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

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

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

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

Assumptions

We assume models $P_\Phi(z)$ and $P_\Phi(y|z)$ 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 $P_\Phi(y|z)$.

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

Modeling y

We would like to use cross-entropy.

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

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

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

Modeling y

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

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

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

The Evidence Lower Bound (The ELBO)

$$\ln P_\Phi(y) = E_{z\sim\hat{P}_\Psi(z|y)} \ln \frac{P_\Phi(y)P_\Phi(z|y)}{P_\Phi(z|y)}$$

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

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

$$\geq E_{z\sim\hat{P}_\Psi(z|y)} \ln \frac{P_\Phi(z,y)}{\hat{P}_\Psi(z|y)} \quad \text{The ELBO}$$

Variational Autoencoders

$$\text{ELBO:} \qquad \ln P_\Phi(y) \;\geq\; E_{z\sim\hat{P}_\Psi(z|y)} \ln \frac{P_\Phi(z,y)}{\hat{P}_\Psi(z|y)}$$

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

$$\text{VAE:} \quad -\ln P_\Phi(y) \;\leq\; E_{z\sim\hat{P}_\Psi(z|y)} \ln \frac{\hat{P}_\Psi(z|y)}{P_\Phi(z)} - \ln P_\Phi(y|z)$$

Here $\hat{P}_\Psi(z|y)$ is the encoder and $P_\Phi(y|z)$ is the decoder and the "rate term" $E_{z|y}\ \ln \hat{P}_\Psi(z|y)/P_\Phi(z)$ is a KL-divergence.

The Re-Parameterization Trick

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

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

The VAE training equation can then be written as

$$\Phi^*, \Psi^* = \underset{\Phi,\Psi}{\operatorname{argmin}} \; E_{y,\epsilon} \; \ln \frac{\hat{P}_\Psi(z|y,\epsilon)}{P_\Phi(z)} - \ln P_\Phi(y|z)$$

EM is Alternating Optimization of the ELBO

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

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

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

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

Posterior (Encoder) Collapse

$$\Phi^*, \Psi^* = \underset{\Phi,\Psi}{\operatorname{argmin}} \; E_{y,z} \; \ln \frac{\hat{P}_\Psi(z|y)}{P_\Phi(z)} - \ln P_\Phi(y|z)$$

Consider a trivial encoder with $P_\Psi(z^*|y) = 1$ and $\hat{P}_\Phi(z^*) = 1$ for a fixed value z^* independent of y and the first term is zero.

Under universal expressiveness we have $\hat{P}_{\Phi^*}(y|z) = \text{Pop}(y)$ yielding $\hat{H}_{\Psi,\Phi}(y|z) = H(y)$.

Therefore, under universal expressiveness **there exists an optimal solution where the posterior (encoder)** $P_\Psi(z|y)$ **collapses**.

The β-VAE

$$P_\Psi(y,z) = \mathrm{Pop}(y) P_\Psi(z|y) \quad \text{The sampling distribution on } y,\ z$$

$$\text{VAE:} \quad \Phi^*, \Psi^* = \underset{\Phi,\Psi}{\operatorname{argmin}}\ E_{y,z}\ \ln \frac{\hat{P}_\Psi(z|y)}{P_\Phi(z)} - \ln P_\Phi(y|z)$$

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

$\beta < 1$ may avoid posterior collapse. $\beta > 1$ may improve interpretability.

Autonomous Encoder VAE (AE-VAE)

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

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

$$\text{AE-VAE:} \quad \Phi^*, \Psi^* = \underset{\Phi,\Psi}{\operatorname{argmin}} \; E_{y,z} \; \beta \left(\ln \frac{\hat{P}_\Psi(z|y)}{P_\Phi(z)} \right) - \ln P_\Phi(y|z) + \mathcal{L}(\Psi)$$

Autonomous Encoder VAE

$$\text{AE-VAE:} \quad \Phi^*, \Psi^* = \underset{\Phi,\Psi}{\operatorname{argmin}} \; E_{y,z} \; \beta \left(\ln \frac{\hat{P}_\Psi(z|y)}{P_\Phi(z)} \right) - \ln P_\Phi(y|z) + \mathcal{L}(\Psi)$$

We can hold the encoder $\hat{P}_\Psi(z|y)$ fixed and optimize $P_\Phi(z)$ and $P_\Phi(y|z)$ independently.

Assuming universality it can be shown that the optimum for $P_\Phi(z)$ is the true marginal on z under the distribution defined by the fixed encoder.

Assuming universality the optimum for $P_\Phi(y|z)$ is the true conditional probability under the distribution defined by the fixed encoder.

We then get that $P_{\Phi^*}(y)$ is the population distribution.

Gaussian VAEs

$$\text{VAE:} \quad \Phi^*, \Psi^* \;=\; \underset{\Phi,\Psi}{\operatorname{argmin}} \; E_{y,z} \; \ln \frac{\hat{p}_\Psi(z|y)}{p_\Phi(z)} - \ln p_\Phi(y|z)$$

All models are Gaussian densities.

$$p_\Phi(z[i]) \;\propto\; \exp((z[i] - \mu_\Phi[i])^2 / 2\sigma_\Phi^2[i])$$

$$p_\Psi(z[i]|y) \;\propto\; \exp((z[i] - \hat{z}_\Psi(y)[i])^2 / 2\sigma_\Psi^2(y)[i])$$

$$p_\Psi(y[i]|z) \;\propto\; \exp((y[i] - \hat{y}_\Phi(z)[i])^2 / 2\sigma^2(z)[i])$$

$$\textbf{WLOG } \hat{p}_\Phi(z) = \mathcal{N}(0, I)$$

The prior $p_\Phi(z)$ only appears in a KL term $KL(p_\Psi(z|y), p_\Phi(z))$.

We can reparameterize $\hat{z}_\Psi(y)$ to $\hat{z}_{\Psi'}(y)$ such that

$$KL(p_\Psi(z|y), p_\Phi(z)) = KL(p_{\Psi'}(z|y), \mathcal{N}(0, I)))$$

Gaussian VAEs for Faces 2014

We can sample faces from the VAE by sampling noise z from $p_\Phi(z)$ and then sampling an image y from $p_\Phi(y|z)$.

[Alec Radford]

END