中国科学技术大学
University of Science and Technology of China

Interpreting the Latent Space of GANs via Correlation Analysis for Controllable Concept Manipulation

Ziqiang Li[1*], Rentuo Tao[1*], Hongjing Niu[1], Mingdao Yue[2], Bin Li[1]

[1]University of Science and Technology of China
[2]Suzhou University
[*]Enqual contribution

Background

1. For the interpretability of GANs, due to the black-box property of deep neural models, hardly can we understand how the latent variables affect the generation process.

2. We use t-SNE to analyze the latent representations of Fashion-MNIST samples and find that the latent representations of samples from different classes can be well-separated.

3. It motivated us to quantify the importance of different latent dimensions for specific concept generation.

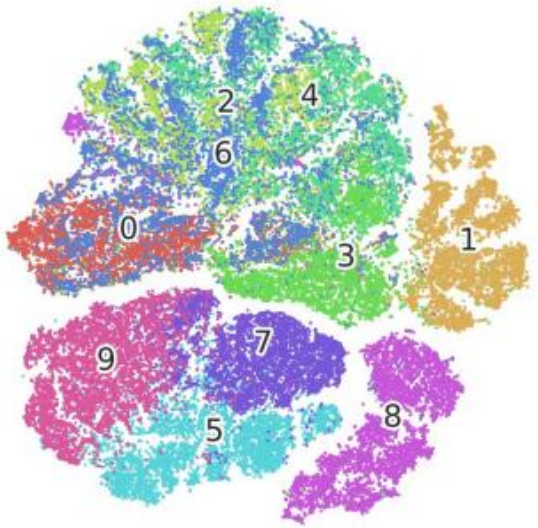

Fig. 1: t-SNE analysis on latent representations of Fashion-MNIST dataset. Points in different color are 2D features of latent representations belong to different classes.

Contributions

• We first propose to interpret the latent space of GANs by quantifying the correlation between the latent inputs and the generated outputs.

• We demonstrate that for generating contents of specific concept, the importance of different latent variables may varies greatly. Moreover, we propose an optimization-based method to find controlling latent variables for specific concept.

• The proposed method can fulfill controllable concept manipulation in generated images via controlling variables discovering and intervention.

