TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2022

Variational Auto-Encoders (VAEs)

Image Compression and Image Generation

Suppose that we want to model a population distribution on y, for example the distribution of "natural images".

Shannon's source coding theorem implies that there exists a coding function with an inverse decoding function such that decoding a random string samples an image from the population distribution on images.

While we cannot optimally compress images, it can be useful to represent a population distribution on y in terms of a latent (unabserved) variable $z(y)$ loosely analogous to a compressed form.

The Encoder, Decoder and the Prior

Consider a probabilistic encoder algorithm $P_{\mathrm{enc}}(z|y)$ — perhaps a stochastic image compression algorithm.

The encoder $P_{\mathrm{enc}}(z|y)$ defines a joint probability distribution on pairs (y, z).

We will also introduce a decoder model (decompressor) $P_{\mathrm{dec}}(y|z)$ and a prior probability model $P_{\mathrm{pri}}(z)$ which are to be trained using a cross-entropy loss to $P(y|z)$ and and $P(z)$ as defined by the encoder.

The ELBO

$$
\begin{aligned}
H(y,z) &= H(y) + H(z|y) \;=\; H(z) + H(y|z) \\
H(y) &= H(z) + H(y|z) - H(z|y) \\
&\leq CE(P(z), P_{\mathrm{pri}}(z)) + CE(P(y|z), P_{\mathrm{dec}}(y|z)) - H_{\mathrm{enc}}(z|y) \\
&= E_{y\sim\mathrm{Pop},z\sim P_{\mathrm{enc}}(z|y)} \; -\ln \frac{P_{\mathrm{pri}}(z)P_{\mathrm{dec}}(y|z)}{P_{\mathrm{enc}}(z|y)}
\end{aligned}
$$

The last line is the (negative) ELBO. ELBO stands for "Evidence Lower Bound" but this terminaology is obscure and unhelpful.

The ELBO Loss

We now interpret the ELBO as a loss on a given value y.

$$
\begin{aligned}
H(y) \;&\leq\; E_{y\sim\mathrm{Pop},z\sim P_{\mathrm{enc}}(z|y)} \;-\ln\frac{P_{\mathrm{pri}}(z)P_{\mathrm{dec}}(y|z)}{P_{\mathrm{enc}}(z|y)} \\
&=\; E_{y\sim\mathrm{Pop}}\;\mathcal{L}_E(y)
\end{aligned}
$$

$$
\mathcal{L}_E(y) \;=\; E_{z\sim P_{\mathrm{enc}}(z|y)} \;-\ln\frac{P_{\mathrm{pri}}(z)P_{\mathrm{dec}}(y|z)}{P_{\mathrm{enc}}(z|y)}
$$

A Third Fundamental Equation

$$\mathrm{pri}^*, \mathrm{dec}^* = \underset{\mathrm{pri,dec}}{\mathrm{argmin}} \quad E_y\ \mathcal{L}_E(y)$$

$$E_y\ \mathcal{L}_E(y)$$

$$= \ E_{y\sim\mathrm{Pop},\ z\sim P_{\mathrm{enc}}(z|y)} \quad \ln \frac{P_{\mathrm{enc}}(z|y)}{P_{\mathrm{pri}}(z)P_{\mathrm{dec}}(y|z)}$$

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

$$= \ H(y) + KL(\mathrm{Pop}(y)P_{\mathrm{enc}}(z|y), \quad P_{\mathrm{pri}}(z)P_{\mathrm{dec}}(y|z))$$

The ELBO Loss

$$E_y \, \mathcal{L}_E(y) \; = \; H(y) + KL(\mathrm{Pop}(y) P_{\mathrm{enc}}(z|y), \;\; P_{\mathrm{pri}}(z) P_{\mathrm{dec}}(y|z))$$

Minimization occurs when the prior and the docoder satisfy

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

The ELBO Loss

$$E_y \, \mathcal{L}_E(y) \; = \; H(y) + KL(\mathrm{Pop}(y) P_{\mathrm{enc}}(z|y), \;\; P_{\mathrm{pri}}(z) P_{\mathrm{dec}}(y|z))$$

Minimizing gives

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

and hence

$$\mathcal{L}_E(y) = \;\; E_z \; \ln \frac{P(z|y)}{P(z,y)} = E_z \; \ln \frac{P(z,y)/P(y)}{P(z,y)} \;\; = -\ln \mathrm{Pop}(y)$$

After optimization one can interpret $\mathcal{L}_E(y)$ as $-\ln \mathrm{Pop}(y)$.

Optimizing the Encoder

Although in princile the encoder need not be trained, it is sometimes jointly optimized with the prior and the decoder.

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

This is necessary if want to interpret z as some kind of "understanding" of the distribution on y that facilitates representing the prior and the decoder.

Optimizing the Encoder

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

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

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

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

The Re-Parameterization Trick

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

To get gradients we must have that $z_{\mathrm{enc}}(y, \epsilon)$ is a smooth function of the encoder parameters and all probabilties must be a smooth function of z.

Most commonly $\epsilon \in R^d$ with $\epsilon \sim \mathcal{N}(0, I)$ and

$$z^i_{\mathrm{enc}}(y, \epsilon) = \hat{z}^i_{\mathrm{enc}}(y) + \sigma^i \epsilon^i.$$

Optimizing the encoder is tricky for discrete z. Discrete z is handled effectively in EM algorithms and in VQ-VAEs.

EM is Alternating Optimization of the ELBO Loss

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

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

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

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

END