TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2018

Variational Autoencoders

The Latent Variable Cross-Entropy Objective

We will now drop the negation and switch to argmax.

$$\Phi^* = \operatorname*{argmax}_{\Phi} \;\; E_{y \sim \mathrm{Pop}} \ln Q_\Phi(y)$$

$$Q_\Phi(y) = \sum_{\hat{z}} \; Q_\Phi(\hat{z}, y)$$

EG Identity: $\nabla_\Phi \ln Q_\Phi(y) = E_{\hat{z} \sim Q_\Phi(\hat{z}|y)} \, \nabla_\Phi \; \ln Q_\Phi(\hat{z}, y)$

Variational Autoencoders

$$\nabla_\Phi \ln Q_\Phi(y) = E_{\hat{z} \sim Q_\Phi(\hat{z}|y)} \nabla_\Phi \; \ln Q_\Phi(\hat{z}, y)$$

Except for directed tree models, this gradient must be approximated — exact computation is #-P hard.

Variational autoencoders approximate $Q_\Phi(\hat{z}|y)$ with a model supporting easy sampling of $\hat{z}$.

Generative Models

A model for which sampling is easy will be called **generative**.

In Variational autoencoders we assume that $Q_\Phi(y|\hat{z})$ is generative but that $Q_\Phi(\hat{z}|y)$ is not — that sampling from $Q_\Phi(\hat{z}|y)$ is hard.

We approximate $Q_\Phi(\hat{z}|y)$ with a generative model.

Generation Replaces Search

"Generation replaces search" can be viewed as a general principle of Deep leaning.

Rather than search for a $\hat{z}$ that generates y we strive to directly calculate — to generate — a $\hat{z}$ that generates y.

"Generation replaces search" is exemplified in current parsing architectures.

Variational Autoencoders

$$\nabla_\Phi \ln Q_\Phi(y) = E_{\hat{z}\sim Q_\Phi(\hat{z}|y)}\, \nabla_\Phi \; \ln Q_\Phi(\hat{z}, y)$$

$$\Phi^*, \Psi^* = \operatorname*{argmax}_{\Phi,\Psi} \; E_{y\sim \mathrm{Pop}}\, E_{\hat{z}\sim {\color{red}P_\Psi(\hat{z}|y)}} \; \ln Q_\Phi(\hat{z}, y) + {\color{red}H(P_\Psi(\hat{z}|y))}$$

Here $P_\Psi(\hat{z}|y)$ is a generative approximation of $Q_\Phi(\hat{z}|y)$.

The quantity being optimized is called the evidence lower bound (ELBO).

Variational Autoencoders

$$\nabla_\Phi \ln Q_\Phi(y) = E_{\hat{z}\sim Q_\Phi(\hat{z}|y)}\, \nabla_\Phi\; \ln Q_\Phi(\hat{z}, y)$$

$$\begin{aligned}\Phi^*, \Psi^* &= \operatorname*{argmax}_{\Phi,\Psi}\; E_{y\sim\mathrm{Pop}}\, E_{\hat{z}\sim P_\Psi(\hat{z}|y)}\; \ln Q_\Phi(\hat{z}, y) + H(P_\Psi(\hat{z}|y)) \\ &= \operatorname*{argmax}_{\Phi,\Psi}\; E_{y\sim\mathrm{Pop}}\; \ln Q_\Phi(y) - KL(P_\Psi(\hat{z}|y),\; Q_\Phi(\hat{z}|y))\end{aligned}$$

The equivalence of the two ELBO expressions is proved below.

The first expression supports SGD training through sampling.

The second expression establishes that the ELBO is a lower bound on the "evidence" $\ln Q_\Phi(y)$ and that $P_\Psi(\hat{z}|y)$ should approximate $Q_\Phi(\hat{z}|y)$.

