TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2023

Variational Auto-Encoders (VAEs)

Fundamental Equations of Deep Learning

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

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

- VAE (including diffusion models)

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

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

Generative AI for Continuous Data: VAEs

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

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

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

VAE training uses a "GAN inverter" $P_{\text{enc}}(z|y)$.

Fixed Encoder Training

$$\text{pri}^*, \text{gen}^* = \underset{\text{pri,gen}}{\text{argmin}} \; E_{y\sim \text{Pop}(y), z\sim \text{enc}(z|y)} \left[-\ln P_{\text{pri}}(z) P_{\text{gen}}(y|z) \right]$$

This is cross-entropy loss from $\text{Pop}(y) P_{\text{enc}}(z|y)$ to $P_{\text{pri}}(z) P_{\text{gen}}(y|z)$

Universality gives

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

Hence sampling from $P_{\text{pri}^*}(z) P_{\text{gen}^*}(y|z)$ samples y from the population.

Degrees of Freedom

$$P_{\text{pri}}(z)P_{\text{gen}}(y|z) = \text{Pop}(y)P_{\text{enc}}(z|y)$$

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

Bayesian Encoder Training

We consider the case of a probabilitic model with a small number of parameters (by deep learning standards).

For example a Gaussian mixture model (GMM) or a probabilistic context free grammar (PCFG).

Such models impose a strong structural constraint and are far from universal.

For such models we clearly need to train the encoder. More generally, training the encoder can improve the model (reduce an upper bound on $H(y)$).

Training the Encoder (the GAN Inverter)

Define the ELBO loss as follows (acronym described later).

$$\mathcal{L}(y,z) \;=\; -\ln \frac{P_{\text{pri}}(z)P_{\text{gen}}(y|z)}{P_{\text{enc}}(z|y)}$$

Recall cross-entropy loss: $H(y) \leq E_{y\sim \text{Pop}}\left[-\ln P_\Phi(y)\right]$

We will show $H(y) \leq E_{y\sim \text{Pop}, z\sim P_{\text{enc}}(z|y)}\, \mathcal{L}(y,z)$

A Bayesian Interpretation

VAEs were originally motivated by a Bayesian interpretation:

- $P_{\text{pri}}(z)$ is the Bayesian prior on hypothesis z.
- $P_{\text{gen}}(y|z)$ is the propability of the "evidence" y given hypothesis z.
- $P_{\text{enc}}(z|y)$ is a model approximating the Bayesian posterior on hypothesis z given evidence y.

The Bayesian motivation is to train $P_{\text{enc}}(z|y)$ to approximate Bayesian inference.

Bayesian Interpretation

$$
\begin{aligned}
H(\text{Pop}) &\leq E_{y\sim\text{Pop}}\left[-\ln P_{\text{pri,gen}}(y)\right] \\
\ln P_{\text{pri,gen}}(y) &= \ln \frac{P_{\text{pri}}(z)P_{\text{gen}}(y|z)}{P_{\text{pri,gen}}(z|y)} \\
&= E_{z\sim P_{\text{enc}}(z|y)}\left[\ln \frac{P_{\text{pri}}(z)P_{\text{gen}}(y|z)}{P_{\text{enc}}(z|y)}\right] + KL(P_{\text{enc}}(z|y), P_{\text{pri,gen}}(z|y)) \\
&\geq E_{z\sim P_{\text{enc}}(z|y)}\left[\ln \frac{P_{\text{pri}}(z)P_{\text{gen}}(y|z)}{P_{\text{enc}}(z|y)}\right]
\end{aligned}
$$

A Bayesian thinks of y as "evidence" for hypothesis z.

$E_{z\sim P_{\text{enc}}(z|y)}[-\mathcal{L}(y,z)]$ is called the evidence lower bound (ELBO).

Expectation Maximization (EM)

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

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

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

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

Posterior Collapse

$$P_{\text{pri}}(z)P_{\text{gen}}(y|z) = \text{Pop}(y)P_{\text{enc}}(z|y)$$

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.

