Hierarchical Modes Exploring in Generative Adversarial Networks

Mengxiao Hu, Jinlong Li, Maolin Hu, Tao Hu
University of Science and Technology of China
m_x_hu@126.com, jlli@ustc.edu.cn, {humaolin, Skyful}@mail.ustc.edu.cn

Abstract

In conditional Generative Adversarial Networks (cGANs), when two different initial noises are concatenated with the same conditional information, the distance between their outputs is relatively smaller, which makes minor modes likely to collapse into large modes. To prevent this happen, we proposed a hierarchical mode exploring method to alleviate mode collapse in cGANs by introducing a diversity measurement into the objective function as the regularization term. We also introduced the Expected Ratios of Expansion (ERE) into the regularization term, by minimizing the sum of differences between the real change of distance and ERE, we can control the diversity of generated images w.r.t specific-level features. We validated the proposed algorithm on four conditional image synthesis tasks including categorical generation, paired and un-paired image translation and text-to-image generation. Both qualitative and quantitative results show that the proposed method is effective in alleviating the mode collapse problem in cGANs, and can control the diversity of output images w.r.t specific-level features.

Introduction

With the potentiality of capturing high dimensional probability distributions, Generative Adversarial Networks (GANs) (Goodfellow 2016) are broadly used in synthesizing text (Yu et al. 2017a), videos (Zhang and Peng 2018) and images (Ge et al. 2018). Conditional GANs (cGANS) (Mirza and Osindero 2014) are one of the early variants of GANs and have been applied in many tasks of image synthesis (Doan et al. 2019) because of the ability to synthesizing images with given information (e.g, generating images of bird with given description of colors).

Many generation tasks adopt GANs for its simple setting and impressive result, but we often suffer from the problem that the generator can only synthesize samples from few modes of the real data distribution, which is called "mode collapse". The formal definition of mode collapse (Lin et al. 2018) provides a theoretical measure of mode collapse.

In image synthesis, mode collapse means the output images are less diverse than the real ones. Therefore, it might be very important to quantify the diversity of output images for addressing the mode collapse problem. The Learned Perceptual Image Path Similarity (LPIPS) (Zhang et al. 2018) and the Fréchet Inception Distance (FID) (Heusel et al. 2017) are often used to measure the diversity and the quality of output images, respectively. The Number of Statistically-Different Bins (NDB) (Richardson and Weiss 2018) is also used for estimation of mode missing in the generated distribution (Mao et al. 2019). Unlike previous works only use the diversity metrics for evaluation, in this work, we proposed hierarchical Modes Exploring Generative Adversarial Networks to alleviate the mode collapse problem in cGANs by introducing a diversity measurement into the objective function as a regularization term.

The regularization term was employed to expand the difference of an output image pair which is obtained by feeding an input pair from the same batch to the generator (Mao et al. 2019; Yang et al. 2019). Here, we firstly compute the ratio of the distance between an input pair to the distance between an output feature pair and use it as a coefficient of the expansion at each convolutional layer of the generator. Then, we calculate the absolute difference between the computed ratio and a predefined ratio which is used to control the generated images with different features. At last, we sum the absolute differences across all layers as the regularization term of the objective function. Since our regularization method requires no modification to the structure of original networks, it can be used in cGANs for various tasks.

In this work, our primary contributions are:

- We proposeed a hierarchical mode exploring method to alleviate mode collapse in cGANs by introducing a diversity measurement into the objective function as the regularization term.
- We introduced the Expected Ratio of Expansion (ERE) into the regularization term. With different ERE, we can control the diversity of generated images w.r.t specific-level features.
- We demonstrated the proposed regularization method on different datasets in three image synthesis tasks, and experimental results show that our method can generate images with higher diversity, compared with the baseline models.

Backgrounds

Generative adversarial networks (GANs) are composed of two players: the generator *G* and the discriminator *D*. The training process of GANs is a minimax game between *D* and *G* . *G* is learning to transform the initial noise to data with same dimension as real data's, such that *D* cannot tell whether the output was drawn from the true distribution or was generated by *G*. The solution of this minimax game is a Nash equilibrium in which neither *G* nor *D* can improve unilaterally.

In cGANs, because with the same concatenated information vectors, the distance between two inputs is smaller than the distance between two noise vectors, and the shrinkage of distance is very likely to be persevered through a upsampling layer (e.g., a fractionally-strided convolution layer) which is necessary for *G*. So, we can expect the distance between two outputs of *G* is smaller,compared to the one in standard GANs. To offset the shrinkage of distance, people used a regularization term $\mathcal{L}_d$ to maximize the distance between two output images (Mao et al. 2019; Yang et al. 2019).

$$\mathcal{L}_d = \left\| \frac{d^{(1)}(\mathbf{z_1}, \mathbf{z_2})}{d^{(n)}(G^{(n)}(\mathbf{z_1}), G^{(n)}(\mathbf{z_2}))} \right\|_1 \tag{1}$$

Where, $\mathbf{z}$ is the latent code, $d^{(i)}(\cdot)$ refers the distance metric, $G^{(i)}(\mathbf{z})$ is the output of i-th convolutional layer in the generator when the input is $\mathbf{z}$, and n is the number of convolutional layers in the generator.

Methods

Regularization term

Given an input pair $(\mathbf{z_1}, \mathbf{z_2})$, we firstly measure the distance between the output pair of every convolutional layer, then, we compute the ratio of the distance at one layer to the distance at the next layer:

$$ratio^{(i)} = \frac{d^{(i-1)}(G^{(i-1)}(\mathbf{z_1}), G^{(i-1)}(\mathbf{z_2}))}{d^{(i)}(G^{(i)}(\mathbf{z_1}), G^{(i)}(\mathbf{z_2}))} \tag{2}$$

As shown by Eq. (2), we denote the sum of L_1 norms of the difference between the computed ratio and a target ratio as the regularization term:

$$\mathcal{L}_h = \sum_{i=2}^{n} \left\| ratio^{(i)} - \lambda^{(i)} \right\|_1 \tag{3}$$

Where, $\lambda^{(i)}$ is the hyper-parameter to control the diversity gain though i-th layer. When $\lambda^{(i)} = 0$ for $\forall i \in \{2, ..., n\}$, we maximize the diversity of the output images by minimizing the regularization term $\mathcal{L}_h$.

