

ANALYSIS OF METRICS FOR GAN EVALUATION

Generative-adversarial networks have become quite popular in recent years. In general, these networks are based on convolutional neural networks used in classification problems. In recent years, researchers have proposed and developed many variations of GAN network architectures and techniques for their optimization, as the learning process is quite complex and unstable. Despite great theoretical advances in improving network data, evaluating and comparing GANs remains a challenge. Although several metrics have been introduced to evaluate these networks, there is currently no consensus on which metrics best reflect the strengths and limitations of models and should be used to compare models and evaluate synthesized images. This paper discusses the two most popular metrics, Inception Score (IS) and Frechet Inception Distance (FID), which are used to estimate GAN networks.

Because these metrics are based on a pre-built Google Inception model used as a classifier for IS metrics and a feature extractor for FID metrics, the goal is to develop a program module to compare metric data using the base model (Inception) and custom models.

The scientific novelty is that these metrics were first used to compare cytological images using a model different from the one proposed by the authors - Google Inception.

The practical significance of the work is the development of a software module for calculating metric data for GAN networks used for the synthesis of cytological images.

As a result, two basic models (BioCNN-1 and BioCNN-2) and a Python module for calculating IS and FID metrics for cytological images were developed. The developed module works with color images with a resolution of 64 x 64 pixels. Comparisons of metrics based on the base model and the developed models for estimating GAN networks for cytological image synthesis were compared.

It was shown that the metrics based on the developed models show better results. The FID score reduced from 31.20 to 0.034 and the IS score increased from 3.52 to 3.81. A total metric calculation time reduced from 2 minutes to 15 seconds.

Keywords: GAN evaluation, metrics, inception score, frechet inception distance.

Петро ЛЯЩИНСЬКИЙ, Павло ЛЯЩИНСЬКИЙ
Західноукраїнський національний університет

АНАЛІЗ МЕТРИК ДЛЯ ОЦІНКИ GAN МЕРЕЖ

Генеративно-змагальні мережі стали досить популярними в останні роки. Загалом ці мережі побудовані на основі згорткових нейронних мереж, що застосовуються у завданнях класифікації. В останні роки дослідниками запропоновано та розроблено дуже багато варіацій самих архітектур GAN мереж та технік для їх оптимізації, оскільки процес навчання є досить складним та нестабільним. Незважаючи на великі теоретичні успіхи в покращенні даних мереж, оцінка та порівняння GAN залишається складним завданням. Не дивлячись на те, що було введено кілька метрик для оцінки цих мереж, наразі немає консенсусу щодо того, яка метрика найкраще відображає сильні сторони та обмеження моделей і повинна використовуватися для порівняння моделей та оцінки синтезованих зображень. У даній роботі розглянуто дві найпопулярніші метрики Inception Score (IS) та Frechet Inception Distance (FID), які застосовуються для оцінки GAN мереж.

Оскільки дані метрики базуються на використанні попередньо підготовленої моделі Google Inception, яка застосовується в якості класифікатора для метрики IS та екстрактора ознак для метрики FID, то метою роботи є розробка програмного модуля для порівняння даних метрик із використанням базової моделі (Inception) та користувацьких моделей.

Наукова новизна полягає в тому, що дані метрики вперше застосовано для порівняння цитологічних зображень з використанням моделі, що відрізняється від запропонованої авторами - Google Inception.

Практичним значенням роботи є розробка програмного модуля для обчислення даних метрик для GAN мереж, що застосовуються для синтезу цитологічних зображень.

В результаті було розроблено дві базові моделі (BioCNN-1 та BioCNN-2) та модуль на мові Python для обчислення метрик IS та FID для цитологічних зображень. Розроблений модуль працює із кольоровими зображеннями роздільною здатністю 64 x 64 пікселі. Здійснено порівняння метрик на основі базової моделі та на основі розроблених моделей для оцінки GAN мереж для синтезу цитологічних зображень.

Метрики на основі розроблених моделей показують кращі результати. Значення метрики FID зменшилося з 31.20 до 0.034, а значення метрики IS збільшилося з 3.52 до 3.81. Також загальний час обчислення метрик зменшився з 2 хвилин до 15 секунд.

Ключові слова: оцінка GAN мереж, метрики, inception score, frechet inception distance.

Introduction

In 2014, a completely new approach for image synthesis using generative adversarial networks (GAN) was invented [1]. After that, a lot of new architectures were proposed [2,3,4]. Despite the fact that a significant amount of research studies are focused mainly on the theory behind GANs, currently there are a few studies that are related to the evaluation of GAN networks [5]. The purpose of such evaluation is to measure the distance between synthesized and real images. Most existing methods use the initial Inception model to represent images in a lower dimensional space. The most popular metric at the moment is the Inception Score (IS), which measures the distance using Kullback-Leibler divergence (KL) [5]. However, this metric is based on the probability of an image belonging to one of the classes and cannot show the model overfitting. Frechet Inception Distance metric is proposed as a

better alternative. This metric directly measures the Frechet distance on a feature space by approximating a single-varying Gaussian distribution.

