CS6360: Assignment 4

Extensions

Pranav K Nayak

The Lie Group Assumption

The paper's central assumption is that the group of transformations acting on the world-state space forms a differentiable manifold (i.e. that it is a Lie group). This assumption is surprisingly lenient. Transformations captured by this assumption include:

- All transformations representable through an invertible $n \times n$ matrix with determinant 1[1] (i.e. the group $\mathbb{SL}(\mathbb{R}, n)$).
- All rotations ($\mathbb{SO}(n)$).

Note that matrix multiplication by $n \times n$ matrices cannot capture affine transformations in $\mathbb{R}^n$, but this can be taken care of via an injective mapping into the space $\mathbb{R}^{n+1}$, followed by a matrix multiplication. Similar to what is done when folding in a bias term into a matrix multiplication, :

$$Ax + b \mapsto A'x'.$$

[1] $n \leq 2$

The Group-Algebra Correspondence Assumption

A relatively restrictive assumption is that the Lie group be closed and connected. This is required so that the mapping from the Lie algebra to the Lie group is invertible.
However, this excludes a very important group of transformations, $\mathbb{GL}(\mathbb{R}, n)$, the **general linear group** consisting of all invertible $n \times n$ matrices.
If we relax this assumption, then the algebra captures the structure of the group only in a neighborhood around the identity.
Incorporating the general linear group hinges on proving that the deviation between the group and its linear approximation (the algebra) is "small enough" within a neighborhood of the identity, such that the neighborhood actually contains meaningfull transformations.

Diffusion

Similarities between the diffusion process and the HAE architecture are largely from two observations:

- The step-wise, iterative application of transformations.
- The transformations can be composed in representation space to, in a single step, predict the encoding of a multi-step transformation without performing the actual encoding process.

Obvious Question: How well does diffusion *actually* fit into this autoencoding/transformation representation learning framework?

Additive Noise

It is important to know if the addition of noise permits a group structure. If not, then it's unlikely that the two frameworks can be reconciled.
If additive noise is a group, then the structure of the manifold of the group should then be understood, to know if the group is a Lie group, and if it is closed and/or compact.
Understanding this manifold should be tractable, since additive noise is a translation operation, and thus lies on a subspace of the augmented matrix's space.

HAE + Diffusion

Three immediate directions for combining the two architectures:

- **Replace the translation group with the Gaussian group**. Repeat all experiments from the paper, and attempt to understand what the representation space looks like.
- **Interleave the two processes**. Add a HAE to a diffusion model's architecture[2], optimize for some combined goal.
- **Train a HAE on a pretrained diffusion model**: Have the diffusion model go through either/both of the noising/denoising processes, and train the HAE to learn a representation corresponding to each stage of the process. This should give us a representation of the diffusion process, and with intelligent design it could give benefits similar to latent diffusion models[3].

[2] *how?*

[3] speculation

Variational HAE

The authors of the paper assert that the encoder and decoder of the HAE be deterministic. One very simple way to extend the HAE is to make the autoencoding process variational, allowing greater control over the space of representations.

Changing Losses

The current formulation of the loss, namely

$$\sum_{t=2}^{N+1}\left|\left|h(o_t)-\left(\prod_{i=1}^{t-1}\rho(g_i)\right)h(o_1)\right|\right|_2^2+\sum_{t=1}^{N+1}\left|\left|o_t-d\left(\left(\prod_{i=1}^{t-1}\rho(g_i)\right)h(o_1)\right)\right|\right|_2^2.$$

can be modified. The authors state that any positive function can be used in place of either l_2 norm, leading to two possible extensions:

- **Using distributional losses:** This also plays into the variational HAE idea.
- **Discounting based loss:** If we can theoretically motivate earlier/later predictions/reconstructions being more important, than a discounting/compounding loss could be used, with different stages of the HAE's pass being weighted differently.

Piyushi's Ideas

1. Inducing sparsity through submodularity instead of via the block-diagonal regularizer.
2. There might be benefits to framing latent OT in terms of the HAE architecture.

