TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2019

Expectation Maximization (EM)

The Evidence Lower Bound (the ELBO)

Variational Autoencoders (VAEs)

Latent Variable Models

We are often interested in models of the form

$$P_\Phi(y) = \sum_z P_\Phi(z) P_\Phi(y|z).$$

$$P_\Phi(y|x) = \sum_z P_\Phi(z|x) P_\Phi(y|z).$$

For example, CTC and probabilistic grammar models.

Expectation Maximization (EM)

Mixture of Gaussian Modeling

$$\Phi = (\pi_1, \mu_1, \Sigma_1, \ldots, \pi_k, \mu_k, \Sigma_k)$$

$$\begin{aligned} p_\Phi(y) &= \sum_i P(i)P(y|i) \\ &= \sum_i \pi_i \frac{1}{Z_i} \exp\left(-\frac{1}{2}(y-\mu_i)^\top \Sigma_i^{-1}(y-\mu_i)\right) \end{aligned}$$

i is the latent variable.

Expectation Maximization (EM)

Mixture of Gaussian Modeling

$$\Phi = (\pi_1, \mu_1, \Sigma_1, \ldots, \pi_k, \mu_k, \Sigma_k)$$

$$\mathrm{Train} = \{y_1, \ldots, y_N\}$$

Until Convergence:

$$w[i,j] = P_\Phi(i|y_j) \quad \text{Inference (E step)}$$

$$\left.\begin{array}{lcl} \pi_i & = & \frac{1}{N} \sum_j w[i,j] \\ \mu_i & = & \frac{1}{N} \sum_j w[i,j] y_j \\ \Sigma_i & = & \frac{1}{N} \sum_j w[i,j] y_j y_j^\top \end{array}\right\} \text{Model Update (M step)}$$

General EM

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

$$P_\Phi(y) = \sum_z P_\Phi(z)P_\Phi(y|z).$$

$$\Phi^{t+1} = \underbrace{\operatorname*{argmin}_{\Phi}}_{\substack{\text{M Step}\\ \text{Update}}} E_{y\sim\mathrm{Train}}\; \underbrace{E_{z\sim P_{\Phi^t}(z|y)}}_{\substack{\text{E Step}\\ \text{Inference}}} - \ln P_\Phi(z,y)$$

Superpixel Colorization

SLIC superpixels, Achanta et al.

x is a black and white image.

y is a color image drawn from $\mathrm{Pop}(y|x)$.

$\hat{y}$ is an arbitrary color image.

$P_\Phi(\hat{y}|x)$ is the probability that model Φ assigns to the color image $\hat{y}$ given black and white image x.

Latent Semantic Segmentation (TZ)

$$P_\Phi(y|x) = \sum_z P_\Phi(z|x) P_\Phi(y|z).$$

intput x

$P_\Phi(z|x) = \ldots$ semantic segmentation (friendly — pixel RNN?)

$P_\Phi(\hat{y}|z) = \ldots$ colorization (friendly — table lookup?)

$\mathcal{L}(y, \hat{y}) = -\ln\, p_\Phi(y|\hat{y})$ distortion (friendly — Gaussian)

The composition is unfriendly.

Maybe EM?

$$
\begin{aligned}
&\text{intput } x \\
P_\Phi(z|x) &= \ldots \quad \text{semantic segmentation (friendly)} \\
P_\Phi(\hat{y}|z) &= \ldots \quad \text{colorization (friendly)} \\
\mathcal{L}(y,\hat{y}) &= -\ln\, p_\Phi(y|\hat{y}) \quad \text{distortion (friendly)}
\end{aligned}
$$

$$
\Phi^{t+1} = \underset{\Phi}{\operatorname{argmin}}\; E_{y\sim\mathrm{Train}}\; E_{z\sim P_{\Phi^t}(z|y)}\; -\ln P_\Phi(z,y)
$$

Update Inference

The inference is intractible!

