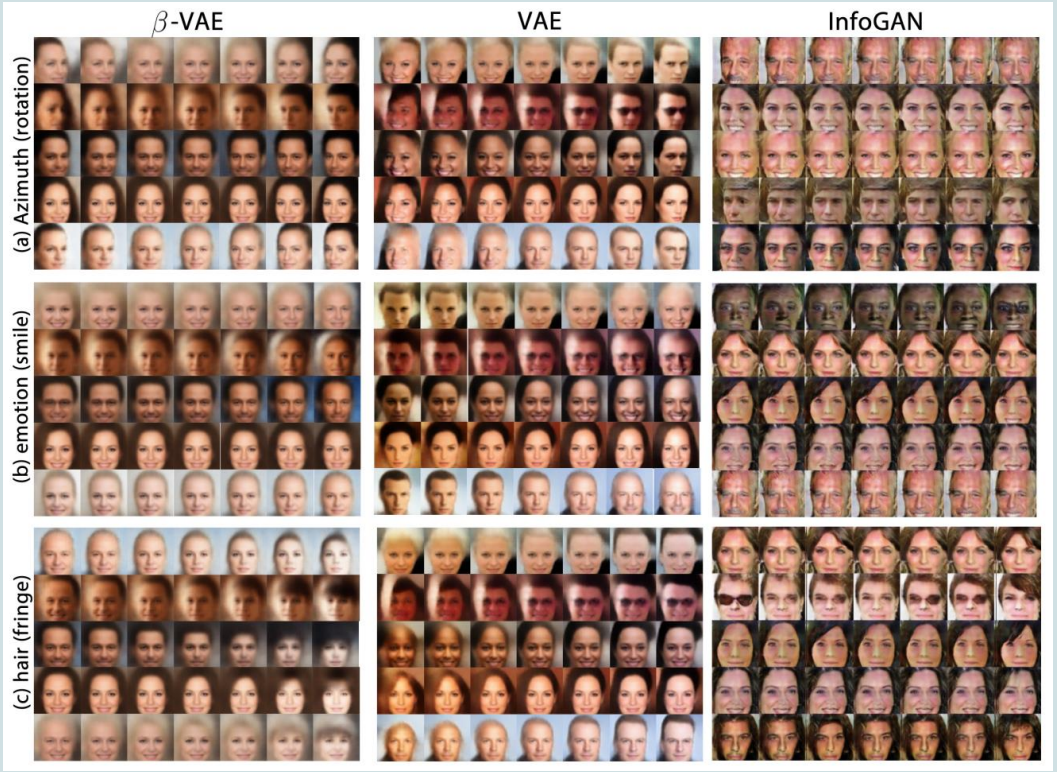
beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework

Paper Presentation #1 by Arushi Sahai

COMP690-204 Fall 2026

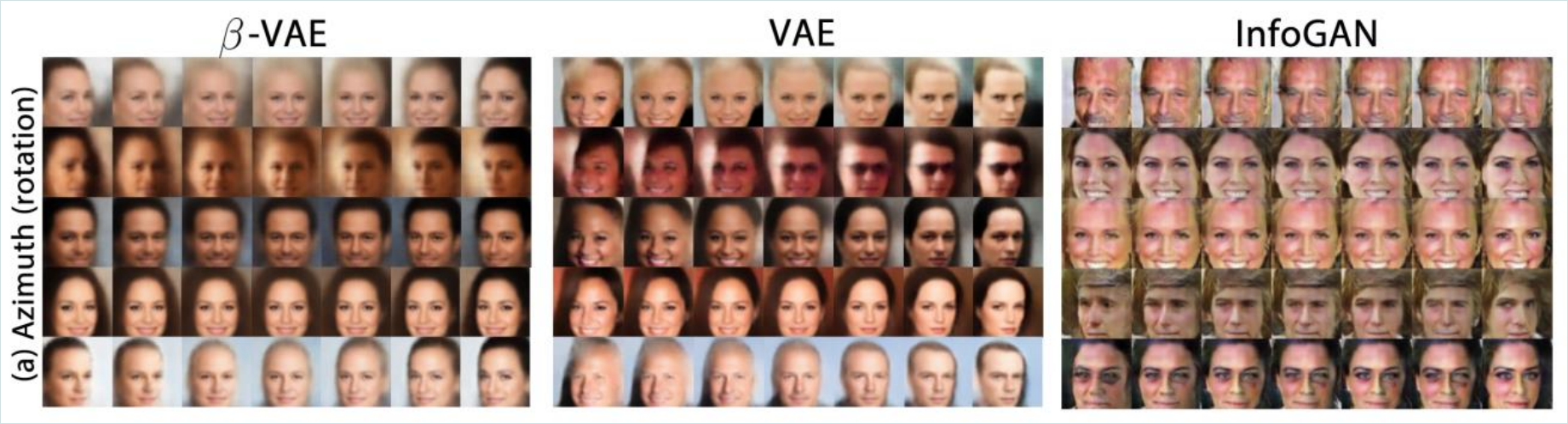
beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework

