# Closed-Form Factorization of Latent Semantics in GANs

Seungmin Lee (profile2697@gmail.com; 2020-20866),
Dept. of Electrical and Computer Engineering, Seoul National University

## 1. Introduction

The latent space of Generative Adversarial Network (GAN) has a rich set of interpretable directions that we can use to edit synthesized images. However, previous methods to find the interpretable directions require human annotations on a collection of synthesized images. In this paper, the authors propose a closed-form algorithm that identifies the semantic directions without using human annotations. More specifically, the proposed method discovers semantics by only using the weights of a pre-trained generator.

## 2. Preliminaries: Manipulating Generator in GAN Latent Space

A generator $G(\cdot)$ takes a $d$-dimensional latent vector $\mathbf{z}$ from the latent space $\mathcal{Z} \in \mathbf{R}^d$ and produces an image $\mathbf{I} = G(\mathbf{z})$. The authors focus on the first layer of the generator ($G_1 : \mathbf{R}^d \rightarrow \mathbf{R}^m$) since it directly acts on the latent space. Like most many GANs have done, the authors assume $G_1$ is an affine transformation:

$$\mathbf{y} := G_1(\mathbf{z}) = \mathbf{W}\mathbf{z} + \mathbf{b},$$

where $\mathbf{W} \in \mathbf{R}^{m \times d}$ and $\mathbf{b} \in \mathbf{R}^m$ denote the weights and bias, respectively.

We can manipulate the image generation by adding a particular direction $\mathbf{n} \in \mathbf{R}^d$ that represents a semantic concept to a given input vector:

$$\text{edit}(G(\mathbf{z})) = G(\mathbf{z} + \alpha\mathbf{n})$$

where $\alpha$ controls the manipulation intensity.

## 3. Method

At the first layer, we can simplify the manipulation as follows:

$$\begin{aligned} \mathbf{y}' &= G_1(\mathbf{z} + \alpha\mathbf{n}) = \mathbf{W}(\mathbf{z} + \alpha\mathbf{n}) + \mathbf{b} \\ &= (\mathbf{W}\mathbf{z} + \mathbf{b}) + \alpha\mathbf{W}\mathbf{n} \\ &= \mathbf{y} + \alpha\mathbf{W}\mathbf{n}. \end{aligned} \tag{1}$$

By observing Eq. (1), the authors claim that $\mathbf{W}$ already contains essential information about the image edition, and thus, we can discover useful latent directions by decomposing $\mathbf{W}$.

Assume that we want to find $k$ most significant directions $\mathbf{N}^* \in \mathbf{R}^{d \times k}$. Then, $\mathbf{N}^*$ should satisfy the following equation:

$$\mathbf{N}^* = \underset{\{\mathbf{N} \in \mathbf{R}^{d \times k}:\ \mathbf{n}_i^T \mathbf{n}_i = 1\ \forall i = 1, \cdots, k\}}{\arg\max} \sum_{i=1}^{k} ||\mathbf{W}\mathbf{n}_i||_2^2, \tag{2}$$

where $\mathbf{n}_i \in \mathbf{R}^d$ denotes the $i$-th column of $\mathbf{N}$, and $||\cdot||_2$ corresponds to $l_2$ norm.

The authors use the Lagrange multipliers $\{\lambda_i\}_{i=1}^k$ to solve Eq. (2). With the multipliers, Eq. (2) can be written as:

$$\begin{aligned} \mathbf{N}^* &= \underset{\mathbf{N} \in \mathbf{R}^{d \times k}}{\arg\max} \sum_{i=1}^{k} ||\mathbf{W}\mathbf{n}_i||_2^2 - \sum_{i=1}^{k} \lambda_i (\mathbf{n}_i^T \mathbf{n}_i - 1) \\ &= \underset{\mathbf{N} \in \mathbf{R}^{d \times k}}{\arg\max} \sum_{i=1}^{k} (\mathbf{n}_i^T \mathbf{W}^T \mathbf{W} \mathbf{n}_i - \lambda_i \mathbf{n}_i^T \mathbf{n}_i + \lambda_i). \end{aligned} \tag{3}$$

By taking partial derivative of Eq. (3), we get $\mathbf{n}_i$:

$$2\mathbf{W}^T\mathbf{W}\mathbf{n}_i - 2\lambda_i \mathbf{n}_i = 0. \tag{4}$$

## 4. Results

The authors conduct experiments using various GANs. The proposed method consistently finds useful directions, as we can observe from Fig. (1).

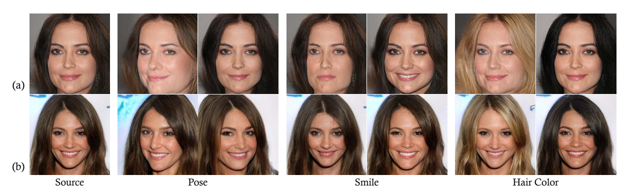


Figure 1. Some qualitative results.

## 5. Personal Note

The authors' claims are intuitive and useful. Moreover, the proposed method seems practical since it does not require any data samples and human annotations.