TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2022

Variational Auto-Encoders (VAEs)

Meaningful Latent Variables: Learning Phonemes and Words

A child exposed to speech sounds learns to distinguish phonemes and then words.

The phonemes and words are "latent variables" learned from listening to sounds.

We will use y for the raw input (sound waves) and z for the latent variables (phonemes).

Other Examples

z might be a parse tree, or some other semantic representation, for an observable sentence (word string) y.

z might be a segmentation of an image y.

z might be a depth map (or 3D representation) of an image y.

z might be a class label for an image y.

Here we are interested in the case where z is **latent** in the sense that we do not have training labels for z.

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" $P_{\text{enc}}(z|y)$.

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" $P_{\text{pri}}(z)$ and model $H(y|z)$ with a "decoder model" $P_{\text{dec}}(y|z)$.

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

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 $P_{\text{enc}}(z|y)$.

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

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

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

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

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

An Encoder First Treatment of VAEs

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

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

This encoder first formulation will be particularly relevant to the diffusion models underlying DALL·E-2.

An Encoder First Treatment of VAEs

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

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

Solving for $H(y)$ gives

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

An Encoder First Treatment of VAEs

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

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

VAE

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

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

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

VAE

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

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

Under the universality assumption for the prior and the decoder we do not need to optimize the encoder.

Before the deep revolution the structure of the prior and the decoder was typically highly restricted. In such cases we are far from universality and we need to optimize the encoder. This is related to EM as discussed below.

The ELBO

$$
\begin{aligned}
P_{\text{pri,dec}}(y) &= \sum_{z} P_{\text{pri}}(z) P_{\text{dec}}(y|z) \\
&= E_{z \sim P_{\text{pri}}(z)} \quad P_{\text{dec}}(y|z)
\end{aligned}
$$

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

The ELBO

We will show

$$\text{ELBO}: \quad \ln \; P_{\text{pri,dec}}(y) \geq E_{z \sim P_{\text{enc}}(z|y)} \; \left[\ln \frac{P_{\text{pri}}(z) P_{\text{dec}}(y|z)}{P_{\text{enc}}(z|y)} \right]$$

The left hand side is the log-probability of the evidence y and the right hand side is the **evidence lower bound** or ELBO.

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

The negative ELBO is the loss function of the VAE.

Deriving the ELBO

But even when $P_{\text{pri}}(z)$ and $P_{\text{dec}}(y|z)$ are samplable, if z is a structured value we cannot typically compute $P_{\text{pri,dec}}(y)$.

$$P_{\text{pri,dec}}(y) = \sum_z P_{\text{pri}}(z) P_{\text{dec}}(y|z) = E_{z \sim P_{\text{pri}}(z)} P_{\text{dec}}(y|z)$$

The sum is too large and sampling z from $P_{\text{pri}}(z)$ is unlikely to sample the values that dominate the sum.

Deriving the ELBO

A much better estimate could be achieved by importance sampling — sampling z from the posterior $P_{\text{pri,dec}}(z|y)$.

$$
\begin{aligned}
P_{\text{pri,dec}}(y) &= \sum_z P_{\text{pri}}(z) P_{\text{dec}}(y|z) \\
&= \sum_z P_{\text{pri,dec}}(z|y) \, \frac{P_{\text{pri}}(z) P_{\text{dec}}(y|z)}{P_{\text{pri,dec}}(z|y)} \\
&= E_{z \sim P_{\text{pri,dec}}(z|y)} \quad \frac{P_{\text{pri}}(z) P_{\text{dec}}(y|z)}{P_{\text{pri,dec}}(z|y)}
\end{aligned}
$$

Deriving the ELBO

$$P_{\text{pri,dec}}(y) = E_{z \sim P_{\text{pri,dec}}(z|y)} \frac{P_{\text{pri}}(z) P_{\text{dec}}(y|z)}{P_{\text{pri,dec}}(z|y)}$$

Unfortunately the conditional distribution $P_{\text{pri,dec}}(z|y)$ also cannot be computed or sampled from.

Variational Bayes side-steps the intractability problem by introducing another model component — a model $P_{\text{enc}}(z|y)$ to approximate the intractible $P_{\text{pri,dec}}(z|y)$.

The Evidence Lower Bound (The ELBO)

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

$$= E_{z\sim P_{\text{enc}}(z|y)} \left(\ln \frac{P_{\text{pri,dec}}(z,y)}{P_{\text{enc}}(z|y)} + \ln \frac{P_{\text{enc}}(z|y)}{P_{\text{pri,dec}}(z|y)} \right)$$

$$= \left(E_{z\sim P_{\text{enc}}(z|y)} \ln \frac{P_{\text{pri,dec}}(z,y)}{P_{\text{enc}}(z|y)} \right) + KL(P_{\text{enc}}(z|y), P_{\text{pri,dec}}(z|y))$$

$$\geq E_{z\sim P_{\text{enc}}(z|y)} \ln \frac{P_{\text{pri,dec}}(z,y)}{P_{\text{enc}}(z|y)} \quad \text{The ELBO}$$

The Re-Parameterization Trick

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

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 $z_{\text{enc}}(y, \epsilon)$

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

The Re-Parameterization Trick

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

To get gradients we must have that $z_{\text{enc}}(y, \epsilon)$ 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

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

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 $P_{\text{pri,dec}}(z|y)$ is samplable and computable. EM alternates exact optimization of enc and the pair (pri, dec) in:

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

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

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

END