- **Authors:** Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, Alexander Lerchner — all from Google DeepMind
- **Published** as a conference paper at ICLR 2017
- **Cite: Figure 1.** Manipulating latent variables on celebA



The Problem: “Entangled” Representations

- Learned latent codes are usually entangled: one dimension mixes multiple factors together
- Goal: one latent axis ↔ one independent generative factor
- Cite: Figure 1. Manipulating latent variables on celebA

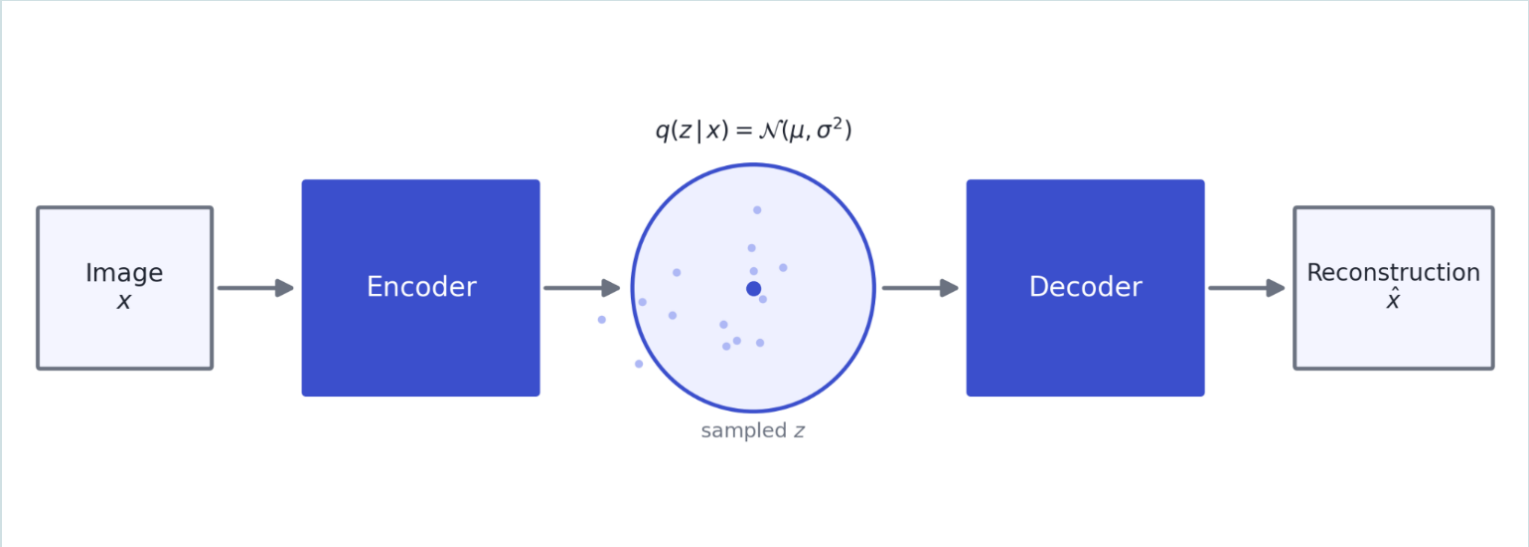


Why is this interesting, important, and hard?

