TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2024

Variational Auto-Encoders (VAEs)

Fundamental Equations of Deep Learning

- Cross Entropy Loss: $\Phi^* = \mathrm{argmin}_\Phi\, E_{(x,y)\sim \mathrm{Pop}}\left[-\ln P_\Phi(y|x)\right]$.

- GAN: $\mathrm{gen}^* = \mathrm{argmax}_{\mathrm{gen}}\, \min_{\mathrm{disc}}\, E_{i\sim\{-1,1\},y\sim P_i}\left[-\ln P_{\mathrm{disc}}(i|y)\right]$.

- VAE (including diffusion models)

$$\mathrm{pri}^*, \mathrm{dec}^*, \mathrm{enc}^*$$

$$= \operatorname*{argmin}_{\mathrm{pri,dec,enc}}\; E_{y\sim\mathrm{Pop},z\sim P_{\mathrm{enc}}(z|y)}\left[-\ln \frac{P_{\mathrm{pri}}(z)P_{\mathrm{dec}}(y|z)}{P_{\mathrm{enc}}(z|y)}\right]$$

VAEs

A variational autoencoder (VAE) is defined by three parts:

- An encoder distribution $P_{\mathrm{enc}}(z|y)$.
- A decoder distribution $P_{\mathrm{dec}}(y|z)$
- A "prior" distribution $P_{\mathrm{pri}}(z)$

VAE generation uses $P_{\mathrm{pri}}(z)$ and $P_{\mathrm{dec}}(y|z)$.

VAE training uses the encoder $P_{\mathrm{enc}}(z|y)$.

Two Joint Distributions

A VAE defines two joint distributions on y and z, namely $P_{\mathrm{Bayes}}(y,z)$ and $P_{\mathrm{enc}}(y,z)$ defined by

$$P_{\mathrm{Bayes}}(y,z) = P_{\mathrm{pri}}(z)P_{\mathrm{dec}}(y|z)$$

$$P_{\mathrm{enc}}(y,z) = \mathrm{Pop}(y)P_{\mathrm{enc}}(z|y)$$

Training the Bayesian Model

Fix the encoder arbitrarily and train P_{Bayes} by cross entropy.

$$\mathrm{Bayes}^* = \operatorname*{argmin}_{\mathrm{Bayes}} \; E_{(y,z)\sim P_{\mathrm{enc}}(y,z)} \left[- \ln P_{\mathrm{Bayes}}(y, z)\right]$$

Under Universality we have that generating y from P_{Bayes^*} now samples y from Pop.

Training the Encoder

If the Bayes model is not universal then the choice of encoder matters.

$$\mathrm{Pop}(y) \;=\; \frac{\mathrm{Pop}(y)P_{\mathrm{enc}}(z|y)}{P_{\mathrm{enc}}(z|y)} \;=\; \frac{P_{\mathrm{enc}}(y,z)}{P_{\mathrm{enc}}(z|y)}$$

$$H(y) \;\leq\; E_{(y,z)\sim P_{\mathrm{enc}}}\left[-\ln\frac{P_{\mathrm{Bayes}}(y,z)}{P_{\mathrm{enc}}(z|y)}\right]$$

$$\mathrm{enc}^* \;=\; \operatorname*{argmin}_{\mathrm{enc}}\; E_{(y,z)\sim P_{\mathrm{enc}}}\left[-\ln\frac{P_{\mathrm{Bayes}}(y,z)}{P_{\mathrm{enc}}(z|y)}\right]$$

VAEs Evolved from Variational Bayesian Inference

Here y is the evidence about z under the Bayesian model.

$$\begin{aligned}
\ln P_{\mathrm{Bayes}}(y) &= \ln \frac{P_{\mathrm{Bayes}}(y)P_{\mathrm{Bayes}}(z|y)}{P_{\mathrm{Bayes}}(z|y)} \\
&= E_{z\sim P_{\mathrm{enc}}(z|y)}\left[\ln \frac{P_{\mathrm{Bayes}}(y,z)}{P_{\mathrm{Bayes}}(z|y)}\right] \\
&\geq E_{z\sim P_{\mathrm{enc}}(z|y)}\left[\ln \frac{P_{\mathrm{Bayes}(y,z)}}{P_{\mathrm{enc}}(z|y)}\right]
\end{aligned}$$

Here we have replaced a cross-entropy by an entropy.

Variational Bayesian Inference

y is the evidence about z under the Bayesian model.

$$\ln P_{\mathrm{Bayes}}(y) \;\geq\; E_{z \sim P_{\mathrm{enc}}(z|y)} \left[\ln \frac{P_{\mathrm{Bayes}}(y,z)}{P_{\mathrm{enc}}(z|y)}\right]$$

This is the **evidence lower bound** or **ELBO**.

