Comparative Analysis the Super-Resolution Image Generation Performance Based on BigGAN and VQ-VAE-2

Lingtao Cai

University of Bristol, Bristol, BS8 1UG, United Kingdom

xo19066@bristol.ac.uk

Abstract. Super-resolution image reconstruction has always been a popular research direction in the field of computer vision, which aims to recover high-resolution clear images from low-resolution images. Traditional super-resolution reconstruction algorithms mainly rely on the construction of constraints and the accuracy of registration between images to achieve the reconstruction effect, but their accuracy cannot meet the needs of practical applications with large multiples. Thanks to the rapid development of deep learning field, super-resolution image reconstruction based on deep learning has become the mainstream, and has achieved great success in reconstruction accuracy and speed. According to the different generative models used, the existing super-resolution image reconstruction methods mainly include two categories: GAN-based and VAEs-based. To quantitatively compare the limits of the two approaches' performance, this study selects two representative algorithms, BigGAN and VQ-VAE-2, and introduces the theoretical details and training process of these two methods, respectively. Furthermore, the reconstruction results of BigGAN and VQ-VAE-2 are further compared. Finally, this paper discusses the future development trend of super-resolution picture reconstruction with the current potential problems of BigGAN and VQ-VAE-2.

Keywords: Super-resolution image reconstruction, deep learning, BigGAN, VQ-VAE-2.

1. Introduction

The ability of an imaging system to accurately reflect the fine details of an item is measured by the richness of the detailed information present in an image, which is known as image resolution. Compared to low-resolution photos, high-resolution images often have higher pixel densities, richer texture details, and higher dependability. However, due to the constraints of acquisition equipment, network transmission medium and bandwidth, image degradation and many other factors, the acquired high-resolution images often cannot meet the actual application requirements, such as Chest CT image in clinic testing [1] and images in thermal imagery [2]. To this end, the technology of super-resolution image reconstruction has gradually grown into a research hotspot in the field of computer vision with the aim of recovering high-resolution clear images from low-resolution images.

Traditional super-resolution reconstruction algorithms mainly rely on the construction of constraints and the accuracy of registration between images to achieve reconstruction effects, incorporating super-resolution reconstruction based on interpolation, super-resolution reconstruction based on degradation models, and super-resolution reconstruction based on learning. Rate reconstruction in three categories. However, with the increase of the magnification factor, the information provided by the artificially defined prior knowledge and observation model for super-resolution reconstruction becomes less and less, and the reconstruction effect cannot meet the needs of practical applications with large multiples. Convolutional neural networks' quick progress has made it possible for deep learning-based image super-resolution to steadily gain popularity and experience remarkable success in recent years. The need to improve image quality requires more research on the image generation process. According to the differences in generative models, the existing deep learning-based image re-algorithms mainly include generative adversarial networks [3], variational autoencoders [4] and autoregressive models [5] et al. Among all of these models, Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) are gaining popularity and showing more distinct features in image production. Additionally, there is more

research value in terms of the calibre and variety of the data obtained due to the parallels and contrasts between the two models and their underlying structures.

Super-resolution reconstruction based on GANs. GAN is one of the most famous models of image generation, which set a minimax objective between generators for producing images from random noise to generated images and discriminators for classifying if the images generated by generators are real or fake and update calculating the loss functions of the generators based on the results of calculation. The specific implementations of GANs have different features and different situations of application. The larger-scale GAN(BigGAN) shows outstanding effects on high resolution and high-resolution image generation. However, the GAN models are changeling with the problem of missing variety and mode collapse which influence the stability and diversity of high-fidelity image generation respectively. Furthermore, it is challenging to assess the diversity of the BigGAN results obtained. With random features, things look incorrectly, meaning that the features produced by the generator might not match the expected outcomes or be unable to match the genuine images used to train the discriminator. The number of pieces in the photographs or where they are located does not match the real images.