Extending Homomorphism AE:
On Improving the Disentanglement (Idea 1)

cs18m20p100002

Recap: Loss Functions

$$\rho = \exp \circ \phi$$

$$\mathcal{L}^N_{rec}(\rho, h, d) = \sum_{t=1}^{N+1} \left\| o_t - d\left(\left(\prod_{i \geq 1}^{t-1} \rho(g_i) \right) h(o_1) \right) \right\|_2^2$$

$$\mathcal{L}^N_{pred}(\rho, h) = \sum_{t=2}^{N+1} \left\| h(o_t) - \left(\prod_{i=1}^{t-1} \rho(g_i) \right) h(o_1) \right\|_2^2$$

Recap: Loss Functions

$$\rho = \exp \circ \phi$$

$$\mathcal{L}^N_{rec}(\rho, h, d) = \sum_{t=1}^{N+1} \left\| o_t - d\left(\left(\prod_{i \geq 1}^{t-1} \rho(g_i) \right) h(o_1) \right) \right\|_2^2$$

$$\mathcal{L}^N_{pred}(\rho, h) = \sum_{t=2}^{N+1} \left\| h(o_t) - \left(\prod_{i=1}^{t-1} \rho(g_i) \right) h(o_1) \right\|_2^2$$

For Disentanglement:

$$\mathcal{L}_{sparse}(\rho) = \sum_t \sum_{i \geq 0} \sqrt{\sum_{j \geq i+1,\, k \leq i} \rho_{kj}(g_t)^2 + \rho_{jk}(g_t)^2}$$

$$\phi(\varphi(g)) = \begin{pmatrix} \phi_1(\varphi(g)) & 0 & \dots & 0 \\ 0 & \phi_1(\varphi(g)) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \phi_1(\varphi(g)) \end{pmatrix} \quad (5)$$

While we do not prove that this block diagonal constraint leads to disentanglement, we show through experiments that

Proposal for Disentanglement

$$\rho = \exp \circ \phi$$

$$\mathcal{L}^N_{rec}(\rho, h, d) = \sum_{t=1}^{N+1} \left\| o_t - d\left(\left(\prod_{i \geq 1}^{t-1} \rho(g_i) \right) h(o_1) \right) \right\|_2^2$$

$$\mathcal{L}^N_{pred}(\rho, h) = \sum_{t=2}^{N+1} \left\| h(o_t) - \left(\prod_{i=1}^{t-1} \rho(g_i) \right) h(o_1) \right\|_2^2$$

- If we show that $\mathcal{L}^N_{rec}(\rho, h, d) + \mathcal{L}^N_{pred}(\rho, h)$ is strongly convex in terms of ϕ.
- If exp preserves the strong convexity.

Proposal for Disentanglement

$$\rho = \exp \circ \phi$$

$$\mathcal{L}^N_{rec}(\rho, h, d) = \sum_{t=1}^{N+1} \left\| o_t - d\left(\left(\prod_{i \geq 1}^{t-1} \rho(g_i) \right) h(o_1) \right) \right\|_2^2$$

$$\mathcal{L}^N_{pred}(\rho, h) = \sum_{t=2}^{N+1} \left\| h(o_t) - \left(\prod_{i=1}^{t-1} \rho(g_i) \right) h(o_1) \right\|_2^2$$

- If we show that $\mathcal{L}^N_{rec}(\rho, h, d) + \mathcal{L}^N_{pred}(\rho, h)$ is strongly convex in terms of ϕ.
- If exp preserves the strong convexity.
- Then we will get an equivalent weakly submodular optimization:

