Disentanglement via Latent Quantization

Kyle Hsu[†], Will Dorrell[‡], James C. R. Whittington[†§], Jiajun Wu[†], Chelsea Finn[†]

[†] Stanford University [‡] UCL [§] University of Oxford

NEURAL INFORMATION PROCESSING SYSTEMS

Nonlinear ICA Problem

$$\hat{p}(\mathbf{x}) = \frac{1}{|\{\mathbf{x}\}|} \qquad \mathbf{x} = g(\mathbf{s}) \qquad p(\mathbf{s}) = \prod_{i=1}^{n_s} p(\mathbf{s}_i)$$

Given unlabelled data generated noiselessly from independent sources,

$$\mathbf{z} = \hat{g}^{-1}(\mathbf{x}) = \hat{g}^{-1} \circ g(\mathbf{s})$$ permutations, dimensionwise invertible functions

learn latents that recover the sources up to allowable indeterminacies.

InfoMEC: Information-Theoretic Metrics for Disentanglement

$$\mathrm{NMI}[i,j] = \frac{I(\mathbf{s}_i;\mathbf{z}_j)}{H(\mathbf{s}_i)} \qquad \mathrm{NMI} \in [0,1]^{n_s \times n_z}$$

InfoModularity: average max/sum in columns of NMI

InfoCompactness: average max/sum in rows of NMI

$$\mathrm{NMI}_{\mathcal{V}}[i] = \frac{I_{\mathcal{V}}(\mathbf{z} \to \mathbf{s}_i)}{H_{\mathcal{V}}(\mathbf{s}_i \mid \varnothing)} \qquad \mathrm{NMI}_{\mathcal{V}} \in [0,1]^{n_s}$$

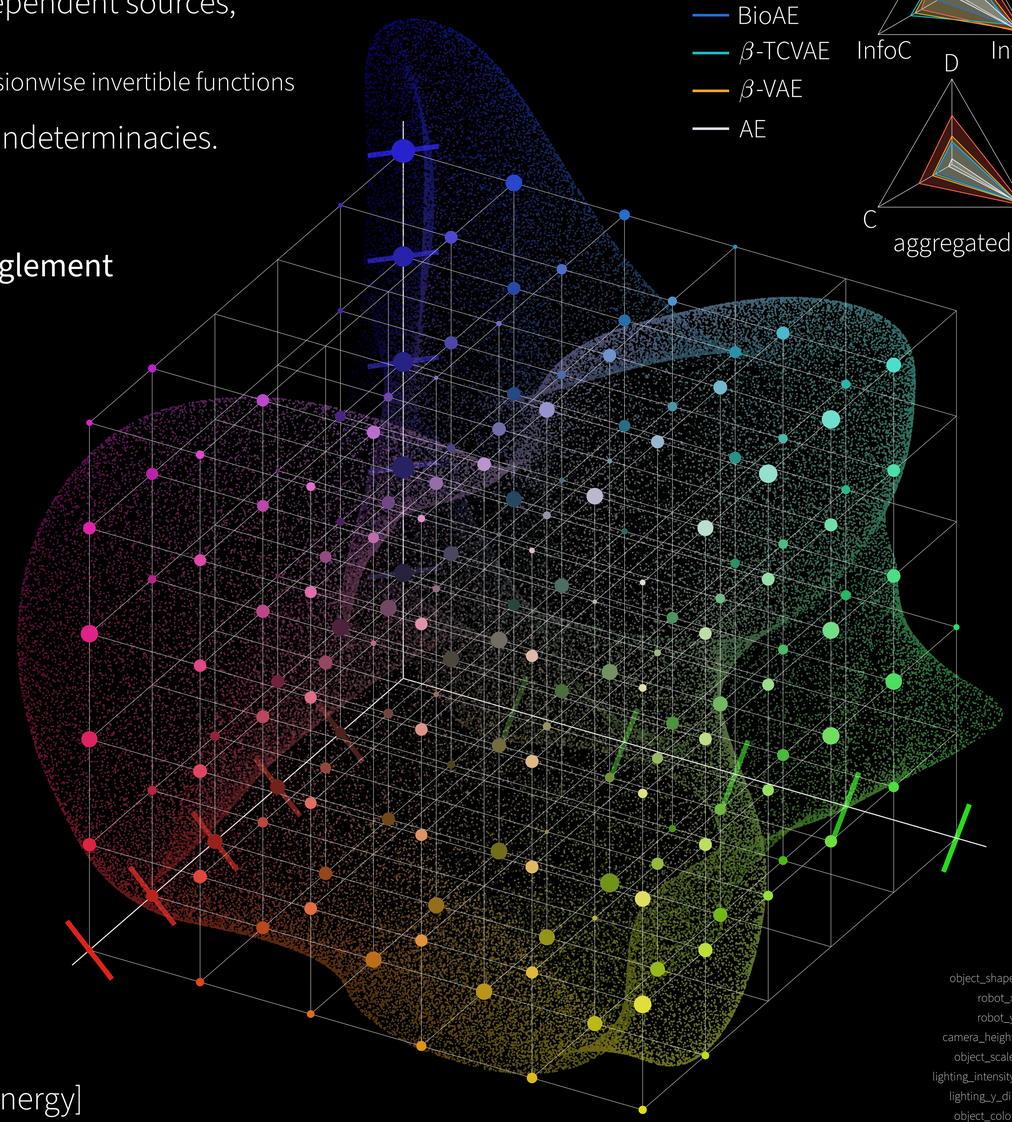
InfoExplicitness: average $\mathrm{NMI}_{\mathcal{V}}$, $\mathcal{V} =$ linear models