Method: Analyzing the Latent Space

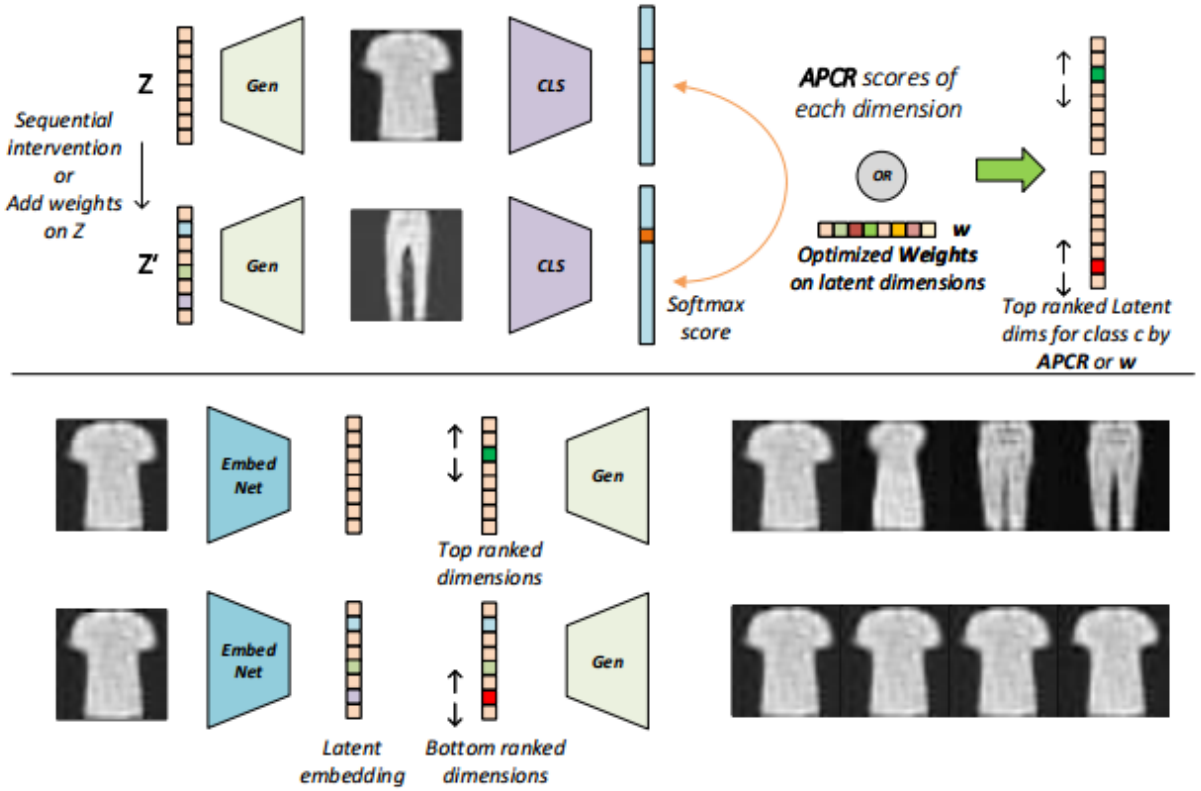

Fig. 2: The proposed method for analyzing the correlation between latent space and output image space of GANs. Top part illustrate the process of finding high-correlated latent dimensions by sequential intervention or adding weights on latent variables. Bottom part denote the process of latent intervention on top or bottom ranked latent dimensions.

Experiments: Sequential Latent Intervention

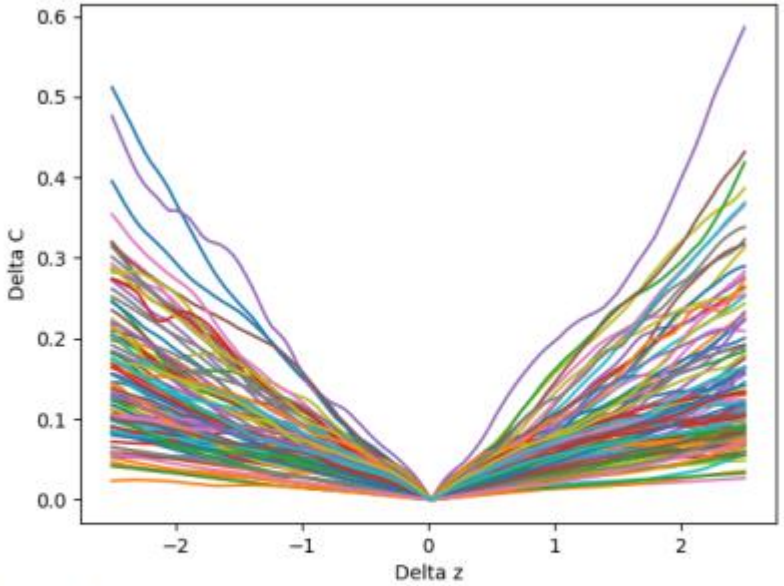

Fig. 3: Classification score change with respect to intervention on different latent dimensions. Each color represent a latent dimension.

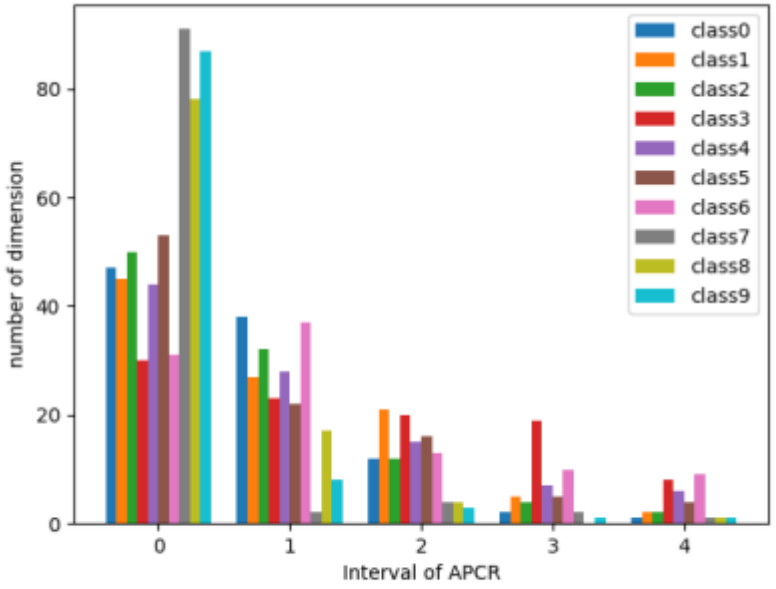

Fig. 4: Number distribution of latent dimensions with respect to different APCR value range.

