256, 3]**

"A picture of a cat"

**A line of words.**

# Intro

## What makes an image a 'cat'? What defines 'cat-ness'?



**[256, 256, 3]**

“A picture of a cat”

**A line of words.**

I aim to offer my personal perspective on understanding Variational Autoencoders (VAEs) by relating them to established statistical frameworks.

# Making connection to Variational Inference (VI)

- Probabilistic models aim to capture the **underlying distribution of the data**, $p(x)$, as well as **relationships between variables**. In many cases, this involves not just observed variables $x$ but also hidden or latent variables $z$.
- **Challenge of inference**: In practice, we're often interested in understanding how the latent variables $z$ behave given some observed data $x$. This is captured by the posterior distribution $p(z \mid x)$.

$$p(z \mid x) = \frac{p(x \mid z) \cdot p(z)}{p(x)}$$

  where
  - $p(x \mid z)$ is the likelihood of observing $x$ given $z$.
  - $p(z)$ is the prior distribution over the latent variables.
  - $p(x)$ is the evidence, which acts as a normalizing constant and is computed as $p(x) = \int p(x, z) \, \mathrm{d}z$.

# Analytical Intractability

The challenge arises because the evidence term $p(x)$ often involves an integral over all possible configurations of $z$ that is analytically intractable. That is:

$$p(x) = \int p(x \mid z) \cdot p(z) \, \mathrm{d}z$$

This integral is computationally expensive or even infeasible to compute directly, especially when $z$ is high-dimensional.

# Variational Inference (VI) as a solution

Variational Inference (VI) offers a solution by turning the intractable inference problem into an **optimization** problem.
We specify a family $\mathcal{Z}$ of densities over the latent variables. Each $q(z) \in \mathcal{Z}$ is a candidate approximation to the conditional. The inference problem is equivalent to solving the following optimization problem

$$q^*(z) = \underset{q(z)\in\mathcal{Z}}{\arg\min}\, \mathrm{D}_{\mathrm{KL}}(q(z) \parallel p(z \mid x)), \tag{1}$$

where

$$\mathrm{D}_{\mathrm{KL}}(q(z) \parallel p(z \mid x)) = \mathbb{E}_q[\log q(z)] - \mathbb{E}_q[\log p(z \mid x)] \tag{2}$$

$$= \mathbb{E}_q[\log q(z)] - \mathbb{E}_q[\log p(z, x)] + \log p(x). \tag{3}$$

# the Evidence Lower BOund (ELBO)

rewrite

$$\mathrm{D}_{\mathrm{KL}}(q(z) \parallel p(z \mid x)) = \underbrace{\mathbb{E}_q[\log q(z)] - \mathbb{E}_q[\log p(z, x)]}_{-\mathsf{ELBO}} + \log p(x)$$

such that

$$\log p(x) = \mathrm{D}_{\mathrm{KL}}(q(z) \parallel p(z \mid x)) + \mathsf{ELBO}(q). \tag{4}$$

We have $\log p(x) \geq \mathsf{ELBO}(q)$ for any $q(z)$.
Maximizing the ELBO is equivalent to minimizing the $\mathrm{KL}$ divergence.

# ELBO

$$\begin{aligned}\text{ELBO}(q) &= \mathbb{E}_q[\log p(z)] + \mathbb{E}_q[\log p(x|z)] - \mathbb{E}[\log q(z)] \\ &= \mathbb{E}[\log \underbrace{p(x|z)}_{\text{observed likelihood}}] - \mathrm{D}_{\mathrm{KL}}(\underbrace{q(z)}_{\text{variational distribution}} \,\|\, \underbrace{p(z)}_{\text{prior distribution}})\end{aligned}$$

We get $q(z)$ via optimization, therefore solving the intractable inference problem in a computationally tractable manner.

# Latent variables in traditional models

- **Role**: The latent variables $z$ simplify complex data distributions into lower-dimensional, interpretable, representations.
- **Example**: Gaussian Mixture Model (GMM)
  1. Latent Variable: Cluster assignment for each data point.
  2. Assumption: Data is generated from multiple Gaussian distributions.
  3. Interpretable: Each cluster can be characterized by its mean and covariance, giving insight into data structure.

# Variational Autoencoder (VAE)

*a*

From a coding theory perspective, the unobserved variables $\mathbf{z}$ have an interpretation as a latent representation or *code*. In this paper we will therefore also refer to the recognition model $q_\phi(\mathbf{z}|\mathbf{x})$ as a probabilistic *encoder*, since given a datapoint $\mathbf{x}$ it produces a distribution (e.g. a Gaussian) over the possible values of the code $\mathbf{z}$ from which the datapoint $\mathbf{x}$ could have been generated. In a similar vein we will refer to $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ as a probabilistic *decoder*, since given a code $\mathbf{z}$ it produces a distribution over the possible corresponding values of $\mathbf{x}$.

