TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2022

Variational Auto-Encoders (VAEs)

Rate-Distortion Autoencoders

Consider image compression where we compress an image y into a compressed file z.

We will assume a stochastic compression algorithm which we will call the "encoder" $P_{\mathrm{enc}}(z|y)$.

The number of bits needed for the compressed file is given by $H(z)$. $H(z)$ is the "rate" (bits per image) for transmitting compressed images.

The number of unknown additional bits needed to exactly recover y is $H(y|z)$. $H(y|z)$ is a measure of the "distortion" of y when y is decoded without the missing bits.

Rate-Distortion Autoencoders

In practice we model $H(z)$ with a "prior model" $P_{\mathrm{pri}}(z)$ and model $H(y|z)$ with a "decoder model" $P_{\mathrm{dec}}(y|z)$.

So the rate-distortion auto-encoder has three parts $P_{\mathrm{enc}}(z|y)$, $P_{\mathrm{pri}}(z)$, and $P_{\mathrm{dec}}(y|z)$.

The **variational autoencoder (VAE)** with latent variable z is mathematically the same as a rate-distortion autoencoder with compressed form z.

An "Encoder First" Treatment of VAEs

Fix an arbitrary encoder model $P_{\mathrm{enc}}(z|y)$.

For $y \sim \mathrm{Pop}$ and $z \sim P_{\mathrm{enc}}(z|y)$ train models pri and dec.

$$\text{Prior Model: } \mathrm{pri}^* = \underset{\mathrm{pri}}{\mathrm{argmin}} \; E_{y,z} \; -\ln P_{\mathrm{pri}}(z)$$

$$\text{Decoder Model: } \mathrm{dec}^* = \underset{\mathrm{dec}}{\mathrm{argmin}} \; E_{y,z} \; -\ln P_{\mathrm{dec}}(y|z)$$

For any $P_{\mathrm{enc}}(z|y)$ the universality assumption for pri* and dec* gives

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

Upper Bounding $H(y)$

Cross-entropy upper bounds $H(y)$ and equals $H(y)$ assuming universality.

The ELBO plays the role of cross-entropy for latent variable models.

The negative ELBO uppr bounds $H(y)$ and equals $H(y)$ assuming universality.

Deriving the ELBO

Sample $y \sim \mathrm{Pop}$ and $z \sim P_{\mathrm{enc}}(z|y)$.

$$H(y,z) = H(y) + H(z|y) = H(z) + H(y|z)$$

Solving for $H(y)$ gives

$$H(y) = H(z) + H(y|z) - H(z|y)$$

An Encoder First Treatment of VAEs

Replace the first two entropies by cross entropies.

$$
\begin{aligned}
H(y) &= H(z) + H(y|z) - H_{\mathrm{enc}}(z|y) \\
&\leq H_{\mathrm{pri}}(z) + H_{\mathrm{dec}}(y|z) - H_{\mathrm{enc}}(z|y) \\
H_{\mathrm{pri}}(z) &= E_{y,z}\left[-\ln P_{\mathrm{pri}}(z)\right] \\
H_{\mathrm{dec}}(y|z) &= E_{y,z}\left[-P_{\mathrm{dec}}(y|z)\right]
\end{aligned}
$$

VAE

$$
\begin{aligned}
H(y) &= H(z) + H(y|z) - H_{\mathrm{enc}}(z|y) \\
&\leq H_{\mathrm{pri}}(z) + H_{\mathrm{dec}}(y|z) - H_{\mathrm{enc}}(z|y) \\
&= E_{y,z}\ \left[-\ln \frac{P_{\mathrm{pri}}(z)P_{\mathrm{dec}}(y|z)}{P_{\mathrm{enc}}(z|y)}\right]
\end{aligned}
$$

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

The Re-Parameterization Trick

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

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

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

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 VAE

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 $(\mathrm{pri}, \mathrm{dec})$ in:

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

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

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