The proposed regularization method offsets the shrinkage of distance between two outputs at every layer in the generator. When we minimize $\mathcal{L}_h$ with $\lambda^{(i)}$ being set to an appropriate value, it can alleviate the mode collapse problem at every layer of the generator.

To illustrate the advantage of $\mathcal{L}_h$, we compared $\mathcal{L}_h$ with $\mathcal{L}_d$ proposed by Mao *et al.*, and there is a true proposition.

Proposition 1. Denotes the rate of $ratio_{(i)}$ converging to 0 as $r^{(i)}$. If $\exists i \in \{2, ..., n\}$ such that $r^{(i)} \gg r^{(i')}$ for $\forall i' \in \{2, ..., n\} \setminus \{i\}$, in both training processes using $\mathcal{L}_h$ and $\mathcal{L}_d$, namely, $r_l^{(i)} \gg r_{l'}^{(i')}$, for $\forall l, l' \in \{\mathcal{L}_h, \mathcal{L}_d\}$, then, $\exists H \in (0, +\infty), (\mathcal{L}_h \leq H, and, \mathcal{L}_d \leq H) \implies l(t_h) < l(t_d)$. Here, when $l \leq H$, the corresponding training process is immediately stopped, t_h is the stop time when the regularization term is $\mathcal{L}_h$.

Proof. Assume $ratio^{(i)}_{\mathcal{L}_h}(t) = ratio^{(i)}_{\mathcal{L}_d}(t) = e^{-r^{(i)}_{\mathcal{L}_h} t}$. According to the condition in the proposition, assume $ratio^{(i')}_{\mathcal{L}_h}(t) \approx ratio^{(i')}_{\mathcal{L}_d}(t) = e^{-r^{(i')}_{\mathcal{L}_h} t}$, and $r^{(i)}_{\mathcal{L}_h} \gg r^{(i')}_{\mathcal{L}_h}$, such that $e^{-r^{(i')}_{\mathcal{L}_h} \bar{t}} > 2e^{-r^{(i)}_{\mathcal{L}_h} \bar{t}}$ for a big $\bar{t}$, such that $e^{-r^{(i)}_{\mathcal{L}_h} \bar{t}} \approx 0$.

Denotes $a = e^{-r^{(i')}_{\mathcal{L}_h} \bar{t}} - e^{-r^{(i)}_{\mathcal{L}_h} \bar{t}}$, so, $a > e^{-r^{(i)}_{\mathcal{L}_h} \bar{t}} > 0$, $a \in (0, 1)$, and, $e^{-r^{(i')}_{\mathcal{L}_h} \bar{t}} = a + e^{-r^{(i)}_{\mathcal{L}_h} \bar{t}}$.

It is convenient for the proof to assume $ratio^{(i)}_{\mathcal{L}_h}(t) = ratio^{(i)}_{\mathcal{L}_d}(t)$, we will discuss the other cases later.

According to Eq. (1),

$$\begin{aligned} \mathcal{L}_d(\bar{t}) &= \left\| \frac{d^{(1)}(\mathbf{z_1}, \mathbf{z_2})}{d^{(n)}(G^{(n)}(\mathbf{z_1}), G^{(n)}(\mathbf{z_2}))} \right\|_1 = \prod_{j=2}^{n} \left\| ratio^{(j)}_{\mathcal{L}_d}(\bar{t}) \right\|_1 \\ &= (a + e^{-r^{(i)}_{\mathcal{L}_h} \bar{t}})^{n-2} e^{-r^{(i)}_{\mathcal{L}_h} \bar{t}} \approx 0 \approx e^{-r^{(i)}_{\mathcal{L}_h} \bar{t}} \end{aligned} \tag{4}$$

According to Eq. (2), when $\lambda_i = 0$,

$$\begin{aligned} \mathcal{L}_h(\bar{t}) &= \sum_{j=2}^{n} \left\| ratio^{(j)}_{\mathcal{L}_h}(t)(\bar{t}) \right\|_1 \\ &= (n-2)(a + e^{-r^{(i)}_{\mathcal{L}_h} \bar{t}}) + e^{-r^{(i)}_{\mathcal{L}_h} \bar{t}} \\ &> (n-2)a > 0 \end{aligned} \tag{5}$$

Denotes $H = (n-2)a$, so, $H \in (0, +\infty)$, and, at time $\bar{t}$, $\mathcal{L}_d(\bar{t}) < H$, the corresponding training process has been stopped, suppose it stopped at time t_d, $t_d \leq \bar{t}$; and at time $\bar{t}$, $\mathcal{L}_h(\bar{t}) > H$, the corresponding training process will be stopped at time t_h, $t_h > \bar{t} > t_d$; therefore, according to the monotonicity of l, $l(t_h) < l(t_d)$.

If $ratio^{(i)}_{\mathcal{L}_h}(t) > ratio^{(i)}_{\mathcal{L}_d}(t)$, namely, $r^{(i)}_{\mathcal{L}_h} < r^{(i)}_{\mathcal{L}_d}$, denotes the new stop time as $t_d^{>}$ and the new regularization term as $\mathcal{L}_d^{>}$, so, when $\mathcal{L}_d(t_d) = \mathcal{L}_d^{>}(t_d^{>})$, according to Eq. (4), $e^{-r^{(i)}_{\mathcal{L}_h} t_d} = e^{-r^{(i)}_{\mathcal{L}_d} t_d^{>}}$, $t_d^{>} = (r^{(i)}_{\mathcal{L}_h} / r^{(i)}_{\mathcal{L}_d}) t_d < t_d$, so, $t_h > t_d > t_d^{>}$; therefore, $l(t_h) < l(t_d^{>})$.

