Regional attention generative adversarial network

Wei Wang, Haifeng Hu, Yi Huang, Chongchong Ruan and Dihu Chen✉

In this Letter, the authors propose a novel attention mechanism combined with a classical generative adversarial network (GAN) model to improve the visual quality of generated samples. This novel attention model is named regional attention GAN. The proposed mechanism can build dependencies between the high-level representations extracted from attention regions of real images and corresponding feature maps of the generative network. By modelling these dependencies, the generative network can be facilitated to learn feature mapping and fit the distribution of real data. They conduct extensive experiments on widely used datasets CIFAR-10, STL-10, and CelebA. The quantitative and qualitative performance improvement over state-of-the-art methods demonstrates the validity of the proposed attention mechanism in improving the quality of generated images.

Introduction: Generative adversarial networks (GANs) [1] have become well known for their superiority in image generation. One of the basic problems in image generation is the poor visual quality of the generated samples. A very significant reason is that, in high-dimensional spaces, the density ratio estimation by the discriminator is often inaccurate and unstable, as a result, it is hard for the generative network to capture some feature modes of real data distribution or fit the multi-modal structure of the target data distribution. To solve this problem, the normal practice tends to improve performance at the cost of huge computation, which means the framework of GANs model or the loss functions become more and more complex. Benefiting from the strong capability of building global dependencies, the attention model has drawn great attention on image generation task for GANs. By modelling long-range dependency between distance positions, the self-attention GANs (SAGANs) [2] improve the quality of samples a lot at the price of little increase in computation.

In this Letter, we propose a novel attention GAN model named regional attention GAN (RAGAN) to improve the visual quality of generated samples. Unlike SAGAN focusing on modelling dependencies on different locations of the feature maps, we hope to build dependencies between high-level representations extracted from the attention regions of real images and the corresponding feature maps of the generative network. To achieve this, we embed a deep attention encoder (DAE) into the basic GAN model. By using a localisation function, attention regions are predicted from the real images, and then high-level representations are extracted from these regions. By combining a weighted sum of these representations with the corresponding feature maps of the generative network, specific dependencies can be built between them, which can facilitate mapping learning that matches the data distribution of real images. To the best of our knowledge, this is the first time, we introduce an attention mechanism focusing on abstract representations of some attention regions to model dependencies for the GANs model. The proposed approach is verified to improve the visual quality of generated samples using extensive evaluations. Fig. 1 shows the basic architecture of the original GANs model and our proposed RAGAN model. In a word, our main contributions can be summarised into three: (i) a novel attention mechanism is proposed to model latent dependencies between the high-level representations of some attention regions and the corresponding feature maps of the generative network, which can help generative network learn mappings and fit the target data distribution. (ii) We introduce a novel framework named RAGAN for image generation, which embeds the new attention mechanism into the classical GANs network. (iii) Extensive experiments are conducted on widely used natural datasets to verify the validity of our proposed attention mechanism in improving the visual quality of generated samples.

Regional attention generative adversarial networks: In this section, we employ a DAE model of [3] to build regional attention to the GAN framework, enabling the generator to efficiently model latent dependencies between internal feature maps and attention regions of real images. As shown in Fig. 2, given a feature map $E(X)$ of an input image X, a designed localisation function F is used to predict the attention region locations firstly, which can be represented by

$$F(E(X)) = [x_i, y_i]_{i=1}^{N}, \tag{1}$$

where $[x_i, y_i]$ denote region's centre coordinates and N denotes the number of regions predicted. We utilise the locations of the attention regions to generate the corresponding masks $[M_i]_{i=1}^{N}$. A multiply operation is applied to mask M_i by X, thus we can obtain the desired attention regions $[R_i = X \times M_i]_{i=1}^{N}$. Next, we use a designed encoder E to extract the high-structured representations from these attention regions and get the corresponding representation maps $[A_i = E(R_i)]_{i=1}^{N}$. To establish connections of different representation maps, we sum them up using the parameters $[\lambda_i]_{i=1}^{N}$ and get a weighted attention layer $\sum_{i=1}^{N}[\lambda_i \times A_i]$, indicating the different levels of importance when the generative network outputs diverse images. Finally, we further add the attention layer back to the corresponding feature maps g of the generative network. Therefore, the final output is given by

$$O = \sum_{i=1}^{N}[\lambda_i \times A_i] + g. \tag{2}$$

As a result, the original weights of the input g can be updated according to the specific dependencies built between the attention layer and g. The $[\lambda_i]_{i=1}^{N}$ are initialised as 0, which can progressively increase the weights and complexity to stabilise the training.

