Improving Multimodal Joint Variational Autoencoders through Normalizing Flows and Correlation Analysis

Week 2

Konstantin Yakovlev [1]

[1]MIPT
Moscow, Russia

MIPT 2023

Background[1]

Joint Multimodal Variational Autoencoders

$$p_\theta(x_1, x_2, z) = p(z)p_\theta(x_1|z)p_\theta(x_2|z),$$

$$\log p(x_1, x_2) \geq \mathbb{E}_{q_\phi(z|x_1,x_2)} \log \frac{p(x_1, x_2|z)}{q_\phi(z|x_1, x_2)} = \mathcal{L}(x_1, x_2).$$

If we want to generate one modality from the other: $z \sim q_{\phi_j}(z|x_j)$, $x_i \sim p_\theta(x_i|z)$. Auxiliary distributaions are optimized using:

$$\mathcal{L}_{\text{JM}}(x_1, x_2) = \sum_{i=1}^{2} \text{KL}(q_\phi(z|x_1, x_2)||q_{\phi_i}(z|x_i)) \to \min_{\phi_{1,2}}$$

The final objective: $\mathcal{L}(.,.) - \alpha\mathcal{L}_{\text{JM}}(.,.)$.

[1] Senellart A. et. al, Improving Multimodal Joint Variational Autoencoders through Normalizing Flows and Correlation Analysis, 2023

The proposed model

Challenge 1: α induces a trade-off between the reconstruction and conditional generation.
Solution: Two-stage training: 1) $q_\phi(z|x_1, \ldots, x_m)$. 2) $q_{\phi_i}(z|x_i)$.
Challenge 2: the unimodal postriors need a lot of flexibility
Solution: Enrich the unimodal posterior with Normalizng Flows.
Observation 3: A second observation is that to generate a modality from another one we only need the information shared by both and not the entire data.
Solution: Extract the shared information using DCCA (trained with minimizing **the sum of the pairwise CCA** objectives). As a nice bonus, we reduce the dimensionality of the modalities spaces which simplifies the task of modeling the unimodal posteriors. ***General DCCA embeddigns might be replaced by another data-specific method***
Challenge 4: JMVAE is not scalabele to more than 2 modalities, since for each subset of modalities S we need to have its encoder $q_{\phi_i}(z|(x_i)_{i\in S})$.
Solution: Product of experts:

$$q(z|(x_i)_{i\in S}) \propto p(z)^{1-|S|} \prod_{i\in S} q_{\phi_i}(z|x_i) \Rightarrow \text{sampling with HMC}$$