Disentangled Representations of the Diffusion Process

CS6360 - Assignment 6

Pranav K Nayak

IIT Hyderabad

The Homomorphism Autoencoder

The HAE involves samples of observation-action sequences $(o_1, g_1, ..., g_{N-1}, o_N)$.

Purpose: Learn representations of observations and actions with structural guarantees:

The Homomorphism Autoencoder

The HAE involves samples of observation-action sequences $(o_1, g_1, ..., g_{N-1}, o_N)$.

Purpose: Learn representations of observations and actions with structural guarantees:

- The representation of observations $h(o) : \mathcal{O} \to Z$ is *group-structured*. It preserves equivariance across commutative diagrams that are well-defined.
- The representation of actions $\rho(g) : G \to GL(Z)$ is *disentangled*. This requires that the space learned by h is decomposable as $Z = Z_1 \oplus Z_2 \oplus Z_3 \oplus ... \oplus Z_n$.
- A decoder $d : Z \to \mathcal{O}$, included to ensure that ρ isn't trivial.

The Homomorphism Autoencoder

The HAE involves samples of observation-action sequences $(o_1, g_1, ..., g_{N-1}, o_N)$.

Purpose: Learn representations of observations and actions with structural guarantees:

- The representation of observations $h(o) : \mathcal{O} \to Z$ is *group-structured*. It preserves equivariance across commutative diagrams that are well-defined.
- The representation of actions $\rho(g) : G \to GL(Z)$ is *disentangled*. This requires that the space learned by h is decomposable as $Z = Z_1 \oplus Z_2 \oplus Z_3 \oplus ... \oplus Z_n$.
- A decoder $d : Z \to \mathcal{O}$, included to ensure that ρ isn't trivial.

The Homomorphism Autoencoder

The HAE involves samples of observation-action sequences $(o_1, g_1, ..., g_{N-1}, o_N)$.

Purpose: Learn representations of observations and actions with structural guarantees:

- The representation of observations $h(o) : \mathcal{O} \to Z$ is *group-structured*. It preserves equivariance across commutative diagrams that are well-defined.
- The representation of actions $\rho(g) : G \to GL(Z)$ is *disentangled*. This requires that the space learned by h is decomposable as $Z = Z_1 \oplus Z_2 \oplus Z_3 \oplus ... \oplus Z_n$.
- A decoder $d : Z \to \mathcal{O}$, included to ensure that ρ isn't trivial.

The HAE Losses

Definition (Latent Prediction Loss)

$$\mathcal{L}^N_{\text{pred}}(\rho, h) = \sum_{t=2}^{N+1} \Big|\Big| h(o_t) - \Big(\prod_{i=1}^{t-1} \rho(g_i)\Big) h(o_1) \Big|\Big|.$$

Definition (Reconstruction Loss)

$$\mathcal{L}^N_{\text{rec}}(\rho, h, d) = \sum_{t=1}^{N+1} \Big|\Big| o_t - d\Big(\Big(\prod_{i\geq 1}^{t-1} \rho(g_i)\Big) h(o_1) \Big) \Big|\Big|.$$

The total loss is a weighted linear combination of the above two.

Theoretical Guarantees

If certain conditions on the underlying world-state space and the group of actions are met, any representations (ρ, h) learned by the HAE that minimize the expected loss are guaranteed to meet the structural requirements specified.

The Forward and Reverse Processes

Diffusion consists of two processes:

- The forward process, characterized by the distribution $q(x_{t+1}|x_t)$
- The reverse process, chatacterized by the distribution $p_\theta(x_{t-1}|x_t)$

Note that the reverse process is what is learned, via a neural network that outputs the parameters of the distribution.

Expanding the ELBO

For diffusion, the evidence is considered to be the entire run of decoded samples $p(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_t - 1|\mathbf{x}_t)$. If we express this as the result of marginalizing the joint over q, we get:

$$\log p(\mathbf{x}) \geq \mathbb{E}_{q(\mathbf{x}_1|x_0)}[\log p_\theta(x_0|x_1)] - D_{KL}(q(\mathbf{x}_T|\mathbf{x}_0) - \sum_{t=2}^{T} \mathbb{E}_{q(\mathbf{x}_T|\mathbf{x}_0)}[D_{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)||p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)]$$

The last term indicates that the ELBO can be maximized by making sure that the learned denoising distribution, $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ is as close as possible to the true denoising distribution, $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$.

$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$

Using Bayes' rule:

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \frac{q(\mathbf{x}_t|\mathbf{x}_{t-1})q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)}$$

From the reparameterization trick, we get that

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0) - \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_{t-1}, (1-\alpha_t)\mathbf{I}),$$

$$q(\mathbf{x}_{t-1}|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0, (1-\bar{\alpha}_{t-1})\mathbf{I}), \text{ and}$$

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}).$$

Plugging these into the expression for $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$, expanding the normals into their functional forms, tells us that $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ is proportional to a normally distributed random variable, with mean $\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\mathbf{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\mathbf{x}_0}{1-\bar{\alpha}_t}$[1] and variance $\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{I}$.
This functional form allows us to impose Gaussian structure on the learned denoising distribution, $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$. Note that the variance for the learned denoiser can be computed once the step-size scheme $\{\alpha_t\}$ is fixed.

[1] Detailed derivations are available at [Luo22]

Denoising is Group-Structured

The mean of the learned denoising distribution cannot be directly computed, since we do not have access to $\mathbf{x}_0$ during denoising. Thus, the mean must come from a learned function μ_θ of $\mathbf{x}$
Therefore, we arrive at $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(x_t), \sigma_t^2)$. Performing reparameterization, we get

$$\mathbf{x}_{t-1} = \mu_\theta(\mathbf{x}_t) + \sigma_t \epsilon, \text{ with } \epsilon \sim \mathcal{N}(\epsilon; \mathbf{0}, \mathbf{I})$$

Thus, we have that the learned denoising step is also an affine transformation, and thus is group-structured with respect to the space of images.

The World-State space

If the observation space (the space of images) is $\mathbb{R}^{n \times n}$, then a simple candidate for the world-state space is $W = \mathbb{R}^{n^2}$, with the mapping b between observation and world-state spaces being the linearization/rasterization operation. For such a mapping:

- Invertibility is guaranteed.
- Diffeomorphism is not proven, but is likely.[2]

[2] If we set $W = \mathbb{R}^{n \times n}$, then this is guaranteed.[Lee03]

Group-Actions in the World-State space

We need a space $W*$ that is diffeomorphic to W, but which permits a *continuous*, *injective* group representation $\rho* : G \to GL(W*)$. One such candidate is the space $\mathbb{R}^{n^2+1}$. In this space, the augmented-matrix form of affine transformations lies in $GL(W*)$, making it a perfectly valid group representation.

Conclusion

Thus, all the requisite spaces and forms of transformations have been identified, and the result is that $\rho(g_i)$, the representation of the i^{th} denoising step, is guaranteed to be group structured, and the representation $h(x_i)$ is guaranteed to be disentangled with respect to the decomposition of $\rho(g_i)$.

references


J.M. Lee, *Introduction to smooth manifolds*, Graduate Texts in Mathematics, Springer, 2003.


Calvin Luo, *Understanding diffusion models: A unified perspective*, 2022.