Improved Training of Generative Adversarial Networks using Representative Features

Duhyeon Bang [1] **Hyunjung Shim** [1]

Abstract

Despite the success of generative adversarial networks (GANs) for image generation, the trade-off between visual quality and image diversity remains a significant issue. This paper achieves both aims simultaneously by improving the stability of training GANs. The key idea of the proposed approach is to implicitly regularize the discriminator using representative features. Focusing on the fact that standard GAN minimizes reverse Kullback-Leibler (KL) divergence, we transfer the representative feature, which is extracted from the data distribution using a pre-trained autoencoder (AE), to the discriminator of standard GANs. Because the AE learns to minimize forward KL divergence, our GAN training with representative features is influenced by both reverse and forward KL divergence. Consequently, the proposed approach is verified to improve visual quality and diversity of state of the art GANs using extensive evaluations.

1. Introduction

Generative models aim to solve the problem of density estimation by learning the model distribution P_{model}, which approximates the true but unknown data distribution of P_{data} using a set of training examples drawn from P_{data} (Goodfellow, 2016). The generative adversarial networks (GANs) (Goodfellow et al., 2014) family of generative models implicitly estimate a data distribution without requiring an analytic expression or variational bounds of P_{model}. GANs have been mainly used for image generation, with impressive results, producing sharp and realistic images of natural scenes. The flexibility of the model definition and high quality outcomes has seen GANs applied to many real-world applications, including super-resolution, colorization, face generation, image completion, etc. (Bao et al., 2017; Ledig et al., 2016; Yeh et al., 2017; Cao et al., 2017).

Training a GAN requires two separate networks with competitive goals: a discriminator, D, to distinguish between the real and fake data; and a generator, G, to create as real as possible data to fool the discriminator. Consequently, the generator implicitly models P_{model}, which approximates P_{data}. This problem may be formulated as a minimax game (Goodfellow et al., 2014),

$$\min_G \max_D \mathop{\mathbb{E}}_{\mathrm{x}\sim \mathrm{P}_{\mathrm{data}}} [\log(\mathrm{D}(\mathrm{x}))] + \mathop{\mathbb{E}}_{z\sim \mathrm{P}_z} [\log(1-\mathrm{D}(\mathrm{G}(z))] \ ,$$

where $\mathbb{E}$ denotes expectation, x and z are samples drawn from P_{data} and P_{model} respectively.

When the generator produces perfect samples (i.e., $P_{model} \equiv P_{data}$), the discriminator cannot distinguish between real and fake data, and the game ends because it reaches a Nash equilibrium.

Although GANs have been successful in the image generation field, training process instabilities, such as extreme sensitivity of network structure and parameter tuning, are well-known disadvantages. Training instability produces two major problems: gradient vanishing and mode collapse. Gradient vanishing becomes a serious problem when any subset of P_{data} and P_{model} are disjointed such that the discriminator separates real and fake data perfectly; i.e., the generator no longer improves the data because the discriminator has reached its optimum (Arjovsky & Bottou, 2017). This produces poor results, because training stops even though P_{model} has not learned P_{data} properly. Mode collapse is where the generator repeatedly produces the same or similar output because P_{model} only encapsulates the major or single modes of P_{data} to easily fool the discriminator.

The trade-off between image quality and mode collapse has been theoretically and empirically investigated in previous studies (Berthelot et al., 2017; Fedus et al., 2017), and generally either visual quality or image diversity has been achieved, but not both simultaneously. Visual quality can be achieved by minimizing reverse Kullback-Leibler (KL) divergence, which is suggested in standard GANs including (Goodfellow et al., 2014). Meanwhile, image diversity is strongly correlated with minimizing forward KL divergence

[1]School of Integrated Technology, Yonsei University, South Korea. Correspondence to: Hyunjung Shim <kateshim@yonsei.ac.kr>.

Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

(Arjovsky & Bottou, 2017). Recent techniques (Kodali et al., 2017; Gulrajani et al., 2017; Fedus et al., 2017) have introduced a gradient penalty to regularize the divergence (or distance) for training GANs, and break the trade-off. The gradient penalty smooths the learning curve, improving training stability. Consequently, the gradient penalty is effective to improve both visual quality and image diversity, and has been evaluated for various GAN architectures.

We propose an unsupervised approach to *implicitly* regularize the discriminator using representative features. This approach is similar to the gradient penalty, in that it also aims to stabilize training and break the trade-off between visual quality and image diversity, but does not modify the GAN objective function (i.e., the same divergence or loss definition are employed as a baseline GAN). Rather, we introduce representative features from a pre-trained autoencoder (AE) and transfer them to a discriminator to train the GAN. Because the AE learns to minimize forward KL divergence, adding its representative features to the discriminator of standard GAN lead the discriminator to consider two divergences (i.e., reverse and forward KL). Since forward KL tends to average the overall modes of data distributions during training (Goodfellow, 2016), our representation features provide the overall mode information. Meanwhile, the objective of baseline discriminator pursues the reserve KL, thus tends to choose a single (few) mode of the data distribution. In other words, the discriminator is implicitly interrupted by representative features for discrimination, and encouraged to consider the overall data distribution.