Since these metrics are based on a pre-trained Inception model, then their values might degrade when applied to other datasets that differ from ImageNet (this dataset was used to train the Inception model). Accordingly, an urgent problem is the development of basic user models for IS and FID metrics for a specific dataset, which will allow improving the value of these metrics.

Related works

Comparing how similar two images can be is a common problem in image analysis. For this task, a variety of metrics are used.

A metric is a specific function of the distance between any two components of a collection. A metric function has to conform to three axioms. The metric has to meet the triangle inequality and be identical and symmetric. There are two types of metrics: qualitative and quantitative. Quantitative measurements are the most often utilized metrics in research [6, 7]. Qualitative metrics are metrics that are not numerical and often involve a person's subjective evaluation or evaluation by comparison. The most popular methods are Nearest Neighbors (similar images are grouped into clusters) and Rapid Scene Categorization [8]. The last one is that the experts have to make a choice between a real and a synthesized image in a short period of time. The main disadvantage of the approach based on expert evaluations is that experts can improve their skills over time [9]. For example, experts can receive feedback from other experts and receive tips on how to better detect the synthesized image.

Quantitative metrics are based on the calculation of specific numerical scores that are used to summarize the quality of synthesized images. In [10], researchers refer to such metrics as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR), Coverage Metric, Inception Score, FID and others.

To evaluate images synthesized using GAN networks, researchers have developed several metrics that can be divided into model-dependent and model-independent. Model-dependent metrics usually require either an estimate of the distribution density or an analysis of the internal structure of the network used. So model-independent metrics are more popular in GAN researches [11]. Most of the independent metrics map the image to the feature space using a pre-trained model and measure the similarity of the distribution between the used dataset and the synthesized images.

Among all metrics, Inception Score (IS) and Frechet Inception Distance (FID) are the most popular and relevant metrics for evaluating the quality of images synthesized using GAN networks [12, 13]. It is necessary to perform a detailed analysis of these metrics, since they have proven themselves quite well in many studies and have shown a good correlation with experts' assessments.

Metrics overview

Inception Score. This metric is based on the Google Inception V3 image classification model. This model is designed to classify color images. The ImageNet dataset, which includes about 1.2 million RGB images divided into 1000 classes, was used as a training dataset.

This metric showed good correlation with human-made estimates on the CIFAR-10 dataset.

$$IS(G) \approx \exp(E_{x \sim p_g}[D_{KL}(p(y|x} || p(y))]) = \exp(H(y) - E_{x \sim p_g}[H(y|x)]),$$

where E – expected value,

$x \sim p_g$ shows that x is an image synthesized from the distribution p_g (*distribution of the generator*),

D_{KL} is the Kullback-Leibler divergence between the conditional probability distribution $p(y|x)$ and marginal distribution $p(y) = E_{x \sim p_g}[p(y|x)]$,

H – entropy.

It is assumed that the conditional distribution of data, which contains significant objects, should have low entropy, and the marginal distribution (synthesized images are diverse) should have high entropy [11].

Inception Score works as follows. For example, let's take 5000 synthesized images. In order to obtain a conditional distribution of classes, it is required to classify the image data with the Inception network, which will return a vector of probabilities $p(y|x)$. In order to obtain the marginal distribution, the conditional distribution for each image should be summarized as follows $p(y) = \frac{1}{5000} \sum_{i=1}^{5000} p(y|x_i)$. Next step is to calculate the Kullback-Leibler distance between the conditional distribution of each synthesized image and the overall marginal distribution. The average value of these distances will be the value of the IS metric [12].

Therefore, IS measures the average Kullback-Leibler divergence between the conditional distribution $p(y|x)$ and the marginal class distribution $p(y)$. That is, this metric does not consider the distribution of the original samples at all, and therefore cannot assess how well the images synthesized by the generator are similar to the original samples. This metric evaluates only images diversity. The disadvantages of this metric are sensitivity to the resolution of the images themselves and to changes in the network, which is used for classification.

The minimum value of this metric is 1, and the maximum value is the number of classes that the Inception network can classify. In this case – 1000.

In order to obtain a high IS value, it is necessary that the synthesized images contain clear objects (for example, the images are not blurred) and that the generator synthesizes a variety of images from all classes [13]. Accordingly, if at least one of these conditions is unsatisfactory, the score will be low.

