Evaluation Metrics and Multi-level GAN Approach for Medical Images

Galina Kovbasa
Belarusian State University of Informatics and Radioelectronics
Minsk, Belarus
g.kovbasa@gmail.com

***Abstract*—This article examined methods for using GANs in medicine, their prospects, as well as problems with training generative adversarial networks associated with the increasing use of generated images for training other networks. The analysis of single-layer and multi-layer GANs concluded that although multi-layer GANs perform better statistically, they do not exactly match the distribution of the original dataset and, without medical supervision, such synthetic data should not be used when training new networks. Problems associated with the phenomenon of recursive learning, biased assessments of image realism, and non-optimized structures are considered. Approach is described in context of integrating generative adversarial network models into the OSTIS Technology based hybrid computer systems.**

***Keywords*—Multi-level GANs, recursive learning, synthetic data**

I. Introduction

Generative adversarial networks (GANs) are a remarkably popular technique for generating realistic synthetic data. Modern GANs can have different layers, backgrounds , complexity and be trained by semi-supervised and unsupervised learning. They gain their popularity because of the ability to implicitly modeling high-dimensional data distributions. [1] GANs are of particular interest in the processing, classification and evaluation of medical images, in the future making it possible to speed up and improve the analysis of the results of magnetic resonance imaging, computed tomography, X-rays and others. This can be solved by integrating a GAN neural networks into the OSTIS Technology. [2]

Integration with third-party technologies based on neural networks allows the development of universal hybrid systems. GANs are capable of not only processing images, but also creating new synthetic data on demand, which makes them valuable for creating datasets, anonymous educational materials, etc. GANs are actively changing during development and all these changes can be formalized in the OSTIS system using SC code. [3]

Thus, network artifacts of the processes of creation, training and further tuning, such as numbers and types of layers, weights, activation functions, can be stored in a general form and saved, supporting further replenishment and expansion of the general knowledge base. [4] The OSTIS intelligent system is also capable of saving several versions of the same model for later use, even restoration from a previous point. [5]

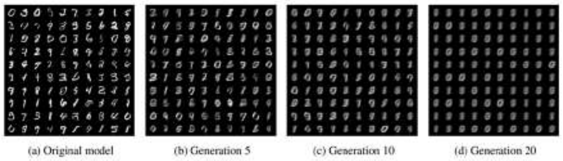
(a) Original model
(b) Generation 5
(c) Generation 10
(d) Generation 20

Figure 1. Model collapsing. Over generations, the generated data begins to look unimodal.

But one of the worst challenges facing GANs in medicine is that medical images are susceptible to various noise and artifacts common to different modalities and, what most important is, have a very little variety of datasets to be used. It should also be noted that much medical information is 3D structures, which can make it difficult to train a GAN on only 2D images.

A. The recursive learning problem

No generative adversarial network can reliably recreate the distribution of the original sample data. It may make mistakes or recreate the same data, that is, clone it. We must train the network so that it does not go beyond the distribution, but also does not repeat the picture. This solution allows us to avoid modern problems with GAN.

Generative adversarial networks are able to generate high-quality images on demand using the same distribution that is given, the score is also high. However, this does not mean that GAN can be used as a universal method or classifier, since it is impossible to assume that the distribution will be reliable or true. Likewise, the images produced by networks cannot be considered representative for training other networks.

However, more and more people are using GANs for reconstruction and generation, but this only leads to the fact that it can be used for a training set, since the generated materials are in the public domain. As a result, the networks degenerate, and the results of reconstruction and generation deteriorate. And as a result, the so-called mode collapse occurs (Fig. 1). [6]

This phenomenon (using generated images as training materials) is most dangerous for image generation, because in this case the networks mutually degrade. The problem of degradation of the second derivative of the loss of the original distribution leads to the fact that the learning network will take everything that the network gives without a critical attitude. It is particularly dangerous for GAN involved in the reconstruction of medical images. It is main case when a medical specialist plays a leading decisive role.

At the moment there is no complete solution to this problem. The search for ideal hyperparameters or best evaluation metric is still ongoing. Degradation can only be prevented by correct data preparation and correct interpretation of data. Even in the case of realistic results, a specialist's hand is still needed, otherwise it leads to the original problem. Partly, the problem of network degradation is related to the lack of training materials in the public domain, because all materials in one way or another relate to personal data, although they are as depersonalized as possible (but there is still a diagnosis verified by a specialist), which limits the number of networks available for training databases. In view of this fact, any decision of the GAN is only to some extent true.

B. Evaluation of GAN

1) Density estimation: Density estimation is a challenging unsupervised learning problem. Existing maximum likelihood approaches for density estimation are either limited or unable to generate high-quality samples. However, the lack of density estimation limits the application of sample data.