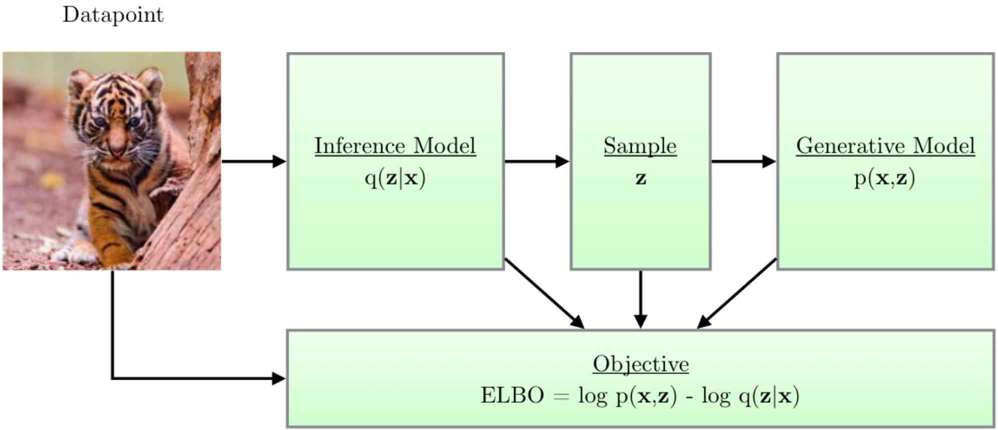
'Equivalent in concept'

- $p(x \mid z; \theta)$: 'decoder' - 'observed likelihood';
- $q(z \mid x; \phi)$: 'encoder' - 'variational distribution'.

[a] Kingma, D. P. and Welling, M. (2022). Auto-Encoding Variational Bayes.

# VAEs

[1]




[1] Kingma, D. P. and Welling, M. (2019). An Introduction to Variational Autoencoders. FNT in Machine

# Neural networks representations

In VAEs, the encoder and decoder can be parameterized by neural networks. For example,

$$p(x \mid z; \theta) = \mathcal{N}(x; \mu(z; \theta), \sigma^2(z; \theta)\boldsymbol{I}), \quad \textbf{decoder}$$

and

$$q(z \mid x; \phi) = \mathcal{N}(z; \mu(x; \phi), \sigma^2(x; \phi)\boldsymbol{I}). \quad \textbf{encoder}$$

These neural networks are trained jointly to maximize the ELBO.

# Making connections

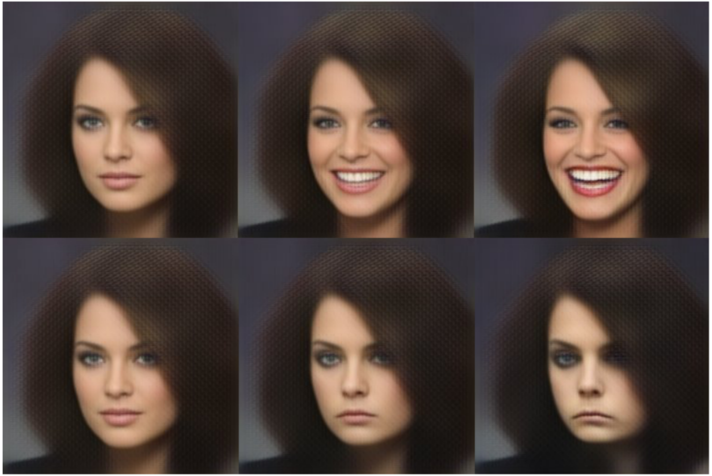

- The EM algorithm can be thought as a special case of VI where we assume $q(z)$ are exactly the posterior distribution $p(z \mid x)$.
- VAEs are flexibly parameterized by neural networks. The 'magic' that can model complex, high-dimensional data, come from the use of neural networks (in my opinion).

# What is the latent space in VAEs?

- The latent space in VAEs is a lower-dimensional representation where each point encodes essential information about a corresponding high-dimensional data point.
- Unlike traditional models where latent variables are based on **explicit model assumptions** (e.g., Gaussian clusters in GMMs), **VAEs learn the latent space from the data**.
- **Opacity**: The latent space in VAEs is often less interpretable than in traditional models. The dimensions do not necessarily have a clear, independent meaning.
- Despite its opacity, this latent space can be highly useful for tasks like image (data) generation, and other unsupervised learning tasks.

# Latent space application

## "ad-hoc" to some extent?



The modification of images in latent space along a 'smile vector' in order to make them look more happy, or more sad looking. [a]

[a] White, T. (2017). SAMPLING GENERATIVE NETWORKS.