Supplementary material

1 Product of experts

As opposed to mixture of experts, product of experts is designed for “and” operations. Intuitively, the probability of a product of experts will be high only when all experts agree on high probability. The formula of a product of experts is:

$$p(x|\theta) = \frac{1}{Z} \prod_i p(x|\theta_i)$$  \hspace{1cm} (1)

When each expert is a Gaussian distribution, we can derive the analytic solution as follows:

$$C^{-1} = \sum_i C_i^{-1}$$  \hspace{1cm} (2)

$$\mu = C(\sum_i C_i^{-1}\mu_i)$$  \hspace{1cm} (3)

where $C$ and $\mu$ are covariance matrix and mean for the Gaussian $p(x|\theta)$ respectively. See for derivations. This formulation shows that when composing multiple distributions, the resulting Gaussian distribution will always be narrower than each original Gaussian.

Potential problem under variational autoencoder

Under a VAE setting, a global prior distribution $p(0, I)$ contains "no information". For example, when an encoder distribution $q(z|x)$ is compressed, the KL divergence on some dimensions will be approximately zero, namely information will be carried by other dimensions to reconstruct the data. Those no information dimensions can have approximately $p(0, I)$ distributions. However, the products of this prior distribution will actually cause more and more narrow Gaussian distributions, which leads to a "ghost" information $KL(\frac{1}{2} \prod p(0, I)||p(0, I))$. This means, when multiple concepts are fused by PoE, those dimensions carrying no information will be more and more narrow due to this effect.

The same situation happens for other similar cases: when "green" concepts are fused multiple times, the information carried by text is only "green", but the fused distribution will be more peaky. The mean of product distribution will probably also be moved based on the number of identical experts. Or in general in all cases, any dimensions, after being multiplied $k$ times, will have extra information.

2 Gated product of experts

Instead of directly using PoE on the original distributions, we do a simple modification by learning a conditional distributions based on distributions to be composed. Specifically, we take the distribution information (mean and variance in our method) and generate a set of gates $\{g\}$ by a neural network $f(\{p_i\})$. Then we gate the information on the original distribution $i$ and derive the following distribution $p(z_i|g_i^\mu \circ \mu_i, g_i^\sigma \circ \sigma_i)$, where $\circ$ is the element-wise product. Then we follow the standard PoE Eqn. 2 and Eqn. 3 to compose the distributions.