- Interesting/important: interpretability, transfer learning, controllable generation
- Hard: no labels for the true factors, no ground truth, and (until this paper) no way to even measure disentanglement

Quick Recap: What is a VAE?

- Encoder → distribution over codes → decoder
- Latent space is regularized to be smooth and sample-able (that's what makes it generative, not just compressive)

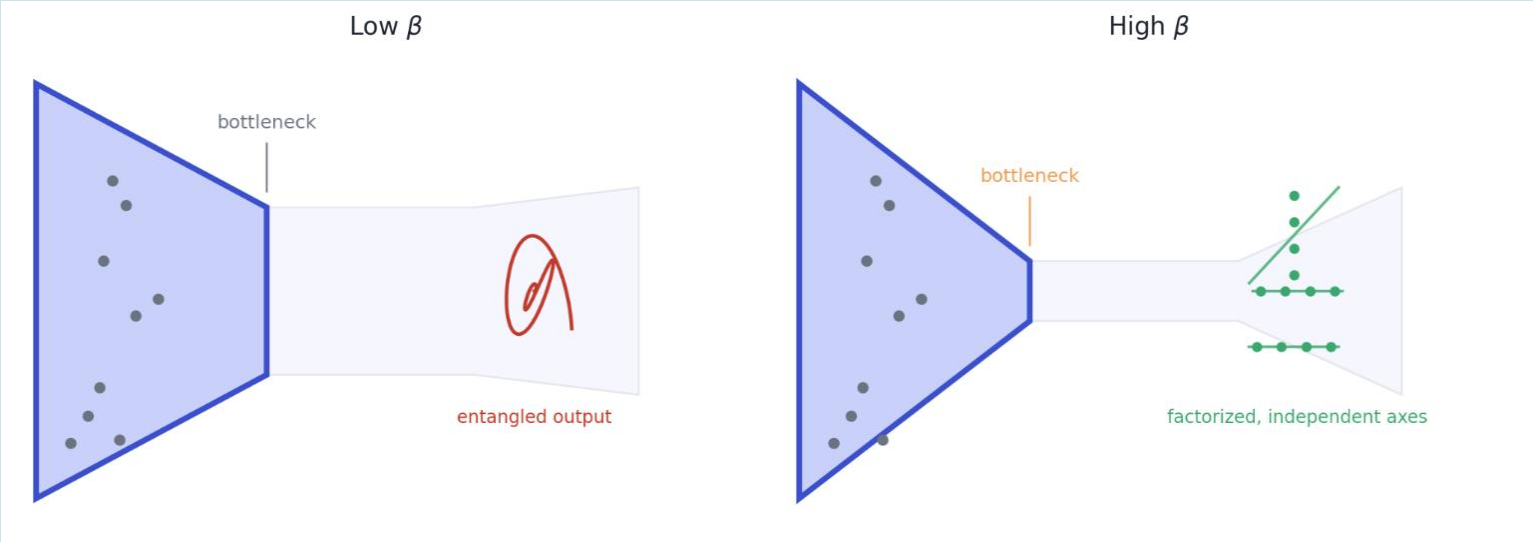


Contribution #1: the β-VAE objective

- Standard VAE: reconstruction − KL(q(z|x) ‖ p(z))
- β-VAE: reconstruction − β · KL(q(z|x) ‖ p(z)), with β > 1