Super-resolution reconstruction based on VAEs. Compared with GANs, Variational Autoencoders (VAEs) which is an upgraded model of auto-encoder applies the way of variational inference to construct the generation model. VAEs is another type of generated model which based on negative likelihood optimization of the training which compare all generated samples with the original data by calculating the Kullback-Leibler divergence (KL-divergence, [6]) of model and data distributions. Unlike GAN models, VAE is more powerful to present latent codes. Moreover, VAE avoid the problems in GANs such as mode collapse and weak diversity since it allows the reconstruction error of samples. However, there are still some disadvantages of VAE in image generation which contain the relatively low quality of generated images because of the lack of Adversarial Networks. Furthermore, there is also potential risk of posterior collapse when the decoder becomes too strong in the training of models. There is a model called VQ-VAE-2 which improve the quality of the images so that the generated images are in high resolution, but further measurements of the quality of the images and diversity are still remain to be verified.

For both GANs and VAEs, they apply different solutions to extract features from training data set and make attempts to generate images in high quality and diversity. Although for both of new models i.e., BigGAN and VQ-VAE-2 respectively, they show high performance in super-resolution image generation, it is still hard to compare them in a comprehensive way. Moreover, it is difficult for a generative model to maintain both quality and diversity at the same time in most cases. Normally different models have unique features in different area of application. Therefore, the target of this paper is to compare the performance of GANs and VAEs in high resolution image generation with their latest model namely BigGAN and VQ-VAE-2 respectively. the criteria of evaluation applied in this paper includes the Inception Score (IS) [7] and Fréchet Inception Distance (FID) [8] to measure the quality of generated images.

2. Method

2.1. Generative Adversarial Networks

GANs are featured and emerging models which are suitable for both semi-supervised learning and unsupervised learning. The basic theory of GANs is to model high-dimensional data implicitly through a pair of networks in the relationship of competition [3]. One of the two types of networks is known as generator(G), which is aiming to generate “fake” data as real as possible. In contrast, the other type of networks, discriminator(D), is trained with real data and designed to classify the generated data and labelled them with fake or real. For example, in image generation, the generator will take noises, and generates forgeries, i.e., fake images to deceive the discriminator, and the discriminator will be trained with both real images and the image generated from generators in order to make judgement correctly. The relationship between generators and discriminators are in

competition, the symbol of a success GAN model is that the discriminator is completely fooled by the generator so that the fake images generated from generators are as real as possible. The architecture of a classic GAN model is shown in Figure1.

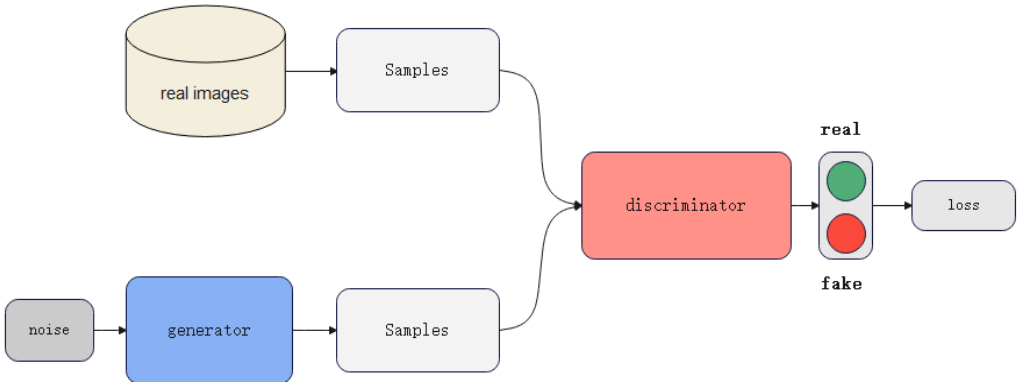

Figure 1. The architecture of a basic GAN model. The samples from generator will be passed to the discriminator so that it will provide a result of true or false as input for the loss function. The loss function is used to constrain the training of this discriminator in order to optimize it for the next iteration

However, the problem of the generator is that they can not access the real images directly, so the generator has to "guess" how to produce a fake image but extremely similar to the real image. The possible way is to update itself gradually in a proper way. The training of both generators and discriminator are in a recursive process, which allows the generator and discriminator update each other depends on the results produced in the last iteration, i.e., the generated images and the judgement from the discriminator. Therefore, the discriminator firstly gains the sample produced by the generator and provide a signal that indicates if the sample was judged as real or fake based on the real samples from the training data set. The generator can be trained to produce better samples in the following iteration by using this error signal.

