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

[*]Equal contribution
`iceli, trtmelon,sasori}@mail.ustc.edu.cn, {ymdustc,binli}@ustc.edu.cn`

Abstract

Generative adversarial nets (GANs) have been successfully applied in many fields like image generation, inpainting, super-resolution, and drug discovery, etc. By now, the inner process of GANs is far from being understood. To get a deeper insight into the intrinsic mechanism of GANs, in this paper, a method for interpreting the latent space of GANs by analyzing the correlation between latent variables and the corresponding semantic contents in generated images is proposed. Unlike previous methods that focus on dissecting models via feature visualization, the emphasis of this work is put on the variables in latent space, i.e. how the latent variables affect the quantitative analysis of generated results. Given a pre-trained GAN model with weights fixed, the latent variables are intervened to analyze their effect on the semantic content in generated images. A set of controlling latent variables can be derived for specific content generation, and the controllable semantic content manipulation is achieved. The proposed method is testified on the datasets Fashion-MNIST and UT Zappos50K, experiment results show its effectiveness.

Conclusions

- 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 optimizationbased 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.

Background

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. To investigate if the variables in latent space contain the necessary information to distinguish different semantic contents in generated images, we use t-SNE[1] to analyze the latent representations of Fashion-MNIST samples and find that the latent representations of samples from different classes can be well-separated, as shown in Figure 1. This indicates that for samples from the same one class, there may be exist closely-related latent variables that made their latent representations distinguishable.

