TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2020

The Evidence Lower Bound (ELBO)

and Variational Auto Encoders (VAEs)

Latent Variable Assumptions

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

Specifically, for $P_\Phi(y)$ defined by a GAN generator we cannot compute $P_\Phi(y)$ for a test image y.

Hence it is not obvious how to optimize the fundamental equation.

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

The Evidence Lower Bound (The ELBO)

We introduce a samplable and computable model $Q_\Phi(z|y)$ to approximate $P_\Phi(z|y)$.

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

The Variational Auto-Encoder (VAE)

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

VAE generalizes EM

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 Q and P.

VAE: $\Phi^* = \mathrm{argmin}_\Phi \, E_{y\sim \mathrm{Train},\; z\sim Q_\Phi(z|y)} \;\; -\ln \frac{P_\Phi(z,y)}{Q_\Phi(z|y)}$

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

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

The Reparameterization Trick

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

$$=\; E_\epsilon\; -\ln \frac{P_\Phi(z)P_\Phi(y|z)}{Q_\Phi(z|y)} \quad z := f_\Phi(y,\epsilon)$$

ϵ is parameter-independent noise.

This supports SGD: $\nabla_\Phi\; E_{y,\epsilon}\;[\ldots] = E_{y,\epsilon}\;\nabla_\Phi\;[\ldots]$

Gaussian VAEs

$$\Phi^* = \underset{\Phi}{\mathrm{argmin}}\, E_{y,\epsilon}\ -\ln \frac{p_\Phi(z)p_\Phi(y|z)}{q_\Phi(z|y)}$$

$$z\ =\ z_\Phi(y) + \sigma_\Phi(y) \odot \epsilon \quad \epsilon \sim \mathcal{N}(0, I)$$

$$q_\Phi(z[i]|y)\ =\ \mathcal{N}(z_\Phi(y)[i], \sigma_\Phi(y)[i])$$

$$p_\Phi(z[i])\ =\ \mathcal{N}(\mu_p, \sigma_p[i])\ \ \mathrm{WLOG} = \mathcal{N}(0, 1)$$

$$p_\Phi(y|z)\ =\ \mathcal{N}(y_\Phi(z),\ \sigma^2 I)$$

$-\ln p_\Phi(y|z)$ as Distortion

For $p_\Phi(y|z) \propto \exp(-||y - y_\Phi(z)||^2/(2\sigma^2))$ we get

$$
\begin{aligned}
\Phi^* &= \operatorname*{argmin}_{\Phi} E_{y,\epsilon} \quad -\ln \frac{p_\Phi(z)}{q_\Phi(z|y)} - \ln p_\Phi(y|z) \\
&= \operatorname*{argmin}_{\Phi,\sigma} E_{y,\epsilon} \quad -\ln \frac{p_\Phi(z)}{q_\Phi(z|y)} + \left(\frac{1}{2\sigma^2}\right) ||y - y_\Phi(z)||^2 + d \ln \sigma
\end{aligned}
$$

where

d is the dimension of y and $\sigma^* = \sqrt{\frac{1}{d}\, E_{y,\epsilon}\, ||y - y_\Phi(z)||^2}$

Posterior Collapse

Assume Universal Expressiveness for $P_\Phi(y|z)$.

This allows $P_\Phi(y|z) = \mathrm{Pop}(y)$ independent of z.

We then get a completely optimized model with z taking a single (meaningless) determined value.

$$Q_\Phi(z|y) = P_\Phi(z|y) = 1$$

Colorization with Latent Segmentation

x $\hat{y}$ y

Larsson et al., 2016

Can colorization be used to learn latent segmentation?

We introduce a latent segmentation into the model.

In practice the latent segmentation is likely to "collapse" because the colorization can be done just as well without it.

Independent Universality

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

It is natural to assume that Φ has independent parameters for each distribution. In practice parameters are often shared.

Since Φ can indepedently parameterize each distribution, we will often use an independent universality assumption that Φ can represent any triple of distributions $Q(z|y)$, $P(z)$ and $P(y|z)$.

Independent Universality

More formally, we will often assume that for any triple of distributions $Q(z|y)$, $P(z)$ and $P(y|z)$ there exists a Φ that simultaneously satisfies

$$
\begin{aligned}
Q_\Phi(z|y) &= Q(z|y) \\
P_\Phi(z) &= P(z) \\
P_\Phi(y|z) &= P(y|z)
\end{aligned}
$$

This assumption allows each distribution to be independently optimized while holding the others fixed.

The β-VAE

β-VAE: Learning Basic Visual Concepts With A Constrained Variational Framework, Higgins et al., ICLR 2017.

The β-VAE introduces a parameter β allow control of the rate-distortion trade off.