The network which constructs the generator and the discriminator can be implemented by multilayer networks which includes convolution layers. The networks of the generator and discriminator are normally invertible between each other since the tasks of them are opposite. For example, if a generator network is to map from a late space to images, then the generator can be presented as $G: G(z) \rightarrow R^{|x|}$, where z is a sample from latent space and $X \in R^{|x|}$ is a picture with the dimensions x which is as same as the real images. In contrast, the discriminator of classic GAN may be characterized in an opposite direction which maps from image data to the results of judgement. However, in basic GAN, the results produced from discriminator, D, can become the probability that if the generated sample, which is the image in image generation, comes from real image dataset distribution. For example, the discriminator can be trained to apply binary classification: $D: D(x) \rightarrow (0,1)$ when the generator is fixed. The images can be classified to one if it is seen as real image or close to zero if it is discriminate as fake image. After updating the discriminator, the generator can be trained with fixed discriminator to fool the rate of correct judgements of discriminator. When the distribution of the generator is matched with the real data, the discriminator will be fooled which means the generator is well-trained.

The process of interactively training generator and discriminator can be named as "Minimax game". In order to evaluate the quality of the images generated by generator, there is a value to calculate which is the loss value, i.e., the value represents the extend of how much the image is different from the real image. The loss function of basic GAN model can be shown as follow:

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}[1 - \log D(G(z))] \quad (1)$$

Where the D, G are the network of discriminator and generator respectively, the discriminator is aiming to maximize the equation at the right side rather that generator target to minimize the equation.

Additionally, $D(x)$ is the probability that the discriminator D considers the sample comes from the original distribution, and $D(G(z))$ is the probability that the discriminator D mistakenly discriminates the fake samples from the generator as real. As a result, the discriminator's job is to maximise D(x) while minimising D(G(z)).

In comparison to basic GAN, BigGAN [10], also known as large scale GAN, demonstrates the high performance of GAN training with even as many as four times the parameters and seven times the batch size. The architecture of generator G of BigGAN can be presented as follows in Figure 2. As shown in Figure 2(a), Each layer of the generation network receives the noise vector z after it has been broken into several blocks, associated with the conditional label c, and transmitted. The structure of the residual block in Figure 2's (b) can be further expanded for each residual block of the generation network.

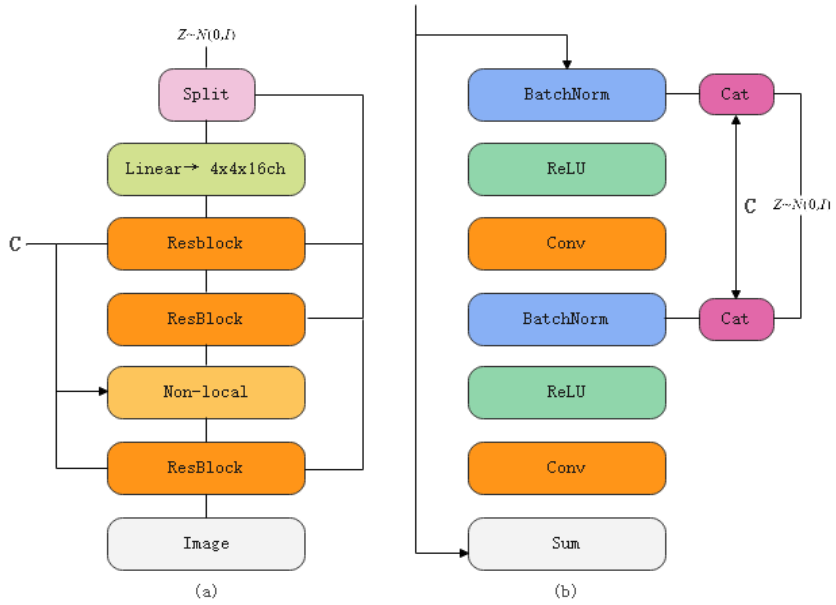

Figure 2. (a) The architecture layout for generator with the labels in details (b) A residual block in generator G