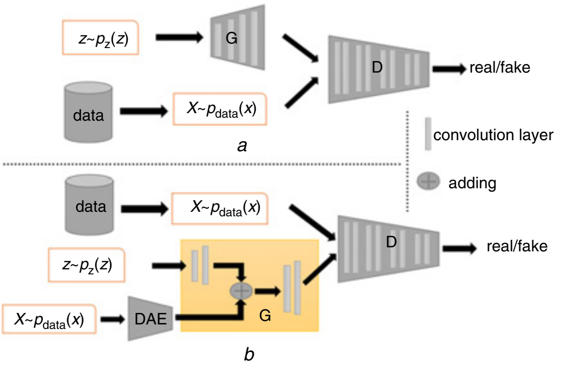

Fig. 1 *The comparison of basic architecture between original GANs and our RAGAN*

a Original GANs, including generator and discriminator.
b Our proposed RAGAN, including generator, discriminator and attention model DAE. *z* denotes input noise and *x* denotes real image

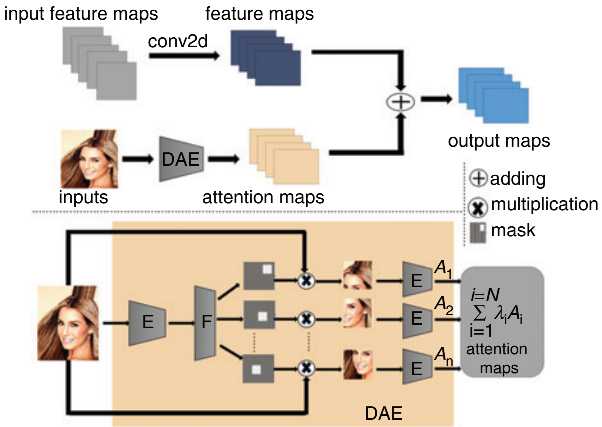

Fig. 2 *Detailed architecture of attention model DAE. Given input image X, localisation function F will first predict N attention regions' coordinates from real image of X. Then N attention masks are generated and activated on X to produce N attention regions. Finally, high-level representations are extracted from attention regions and weighted sum by parameters $[\lambda_i]_{i=1}^{N}$, where E is encoder that can be designed to coordinate with outputs you want*

Experiments: In the following, experiments are carried out to investigate the validity of our proposed attention method on the image datasets from natural scenes. We first present experimental results using specific values to analyse the effectiveness of our proposed attention mechanism on improving the visual quality of generated images quantitatively. Then, we show some randomly generated samples to give a qualitative analysis.

Datasets: Three widely-adopted datasets are used: CIFAR-10, STL-10, and CelebA. In our experiments, the STL-10 images are resized from 96 × 96 to 32 × 32. The CelebA images are centre cropped to 64 × 64 and 128 × 128. *Evaluation metric*: The inception score (IS) is widely used as an assessment for sample quality from a labelled dataset and Fréchet inception distance (FID) is often employed for quantitative evaluation on image quality by measuring the distance of different data distributions. The higher the IS is or the lower the FID is, the better the visual quality the samples have. We show the best values after 200 K iterations training on CIFAR-10 and STL-10 datasets, and after 20 epochs training on CelebA dataset. *Baselines*: Deep convolutional GAN (DCGAN) [4] is the first model that uses fully convolutional networks to implement the basic GANs model stably. Wasserstein GAN (WGAN) [5] uses Wasserstein distance instead of traditional Kullback–Leibler divergence to measure the distance of different data distributions. Recently, SAGAN [2] successfully applied a new attention mechanism to the basic GANs model. All of them achieve great success in image generation. *Network structures*: The generator and discriminator we used in our model originated from the DCGAN-based network [4] and the attention model is only applied in the generative network in all experiments. The loss function we employed is a traditional DCGAN [4] adversarial loss or WGAN [5] adversarial loss.