Variational Inference:
The Evidence Lower Bound (The ELBO)

$$P_\Phi(y) = \sum_z P_\Phi(z)P_\Phi(y|z) \quad \text{unfriendly composition of friendlies}$$

$$\begin{aligned}
\ln P_\Phi(y) &= E_{z\sim P_\Psi(z|y)} \ln P_\Phi(y) \quad P_\Psi(z|y) \text{ friendly} \\
&= E_{z\sim P_\Psi(z|y)} \left(\ln P_\Phi(y)P_\Phi(z|y) + \ln \frac{P_\Psi(z|y)}{P_\Phi(z|y)} + \ln \frac{1}{P_\Psi(z|y)} \right) \\
&= \left(E_{z\sim P_\Psi(z|y)} \ln P_\Phi(z,y) \right) + KL(P_\Psi(z|y), P_\Phi(z|y)) + H(P_\Psi(z|y)) \\
&\geq E_{z\sim P_\Psi(z|y)} \left(\ln P_\Phi(z)P_\Phi(y|z) - \ln P_\Psi(z|y) \right) \quad \text{ELBO}
\end{aligned}$$

Variational Inference:
The Evidence Lower Bound (The ELBO)

$$P_\Phi(y) = \sum_z P_\Phi(z)P_\Phi(y|z) \quad \text{unfriendly composition of friendlies}$$

$$\ln P_\Phi(y) \;=\; \mathrm{ELBO} + KL(P_\Psi(z|y), P_\Phi(z|y))$$

If the friendly $P_\Psi(z|y)$ can match the unfriendly $P_\Phi(z|y)$ then the ELBO is exact.

Measuring the ELBO

$$\mathrm{ELBO}(y, \Phi, \Psi) = E_{z \sim P_\Psi(y)} \ln P_\Phi(z) P_\Phi(y|z) - \ln P_\Psi(z)$$

If $P_\Phi(z)$, $P_\Phi(y|z)$, and $P_\Psi(z|y)$ are friendly (even whwn when $P_\Phi(y)$ is not friendly) we can measure ELBO loss through sampling.

If we can measure it, we can do gradient descent on it (but perhaps with difficulty).

Colorization

x is a black and white image, y a color image, and z a semantic segmentation.

$P_\Phi(z|x)$ is friendly and $P_\Phi(y|z,x)$ is friendly but $P(y|x)$ is not friendly.

$P_\Psi(z|y,x)$ computes a friendly graphical model for z given y.

A General ELBO Architecture

$P_\Phi(y|x)$

$$
\begin{aligned}
& \text{intput } x \\
& \vdots \\
z \;&=\; \underset{z}{\text{expsoftmax}} \ldots \\
& \text{input } z \\
& \vdots \\
y \;&=\; \underset{y}{\text{expsoftmax}} \;\ldots
\end{aligned}
$$

$P_\Psi(z|y,x)$

$$
\begin{aligned}
& \text{intput } x, y \\
& \vdots \\
z \;&=\; \underset{z}{\text{expsoftmax}} \ldots
\end{aligned}
$$

The exponential softmaxes are friendly (they produce a friendly graphical model).

Two Expressions for the ELBO

$$\ln P_\Phi(y) = ELBO(y, \Phi, \Psi) + KL(P_\Psi(z|y), P_\Phi(z|y))$$

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

$$= \ln\ P_\Phi(y) - KL(P_\Psi(z|y), P_\Phi(z|y)) \qquad (2)$$

EM is Alternating Maximization of the ELBO

Forward-backward EM for HMMs and inside-outside EM for PCFGs (or any EM) can be written as

$$\mathrm{ELBO} = E_{z\sim P_\Psi(z|y)} \ln P_\Phi(z,y) + H(P_\Psi(z|y)) \quad (1)$$

$$= \ln\ P_\Phi(y) - KL(P_\Psi(z|y), P_\Phi(z|y)) \quad (2)$$

by (2) $\quad \Psi^{t+1} = \underset{\Psi}{\mathrm{argmin}}\ E_{y\sim \mathrm{Train}}\, KL(P_\Psi(z|y), P_{\Phi^t}(z|y)) = \Phi^t$

by (1) $\quad \Phi^{t+1} = \underset{\Phi}{\mathrm{argmax}}\, E_{y\sim \mathrm{Train}}\, E_{z\sim P_{\Phi^t}(z|y)} \ln P_\Phi(z,y)$

We want Ψ to adapt to Φ

$$\mathcal{L}_{\mathrm{ELBO}}(y, \Phi, \Psi) = KL(P_\Psi(z|y), P_\Phi(z|y)) - \ln\ P_\Phi(y)$$

$$Q^*(z|y) = P_\Phi(z|y)$$

$$E_{y \sim \mathrm{Pop}}\ \mathcal{L}_{\mathrm{ELBO}}(y, \Phi, Q^*) = H(\mathrm{Pop}, P_\Phi)$$

However, Φ can ignore Ψ

$$\mathcal{L}_{\mathrm{ELBO}}(y,\Phi,\Psi) = KL(P_\Psi(z|y), P_\Phi(z|y)) - \ln\ P_\Phi(y)$$

$$P^*(z) = P_\Psi(z)$$
$$P^*(y|z) = P_\Phi(y)$$

$$E_{y\sim\mathrm{Pop}}\ \mathcal{L}_{\mathrm{ELBO}}(y, P^*, \Psi) = H(\mathrm{Pop}, P_\Phi)$$

It seems important that $P_\Phi(y|z)$ have limited expressive power.