If $ratio^{(i)}_{\mathcal{L}_h}(t) < ratio^{(i)}_{\mathcal{L}_d}(t)$, according to the condition in the proposition, assume $ratio^{(i')}_{\mathcal{L}_h}(t) \approx ratio^{(i')}_{\mathcal{L}_d}(t)$, namely, $r^{(i')}_{\mathcal{L}_d} \approx r^{(i')}_{\mathcal{L}_h}$, and $r^{(i)}_{\mathcal{L}_d} \gg r^{(i')}_{\mathcal{L}_d}$, such that $e^{-r^{(i')}_{\mathcal{L}_d} \bar{t}^{<}} > 2e^{-r^{(i)}_{\mathcal{L}_d} \bar{t}^{<}}$ for a big $\bar{t}^{<}$, such that $e^{-r^{(i)}_{\mathcal{L}_d} \bar{t}^{<}} \approx 0$, and Eq. (4) and Eq. (5) are still satisfied when $t = \bar{t}^{<}$, the only difference between the old equations and the new equations is that a in the new Eq. (4) and Eq. (5) is not the same, for clarity, denotes a in the new Eq. (4) and Eq. (5) as a_4 and

a_5, $a_4 = e^{-r_{\mathcal{L}_d}^{(i')}\bar{t}^<} - e^{-r_{\mathcal{L}_d}^{(i)}\bar{t}^<}$, $a_5 = e^{-r_{\mathcal{L}_h}^{(i')}\bar{t}^<} - e^{-r_{\mathcal{L}_h}^{(i)}\bar{t}^<}$; because $e^{-r_{\mathcal{L}_d}^{(i)}\bar{t}^<} > e^{-r_{\mathcal{L}_h}^{(i)}\bar{t}^<}$ and $r_{\mathcal{L}_d}^{(i')} \approx r_{\mathcal{L}_h}^{(i')}$, $a_5 > a_4$; similarly, we can choose $H^< = (n-2)a_5$ and $t_h^<, t_d^<$ as stop time, then, the same conclusion can be deduced when $ratio_i^{\mathcal{L}_h}(t) = ratio_i^{\mathcal{L}_d}(t)$, namely, $l(t_h^<) < l(t_d^<)$.

■

Proposition 1 shows that, when there exists a layer whose $ratio^{(i)}$ converges much faster than other layers', $\mathcal{L}_d$ converges more quickly than $\mathcal{L}_h$, which means the training process supervised by $\mathcal{L}_d$ stops earlier than the one supervised by $\mathcal{L}_h$, even though their cost are the same in any form of $\mathcal{L}_h$ or $\mathcal{L}_d$.

Since we use $\lambda^{(i)}$ to independently control $ratio^{(i)}$ at each layer, we need not assign different weights to them, therefore, the computation cost of searching the only weight of $\mathcal{L}_h$ is $\mathcal{O}(n)$. However, if we use $\mathcal{L}_d$ as the term to adjust the change of distance through specific layer of G, the computation cost for searching the weights of terms grows exponentially with n.

Expected Ratio of Expansion

In Eq. (3), $\lambda^{(i)}$ controls the diversity of the output. For example, when $\lambda^{(i)} = 1$, the i-th convolutional layer is encouraged to not change the distance; when $\lambda^{(i)} = 0$ for $\forall i \in \{2, ..., n\}$, every layer of the generator is to maximize the distance between the output images. $\boldsymbol{\lambda}$ is set as the target ratio of the expansion, we call it the Expected Ratio of Expansion (ERE) in this work.

In practice, it is important to determine the value of $\lambda^{(i)}$, since when $\lambda^{(i)}$ is larger than 1, the diversity is encouraged not to be increased, and because the distance cannot be $+\infty$, we cannot increase the diversity by setting $\lambda^{(i)}$ lower than its lower bound. Therefore, we restrict $\lambda^{(i)} \in [b^{(i)}, 1]$.

To compute $b^{(i)}$, there are two steps, firstly, we pre-trained the cGANs using:

$$\mathcal{L}_{fin} = \mathcal{L}_{ori} + \beta\mathcal{L}_h^0 \tag{6}$$

Where, β is the weight to manipulate the importance of the regularization. $\mathcal{L}_h^0$ denotes the $\mathcal{L}_h$ with all $\lambda^{(i)} = 0$, and $\mathcal{L}_{ori}$ is the objective function used in the cGANs framework into which we integrated the proposed method. Then, we fed the whole dataset X into the generator to calculate the ratio matrix:

$$A^{(i)} = \begin{bmatrix} d_{11} & \dots & d_{1m} \\ \vdots & \ddots & \vdots \\ d_{m1} & \dots & d_{mm} \end{bmatrix} \tag{7}$$

Here, m is the size of X, $d_{uv} = \frac{d^{(i-1)}(G^{(i-1)}(\mathbf{z_u}), G^{(i-1)}(\mathbf{z_v}))}{d^{(i)}(G^{(i)}(\mathbf{z_u}), G^{(i)}(\mathbf{z_v}))}$. If we choose L_1 norm as $d^{(i)}(\cdot)$ for $\forall i$, then $A^{(i)}$ can be calculated by:

$$(A^{(i)})^u{}_v = \frac{|o^u M_{v,p} - o^v M_{u,p}|}{|o^u N_{v,q} - o^v N_{u,q}|} \tag{8}$$

Here, M and N are the $1 \times m \times f^{(i)}$ matrices output by $(i-1)$-th layer and i-th layer, respectively, $f^{(i)}$ is the dimension of output by $G^{(i)}(\cdot)$, o is $\vec{1}_{m\times 1}$.