Restricted Strong Convexity
Implies Weak Submodularity*

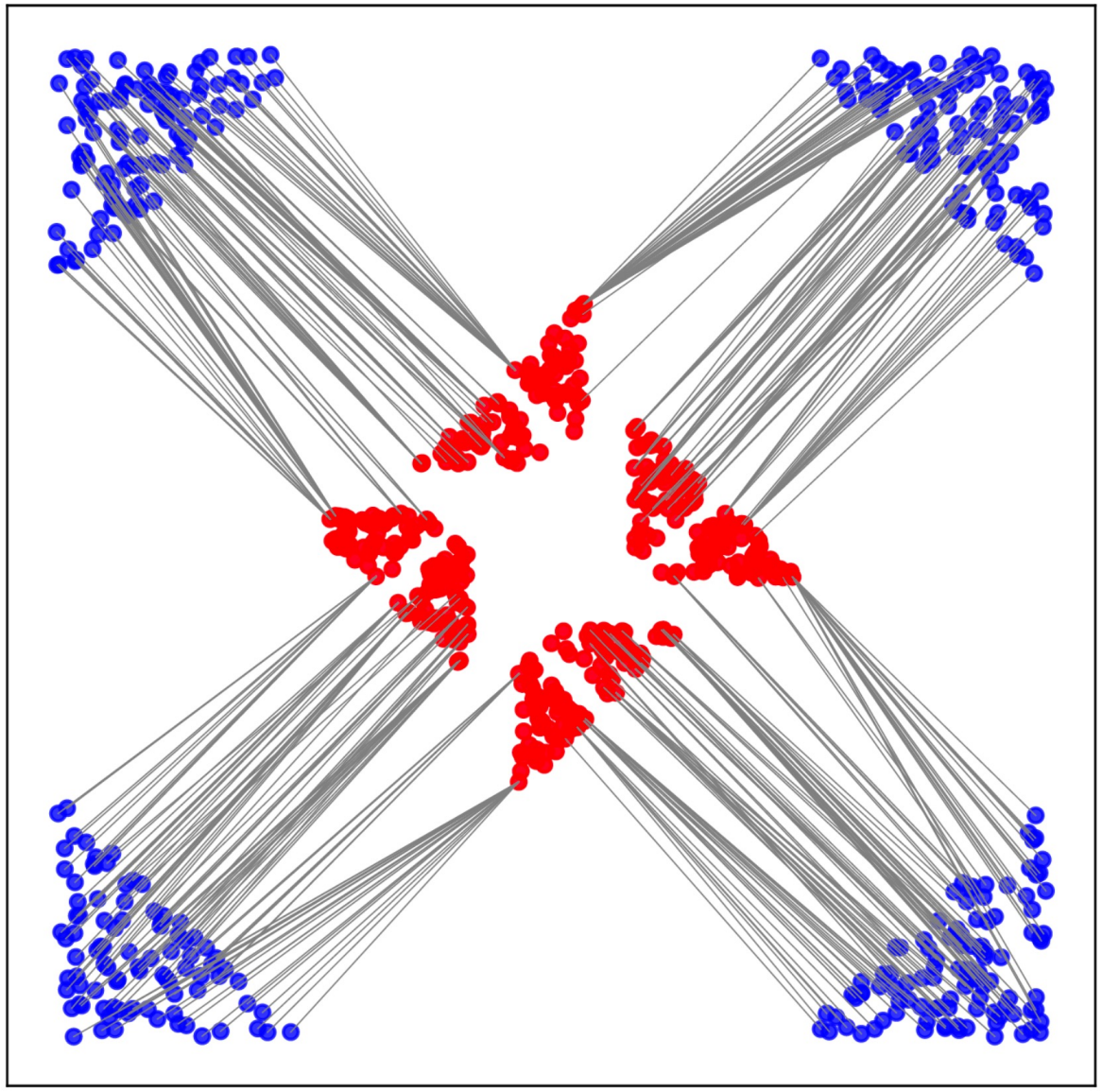
Ethan R. Elenberg[1], Rajiv Khanna[1], Alexandros G. Dimakis[1],
and Sahand Negahban[2]

- Greedy algos can enforce the block-diagonal constraint with a constant-factor approximation guarantee for its optimality.

