TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2021

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.

We want reconstructions of z from y to emerge from observations of y alone.

Latent Variables

Here we often think of z as the causal source of y.

z might be a physical scene causing image y.

z might be a word sequence causing speech sound y.

Latent Variables

$$P_{\Phi,\Theta}(y) = \sum_z P_\Phi(z) P_\Theta(y|z) = E_{z \sim P_\Phi(z)} P_\Theta(y|z)$$

$P_\Phi(z)$ is typically called the prior.

$P_{\Phi,\Theta}(z|y)$ is the posterior where y is the "evidence".

Assumptions

We assume models $P_\Phi(z)$ and $P_\Theta(y|z)$ are both samplable and computable.

In other words, we can sample from these distributions and for any given z and y we can compute $P_\Phi(z)$ and $P_\Theta(y|z)$.

These assumptions hold for auto-regressive models (language) and for Gaussian densities.

Modeling y

We would like to use cross-entropy from the population to the model probability $P_{\Phi,\Theta}(y)$.

$$\Phi^* = \operatorname*{argmin}_{\Phi} \; E_{y\sim\text{pop}} \; -\ln P_{\Phi,\Theta}(y)$$

$$P_{\Phi,\Theta}(y) = E_{z\sim P_\Phi(z)} \, P_\Theta(y|z)$$

But even when $P_\Phi(z)$ and $P_\Theta(y|z)$ are samplable and computable we cannot typically compute $P_{\Phi,\Theta}(y)$ or $P_{\Phi,\Theta}(z|y)$.

Modeling y

$$\Phi^* = \operatorname*{argmin}_{\Phi} \; E_{y \sim \text{pop}} \; -\ln P_{\Phi,\Theta}(y)$$

$$P_{\Phi,\Theta}(y) = E_{z \sim P_\Phi(z)} \, P_\Theta(y|z)$$

VAEs side-step the intractability problem by introducing another model component — a model $P_\Psi(z|y)$ to approximate the intractible $P_{\Phi,\Theta}(z|y)$.

The Evidence Lower Bound (The ELBO)

$$\ln P_{\Phi,\Theta}(y) = E_{z\sim P_\Psi(z|y)} \ln \frac{P_{\Phi,\Theta}(y)P_{\Phi,\Theta}(z|y)}{P_{\Phi,\Theta}(z|y)}$$

$$= E_{z\sim P_\Psi(z|y)} \left(\ln \frac{P_{\Phi,\Theta}(z,y)}{P_\Psi(z|y)} + \ln \frac{P_\Psi(z|y)}{P_{\Phi,\Theta}(z|y)} \right)$$

$$= \left(E_{z\sim P_\Psi(z|y)} \ln \frac{P_{\Phi,\Theta}(z,y)}{P_\Psi(z|y)} \right) + KL(P_\Psi(z|y), P_{\Phi,\Theta}(z|y))$$

$$\geq E_{z\sim P_\Psi(z|y)} \ln \frac{P_{\Phi,\Theta}(z,y)}{P_\Psi(z|y)} \qquad \text{The ELBO}$$

The ELBO

$$\ln P_{\Phi,\Theta}(y) \;\geq\; E_{z\sim P_\Psi(z|y)} \ln \frac{P_{\Phi,\Theta}(z,y)}{P_\Psi(z|y)} \tag{1}$$

$$= \; E_{z\sim P_\Psi(z|y)} \ln \frac{P_\Phi(z) P_\Theta(y|z)}{P_\Psi(z|y)}$$

$$H(y) \;\leq\; E_{y\sim \text{Pop},\, z\sim P_\Psi(z|y)} \; \ln \frac{P_\Psi(z|y)}{P_\Phi(z)} - \ln P_\Theta(y|z) \tag{2}$$

(1) holds with equality when $P_\Psi(z|y)$ equals $P_{\Phi,\Theta}(z|y)$.

(2) holds with equality when we also have $P_{\Phi,\Theta}(y) = \text{Pop}$.

The Variational Auto-Encoder (VAE)

$$\Phi^*, \Theta^*, \Psi^* = \underset{\Phi,\Theta,\Psi}{\operatorname{argmin}} \; E_{y\sim \mathrm{Pop},\, z\sim P_\Psi(z|y)} \; \ln \frac{P_\Psi(z|y)}{P_\Phi(z)} - \ln P_\Theta(y|z)$$

Here $P_\Psi(z|y)$ is the encoder and $P_\Theta(y|z)$ is the decoder and the "rate term" $E\left[\ln P_\Psi(z|y)/P_\Phi(z)\right]$ is a KL-divergence.

The Re-Parameterization Trick

$$\Phi^*, \Theta^*, \Psi^* = \underset{\Phi,\Theta,\Psi}{\operatorname{argmin}} \; E_{y\sim \text{Pop},\, z\sim P_\Psi(z|y)} \; \ln \frac{P_\Psi(z|y)}{P_\Phi(z)} - \ln P_\Theta(y|z)$$

We cannot do gradient descent into Ψ to handle the dependence of the loss on the sampling compute $z \sim P_\Psi(z|y)$.