Derivation of Equivalence I

$$E_{\hat{z}\sim P_\Psi(\hat{z}|y)}\;\ln Q_\Phi(\hat{z},y)$$

$$= \; E_{\hat{z}\sim P_\Psi(\hat{z}|y)}\;\left(\;\ln Q_\Phi(y) + \ln Q_\Phi(\hat{z}|y)\;\right)$$

$$= \; \ln Q_\Phi(y) + \; E_{\hat{z}\sim P_\Psi(\hat{z}|y)}\;\ln Q_\Phi(\hat{z}|y)$$

$$= \; \ln Q_\Phi(y) - H(P_\Psi(\hat{z}|y), Q_\Phi(\hat{z}|y))$$

Derivation of Equivalence II

$$E_{\hat{z}\sim P_\Psi(\hat{z}|y)}\ \ln Q_\Phi(\hat{z},y) + H(P_\Psi(\hat{z}|y))$$

$$= \ln Q_\Phi(y) - H(P_\Psi(\hat{z}|y), Q_\Phi(\hat{z}|y)) + H(P_\Psi(\hat{z}|y))$$

$$= \ln Q_\Phi(y) - KL(P_\Psi(\hat{z}|y), Q_\Phi(\hat{z}|y))$$

EM is Alternating Optimization of the ELBO

$$E_{\hat{z}\sim P_\Psi(\hat{z}|y)}\;\ln Q_\Phi(\hat{z},y) + H(P_\Psi(\hat{z}|y)) \quad (1)$$

$$= \ln\; Q_\Phi(y) - KL(P_\Psi(\hat{z}|y), Q_\Phi(\hat{z}|y)) \qquad (2)$$

$$\text{by (2)} \quad \Psi^* = \operatorname*{argmin}_\Psi E_{y\sim\mathrm{Pop}}\; KL(P_\Psi(\hat{z}|y), Q_\Phi(\hat{z}|y))$$

$$\text{by (1)} \quad \Phi^* = \operatorname*{argmax}_\Phi E_{y\sim\mathrm{Pop}}\; E_{\hat{z}\sim P_\Psi(\hat{z}|y)}\;\ln Q_\Phi(\hat{z},y)$$

$$\text{EM: } \Phi^{t+1} = \mathrm{argmax}_\Phi\;\; E_{y\sim\mathrm{Pop}}\; E_{\hat{z}\sim Q_{\Phi^t}(\hat{z}|y)}\;\; \log Q_\Phi(\hat{z},y)$$

The Reparameterization Trick

$$\Psi^* = \operatorname*{argmax}_{\Psi} \; E_{y\sim\mathrm{Pop}} \; E_{\hat{z}\sim P_\Psi(\hat{z}|y)} \; \ln Q_\Phi(\hat{z}, y) + H(P_\Psi(\hat{z}|y))$$

How do we differentiate the sampling?

The Reparameterization Trick

$$\Psi^* = \operatorname*{argmax}_{\Psi} \; E_{y \sim \text{Pop}} \; E_{\hat{z} \sim P_\Psi(\hat{z}|y)} \; \ln Q_\Phi(\hat{z}, y) + H(P_\Psi(\hat{z}|y))$$

We note that in practice all sampling is computed by a deterministic function of (pseudo) random numbers.

We can make this explicit.

Model $P_\Psi(\hat{z}|y)$ by $\epsilon \sim$ noise, $\hat{z} = \hat{z}_\Psi(y, \epsilon)$

The Reparameterization Trick

$$\Psi^* = \operatorname*{argmax}_{\Psi} \; E_{y\sim\mathrm{Pop}} \; E_{\epsilon\sim\mathrm{noise}} \; \ln Q_\Phi(\hat{z}_\Psi(y,\epsilon), y) + H(P_\Psi(\hat{z}|$$

$$H(P_\Psi(\hat{z}|y)) = E_{\epsilon\sim\mathrm{noise}} \; \ln P_\Psi(\hat{z}_\Psi(y,\epsilon)|y)$$

For VAEs we typically we have $\hat{z}(y,\epsilon) \in \mathbb{R}^d$ with

$$\begin{aligned} \hat{z}(y,\epsilon)[i] &= \mu_\Psi(y)[i] + \sigma_\Psi(y)[i]\;\epsilon[i] \\ \epsilon[i] &\sim \mathcal{N}(0,1) \end{aligned}$$

This supports easy calculation of $P_\Psi(\hat{z}_\Psi(y,\epsilon)|y)$.

Decoding with L_2 Distortion

An autoencoder **encodes** and **decodes**.

We can view $\hat{z}_\Psi(y, \epsilon)$ as the encoding of y.

