Cross Sample Similarity for Stable Training of GAN

Jung Eun Lee
Department of Computer Science and Engineering
Kyunghee University, Republic of Korea
jeunlee0306@gmail.com

Seungkyu Lee
Department of Computer Science and Engineering
Kyunghee University, Republic of Korea
seungkyu@khu.ac.kr

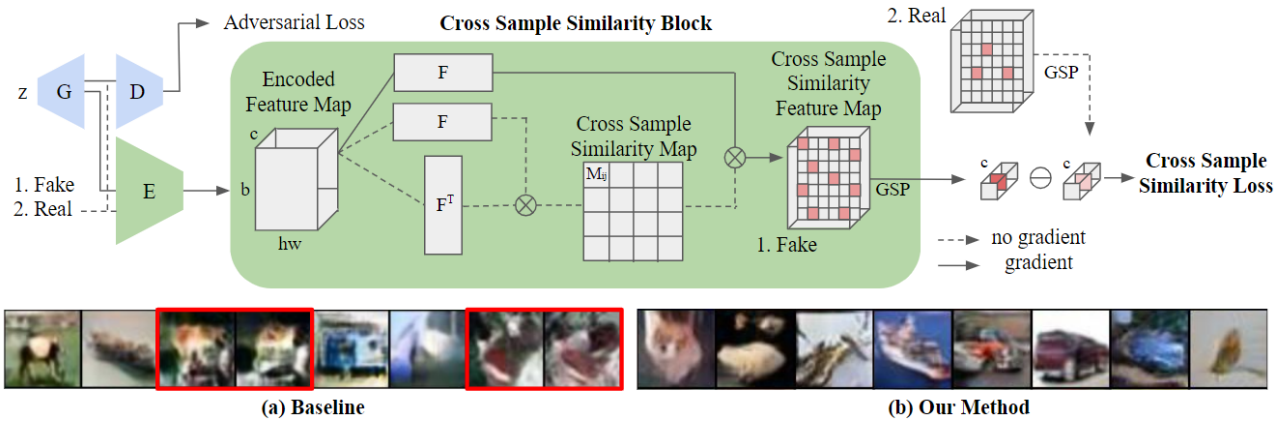

Figure 1: Cross sample similarity: It sees the similarity to penalize features creating bigger similarity from fake samples than reals. (a) and (b) are sample results without and with cross sample similarity loss in unconditional BigGAN with cifar10.

ABSTRACT

Recently attention network finding similarity in non-local area within a 2D image has shown outstanding improvement in multi-class generation task in GAN. However it frequently shows unstable training state sometimes falling in mode collapse. We propose cross sample similarity loss to penalize similar features of fake samples that are rarely observed in reals. Proposed method shows improved FID score compared to baseline methods on CelebA, LSUN, and decreased mode collapse on Cifar10[Krizhevsky 2009].

CCS CONCEPTS

• **Computing methodologies** → **Image representations**; *Unsupervised learning*.

KEYWORDS

deep generative model, generative adversarial nets(GAN), mode collapse, autoencoder

ACM Reference Format:
Jung Eun Lee and Seungkyu Lee. 2021. Cross Sample Similarity for Stable Training of GAN. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Posters (SIGGRAPH '21 Posters), August 09-13, 2021.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3450618.3469169

SIGGRAPH '21 Posters, August 09-13, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8371-4/21/08.
https://doi.org/10.1145/3450618.3469169

1 INTRODUCTION

Image synthesis with GAN (Generative Adversarial Networks) has shown tremendous improvements, especially in generating both globally and locally meaningful images. One of major approaches keeping details of long range dependency is employing non-local block in 2D images. It boosts multi-class generation performance in self-attention GAN [Zhang et al. 2019] and shows effectiveness in large-sized image generation [Brock et al. 2018]. We propose a cross sample similarity loss which calculates the similarity among batch samples and assigns penalty to image regions where outstanding similarity is observed among fake images but little among real images. By calculating the similarity within the batch unit not within a single image, it detects patterns that are frequently found across generated images even though they are hardly found in any single image. We investigate the similarity with pre-trained encoder in order to be free from discriminator's performance and mini-batch's distribution. Proposed cross sample similarity mitigates mode collapse as the training of generator network proceeded across several GANs. Figure 1 (a) shows mode collapse in unconditional BigGAN, and (b) shows results of our method.

2 PROPOSED METHOD

Our aim is to find features that are generated repeatedly in fake samples but are not observed much in real images. We give penalty to corresponding region of an image. We focus on the difference between real and fake similarity, not just the similarity of fake samples. Even though there is certain similarity frequently observed in fake samples, penalty should not be given if it is also observed a lot from real samples (for example, uniform and general background such as sky or sea). In practice, the distribution of real data in a

mini batch does not reflect the distribution of entire real samples. Furthermore, discriminator of GANs changes its encoded space in training progress. Therefore, we use pre-trained encoder for each batch to obtain corresponding distribution and accumulate cross sample similarities: $A_{t+1} = \frac{A_t \times t + E_{t+1}}{t+1}$ where t is the index of learning step, E is the cross sample similarity feature of real samples obtained from the encoder, and A is the accumulated cross sample similarity feature of real samples.