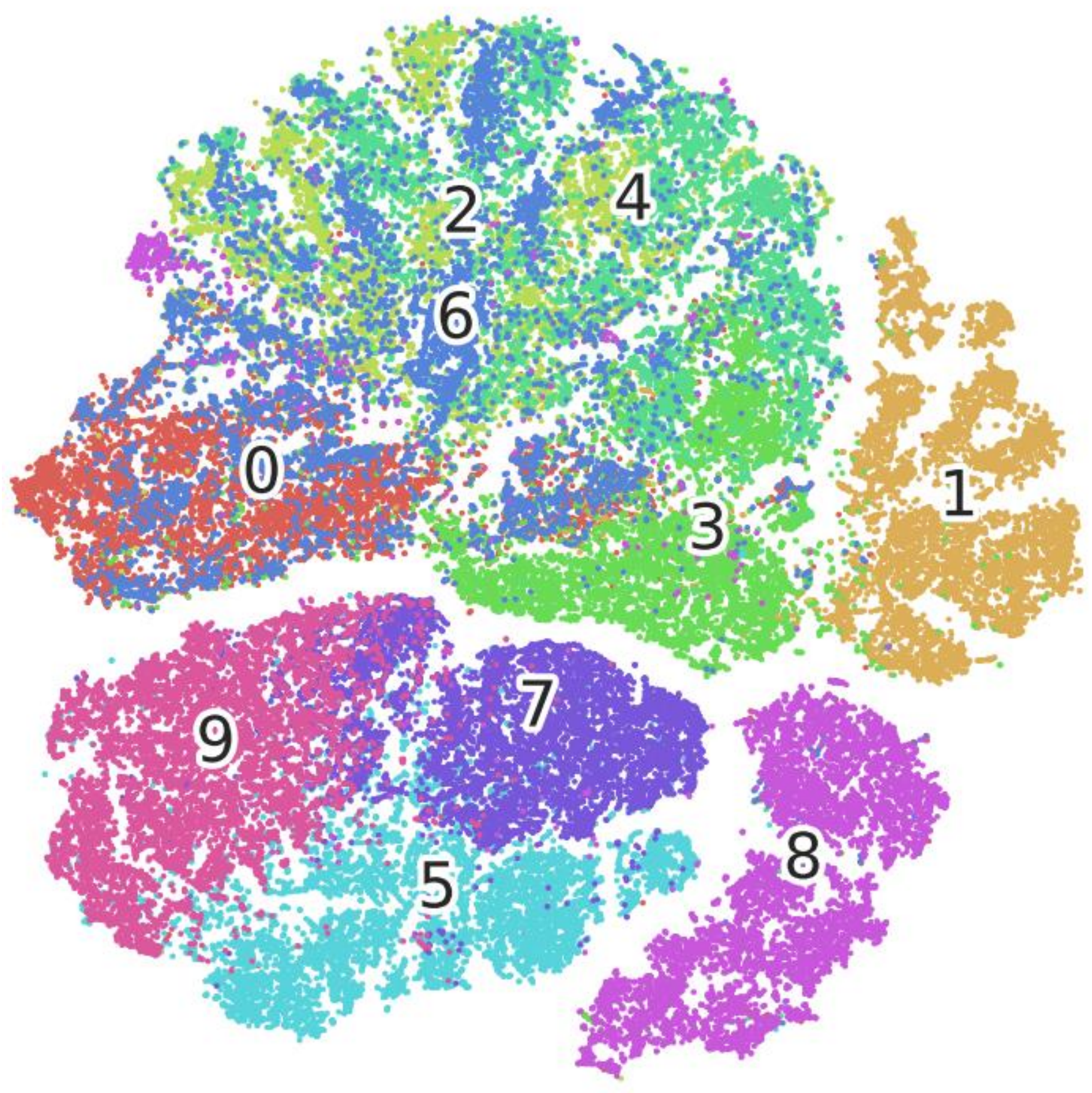

Figure 1: T-SNE analysis on latent representations of Fashion-MNIST dataset. Points in different color are 2D features of latent representations belong to different classes.

Correlation Analysis between Latent and Output Spaces

In this paper, we propose to interpreting the latent space of GANs by analyzing the correlation between latent dimensions and the corresponding content changes in generated images. The target was to semantically interpret the latent space and quantify the importance of different latent dimensions, below are the details of the proposed method. The proposed method was depicted in Figure 2.

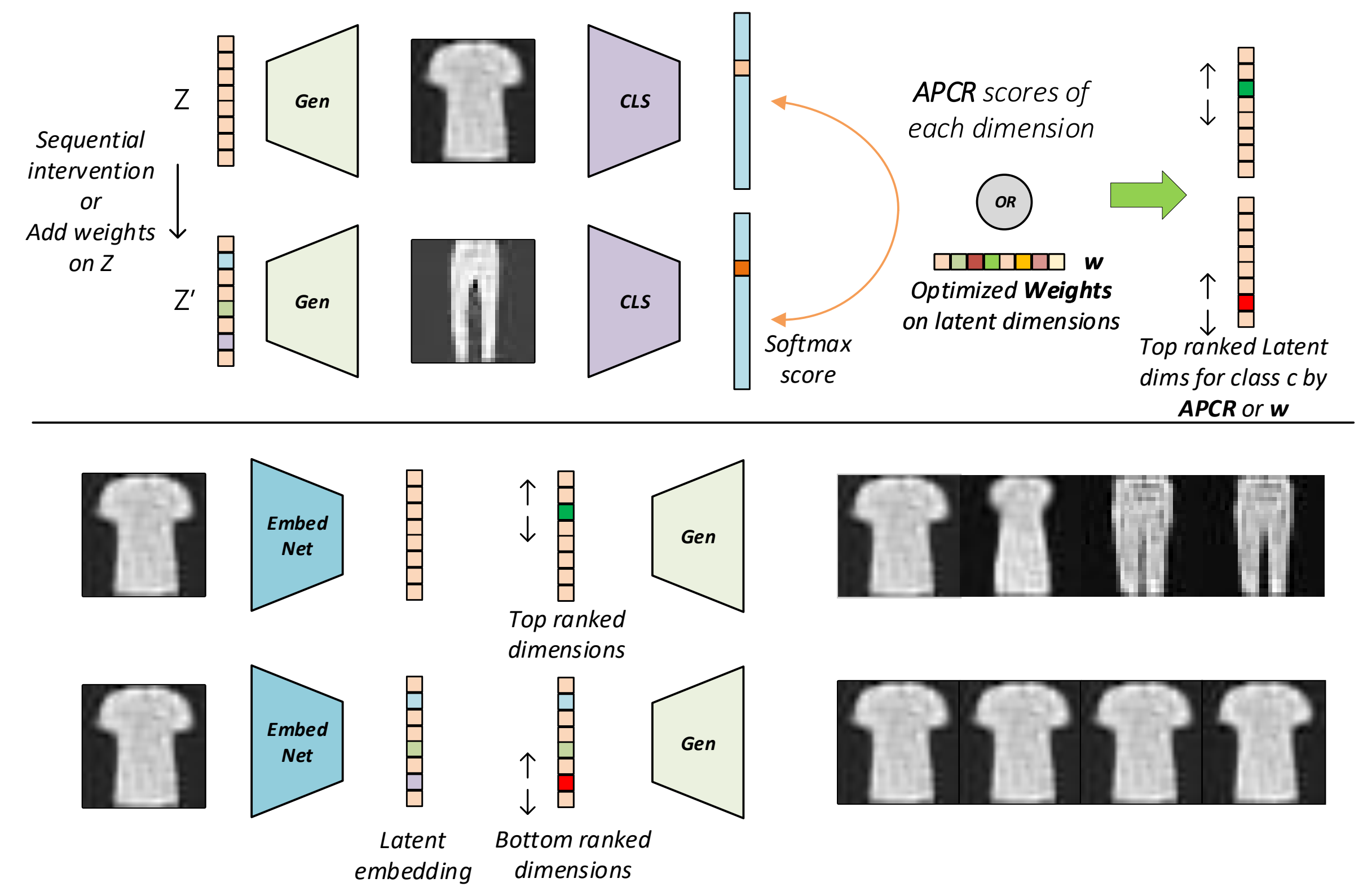
Z
Gen
CLS
Sequential intervention or Add weights on Z
Z′
Gen
CLS
Softmax score
APCR scores of each dimension
OR
w
Optimized Weights on latent dimensions
Top ranked Latent dims for class c by APCR or w
Embed Net
Gen
Top ranked dimensions
Embed Net
Gen
Latent embedding
Bottom ranked dimensions