Frechet Inception Distance. FID compares the distributions of the original and synthesized data. In order to calculate the FID between real and synthesized images the data is transformed into a feature space using a specific layer of the Inception model, namely the *pool3* layer. Feature space is used to represent images in a lower dimensional space where similar images are represented in relatively same regions. At the output, we receive activation maps (also known as feature maps). FID metric assumes that these feature maps can be approximated using two Gaussian distributions. Then the distance between them is calculated as follows:

$$d^2((m_r, C_r), (m_g, C_g)) = \|m_r - m_g\|^2 + \text{Tr}(C_r + C_g - 2(C_r C_g)^{\frac{1}{2}}),$$

where (m_r, C_r) та (m_g, C_g) – average value and covariance matrix of the real and synthesized data distributions, respectively,

Tr – trace of the matrix (the sum of the diagonal elements).

The lower the value of the metric, the smaller the distance between the distributions is. Therefore, the distributions are more similar to each other [14]. The FID metric is quite sensitive to image distortions (rotation, displacement, shift, noise, etc.). The more distortions, the greater the value of the metric is [15].

A low FID value indicates that the distributions of real and synthesized images are similar to each other. However, in practice, if a model has a low FID value, it indicates that the images are of high quality or diversity, or both. This behavior can significantly complicate the diagnosis of the model.

The authors also show that this metric more closely matches human estimates and is more robust to noise than IS [11, 16].

These metrics are quite popular in the field of image synthesis using GAN networks. But they have their drawbacks [17, 18, 19].

Inception Score has the following limitations:

- 1) The value of the metric strongly depends on what the Inception model can classify.
- 2) Synthesis of images of a different set of classes that are not present in the original ImageNet dataset may cause a low IS value.
- 3) If the classifier cannot identify the features that belong to the training dataset, then low-quality images may receive high scores. The Inception network is trained on the ImageNet dataset. If IS is used on a completely different dataset, then the classifier may not be able to identify some features well enough, and therefore low-quality images will receive high scores.

Frechet Inception Distance is also based on the Google Inception model. But unlike IS, this metric can define dependencies between classes. That is, if the model generates only one image per class, then the IS can be quite high, but the FID will be low. Also, the FID metric degrades when various artifacts are added to the image.

The Inception Score does show a correlation with the quality and variety of images produced, which explains its widespread use in practice. However, this metric only evaluates the distribution of the synthesized images, but does not take into account how similar the synthesized and original images are. As a consequence, this may induce models to simply learn distinct and varied images (or even some noise) instead of the distribution of the original data [13].

Inception Score is limited to measuring how diverse the synthesized images are, while FID measures the distance between the distribution of synthesized and real data [14].

IS and FID calculation based on custom classification model for biomedical images

Since both metrics are based on the Inception model to obtain conditional probabilities (IS metric) and feature maps (FID metric), this can significantly affect the results when calculating these metrics for data that is not included in the ImageNet dataset on which the Inception network was trained.

A classifier architecture for biomedical images was developed, which ensures obtaining more relevant conditional probabilities for the IS metric and activation maps for the FID metric, in order to compare the values of the IS and FID metrics calculated using the Inception model and metrics calculated using a different model.

Both networks take as input color images of size 64 by 64 pixels according to the resolution of the images in the training dataset and are named BioCNN-1 and BioCNN-2.

