TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2020

Posterior Collapse

VAE Non-Identifiability

β-VAEs

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.

$$\hat{P}_\Phi(z|y) = P_\Phi(z|y) = 1$$

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

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

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

Since Φ can independently parameterize each distribution, we will here assume independent universality — that Φ can represent any triple of distributions $\hat{P}(z|y)$, $P(z)$ and $P(y|z)$.

Independent Universality

More formally, we assume that for any triple of distributions $\hat{P}(z|y)$, $P(z)$ and $P(y|z)$ there exists a Φ that simultaneously satisfies

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

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

VAE Non-Identifiability

A model is non-identifiable if different model parameters yield the same data distribution and hence cannot be distinguished based on the data.

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

We will now hold $\hat{P}_\Phi(z|y)$ fixed at an arbitrary distribution and optimize $P_\Phi(z)$ and $P_\phi(y|z)$ assuming independent universality.

VAE Non-Identifiability

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

We will show that the optimal distributions for $P_\Phi(z)$ and $P_\Phi(y|z)$ occur when these are the distributions defined by $y \sim \mathrm{Pop}$ and $z \sim \hat{P}_\Phi(z|y)$.

$$P^*(z) = E_{y\sim\mathrm{Pop}}\; \hat{P}_\Phi(z|y)$$

$$P^*(y|z) = \frac{\mathrm{Pop}(y)\hat{P}_\Phi(z|y)}{P^*(z)}$$

VAE Non-Identifiability

$$E \;\; -\ln \frac{P^*(z)P^*(y|z)}{\hat{P}_\Phi(z|y)}$$

$$= \; E \;\; -\ln \frac{P^*(z)P^*(y|z)}{\mathrm{Pop}(y)\hat{P}_\Phi(z|y)} - \ln \mathrm{Pop}(y)$$

$$= \; E \;\; -\ln \mathrm{Pop}(y) \;\; = \;\; H(y)$$

Hence any choice of $\hat{P}_\Phi(z|y)$ gives optimal modeling of y.

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.

The β-VAE

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

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

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

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

Noisy channel RDAs and β-VAEs are essentially the same.

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

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

END