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)

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

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

VAEs

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

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

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

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

Fixing an Arbitray Encoder

Fix the encoder arbitrarily and train P_{pri} and P_{dec} by cross entropy loss.

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

Universality gives

$$P_{\mathrm{pri}^*}(z) P_{\mathrm{dec}^*}(y|z) = \mathrm{Pop}(y) P_{\mathrm{enc}}(z|y)$$

Sampling from $P_{\mathrm{pri}^*}(z) P_{\mathrm{dec}^*}(y|z)$ now samples y from the population.

Training the Encoder — the ELBO

In practice the choice of encoder matters.

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

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

The inequality follows from the fact that cross-entropy (using the models P_{pri} and P_{dec}) upper bounds entropy.

This upper bound on $H(y)$ is called **the ELBO** (Acronym described later).

Difficulties in Training the Encoder

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

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).

Often the encoder collapses to fixing $z = 0$. $P_{\mathrm{dec}}(y|z)$ 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 done in Chat-GPT 4o and other multi-modal language models.

We will consider each these approaches.

Gaussian VAEs

We sample noise ϵ from a Gaussian distribution on R^d.

$$\mathrm{enc}^* = \operatorname*{argmin}_{\mathrm{enc}} \; E_{y,\epsilon\sim\mathcal{N}(0,\sigma I)} \left[-\ln \frac{P_{\mathrm{pri}}(z)P_{\mathrm{dec}}(y|z)}{P_{\mathrm{enc}}(z|y)}\right] \quad z = \hat{z}(y) + \epsilon$$

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

$$= \operatorname*{argmin}_{\mathrm{enc}} \; \frac{||\hat{z}_{\mathrm{enc}}(y) - \hat{z}_{\mathrm{pri}}||^2}{2\sigma^2} + E_\epsilon \, \frac{||y - \hat{y}_{\mathrm{dec}}(\hat{z}_{\mathrm{enc}}(y,\epsilon))||^2}{2\sigma^2}$$

A closed-form expression for the KL term avoids sampling noise.

VAEs Evolved from Variational Bayesian Inference

$P_{\mathrm{pri}}(z)$ is interpreted as the Bayesian prior on “hypothesis” z.

$P_{\mathrm{dec}}(y|z)$ is interpreted as the propability of the “evidence” y given hypothesis z.

We consider the Bayesian distribution defined by $P_{\mathrm{pri}}(z)$ and $P_{\mathrm{dec}}(y|z)$ and we want to compute $P_{\mathrm{pri,dec}}(z|y)$ under this Bayesian distribution. I will write this $P_{\mathrm{Bayes}}(z|y)$

$P_{\mathrm{enc}}(z|y)$ is interpreted as an approximation for $P_{\mathrm{pri,dec}}(z|y)$.

Bayesian Interpretation

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

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

Bayesian Interpretation

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

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

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

Expectation Maximization (EM)

Expectation Maximimization (EM) applies in the (highly special) case where the exact posterior $P_{\text{pri,dec}}(z|y)$ is samplable and computable. EM alternates exact optimization of enc and the pair (pri, dec) in:

$$\text{VAE:} \qquad \text{pri}^*, \text{dec}^* = \operatorname*{argmin}_{\text{pri,dec}} \min_{\text{enc}} E_{y,\; z \sim P_{\text{enc}}(z|y)} \; -\ln \frac{P_{\text{pri,dec}}(z,y)}{P_{\text{enc}}(z|y)}$$

$$\text{EM:} \qquad \text{pri}^{t+1}, \text{dec}^{t+1} = \operatorname*{argmin}_{\text{pri,dec}} \quad E_{y,\; z \sim P_{\text{pri}^t,\text{dec}^t}(z|y)} \; -\ln P_{\text{pri,dec}}(z,y)$$

Inference	Update
(E Step)	(M Step)
$P_{\text{enc}}(z\vert y) = P_{\text{pri}^t,\text{dec}^t}(z\vert y)$	Hold $P_{\text{enc}}(z\vert y)$ fixed

slidePosterior Collapse

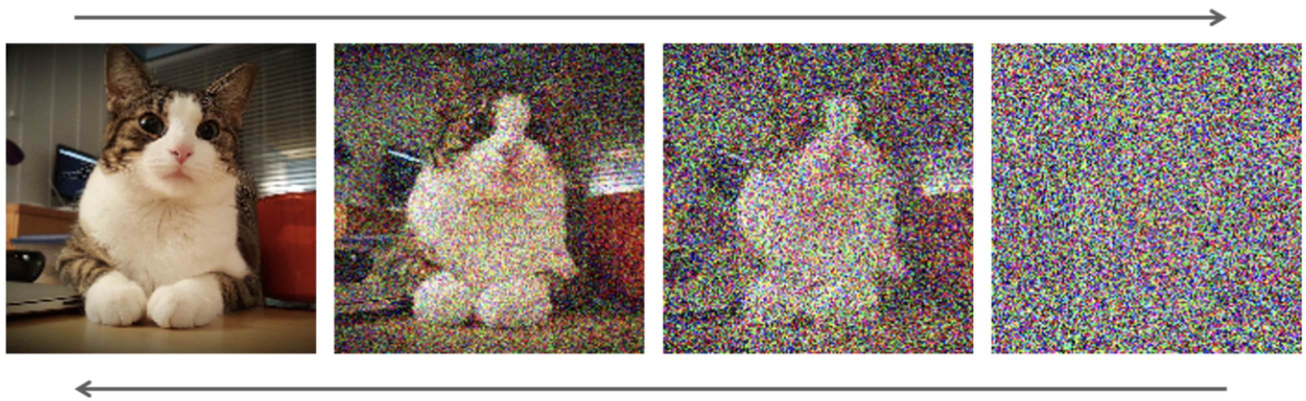
$$P_{\mathrm{pri}}(z)P_{\mathrm{dec}}(y|z) = \mathrm{Pop}(y)P_{\mathrm{enc}}(z|y)$$

Any joint distribution on (y, z) with the desired marginal on y optimizes the bound.

This allows the prior and the encoder (the posterior) to both degenerate to having no mutual information with y.

This often happens in language modeling.

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

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

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

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

END