Disentanglement of Latent Spaces: VAEs vs StyleGAN (Face Generation Task)

Hajar HAMDOUCH

1. Introduction

In latent variable models, a low-dimensional vector $z \in \mathbb{R}^d$ is mapped to complex data such as images. The objective is often *disentanglement*: each latent direction ideally controls a single semantic factor (e.g. identity, pose, hairstyle).

My observation with a convolutional VAE: By building a convolutional VAE , after training when I sampled from the latent space, I consistently obtained the same blurry face, with only the hair texture or noise pattern changing.

2. Mathematical Framework

2.1 Generator and perceptual map

To formalize this:

- Generator: $G : Z \to X$, maps latent z to an image x.
- Perceptual feature extractor: $\phi : X \to \mathbb{R}^m$ (e.g. VGG16).
- Composition: $F = \phi \circ G : Z \to \mathbb{R}^m$.

2.2 Local linearization

For a small perturbation $\delta \in \mathbb{R}^d$:

$$F(z + \delta) \;\approx\; F(z) + J_F(z)\,\delta,$$

where $J_F(z) \in \mathbb{R}^{m \times d}$ is the Jacobian.

2.3 Quadratic form of perceptual change

$$\|F(z + \delta) - F(z)\|^2 \;\approx\; \delta^\top M(z)\,\delta, \quad M(z) := J_F(z)^\top J_F(z).$$

- Eigenvectors of $M(z)$ give the main latent directions.
- Eigenvalues λ_i are squared sensitivities of those directions.

Large λ_i mean hypersensitive directions (e.g. hair/noise), while small λ_i mean suppressed directions (e.g. identity).

3. Entanglement and Disentanglement

- **Disentangled**: eigenvectors align with semantic factors, eigenvalues are balanced.
- **Entangled**: eigenvectors are mixtures, eigenvalues are highly uneven.

This explains what I saw in my VAE: identity factors correspond to small eigenvalues (nearly unchanged), while hair/noise factors dominate.

4. Perceptual Path Length (PPL)

4.1 Definition in Z

$$l_Z = \mathbb{E}\left[\frac{1}{\epsilon^2}\, d\Big(G(\mathrm{slerp}(z_1, z_2; t)),\ G(\mathrm{slerp}(z_1, z_2; t+\epsilon))\Big)\right],$$

where

- $z_1, z_2 \sim \mathcal{N}(0, I)$,
- slerp = spherical linear interpolation,
- Gaussian samples concentrate on a sphere of radius $\sqrt{d}$.

4.2 Linearization along the path

Let $z(t) = \mathrm{slerp}(z_1, z_2; t)$. Define the tangent direction

$$u = \frac{z(t+\epsilon) - z(t)}{\epsilon}, \quad \|u\| \approx 1.$$

Then

$$\frac{1}{\epsilon^2}\|F(z(t+\epsilon)) - F(z(t))\|^2 \ \approx\ u^\top M(z(t))u.$$

5. Averaging Over Random Directions

5.1 Key identity

For u uniform on the unit sphere S^{d-1}:

$$\mathbb{E}[uu^\top] = \tfrac{1}{d} I_d.$$

Reasoning:

- By rotational symmetry, the expectation must be a multiple of the identity.
- Trace condition: $\mathrm{trace}(uu^\top) = \|u\|^2 = 1$.
- Hence $c \cdot d = 1 \implies c = 1/d$.

5.2 Resulting expectation

$$\mathbb{E}_u[u^\top M u] = \tfrac{1}{d}\,\mathrm{trace}(M) = \tfrac{1}{d}\sum_{i=1}^{d} \sigma_i^2(J_F).$$

So, PPL measures the average squared singular values of the Jacobian.

6. Geometric Picture

At each z, $M(z)$ defines a *tangent ellipsoid*:

$$\{\delta : \delta^\top M(z)\delta = 1\}.$$

- Long axes $\to$ insensitive directions (small eigenvalues).
- Short axes $\to$ hypersensitive directions (large eigenvalues).
- Disentanglement means the ellipsoid is close to a sphere, aligned with semantic axes.

7. StyleGAN's Solution: Mapping Network

7.1 Architecture

$$G(z) = g(f(z)), \quad f : Z \to W \text{ (8-layer MLP)}, \quad g : W \to X.$$

7.3 Interpolation change

- In Z: interpolation uses **slerp**.
- In W: interpolation uses **lerp**:

$$l_W = \mathbb{E}\left[\tfrac{1}{\epsilon^2} d\Big(g(\mathrm{lerp}(f(z_1), f(z_2); t)),\ g(\mathrm{lerp}(f(z_1), f(z_2); t+\epsilon))\Big)\right].$$

8. Conclusions

1. The Jacobian tells how sensitive each latent direction is in perceptual space.
2. PPL quantifies the curvature of the latent manifold: higher values mean more entanglement.
3. StyleGAN's mapping network reduces this anisotropy and aligns semantic axes in W.