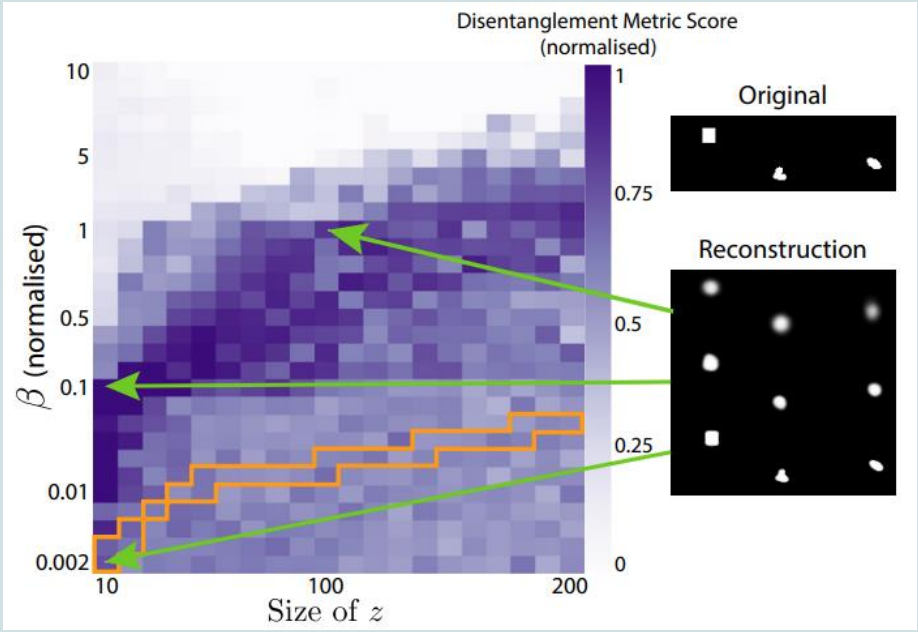
Contribution #2: why this causes disentanglement

- Higher β = tighter information bottleneck
- Forces a more efficient code → efficient coding favors independent, factorized axes
- Framed formally as constrained optimization (β acts like a Lagrange multiplier)



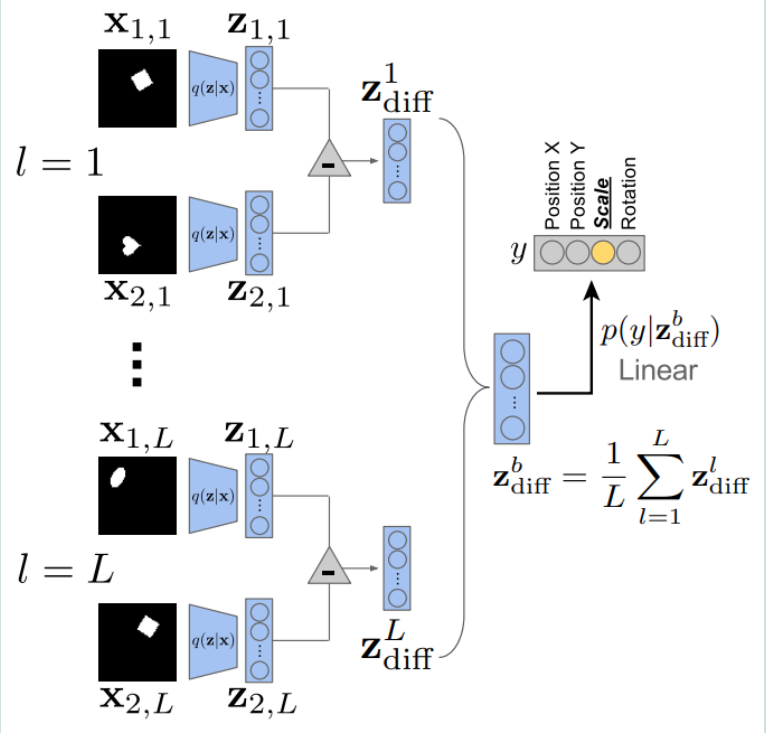
The Tradeoff

- Reconstruction quality drops as β increases (blurrier outputs)
- There's an optimal β for disentanglement, and it depends on latent size — not simply "higher is better"
- β is a real tuning knob — no free lunch



Contribution #3: a new disentanglement metric

- No prior metric existed — log-likelihood doesn't measure disentanglement
- Procedure: fix one ground-truth factor, vary the rest → look at resulting latent codes → train a simple linear classifier to guess which factor was held fixed
- High classifier accuracy = well-disentangled representation
- Cite: Figure 5. Schematic of the proposed disentanglement metric