Hard ELBO

Hard ELBO is to ELBO as hard EM is to EM.

$$\mathcal{L}_{\mathrm{ELBO}}(y,\Phi,\Psi) \;=\; KL(P_\Psi(z|y), P_\Phi(z|y)) - \ln P_\Phi(y)$$

$$\mathcal{L}_{\mathrm{ELBO}}(y,\Phi,\Psi) \;=\; E_{z\sim P_\Psi(z|y)} - \ln P_\Phi(z,y) + \ln P_\Psi(z|y)$$

$$\mathcal{L}_{\mathrm{HELBO}}(y,\Phi,\Psi) \;=\; E_{z\sim P_\Psi(z|y)} - \ln P_\Phi(z,y)$$

Hard ELBO and Rate-Distortion Autoencoding

$$\mathcal{L}_{\mathrm{HELBO}}(y, \Phi, \Psi) \;=\; E_{z \sim P_\Psi(z|y)} - \ln P_\Phi(z, y)$$

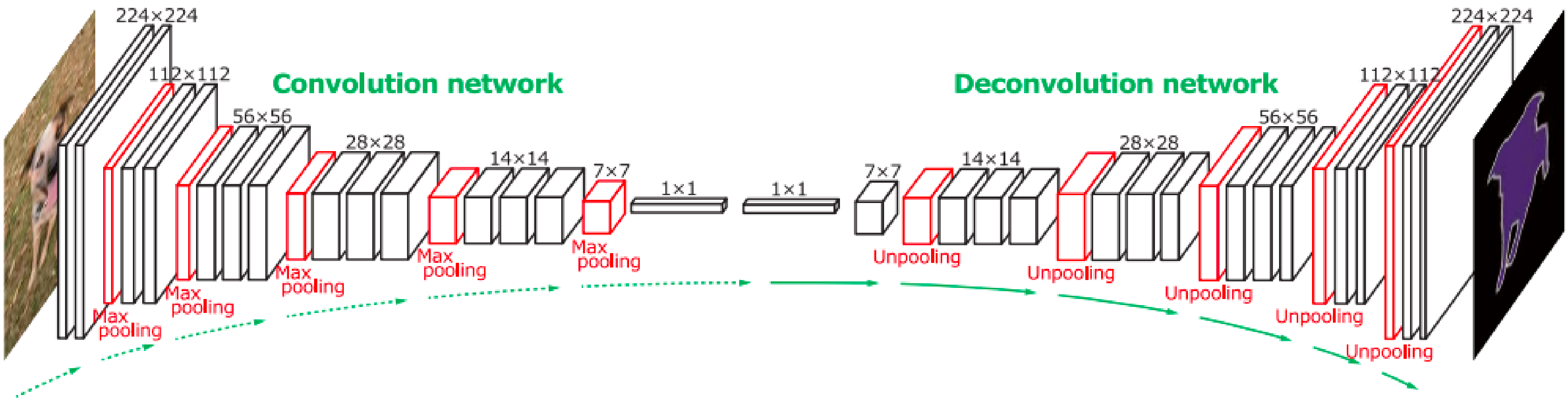
$$\min_{P,Q} E_{y \sim \mathrm{Pop}}\, \mathcal{L}_{\mathrm{HELBO}}(y, P, Q) \leq H(\mathrm{Pop}) + \ln 2$$

This can be proved from Shannon's source coding theorem where z is the code for y.

A VAE for Images

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

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



[Hyeonwoo Noh et al.]