The "truncation trick" involves defining a threshold and resampling out-of-range values to come within it in order to truncate the sampling of z by sampling from the earlier distribution of z. The generation quality indicators IS and FID can be used to calculate this threshold. The quality of the generation improves as the truncation threshold drops, but the generation also veers toward singularity. Therefore, it is a decision to weigh the generation quality and the generation diversity in accordance with the generation requirements of the experiment. Additionally, certain larger models are unsuitable for truncation and will exhibit saturation problems when truncation noise is embedded in them. To prevent this, the BigGAN designer enforces truncation adaptation by modifying G's smoothness qualities so that every point in the z-space maps to useful output samples. In this case, the BigGAN applies orthogonal regularization, which directly enforces the orthogonality condition [11]:

$$R_\beta(W) = \beta \|W^T W - I\|_F^2 \tag{2}$$

This kind of regularization is often too restrictive. In order to lower the constraints while achieving the desired smoothness of the model, it is found that the best version deletes the diagonal parameters from the regularization. Therefore, the equation is improved in BigGAN as:

$$R_\beta(W) = \beta \|W^T W \odot (1 - I)\|_F^2 \tag{3}$$

Where W stands for weight matrix, β represents the hyperparameter. By eemploying a "truncation trick" of the prior distribution z, it allows fine-grained control over sample diversity and fidelity.

Overall, BigGAN is designed to increase the batch size, "truncation technique" and control the stability of the model so that the training performance can keep in a relatively high level even with large scaling.

2.2. Variational Autoencoders

The Variational Autoencoder is a improved version of Autoencoder (AE) which is a self-encoder that can obtain a latent feature code from the original feature, realize automatic feature engineering, and achieve the purpose of dimensionality reduction and generalization through self-supervised training.

The reason for self-supervision is because the target of the network is the input itself, so there is no need for additional labelling work. Although it consists of two parts, the encoder and the decoder, it is clear from that the focus of AE is on encoding, i.e., getting the vector of this hidden layer as a potential feature of input, which is a common way of embedding. The result of the decoding, based on the training target, if the loss is small enough, will be the same as the input, and from this point of view the decoded value has no practical significance, except to add to the error to supplement and smooth some of the initial zero values or some use. Because, the entire process from input to output is based on the mapping of the existing training data, although the dimension of the hidden layer is usually much smaller than the input layer, but the probability distribution of the hidden layer still depends only on the distribution of the training data, which leads to the distribution of the hidden state space is not continuous, so if we randomly generate the state of the hidden layer, then it is likely to no longer have the characteristics of the input feature after decoding, so it is a bit difficult to generate data through the decoder.

In order to address the drawbacks of AE, VAE assumes that the hidden layer encoded by the neural network is represented by a standard Gaussian distribution. It then samples a feature from this distribution and decodes it using this feature, anticipating the same outcome as the original input. The loss of VAE and AE are almost the same, only to increase the coding inference distribution and the standard Gaussian distribution of the KL divergence[6] of the regular term, obviously the purpose of increasing this regular term is to prevent the model from degenerating into ordinary AE, because in order to minimize the reconstruction error during network training, the variance will inevitably be gradually reduced to 0, so that there will no longer be random sampling noise, and it will become ordinary AE. The loss function of the VAE can be presented as follows:

$$loss = \|x - x'\| + KL(N(\mu, \sigma), N(0,1)) = \|x - d(z)\| + KL(N(\mu, \sigma), N(0,1)) \quad (4)$$

Where ||x-x'|| is the loss of VE. Since multivariate normal distribution with each component are considered independently, only the case of a unary normal distribution is needed to derive. The KL divergence between the independent multivariate normal distributions of each component is:

$$KL(N(\mu, diag(\sigma^2))||N(0,1)) = \frac{1}{2}\sum_{i=1}^{d}(\mu_i^2 + \sigma_i^2 - \log \sigma_i^2 - 1) \quad (5)$$

Where d is the dimension of the hidden variable Z.

