

TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2020

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 .

To initially simplify the discussion, we consider models $P_{\Phi}(z)P_{\Phi}(y|z)$ where all variables are discrete.

For example, z might be a parse tree and y the resulting word sequence.

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 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).

However, they fail for loopy graphical models where approximations must be used.

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 $\hat{P}_{\Phi}(z|y)$ to approximate the intractible $P_{\Phi}(z|y)$.

The Evidence Lower Bound (The ELBO)

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

EM is Alternating Optimization of the ELBO

Expectation Maximization (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:

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

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

Inference

(E Step)

$$\hat{P}_{\Psi}(z|y) = P_{\Phi^t}(z|y)$$

Update

(M Step)

Hold $\hat{P}_{\Psi}(z|y)$ fixed

Variational Autoencoders

$$\begin{aligned}\text{ELBO:} \quad \ln P_{\Phi}(y) &\geq E_{z \sim \hat{P}_{\Phi}(z|y)} \ln \frac{P_{\Phi}(z, y)}{\hat{P}_{\Phi}(z|y)} \\ &= E_{z \sim \hat{P}_{\Phi}(z|y)} \ln \frac{P_{\Phi}(z) P_{\Phi}(y|z)}{\hat{P}_{\Phi}(z|y)}\end{aligned}$$

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

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

VAE = RDA

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

$P_{\Phi}(z)$, $P_{\Phi}(y|z)$ and $\hat{P}_{\Phi}(z|y)$ are model components and we can switch the notation to $\hat{P}_{\Phi}(z)$ $\hat{P}_{\Phi}(y|z)$ and $P_{\Phi}(z|y)$ with no change in the model.

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

In an RDA we take $P_{\Phi}(y, z)$ to be $\text{Pop}(y)P_{\Phi}(z|y)$ so that the rate term is an upper bound on $I_{\Phi}(y, z)$.

$$\mathbf{VAE} = \mathbf{RDA}$$

to be continued ...

END