Variational Bayesian Inference

$$\ln P_{\text{Bayes}}(y) \;=\; \ln \frac{P_{\text{Bayes}}(y)P_{\text{Bayes}}(z|y)}{P_{\text{Bayes}}(z|y)}$$

$$\geq\; E_{z\sim P_{\text{enc}}(z|y)}\left[\ln \frac{P_{\text{Bayes}(y,z)}}{P_{\text{enc}}(z|y)}\right]$$

$$\text{enc}^* \;=\; \underset{\text{enc}}{\text{argmax}}\, E_{z\sim P_{\text{enc}}(z|y)}\left[\ln \frac{P_{\text{Bayes}(y,z)}}{P_{\text{enc}}(z|y)}\right] \;=\; P_{\text{Bayes}}(z|y)$$

Expectation Maximization (EM)

EM is used when $P_{\mathrm{enc}}(z|y)$ can be set to $P_{\mathrm{Bayes}}(z|y)$ but $P_{\mathrm{Bayes}}(y,z)$ is highly restricted and cannot express $P_{\mathrm{enc}}(y,z)$.

$$\text{E step: } P^*_{\mathrm{enc}}(z|y) \;=\; P_{\mathrm{Bayes}}(z|y)$$

$$\text{M step: } P^{t+1}_{\mathrm{Bayes}}(y,z) \;=\; \operatorname*{argmin}_{\mathrm{Bayes}} E_{y\sim\mathrm{Train},z\sim P^t_{\mathrm{Bayes}}(z|y)}\left[-\ln P_{\mathrm{Bayes}}(y,z)\right]$$

Difficulties in Training the Encoder

$$\mathrm{enc}^* = \operatorname*{argmin}_{\mathrm{enc}} \; E_{y\sim \mathrm{Pop}(y),\, z\sim P_{\mathrm{enc}}(z|y)} \left[-\ln \frac{P_{\mathrm{Bayes}}(y,z)}{P_{\mathrm{enc}}(z|y)}\right]$$

Gradient descent on the encoder parameters must take into account the fact that we are sampling from the encoder.

Training a sampling distribution typically suffers from **mode collapse** (as in GANs).

The encoder can collapses to a fixed $z = 0$. $P_{\mathrm{dec}}(y|z)$ can always just ignore z. We are then back to standard cross-entropy loss. This is called **posterior collapse**.

Types of VAEs

In **a Gaussian VAE** the we have $P_{\mathrm{pri}}(z)$ and $P_{\mathrm{enc}}(z|y)$ are both Gaussian distributions on R^d. A diffusion model involves a Gaussian VAE at each incremental step of diffusion.

A Vector Quantized VAE (VQ-VAE) defines $P_{\mathrm{enc}}(z|y)$ in terms of vector quantization analogous to K-means clustering. VQ-VAEs provide a translation from continuous data, such as images, to token data that can be modeled with a transformer. This is used in the image understanding abilities of GPT-4o and in autoregressive image generation which is competative with diffusion image generation.

We will first consider Gaussian VAEs and discuss VQ-VAEs later.

Gaussian VAEs

As an example take

$$P_{\mathrm{pri}}(z) = \mathcal{N}(0, I)$$

$$P_{\mathrm{enc}}(z|y) = \mathcal{N}(\hat{z}(y), I)$$

$$P_{\mathrm{dec}}(y|z) = \mathcal{N}(\hat{y}(z), I)$$

In general we can use arbitrary Gaussians but this example makes the math simple.