Gaussian Distributions

$$p_\Phi(z) \propto \exp\left(\sum_i (z[i] - \mu[i])^2/(2\sigma[i]^2)\right)$$

$$p_\Phi(y|z) \propto \exp\left(\sum_j (y[j] - y_\Phi(z)[j])^2/(2\gamma[j]^2)\right)$$

$$p_\Psi(z|y) \propto \exp\left(\sum_i (z[i] - z_\Psi(y)[i])^2/(2\sigma_\Psi(y)[i]^2)\right)$$

KL-Divergence Form for the ELBO

$$E_{z \in p_\Psi(z|y)} \; \ln \; p_\Psi(z|y) - \ln \; p_\Phi(z)p_\Phi(y|z) \quad \mathcal{L}_{\mathrm{ELBO}}$$

$$= \; KL(p_\Psi(z|y), p_\Phi(z)) + E_{z \in P_\Psi(z|y)} - \ln p_\Phi(y|z)$$

The ELBO is a KL-divergence + a cross entropy

Continuous KL-divergence is ok.

Continuous cross-entropy has issues — we will come back to that later.

Closed Form KL-Divergence

$$KL(p_\Psi(z|y), p_\Phi(z))$$

$$= \sum_i \frac{\sigma_\Psi(y)[i]^2 + (z_\Psi(y)[i] - \mu[i])^2}{2\sigma[i]^2} + \ln \frac{\sigma[i]}{\sigma_\Psi(y)[i]} - \frac{1}{2}$$

Standardizing $p_\Phi(z)$

The KL-divergence term is

$$\sum_i \frac{\sigma_\Psi(y)[i]^2 + (z_\Psi(y)[i] - \mu[i])^2}{2\sigma[i]^2} + \ln \frac{\sigma[i]}{\sigma_\Psi(y)[i]} - \frac{1}{2}$$

We can adjust Ψ to Ψ' such that

$$\begin{aligned} z_{\Psi'}(y)[i] &= z_\Psi(y)[i]/\sigma[i] + \mu[i] \\ \sigma_{\Psi'}(y)[i] &= \sigma_\Psi(y)/\sigma[i] \end{aligned}$$

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

Standardizing $p_\Phi(z)$

Without loss of generality the VAE becomes.

$$\min_{\Phi,\Psi}\; E_y\; KL(P_\Psi(z|y), \mathcal{N}(0,I)) + E_{z\in P_\Psi(z|y)} - \ln p_\Phi(y|z)$$

Reparameterization Trick for the Cross-Entropy

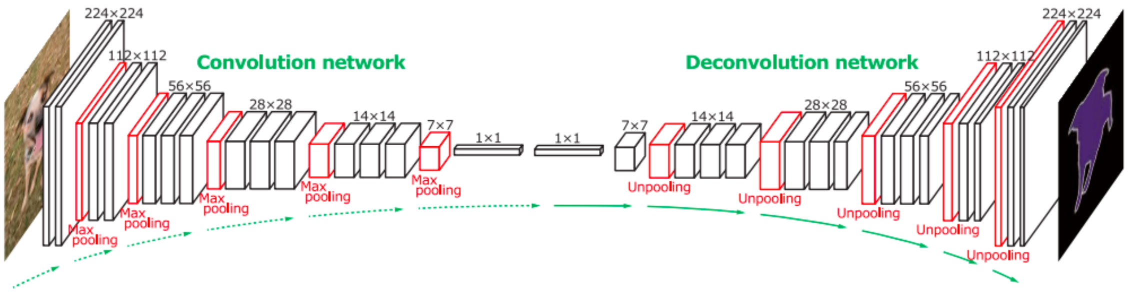
$$p_\Psi(z|y) \propto \exp\left(\sum_i (z[i] - z_\Psi(y)[i])^2/(2\sigma_\Psi(y)[i]^2)\right)$$

$$E_{z \in p_\Psi(z|y)} \ln p_\Phi(y|z)$$

$$= E_{\epsilon \sim \mathcal{N}(0,I)}\; z[i] = z_\Psi(y)[i] + \sigma_\Psi(y)[i]\epsilon[i]; \quad \ln p_\Phi(y|z)$$

Sampling

$$P_\Psi(z|y) \qquad z \qquad P_\Phi(z,y)$$



[Hyeonwoo Noh et al.]

Sampling uses just the second half $P_\Phi(z,y)$.

Sampling

[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.

However, training on L_2 distortion can produce sharp images in rate-distortion autoencoders.

Noisy-Channel Rate-Distortion Autoencoders

The twilight zone is material for which I do not know of a reference.

