Multimodal Variational Autoencoders

Achint Kumar

Shell USA

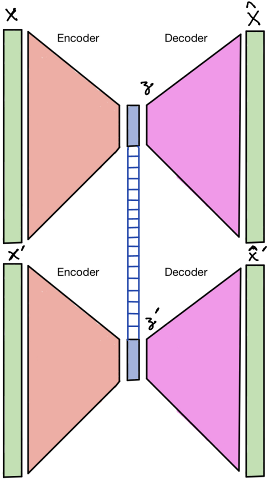
November 13, 2023

Desiderata

Desiderata

Multi-modal Variational Autoencoders

1. Multi-modal VAE is a variant of vanilla VAE in which multiple datasets can be jointly input to the network.
2. It can model nonlinear correlations between modalities.
3. In our case, the two modalities would be the same but it would be conditioned on AI and SI



Three main multimodal VAE frameworks

- MVAE (2018): Uses Product of Experts
- MMVAE (2019): Uses Mixture of Experts
- MoPoE-VAE (2021): Uses Mixture of Product of Experts (current SOTA)

Constructing Multimodal VAE

For any multi-modal VAE variational free energy is given by,

$$\begin{aligned}\mathcal{F}_{var}(x, x') &= \underbrace{||x - \hat{x}||_2^2 + ||x' - \hat{x'}||_2^2}_{reconstruction} + \underbrace{D_{KL}(q_\phi(z|x, x')||p(z))}_{regularizer} \\ &= -\mathbb{E}_{q(z|x,x')}[p(x, x'|z)] + D_{KL}(q_\phi(z|x, x')||p(z)) \\ &= -\mathbb{E}_{q_\phi(z|x,x')}\left[\log \frac{p_\theta(z, x, x')}{q_\phi(z|x, x')}\right]\end{aligned}$$

This equation will be the starting point for constructing our multimodal VAE.

Desiderata

Step 1: Constructing Importance Weighted Autoencoders (IWAE)

Recall we had,

$$\mathcal{F}_{var}(x_1, x_2) = -\mathbb{E}_{z\sim q_\phi}\left[\log \frac{p_\theta(z, x_1, x_2)}{q_\phi(z|x_1, x_2)}\right] = -\mathbb{E}_{z\sim q_\phi}[\log w]$$

where $w = \frac{p_\theta(z,x_1,x_2)}{q_\phi(z|x_1,x_2)}$. We can lower the energy by using K samples instead of 1

$$IWAE(K) = -\mathbb{E}_{z_{1:K}}\left[\log(\frac{1}{K}\sum_{i=1}^{K} w_i)\right]$$

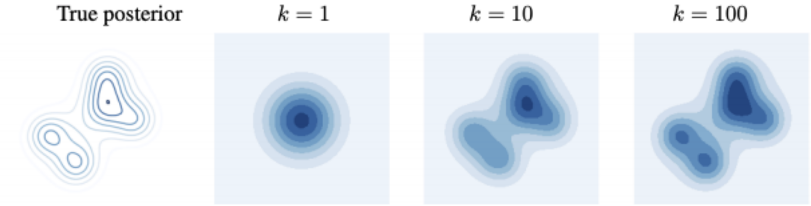
It turns out,

$$\mathcal{F}_{true} \leq IWAE(K) \leq \cdots \leq IWAE(1) \leq \mathcal{F}_{var}$$

Pros and Cons of IWAE

Pros: As K increases,

- IWAE(K) bound with $\mathcal{F}_{true}$ becomes tighter
- The effect of overly simplistic q_ϕ diminishes



Cons: As K increases,

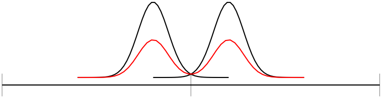
- ϕ gradient of IWAE(K) becomes more variable

Solution: Use Doubly Reparametrized Gradient Estimator (DReG)

Step 2: Formulating Posterior Distribution

There are two ways to define the multimodal posterior, $q_\phi(z|x_1, x_2)$ in terms of unimodal posterior, $q_{\phi_1}(z_1|x_1)$ and $q_{\phi_2}(z_2|x_2)$:

Mixture of Experts



Black: Unimodal posterior
Red: Multimodal posterior

- Posterior is written as sum:

$$q_\phi(z|x_1, x_2) = \sum_i \alpha_i q_{\phi_i}(z_i|x_i)$$

- Like a healthy relationship (each expert can make decision). Does what either likes.

Product of Experts

Black: Unimodal posterior
Red: Multimodal posterior

- Posterior is written as product:

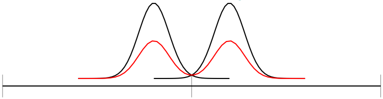
$$q_\phi(z|x_1, x_2) = p(z) \prod_{i=1}^{2} q_{\phi_i}(z_i|x_i)$$

- Like UN Security Council (each expert has veto power). Does what neither likes.

Step 2: Formulating Posterior Distribution

There are two ways to define the multimodal posterior, $q_\phi(z|x_1, x_2)$ in terms of unimodal posterior, $q_{\phi_1}(z_1|x_1)$ and $q_{\phi_2}(z_2|x_2)$:

Mixture of Experts



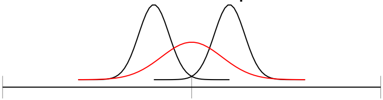
Black: Unimodal posterior
Red: Multimodal posterior

- Posterior is written as sum:

$$q_\phi(z|x_1, x_2) = \sum_i \alpha_i q_{\phi_i}(z_i|x_i)$$

- Like a healthy relationship (each expert can make decision). Does what either likes.

Product of Experts



Black: Unimodal posterior
Red: Multimodal posterior

- Posterior is written as product:

$$q_\phi(z|x_1, x_2) = p(z)\prod_{i=1}^{2} q_{\phi_i}(z_i|x_i)$$

- Like UN Security Council (each expert has veto power). Does what neither likes.

Free energy Formulation

Recall we had,

$$\mathcal{F}_{var}(x_1, x_2) = -\mathbb{E}_{q_\phi(z|x_1,x_2)}\left[\log \frac{p_\theta(z, x_1, x_2)}{q_\phi(z|x_1, x_2)}\right]$$

After using mixture of experts,
$q_\phi(z|x_1, x_2) \sim [q_{\phi_1}(z_1|x_1) + q_{\phi_2}(z_2|x_2)]$ We write,

$$\mathcal{F}_{var}(x_1, x_2) = -\sum_{m=1}^{2} \mathbb{E}_{z_m \sim q_{\phi_m}(z_m|x_m)}\left[\log \frac{p_\theta(z_m, x_1, x_2)}{q_\phi(z_m|x_1, x_2)}\right]$$

1. MMVAE uses IWAE with DReG estimator combined with MoE
2. Prior and Posterior are assumed to follow Laplace distribution

Final Thoughts

The code would require significant modifications from what we are currently using. I suggest following order:

1. Implement IWAE. Minor modifications to code.
2. Improve IWAE with DReG estimator (or alternatives).
3. Pursue multimodal VAEs