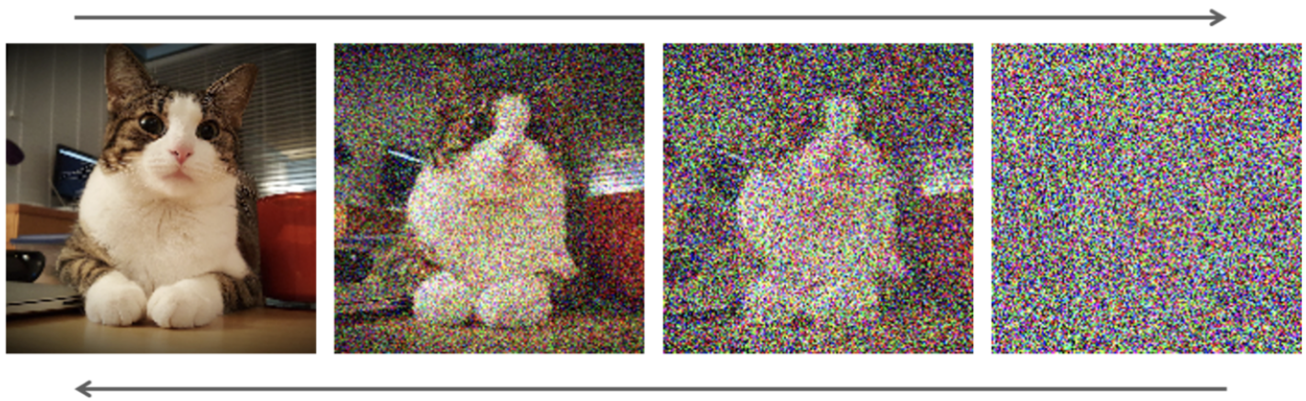
Gaussian VAEs

$$E_{(y,z)\sim P_{\mathrm{enc}}}\;\left[-\ln\frac{P_{\mathrm{pri}}(z)P_{\mathrm{dec}}(y|z)}{P_{\mathrm{enc}}(z|y)}\right]$$

$$=\;E_{y\sim\mathrm{Pop}}\left[KL(P_{\mathrm{enc}}(z|y),P_{\mathrm{pri}}(z))+E_{z\sim P_{\mathrm{enc}}(z|y)}\left[-\ln P_{\mathrm{dec}}(y|z)\right]\right]$$

$$=\;E_{y\sim\mathrm{Pop}}\left[\frac{1}{2}||\hat{z}_{\mathrm{enc}}(y)||^2+E_{\epsilon}\left[\frac{1}{2}||y-\hat{y}_{\mathrm{dec}}(\hat{z}_{\mathrm{enc}}(y)+\epsilon))||^2\right]\right]$$

Hierarchical VAEs



[Sally talked to John] $\rightleftarrows$ [Sally talked to] $\rightleftarrows$ [Sally talked] $\rightleftarrows$ [Sally] $\rightleftarrows$ []

$$y \rightleftarrows z_1 \rightleftarrows \cdots \rightleftarrows z_N$$

Hierarchical VAEs

$$y \overrightarrow{\leftarrow} z_1 \overrightarrow{\leftarrow} \cdots \overrightarrow{\leftarrow} z_N$$

Encoder: $\mathrm{Pop}(y)$, $P_{\mathrm{enc}}(z_1|y)$, and $P_{\mathrm{enc}}(z_{\ell+1}|z_\ell)$.

Generator: $P_{\mathrm{pri}}(z_N)$, $P_{\mathrm{dec}}(z_{\ell-1}|z_\ell)$, $P_{\mathrm{dec}}(y|z_1)$.

The encoder and the decoder define distributions $P_{\mathrm{enc}}(y,\ldots,z_N)$ and $P_{\mathrm{dec}}(y,\ldots,z_N)$ respectively.

Hierarchical VAEs

$$y \rightleftarrows z_1 \rightleftarrows \cdots \rightleftarrows z_N$$

- autoregressive models

- diffusion models

Hierarchical (or Diffusion) ELBO

$$
\begin{aligned}
H(y) &= E_{\mathrm{enc}}\left[-\ln \frac{P_{\mathrm{enc}}(y)P_{\mathrm{enc}}(z_1,\ldots,z_N|y)}{P_{\mathrm{enc}}(z_1,\ldots,z_N|y)}\right] \\
&= E_{\mathrm{enc}}\left[-\ln \frac{P_{\mathrm{enc}}(y|z_1)P_{\mathrm{enc}}(z_1|z_2)\cdots P_{\mathrm{enc}}(z_{N-1}|z_N)P_{\mathrm{enc}}(z_N)}{P_{\mathrm{enc}}(z_1|z_2,y)\cdots P_{\mathrm{enc}}(z_{N-1}|z_N,y)P_{\mathrm{enc}}(z_N|y)}\right] \\
&\leq E_{\mathrm{enc}}\left[-\ln \frac{P_{\mathrm{dec}}(y|z_1)P_{\mathrm{dec}}(z_1|z_2)\cdots P_{\mathrm{dec}}(z_{N-1}|z_N)P_{\mathrm{dec}}(z_N)}{P_{\mathrm{enc}}(z_1|z_2,y)\cdots P_{\mathrm{enc}}(z_{N-1}|z_N,y)P_{\mathrm{enc}}(z_N|y)}\right] \\
&= \left\{\begin{array}{l}
E_{\mathrm{enc}}\left[-\ln P_{\mathrm{dec}}(y|z_1)\right] \\
\\
+\sum_{i=2}^{N}\ E_{\mathrm{enc}}\ KL(P_{\mathrm{enc}}(z_{i-1}|z_i,y),\ P_{\mathrm{dec}}(z_{i-1}|z_i)) \\
\\
+E_{\mathrm{enc}}\ KL(P_{\mathrm{enc}}(Z_N|y), p_{\mathrm{dec}}(Z_N))
\end{array}\right.
\end{aligned}
$$

END