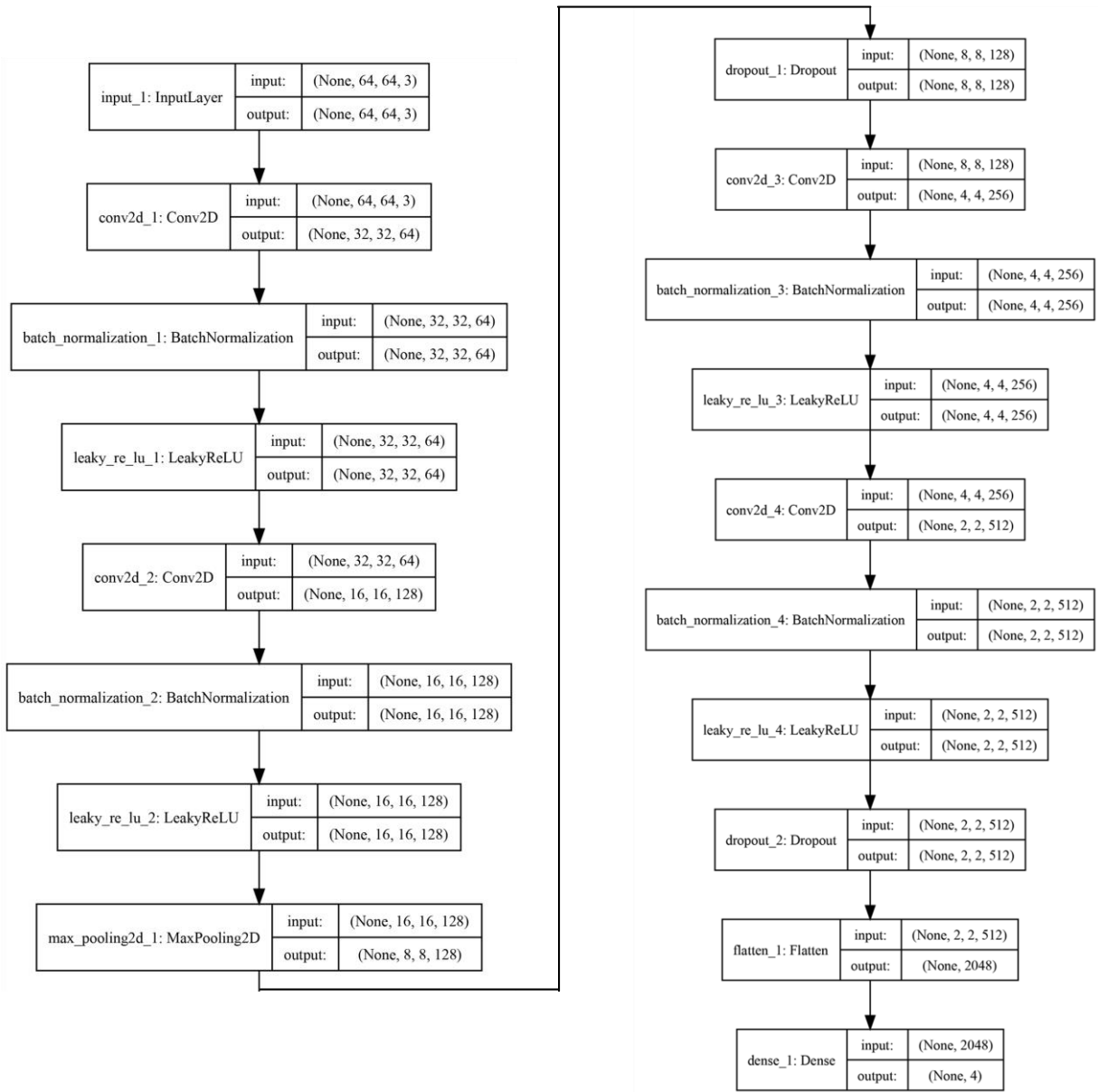


Fig. 1. Architecture of BioCNN-1

These networks are convolutional neural networks (CNNs). This type of networks is widely used in classification and pattern recognition tasks [20]. The BioCNN-1 architecture consists of a sequence of Conv, BatchNorm, and LeakyRelu activation layers. One set of these layers can be called a convolution block. BioCNN-1 consists of four such blocks.

The BioCNN-2 architecture is built using alternating VGG and ResNet blocks. These blocks are separate elements of the architecture of popular convolutional neural networks VGG and ResNet, respectively [22-25].

In general, VGG consists of a sequence of convolutional layers using a small convolutional window size (3 by 3). A subsampling (pooling) layer is placed at the end of such a block.

The ResNet block consists of two convolutional layers with the same number of filters, where the output of the second layer is added to the input of the first.

In the future, the architectures can be improved by optimizing hyperparameters, which is described in [21].