The Reparameterization Trick

$$\text{enc}^* = \underset{\text{enc}}{\text{argmin}} \quad E_{y\sim \text{Pop}(y), z\sim P_{\text{enc}}(z|y)} \left[-\ln \frac{P_{\text{pri}}(z) P_{\text{gen}}(y|z)}{P_{\text{enc}}(z|y)} \right]$$

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

To handle this we sample noise ϵ from a fixed noise distribution and replace z with a determinstc function $z_{\text{enc}}(y, \epsilon)$

$$\text{enc}^*, \text{pri}^*, \text{gen}^* = \underset{\text{enc,pri,gen}}{\text{argmin}} \quad E_{y, \epsilon, z=\hat{z}_{\text{enc}}(y,\epsilon)} \left[-\ln \frac{P_{\text{pri}}(z) P_{\text{gen}}(y|z)}{P_{\text{enc}}(z|y)} \right]$$

The Reparameterization Trick

$$\text{enc}^*, \text{pri}^*, \text{gen}^* = \underset{\text{enc,pri,gen}}{\text{argmin}} \quad E_{y,\epsilon,z=\hat{z}_{\text{enc}}(y,\epsilon)} \left[-\ln \frac{P_{\text{pri}}(z) P_{\text{gen}}(y|z)}{P_{\text{enc}}(z|y)} \right]$$

To get gradients we must have that $\hat{z}_{\text{enc}}(y, \epsilon)$ is a differentiable function of the encoder parameters.

Optimizing the encoder is tricky for discrete z. Discrete z is handled effectively in EM algorithms and general vector quantization (VQ) methods.

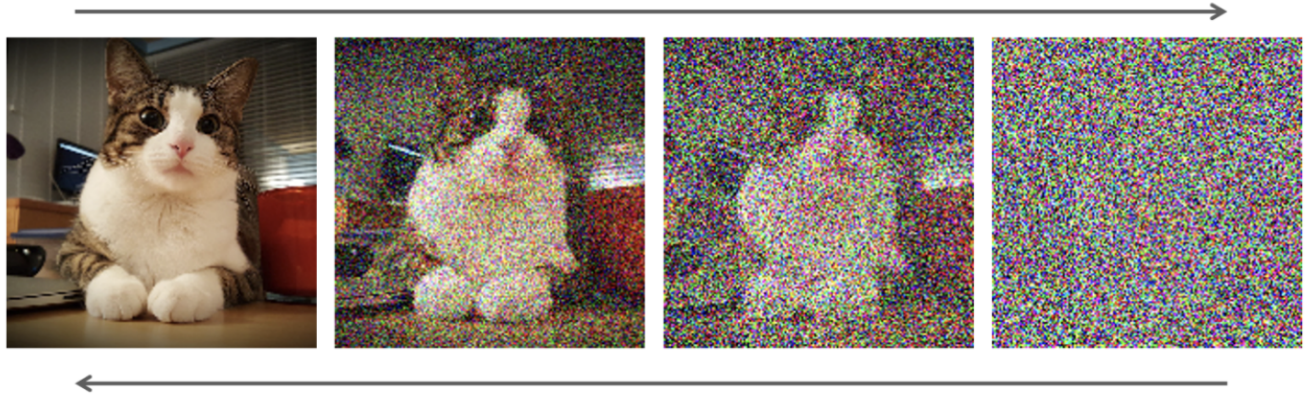
The KL-divergence Optimization

For Gaussian Models we have

$$\begin{aligned}
\mathcal{L}(y) &= E_{z\sim P_{\text{enc}}(z|y)}\left[-\ln \frac{P_{\text{pri}}(z)P_{\text{gen}}(y|z)}{P_{\text{enc}}(z|y)}\right] \\
&= KL(P_{\text{enc}}(z|y), P_{\text{pri}}(z)) + E_{z\sim P_{\text{enc}}(z|y)}\left[-\ln P_{\text{gen}}(y|z)\right] \\
&= \frac{||\hat{z}_{\text{enc}}(y) - \hat{z}_{\text{pri}}||^2}{2\sigma^2} + E_\epsilon \frac{||y - \hat{y}_{\text{gen}}(\hat{z}_{\text{enc}}(y,\epsilon))||^2}{2\sigma^2}
\end{aligned}$$

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

Hierarchical VAEs



[Sally talked to John] $\overrightarrow{\leftarrow}$ [Sally talked to] $\overrightarrow{\leftarrow}$ [Sally talked] $\overrightarrow{\leftarrow}$ [Sally] $\overrightarrow{\leftarrow}$ []

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

Hierarchical VAEs

$$y \overset{\rightarrow}{\leftarrow} z_1 \overset{\rightarrow}{\leftarrow} \cdots \overset{\rightarrow}{\leftarrow} z_N$$

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

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

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

Hierarchical VAEs

$$y \overset{\rightarrow}{\leftarrow} z_1 \overset{\rightarrow}{\leftarrow} \cdots \overset{\rightarrow}{\leftarrow} z_N$$

- autoregressive models

- diffusion models

Hierarchical (or Diffusion) ELBO

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

END