Enhancing Variational Autoencoders with Mutual Information Neural Estimation for Text Generation

1 Appendix

We define the *variational joint distribution* over $\mathbf{x}$ and $\mathbf{z}$ as:

$$q_\phi(\mathbf{x}, \mathbf{z}) = \frac{e^{f_\psi(\mathbf{x},\mathbf{z})}}{Z_\psi} q(\mathbf{x}) q_\phi(\mathbf{z}) \tag{1}$$

where $f_\psi(\mathbf{x}, \mathbf{z})$ is an energy function with parameters ψ and the partition function Z_ψ is defined as $\mathbb{E}_{q(\mathbf{x})q_\phi(\mathbf{z})}[e^{f_\psi(\mathbf{x},\mathbf{z})}]$. In particular, $f_\psi(\mathbf{x}, \mathbf{z})$ is approximated by neural networks.

The lower bound on the mutual information $\mathbb{I}_q[\mathbf{x}, \mathbf{z}]$ is given by:

$$\begin{aligned}
\mathbb{I}_q[\mathbf{x}, \mathbf{z}] &= \mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z})}\left[\log \frac{q_\phi(\mathbf{x}, \mathbf{z})}{q(\mathbf{x}) q_\phi(\mathbf{z})}\right] \\
&= \mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z})}[f_\psi(\mathbf{x}, \mathbf{z})] - \mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z})}[\log Z_\psi] \\
&= \mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z})}[f_\psi(\mathbf{x}, \mathbf{z})] - \log Z_\psi \\
&= \mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z})}[f_\psi(\mathbf{x}, \mathbf{z})] - \log\big[\mathbb{E}_{q(\mathbf{x})q_\phi(\mathbf{z})}\left[e^{f_\psi(\mathbf{x},\mathbf{z})}\right]\big] \\
&\geq \mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z})}[f_\psi(\mathbf{x}, \mathbf{z})] - \xi \cdot \mathbb{E}_{q(\mathbf{x})q_\phi(\mathbf{z})}\big[e^{f_\psi(\mathbf{x},\mathbf{z})}\big] + \log(\xi) + 1
\end{aligned} \tag{2}$$

Proposition 1. With the fixed ϕ and ξ, the optimal energy function $f^*_\psi(\mathbf{x}, \mathbf{z})$ according to the objective in Eq.(2) is given by

$$\begin{aligned}
f^*_\psi(\mathbf{x}, \mathbf{z}) &= \operatorname*{argmax}_{f_\psi(\mathbf{x},\mathbf{z})} \big\{\mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z})}[f_\psi(\mathbf{x}, \mathbf{z})] - \xi \cdot \mathbb{E}_{q(\mathbf{x})q_\phi(\mathbf{z})}[e^{f_\psi(\mathbf{x},\mathbf{z})}]\big\} \\
\nabla_\psi f_\psi(\mathbf{x}, \mathbf{z}) &= q_\phi(\mathbf{x}, \mathbf{z}) - \xi \cdot [q(\mathbf{x}) q_\phi(\mathbf{z})] e^{f_\psi(\mathbf{x},\mathbf{z})} = 0 \\
f^*_\psi(\mathbf{x}, \mathbf{z}) &= \log q_\phi(\mathbf{x}, \mathbf{z}) - \log[q(\mathbf{x}) q_\phi(\mathbf{z})] - \log(\xi)
\end{aligned} \tag{3}$$

When $\xi = 1$, $f^*_\psi(\mathbf{x}, \mathbf{z})$ becomes essentially pointwise mutual information. This means that the energy function assigns zero probability to the samples independently from $q(\mathbf{x}) q_\phi(\mathbf{z})$.

With the optimal function $f^*_\psi(\mathbf{x}, \mathbf{z})$ defined, the max-max objective in Equation (2) can be reformulated as:

