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 Models

$$
\begin{aligned}
P_{\Phi,\Theta}(z,y) &= P_\Phi(z)P_\Theta(y|z) \\
P_{\Phi,\Theta}(y) &= \sum_z P_{\Phi,\Theta}(z,y) \\
P_{\Phi,\Theta}(z|y) &= P_{\Phi,\Theta}(z,y)/P_{\Phi,\Theta}(y)
\end{aligned}
$$

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

$P_\Theta(y|z)$ is the "decoder"

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

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.

Computing $P_{\Phi,\Theta}(y)$

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

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

Computing $P_{\Phi,\Theta}(y)$

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

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

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

Computing $P_{\Phi,\Theta}(y)$

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

$$
\begin{aligned}
P_{\Phi,\Theta}(y) &= \sum_z P_\Phi(z) P_\Theta(y|z) \\
&= \sum_z P_{\Phi,\Theta}(z|y) \, \frac{P_\Phi(z) P_\Theta(y|z)}{P_{\Phi,\Theta}(z|y)} \\
&= E_{z \sim P_{\Phi,\Theta}(z|y)} \, \frac{P_\Phi(z) P_\Theta(y|z)}{P_{\Phi,\Theta}(z|y)}
\end{aligned}
$$

Computing $P_{\Phi,\Theta}(y)$

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

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

Variational Bayes side-steps 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)

$$
\begin{aligned}
\ln P_{\Phi,\Theta}(y) &= E_{z\sim P_\Psi(z|y)} \ln \frac{P_{\Phi,\Theta}(y)P_{\Phi,\Theta}(z|y)}{P_{\Phi,\Theta}(z|y)} \\
&= E_{z\sim P_\Psi(z|y)} \left(\ln \frac{P_{\Phi,\Theta}(z,y)}{P_\Psi(z|y)} + \ln \frac{P_\Psi(z|y)}{P_{\Phi,\Theta}(z|y)} \right) \\
&= \left(E_{z\sim P_\Psi(z|y)} \ln \frac{P_{\Phi,\Theta}(z,y)}{P_\Psi(z|y)} \right) + KL(P_\Psi(z|y), P_{\Phi,\Theta}(z|y)) \\
&\geq E_{z\sim P_\Psi(z|y)} \ln \frac{P_{\Phi,\Theta}(z,y)}{P_\Psi(z|y)} \qquad \text{The ELBO}
\end{aligned}
$$

The ELBO

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

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

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

The inequalities hold with equality when $P_\Psi(z|y)$ equals $P_{\Phi,\Theta}(z|y)$.

The Variational Auto-Encoder (VAE)

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

Here $P_\Phi(z)$ is **the prior**, $P_\Psi(z|y)$ 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

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

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

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

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:

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

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

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

Encoder Autonomy

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

But consider computing Φ^* and Θ^* for a fixed Ψ:

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

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

Independent of the encoder Ψ if $P_{\Phi^*}(z) = P_{\mathrm{Pop},\Psi}(z)$ and $P_{\Theta^*}(y|z) = P_{\mathrm{Pop},\Psi}(y|z)$ then the value of the objective function is $H(y)$ (the minimum possible) and $\mathrm{Pop}(y) = P_{\Phi,\Theta}(y)$.

Two-Pass Optimization

Fix the prior $P_\Psi(z)$ at a simple naive distribution and optimize the encoder $P_\Psi(z|y)$ and the decoder $P_\Theta(y|z)$.

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

We can think of this as lossy data compression under a simple fixed prior (coding) on the compressed file z.

While the fixed prior $P_\Phi(z)$ can be taken to be very simple, the decoder $P_\Theta(y|z)$ should be optimized aggressively.

Two-Pass Optimization

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

Only the last term depends on the decoder and so we get an optimal decoder for the encoder $P_{\Psi^*}(z)$.

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

Then train the prior $P_\Phi(z)$ aggressively holding the encoder and decoder fixed.

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

Two-Pass Optimization

$$\Theta^* = \operatorname*{argmin}_{\Theta}\; E_{y \sim \mathrm{Pop}, z \sim P_{\Psi^*}(z|y)}\left[-\ln P_\Theta(y|z)\right]$$

$$\Phi^* \;=\; \operatorname*{argmin}_{\Phi}\; E_{y \sim \mathrm{Pop}, z \sim P_{\Psi^*}(z|y)}\left[-\ln P_\Phi(z)\right]$$

Under a universality assumption for Φ and Θ we have that a perfect model of y can be achieved by optimizing the prior $P_\Phi(z)$ in a final pass for pre-trained Ψ and Θ.

Joint Training of Φ with Ψ and Θ is not required.

END