Firstly, we report the IS and FID results we obtained, which are shown in Table 1. From the results, we can see that the basic GAN models can achieve better results by adding our proposed attention model. The comparison between our attention model and self-attention model [2] validates that our method can also achieve better performance, which verifies the great effect it has on improving the quality of generated samples. We also show some training curves in Fig. 3 to display the training details.

Table 1: IS and FID

Evaluation metric	IS		FID	
Dataset	CIFAR-10	STL-10	CIFAR-10	CelebA
Real data	11.24 ± 0.16	26.08 ± 0.26	0	0
baseline DCGAN [4]	7.25 ± 0.09	7.86 ± 0.11	28.29	14.04
+self-attention [2]	7.35 ± 0.04	7.88 ± 0.03	27.51	12.34
+attention (ours)	*7.52 ± 0.05*	*8.10 ± 0.01*	*25.43*	*11.74*
baseline WGAN [5]	7.27 ± 0.06	7.88 ± 0.11	26.15	14.87
+self-attention [2]	7.34 ± 0.1	7.92 ± 0.07	27.33	14.99
+attention (ours)	*7.44 ± 0.01*	*8.00 ± 0.1*	*25.89*	*13.14*

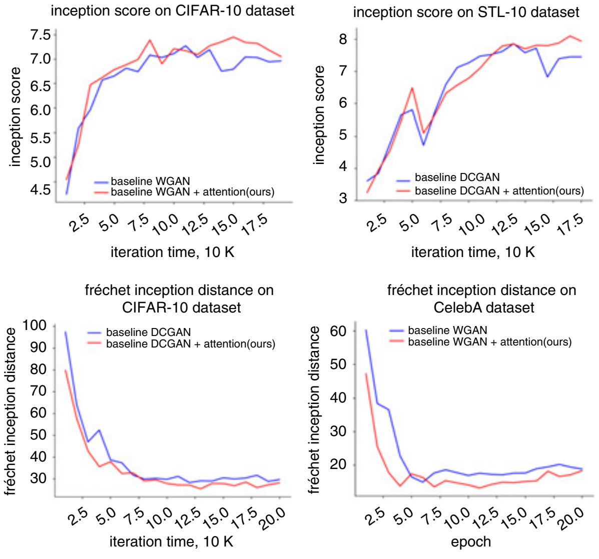

Fig. 3 *Some training curves for baseline models and our model RAGAN without any stabilisation techniques*

Next, we show samples randomly generated by our RAGAN trained on three datasets for qualitative assessment. As shown in Fig. 4, although the visual quality of the images generated by baseline DCGAN [4] and WGAN [5] seems good, the profiles of many images are blurry. However, the profile of the samples generated by our RAGAN is easier to see clearly. Fig. 5 shows the generated samples trained on CelebA dataset, from which we can see our model can achieve higher visual quality with fine details. This verifies the great effect of our model on improving the visual quality of generated samples.