Figure 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.

Calculate the APCR by sequential intervention

For pre-trained classifier $Q(x)$, the i-th latent dimension and j-th class have:

$$\mathbf{Z}^k = \mathbf{Z} + k \cdot \delta \cdot [0, \ldots, 1, \ldots, 0] \quad k \in [-m, m] \tag{1}$$

$$\mathbf{X}^k = \mathbf{G}\left(\mathbf{Z}^k\right) = \mathbf{G}\left([z_1, \ldots, z_i + \delta \cdot k, \ldots, z_N]\right) \tag{2}$$

$$\mathbf{S}_{i,j}^k = \mathbf{Q}_j\left(\mathbf{X}^k\right) = \mathbf{Q}_j\left(\mathbf{G}\left(\mathbf{Z}^k\right)\right) \tag{3}$$

To measure the i-th latent dimension's effect on the generation of j-th class contents, we use the averaged probability change ratio (APCR) as the quantitative evaluation metric.

$$APCR_{i,j} = \left\| \frac{\sum_{k=1}^{m}\left(\mathbf{S}_{i,j}^k - \mathbf{S}_{i,j}^{k-1}\right)}{2 \cdot m} \right\|_1 + \left\| \frac{\sum_{k=-1}^{-m}\left(\mathbf{S}_{i,j}^k - \mathbf{S}_{i,j}^{k-1}\right)}{2 \cdot m} \right\|_1 \tag{4}$$

Obtain the weight vector of the latent by optimization

To derive the controlling dimensions for j-th class, we firstly add differentiated distortions to the latent dimensions by utilizing a weight vector $\mathbf{w} = [w_1, \ldots, w_N]$:

$$\mathbf{Z}' = \mathbf{Z} + \mathbf{w} * \xi = [z_1 + w_1 \cdot \xi, \ldots, z_N + w_N \cdot \xi] \tag{5}$$

Then we can derive the optimization objective of the weight vector $\mathbf{w}$:

$$\Delta\mathbf{S}_j = \mathbf{S}_j' - \mathbf{S}_j = \mathbf{Q}_j\left(\mathbf{G}\left(\mathbf{Z}'\right)\right) - \mathbf{Q}_j\left(\mathbf{G}\left(\mathbf{Z}\right)\right) \tag{6}$$

$$\mathbf{w}^* = \arg\max \mathbf{L_w} = \arg\max\left(\left|\Delta\mathbf{S}_j\right|\right) \tag{7}$$

$$\mathbf{w}^* = \arg\min\left(1 - \left|\Delta\mathbf{S}_j\right| + \lambda \cdot \|\mathbf{w}\|_2\right) \tag{8}$$

$$\mathbf{w}_{>t_j} \Longleftarrow \left[\left|w_{j,1}^*\right| > t_j, \ldots, \left|w_{j,N}^*\right| > t_j\right] \tag{9}$$

Experiments

Find High-Correlated Latent Dimensions by Sequential Latent Intervention

Given a specific class concept, we intervene in the latent dimensions sequentially and get corresponding APCR scores for each latent dimension.