Prior Works Disentangle through Regularization

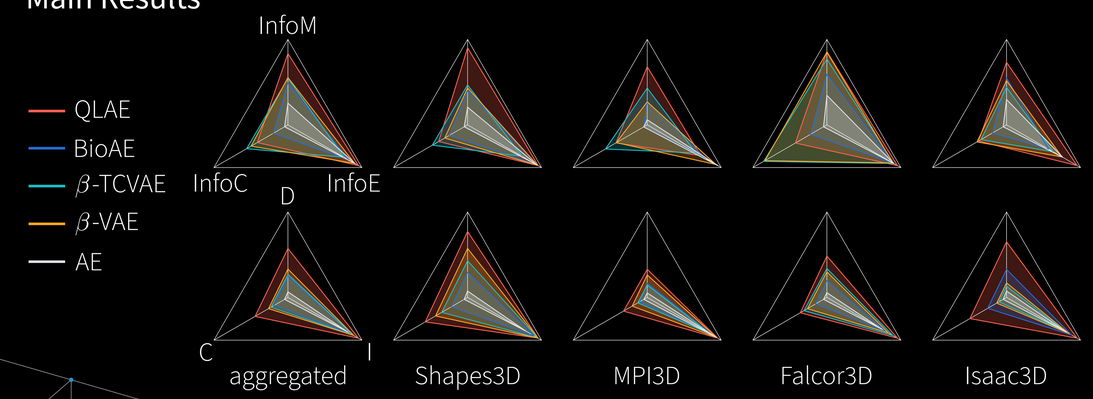
β-VAE: minimize $\mathbb{E}_{\hat{p}(\mathbf{x})}\left[\beta D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \,\|\, p(\mathbf{z}))\right]$

β-TCVAE: minimize $\mathbb{E}_{\hat{p}(\mathbf{x})}\left[\beta D_{\mathrm{KL}}(q_\phi(\mathbf{z}) \,\|\, \prod_{j=1}^{n_z} q_\phi(\mathbf{z}_j))\right]$ (latent multiinformation)

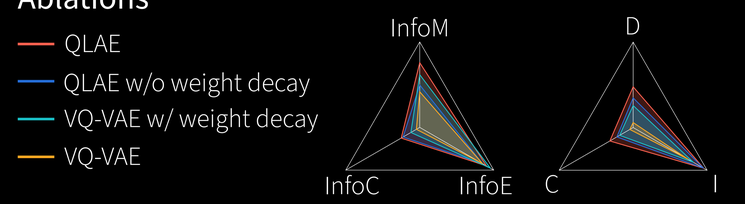
BioAE: minimize $\mathbb{E}_{\hat{p}(\mathbf{x})}$ [latent negativitity & latent energy]



Main Results



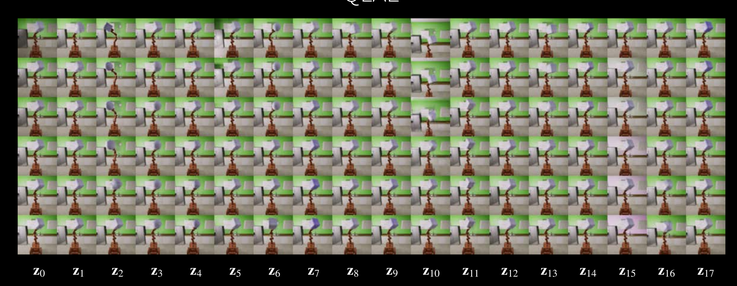
Ablations

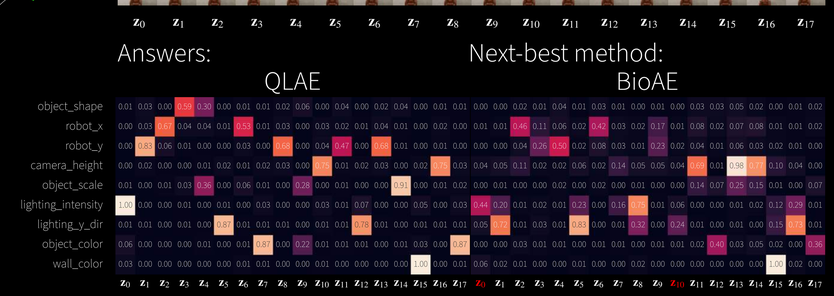


Visualizations (Isaac3D)

Can you interpret what source(s) each latent dimension encodes?

QLAE





$$z_j = \arg\min_{v_{jk} \in V_j} |\hat{g}^{-1}(x)_j - v_{jk}|, \quad j = 1, \ldots, n_z$$

Quantizing latent representations severely limits a model's channel capacity **by construction** (cf. regularization), forcing autoencoders and InfoGANs to represent high-dimensional data in a more disentangled manner.

paper