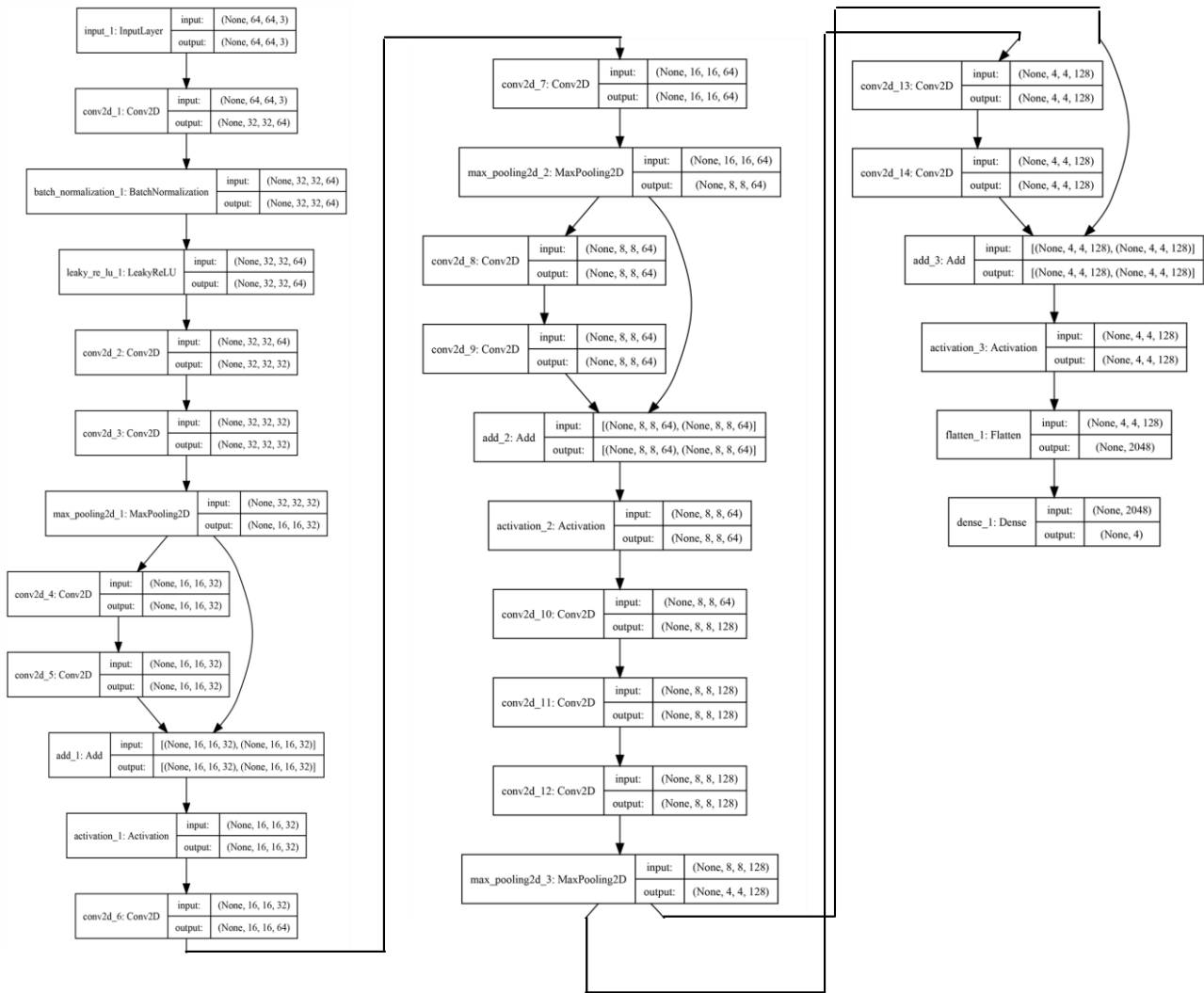


Fig. 2. Architecture of BioCNN-2

Experiments

For performing experiments, an artificial set of cytological images with a size of 64 by 64 pixels was synthesized using the GAN network [26]. Cytological images are a subset of biomedical images, which are structural and functional images of human organs and are intended for the diagnosis of diseases [27]. In general, biomedical images can be divided into three groups: cytological (images of cells), histological (images of tissues), and immunohistochemical (images of cells and their reactions and specific markers) [28, 29]. Examples of cytology images from the original and synthesized samples are shown in the figures below.