The pre-trained AE learns from P_{data} samples and is then fixed. Isolating representative feature extraction from GAN training guarantees that the pre-trained AE embedding space and corresponding features have representative power. Since the representative features are derived from the pre-trained network, they are more informative during early stage discriminator training, which accelerates early stage GAN training. In addition, representative features provide the overall mode information as mentioned earlier, thus preventing GANs from mode collapse. Although the representative features no longer distinguish real and fake images in the second half of training, the discriminative features continue to learn toward improving the discrimination power. Note that the total loss of the proposed model consists of loss of representative and discriminative features, and the discriminator learns the balance between them from the training data automatically. Therefore, the proposed approach stably improve both visual quality and image diversity of generated samples. We call this new architecture a representative feature based generative adversarial network (**RFGAN**).

The major contributions of this paper are as follows.

1. We employ additional representative features extracted from a pre-trained AE to implicitly constrain discriminator updates. This can be interpreted as effectively balancing reverse and forward KL divergences, thus GAN training is stabilized. Consequently, we simultaneously achieve visual quality and image diversity in an unsupervised manner.
2. The proposed RFGAN framework can be simply extended to various GANs using different divergences or structures, and is also robust against parameter selections. The approach employs the same hyperparameters suggested by a baseline GAN.
3. Extensive experimental evaluations show RFGAN effectiveness, improving existing GANs including those incorporating gradient penalty (Kodali et al., 2017; Gulrajani et al., 2017; Fedus et al., 2017).

Section 2 reviews recent studies and analyzes how the proposed RFGAN relates to them. Section 3 discusses RFGAN architecture and distinctive characteristics, and Section 4 summarizes the results of extensive experiments including simulated and real data. The quantitative and qualitative evaluations show that the proposed RFGAN simultaneously improved image quality and diversity. Finally, Section 5 summarizes and concludes the paper, and discusses some future research directions.

2. Related Work

Various techniques have been proposed to improve GAN training stability, which mostly aim to resolve gradient vanishing and mode collapse. Previous studies can be categorized into two groups as discussed below.

1. GAN training by modifying the network design

   To avoid gradient vanishing, the minimax game based GAN formulation was modified to a non-saturating game (Goodfellow et al., 2014), changing the generator objective function from $J(G) = \mathbb{E}_{z\sim P_z} \log(1 - D(G(z)))$ to $J(G) = -\frac{1}{2}\mathbb{E}_{z\sim P_z} \log(D(G(z)))$. This relatively simple modification effectively resolved the gradient vanishing problem, and several subsequent studies have confirmed this theoretically and empirically (Arjovsky & Bottou, 2017; Fedus et al., 2017).

   (Radford et al., 2015) first introduced GAN with a stable deep convolutional architecture (DCGAN), and their visual quality was quantitatively superior to a variant of GANs proposed later, according to (Lucic et al., 2017). However, mode collapse was a major DCGAN weakness, and unrolled GANs were proposed to adjust the generator gradient update by introducing a surrogate objective function that simulated the discriminator response to generator changes (Metz et al., 2016). Consequently, unrolled GANs successfully solved model collapse.

   InfoGAN (Chen et al., 2016) achieved unsupervised

disentangled representation by minimizing the mutual information of auxiliary (i.e., matching semantic information) and adversarial loss. Additionally, (Salimans et al., 2016) proposed various methods to stabilize GAN training using semi-supervised learning and smoothed labeling.

2. Effects of various divergences.

   In (Nowozin et al., 2016), the authors showed that the Jensen-Shannon divergence used in the original GAN formulation (Goodfellow et al., 2014) can be extended to different divergences, including f-divergence (f-GAN). KL divergence has been theoretically shown to be one of the causes of GAN training instability (Arjovsky & Bottou, 2017; Arjovsky et al., 2017), and the Wasserstein distance was subsequently to measure the similarity between P_{model} and P_{data} to overcome this instability. Weights clipping was introduced into the discriminator to implement the Wasserstein distance (WGAN) (Arjovsky et al., 2017) to enforce the k-Lipschitz constraint. However, weight clipping often fails to capture higher moments of P_{data} (Gulrajani et al., 2017), and a gradient penalty was proposed to better model P_{data}. The discriminator becomes closer to the convex set using the gradient penalty as a regularization term (Kodali et al., 2017), which effectively improved Gan training stability. Least squares GAN (LSGAN) replaces the Jenson-Shannon divergence, defined by the sigmoid cross-entropy loss term, with a least squares loss term (Mao et al., 2017), which can be essentially interpreted as minimizing Pearson $\tilde{\chi}^2$ divergence.

Most previous GAN approaches have investigated stable architecture or additional layers to stabilize discriminator updates, changing the divergence, or adding a regularization term to stabilize the discriminator. The proposed RFGAN approach can be classified into the first category, modifying GAN architecture, and is distinct from previous GANs in that features from the encoder layers of the pre-trained AE are transferred while training the discriminator.