Variational inference is a featured algorithm of VAEs, which considers a Bayesian inference problem, given the observation variable $x \in Rk$ and the latent variable $z \in Rd$ whose joint probability distribution is p(z, x) = p(z)p(x|z). In order to calculate the posterior distribution p(z|x), It can be assumed that a variational distribution q(z) comes from the distribution family Q, by minimizing the KL divergence to approximate the posterior distribution p(z|x):

$$q* = \underset{q(z)\in Q}{\operatorname{argmin}} KL(q(z)||p(z|x)) \quad (6)$$

Based on Bayesian formula:

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)} \quad (7)$$

The optimized q* can be presented as follows:

$$q* = \underset{q(z)\in Q}{\operatorname{argmin}} E_{q(z)}[-\log p(x|z) + KL(q(z)||p(z))] \tag{8}$$

Which is exactly the loss function presented above in (4), the main target of the VAE model is to minimize the loss value so that the images generated throng decoding can be as similar as the real images.

VQ-VAE [12] and the VQ-VAE-2 [13] is designed to solve the problem from Autoregressive models which fail due to a surge in computational demand when the image is relatively large. The idea of VQ-VAE is that if you can compress the photo into a low-dimensional space, train an autoregressive neural network in a low-dimensional space, and then decode it to a high-dimensional space, i.e., using autoregressive neural networks on compressed space z:

$$p(z) = p(z_0)P(z_1|z_0)\ P(z_2|z_0z_1)\dots P(z_m|z_0z_1z_2\dots z_{m-1})) \tag{9}$$

The number of elements in the z-vector m is generally much smaller than the number of pixels in x N. Another advantage is that the original image has a lot of redundant information, such as large blocks of homochromatic pixels, for the use of autoregression, you can ignore this redundant information, to get a structured global semantic probability distribution.

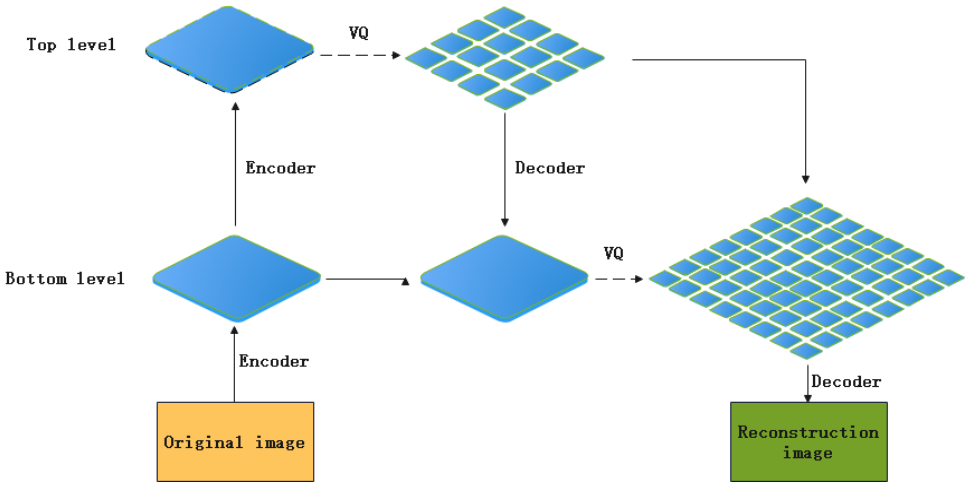

Figure 3. The architecture of VQ-VAE-2 with a more hierarchical structure

Figure 3 shows the training process of VQ-VAE-2 which is divided into two layers. The upper potential space is 32x32, and the lower potential space size is 64 x 64. The upper layer first performs hierarchical quantization to obtain the quantized dictionary vector:

Using this dictionary vector as a condition, along with the input x, calculates the quantized form of the underlying potential space. Finally, the upper and lower layers are quantized into the selected dictionary vectors etop and ebottom into the decoder at the same time, calculating the loss function defined earlier, updating networks of encoder and decoder, and the weights of the vector.

Compared with VQ-VAE, in VQ-VAE-2, for the upper layer, the authors used a multi-headed self-attention mechanism. The self-attention mechanism has a relatively good long-range correlation because it considers the association of any one position with all other positions. However, because the computational complexity is $O(n^2)$, for the underlying feature map, n=64x64, the memory overflows, so in the actual calculation, only the global attention mechanism is used for the upper layer. Conditional probabilities from the top layer can help the bottom layer generate a good partial map.