$$
\begin{aligned}
C(\phi, \psi^*) &= \mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z})}[f^*_\psi(\mathbf{x},\mathbf{z})] - \xi \cdot \mathbb{E}_{q(\mathbf{x})q_\phi(\mathbf{z})}\left[e^{f^*_\psi(\mathbf{x},\mathbf{z})}\right] + \log(\xi) + 1 \\
&= \mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z})}\left[\log \frac{q_\phi(\mathbf{x},\mathbf{z})}{\xi \cdot q(\mathbf{x})q_\phi(\mathbf{z})}\right] - \xi \cdot \mathbb{E}_{q(\mathbf{x})q_\phi(\mathbf{z})}\left[\frac{q_\phi(\mathbf{x},\mathbf{z})}{\xi \cdot q(\mathbf{x})q_\phi(\mathbf{z})}\right] + \log(\xi) + 1 \\
&= -\log(\xi) + \mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z})}\left[\log \frac{q_\phi(\mathbf{x},\mathbf{z})}{q(\mathbf{x})q_\phi(\mathbf{z})}\right] - \mathbb{E}_{q(\mathbf{x})q_\phi(\mathbf{z})}\left[\frac{q_\phi(\mathbf{x},\mathbf{z})}{q(\mathbf{x})q_\phi(\mathbf{z})}\right] + \log(\xi) + 1 \\
&= \mathrm{KL}[q_\phi(\mathbf{x},\mathbf{z}) || q(\mathbf{x})q_\phi(\mathbf{z})] \qquad (4)
\end{aligned}
$$

Thus, maximizing the lower bound on the MI with respect to ϕ and ψ is equivalent to maximizing the KL divergence between the joint distribution and two marginal distributions.

We can then compute the gradients of Eq.(4) with respect to ϕ and θ for the optimization. While it is easy to compute the gradient with respect to θ, the gradient with respect to ϕ is hard to compute since $C(\phi, \psi^*)$ itself depends on ϕ. Actually, when the function $f_\psi(\mathbf{x},\mathbf{z})$ is optimal, the expectation of the gradients becomes zero, that is

$$
\mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z})}[\nabla_\phi f^*_\psi(\mathbf{x},\mathbf{z})] - \mathbb{E}_{q(\mathbf{x})q_\phi(\mathbf{z})}\left[\nabla_\phi e^{f^*_\psi(\mathbf{x},\mathbf{z})}\right] = 0 \qquad (5)
$$

Thus, we can ignore gradients in the optimization when $f_\psi(\mathbf{x},\mathbf{z})$ is optimal.

Proof.

$$
\mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z})}[\nabla_\phi f^*_\psi(\mathbf{x},\mathbf{z})] = \underbrace{\mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z})}[\nabla_\phi \log q_\phi(\mathbf{z}|\mathbf{x})]}_{\circledast} - \underbrace{\mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z})}[\nabla_\phi \log q_\phi(\mathbf{z})]}_{\circledcirc} \qquad (6)
$$

For the $\circledast$ part,

$$
\begin{aligned}
\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\nabla_\phi \log q_\phi(\mathbf{z}|\mathbf{x})] &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\frac{1}{q_\phi(\mathbf{z}|\mathbf{x})}\nabla_\phi q_\phi(\mathbf{z}|\mathbf{x})\right] \\
&= \int_{\mathbf{z}} \nabla_\phi q_\phi(\mathbf{z}|\mathbf{x}) \\
&= 0 \qquad (7)
\end{aligned}
$$

For the $\circledcirc$ part,

$$
\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\nabla_\phi \log q_\phi(\mathbf{z})] = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\frac{1}{q_\phi(\mathbf{z})}\nabla_\phi q_\phi(\mathbf{z})\right] \qquad (8)
$$

$$
\begin{aligned}
\mathbb{E}_{q_\phi(\mathbf{z})}[\nabla_\phi e^{f^*_\psi(\mathbf{x},\mathbf{z})}] &= \mathbb{E}_{q_\phi(\mathbf{z})}\left[\nabla_\phi \frac{q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z})}\right] \\
&= \mathbb{E}_{q_\phi(\mathbf{z})}\left[\frac{\nabla_\phi q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z})} - \frac{q_\phi(\mathbf{z}|\mathbf{x})\nabla_\phi q_\phi(\mathbf{z})}{q_\phi(\mathbf{z})^2}\right] \\
&= -\mathbb{E}_{q_\phi(\mathbf{z})}\left[\frac{q_\phi(\mathbf{z}|\mathbf{x})\nabla_\phi q_\phi(\mathbf{z})}{q_\phi(\mathbf{z})^2}\right] \\
&= -\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\frac{1}{q_\phi(\mathbf{z})}\nabla_\phi q_\phi(\mathbf{z})\right] \qquad (9)
\end{aligned}
$$

Therefore,

$$\mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z})}[\nabla_\phi f_\psi^*(\mathbf{x},\mathbf{z})] - \mathbb{E}_{q(\mathbf{x})q_\phi(\mathbf{z})}[\nabla_\phi e^{f_\psi^*(\mathbf{x},\mathbf{z})}] = 0 \tag{10}$$