Disentanglement of Latent Spaces: From VAEs to StyleGAN

Lecture Notes

1. Introduction

Latent variable models map a low-dimensional representation $z \in \mathbb{R}^d$ to complex data such as images. A central goal is *disentanglement*: each latent direction controls a single semantic factor (e.g. identity, pose, hairstyle). In practice, standard VAEs often yield *entangled* latents, where perturbing one coordinate changes multiple factors in a non-linear way.

Observation from a convolutional VAE: Sampling from the latent space consistently generated the same blurry identity, with only hair texture changing. This illustrates entanglement: "identity" is suppressed while "hair/noise" dominates.

2. Mathematical Framework

2.1 Generator and perceptual map

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

2.2 Local linearization

For small perturbations $\delta \in \mathbb{R}^d$,

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

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 $M(z)$: principal latent directions.
- Eigenvalues λ_i: squared sensitivities in those directions.

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

3. Entanglement and Disentanglement

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

In VAEs, identity factors correspond to small eigenvalues, hence remain nearly unchanged, while hair/noise dominates.

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 lie near 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

If u is uniform on the unit sphere S^{d-1},

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

Reasoning:

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

5.2 Resulting expectation

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

Thus, 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: insensitive directions (small eigenvalues).
- Short axes: hypersensitive directions (large eigenvalues).
- Disentanglement corresponds to ellipsoid $\approx$ sphere, aligned with semantic axes.

7. StyleGAN's Solution: Mapping Network

7.1 Architecture

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

7.2 Motivation

- In Z: constrained Gaussian prior $\to$ entangled directions.
- In W: distribution unconstrained $\to$ mapping f can unwarp latent space.

7.3 Interpolation change

- Eq. (2): in Z, interpolation uses **slerp**.
- Eq. (3): 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. Key Insights

1. Jacobian analysis reveals local perceptual sensitivity.
2. PPL quantifies latent curvature: higher values $\to$ more entangled.
3. VAE: anisotropic Jacobian spectrum $\to$ same face, noisy hair.
4. StyleGAN: mapping network reduces anisotropy, aligns semantic axes in W.
5. Intuition: VAE latent space = tangled ball of strings; StyleGAN untangles them.