Density estimates are needed in a wide range of practical computer vision problems, especially when the likelihood of the generated samples is critical. This occurs, for example, when it is important to simultaneously explore and optimize the search space, when confidence estimates of a hypothesis are required, or when control over the level of generalization is important. Typical sampling quality metrics are inadequate because the generative model may simply remember the data or miss important modes.

Such methods are effective for determining whether a generative model has learned the correct statistics, but they are somewhat limited. Most techniques define the statistic to be zero if the generated GAN and the true samples belong to the same distribution. Difference between distributions is measured only on the basis of certain statistics, ignoring other. In particular, the manifold representation ignores the densities that the generator assigns to different parts of space, as well as whether the manifold is more abundant in regions around the true distribution.

Alternative density estimators, such as auto-regressive models, stream methods, or non-parametric methods such as kernel density estimation (KDE), are either too computationally intensive or require significant neural network engineering [7].

2) Log-likelihood: Log-likelihood is widely used to evaluate other families of generative models. On top of that, log-likelihood has been used before to demonstrate that a wide family of generative models assigns a greater likelihood to images outside of the training distribution. In [8] states that probability-controlled models that have much worse FID show better performance and overall distribution evaluation than state-of-the-art GANs. By evaluating GANs with low FIDs, we show that multi-level GANs are superior to single-level models in terms of average test log likelihood and generate subjectively better images on medical datasets. [9] indicates that AIS (Monte Carlo method that estimates an equation's integral by utilizing various intermediate distributions) is accurate enough to make reliable comparisons between models and can compete with other alternative density estimators. There is no guarantee that such approximation metrics will hold for real data, although [9] found that the behavior of AIS closely matches real and simulated data.

C. Multi-layer GAN

A multi-layer GAN, also can be called hierarchical or nested GAN, is a type of generative adversarial network (GAN), comprising with multiple networks of generators and discriminators organized in a hierarchical or nested structure. The main idea of multi-layer GANs is to improve the quality and variety of generated samples by using multiple layers of abstraction. By training generator and discriminator networks at different levels of abstraction, multi-layer GANs are able to learn more complex and diverse data distributions. In such a GAN, the generator network creates a sequence of samples that become increasingly refined and detailed, and the discriminator network evaluates the quality of these samples at each level of abstraction. Multi-layer GANs can be implemented using various architectures and techniques such as progressive growth, ladder networks, and recursive networks. These approaches differ in how they organize hierarchical or nested connections between networks of generators and discriminators, and in how they convey gradients and losses between layers.

II. Related work

A. Preparing of the data

In order to evaluate the realism of images based on medical data, two sets of datasets, The Automated Cardiac Diagnosis Challenge (ACDC) [10] and The Indian Diabetic Retinopathy Image Dataset (IDRID) [11], were selected to evaluate the quality of synthetic image reproduction using GAN. These datasets were chosen not only because of the high quality of the various types of images, but also because of the low number of uses in GAN research.

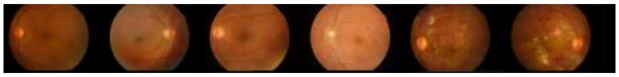

Figure 2. IDRiD images. From not affected to severe cases

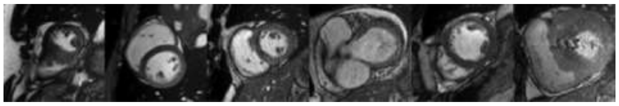

Figure 3. ACDC MRI image slices

ACDC dataset contains 150 studies of short-axis cardiac cine MRI of the University Hospital of Dijon patience, 1902 2D slices for training and 1078 2D slices for testing [10]. The 516 photos in the IDRID collection include both pathological and normal retinal fundus [11].

For the purpose of metrics analysis, we sampled a large number of images from each of our trained GANs to see how well it could learn the original data distribution. According to [12], although GANs are characterized by sensitivity to hyperparameters, a subset of GANs are characterized by their oversensitivity, as shown by the sharp change in FID score under different sets of hyperparameters.

Firstly, evaluating set of hyperparameters require some computational time to acxhive best performs for all the GANs described below.

Secondly, GANs are sensitive to the reference dataset: as it increases, image quality improves. To represent difference in GAN characteristics due to dataset size difference all estimation metrics will be evaluated on each dataset with different numbers of training samples.

Each GAN described below produced a batch of two thousand images without additional post processing.