Several previous approaches have also considered AE or encoder architectures. ALI (Dumoulin et al., 2016), BiGAN (Donahue et al., 2016) and MDGAN (Che et al., 2016) proposed that P_{data} samples should be mapped to the generator latent space generator using the encoder structure. This would force the generator latent space to learn the entire P_{data} distribution, solving mode collapse. Although these are similar approaches, in that encoder layers are employed to develop the GAN, RFGAN uses an AE to provide a completely different method to extract representative features, and those features stabilize the discriminator.

EBGAN (Zhao et al., 2016) and BEGAN (Berthelot et al., 2017) proposed an energy based function to develop the discriminator, and such networks showed stable convergence and less sensitivity to parameter selection. They adopted AE architecture to define the energy based function, which served the discriminator. In contrast, RFGAN employs representative features from the encoder layers and retains conventional discriminator architecture to maintain its discriminative power. Another alternative approach extracted features from discriminator layers, applied a denoising AE, and used the output to regularize adversarial loss (Warde-Farley & Bengio, 2017). The technique improved image generation quality, and a denoising AE was employed to ensure robust discriminator features. However, this differs somewhat from the proposed RFGAN approach, where the AE is used as the feature extractor.

In contrast to previous approaches that trained the AE or encoder layers as part of the GAN architecture, RFGAN separately trains the AE to learn P_{data} in an unsupervised manner. Thus, feedback is disconnected from P_{model} when training the AE, and the focus is on learning the feature space to represent P_{data}. In addition, RFGAN does not utilize the decoder, avoiding problems such as image blur.

3. Representative Feature based GAN

To resolve GAN training instability, we extract representative features from a pre-trained AE and transfer them to the discriminator; implicitly enforcing the discriminator to be updated by effectively considering both reverse and forward KL divergence.

The aim of an AE is to learn a reduced representation of the given data, since it is formulated by reconstructing the input data after passing through the network. Consequently,feature spaces learnt by the AE are powerful representations to reconstruct the P_{data} distribution.

Several studies have utilized AE functionality as a feature extractor for classification tasks through fine-tuning (Zhou et al., 2012; Chen et al., 2015). However, a good reconstruction representation does not guarantee good classification (Alain & Bengio, 2014; Wei et al., 2015), because reconstruction and discriminative model features are derived from different objectives, and hence should be applied to their appropriate tasks for optimal performance.

When training a GAN, the discriminator operates as a binary classifier (Radford et al., 2015), so features extracted from the discriminator specialize in distinguishing whether the input is real or fake. Thus, discriminator features have totally different properties from AE features. Considering these different properties, we denote AE and discriminator features as representative and discriminative features, respectively. Although the original GAN formulation evaluates data generation quality purely based on discriminative features, we

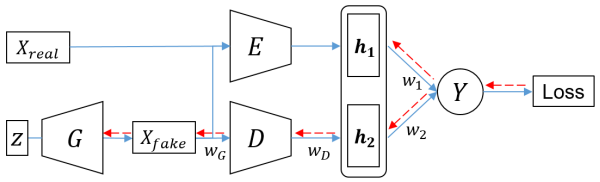


Figure 1. Representative feature based generative adversarial network graphical model. X_{real} and X_{fake} are input and generated images, respectively; E, G, and D are encoder, generator, and discriminator networks, respectively; Z is the latent vector; Y is binary output representing the real or synthesized image; h_1 and h_2 are representative and discriminative features, respectively; and w_1, w_2, w_D, and w_G are network parameters. Blue solid and red dash lines represent forward and backward propagation, respectively.

propose leveraging both representative and discriminative features to implicitly regularize the discriminator, and hence stabilize GAN training.

This section describes the proposed RFGAN model, and the effects of the modified architecture for training the discriminator. We also investigate how this effect could overcome mode collapse and improve visual quality.

3.1. RFGAN Architecture

The main contribution of the RFGAN model is adopting representative features from a pre-trained AE to develop the GAN. Thus, RFGAN can be based on various GAN architectures, and refers to a set of GANs using representative features. For simplicity, we use DCGAN (Radford et al., 2015) employing non-saturated loss as the baseline GAN, and apply representative features to the discriminator to construct DCGAN-RF. Section 4 introduces various GANs as baselines to develop RFGAN variants. We use exactly the same hyper-parameters, metrics, and settings throughout this paper, as suggested for a baseline GAN, to show that the RFGAN approach is insensitive to parameter selection, since representative features extracted are supplied from the encoder layer (part of the pre-trained AE) to the discriminator. The AE is pre-trained unsupervised using samples from P_{data}, and isolated from GAN training.