Indeterminacy of the VAE

VAE: $\Phi^* = \mathrm{argmin}_\Phi \, E_{y,\epsilon} \;\; -\ln \frac{P_\Phi(z)}{Q_\Phi(z|y)} - \ln P_\Phi(y|z)$

Assuming independent universality we can optimize $P_\Phi(z)$ and $P_\Phi(y|z)$ while holding $Q_\Phi(z|y)$ fixed. This gives

$$P^*(z) \;=\; = P_{\mathrm{pop}}(z) = E_y \, Q_\Phi(z|y)$$

$$P^*(y|z) \;=\; P^*_{\mathrm{pop}}(y|z) \propto P(y,z) = \mathrm{Pop}(y) Q_\Phi(z|y)$$

Indeterminacy of the VAE

$$
\begin{aligned}
\Phi^* &= \operatorname*{argmin}_{\Phi} E_{y,\epsilon}\ -\ln \frac{P_{\mathrm{pop}}(z)}{Q_\Phi(z|y)} - \ln P_{\mathrm{pop}}(y|z) \\
&= \operatorname*{argmin}_{\Phi} \quad I_\Phi(y,z)\ +\ H_\Phi(y|z) \\
&= \operatorname*{argmin}_{\Phi} \qquad H_{\mathrm{pop}}(y)
\end{aligned}
$$

But $H_{\mathrm{pop}}(y)$ is independent of Φ.

Any choice of $Q_\Phi(z|y)$ gives optimal modeling of y.

Indeterminacy of the VAE

$$\Phi^* = \underset{\Phi}{\operatorname{argmin}} \; I_\Phi(y,z) + H_\Phi(y|z)$$

The choice of $Q_\Phi(z|y)$ does not influence the value of the objective function but controls $I(y,z)$.

We have $0 \leq I(y,z) \leq H(y)$ with the full range possible.

The β-VAE

To control $I(y, z)$ we introduce a weighting β

$$\Phi^* = \operatorname*{argmin}_{\Phi} \; \beta I_\Phi(y, z) + H_\Phi(y|z)$$

$$\beta\text{-VAE} \;\; \Phi^* = \operatorname*{argmin}_{\Phi} E_{y,\epsilon} \; -\beta \ln \frac{P_\Phi(z)}{Q_\Phi(z|y)} - \ln P_\Phi(y|z)$$

For $\beta < 1$ we no longer have an upper bound on $H_{\mathrm{pop}}(y)$ but we can force the use of z (avoid posterior collapse).

For $\beta > 1$ the bound on $H_{\mathrm{Pop}}(y)$ becomes weaker and the latent variables carry less information.

RDAs vs. β-VAEs

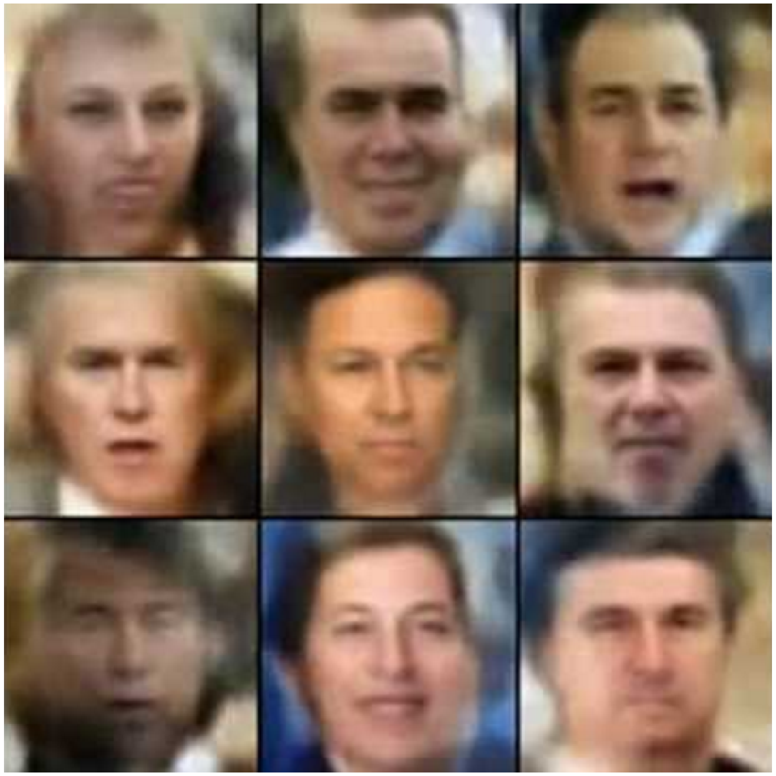
RDAs and β-VAEs are essentially the same.