Extending Homomorphism AE: Applications to Latent OT (Idea 2)

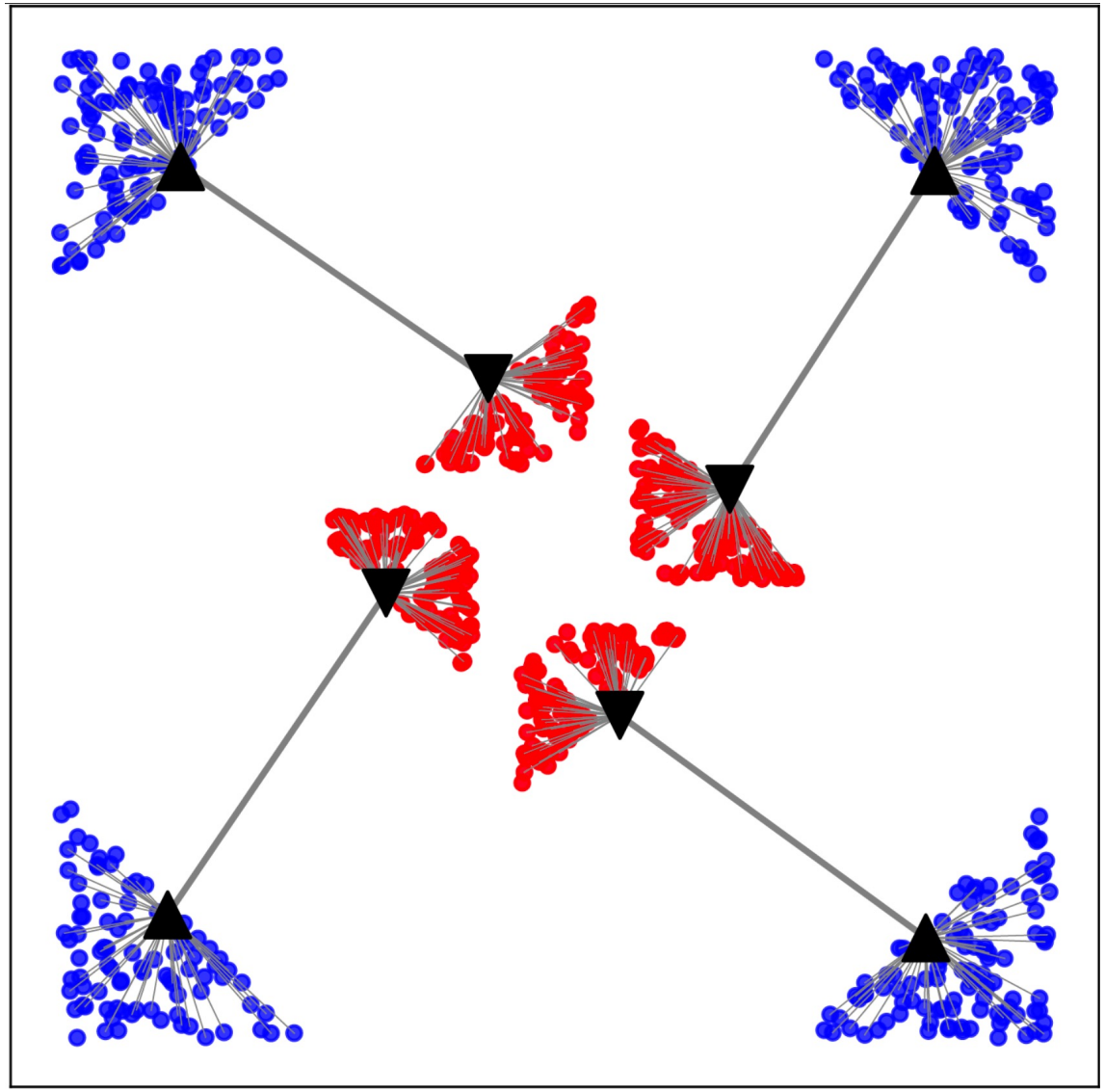
Brief Overview: Latent OT

OT alignments b/w source & target :



Brief Overview: Latent OT

LOT (ICML'21) alignments b/w source & target :



