Interpretable GAN

January 24, 2024

Reviewr : Park Seok Hun

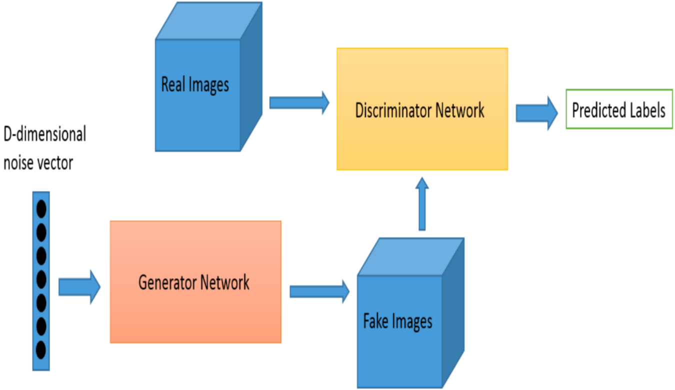
Table of Contents

1. GAN
   - Introduction
2. Interpretable GAN
   - Introduction
   - Algorithm
   - Experiments

Table of Contents

1. GAN
   - Introduction
2. Interpretable GAN
   - Introduction
   - Algorithm
   - Experiments

Introduction



- Generator is trained to generate fake images which is similar to real images.
- Discriminator is trained to discriminate between real and fake images.

Table of Contents

1. GAN
   - Introduction
2. Interpretable GAN
   - Introduction
   - Algorithm
   - Experiments

Introduction

- They proposed the new GAN structure which is interpretable.
- They force filters in the generator to have meaningful visual concepts without any manual annotations for visual concepts.
- They expect each filter in the layer have identical visual concept for any input.

Model design

- Suppose that $z_1, ..., z_N \in \mathbb{R}^d$ is the input latent vector.
- Suppose that there are C unique visual concepts and M filters in the generator.
- Let $\mathrm{Q} = \{q^1, ..., q^M\}$ be a partition of filters. In other words, $q^j \in \{1, ..., C\}$ means j-th filters have q^j visual concept.

Learning Q

- Let $f_G(z_i) = [f_i^1, ..., f_i^M], f_i^j \in \mathbb{R}^K$ be a feature map from j-th filter.
- We denote $F^j = [f_1^j, ..., f_N^j]$ as the feature map of j-th filter from dataset.
- $P_\theta(F^j) = \sum_{c=1,...,C} P_\theta(q^j = c) P_\theta(F^j | q^j = c)$ where $P_\theta(F^j | q^j = c)$ means the probability of j-th filter's feature maps in the c-th group.
- We can obtain MLE $\hat{\theta}$ with $f_i^j | q^j = c \sim N(\mu_c, \sigma_c^2 I)$ where $\theta = (p_c, \mu_c, \sigma_c^2)$.
- Therefore, we can obtain the group set Q by $q^j = \text{argmax}_{q^j} P_{\hat{\theta}}(q^j | F^j)$

Realism of generated images

- Given the partition Q for each filters, the realism of generated images can be decreased.
- To solve this problem, they used energy-based model.
- $L_{real}(W, G) = -\frac{1}{N}\sum_{i=1}^{N} \log P_W(f_G(z_i)|\text{Q})$

where

$$P_W(f_G(z)|\text{Q}) = \frac{1}{Z(W)} exp(g_W(f_G(z))) \quad (1)$$

$$= \frac{1}{Z(W)} exp(\sum_{j=1}^{M}\sum_{c=1}^{C}[W_{jc} \cdot (f^j \odot \bar{f}^c)]) \quad (2)$$

$Z(W) = \int exp(g_W(f_{G'}(z))dz$: normalized constant.

Interpretability of filters

- They expect each filters in the same group to have same visual concept and filters in the different group to have different visual concept.
- In other words, filter f^j in the c group have to be closed to the group $\bar{f}^c$.
- $L_{interpret}(W)$
  $= \sum_{j=1}^{M} \sum_{c=1}^{C} \sum_{k=1}^{K} -\mathbb{I}(q^j = c) W_{jck} + \lambda_1 \mathbb{I}(q_j \neq c) W_{jck}.$

  where $\lambda_1 > 0$.

- If $W_{ijk} > 0$, $g_W(f_G(z))$ forces f_j to be closed to the $\bar{f}^c$ where f_j is the function map which is in the group c.

Loss

- $Loss = L_{GAN} + \lambda_1 L_{real}(W, G) + \lambda_2 L_{interpret}(W)$

Experiments

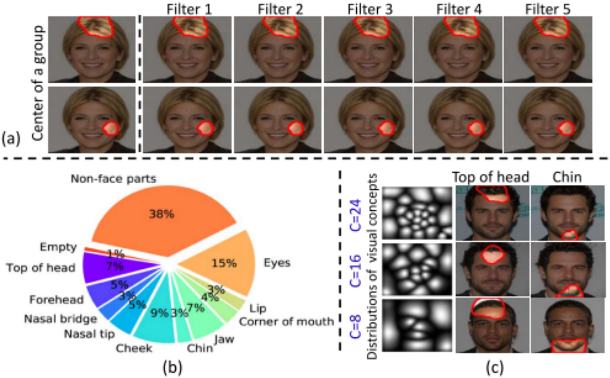
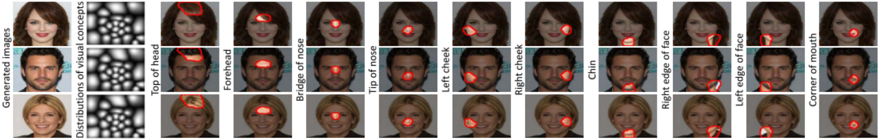
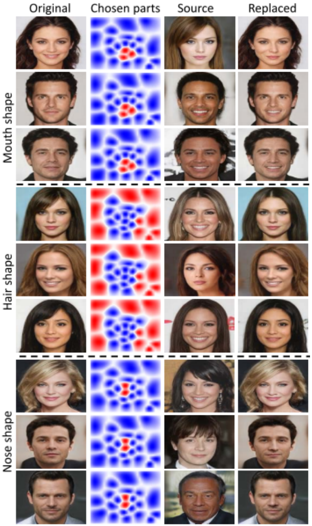


Figure 4: (a) Comparisons of receptive fields (RFs) between the center of a group and each filter in the group. (b) Proportions of filters representing different visual concepts. (c) Filters learned with different values of C.

Experiments



Experiments



Experiments

- They modify specific visual concepts on generated images. To be specific, they exchanged a specific visual concept between paris of images by exchanging the corresponding feature maps in the generator.