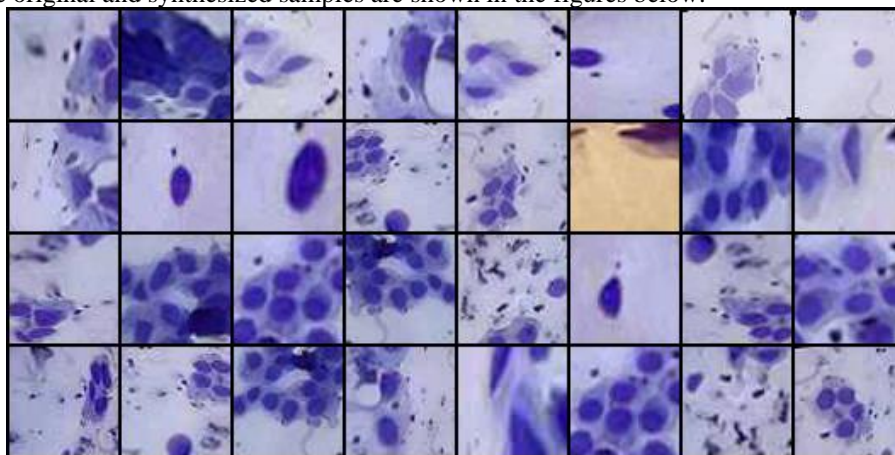


Fig. 3. Real images

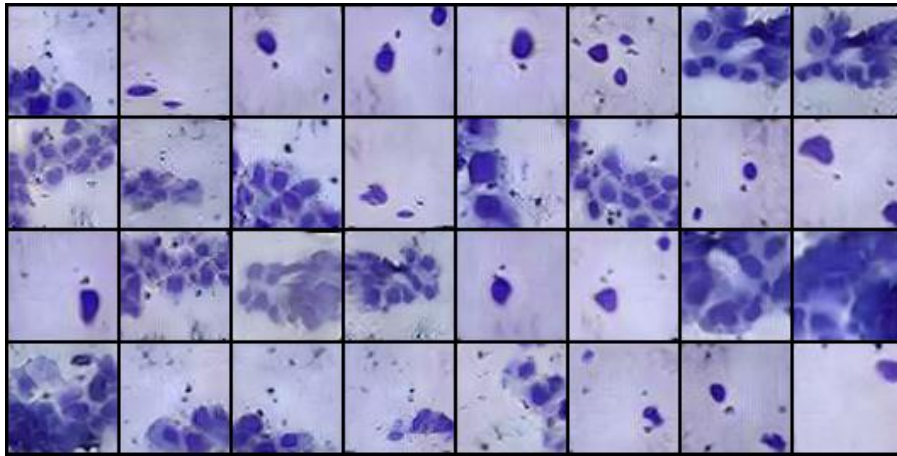


Fig. 4. Synthesized images

IS and FID metrics were used to compare the synthesized images with the original ones. To calculate the metrics based on the custom classifier, the proposed CNN architectures of the BioCNN-1 and BioCNN-2 networks are applied. To build models, train them, and calculate IS and FID metrics, a software module was developed in the Python programming language using the Keras machine learning framework. The experiments were performed on a laptop with an Intel Core i7 2.5GHz CPU and 16GB of RAM. The hyperparameters of training are listed in Table 1.

Table 1

Training parameters

Model name	Loss function	Optimizer	Learning rate	Batch size	Epochs
BioCNN-1	categorical_crossentropy	Adam	0.003	128	40
BioCNN-2	categorical_crossentropy	Adam	0.003	64	100

A sample of color cytological images divided into 4 classes with a total number of approximately 4500 images (resolution of 64 by 64 pixels) was used as a training dataset. This dataset was divided in the ratio of 80-10-10 as a training, test and validation dataset. BioCNN-1 network achieved classification accuracy of 97% and BioCNN-2 - 98.8%. The training time of the first network was approximately 15 minutes, and the second network took 45 minutes. The second network needs more time to train because its architecture is deeper. The ROC curves for both networks are shown in the figures below.

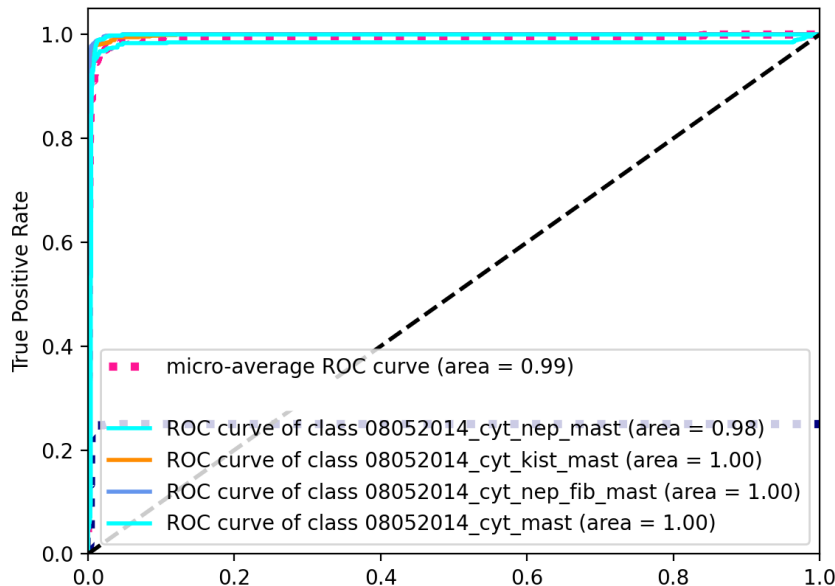
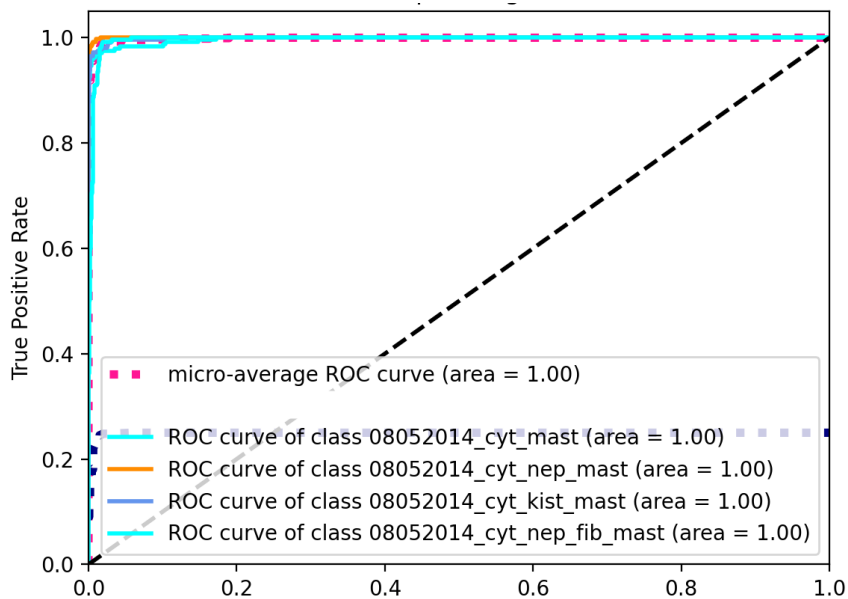


