1 Information Bottleneck Principle for Sequential Data.

Here the aim is to show the relationship between FAVAE and the information bottleneck for sequential data. Consider the information bottleneck object:

$$\max I\left(z; x_{1:T}\right) \qquad \text{s.t. } \left|I\left(\hat{x}_{1:T}; z\right) - C\right| = 0 \tag{1}$$

is expanded from [Alemi *et al.*2016] to sequential data. We need to distinguish between $\hat{x}_{1:T}$ and $x_{1:T}$, where $x_{1:T}$ is the true distribution and $\hat{x}_{1:T}$ is the empirical distribution created by sampling from the true distribution. We maximize the mutual information of true data $x_{1:T}$ and z while constraining the information contained in the empirical data distribution. We do this by using Lagrange multiplier:

$$I\left(z; x_{1:T}\right) + \beta\left|I\left(\hat{x}_{1:T}; z\right) - C\right|, \tag{2}$$

where β is a constant. For the first term,

$$\begin{aligned} I\left(z; x_{1:T}\right) &= \iint p\left(z; x_{1:T}\right) \log \frac{p\left(x_{1:T}|z\right)}{p\left(x_{1:T}\right)} dx_{1:T} dz \\ &= \iint p\left(z; x_{1:T}\right) \log p\left(x_{1:T}|z\right) dx_{1:T} dz \\ &- \iint p\left(x_{1:T}; z\right) \log p\left(x_{1:T}\right) dx_{1:T} dz \\ &= \iint p\left(x_{1:T}\right) p\left(z|x_{1:T}\right) \log p\left(x_{1:T}|z\right) dx_{1:T} dz \\ &+ H\left(x_{1:T}\right) \\ &\sim \frac{1}{N} \sum_i \left[p\left(z|\left(x_{1:T}\right)_i\right) \log p\left(\left(x_{1:T}\right)_i|z\right)\right] \\ &+ H\left(x_{1:T}\right), \end{aligned} \tag{3}$$

where $H\left(x_{1:T}\right)$ is entropy, which can be neglected in optimization. The last line is Monte Carlo approximation. For the second term,

$$\begin{aligned} I\left(\hat{x}_{1:T}; z\right) &= \int p\left(z|\hat{x}_{1:T}\right) p\left(\hat{x}_{1:T}\right) \log \frac{p\left(z|\hat{x}_{1:T}\right)}{p\left(z\right)} dz d\hat{x}_{1:T} \\ &\sim \frac{1}{N} \sum_j p\left(z|\left(x_{1:T}\right)_j\right) \log \frac{p\left(z|\left(x_{1:T}\right)_j\right)}{p\left(z\right)} dz \\ &= \frac{1}{N} \sum_j D_{\mathrm{KL}}\left(p\left(z|\left(x_{1:T}\right)_j\right) || p\left(z\right)\right). \end{aligned} \tag{4}$$

As a result,

$$
\begin{aligned}
& I\left(z;x_{1:T}\right)+\beta\left|I\left(\tilde{x}_{1:T};z\right)-C\right| \\
& \leq \frac{1}{N}\sum_{i}\left[p\left(z|\left(x_{1:T}\right)_{i}\right)\log p\left(\left(x_{1:T}\right)_{i}|z\right)\right] \\
& +\frac{1}{N}\sum_{j}D_{\mathrm{KL}}\left(p\left(z|\left(x_{1:T}\right)_{j}\right)||p\left(z\right)\right). \qquad (5)
\end{aligned}
$$

For convenience of calculation, we use x_i sampled from mini-batch data for both the reconstruction term and the regularization term. This is only an approximation. If the information bottleneck principle is followed completely, it is better to use different batch data for the reconstruction and regularization terms.

References

[Alemi *et al.*2016] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.