3. Experiment

The main target of both BigGAN and VQ-VAE-2 is the same which is to generated images in high fidelity i.e., high quality. The level of images quality is hard to evaluate since not only the similarity between generated images and real images but also more detailed indicators such as picture saturation, picture detail plausibility and the fidelity of the results should also be considered. Under the consideration of the efficiency of the generation process and the effects of the super-resolution image

generation, the training data set in this comparison experiment employ with the images from ImageNet size up to 512×512 resolution, and quantify the analysis of larger pictures resolutions.

For evaluating the quality of the picture being generated quantitatively, the indicators of IS [7] and FID [8] are applied. The IS indicator is an effective way to evaluate clarity: Assume that a cleaned image should have a very high likelihood of belonging to one class and a very low probability of belonging to any other classes. Entropy represents the degree of chaos, the uniform distribution of chaos is the largest, and entropy is the greatest; theoretically, entropy should be minimum. In other words, the more precise the probability distribution function plot's output. For diversity, Generated pictures in all categories of probability with a large p(y) entropy distribution (uniform distribution). Compared with IS score, FID applies the mean and covariance distance evaluation between the real picture and the generation of the image after extracting the feature vector. When the generated picture and the real picture features are closer, the smaller the square of the difference in means, the smaller the covariance, and the smaller the sum (FID). Generally, the lager IS means higher quality in image generation, and the smaller the FID, the better the picture quality, the better the diversity.

The paper of VA-VAE-2 [13] indicate a classifier method based on Rejection Sampling to make the samples closer to the true data manifold by labelling them through a pre-trained classifier. In terms of the BigGAN model, although it shows slightly worse FID score as it grows, it also shows relatively stable shape of curve. Moreover, the BigGAN-critic curve shows in Figure 5 indicate that the images generation process can be apparently improved with the rejection sampling applied. Additionally, the GAN models have the advantage of efficiency in training process since there is no process like comparing all data distributions in VAEs so it can be easier to enlarge the scales in order to make intuitive effects in super-resolution image generation.

Table 1. Evaluation of different models at different resolutions [10]

Model	Resolution	FID / IS	(Min FID) / IS	FID / (valid IS)	FID / (max IS)
SN-GAN	128	27.62/36.80	N/A	N/A	N/A
SA-GAN	128	18.65/52.52	N/A	N/A	N/A
BigGAN	128	8.7/98.8	7.7/126.5	9.6/166.3	25/206
BigGAN	256	8.7/142.3	7.7/178.5	9.3/233.1	25/291
BigGAN	512	8.1/144.2	7.6/170.3	11.8/241.4	27.0/275

As the data shows in Table 1, the third column to sixth column shows the scores without truncation, the best FID scores, the valid IS scores, and the maximum IS scores respectively. Quantitatively, the results are compared with other GAN models such as SN-GAN and SA-GAN, the BigGAN shows apparent lower value of FID which stands for better image quality, and higher IS score which represent dramatically improvements of both quality and diversity. Moreover, with the scores shows between different resolution, it shows stable high performance in with higher resolution. Furthermore, BigGAN also implement with a featured method that improve the quality and efficiency in the generation process through changing of the batch size, which is beneficial for testing and evaluating in qualitative analysis.

4. Discussion

BigGAN and VQ-VAE-2 both focus on super-resolution image generation and all produce high quality of images. However, there are still some potential problems may be caused during the process. For BigGAN, the main problem caused by the initial idea of itself which is to improve the image quality through enlarging the Batch size. Large Batches, large parameters, truncation, and large-scale GAN training stability control were used by BigGAN to accomplish its feat. In contrast, huge amount of computation may also bring the make challenges for both hardware and software. As the computation grows, the requirements of hardware will become more pressured in order to satisfy the growth of computation. In terms of software, the main target is how to maintain the stability and the speed of computation when the batch size increased, it might be suggested that a improved algorithms which can wisely divide large data set and provide a well-structured way to process through whole

data set might solve such a problem. A further suggestion for BigGAN comes from the paper of VQ-VAE-2 [13] which indicate that the application of Rejection Sampling can be beneficial in BigGAN models so that it can improve the performance and results produced from BigGAN.

