Score-Based Multimodal Autoencoders

Week 1

Konstantin Yakovlev [1]

[1]MIPT
Moscow, Russia

MIPT 2023

Score-Based Multimodal Autoencoders

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.

Selecting dependent modalities (proposed)

Task: remove independent modalities from $\mathbf{x}_{1:M}$. *Are we really intended to solve it?*
Solution: learn the structure of $s_\theta(\mathbf{z}_{1:M}) = \sum_{i \in \mathcal{I}} s_\theta(\mathbf{z}_i) + s_\theta(\mathbf{z}_{i:i \in \mathcal{D}})$.
The Method: greedily remove modalities one by one, decreasing the score matching objective. Initialize $\mathcal{D}_0$ with $\{1, \dots, M\}$.