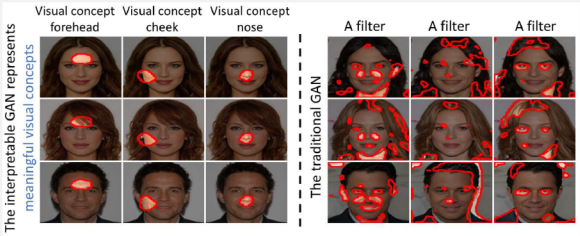
Interpretable Generative Adversarial Networks

Chao Li, Kelu Yao, Jin Wang, Boyu Diao, Yongjun Xu, Quanshi Zhang

Introduction



Traditional GAN: has no self-reflection of its representations.

Our Interpretable GAN:

- each filter consistently represents a meaningful visual concept when generating different images.
- different filters represent different visual concepts.

Algorithm

Given training images without annotations of visual concepts, we aim to train an interpretable GAN in an end-to-end manner. The interpretable GAN should satisfy the following two objectives.

Realism of generated images: the generator of the interpretable GAN still generates realistic images.

Interpretability of filters:

- Each filter is supposed to consistently generate image regions corresponding to the same visual concept when generating different images.
- Different filters are supposed to generate image regions corresponding to different visual concepts.

Method:

Given the generator, we learn partition Q of filters to ensure filters in the same group generate similar image regions.

To ensure the realism of generated images, we use an energy-based model to estimate the realism of the feature maps.

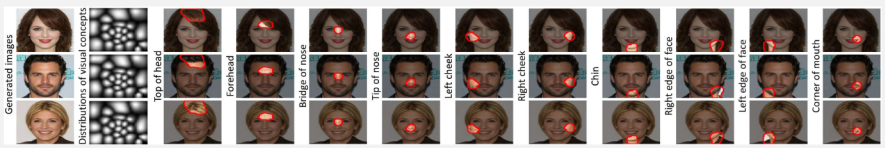
$$\mathcal{L}_{real}(W,G) = -\frac{1}{N}\sum_{i=0}^{n} log P_w(f_G(z_i)|Q) \quad P_w(f_G(z)|Q) = -\frac{1}{Z(W)} exp(g_w(f_G(z)))P_0(z)$$

To increase the interpretability of the filters in the target layer, we design the following loss:

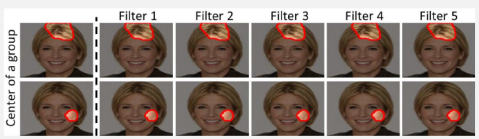
$$\mathcal{L}_{interp}(W) = -\sum_{j=1}^{M}\sum_{c=1}^{C}\sum_{k=1}^{K} I(q_j = c)W_{jck} + \lambda_1 \sum_{j=1}^{M}\sum_{c=1}^{C}\sum_{k=1}^{K} I(q_j \neq c)W_{jck}$$

Experiments

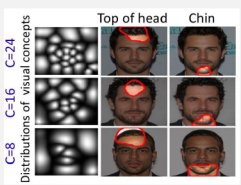
Visualization of feature maps in interpretable GANs



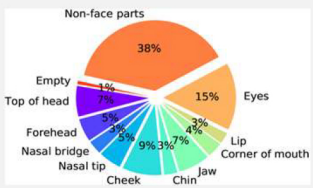
Visualization results show that each filter in an interpretable GAN consistently generated image regions corresponding to the same visual concept. Different filters generated image regions corresponding to different visual concepts.



Comparisons of receptive fields (RFs) between the center of a group and each filter in the group.

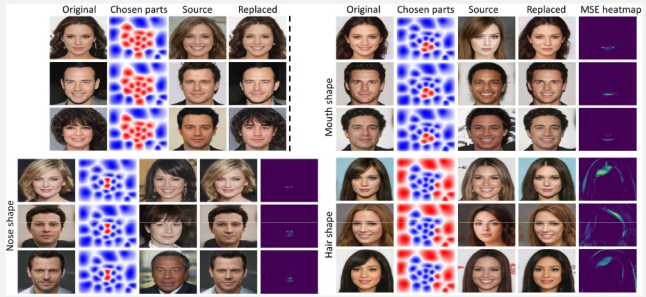


GANs with a larger value of group number learned more detailed concepts.



512 filters totally represented 11 visual concepts when setting C = 24.

Modifying visual concepts on images



Swapping the whole face or a specific visual concept between pairs of images by exchanging the corresponding feature maps in the interpretable layer. Result show our modifications were both perceptible and localized.

Improving the realism of generated / modified images by Langevin dynamics.

Quantitative analysis

Model	Mouth(%)	Eyes(%)	Chin(%)
Editing in Style (Collins et al. 2020)	37.90	34.60	-
Feature Collaging (Suzuki et al. 2018)	56.00	45.40	46.40
Interpretable StyleGAN	83.60	63.70	81.67
Interpretable BigGAN	89.60	82.10	92.30

Human perception evaluation shows the correctness of modifying specific visual concepts.

Model	Face verfication accuracy(%)
SimSwap (Chen et al. 2020)	87.40
FaceShifter [1] (Li et al. 2020)	85.45
FSGAN (Nirkin, Keller, and Hassner 2019)	89.20
Interpretable StyleGAN	90.25

Identity preserving evaluation shows the correctness of swapping the whole face.

Model	Mouth	Eyes	Chin
Editing in Style (Collins et al. 2020)	1.3649	0.9745	-
Feature Collaging (Suzuki et al. 2018)	0.1872	0.1293	0.0576
Interpretable StyleGAN	0.0606	0.0502	0.0163
Interpretable BigGAN	0.0296	0.0197	0.0311

Locality evaluation shows that our method had better localization.