For VQ-VAE, the potential problem caused by the key concepts of the VQ-VAE which is dimensionality reduction followed by modelling the encoding vectors with PixelCNN. Because of the discrete sequence generated by PixelCNN, Modelling the coding vector with PixelCNN means that the coding vector is also discrete. Generating discrete variables also often means that there is a problem of gradient vanishing which is the classic problem for many generative models. Moreover, although the Layering techniques applied in VQ-VAE-2 shows good behaviours in training the encoder and decoder, the suggestion for future development might be modelling information by layering different structures specifically for different data types

Both GANs and VQ-VAE have different advantages and disadvantages, GANs can be easier to build and better efficiency in higher resolution image generation but it suffers from the problems such as mode collapse. GANs may be too creative so that in some cases, may produce some unexpected samples which lead unnatural sense in the generated images. VQ-VAE, on the other hand, VQ-VAE produces samples that are closer to those that exist in the real world because of discretization. VQ-VAE separates the top-level global from the underlying local information to produce a globally self-consistent, locally high-definition image. This method can also be used for image generation, speech generation, continuous video generation. In wide range of area, further potential applications of featured VQ-VAE can make contribution to the development of AI in terms of natural language generation, speech recognition and face recognition [14] which are all the vital and valuable directions in the future.

5. Conclusion

By presenting generative models for generating high resolution images with the most popular models in terms of GANs and VAEs respectively, this survey introduces them for their original models, i.e., GAN and VAE, in details which includes concepts, architectures, features and applications. In order to make a survey based on the most powerful models for generating high resolution images, the new models in image generation namely BigGAN and VQ-VAE-2 are introduced and compared in terms of the quality and diversity of the generated images. A detailed description for these two generative models is introduced with their training process in theory and make comparisons between them. For evaluating these two indicators, IS and FID of BigGAN and VQ-VAE-2 are compared under same situation. The results are shown as a table ana a curve graph to highlight the features of both models. Finally, this essay makes further suggestions and possible trends of development based on the current potential problems in BigGAN and VQ-VAE-2.

References

[1] Umehara, K., Ota, J. and Ishida, T. (2018) “Application of Super-Resolution Convolutional Neural Network for Enhancing Image Resolution in Chest Ct,” Journal of digital imaging, 31(4), pp. 441–450. doi: 10.1007/s10278-017-0033-z.

[2] Mandanici, E. et al. (2019) “A Multi-Image Super-Resolution Algorithm Applied to Thermal Imagery,” Applied Geomatics, 11(3), pp. 215–228. doi: 10.1007/s12518-019-00253-y.

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in neural information processing systems, pages 2672–2680, 2014.

[4] Iederik P. Kingma and Max Welling. Auto-encoding variational bayes. CoRR, abs/1312.6114, 2013.

[5] Hugo Larochelle and Iain Murray. The neural autoregressive distribution estimator. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, pages 29–37, 2011.

[6] Joyce, James M. "Kullback-leibler divergence." International encyclopedia of statistical science. Springer, Berlin, Heidelberg, 2011. 720-722.

[7] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In Advances in neural information processing systems, pages 2234–2242, 2016.

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems 30, pages 6626–6637. Curran Associates, Inc., 2017.

[9] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta and A. A. Bharath, "Generative Adversarial Networks: An Overview," in IEEE Signal Processing Magazine, vol. 35, no. 1, pp. 53-65, Jan. 2018, doi: 10.1109/MSP.2017.2765202.

[10] Brock, A., Donahue, J. and Simonyan, K., 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. [online] arXiv.org. Available at: <https://arxiv.org/abs/1809.11096>.

[11] Andrew Brock, Theodore Lim, J.M. Ritchie, and Nick Weston. Neural photo editing with introspective adversarial networks. In ICLR, 2017.

[12] Oord, A., Vinyals, O. and Kavukcuoglu, K., 2017. Neural Discrete Representation Learning. [online] arXiv.org. Available at: <https://arxiv.org/abs/1711.00937v2>.

[13] Razavi, A., Oord, A. and Vinyals, O., 2019. Generating Diverse High-Fidelity Images with VQ-VAE-2. [online] arXiv.org. Available at: <https://arxiv.org/abs/1906.00446v1>.

[14] Jiang, Y., Li, X., Luo, H. et al. Quo vadis artificial intelligence? Discov Artif Intell 2, 4 (2022). https://doi.org/10.1007/s44163-022-00022-8.