In particular, we construct the AE such that its encoder and decoder share the same architecture as the discriminator and generator, respectively. We then concatenate two feature vectors, one from the last convolution layer of the encoder and the other from the discriminator. Final weights are trained for the concatenated feature vector, to deciding between real or fake input. Figure 1 demonstrates the model for input data passing through encoder, E, and discriminator, D, networks; whereh_1 and h_2 represent the representative and discriminative feature vectors, respectively, which are concatenated and transformed to a single sigmoid output, Y, through a fully connected layer. The output is evaluated with the ground truth label based on sigmoid cross entropy, and then the gradient of the loss function is delivered to the discriminator via backpropagation to update the parameters. This feedback is not propagated to the encoder, because its parameters are already trained and subsequently fixed. The procedure for gradient updates is

$$D(x) = -\log Y \text{ for } x \sim P_{data},\ Y = \sigma(h_1 w_1 + h_2 w_2),$$

$$\nabla w_i = \frac{\partial D(x)}{\partial w_i} = -\frac{1}{Y} \cdot Y(1-Y) \cdot h_i = (Y-1)\, h_i,\ i \in \{1,2\},$$

$$\nabla w_D = \frac{\partial D(x)}{\partial w_D} = (Y-1) \cdot w_2 \cdot u(w_D)\,.$$

The GAN objective function represented by parameters

$$J(\theta_G, \theta_D) = \mathop{\mathbb{E}}_{x \sim P_{data}} [\log D(x; \theta_D)] + \mathop{\mathbb{E}}_{z \sim P_z} [\log (1 - D(G(z; \theta_G); \theta_D))\,]$$

is updated by

$$\theta_D^{t+1} \leftarrow \theta_D^t - \eta_D \frac{dJ\left(\theta_G,\ \theta_D^t\right)}{d\theta_D^t} = \theta_D^t - \eta_D (\nabla w_D + \nabla w_i)\,,$$

$$\theta_G^{t+1} \leftarrow \theta_G^t - \eta_G \frac{dJ\left(\theta_G,\ \theta_D^t\right)}{d\theta_G^t}\,.$$

where σ and u are sigmoid and step functions, respectively.

Since the encoder is pre-determined, we only consider discriminator updates. We can derive the gradient toward the discriminator by calculating the partial derivative of loss term with respect to w_D, which indicates the network parameters as shown in Fig. 1. Thus, ∇w_D depends on h_1, and the representative features affect the discriminator update. The procedure was derived for the case where x is real. In the case of a fake sample, the same conclusion is reached, except that $D(x)$ is now $-\log(1 - \sigma(h_1 w_1 + h_2 w_2))$.

Therefore, the generator is trained by considering both representative and discriminative features, because it should fool the discriminator by maximizing $-\log D(G(z))$. RFGAN representative features retain their properties, such as a global representation for reconstructing the data distribution, by fixing the encoder parameters.

3.2. Mode collapse

The AE decoder estimates the $P(x|En(x))$ distribution parameters based on a probabilistic interpretation, to generate x with high probability formulated by cross-entropy loss (Vincent et al., 2010). It is possible to interpret that the AE follows forward KL divergence between P_{data} and P_{model} (i.e., $\mathrm{KL}(P_{data}\ ||P_{model})$). Since the model approximated by forward KL divergence is evaluated using every true

data sample (i.e., any $x : P_{data}(x) > 0$), it tends to average all P_{data} modes (Goodfellow, 2016). Hence, representative features extracted from the AE are similar, in that they effectively represent entire P_{data} modes (Rosca et al., 2017).

On the contrary, the aim of DCGAN with a non-saturated loss (the base architecture of the RFGAN model) is to optimize reverse KL divergence between P_{data} and P_{model}, i.e., $KL(P_{model} \| P_{data}) - 2JSD$ (Arjovsky & Bottou, 2017), where JSD is Jensen–Shannon divergence. Since the reverse KL objective based model is examined for every fake sample (i.e., any $x : P_{model}(x) > 0$), it has no penalty for covering the entire true data distribution. Hence, it is likely to focus on single or partial modes of the true data distribution, which is the mode collapse problem.

The proposed RFGAN optimizes reverse KL divergence because the framework is built upon a non-saturated GAN. We also introduce AE representative features simultaneously, which encourages the model to cover the entire P_{data} modes, similar to optimizing forward KL divergence. This suppresses the tendency toward mode collapse.

3.3. Improving visual quality

Although representative features are useful to advance the discriminator in the early stage, they become less informative when approaching the second half of training, because the AE has limited performance for discrimination. Since the AE is built by minimizing reconstruction error, e.g. L2 or L1 loss; the model cannot learn multiple different correct answers, which causes the model to choose the average (or median) output (Goodfellow, 2016). While this is useful to distinguish between poor fake and real input, when the generator starts producing good fake input, AE representative features are less discriminative, and hence interfere with decisions by the discriminator in the later training stages. Figure 2 shows the output for several real and fake examples passed through the pre-trained AE. Real or fake inputs are easily distinguished at the beginning of training, but after several iterations they look similar. These experimental results demonstrate AE discriminative power for different levels of fake examples.