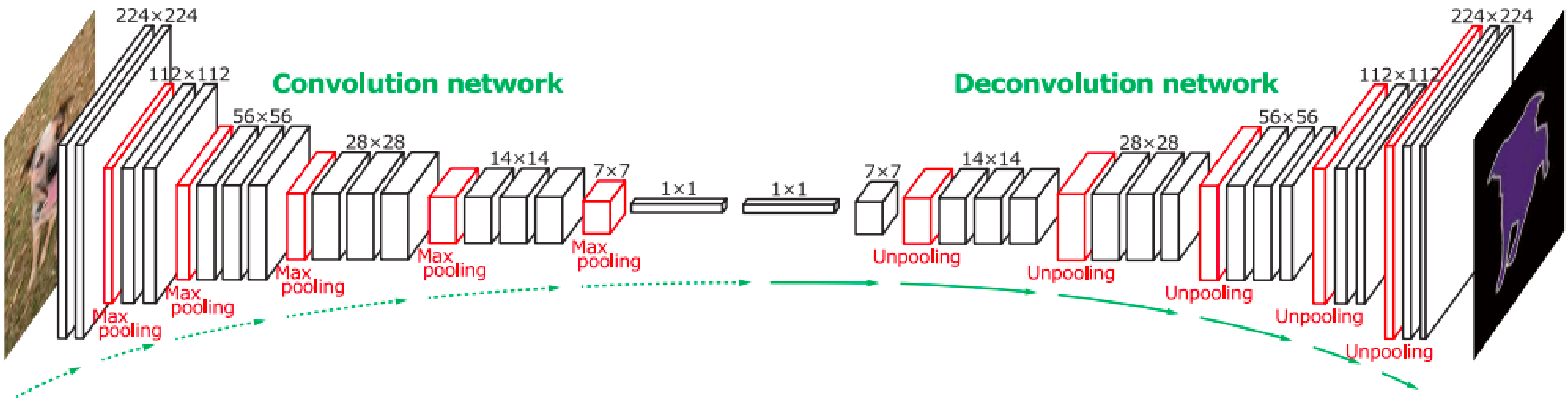
We now consider a deterministic decoder $\hat{y}_\Phi(\hat{z})$ and define a model

$$Q_\Phi(y|\hat{z}) \propto \exp\left(\frac{-||y - \hat{y}_\Phi(\hat{z})||^2}{2\sigma^2}\right)$$

A VAE for Images

Auto-Encoding Variational Bayes, Diederik P Kingma, Max Welling, 2013.

y $\quad$ $\hat{z}_\Psi(y, \epsilon)$ $\quad$ $\hat{z}$ $\quad$ $\hat{y}_\Phi(\hat{z})$ $\quad$ $||y - \hat{y}||^2$



[Hyeonwoo Noh et al.]

Deconvoution: Increasing Spatial Dimension

Consider a stride 2 convolution

$$
\begin{aligned}
y[i,j,c_y] &= W[\Delta i, \Delta j, c_x, c_y] x[2i+\Delta i, 2j+\Delta j, c_x] \\
y[i,j,c_y] &\mathrel{+}= B[c_y]
\end{aligned}
$$

For deconvolution we use stride 1 with 4 times the channels.

$$
\begin{aligned}
\hat{x}[i,j,c_{\hat{x}}] &= W'[\Delta i, \Delta j, c_{\hat{y}}, c_{\hat{x}}] \hat{y}[i+\Delta i, j+\Delta j, c_{\hat{x}}] \\
\hat{x}[i,j,c_{\hat{x}}] &\mathrel{+}= B[c_{\hat{x}}]
\end{aligned}
$$

The channels at each lower resolution pixel $\hat{x}[i,j]$ are divided among four higher resolution pixels.

This is done by a simple reshaping of $\hat{x}$.