Then, $b^{(i)}$ is determined as the minimum element of $A^{(i)}$, $b^{(i)} = min(d_{11}, ..., d_{jk}, ..., d_{mm})$. Since it requires 2 loops to calculate all d_{uv} to determine $A^{(i)}$, the time complexity of naively computing $b^{(i)}$ is $O(m^2)$. Eq. (8) provides a way to compute $A^{(i)}$ in the form of tensors to compute $b^{(i)}$ with complexity $O(m)$, because tensor operation of a batch can be executed on GPU in parallel.

Experiments

To validate our regularization method under an extensive evaluation, we incorporated four baseline models (DCGAN (Yu et al. 2017b), Pix2Pix (Isola et al. 2017), DRIT (Lee et al. 2018) and StackGAN++ (Zhang et al. 2017a)) with it for three conditional image synthesis tasks:

- Categorical generation, it is trained on CIFAR-10 (Szegedy et al. 2015) using DCGAN as the baseline model.
- Image-to-image translation, it can be divided into two subtasks:
  - Paired image-to-image translation, it is trained on facades and maps using Pix2Pix as the baseline model.
  - Unpaired image-to-image translation, it is trained on Yosemite (Zhu et al. 2017a) and cat⇌dog (Lee et al. 2018) using DRIT as the baseline model.
- Text-to-image generation, it is trained on CUB-200-2011 (Wah et al. 2011) using StackGAN++ as the baseline model.

Because the original networks of the baseline model do not change after adding the attention unit and the regularization term, we kept the hyper-parameters of the baseline model original.

We adopted L_1 norm as distance metrics for all $d^{(i)}(\cdot)$ and set the weight of regularization $\beta = 1$ in all experiments.

Evaluation metrics

To evaluate the quality of the generated images, we used FID (Heusel et al. 2017) to measure the difference between the distribution of generated images and the distribution of real images. To compute FID, a pretrained Inception Network (Szegedy et al. 2015) needed for extracting features of images. Lower FID indicate higher quality of the generated images.

To evaluate diversity, we employed LPIPS (Zhang et al. 2018).

$$d(\mathbf{x_1}, \mathbf{x_2}) = \sum_l \frac{1}{H^{(l)}W^{(l)}} \sum_{h,w} \left\| w^{(l)} \odot (E_{hw}^{(l)}(\mathbf{x_1}) - E_{hw}^{(l)}(\mathbf{x_2}) \right\|_2^2 \tag{9}$$

$$diversity_{output} = \sum_{j=1}^{m} \sum_{k=1, k\neq j}^{m} d(\mathbf{x_j}, \mathbf{x_k}) \tag{10}$$

Because deeper convolutional layers detect higher-level features (Zeiler and Fergus 2014), it is natural to measure the diversity w.r.t specific-level feature with a specific-l-th

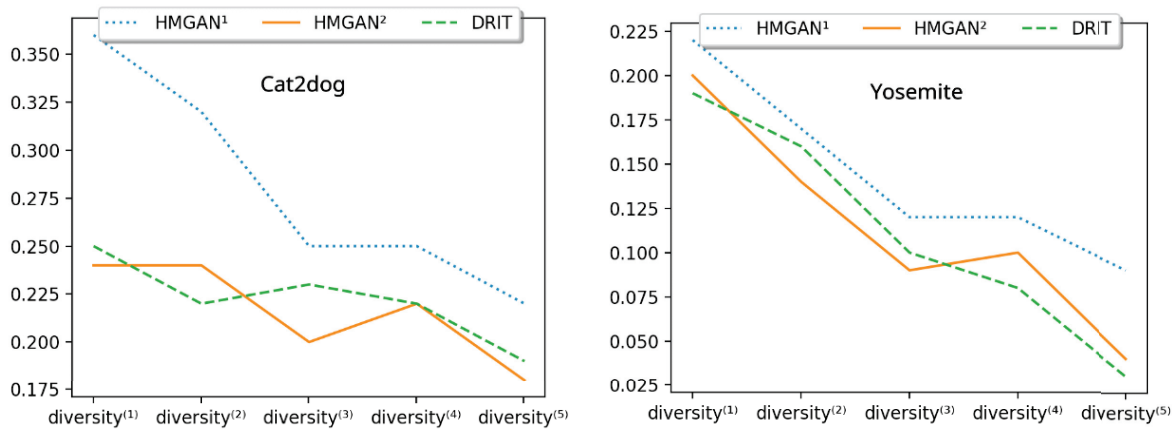

Figure 1: Visualizing diversity of image batch w.r.t different-level features. HMGAN^1 refers $\lambda^{(i)} = 0$ for $\forall i$, and HMGAN^2 refers $\lambda^{(i)} = 1$. We measured the diversity with LPIPS which uses Alexnet as feature decoder, it provides outputs from 5 layers.

Table 1: NDB and JSD results on the CIFAR-10 dataset.