B. GAN approaches for realistic data generation

1) Deep Convolutions in GAN: Deep convolutional GANs [13] were one of the first GANs to use convolutional layers and made significant contributions to balanced GAN learning. Although convolutional layers have been used in GAN architectures before, DCGAN offers an adapted architecture. Several rules have been proposed to create a stable convolutional architecture. Although convolutional networks have been used in GAN architectures before, DCGAN offers an adapted architecture. Several rules have been proposed to create a stable convolutional architecture. These rules are as follows [13]:

- Don't use merge layers. Instead, use straight line convolutions for the discriminator and fractional line convolutions for the generator.
- Use batchnorm in generator and discriminator layers.
- Do not use fully connected structures in hidden layers.
- In the output layer of the generator, use the Tanh function, in other layers - the Relu function.
- Use the LeakyReLU activation function on all discriminator layers.

Deep neural networks with a large number of parameters are very powerful machine learning systems. However, a serious problem for such networks is brute force. Large networks are also slow to use, making it difficult to combat overfitting by combining the predictions of many different large neural networks during testing. To solve this problem, the dropout technique is used [14]. Although DCGAN has a variety of advantages over non-convolutional models, in further chapter we show, that multi-level non-convolutional GANs outperforms it in terms of the quality of generative image for medical data.

2) Supervised vs Unsupervised networks: The variety of image structure is determined by the criteria of realism, which must be described, but at the end of training, such an approach is already meaningless, since the criteria need to be adjusted during the work process. Networks with this approach are called semi-supervised.

In recent years, deep generative models have dramatically pushed forward in generative modelling, achieving state-of-the-art semi-supervised learning results. Unsupervised networks also show good results, but they do not have the same advantages:

- The ability to predict H+1 classes (number of classifiers) during training time, forces class labels to be displayed. This extra class correlates with the outputs of generative models and can produce higher quality outputs. This may affect the results of the generative model. We can pass through this class to the corresponding outputs of the generative model, which allows us to improve performance and quality (fewer epochs). This can be compared to the feedback between a discriminator and a generative model.
- We can submit softmaxes of fake and original images.
- Better pictures and at the same time fewer epochs.
- Possibility of reducing the impact of type I error: fake images that are perceived as real [15].
  The common approaches that have been used in most multi-layer GANs is a progressive growth [16]

3) Progressive growth: A GAN training methodology that starts with low-resolution images and then gradually increases the resolution by adding layers to the network, as shown in the figure 4. This gradual nature of learning allows you to first detect the large-scale structure of the distribution of images (the small size of which, as noted by [16], is characterized by greater stability), and then switch attention to increasingly smaller details, refining and complementing the image, rather than studying all

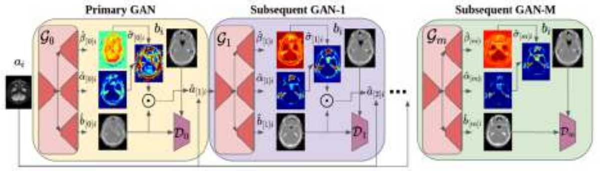

Figure 4. General architecture of Uncertainty-Guided Progressive GANs

scales simultaneously. Accordingly, the following advantages can be identified:

- More stable image generation in the early stages.This is due to less information about classes and fewer modes. The resolution enhancement approach simplifies the task for generating the final images. And this ultimately leads to image stabilization with reliable synthesis of results.
- Reduced training time.When using progressively growing GANs, most iterations are performed at a lower resolution, and comparable result quality is often achieved 2-6 times faster, depending on the final output resolution.
- Relatively low GPU requirement [17]. Because GANs at each scale are trained separately, the training process never exceeds a certain maximum size, so GPU requirements remain consistently low, making it easier to generate arbitrarily large images.

4) Multi-scale discriminator: The [17] used a multi-scale discriminator consisting of two parts of the same structure, which made it possible to solve problems that single-scale discrimination could not handle. The first discriminator, D1, looked at the smaller images to process the overall structure, and the second, D2, helped to reconstruct the finer details of the image. This approach allows you to train a network to find structures in an image more efficiently than single-level GANs. It is also worth noting that this way we can conclude that multi-layer GANs are inherently multi-layer discriminators. In [18], images are generated in several stages (Primary GAN, Subsequent GAN-1, ..., Subsequent GAN-N) (Figure 4).

The output of one phase serves as input to the subsequent GAN in the next phase, explicitly guided by the attention map derived from the uncertainty estimates. Using multi-level generators and discriminators, the authors achieved comprehensive correction of artifacts and noise, potentially replacing additional imaging procedures, which can reduce examination costs and time [18].

A similar approach was used in [19] and [20], where the image was also generated sequentially using enhancement modules (Fig. 6). At the same time, in [21] multi-level generators were used instead of a multilevel discriminator, since it was noted that the instability of the discriminator interferes with focusing on noise removal.