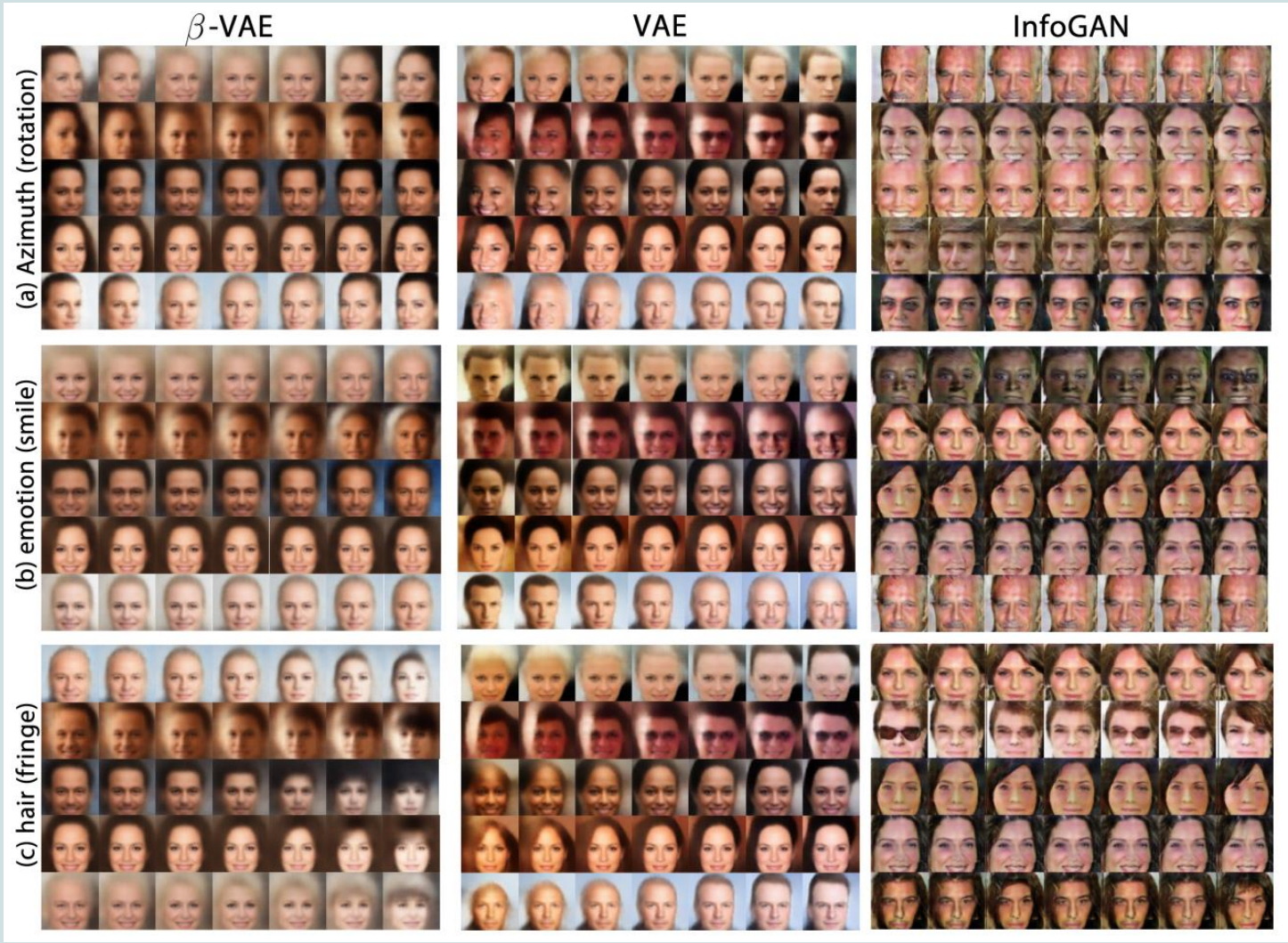


Datasets and Baselines

- Datasets: dSprites (synthetic 2D shapes, ground-truth factors), 3D chairs, 3D faces, CelebA
- Baselines: standard VAE, InfoGAN, PCA, ICA, and semi-supervised DC-IGN

Qualitative Results: Latent Traversals

Cite: Figure 1. Manipulating latent variables on celebA



Quantitative Results

- Disentanglement metric score (dSprites): β-VAE ≈ 99.2%, standard VAE ≈ 61.6%, InfoGAN ≈ 73.5%
- Matches semi-supervised DC-IGN (≈ 99.3%) — despite being fully unsupervised
- Cite: Figure 6. Disentanglement metric classification accuracy for 2D shapes dataset