Metrics	Models	airplane	automobile	bird	cat	deer
NDB ↓	DCGAN	49.60 ± 3.50	53.30 ± 6.34	34.30 ± 5.71	46.00 ± 2.65	43.90 ± 4.12
	HMGAN^1	$\mathbf{45.30 \pm 5.24}$	$\mathbf{51.50 \pm 3.13}$	$\mathbf{33.20 \pm 2.02}$	$\mathbf{42.00 \pm 1.47}$	$\mathbf{42.20 \pm 4.37}$
	HMGAN^2	$\mathbf{48.70 \pm 5.13}$	$\mathbf{52.90 \pm 2.98}$	$\mathbf{34.00 \pm 2.38}$	$\mathbf{45.50 \pm 2.12}$	$\mathbf{43.50 \pm 4.14}$
JS ↓	DCGAN	0.035 ± 0.002	0.035 ± 0.002	0.026 ± 0.002	0.031 ± 0.001	0.033 ± 0.002
	HMGAN^1	$\mathbf{0.028 \pm 0.002}$	$\mathbf{0.029 \pm 0.002}$	$\mathbf{0.024 \pm 0.001}$	$\mathbf{0.026 \pm 0.001}$	$\mathbf{0.029 \pm 0.002}$
	HMGAN^2	$\mathbf{0.033 \pm 0.00}2$	$\mathbf{0.034 \pm 0.001}$	$\mathbf{0.026 \pm 0.001}$	$\mathbf{0.029 \pm 0.001}$	$\mathbf{0.032 \pm 0.002}$
		dog	frog	horse	ship	truck
NDB ↓	DCGAN	51.80 ± 3.92	53.20 ± 4.27	55.00 ± 2.81	43.50 ± 5.00	45.50 ± 5.05
	HMGAN^1	$\mathbf{34.00 \pm 2.80}$	$\mathbf{41.60 \pm 3.55}$	$\mathbf{46.50 \pm 5.83}$	$\mathbf{41.50 \pm 3.04}$	$\mathbf{43.20 \pm 3.01}$
	HMGAN^2	52.00 ± 3.02	$\mathbf{52.10 \pm 3.22}$	$\mathbf{53.90 \pm 4.32}$	$\mathbf{42.80 \pm 3.05}$	$\mathbf{45.30 \pm 5.00}$
JS ↓	DCGAN	0.035 ± 0.002	0.035 ± 0.002	0.036 ± 0.001	0.030 ± 0.002	0.034 ± 0.002
	HMGAN^1	$\mathbf{0.024 \pm 0.001}$	$\mathbf{0.029 \pm 0.001}$	$\mathbf{0.032 \pm 0.002}$	$\mathbf{0.027 \pm 0.002}$	$\mathbf{0.028 \pm 0.002}$
	HMGAN^2	0.035 ± 0.002	$\mathbf{0.034 \pm 0.002}$	$\mathbf{0.031 \pm 0.002}$	$\mathbf{0.028 \pm 0.002}$	0.034 ± 0.002

term in Eq. (9),

$$d^{(l)}(\mathbf{x_1}, \mathbf{x_2}) = \frac{1}{H^{(l)}W^{(l)}} \sum_{h,w} \left\| w^{(l)} \odot (E^{(l)}_{hw}(\mathbf{x_1}) - E^{(l)}_{hw}(\mathbf{x_2}) \right\|_2^2 \tag{11}$$

The $diversity^{(l)}$ is similarly computed as Eq. (10) does. To statistically view the diversity of an image batch w.r.t different-level features, we visualized all $diversity^{(l)}$, as shown in figure 1. Higher LPIPS means the generated images are more diverse.

To test the generated images and the real images are from the same distribution, we employed the NDB score (Richardson and Weiss 2018). To compute NDB score, it first put all real and generated samples into bins, then, the numbers of the real images and the generated images in one bin are used to decide if those two numbers are statistically different, finally, the number of all statistically different bins defines the NDB score. The bins are the result from a K-means clustering. In other words, the K-means clustering finds k modes, so, we can not only estimate the similarity between two distributions by comparing the NDB scores but can also tell which mode has collapsed by referring the indices of statistically different bins. However, there is a trade-off between a less number of bins (less computation for the clustering) and a higher accuracy of the estimation, we presented the Jensen–Shannon divergence to validate the NDB scores, and to find a proper number of bins during the experiment. Lower NDB and JSD mean the generated images are more likely from the real distribution.

Table 2: FID and LPIPS results on the CIFAR-10 dataset.

Model	DCGAN	HMGAN
FID ↓	32.21 ± 0.05	$\mathbf{28.84 \pm 0.05}$
LPIPS ↑	0.208 ± 0.002	$\mathbf{0.209 \pm 0.002}$

Categorical generation

Firstly, we validated the regularization method on categorical generation task. In categorical generation, the generator takes the initial noise concatenated with class labels as input

Table 3: Quantitative results from paired image-to-image translation task.

Datasets	Facades		
	Pix2Pix	HMGAN^1	HMGAN^2
FID ↓	140.00 ± 2.57	$\mathbf{90.00 \pm 3.25}$	$\mathbf{138.80 \pm 2.00}$
NDB ↓	16.00 ± 0.38	$\mathbf{12.30 \pm 0.32}$	16.12 ± 0.59
JSD ↓	0.078 ± 0.003	$\mathbf{0.028 \pm 0.006}$	0.080 ± 0.004
LPIPS ↑	0.005 ± 0.001	$\mathbf{0.192 \pm 0.001}$	$\mathbf{0.007 \pm 0.001}$
Datasets	Maps		
	Pix2Pix	HMGAN^1	HMGAN^2
FID ↓	165.80 ± 3.21	$\mathbf{153.60 \pm 2.50}$	$\mathbf{164.50 \pm 2.39}$
NDB ↓	$47.30 \pm 2,35$	$\mathbf{42.00 \pm 2.52}$	$\mathbf{46.80 \pm 3.52}$
JSD ↓	0.072 ± 0.023	$\mathbf{0.035 \pm 0.003}$	0.076 ± 0.025
LPIPS ↑	0.003 ± 0.001	$\mathbf{0.205 \pm 0.001}$	0.003 ± 0.001

