Score-Based Multimodal Autoencoders

Week 12

Konstantin Yakovlev [1]

[1]MIPT
Moscow, Russia

MIPT 2023

Score-Based Multimodal Autoencoders[1]

Challenge: conditioning on more modalities often reduces the quality of the generated modality.

Solution: instead of learning a joint posterior, try to model a joint prior $p_\theta(\mathbf{z}_{1:M})$. This allows us to better model correlation among modalities.

The Method: Assume that

$$p(\mathbf{x}_{1:M}|\mathbf{z}_{1:M}) = \prod_{k=1}^{M} p(\mathbf{x}_k|\mathbf{z}_k),$$

$$q(\mathbf{z}_{1:M}|\mathbf{x}_{1:M}) = \prod_{k=1}^{M} q(\mathbf{z}_k|\mathbf{x}_k).$$

Then, $\mathrm{ELBO} = \sum_k \mathrm{ELBO}_k$ if the prior is decomposable.

Two-stage training:

- Train the autoencoders separately, assuming that $p(\mathbf{z}_m) = \mathcal{N}(0, \mathbf{I})$.
- Freeze the autoencoders and leran a joint prior $p(\mathbf{z}_{1:M})$. More precisely, we need a score function $s_\theta(\mathbf{z}_{1:M})$ to sample from the prior.

Finally, it becomes trivial to sample from any subset of missing modalities using Langevin dynamics.

[1] Wesego D. et. al, Score-Based Multimodal Autoencoders, 2023

Dimension reduction via score ratio matching[2]

Algorithm 1 Estimate low-dimensional subspace U_r

1: **Input**: Target data $\{x_i\}_{i=1}^n \sim \pi$, and user tolerance $\varepsilon > 0$
2: Center the mean and scale data by the Cholesky factor of the empirical precision matrix.
3: Solve $\min_{s_\theta} F(s_\theta)$ to obtain the score-ratio approximation $s_\theta(x)$.
4: Estimate the diagnostic matrix $\widehat{H} = \frac{1}{n}\sum_{i=1}^n s_\theta(x_i)s_\theta(x_i)^\top$.
5: Compute the eigenpairs of $\widehat{H}$, $(\lambda_i, u_i) \in \mathbb{R}_{\geq 0} \times \mathbb{R}^d$.
6: Set $U = [u_1 \; \dots \; u_n]$ and pick r so that $\widehat{E}_r(U) = \frac{1}{2}(\lambda_{r+1} + \cdots + \lambda_d) < \varepsilon$

$$\pi_r(\mathbf{x}) \propto f(\mathbf{U}_r^\top \mathbf{x})\rho(\mathbf{x}).$$

Proposition: $D_{\mathrm{KL}}(\pi || \pi_r) \leq \frac{1}{2}(\lambda_{r+1} + \ldots + \lambda_d)$.

[2] Baptista R. et. al, Dimension reduction via score ratio matching, 2022

Learning the correlation among the latent variables (proposed)

Challenge: Score-Based Multimodal Autoencoders do not select the dimension of the latent space of each modality properly. Therefore, this makes it more difficult to learn correlations among the modalities.
Solution: reduce the dimension of the latent space of each modality.

$$\pi_r(\mathbf{x}_{1:M}) \propto f(\mathbf{W}^\top \mathbf{x}_{1:M}) \prod_{m=1}^{M} \rho(\mathbf{x}_m), \quad \mathbf{W}^\top = \mathrm{diag}(\mathbf{U}_1^\top, \ldots, \mathbf{U}_M^\top).$$

Proposition: the proposed parametrization does not require an additional computational cost when computing the eigenpairs of $\mathbf{H}$.
Note: other modalities $\mathbf{x}_{\backslash m}$ contribute to the reduction of the dimension of the modality $\mathbf{x}_m$.

Project description

Title: Learning the correlation among the latent variables in the reduced latent space.
Problem: Consider a multimodel generative modeling task. The goal of inference is to sample unobserved modalities given the observed ones.
Data: PolyMnist, CelebAMask-HQ.
Reference: (1) and (2).
Basic solution: Score-Based Multimodal AE.
Proposed solution: Score-Based Multimodal Autoencoders do not select the dimension of the latent space of each modality properly. Therefore, this makes it more difficult to learn correlations among the modalities. Thus, reduce the dimension of the latent space of each modality with Score Ratio Matching.
Novelty: we address the challenge of generative quality degradation when the number of modalities increases.