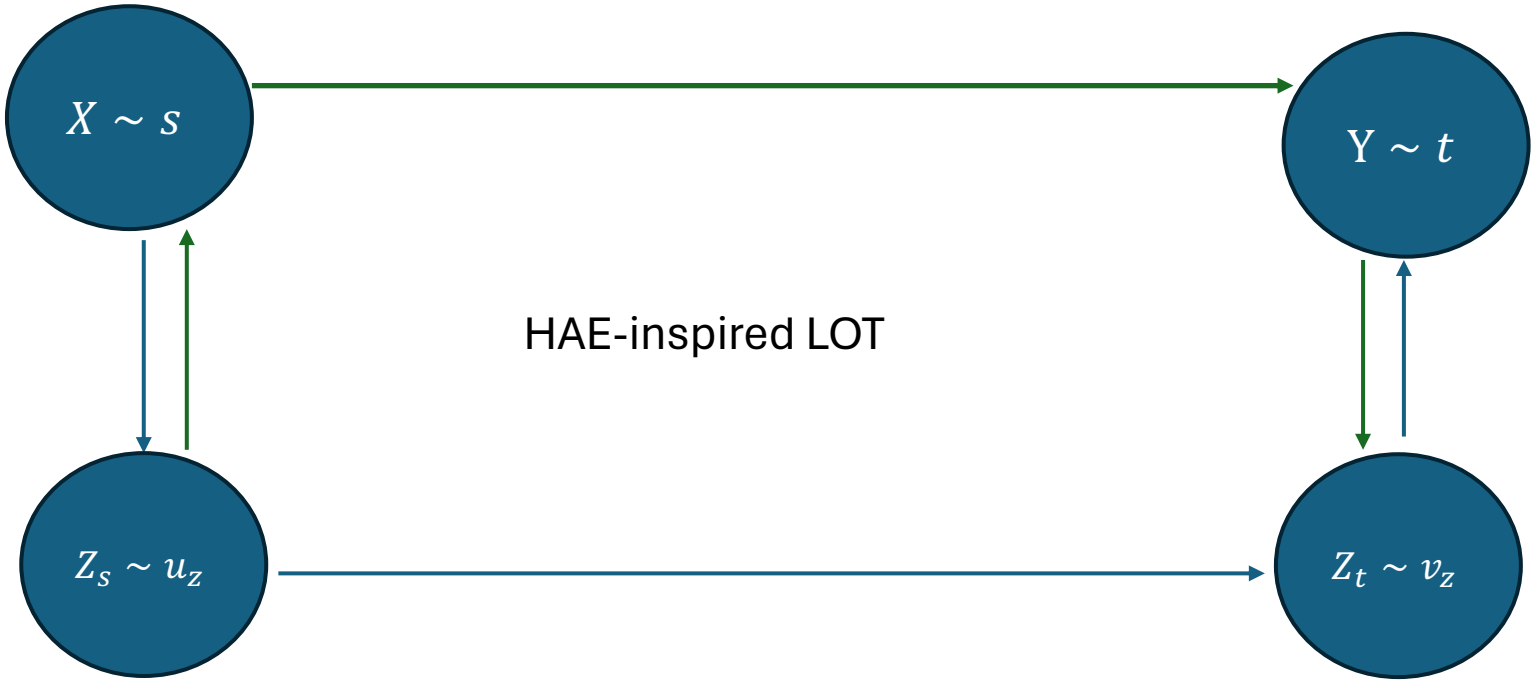
LOT & HAE (High-Level Connection)