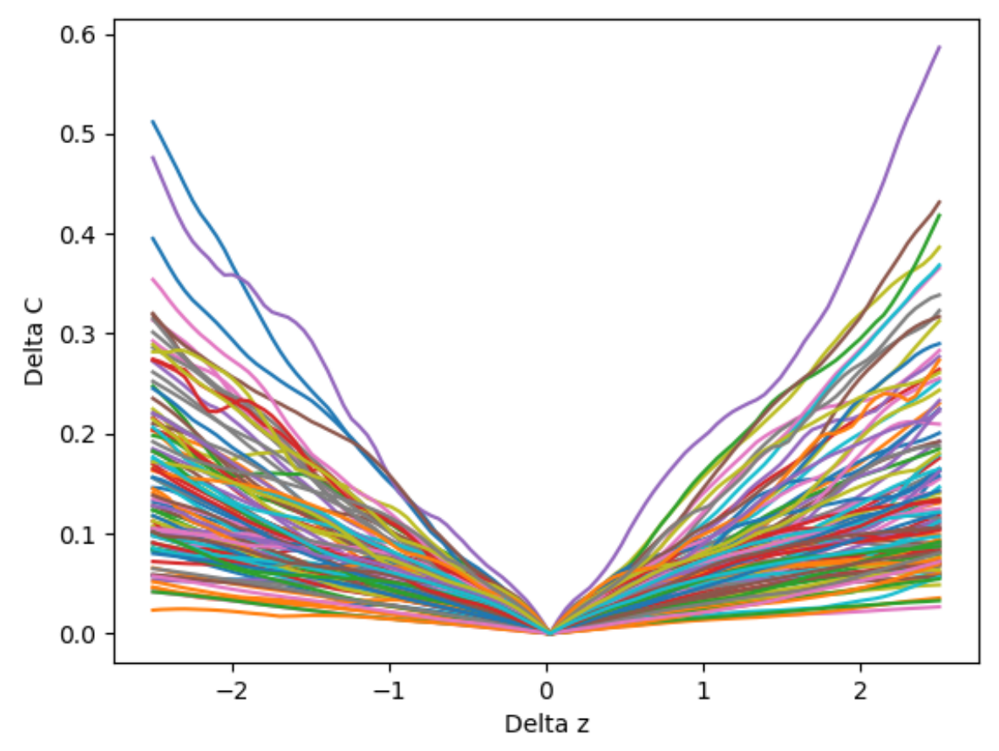

Figure 3: Classification score change with respect to intervention on different latent dimensions. Each color represent a latent dimension.

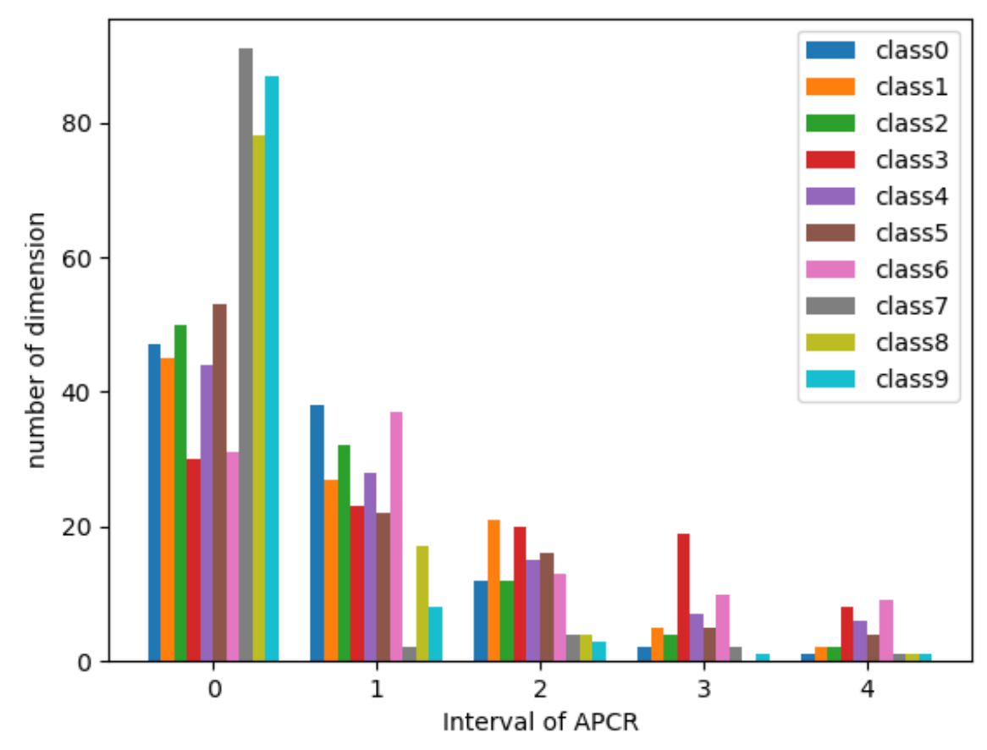

Figure 4: Number distribution of latent dimensions with respect to different APCR value range.