Figure 1 shows overall framework of our neural network. Proposed network is trained with both adversarial and cross sample similarity loss. F is the feature map of all images in the batch obtained through the pre-trained encoder E. It is reshaped to $(c, b \times w \times h)$ and normalized. $\otimes$ denotes matrix multiplication. $M_{i,j}$ in the similarity map indicates partial similarity map for image i and j. For example, $M_{1,3}$ stores the similarity of each pixel in first image and third image. Cross sample similarity map stores the pixel-by-pixel similarity across the feature level. In order to determine which channel of the feature map contribute to activate similar pixels, dot product of F and cross sample similarity map is conducted. The result is named cross sample similarity feature map. Since F is an activation map after ReLU and similarity map is all positive, cross sample similarity map is all positive.

$$D = GSP(S(G(z)) - GSP(S(x)) \tag{1}$$

$$Loss = \sum_{k=0}^{C} \|max(0, D_k)\|_2^2 \tag{2}$$

where S indicates cross sample similarity feature map, C is channel, GSP is Global Sum Pooling, G(z) is generated image and x is real image. If cross sample similarity feature map is obtained for fake image and real image, global sum pooling is conducted respectively. l2 norm is used for the channel-wise difference. Only if fake-real is positive, it is used as a loss. This is to give a penalty only to the features of the generator that create similarity that does not exist in the real image. Otherwise, even if the features are well-generated, they are penalized if the difference between two similarity is large. this aggravate mode collapse from the beginning.

3 EXPERIMENTAL RESULTS

We test on two public data sets (CelebA, the church-outdoor class of the LSUN) for quantitative evaluation and Cifar10 for visualized evaluation that shows noticeably diverse distribution. For quantitative evaluation, we employ Fréchet Inception Distance (FID). We compare our network with Wgan-GP with hinge loss, SNGAN with TTUR, uncoditional BigGAN. For pretrained encoder, we use VQVAE[Oord et al. 2017]. In Wgan-GP, it is modified to apply spectral normalization and hinge loss to balance adversarial and our loss. SNGAN with TTUR is SAGAN[Zhang et al. 2019] without self attention module. In unconditional BigGAN, we use the settings for training unconditional class in UnetGAN[Schonfeld et al. 2020]. We add a cross sample similarity loss to these baselines. Batch size is 64 for CelebA and LSUN, and 512 for Cifar10. Figure 2 shows that our method mitigates mode collapse during the learning progresses. For baseline, FID score starts to increase from certain iteration, but proposed method shows stable training results. To ensure that improved result is not simply because of the addition of pre-trained encoder, we conduct two experiments. First, we apply our module in discriminator instead of encoder. After few training epochs of baseline network, we add ours. Figure3 shows that, for the same epoch with same latent vector, baseline observes mode collapse while ours doesn't. Second, we test feature matching loss with encoder. Figure 2 shows that simple feature matching fails to remove mode collapse even though sometimes it shows better FID score.

Table 1: FID Score of Our Experiments

Method	CelebA	LSUN
SNGAN Baseline	20.91	35.0
SNGAN with Feature Matching	19.2	33.49
SNGAN with Cross Sample Similarity	**20.25**	**32.42**
WGAN-GP Baseline	26.56	52.65
WGAN-GP with Cross Sample Similarity	**23.65**	**55.3**

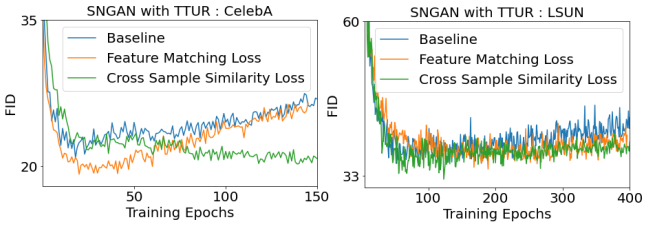

Figure 2: FID scores for baseline, cross sample similarity and feature matching on CelebA and LSUN dataset of 64*64 resolution. Baseline observes mode collapse from certain epoch.

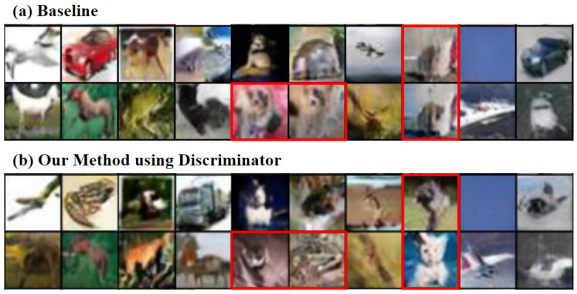

Figure 3: Comparison of baseline and our method using discriminator instead of encoder

4 CONCLUSION

We propose cross sample similarity penalizing similar features of fake samples that are rarely observed in real samples. Our method effectively stabilizes training of GANs preventing mode collapse.

REFERENCES

Andrew Brock, Jeff Donahue, and Karen Simonyan. 2018. Large Scale GAN Training for High Fidelity Natural Image Synthesis. *International Conference on Learning Representations* (2018).

Alex Krizhevsky. 2009. *Canadian Institute for Advanced Research.* Technical Report.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937* (2017).

Edgar Schonfeld, Bernt Schiele, and Anna Khoreva. 2020. A U-Net Based Discriminator for Generative Adversarial Networks. In *CVPR 2020.*

Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. 2019. Self-attention generative adversarial networks. In *International conference on machine learning*. PMLR, 7354–7363.