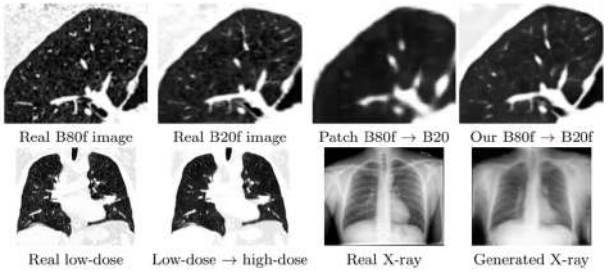
Real B80f image
Real B20f image
Patch B80f → B20
Our B80f → B20f
Real low-dose
Low-dose → high-dose
Real X-ray
Generated X-ray

Figure 5. Results of Multi-Scale GANs (MSGANs)

This fact is important, since medical images are susceptible to various noises and artifacts characteristic of different modalities, so it makes sense to consider GAN network architectures where there are no discriminators at certain levels. The [21] architecture is discussed in more detail in a separate paragraph.

III. Material and Methods

A. Modern multi-layer GAN architectures

Unlike conventional GANs, multi-layer architectures produce higher-quality images. Since, presumably, the discriminator is the largest part responsible for the quality of image synthesis (a similar conclusion can be made based on [22]), in order to avoid errors of the second type. In the case of multi-level GANs, to improve the quality of generation, a load with additional discriminators is used. Also, a significant part of the result depends on the settings of the output layer loss function. In the case of multi-layer GANs, the result of the previous layer goes to the output layer function, which is scaled to the next layer. Thus, such an algorithm is not a multi-level synthesis, but a separation of correct characteristics from incorrect ones. The main advantage of the multi-level GAN architecture is that, unlike single-level ones, it is capable of such separation.

1) Uncertainty-Guided Progressive GANs (ProGANs): In Fig. 4 [18] shows a model consisting of cascaded GANs, where each generator is capable of estimating aleatory uncertainty as well as generating images. This solution removes the above-mentioned limitations of modern methods. Getting rid of them by modeling the underlying distribution of residuals across pixels as an independent but non-identically distributed zero-mean generalized Gaussian distribution (GGD), so
$\widehat{b}_{ij} = b_{ij} + \epsilon_{ij}\epsilon_{ij} \sim GGD(\epsilon; 0, \alpha_{ij}, \beta_{ij}) \equiv \beta_{ij}(2\alpha_{ij}\Gamma(\beta_{ij}^{-1}))^{-1} exp(-\alpha_{ij}^{-1}|\epsilon|^{\beta_{ij}})$

2) Hierarchical GANs (HierGANs): HI-GAN [21] consists of four sub-networks, namely G_b, G_a, discriminator D_a and boost network G_c. Both G_b and G_a generators are DCNNs used for image desaturation. In addition, G_a is trained together with D_a to improve its ability to desaturate damaged images and preserve details. The advantage of G_a is its ability to solve the problem of

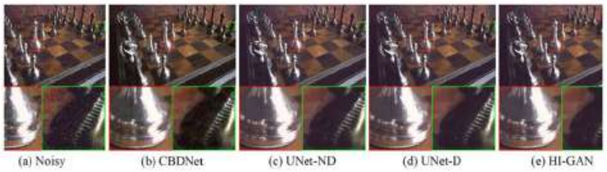

Figure 6. Results of Cascaded Refinement Networks (CRN)

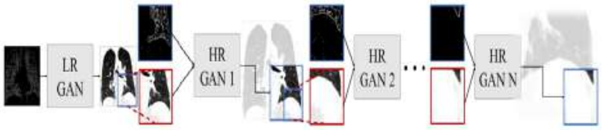

Figure 7. General scheme for image generation using Multi-scale GANs

loss of high-frequency characteristics such as edges and texture, constantly competing with D_a in a repeated zero-sum game. In contrast, G_b learns on its own and does not need to compete with any network. G_b's strategy is to avoid the influence of discriminator instability and focus only on noise removal. Overall, $G_b$and Ga have different strategies and use different criteria to evaluate reconstruction performance, and neither is better than the other. For this reason, G_c is used to help G_b and Ga cooperate more efficiently and improve reconstruction efficiency. G_c takes the bleached images of G_b and Ga as input and generates a synthesized image whose PSNR is close to that of G_b and details are recovered more accurately than that of G_a.