To handle this we sample noise ϵ from a fixed noise distribution and replace $P_\Psi(z|y)$ with $P_\Psi(z|y,\epsilon)$.

The VAE training equation can then be written as

$$\Phi^*, \Theta^*, \Psi^* = \underset{\Phi,\Theta,\Psi}{\operatorname{argmin}} \; E_{y\sim \text{Pop},\epsilon\sim\text{noise}} \; \ln \frac{P_\Psi(z|y,\epsilon)}{P_\Phi(z)} - \ln P_\Theta(y|z)$$

EM is Alternating Optimization of the VAE

Expectation Maximimization (EM) applies in the (highly special) case where the exact posterior $P_{\Phi,\Theta}(z|y)$ is samplable and computable. EM alternates exact optimization of Ψ and the pair (Φ, Θ) in:

$$\text{VAE:} \qquad \Phi^*, \Theta^* = \underset{\Phi,\Theta}{\operatorname{argmin}} \min_{\Psi} E_{y,\, z\sim P_\Psi(z|y)} \; - \ln \frac{P_\Phi(z,y)}{P_\Psi(z|y)}$$

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

Inference (E Step)	Update (M Step)
$P_\Psi(z\|y) = P_{\Phi^t,\Theta^t}(z\|y)$	Hold $P_\Psi(z\|y)$ fixed

An Alternate Derivation of the VAE

Consider the following quantities for the distribution defined by $y \sim \text{Pop}$ and $z \sim P_\Psi(z|y)$.

$$
\begin{aligned}
H(z,y) &= H(y) + H(z|y) = H(z) + H(y|z) \\
H(y) &= H(z,y) - H(z|y) \\
&= H(z) + H(y|z) - H(z|y) \\
&\leq H(z, P_\Phi) + H(y, P_\Theta \mid z) - H(z|y) \quad \text{replacing entropies by cross-entropes} \\
&= E_{y\sim\text{Pop}, z\sim P_\Psi(z|y)} \ln \frac{P_\Psi(z|y)}{P_\Phi(z)} - \ln P_\Theta(y|z)
\end{aligned}
$$

Encoder Autonomy

$$H(y) \quad \leq \quad E_{y\sim \mathrm{Pop},\, z\sim P_\Psi(z|y)} \ \ln \frac{P_\Psi(z|y)}{P_\Phi(z)} - \ln P_\Theta(y|z)$$

Replacing an entropy by a cross-entropy will yield an equality when the model equals the source.

Thus for any encoder Ψ we get equality when $P_\Phi(z) = P_{\mathrm{Pop},\Psi}(z)$ and $P_\Theta(y|z) = P_{\mathrm{Pop},\Psi}(y|z)$.

Encoder Autonomy and Posterior Collapse

$$H(y) \quad \leq \quad E_{y\sim \text{Pop},\, z\sim P_\Psi(z|y)} \ \ln \frac{P_\Psi(z|y)}{P_\Phi(z)} - \ln P_\Theta(y|z)$$

Equality can be achieved for any Ψ.

For example $P_\Psi(z|y)$ can "collapse" to put all its weight on a single value. This is called posterior collapse.

The β-VAE

$$\text{VAE:} \quad \Phi^*, \Theta^*, \Psi^* = \underset{\Phi,\Theta,\Psi}{\operatorname{argmin}} \; E_{y\sim\text{Pop},z\sim P_\Psi(z|y)} \; \ln \frac{P_\Psi(z|y)}{P_\Phi(z)} - \ln P_\Theta(y|z)$$

$$\beta\text{-VAE:} \quad \Phi^*, \Theta^*, \Psi^* = \underset{\Phi,\Theta,\Psi}{\operatorname{argmin}} \; E_{y\sim\text{Pop},z\sim P_\Psi(z|y)} \; \beta \left(\ln \frac{P_\Psi(z|y)}{P_\Phi(z)} \right) - \ln P_\Theta(y|z)$$

$\beta < 1$ may avoid posterior collapse. $\beta > 1$ may improve interpretability. This is widely used but the motivation seems unclear.

Autonomous Encoder VAE (AE-VAE)

In the autonomous encoder VAE we add an arbitrary loss function on the encoder Ψ.

$$\text{AE-VAE:} \quad \Phi^*, \Theta^*, \Psi^* = \underset{\Phi,\Theta,\Psi}{\operatorname{argmin}} \; E_{y\sim\text{Pop},z\sim P_\Psi(z|y)} \quad \ln \frac{P_\Psi(z|y)}{P_\Phi(z)} - \ln P_\Theta(y|z) + \mathcal{L}(\Psi)$$

For any regularization on Ψ the first two terms will converge to $H(y)$ if $P_\Phi(z)$ and $P_\Theta(y|z)$ are sufficiently expressive and optimizable.

END