ICLR REPRODUCIBILITY OF SNGAN

Haoen Feng, Haobo Xu, Wenhao Li
Department Electrical Engineering and Computer Science
University of Southampton
Southampton
UK

ABSTRACT

This report focuses on the theories and implementation of Spectral Normalization for Generative Adversarial Networks (SNGAN) for stabilizing the gradient update for the generator and the discriminator. The main idea of SNGAN is derived from the spectral normalization, which is widely used in regularization in loss function of the neural networks. The report explore the the foundations of SNGAN by analyzing the function of spectral normalization, and discuss its performance with the comparison of different improved Generative Adversarial Networks (GAN).

1 INTRODUCTION

The purpose of the generator is to create samples having the same probability distribution as that in the training data. The discriminator is a binary classifier which is used to determine whether the input image is a real image from the training dataset or a fake sample from the generator. In addition, the discriminator provides the gradient update to guide the generator to create more convincing samples, which is the only way for the generator to optimize its parameters. Therefore, it is an adversarial game where the generator tends to create the samples that are able to deceive the discriminator while the discriminator tends to make a more accurate classification. The optimal solution for the game is to reach the Nash equilibrium for both the generator and the discriminator. However, since that the generator and the discriminator share the same gradient, the unbalance between them leads to unstable training and poor performance. To achieve further improvement, the Spectral Normalization for Generative Adversarial Networks (SNGAN) is proposed to stabilize the gradient update. Its theories and implementation are explored in the following content.

2 MATHEMATICAL ANALYSIS OF SNGAN

The article proposed that the performance of the generator significantly influences the overall GAN performance. Based on that, stabilizing the training of the generator network is beneficial to improve the GAN network. The main purpose of SNGAN is to ensure that the norm of the gradient for all parameters is less than or equal to one.Miyato et al. (2018) In neural networks, keeping the norm of the gradient in a limited range would be of benefit to the function smoothing where the best example for it is the regularization. In order to make the regularization into specific areas, SNGAN (spectral normalization GAN) introduces spectral normalization to normalize the parameters of the neural network. A feedforward neural network can be represented as a cascade calculation: $x^l = f^l(W^l x^{l-1} + b^l)$, where l represents the number of layers, x is the input of l th layer, f^l is a unlinear activation function, w and b represent the weights and biases in l th layer respectively. Firstly, for a better understanding, we make all parameters including the weights and biases as a set θ, $\theta = \{W^l, b^l\}_{l=1}^{L}$, and the whole neural network stands for a function f, $f_\theta(x^0) = x^l$. Given K sets of training datasets $(x^i, y^i)_{i=1}^{K}$, corresponding loss function is defined as $\frac{1}{K}\sum_{i=1}^{K} L(f_\theta(x_i), y_i)$, where L usually is the Euclidean distance loss or the cross-entropy loss. Since the activation function is usually a non-linear function such as ReLU, the neural network f can be seen as a non-linear function. However, if we just consider a small region of input x, we can approximate f as a linear function. In other words, we can express the function by affine-mapping, $x \rightarrow W_{\theta,x} + b_{\theta,x}$, where $W_{\theta,x}$ and $b_{\theta,x}$ are the weight and bias respectively. In order to mitigate the noise sensitivity from

the input, the Euclidean distance of adjacent data need to be small. Then, we have

$$\frac{||f_\theta(x+\xi)-f(x)||_2}{||\xi||_2}=\frac{||(W_{\theta,x}(x+\xi)+b_{\theta,x}-W_{\theta,x}x+b_{\theta,x})||_2}{||\xi||_2}=\frac{||W_{\theta,x}\xi||_2}{||\xi||_2}\leq\sigma(W_{\theta,x}) \tag{1}$$

, where $\sigma(W_{\theta,x})$ is the formula for the norm of $w_{\theta,x}$ which is mathematically equal to compute the maximum singular value of matrix $w_{\theta,x}$.Wikipedia contributors (2019) The formula for the maximum singular value is

$$\sigma(A)=\max_{\xi\in\mathbb{R},\xi\neq 0}\frac{||A\xi||_2}{||\xi||_2} \tag{2}$$

In view of the above reasoning, we should train the model parameters θ so that the norm of any x and w is small. If all activation functions are using the ReLU function f^l, from the definition of ReLU function, f^l can be approximated as a diagonal matrix $D^l_{\theta,x}$ where if the corresponding parameters in x^{l-1} is positive, the parameters in the diagonal are equal to 1. Thus, we can rewrite the formula $w_{\theta,x}$ into:

$$W_{\theta,x}=D^L_{\theta,x}W^LD^{L-1}_{\theta,x}W^{L-1}...D^1_{\theta,x}W^1 \tag{3}$$

, then for every $l\in 1,...,L$, we have $\sigma(D^l_{\theta,x})\leq 1$ and thus:

$$\sigma(W_{\theta,x})=\sigma(D^L_{\theta,x})\sigma(W^L)\sigma(D^{L-1}_{\theta,x})\sigma(W^{L-1})...\sigma(D^1_{\theta,x})\sigma(W^1)\leq\prod_{l=1}^{L}\sigma(W^l) \tag{4}$$