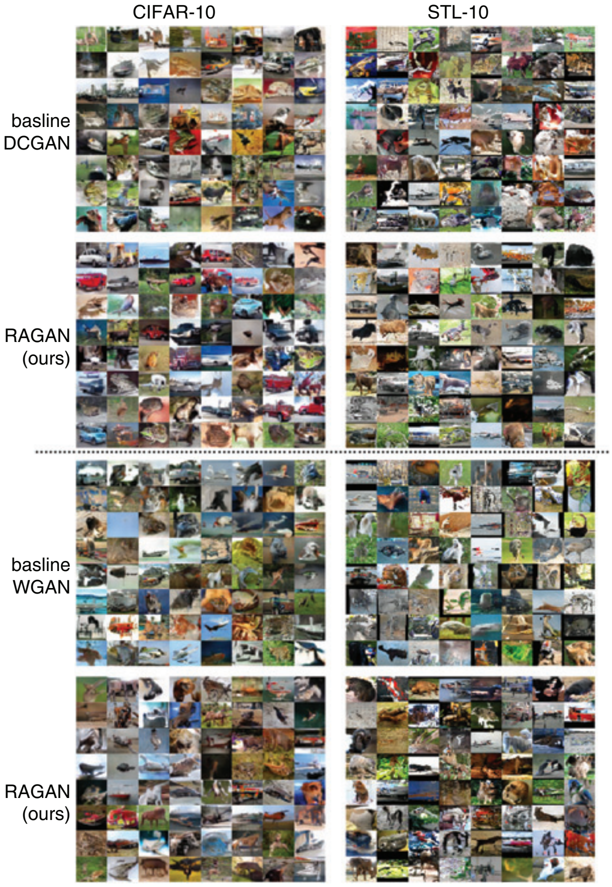

Fig. 4 *Comparison of randomly generated samples after 200k iterations training on CIFAR-10 and STL-10 datasets*

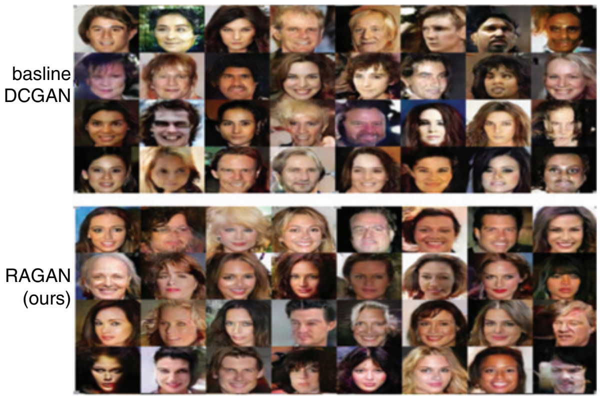

Fig. 5 *Comparison of randomly generated samples after 20 epoches training on CelebA dataset*

Conclusion: In this Letter, we introduce a new attention mechanism to improve the quality of generated samples by building dependencies between high-level representations extracted from some attention regions of real images and corresponding feature maps of the generative network. As a result, some feature modes of the data distribution can be stressed and some high-level targets can be configured to guide the generation network fitting the distribution of real data. We conduct extensive experiments on widely used nature datasets. The improvement in IS and FID confirms the effectiveness of our proposed attention model. The samples generated by our RAGAN present high visual quality with fine details and clearer profiles.

Acknowledgment: This work was supported in part by the Natural Science Foundation of Guangdong Province (2016B010123005, 2017B090909005).

Submitted: *31 December 2018* E-first: *4 March 2019*
doi: 10.1049/el.2018.8251
One or more of the Figures in this Letter are available in colour online.

Wei Wang, Haifeng Hu, Yi Huang and Chongchong Ruan (*Sun Yat-Sen University, 26469, Guangzhou, 510275, People's Republic of China*)

Dihu Chen (*School of Electronics and Information Technology, Sun Yat-sen University, 510275, People's Republic of China*)

✉ E-mail: stscdh@mail.sysu.edu.cn

References

1 Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., *et al.*: 'Generative adversarial nets', Advances in 27th NeuralInformation Processing Systems (NIPS), Montreal, Canada, December 2014, pp. 2672–2680

2 Zhang, H., Goodfellow, I., Metaxas, D., *et al.*: 'Self-attention generative adversarial networks', 2018, arXiv:1805.08318[stat.ML]

3 Shuang, M., Jianlong, F., Chen, C.W., *et al.*: 'DA-GAN: instance-level image translation by deep attention generative adversarial networks'. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, June 2018, pp. 5657–5666

4 Radford, A., Metz, L., and Shintala, S.: 'Unsupervised representation learning with deep convolutional generative adversarial networks'. Int. Conf. on Learning Representations, San Juan, Puerto Rico, May 2016

5 Arjovsky, M., Chintala, S., and Bottou, L.: 'Wasserstein GAN', 2017, arXiv preprint arXiv:1701.07875