Fig. 5 ROC for BioCNN-1



After training both networks, the values of IS and FID metrics were calculated to compare the validation dataset with the synthesized images. To obtain the activation maps used in the FID metric, the fourth layer from the end (*leaky_re_lu_4*) of the BioCNN-1 model and the third layer from the end (*activation_3*) of the BioCNN-2 model were taken. The summarized results are given in Table 2.

Table 2

IS and FID scores

Inception Score, higher is better	Frechet Inception Distance, lower is better	Classification model	Total metric calculation time
3.52	31.20	Google Inception V3	~ 2 minutes
3.64	23.41	BioCNN-1	~ 8 seconds
3.81	0.034	BioCNN-2	~ 15 seconds

Discussion

As a result of the experiments, it is shown that the value of the metrics has improved when applying the developed models. There is a slight improvement in the IS metric. This indicates that the IS metric is not so dependent on the model used. The reason for this is that this metric is calculated based on the probabilities of an image belonging to one of the classes. The theoretical explanation is that similar images will be assigned to the same class regardless of the model used. However, the use of custom models did improve the IS metric, as the custom model classifies cytological images better than the Inception model.

When the BioCNN-1 model was used to calculate the FID metric compared to the Inception model, the FID value decreased from 31.20 to 23.41. However, when using the BioCNN-2 model, the metric value decreased to 0.034. To calculate this metric, feature maps obtained from a specific layer of the base model are used. The improvement of the metric values when applying the developed models indicates that the developed models provide more relevant feature maps for cytological images, since they were trained on images from this domain.

The significant difference between the values of the FID metric when using BioCNN-1 and BioCNN-2 can be explained by the architectural details of the networks themselves. Despite the fact that both networks achieved approximately the same classification accuracy on the test dataset during training, the second network is much deeper than the first. During the experiments, we also noticed a tendency for the FID value to increase significantly as the layer used as a feature extractor approaches the network input. The BioCNN-2 network demonstrates this trend in a less pronounced manner.

The fact that there is a significant difference in the FID value when using the developed networks, considering that these networks were trained on the same dataset, suggests that the deeper network (BioCNN-2) can represent the input image much better in a low-dimensional space, leading to more relevant and "informative" feature maps. In contrast to the IS metric, the FID metric is thus considerably dependent on the network utilized as a feature extractor.

Conclusions

The main results of this work are:

1. A comparison of IS and FID metrics was made for evaluating GAN networks for the synthesis of cytological images using the basic Inception model and the developed BioCNN-1 and BioCNN-2 models.

2. A Python module was developed to calculate IS and FID metrics for cytological images using the developed models.
3. The usage of the developed models, as opposed to the Inception network, greatly reduces the time required to calculate these metrics, according to actual experiments. The calculation took 15 seconds instead of 2 minutes.
4. Significant reduction in the calculation time and improvement in the values of the metrics themselves makes it possible to develop this study in the direction of using the FID metric as an additional parameter in the GAN network loss function, which would theoretically improve the quality of synthesized images.