Therefore, it can be seen that restricting the norm of w of each layer l is able to restrict the norm of w. The SNGAN adopts many ideas from the spectral normalization to make the discriminator satisfy the Lipschitz hypothesis. The spectral normalization is to use the regularization to restrict the norm of parameters w. In comparison, SNGAN is more effectively. It only changes the maximum singular value of the weight matrix so that the original information can be significantly retained.

$$\begin{aligned}W_{SN}(W)&=W/\sigma(W)\\ \sigma(W_{\theta,x})&\leq\prod_{l=1}^{L}\sigma(W^l_{SN})\leq 1\end{aligned} \tag{5}$$

This inequality achieves the effect of limiting the Lipschitz norm of the discriminator D, which is the realization of spectral normalization.

To efficiently constrain the spectral norm of each weight matrix, a new minimization is introduced. When performing a standard gradient descent, we need to calculate the maximum singular value for each layer of W. Since that the total computational time is quite large if the singular value decomposition is performed in each iteration, Miyato introduced an indirect method which is namely power iteration to approximate the maximum singular value σ_1, u_1 and v_1.Miyato et al. (2018) As the errors in the training process are unavoidable, the σ_1 and σ_2, which are the singular values of $\sigma(W^l)$, would be not equal. Thus, we can assume that σ_1 is always different from σ_2 so that $\sigma(w^l)$ is differential and the gradient of $\sigma(w^l)^2/2$ is $d_1u_1v_i^T$ where u_1 and v_1 are left and right eigenvector matrix, respectively. In order to maximize σ_1, u_1 and v_1, we can use v_1 instead of re-initializing v_1 from the second iteration of stochastic gradient descent (SGD). The author of SNGAN found that a sufficient approximation would be obtained with only one iteration in his experiments. In addition, the weights are divided by the current SN estimate so as to normalization.

3 Implementation of SNGAN

The most important purpose for GANs is to achieve the lipschitz limit of the discriminator. Although the spectral normalization is effective, it still cannot guarantee that the gradient of θ is limited to a certain range. Based on spectral normalization, SNGAN constrains the gradient of θ to be less than or equal to 1 by normalizing the weights matrix. SNGAN proves that as long as the spectral norm of each layer is limited to 1, the target function satisfies the lipschitz limitation. Furthermore, the Lipschitz constant is the only hyperparameter in SNGAN to adjust the maximum singular value.

Although SNGAN can effectively constrain the update of the discriminator, the training is still not stable with the accuracy of the generator increases. Also, previous work has shown that the regularization of the discriminator will slow down the training of the GAN. The existing solutions to

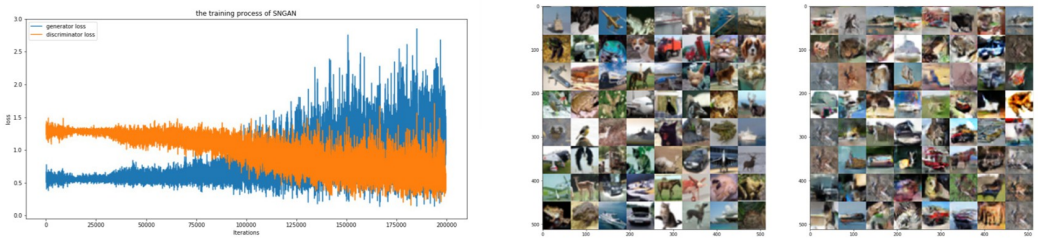

Figure 1: The training is relatively stable in the beginning stage, but the range becomes larger after 100000 iterations. The fluctuation range of the generator loss is almost twice as large as that of the discriminator. The figure shows the comparison of true images(left) and generative images(right) from SNGAN. Some of the generative images can be identified as the real objects, whcih represents great potential in SNGAN.

balance the update rates are to update the discriminator several times more than the generator. There is also an efficient way to solve the problem of unbalanced update rates. Due to the fact that the generator and the discriminator are trained together in GAN, Heusel et al. introduced two time-scale update rules (TTUR) in GAN training, and used different learning rates for generator and discriminator.Heusel et al. (2017) In the paper, the learning rate of the discriminator is four times larger than that for the generator, which is 0.004 and 0.001, respectively. A larger learning rate means that the discriminator can obtain a larger part of the gradient update. Furthermore, a higher learning rate can mitigate the problem that discriminator with regularization learns slowly. In addition, this method can also enable generator and discriminator to be updated at the same rate. In the experiment, a random vector z is input into the generator and is reshaped into 128×128 size of an RGB image. All layers including the fully connected layer use spectral normalization to constrain the value and remove the redundant gradients. The Batch-normalization and ReLU activation functions are used in the generator. The discriminator also uses spectral normalization in all layers. It takes an RGB image sample of size 128×128 as input and outputs a scale-free probability. It uses the leaky ReLU activation function with the alpha parameter set to 0.02.