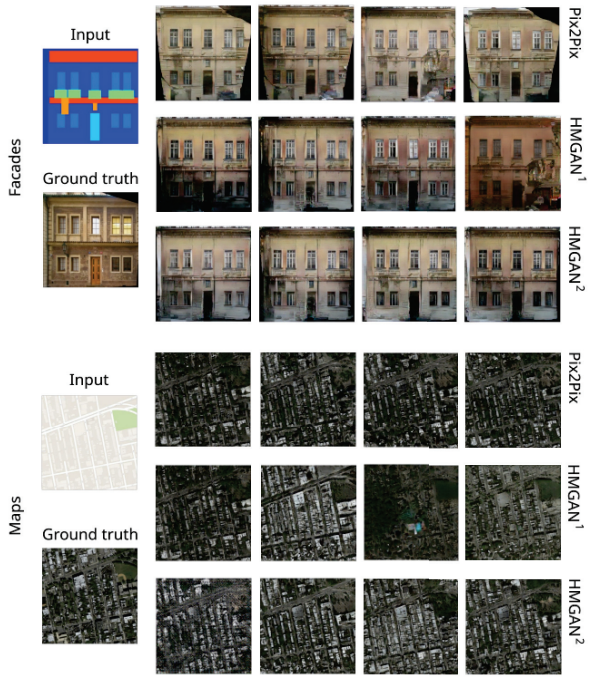

Figure 2: Diversity comparison. HMGAN^1 learns more diverse results and HMGAN^2 learns less diverse results.

to generate images in corresponding categories. This task is conducted on the CIFAR-10 dataset. It has images with size 32×32 in 10 categories. The NDB scores and JSD are reported in Table 1, and Table 2 presents the results FID and LPIPS. The proposed method alleviates the mode collapse problem in most categories and maintains the image quality.

Image-to-image translation

Conditioned on paired images

In this task, we integrated the proposed method into Pix2Pix. In experiments, we kept the original hyper-parameters setting of Pix2Pix for fair comparison. Figure 2 and Table 3 shows the qualitative and quantitative results, respectively. It is shown that the proposed method exceeds Pix2Pix in terms of all metrics when $\lambda^{(i)} = 0$ for $\forall i$, and the output images from the proposed method have comparable diversity to the ones from Pix2Pix when $\lambda^{(i)} = 1$ for $\forall i$. The low quality of generated facades images might be caused by encouraging diversity too much (Yang et al. 2019), since setting $\lambda^{(i)} = 0$ for $\forall i$ regularizes the training more strictly than minimizing $\mathcal{L}_d$.

Conditioned on unpaired images

To generate images when paired images are not available, we chose DRIT as the baseline model. It is pointed out that DRIT can generate diverse images only w.r.t low-level features, in other words, the output images share similar structures. To demonstrate the proposed method can improve the diversity w.r.t high-level features, we conducted this experiment on cat⇌dog dataset whose images has shape variations. We also compared the abilities of generating diverse image w.r.t low-level features between the proposed method and DRIT, this experiment is conducted on shape-invariant Yosemite dataset.

Table 4 shows that the proposed method outperforms DRIT in terms of all metric in both experiments, especially on the cat⇌dog dataset. To quantitively present the difference of ability to generate diverse images w.r.t different-level features, we ploted all $diversity^{(l)}$ in figure 1. Figure 1. shows that our proposed method improved the diversity w.r.t high-level features, and has comparable ability to generate diverse images w.r.t low-level features.

Text-to-image generation

StackGAN++ is proposed to generate diverse images whose contents is corresponding to given descriptive sentences. We chose it as the baseline model in this task, and the task is conducted on the CUB- 200-2011 dataset.

Table 5 presents quantitative comparisons between the proposed method and StackGAN++. And the qualitative results are shown in figure 3, it shows that the proposed method improve the diversity without losing visual quality.

Table 4: Quantitative results from unpaired image-to-image translation task.

Datasets	Summer2Winter		Winter2Summer	
	DRIT	HMGAN[1]	DRIT	HMGAN[1]
FID ↓	55.03 ± 3.26	$\mathbf{50.00 \pm 3.23}$	47.00 ± 4.28	$\mathbf{46.20 \pm 3.38}$
NDB ↓	25.50 ± 3.35	$\mathbf{23.00 \pm 0.25}$	29.00 ± 2.47	$\mathbf{27.50 \pm 2.55}$
JSD ↓	0.062 ± 0.003	$\mathbf{0.052 \pm 0.003}$	0.050 ± 0.007	$\mathbf{0.038 \pm 0.005}$
LPIPS ↑	0.112 ± 0.001	$\mathbf{0.143 \pm 0.001}$	0.112 ± 0.001	$\mathbf{0.119 \pm 0.001}$
Datasets	Cat2Dog		Dog2Cat	
	DRIT	HMGAN[1]	DRIT	HMGAN[1]
FID ↓	22.50 ± 0.35	$\mathbf{16.02 \pm 0.35}$	59.05 ± 0.31	$\mathbf{28.97 \pm 0.54}$
NDB ↓	39.28 ± 3.36	$\mathbf{27.00 \pm 0.50}$	41.32 ± 0.52	$\mathbf{32.23 \pm 0.53}$
JSD ↓	0.125 ± 0.003	$\mathbf{0.085 \pm 0.001}$	0.269 ± 0.002	$\mathbf{0.071 \pm 0.001}$
LPIPS ↑	0.250 ± 0.002	$\mathbf{0.280 \pm 0.002}$	0.100 ± 0.002	$\mathbf{0.220 \pm 0.003}$

Table 5: Quantitative results from text-to-image generation task. HMGAN[3] refers $\lambda^{(i)} = 0.5$ for $\forall i$.