Thus, it is difficult to improve data generation visual quality beyond a certain level using representative features alone. Therefore, the proposed RFGAN model employs both representative and discriminative features to train the discriminator. Although representative features interfere with discrimination between real and fake input as training progresses, the RFGAN discriminator retains discriminative features, which allows training to continue. Consequently, the generator consistently receives sufficient feedback from the discriminator, i.e., the gradient from the discriminator increases, to learn P_{data}. Since these two features are opposing, and disagree with each other, the discriminator is stabilized without abnormal changes. By stabilizing discriminator growth, the RFGAN model generates high quality data, improving the original GAN.

Figure 2. Reconstruction comparison after iteration with the pre-trained AE. The first row shows generated images that are passed through the pre-trained AE along with real images, as shown in the second and third rows.

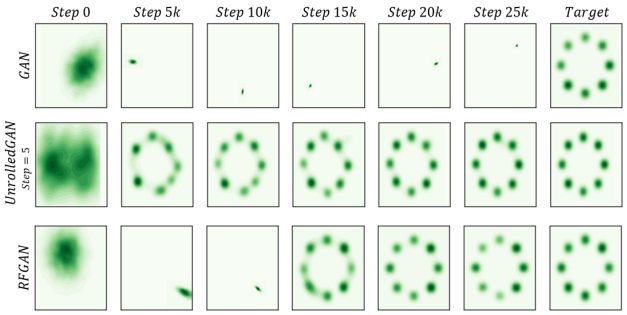


Figure 3. Mode collapse test learning a mixture of eight Gaussian spreads in a circle

4. Experimental results

For quantitative and qualitative evaluations, we include simulated and three real datasets: CelebA (Liu et al., 2015), LSUN-bedroom (Yu et al., 2015), and CIFAR-10 (Krizhevsky & Hinton, 2009), normalizing between -1 and 1. A denoising AE (Vincent et al., 2008) is employed to improve feature extraction robustness, achieving a slight quality improvement compared to conventional AEs. Since the concurrent training of AE and GAN does not improve the performance, we use the pre-trained and then fixed AE for reducing computational complexity.

4.1. Mode collapse

To evaluate how well the RFGAN model could achieve data generation diversity, i.e., solving mode collapse, we train the network with a simple 2D mixture of 8 Gaussians (Metz et al., 2016). The Gaussian means form a ring, and each distribution has standard deviation = 0.1. Figure 3 compares RFGAN, GAN, and unrolled GAN models, and confirms that GAN suffers from mode collapse while unrolled GAN effectively solves this problem (Metz et al., 2016).

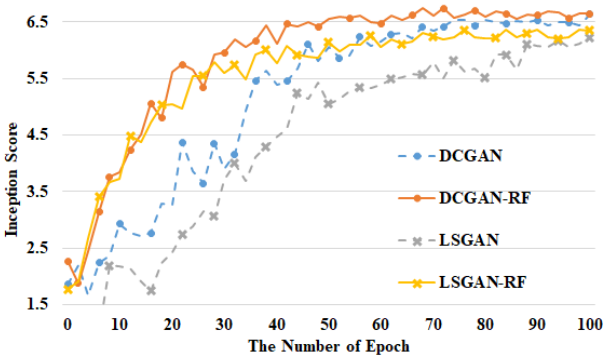
Previous studies solved mode collapse similarly to unrolled GAN by covering the entire distribution region and then gradually localizing the modes (Arjovsky et al., 2017; Donahue et al., 2016). However, RFGAN first learns each mode, and escapes from mode collapse by balancing representative features. This is because RFGAN minimizes reverse KL divergence, but is simultaneously influenced by representative features derived from forward KL divergence. When the representative features no long distinguish between real and fake input, the generator has achieved the representation power of the representative features. In other words, the generator learns the entire mode as well as the representative features, and then escapes mode collapse. Therefore, RFGAN first responds similarly to GAN, then gradually produces the entire mode.

4.2. Quantitative evaluation

Since RFGAN is built upon the baseline architecture and its suggested hyper-parameters, input dimensionality is set at (64, 64, 3), which is acceptable for the CelebA and LSUN datasets. However, we modify network dimensions for the CIFAR-10 dataset, fitting the input into (32, 32, 3) to ensure fair and coherent comparison with previous studies. We also drew 500 k images randomly from the LSUN bedroom dataset for efficient training and comparison.

Two metrics were employed to measure visual quality and data generation diversity, respectively. The inception score (Salimans et al., 2016) measured visual quality for GANs using CIFAR-10 datasets, with larger score representing higher quality. The MS-SSIM metric is often employed to evaluate GAN diversity (Odena et al., 2016), with smaller MS-SSIM implying better the performance in producing diverse samples.

The inception score correlates well with human annotator quality evaluations (Salimans et al., 2016), and hence is widely used to assess visual quality of GAN generated samples. We compute the inception score for 50 k GAN generated samples (Salimans et al., 2016), using DCGAN based architecture to allow direct comparison with previous GANs. To show that the proposed algorithm was extendable to different GAN architectures, we also apply the proposed framework to other state of the art GANs (LSGAN (Mao et al., 2017), DRAGAN (Kodali et al., 2017), and WGAN-GP (Gulrajani et al., 2017)); modifying their discriminators by adding representative features, and training them with their original hyper-parameters. The WGAN-GP generator is updated once after the discriminator is updated five times. Following the reference code[1], other networks are trained by updating the generator twice and the discriminator once. This ensures that discriminator loss did not vanish, i.e., the loss does not become zero, which generally provides better performance.