3) Multi-scale GANs (MSGANs): As shown on Fig. 7 [20], different generator architectures were chosen for G_0 and $G_{1...n}$ because the tasks of generating low-resolution whole images and high-resolution patches differ in a number of requirements. LR GAN uses U-Net architecture, which is able to filter out many irrelevant details and generalize better due to its bottleneck. Its tendency to produce blurrier images is negligible in the context of low-resolution images. ResNet blocks were chosen to generate patches using HR GAN because they are known to produce clear results while maintaining the same resolution of the input image [20]. The risk of overkill in the absence of a bottleneck is reduced by stronger conditioning (on previous scales) and an overall higher number of patches compared to the number of images used. For discriminators $D_{0...n}$ the usual fully convolutional architecture [20] is chosen.

B. Evaluation metrics and network optimization

Let's consider popular metrics for assessing the quality of generative image models.

1) Inception Score (IS): Inception Score is a metric for automatically assessing the quality of generative image models. This metric has been shown to correlate well with human assessment of the realism of generated images from the CIFAR-10 dataset. Inception Score uses an Inception v3 network pre-trained on ImageNet and calculates statistics of the network's outputs when applied to the generated images.

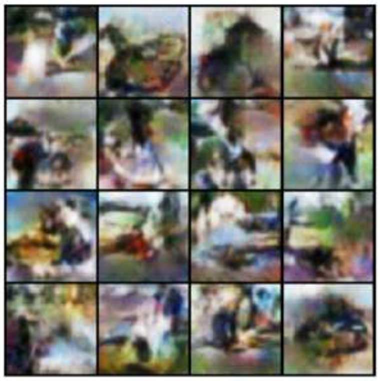

Figure 8. Examples of images with IS equal to 900.15

$IS(G) = exp(\mathbb{E}_{x \sim p_g} D_{KL}(p(y|x)||p(y)))$ where $x \sim p_g$ means that x is an image sampled from p_g, $D_{KL}(p \parallel q)$ — KL-divergence between distributions p and q, $p(y|x)$ is the conditional class distribution, $p(y) = \int_x p(y \mid x) p_g(x)$ — marginal class distribution. exp is present in the expression to make it easier to compare values, so we will ignore it and use $ln(IS(G))$ without loss of generality [23].

In other words, IS can be interpreted as a measure of the dependence between the images generated by G and the marginal distribution of classes in y. The mutual information of two random variables is also related to their entropies:

Inception Score does show a reasonable correlation with the quality and variety of generated images, which explains its widespread use in practice. However, it is not entirely correct because it only evaluates Pg as an image generation model and not its similarity to Pr. Such gross violations as mixing in natural images from a completely different distribution completely deceive the Inception Score. As a result, this may push models to simply learn sharp and varied images (or even some unfavorable noise) instead of Pr. This also applies to Mode Score. Additionally, Inception Score is unable to detect overfitting because it cannot use the validation set [24].

Applying Inception Score to generative models trained on non-ImageNet datasets produces unreliable results. Most often, Inception Score is used for generative models trained not on ImageNet sets, but on CIFAR-10, since it is slightly smaller and more convenient for training than ImageNet.

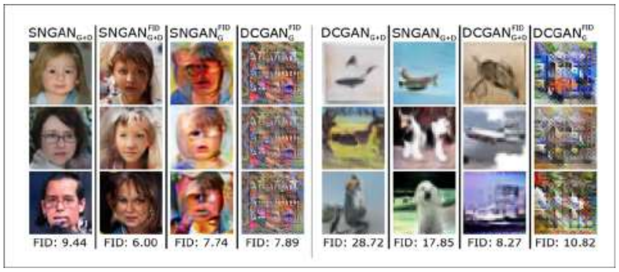

Figure 9. Example of the FID manipulation problem

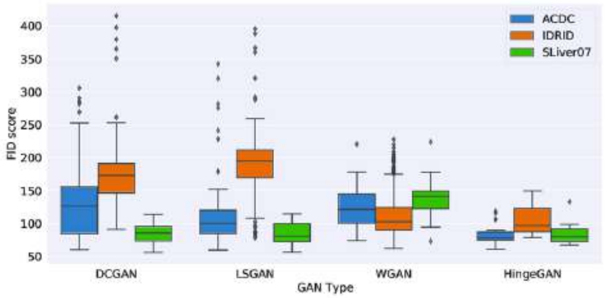

Figure 10. FID score for different GAN types

2) Frechet Inception Distance (FID): Frechet Inception Distance, or FID for short, is a metric for assessing the quality of generated images, specifically designed to evaluate the performance of generative adversarial networks. This metric was proposed as an improvement to the existing Inception Score, or IS. The Inception Score evaluates the quality of a collection of synthetic images based on how well the best Inception v3 image classification model classifies them as one of 1,000 known objects [22]. The scores combine both the confidence in the conditional class predictions for each synthetic image (quality) and the integral of the marginal probabilities of the predicted classes (diversity). The inception score does not reflect the comparison of synthetic images with real ones.