Usual OT

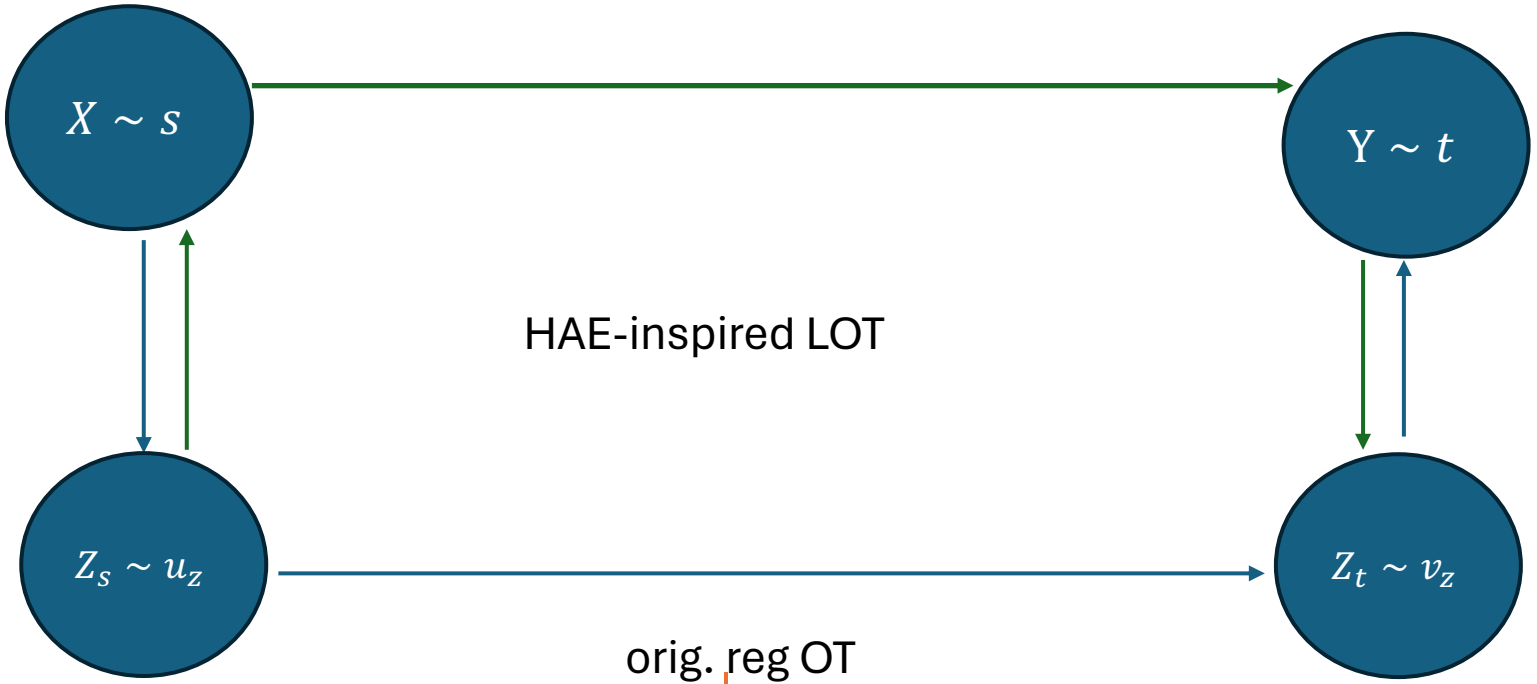
LOT & HAE (High-Level Connection)

LOT

LOT & HAE (High-Level Connection)



LOT & HAE (High-Level Connection)