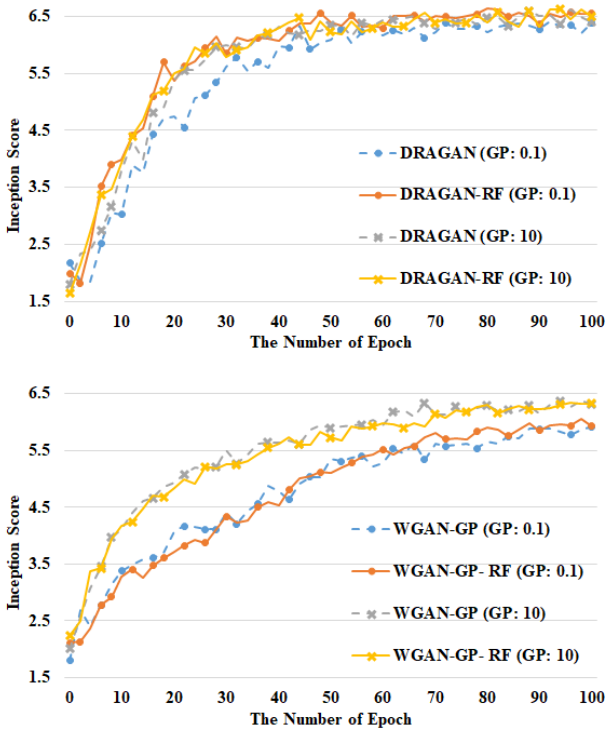
	DCGAN	DCGAN-RF	LSGAN	LSGAN-RF
Inception score	6.5050	6.6349	5.9843	6.2791

Figure 4. CIFAR10 inception score for DCGAN and LSGAN models with and without the representative feature approach

Figure 4 compares inception scores as a function of epoch. We compare DCGAN, DCGAN-RF, LSGAN, and LSGAN-RF, where "-RF " extension refers to the base model with the representative feature. DCGAN-RF and LSGAN-RF outperform DCGAN and LSGAN, respectively, in terms of inception scores. In addition, DCGAN-RF and LSGAN-RF inception scores grow faster than DCGAN and LSGAN, respectively, confirming that the proposed representative feature improves training efficiency. Thus, the proposed algorithm approaches the same visual quality faster than the baseline GAN.

The DRAGAN and WGAN-GP baseline GANs were recently proposed, using gradient penalty as a regularization term to train the discriminator. We also extend these using the proposed representative feature and compared with the original baseline GANs, as shown in Fig. 5. The proposed modification still improves inception scores, although the improvement is not as significant as with DCGAN and LSGAN for coefficient of gradient penalty = 10. Interestingly, this coefficient plays an important role regarding the inception score, with larger coefficients producing stronger gradient penalty. Discriminator training is disturbed when

[1] https://github.com/carpedm20/DCGAN-tensorflow



	GP	DRAGAN	DRAGAN-RF	WGAN-GP	WGAN-GP-RF
Inception Score	0.1	6.3191	6.5314	5.7951	5.9141
	10	6.4783	6.4905	6.2680	6.2699

Figure 5. CIFAR10 inception score for (top) DRAGAN, and (bottom) WGAN-GP, with and without the representative feature for gradient penalty coefficients = 0.1 and 10

the gradient penalty term became sufficiently strong, because gradient update is directly penalized. The score gap between DRAGAN and DRAGAN-RF increases when the coefficient of gradient penalty = 0.1, as expected since the DRAGAN performance approaches that of DCGAN as the gradient penalty decreases. However, since WGAN-GP replaces weight clipping in WGAN with a gradient penalty, it does not satisfy the k-Lipschitz constraint with low gradient penalty, which degrades WGAN-GP performance. Thus, it is difficult to confirm the tendency of the proposed representative feature for various WGAN-GP coefficients.

The DCGAN-RF model produces the best overall score, including previous GANs with or without the proposed representative feature, which is consistent with previous studies that showed DCGAN to be the most effective model for high quality image generation (Lucic et al., 2017). The proposed model achieves 0.128 mean improvement over the relevant baseline GAN, which is significant, and comparable or greater than the differences between different GANs. The improvement is particularly noticeable between LSGAN and LSGAN-RF.

Table 1. GAN diversity using the MS-SSIM metric. Real dataset MS-SIMM = 0.3727. NB: low MS-SSIM implies higher diversity.