Model	Disentanglement metric score
Ground truth	*100%*
Raw pixels	$45.75 \pm 0.8\%$
PCA	$84.9 \pm 0.4\%$
ICA	$42.03 \pm 10.6\%$
DC-IGN	$\mathbf{99.3 \pm 0.1\%}$
InfoGAN	$73.5 \pm 0.9\%$
VAE untrained	$44.14 \pm 2.5\%$
VAE	$61.58 \pm 0.5\%$
β-VAE	$\mathbf{99.23 \pm 0.1\%}$

Strengths

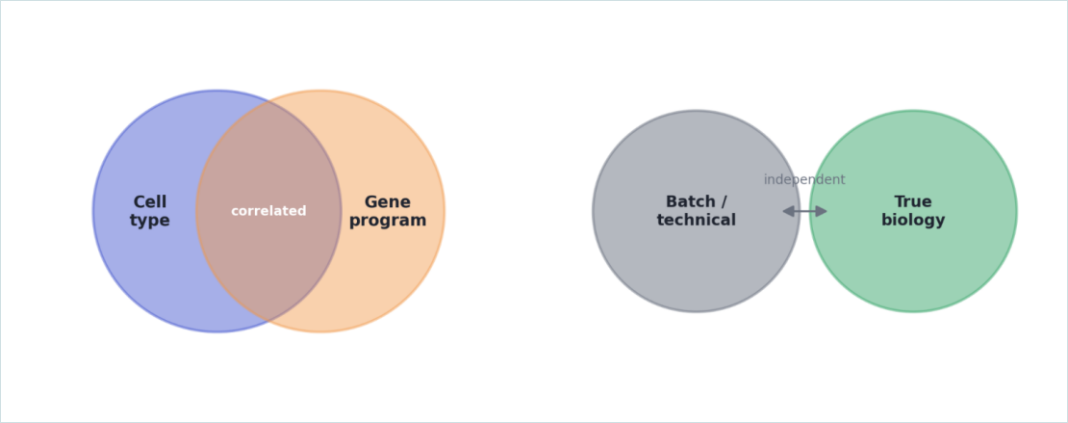
- Extremely simple: one hyperparameter, same architecture
- Fully unsupervised, stable to train
- Comes with a principled way to select β (via the metric)
- State-of-the-art disentanglement at the time

Weaknesses/Limitations

- Reconstruction blurriness at higher β
- β must be tuned per dataset — no universal value
- Metric and method both assume the data's true factors are already fairly independent — doesn't always hold
- Metric requires known ground-truth factors, so it can't be computed on most real-world datasets (only shown qualitatively there)

Open Questions, Extensions, and Applications

- Open question: β-VAE assumes the true generative factors are statistically independent — but are they, ever, in real biological data?
- Example: cell type identity and gene program activity are usually correlated, not independent
- Application: disentangling batch effects from true biological signal in single-cell data



Takeaways

- Problem: learned representations are entangled and hard to interpret
- Fix: one hyperparameter, β, tightening the VAE's information bottleneck
- Cost: reconstruction fidelity, tuning effort
- Legacy: introduced the first real quantitative disentanglement metric and kicked off a whole line of follow-up work

Answering YOUR Questions

1. If I want to find out if β-VAE improve transfer learning. Could I design a experiment like this? β-VAE and standard VAE could first be trained unsupervised on images of several object classes with varying position, scale, and rotation. Their encoders could then be frozen and transferred to previously unseen object classes. Using only a small amount of labeled target data, a linear predictor could be trained to predict position, scale, or rotation. If β-VAE requires fewer labeled examples or achieves higher accuracy than standard VAE, this would provide direct evidence that disentangled representations improve transfer learning.
2. As a paper published almost 10 years ago, is this paper still effective today or is its methods still used?
3. How do you choose the right β when you don't have any labeled data?
4. Would β-VAE still work well if the real-world factors are not independent?
5. Why does making the representation more disentangled often result in blurrier reconstructions?

Thank you!