Differential Entropy and Cross-Entropy are Ill-Defined

$$\mathcal{L}_{\mathrm{VAE}} = \sum_j \frac{E_{z\sim P_\Psi(z|y)}\,(y[j] - \hat{y}_\Phi(z)[j])^2}{2\gamma[j]^2} + \ln \gamma[j]$$

$$+KL(p_\Psi(z|y), p_\Phi(z))$$

Consider a probability density on light intensity.

While the first term is dimensionless, $\gamma[j]$ is an intensity.

The cross-entropy term can be assigned any numerical value depending on the choice units (metric, English, or martian).

Differential Entropy and Cross-Entropy are Ill-Defined

There are also other problems with continuous entropy and cross-entropy.

- Finite continuous entropy violates the source coding theorem — it takes an infinite number of bits to code a real number.

- Finite continuous entropy violates the data processing inequality that $H(f(x)) \leq H(x)$. For a continuous random variable x under finite continuous entropy we can have $H(f(x)) > H(x)$.

For these reasons it seems best to avoid using finite continuous entropy and finite continuous cross entropy.

Distortion

A stochastic encoder $p_\Phi(z|y)$, a decoder $y_\Phi(z)$, and distortion function D define a quantity of distortion.

$$E_{y\sim\mathrm{Pop},\; z\sim p_\Phi(z|y)}\; D(y, y_\Phi(z))$$

For L_2 distortion we can use

$$D(y, y') = ||y - y'||_2$$

Distortion can typically be given the same units as y.

Rate

A stochastic encoder defines a rate.

$$p_\Phi(z) \doteq \sum_y \mathrm{Pop}(y) p_\Phi(z|y)$$

$$I_\Phi(y,z) = E_y \, KL(p_\Phi(z|y), p_\Phi(z))$$

By Shannon's channel capacity theorem, $I_\Phi(y,z)$ is the channel capacity when sending y across the noisy channel z.

For z continuous, a deterministic encoder has an infinite rate.

Here $p_\Phi(z)$ is not friendly.

Bounding the Rate

$$
\begin{aligned}
I_\Phi(y,z) &= E_{y\sim \mathrm{Pop}}\, KL(p_\Phi(z|y), p_\Phi(z)) \\
&= E_{y,z} \ln p_\Phi(z|y) - \ln p_\Psi(z) + \ln p_\Psi(z) - \ln p_\Phi(z) \\
&= E_y\, KL(p_\Phi(z|y), p_\Psi(z)) - KL(p_\Phi(z), p_\Psi(z)) \\
&\leq E_y\, KL(p_\Phi(z|y), p_\Psi(z))
\end{aligned}
$$

We can take $p_\Psi(z)$ to be friendly, and WLOG, fixed at $\mathcal{N}(0, I)$.

The Noisy-Channel Rate-Distortion Autoencoder

$$\Phi^* = \underset{\Phi}{\text{argmin}}\ E_y\ KL(p_\Phi(z|y), \mathcal{N}(0, I)) + \frac{1}{\gamma}\, E_{z \sim p_\Phi(z|y)}\ D(y,\ y_\Phi(z))$$

Here γ has the same units as distortion and controls the trade-off between rate and distortion.

Summary: Rate-Distortion

Rate-Distortion: y, continuous, $\tilde{z}$ a bit string,

$$\Phi^* = \operatorname*{argmin}_{\Phi} E_y \, |\tilde{z}_\Phi(y)| + \lambda D(y, y_\Phi(\tilde{z}_\Phi(y)))$$

Noisy Channel: $\tilde{z} = z_\Phi(y) + \sigma_\Phi(y) \odot \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I)$

$$\Phi^* = \operatorname*{argmin}_{\Phi} E_y \;\; KL(p_\Phi(\tilde{z}|y), \mathcal{N}(0, I)) + E_{\tilde{z}\sim p_\Phi(\tilde{z}|y)} \, \lambda D(y, y_\Phi(\tilde{z}))$$

Summary: ELBO and VAE

ELBO: $P_\Phi(z)$, $P_\Phi(y|z)$, $P_\Psi(z|y)$ friendly graphical models:

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

VAE: $p_\Phi(z|y)$, $p_\Phi(y|z)$ Gaussian:

$$\Phi^* = \operatorname*{argmin}_{\Phi} E_{y\sim\mathrm{Pop}} \; KL(p_\Phi(z|y), \mathcal{N}(0, I)) - E_{z\sim p_\Phi(z|y)} \; \ln p_\Phi(y|z)$$

END