	DCGAN	LSGAN	DRAGAN	WGAN-GP
ORIGINAL	0.4432	0.3907	0.3869	0.3813
WITH RF	0.4038	0.3770	0.3683	0.3773

The MS-SSIM metric computes similarity between image pairs randomly drawn from generated images (Odena et al., 2016), and was introduced it as a suitable measure for image generation diversity. However, MS-SSIM is meaningless if the dataset is already highly diverse (Fedus et al., 2017). Therefore, we use only the CelebA dataset to compare MS-SSIM, since CIFAR-10 is composed of different classes, hence already includes highly diverse samples; and LSUN-bedroom also exhibits various views and structures, so has a diverse data distribution. We choose four previously proposed GANs as baseline algorithms: DCGAN, LSGAN, DRAGAN, and WGAN-GP, and compare them with their RFGAN variants (DCGAN-RF, LSGAN-RF, DRAGAN-RF, and WGAN-GP-RF, respectively), as shown in Table 1. the proposed GANs (RFGANs) significantly improve diversity (i.e., reduced MS-SSIM) compared with the baseline GANs, consistent over all cases, even in the presence of the gradient penalty term.

The LSGAN-RF, WGAN-GP-RF, and DRAGAN-RF scores are close to that of the real dataset diversity, i.e., the generator produces diverse samples reasonably well. DRAGAN-RF achieves the best MS-SSIM performance, generating the most diverse samples, whereas DCGAN-RF demonstrates the most notable improvement over the baseline (DCGAN), since DCGAN frequently suffers from mode collapse. Thus, the experimental study confirms that RFGAN effectively improved generated image diversity.

In addition to four baseline GANs, we also compare our results with ALI/BiGAN and AGE, which utilizes the encoders for GAN training. Since they focus on resolving mode collapse, the MS-SSIM is close to our results, but the inception score of ALI/BiGAN and AGE are much worse than our results; MS-SSIM of ALI/BiGAN and AGE is 3.7938 and 3.8133 respectively, and the inception score of ALI/BiGAN and AGE is and 5.34 and 5.90 respectively when our model achieves around 0.3816 of MS-SSIM and more than 6.20 of the inception score.

4.3. Qualitative evaluation

We compare DCGAN and DCGAN-RF generated images from the same training iteration, as shown in Fig. 6. The proposed RFGAN produces significantly enhanced results, and also speeds up the training process, with RFGAN visual quality being similar to results from later DCGAN iterations, which is consistent with Fig. 4. .

Figure 6. Stepwise visual quality comparison between generated images using DCGAN and DCGAN-RF trained with (left) CelebA and (right) LSUN

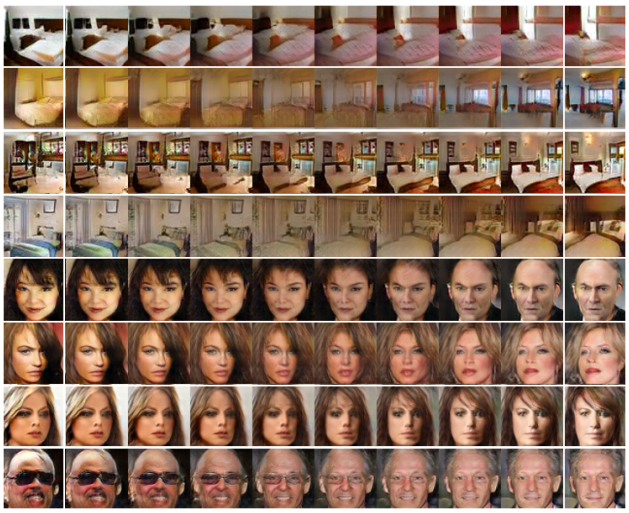


Figure 7. Latent space interpolations from LSUN and CelebA datasets. Left and right-most columns are samples randomly generated by DCGAN-RF, and intermediate columns are linear interpolations in the latent space between them.

Since we reuse the training data to extract representative features, it is possible the performance enhancement came from overfitting the training data. To demonstrate that the enhancements are not the result of data overfitting, we generate samples by walking in latent space, as shown in Fig. 7. Interpolated images between two images in latent space do not have meaningful connectivity, i.e., there is a lack of smooth transitions (Radford et al., 2015; Bengio et al., 2013; Dinh et al., 2016). This confirms that RFGAN learns the meaningful landscape in latent space, because it produces natural interpolations of various examples. Thus, RFGAN does not overfit the training data.

5. Conclusions

This study proposes an improved technique for stabilizing GAN training and breaking the trade-off between visual quality and image diversity. Previous GANs explicitly add regularization terms, e.g. gradient penalty, to improve training stability, whereas the proposed RFGAN approach implicitly hinders fast discriminator update growth, thus achieving stable training. RFGAN employs representative features from an AE pre-trained with real data. Our model achieves stabilizing and improving GAN training because RFGAN is influenced by two different characteristics of reverse and forward KL; learning the average mode and choosing a single mode. Consequently, we successfully improve generated sample visual quality and solve mode collapse. We also show that the proposed RFGAN approach is easily extendable to various GAN architectures, and robust to parameter selection.

In the future, our framework can be extended to various directions. For example, it is possible to utilize other types of features or more proper architectures, or training schemes that could further improve GAN performance. Specifically, replacing the convolution layer of the encoder to the residual block improves the visual quality; the inception score is increased from 6.64 to 6.73 for DCGAN-RF. The current study provides a basis for work employing various features or prior information to better design GAN discriminators.