	StackGAN++	HMGAN[1]	HMGAN[2]	HMGAN[3]
FID ↓	26.00 ± 4.23	$\mathbf{25.40 \pm 2.00}$	27.00 ± 1.25	$\mathbf{25.55 \pm 1.50}$
NDB ↓	37.80 ± 2.44	$\mathbf{29.90 \pm 2.55}$	$\mathbf{37.55 \pm 1.83}$	$\mathbf{30.00 \pm 3.82}$
JSD ↓	0.091 ± 0.005	$\mathbf{0.070 \pm 0.005}$	0.093 ± 0.005	$\mathbf{0.072 \pm 0.003}$
LPIPS ↑	0.364 ± 0.005	$\mathbf{0.376 \pm 0.005}$	0.358 ± 0.005	$\mathbf{0.374 \pm 0.002}$

Controlling diversity

To control diversity w.r.t specific level features, we chose $\lambda^{(5)}$ as the control variable in this experiment. We firstly computed the lower bound of $\lambda^{(5)}$ by choosing the minimum element in $A^{(5)}$ computed by Eq. (8). Figure 5 shows the results in text-to-image synthesis task, we can see $diversity^{(5)}$ is bigger with smaller $\lambda^{(5)}$, and reaches the limit when $\lambda^{(5)} < b^{(5)}$. Figure 1 and figure 4 show that, in image translation and text-to-image generation, our method can generate outputs with different distributions of $diversity^{(l)}$, which is unachievable to the previous method. We also noticed that, in figure 3, when $\lambda^{(j)} = 1, \lambda^{(k)} = 0$, the proposed method tends to change the observation angle to the bird, or to change the posture of the bird.

Supplementary results

We also conducted the three conditional image synthesis tasks using Eq. (1). The results show that our method outperforms the one using Eq. (1) in tasks of categorical generation (85% of the results are better), paired image-to-image generation (50% of the results are better), unpaired image-to-image generation (69% of the results are better) and text-to-image generation (100% of the results are better).

Related Work

Unlike standard GANs only require an initial noise as input for the generator, cGANs concatenates external information (e.g. , the number of age) with the initial noise, during training, the correspondence between perceptual features (e.g., wrinkles of a face) and the additional information can be learned, as a result, an image with specific feature can be synthesized by a generator conditioned on the external information. However, it does not only inherit the mode collapse problem in standard GANs, but also worsen it when the input has a high-dimension information part (Yang et al. 2019) . And it is pointed out that the noise vector is responsible for generating various images, due to its comparative low dimension, it is often ignored by the generator (Mao et al. 2019). More specifically, because the input pair has the same external information, once it is propagated through a convolutional layer, the distance between an output pair is smaller, especially when the external information has a high dimension. In this situation, two modes are prone to be collapsed into one if their initial noises are close.

To alleviate mode collapse in cGANs, some approaches are proposed by recent works. In text-to-image tasks, (Zhang et al. 2017b) uses a fully connected layer to sample additional noise from a Gaussian, the noise is then combined with the feature of an image as a whole conditional context to obtain more training pairs for augmentation. A different approach proposes an extra encoder network which can generate the noise vector given the generated image to help the generator construct a one-to-one mapping between the input and output, this approach was employed in image-to-image translation (Zhu et al. 2017b). However, the two approaches require extra time to generate augmentation pairs or to train an additional encoder, not to mention they are substantial task-specific modifications to cGANs, that is to say, they are less generalizable and charge more computational resource. Recently, (Mao et al. 2019) and (Yang et al. 2019) propose a regularization method (Diversity-sensitive cGANs) to amplify the diversity of the outputs, more specifically, the regularization term encourages the generator to maximize the ratio of the distance between a noise pair to the distance be-

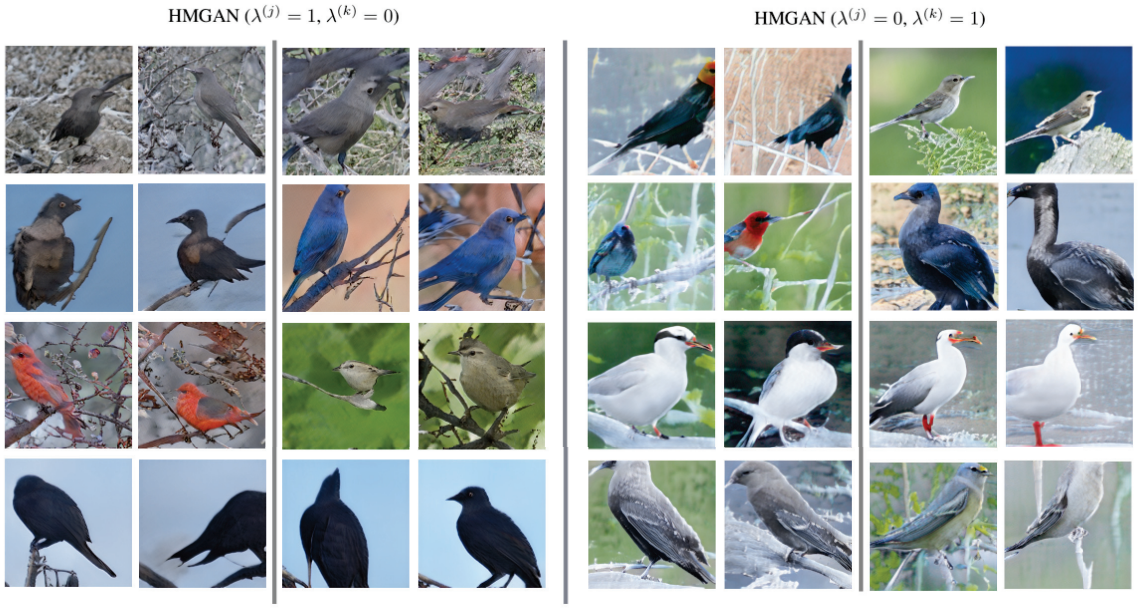

Figure 3: Diversity comparison. Each pair is conditioned by the same sentence. Since StackGAN++ has 15 convolutional layers, the 6 15th layers are designed to improved the resolution of the 5th layer's output, the diversity w.r.t high-level features is controlled by 4th and 5th layer. Here, $j \in \{1, 2, 3, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15\}$ and $k \in \{4, 5\}$.