Experiments: Optimization

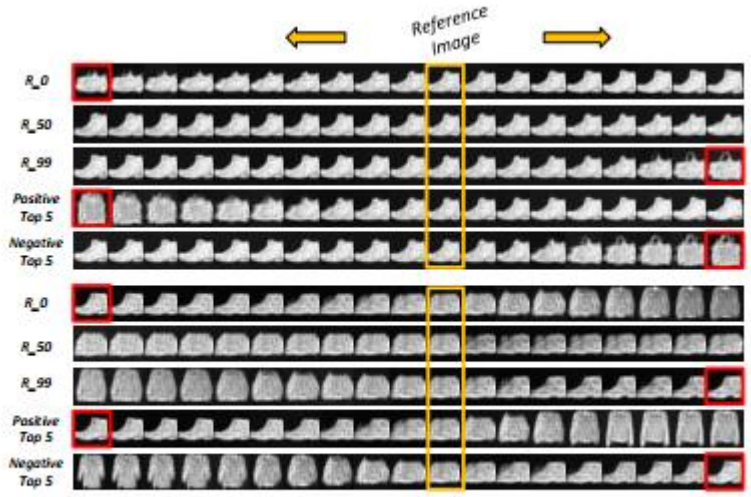

Fig. 5: Intervene on controlling set of latent dimensions.

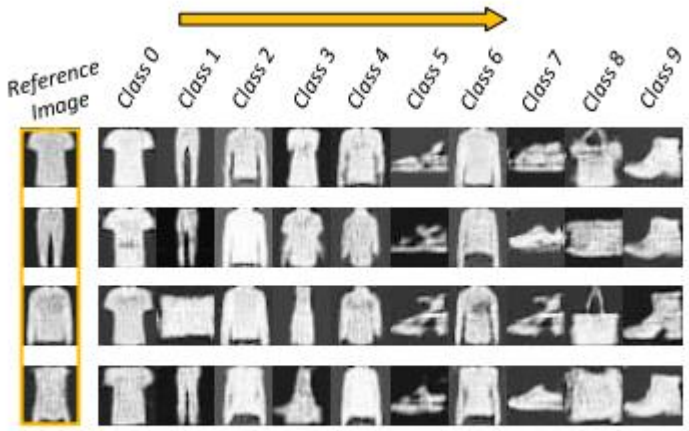

Fig. 6: Controllable concept manipulation on Fashion-MNIST through intervening on controlling set of latent dimensions (final manipulation results).

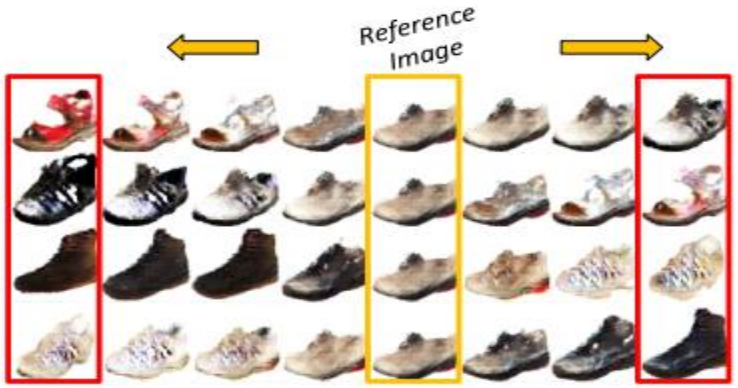

Fig. 7: Controllable concept manipulation on UT Zappos50k through intervening on controlling set of latent dimensions.

Classes	class0	class1	class2	class3	class4
IR_{ctrl}	0.7	0.9	0.7	0.8	0.4
Classes	class5	class6	class7	class8	class9
IR_{ctrl}	1	0.7	0.9	0.7	0.7

TABLE I: Intersection ration of high-correlated latent dimensions derived by sequential intervention and optimization

中国科学技术大学
University of Science and Technology of China

Thank you for listening !