Decoding with L_2 Distortion

$$\Phi^*, \Psi^* = \operatorname*{argmax}_{\Phi,\Psi} \; E_{y\sim \mathrm{Pop}} \, E_{\hat{z}\sim P_\Psi(\hat{z}|y)} \; \ln Q_\Phi(\hat{z}, y) + H(P_\Psi(\hat{z}|y))$$

The objective now becomes

$$E_{y\sim \mathrm{Pop}} \, E_{\hat{z}\sim P_\Psi(\hat{z}|y)} \; \left(\ln P_\Phi(\hat{z}) - \frac{1}{2\sigma^2}\,||y - \hat{y}_\Phi(\hat{z})||^2\right) + H(P_\Psi(\hat{z}|y))$$

Decoding with L_2 Distortion

Switching back to minimization, we can now rewrite the objective as

$$\min \; E_{y,\epsilon} \;\; \left(|\hat{z}_\Psi(y,\epsilon)|_\Phi + \frac{1}{2}\lambda ||y - \hat{y}_\Phi(\hat{z}_\Psi(y,\epsilon))||^2 \right) - |\hat{z}_\Psi(y,\epsilon)|_{\Psi,y}$$

$$|\hat{z}|_\Phi = -\log_2 P_\Phi(\hat{z}) \qquad\qquad |\hat{z}|_{\Psi,y} = -\log_2 P_\Psi(\hat{z}|y)$$

For $\hat{z}$ discrete, $|\hat{z}|_\Phi$ is the code length of $\hat{z}(y,\epsilon)$ under an optimal code for P_Φ.

$|\hat{z}|_{\Psi,y}$ is the code length for $\hat{z}$ under the code for $P_\Psi(\hat{z}|y)$.

Soft EM is to Hard EM as VAE is to Rate-Distortion

(Soft) EM: $\Phi^{t+1} = \mathrm{argmax}_{\Phi}\, E_{y\sim\mathrm{Pop}} \quad E_{\hat{z}\sim Q_{\Phi^t}(\hat{z}|y)} \quad \log Q_{\Phi}(\hat{z}, y)$

Hard EM: $\Phi^{t+1} = \mathrm{argmax}_{\Phi}\;\; E_{y\sim\mathrm{Pop}}\, Q_{\Phi}(\hat{z}(y), y)$

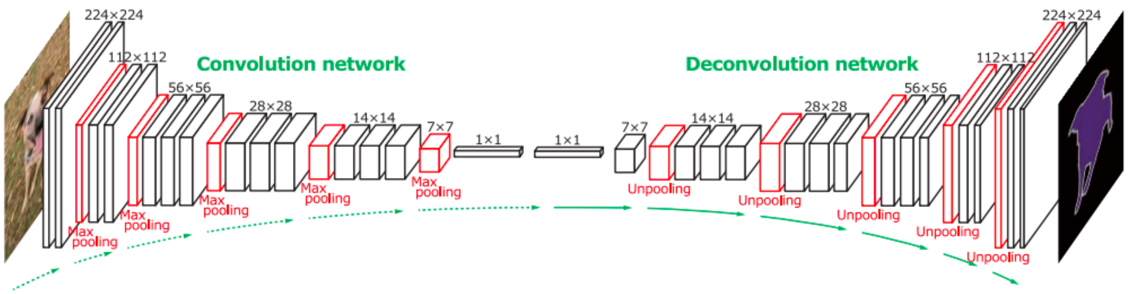
$$\hat{z}(y) = \underset{\hat{z}}{\mathrm{argmax}}\;\; Q_{\Phi^t}(\hat{z}|y)$$

VAE: $\min E_{y,\epsilon}\; |\hat{z}_{\Psi}(y,\epsilon)|_{\Phi} + \frac{1}{2}\lambda||y - \hat{y}_{\Phi}(\hat{z}_{\Psi}(y,\epsilon))||^2 - |\hat{z}_{\Psi}(y,\epsilon)|_{\Psi,y}$

RD: $\min E_y\; |\hat{z}_{\Psi}(y)|_{\Phi} + \frac{1}{2}\lambda||y - \hat{y}_{\Phi}(\hat{z}_{\Psi}(y))||^2$

Sampling

$$P_\Psi(\hat{z}|y) \qquad \hat{z} \qquad Q_\Phi(\hat{z}, y)$$



[Hyeonwoo Noh et al.]

Sampling uses just the second half $Q_\Phi(\hat{z}, y)$.

Sampling from Gaussian Variational Autoencoders

[Alec Radford]

Why Blurry?

A common explanation for the blurryness of images generated from VAEs is the use of L_2 as the distortion measure.

It does seem that L_1 works better (see the slides on image-to-image GANs).

However, training on L_2 distortion can produce sharp images in rate-distortion autoencoders (see the slides on rate-distortion autoencoders).

END