6. Acknowledgement

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ICT Consilience Creative Program (IITP-2018-2017-0-01015) supervised by the IITP(Institute for Information & communications Technology Promotion), the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2018-2016-0-00288) supervised by the IITP(Institute for Information & communications Technology Promotion), the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the MSIP (NRF-2016R1A2B4016236), and ICT R&D program of MSIP/IITP. [R7124-16-0004, Development of Intelligent Interaction Technology Based on Context Awareness and Human Intention Understanding] We thank Dr. Jongwuk Lee for offering constructive feedback about the autoencoder training and writing.

References

Alain, G. and Bengio, Y. What regularized auto-encoders learn from the data-generating distribution. *The Journal of Machine Learning Research*, 15(1):3563–3593, 2014.

Arjovsky, M. and Bottou, L. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

Bao, J., Chen, D., Wen, F., Li, H., and Hua, G. Cvae-gan: Fine-grained image generation through asymmetric training. *arXiv preprint arXiv:1703.10155*, 2017.

Bengio, Y., Mesnil, G., Dauphin, Y., and Rifai, S. Better mixing via deep representations. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 552–560, 2013.

Berthelot, D., Schumm, T., and Metz, L. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

Cao, Y., Zhou, Z., Zhang, W., and Yu, Y. Unsupervised diverse colorization via generative adversarial networks. *arXiv preprint arXiv:1702.06674*, 2017.

Che, T., Li, Y., Jacob, A. P., Bengio, Y., and Li, W. Mode regularized generative adversarial networks. *arXiv preprint arXiv:1612.02136*, 2016.

Chen, L., Rottensteiner, F., and Heipke, C. Feature descriptor by convolution and pooling autoencoders. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 40(3):31, 2015.

Chen, X., Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 2172–2180. Curran Associates, Inc., 2016.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

Donahue, J., Krähenbühl, P., and Darrell, T. Adversarial feature learning. *CoRR*, abs/1605.09782, 2016. URL http://arxiv.org/abs/1605.09782.

Dumoulin, V., Belghazi, I., Poole, B., Lamb, A., Arjovsky, M., Mastropietro, O., and Courville, A. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016.

Fedus, W., Rosca, M., Lakshminarayanan, B., Dai, A. M., Mohamed, S., and Goodfellow, I. Many paths to equilibrium: Gans do not need to decrease adivergence at every step. *arXiv preprint arXiv:1710.08446*, 2017.

Goodfellow, I. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 27*, pp. 2672–2680. Curran Associates, Inc., 2014.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028*, 2017.

Kodali, N., Abernethy, J., Hays, J., and Kira, Z. On convergence and stability of gans. *arXiv preprint arXiv:1705.07215*, 2017.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*, 2016.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015.

Lucic, M., Kurach, K., Michalski, M., Gelly, S., and Bousquet, O. Are gans created equal? a large-scale study. *arXiv preprint arXiv:1711.10337*, 2017.

Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., and Smolley, S. P. Least squares generative adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2813–2821. IEEE, 2017.

Metz, L., Poole, B., Pfau, D., and Sohl-Dickstein, J. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*, 2016.

Nowozin, S., Cseke, B., and Tomioka, R. f-gan: Training generative neural samplers using variational divergence minimization. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, pp. 271–279. Curran Associates, Inc., 2016.

Odena, A., Olah, C., and Shlens, J. Conditional image synthesis with auxiliary classifier gans. *arXiv preprint arXiv:1610.09585*, 2016.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Rosca, M., Lakshminarayanan, B., Warde-Farley, D., and Mohamed, S. Variational approaches for auto-encoding generative adversarial networks. *arXiv preprint arXiv:1706.04987*, 2017.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., and Chen, X. Improved techniques for training gans. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 2234–2242. Curran Associates, Inc., 2016.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pp. 1096–1103, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-205-4. doi: 10.1145/1390156.1390294.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(Dec):3371–3408, 2010.

Warde-Farley, D. and Bengio, Y. Improving generative adversarial networks with denoising feature matching. In *International Conference on Learning Representations*, 2017.

Wei, H., Seuret, M., Chen, K., Fischer, A., Liwicki, M., and Ingold, R. Selecting autoencoder features for layout analysis of historical documents. In *Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing*, pp. 55–62, New York, NY, USA, 2015. ACM.

Yeh, R. A., Chen, C., Lim, T. Y., Schwing, A. G., Hasegawa-Johnson, M., and Do, M. N. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5485–5493, 2017.

Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., and Xiao, J. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Zhao, J., Mathieu, M., and LeCun, Y. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016.

Zhou, G., Sohn, K., and Lee, H. Online incremental feature learning with denoising autoencoders. In Lawrence, N. D. and Girolami, M. (eds.), *Artificial Intelligence and Statistics*, pp. 1453–1461, La Palma, Canary Islands, 2012. PMLR.