4 THE PERFORMANCES OF DIFFERENT GANS

Both Inception score and FID are used to evaluate the performance of different GAns. Also, we used three different parameters to experiment with these models, and all models ran 300 epochs under these parameters and the resulting image was 28×28. The detail is shown in the table 1, where A,B,C are the hyperparameters in 3 different papers respectively. Since the epochs for training is 300, smaller than that in the original paper (above 10^5), the figures just show the trends of differnet GANs. The results show that SNGAN still performs better than the others. The inception score and FID obtained using the three parameters in the previously described GANs are as follows: A($\alpha = 0.001, \beta_1 = 0.5, \beta_2 = 0.9$), B($\alpha = 0.001, \beta_1 = 0.5, \beta_2 = 0.999$), C($\alpha = 0.002, \beta_1 = 0.5, \beta_2 = 0.999$). Full code is available at `https://github.com/COMP6248-Reproducability-Challenge/SNGAN`

REFERENCES

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a nash equilibrium. *CoRR*, abs/1706.08500, 2017. URL `http://arxiv.org/abs/1706.08500`.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *CoRR*, abs/1802.05957, 2018. URL `http://arxiv.org/abs/1802.05957`.

Wikipedia contributors. Matrix norm — Wikipedia, the free encyclopedia, 2019. URL `https://en.wikipedia.org/w/index.php?title=Matrix_norm&oldid=889037644`. [Online; accessed 16-May-2019].

Method	Inception Score	FID
Real data	9.6727	17.269
BEGAN	2.8899	193.063
EBGAN	2.2915	232.225
WGAN-GP	3.6489	186.357
BN	2.8561	243.652
LN	3.3469	219.828
SNGAN	3.8849	173.387

Table 1: The best Inception score and FID for different models in different hyperparameters

As can be seen from Table 2, the difference between the image generated by the model we used and the real data is still relatively large. We believe that an important reason is that the number of iterations is too small (we only run 300 epoch). At the same time, we can still see that SNGAN is the model closest to real data in both inception score and FID.

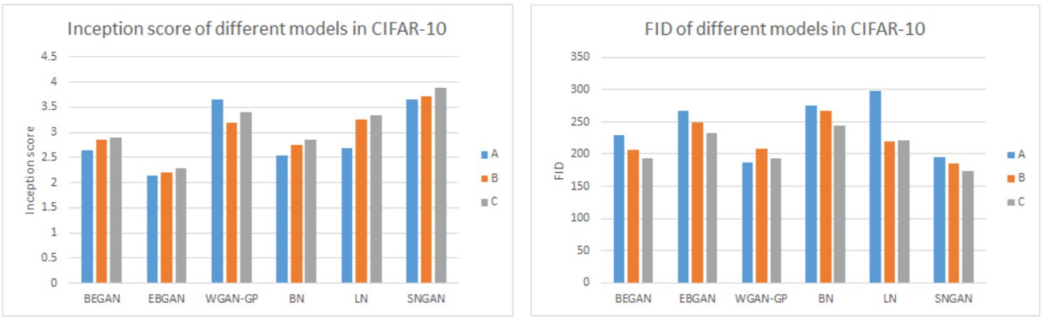

Figure 2: It can be seen that the SNGAN's comprehensive ability is higher than other models under three different parameters. Its inception score is between 3.5 and 4, while the inception scores for BEGAN, EBGAN and BN are all below 3. Especially EBGAN, the highest score is lower than 2.5. It can be seen that SNGAN is indeed better than other models.(higher is better) For FID, SNGAN is the only one with a score below 200 under different parameters, followed by WGAN-GP, which only scores greater than 200 at B. The other models are basically above 200 points, and the score of LN is almost close to 300 when it is at A. It can be seen from the FID score that the performance of SNGAN is better in these models.(lower is better))

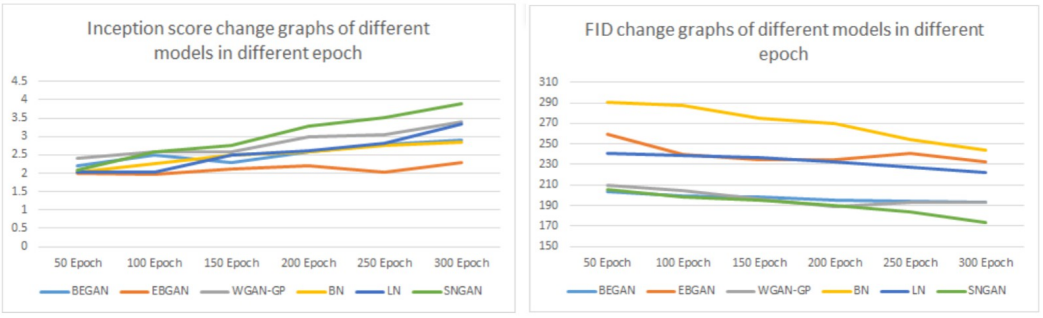

Figure 3: As the epoch increases, the inception score of each model increases, and the inception score of the 50th epoch of all models is between 2 and 2.5 The fastest rising model among these models is SNGAN. With the increase of epoch, the FID of all models has been decreasing, and it can be seen that the FID ranking of the 50th epoch is roughly the same as the ranking of the 300th epoch.