The FID score was designed to evaluate synthetic images based on the statistics of synthetic images datasets compared to the same statistics of real images from the goal distribution. As with inception estimation, FID estimation uses the inception v3 model. In particular, the model's encoding layer (the last pooling layer before output image classification) is used to capture specific features of the image, they will be generalized as a multivariate Gaussian, computing the mean and covariance of the images. These statistics are then calculated correspondingly for the collection of real and generated images.

However, it should be noted that this metric cannot fully evaluate images on an equal level with a person and it does not fully comply with human standards for image evaluation [22].

3) LL computation: As a high test LL corresponds to a low KL divergence between the generative distribution and the genuine data distribution, many generative models, aside from GANs, employ it as an additional evaluation variable. When decoder is not created to be reversible log-likelihood estimation in decoder-based models is typically intractable ([25], [26]). For GANs it is not obvious how to compute a good lower bound, unlike in the case of VAE-based models. Even when lower bounds are available, they have only been calculated in relatively few studies ([9], [27]). LL has received little attention and is never utilized specifically in GANs. [25]

Monte Carlo techniques such as AIS and nonparameteric density estimation methods such as KDE get around this by assuming a Gaussian observation model $p\theta(x|z)$ for the generator. In particular, [28] demonstrated that LL of GANs may be precisely approximated via annealed significance sampling. The AIS implementation in this study adheres to the [29] approaches. We applied a Metropolis-Hastings (MH) adjustment and 10 leapfrog steps with the HMC transition operator. The optimal 65 percent acceptance rate for HM [29] was used to tune the HMC. AIS algorithm used 16 independent chains, approximately 8000 intermediate distributions. It was found that AIS is precise enough to enable thorough comparisons between generative models for the majority of the models we looked at.

IV. Results

We used AIS to estimate log-likelihoods for all models under consideration, also was calculated FID and Inception metrics for all GANs specified at tables below. As the result, the log-likelihood assigned to the GAN does not correlate FID or Inception scores. As GAN tend to favor images with larger regions, with less local variance, there is little to no correlation between actual distribution of GAN generated images and common statistics approaches. All AIS results from represented GANs are listed in the table I. All FID and Inception results, if calculated from represented GANs, are listed in the table below. Examples of generated images from multi-layer GANs are shown in the figure 11.

V. Discussion

Training a GAN requires significant hyperparameter tuning and powerful computing resources. With the OSTIS system, it has become possible to automate the selection of the optimal neural network architecture and protects the core of the intelligent system from the formation of redundant approaches and confusing terms.

Neural network models as components of the ostis system are accelerated by reducing the number of parameters used. In the case of optimization and complexity reduction, it is sufficient to determine the maximum upper value of the number of neurons of each layer without the need to select these parameters during a

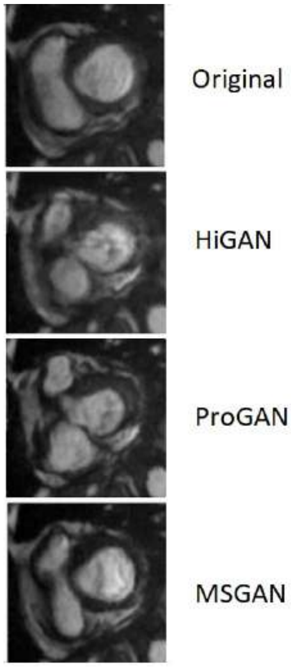

Figure 11. Examples of synthetic images from different nets

series of experiments. Deep models can often be slow and resource-intensive, which impacts overall system performance, leading to severe lags, especially in the absence of efficient and powerful computers and video adapters.

However, the trained model may not produce the desired result when considering the subsequent tasks for which the generated data is intended. Even if ProGAN or MSGAN achieve better results, they can produce unreliable images. There is no objective way to evaluate whether a medical image actually involves in gathering data responsible for the diagnosis with GAN approach without involvement of a medical expert. Due to the smaller scale of medical datasets not every GAN architecture is adapted to capture specific features of medical images.

The OSTIS technology promotes and supports integration of various neural network models. This system helps user achieve compatibility between different AI systems, effectively use knowledge to create code-based transcriptions and graphical interpretations of models. [2]

The OSTIS system also simplifies network development, expands functionality by using and reusing existing designs. This technology is capable to summarize results of a GAN training and testing and store different model versions.