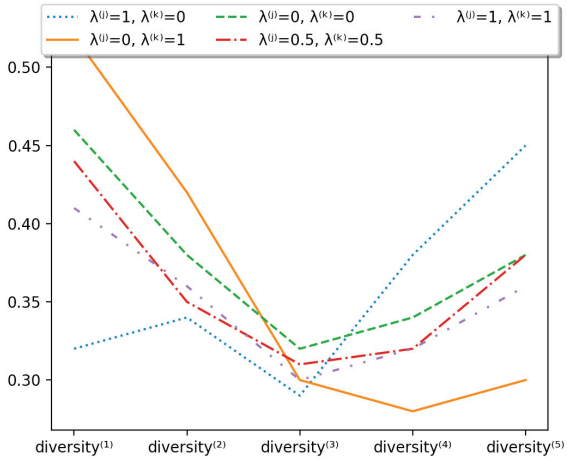

Figure 4: Visualizing diversity of image batch w.r.t different-level features in text-to-image task.

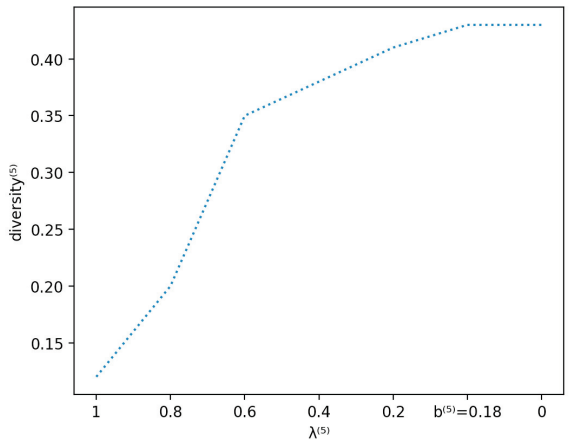

Figure 5: Controlling the diversity w.r.t specific-level features by tuning one term of ERE.

tween an image pair. The method needs no training overheads and can be easily extended to other frameworks of cGANs. But it ignores the diversity in hierarchical feature spaces, one of the results is that Diversity-sensitive GANs can generate the images of bird with various posture but is not able to synthesize different feather textures.

Conclusion

In this work, we applied a regularization term on the generator to address the mode collapse problem in cGANs. And we avoided the computation cost for searching a hyperparameter growing exponentially with the number of layers of the generator. We minimized the differences between the real change of feature distance and a target change at all convolutional layers of the generator to control diversities w.r.t specific-level features. The proposed regularization term could be integrated into the existing different frame-

works of cGANs. Our method is demonstrated on three image generation tasks and experimental results showed that our regularization can increase the diversity without decreasing visual quality. As a future work, we will add more convolutional layers in the generator and validate how to control diversities more precisely. We also hope to conduct more experiments to find the dependencies of $ratio^{(i)}$.

Acknowledgement

Jinlong Li is supported by the National Key Research and Development Program of China (Grant No. 2017YFC0804001) and the National Natural Science Foundation of China (Grant No. 61573328).

References

Doan, T.; Monteiro, J.; Albuquerque, I.; Mazoure, B.; Durand, A.; Pineau, J.; and Hjelm, R. D. 2019. Online adaptative curriculum learning for gans. In *Thirty-Third AAAI Conference on Artificial Intelligence*.

Ge, Y.; Li, Z.; Zhao, H.; Yin, G.; Yi, S.; Wang, X.; et al. 2018. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *Advances in Neural Information Processing Systems*, 1222–1233.

Goodfellow, I. J. 2016. Nips 2016 tutorial: Generative adversarial networks. *CoRR* abs/1701.00160.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 6626–6637.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Lee, H.-Y.; Tseng, H.-Y.; Huang, J.-B.; Singh, M.; and Yang, M.-H. 2018. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 35–51.

Lin, Z.; Khetan, A.; Fanti, G.; and Oh, S. 2018. Pacgan: The power of two samples in generative adversarial networks. In *Advances in Neural Information Processing Systems*, 1498–1507.

Mao, Q.; Lee, H.-Y.; Tseng, H.-Y.; Ma, S.; and Yang, M.-H. 2019. Mode seeking generative adversarial networks for diverse image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Mirza, M., and Osindero, S. 2014. Conditional generative adversarial networks. *Manuscript: https://arxiv.org/abs/1709.02023* 9:24.

Richardson, E., and Weiss, Y. 2018. On gans and gmms. In *Advances in Neural Information Processing Systems*, 5847–5858.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Yang, D.; Hong, S.; Jang, Y.; Zhao, T.; and Lee, H. 2019. Diversity-sensitive conditional generative adversarial networks. In *International Conference on Learning Representations*.

Yu, L.; Zhang, W.; Wang, J.; and Yu, Y. 2017a. Seqgan: Sequence generative adversarial nets with policy gradient. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Yu, Y.; Gong, Z.; Zhong, P.; and Shan, J. 2017b. Unsupervised representation learning with deep convolutional neural network for remote sensing images. In *International Conference on Image and Graphics*, 97–108. Springer.

Zeiler, M. D., and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*, 818–833. Springer.

Zhang, C., and Peng, Y. 2018. Visual data synthesis via gan for zero-shot video classification. In *IJCAI*.

Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; and Metaxas, D. 2017a. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PP.

Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; and Metaxas, D. N. 2017b. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 5907–5915.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 586–595.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017a. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.

Zhu, J.-Y.; Zhang, R.; Pathak, D.; Darrell, T.; Efros, A. A.; Wang, O.; and Shechtman, E. 2017b. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*, 465–476.