Weight optimization of latent variables

We use the optimization-based method described to derive the coefficients vector ($w \in [-1, 1]$) on latent dimensions for each concept.

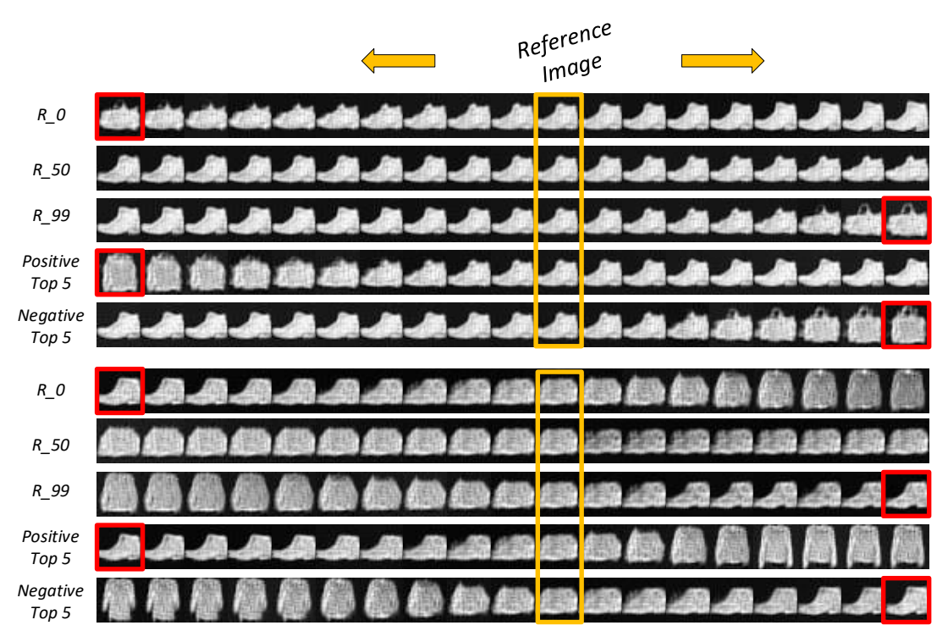

Figure 5: Intervene on controlling set of latent dimensions.

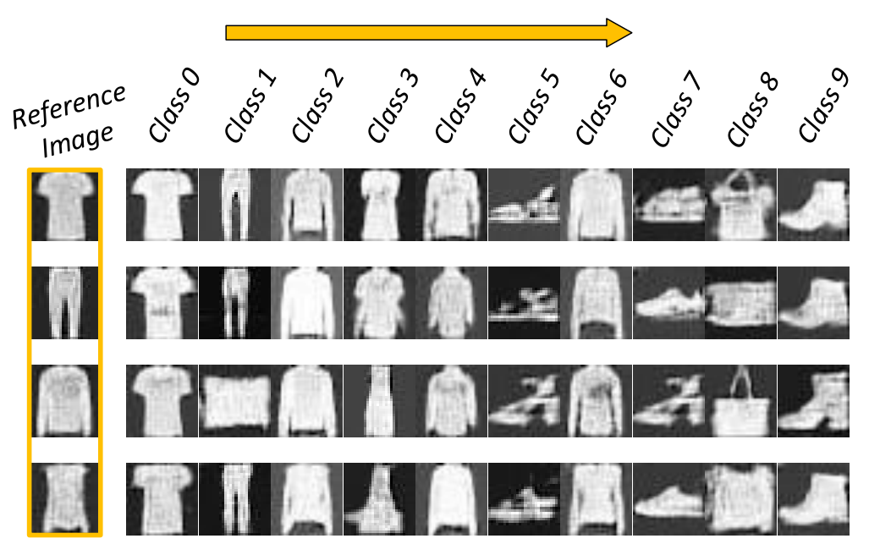

Figure 6: Controllable concept manipulation on Fashion-MNIST through intervening on controlling set of latent dimensions (final manipulation results).

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

References

[1] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 2008.

Acknowledgements

This work was partially supported by Visiting Foundation of USTC (project number: gxgnfx2019049) and the National Natural Science Foundation of China under grand No.U19B2044 and No.61836011.