VI. Conclusions

Table I
AIS ACDC results

(Nats)	AIS Test	AIS Train
DCGAN	448.43±7.40	447.23±2.24
HiGAN	349.17±4.51	546.73±5.08
ProGAN	571.64±10.22	721.37±8.76
MSGAN	651.65±9.10	732.42±2.43

Table II
AIS ACDC results.
Evaluation for half of 2D slices

(Nats)	AIS Test	AIS Train
DCGAN	348.25 ± 6.32	444.13 ± 2.24
HiGAN	294.21 ± 7.52	446.73 ± 9.38
ProGAN	571.64 ± 10.22	652.50 ± 7.5
MSGAN	551.65 ± 9.74	732.42 ± 2.43

As a result of the literature review, it was concluded that the realism of the generated GAN images largely depends on the distribution of the original training data and how similar the output results are to the training ones. Viewing different parameters at different levels allows you to determine the most optimal ones for training multi-level GANs.

It was determined that multi-scale GANs are characterized by sensitivity to changes in hyperparameters, and some multi-scale GANs are hypersensitive to such changes. The main advantages for integrated computer ostis-systems that appear when using reduction as a way to reduce the dimensionality of neural networks are determined. The authors see the further development of the proposed approach in obtaining practical results for known deep architectures of models used to solve problems of computer vision and natural language processing.

The multi-layer GAN architecture coupled with a progressive learning method allows you to better perform with medical dataset, showing improved results in popular evaluation metrics such as FID and Inception. Having considered the problem of recursive learning, we can conclude that it is impossible to obtain a real data distribution using GSN. And when using generated

Table III
AIS IDRiD results

(Nats)	AIS Test	AIS Train
DCGAN	343.11±6.07	352.45±4.7
HiGAN	289.96±5.22	498.21±9.32
ProGAN	434.71±12.54	551.15±6.13
MSGAN	532.23±6.56	678.02±15.21

Table IV
FID and Inception results

Score	FID (IDRiD)	Inception (IDRiD)
DCGAN	77.34	-
HiGAN	66.54	-
ProGAN	27.21	721.10
MSGAN	32.8	832.82
	FID (ACDC)	Inception (ACDC)
DCGAN	59.34	-
HiGAN	46.54	-
ProGAN	47.21	755.16
MSGAN	37.8	712.45

images to train networks, it leads to complete degradation of the neural network. This problem can only be solved with the help of human control of neural networks. Another problem with CT scans data, which is obtained in 3D: common GANs may not produce a good quality image if trained solely on 2D images. This makes the study of GANs specifically built for medical data an interesting avenue of research and could lead to improvements in quality and ultimately clinical usability. Thus, we can conclude that it is impossible to obtain a real pseudo-real image of a medical nature without adjusting their structure, like in multi-level discriminators, and evaluate proper metrics to correctly estimate generated distribution to make reliable synthetic data.

References

[1] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath Generative adversarial networks: An overview. *IEEE signal processing magazine*, 2018, vol. 35, no 1, pp. 53-65.

[2] V. V. Golenkov, N. A. Gulyakina, and D. V. Shunkevich, *Otkrytaya tekhnologiya ontologicheskogo proektirovaniya, proizvodstva i ekspluatatsii semanticheski sovmestimykh gibridnykh intellektual'nykh komp'yuternykh sistem [Open technology of ontological design, production and operation of semantically compatible hybrid intelligent computer systems]*. Minsk, Bestprint, 2021, (In Russ.).

[3] (2024, Feb) Ostis applications. [Online]. Available: https://github.com/ostis-ai/sc-web

[4] (2024, Feb) OSTIS. [Online]. Available: https://github.com/ostis-ai

[5] M. Kovalev. Versioning model of neural network problem-solving methods in intelligent systems *Open Semantic Technologies for Intelligent Systems (OSTIS)*, 2023, Vol. 7, pp. 121–126.

[6] I. Shumailov, Z. Shumaylov, Y. Zhao, Y. Gal, N. Papernot and R. Anderson. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv*:2305.17493, 2023.

[7] M. E. Abbasnejad, Q. Shi, A. V. D. Hengel and L. Liu A generative adversarial density estimator. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10782-10791

[8] Ethan Fetaya, Jörn-Henrik Jacobsen, Will Grathwohl, and Richard Zemel.Understanding the limitations of conditional generative models.*arXiv preprint arXiv*:1906.01171, 2019.

[9] Y. Wu, Y. Burda, R. Salakhutdinov, and R. Grosse. *On the quantitative analysis of decoder-based generative models. arXiv preprint arXiv*:1611.04273, 2016.