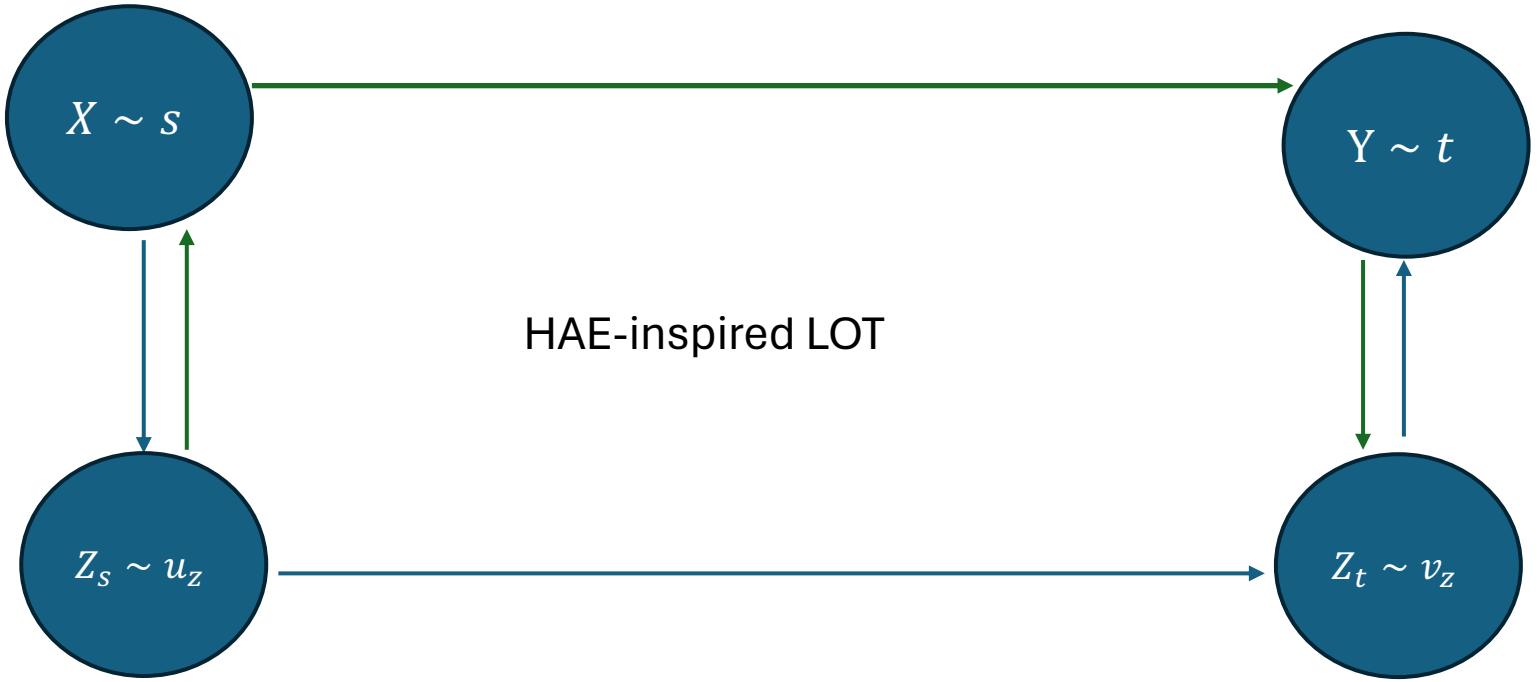
Why should disentanglement help LOT?

orig. reg OT

$$\varepsilon KL(\mathbf{P}\|\mathbf{K}) \leq \varepsilon(KL(\mathbf{P}_x\|\mathbf{K}_x) + KL(\mathbf{P}_z\|\mathbf{K}_z) + KL(\mathbf{P}_y\|\mathbf{K}_y)) + \varepsilon(\mathbf{H}(\mathbf{u}_z) + \mathbf{H}(\mathbf{v}_z)),$$

where $\mathbf{H}(\mathbf{a}) := -\sum_i \mathbf{a}_i \log \mathbf{a}_i$ *denotes the entropy.*

LOT & HAE (High-Level Connection)



Why should disentanglement help LOT?

$$\varepsilon KL(\mathbf{P}\|\mathbf{K}) \leq \varepsilon(KL(\mathbf{P}_x\|\mathbf{K}_x) + KL(\mathbf{P}_z\|\mathbf{K}_z) + KL(\mathbf{P}_y\|\mathbf{K}_y)) + \varepsilon(\mathbf{H}(\mathbf{u}_z) + \mathbf{H}(\mathbf{v}_z)), \quad \leq \mathrm{I}(Z_s; Z_t)$$

where $\mathbf{H}(\mathbf{a}) := -\sum_i \mathbf{a}_i \log \mathbf{a}_i$ *denotes the entropy.*