TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2020

Variational Auto-Encoders (VAEs)

Posterior Collapse

β-VAEs

Encoder Autonomy

VAE

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

$$z = z_\Psi(y,\epsilon) \quad \epsilon \text{ parameter independent noise}$$

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

VAE

$$P_\Psi(y,z) \;=\; \mathrm{Pop}(y)P_\Psi(z|y)$$

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

$$\;=\; \operatorname*{argmin}_{\Psi,\Phi}\; \hat{I}_{\Psi,\Phi}(y,z) \;+\; \hat{H}_{\Psi,\Phi}(y|z)$$

$$I_\Psi(y,z) \le \hat{I}_{\Psi,\Phi}(y,z) \qquad H_\Psi(y|z) \le \hat{H}_{\Psi,\Phi}(y|z)$$

$$H(y) = I_\Psi(y,z) + H_\Psi(y|z)$$

Inequalities hold with equality under universal expressiveness.

Posterior (Encoder) Collapse

$$\Psi^*, \Phi^* = \operatorname*{argmin}_{\Psi,\Phi} \; \hat{I}_{\Psi,\Phi}(y, z) + \hat{H}_{\Psi,\Phi}(y|z)$$

Consider a trivial encoder with $P_\Psi(z^*|y) = 1$ and $\hat{P}_\Phi(z^*) = 1$ for a fixed value z^* independent of y yielding $\hat{I}_{\Psi,\Phi}(y, z) = 0$.

Under universal expressiveness we have $\hat{P}_{\Phi^*}(y|z) = \mathrm{Pop}(y)$ 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

$$
\begin{aligned}
P_\Psi(y,z) &= \mathrm{Pop}(y)P_\Psi(z|y) \quad \text{The sampling distribution on } y,\ z \\
\text{VAE:}\quad \Psi^*,\Phi^* &= \underset{\Psi,\Phi}{\mathrm{argmin}} \left(E_{y,z}\ \ln\frac{P_\Psi(z|y)}{\hat{P}_\Phi(z)}\right) + \left(E_{y,z}\ -\ln\hat{P}_\Phi(y|z)\right) \\
&= \underset{\Psi,\Phi}{\mathrm{argmin}}\ \beta\ \hat{I}_{\Psi,\Phi}(y,z) + \hat{H}_{\Psi,\Phi}(y|z) \\
\text{RDA:}\quad \Psi^*,\Phi^* &= \underset{\Psi,\Phi}{\mathrm{argmin}}\ \hat{I}_{\Psi,\Phi}(y,z) + \lambda\ \mathrm{Dist}_\Phi(y|z)
\end{aligned}
$$

The β-VAE introduces a rate-distortion tradeoff parameter into the VAE. $\beta < 1$ may avoid posterior collapse. $\beta > 1$ may improve interpretability.

The Universality Theorem for β-VAEs

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

$$I_\Psi(y,z) \leq \hat{I}_{\Psi,\Phi}(y,z) \qquad H_\Psi(y|z) \leq \hat{H}_{\Psi,\Phi}(y|z)$$

$$H(y) = I_\Psi(y,z) + H_\Psi(y|z)$$

Assuming universaltiy, optimizing $\hat{P}_\Phi(z)$ while holding $P_\Psi(z|y)$ and $\hat{P}_\Phi(y|z)$ fixed drives the first inequality to equality.

Optimizing $\hat{P}_\Phi(y|z)$ while holding $P_\Psi(z|y)$ and $\hat{P}_\Phi(z)$ fixed drives the second inequality to equality.

Encoder Autonomy

Assuming universality, optimizing Φ for any fixed value of Ψ yields the population distribution on y.

This implies that we can add any loss on Ψ alone and the universality theorem still holds.

$$\text{VAE:} \quad \Psi^*, \Phi^* = \operatorname*{argmin}_{\Psi,\Phi} \; \beta \, \hat{I}_{\Psi,\Phi}(y,z) \; + \; \hat{H}_{\Psi,\Phi}(y|z) + \mathcal{L}(\Psi)$$

$$I_\Psi(y,z) \leq \hat{I}_{\Psi,\Phi}(y,z) \qquad H_\Psi(y|z) \leq \hat{H}_{\Psi,\Phi}(y|z)$$

$$H(y) = I_\Psi(y,z) + H_\Psi(y|z)$$

END