References

1. Goodfellow I. J., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y. Generative Adversarial Networks. arXiv, 2014.
2. Radford A., Metz L., Chintala S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *ICLR*. 2016.
3. Arjovsky M., Chintala S., Bottou L. Wasserstein Generative Adversarial Networks. PMLR, 2017.
4. Berthelot D., Schumm T., Metz L. BEGAN: Boundary Equilibrium Generative Adversarial Networks. arXiv, 2017.
5. Theis L., Oord A. van den, Bethge M. A note on the evaluation of generative models. arXiv, 2016.
6. Im D. J., Ma H., Taylor G., Branson K. Quantitatively Evaluating GANs With Divergences Proposed for Training. arXiv, 2018.
7. Che T., Li Y., Jacob A. P., Bengio Y., Li W. Mode Regularized Generative Adversarial Networks. arXiv, 2017.
8. Kontorovich A., Sabato S., Umer R. Active Nearest-Neighbor Learning in Metric Spaces. arXiv, 2018.
9. Dziugaite G. K., Roy D. M., Ghahramani Z. Training generative neural networks via Maximum Mean Discrepancy optimization. arXiv, 2015.
10. Lopez-Paz D., Oquab M. Revisiting Classifier Two-Sample Tests. arXiv, 2018.
11. Salimans, Tim, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, i Xi Chen. «Improved Techniques for Training GANs». B *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2016. URL: <https://proceedings.neurips.cc/paper/2016/file/8a3363abe792db2d8761d6403605aeb7-Paper.pdf>.
12. Barratt S., Sharma R. A Note on the Inception Score. arXiv, 2018. URL: <http://arxiv.org/abs/1801.01973>
13. Xu Q., Huang G., Yuan Y., Guo C., Sun Y., Wu F., Weinberger K. An empirical study on evaluation metrics of generative adversarial networks. *arXiv:1806.07755 [cs, stat]*. 2018.
14. Borji A. Pros and Cons of GAN Evaluation Measures. *arXiv:1802.03446 [cs]*. 2018. URL: <http://arxiv.org/abs/1802.03446>
15. M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, in: *Advances in Neural Information Processing Systems*, 2017, pp. 66296640.
16. Shmelkov K., Schmid C., Alahari K. How good is my GAN? arXiv, 2018. URL: <http://arxiv.org/abs/1807.09499>
17. Lucic M., Kurach K., Michalski M., Bousquet O., Gelly S. Are GANs created equal? a large-scale study. Red Hook, NY, USA:Curran Associates Inc., 2018.
18. Doan K. D., Manchanda S., Wang F., Keerthi S., Bhowmik A., Reddy C. K. Image Generation Via Minimizing Fréchet Distance in Discriminator Feature Space. *arXiv:2003.11774 [cs, eess]*. 2020.
19. Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z. Rethinking the Inception Architecture for Computer Vision. arXiv, 2015.
20. Manziuk E., Skrypyk T., Hirnyi M. (2020). Determination of recipes constituent elements based on image. *Computer Systems and Information Technologies*, (1), 42-46. <https://doi.org/10.31891/CSIT-2020-1-5>
21. Radiuk P. (2020). An approach to accelerate the training of convolutional neural network by tunin the hyperparameters of learning. *Computer Systems and Information Technologies*, (2), 32-37. <https://doi.org/10.31891/CSIT-2020-2-5>
22. Simonyan K., Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv, 2015.
23. He K., Zhang X., Ren S., Sun J. Deep Residual Learning for Image Recognition. arXiv, 2015.
24. Krizhevsky A., Sutskever I., Hinton G. E. ImageNet Classification with Deep Convolutional Neural Networks. Curran Associates, Inc., 2012.
25. Szegedy C., Liu W., Jia Y., Sermanet P., Reed S., Anguelov D., Erhan D., Vanhoucke V., Rabinovich A. Going Deeper With Convolutions. 2015. URL: https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Szegedy_Going_Deepier_With_2015_CVPR_paper.html
26. Berezsky O. M., Liashchynskiy P. B. Comparison of generative adversarial networks architectures for biomedical images synthesis. *Applied Aspects of Information Technology*. 2021. Vol. 4, № 3. P. 250–260.
27. Berezsky O., Pitsun O., Datsko T., Derysh B., Melnyk G. Breast cancer immunohistological imaging database. *Computer systems and information technologies*. 2022. № 1. P. 75–82.
28. Berezsky, O., Melnyk, G., Datsko, T. & Verbovy, S. “An Intelligent System for Cytological and Histological Image Analysis”. Proceedings of the 13 th International Conference “The Experience of Designing and Application of CAD Systems in Microelectronics” CADSM 2015. 24-27 February 2015. Polyana-Svalyava: Ukraine. 2015. p. 28–31.
29. Berezsky, O., Verbovy, S. & Pitsun, O. “Hybrid Intelligent Information Techology for Biomedical Image Processing”. Proceedings of the IEEE International Conference “Computer Science and Information Technology” CSIT’2018. Lviv: Ukraine. 11-14 September, 2018. p. 420–423.

Petro Liashchynskiy Петро Ляшинський	PhD Student, Department of Computer Engineering, West Ukrainian National University, email: p.liashchynskiy@st.wunu.edu.ua https://orcid.org/0000-0002-3920-6239 Scopus Author ID: 57202448801, ResearcherID: CAG-1836-2022 https://scholar.google.com/citations?user=eJVS9IAAAAJ	аспірант кафедри комп'ютерної інженерії, Західноукраїнський національний університет.
Pavlo Liashchynskiy Павло Ляшинський	PhD Student, Department of Computer Engineering, West Ukrainian National University, email: pavloksmfic@gmail.com , https://orcid.org/0000-0001-8371-1534	аспірант кафедри комп'ютерної інженерії, Західноукраїнський національний університет.