$$\text{RDA: } \Phi^* = \mathrm{argmin}_\Phi \, E_{y,z\sim Q_\Phi(z|y)} \; -\ln \frac{P_\Phi(z)}{Q_\Phi(z|y)} + \lambda \mathrm{Dist}(y, y_\Phi(z))$$

$$\beta\text{-VAE: } \Phi^* = \mathrm{argmin}_\Phi \, E_{y,z\sim Q_\Phi(z|y)} \; -\beta \ln \frac{P_\Phi(z)}{Q_\Phi(z|y)} - \ln P_\Phi(y|z)$$

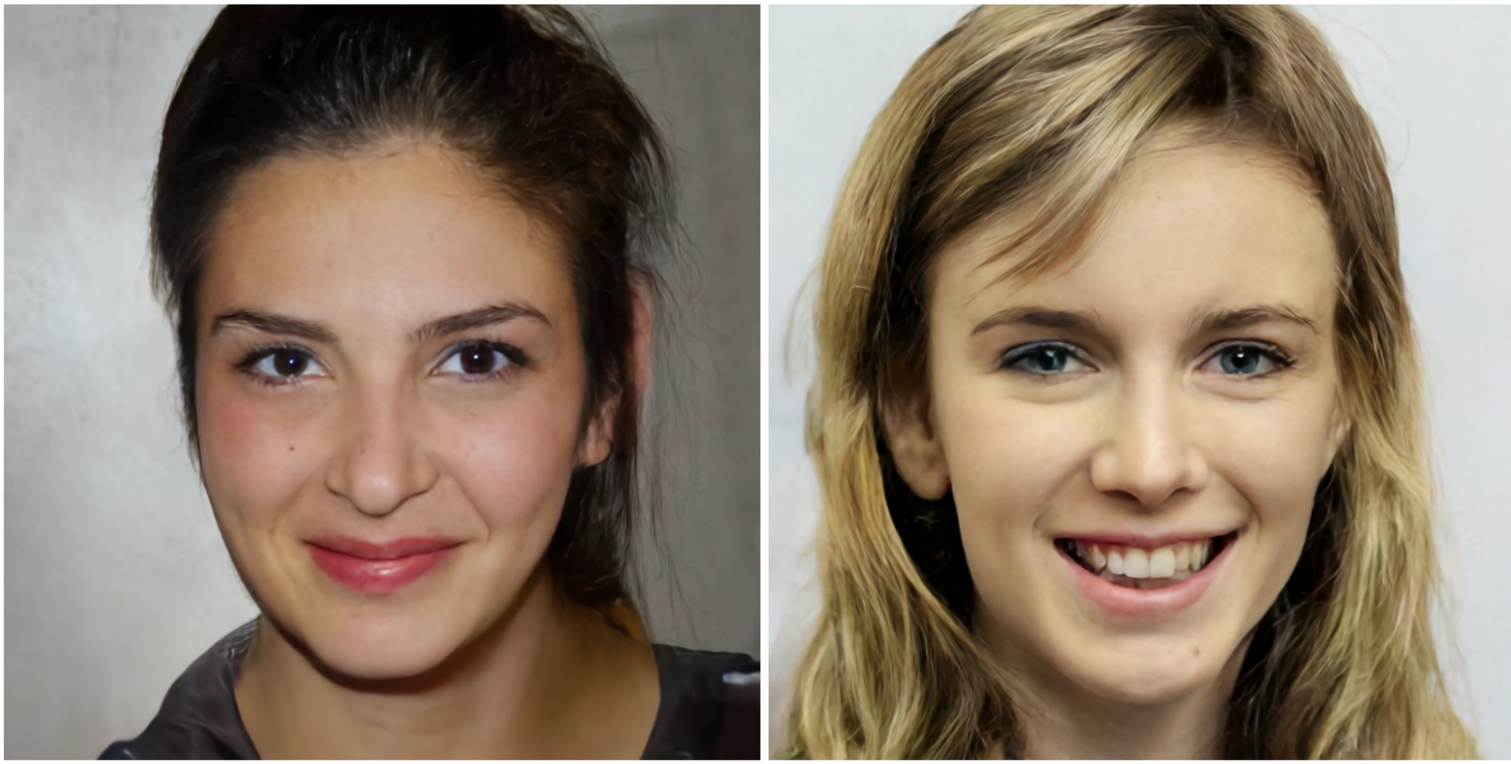
VAEs 2013

Sample $z \sim \mathcal{N}(0, I)$ and compute $y_\Phi(z)$



[Alec Radford]

VAEs 2019



VQ-VAE-2, Razavi et al. June, 2019

VAEs 2019

Figure 1: Class-conditional 256x256 image samples from a two-level model trained on ImageNet.

VQ-VAE-2, Razavi et al. June, 2019

END