[10] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, et al. Deep Learning Techniques for Automatic MRI Cardiac Multi-structures Segmentation and Diagnosis: Is the Problem Solved? *IEEE Transactions on Medical Imaging*. 2018. Vol. 37, no. 11, pp. 2514-2525.

[11] P. Porwal, S. Pachade, R. Kamble, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, F. Meriaudeau. Indian Diabetic Retinopathy Image Dataset (IDRiD). *IEEE Dataport*. 2018.

[12] Y. Skandarani, P. M. Jodoin and A. Lalande Gans for medical image synthesis: An empirical study. *Journal of Imaging*, 2023, vol. 9, no 3, pp. 69.

[13] A. Öcal and L. Özbakır Supervised deep convolutional generative adversarial networks. *Neurocomputing*, 2021, no 449, pp. 389-398.

[14] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 2014, vol. 15, no 1, pp. 1929-1958.

[15] V. Wilmet, S. Verma, T. Redl, H. Sandaker and Z. Li A Comparison of Supervised and Unsupervised Deep Learning Methods for Anomaly Detection in Images. *arXiv preprint arXiv*:2107.09204, 2021.

[16] T. Karras, T. Aila, S. Laine and J. Lehtinen Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv*:1710.10196, 2017.

[17] J. He, X. Li, N. Liu and S. Zhan Conditional generative adversarial networks with multi-scale discriminators for prostate MRI segmentation. *Neural Processing Letters*, 2020, vol. 52, no 2, pp. 1251-1261.

[18] U. Upadhyay, Y. Chen, T. Hepp, S. Gatidis and Z. Akata (2021) Uncertainty-guided progressive GANs for medical image translation. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24 (pp. 614-624). *Springer International Publishing*.

[19] Q. Chen and V. Koltun Photographic image synthesis with cascaded refinement networks. In Proceedings of the IEEE international conference on computer vision, 2017, (pp. 1511-1520).

[20] H. Uzunova, J. Ehrhardt, F. Jacob, A. Frydrychowicz and H. Handels (2019). Multi-scale gans for memory-efficient generation of high resolution medical images. *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17*, 2019, Proceedings, Part VI 22, pp. 112-120. *Springer International Publishing*.

[21] D. M. Vo, D. M. Nguyen, T. P. Le and S. W. Lee HI-GAN: A hierarchical generative adversarial network for blind denoising of real photographs. *Information Sciences*, 2021, 570, 225-240.

[22] Y. Yu, W. Zhang and Y. Deng Frechet inception distance (fid) for evaluating gans. *China University of Mining Technology Beijing Graduate School*, 2021.

[23] S. Barratt and R. Sharma A note on the inception score. *arXiv preprint arXiv*:1801.01973, 2018

[24] Q. Xu, G. Huang, Y. Yuan, C. Guo, Y. Sun, F. Wu and K. Weinberger An empirical study on evaluation metrics of generative adversarial networks. *arXiv preprint arXiv*:1806.07755, 2018

[25] L. Dinh, D. Krueger, and Y. Bengio. Nice: non-linear independent components estimation. *arXiv preprint arXiv*:1410.8516, 2014.

[26] L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using Real NVP. *arXiv preprint arXiv*:1605.08803, 2016.

[27] L. Theis, A. van den Oord, and M. Bethge. A note on the evaluation of generative models. *arXiv preprint arXiv*:1511.01844, 2015.

[28] R. B. Grosse, Z. Ghahramani, and R. P Adams. Sandwiching the marginal likelihood using bidirectional monte carlo. *arXiv preprint arXiv*:1511.02543, 2015.

[29] Neal, R. M. MCMC using Hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2011, Vol. *2*, no 11, pp. 2

МЕТРИКИ ОЦЕНКИ И ПРИМЕНЕНИЕ МНОГОУРОВНЕВЫХ GAN ДЛЯ МЕДИЦИНСКИХ ИЗОБРАЖЕНИЙ

Ковбаса Г. А.

В данной статье были рассмотрены методы использования GAN в медицине, их перспективы, а также проблемы в обучении генеративно-состязательных сетей, связанные с увеличением использования сгенерированных изображений для обучения других сетей. Анализ одноуровневых и многоуровневых GAN пришел к выводу, что, хотя по статистике многоуровневые GAN работают лучше, они не совсем соответствуют распределению исходного набора данных, и без наблюдения медицинского специалиста такие синтетические данные не следует использовать при обучении новых сетей. Рассмотрены проблемы, связанные с явлением рекурсивного обучения, предвзятыми оценками реалистичности изображений и неоптимизированными структурами. Подход описан в контексте интеграции моделей генеративно-состязательных сетей в гибридные компьютерные системы на основе технологий OSTIS.

Received 13.03.2024