The Efficiency of FedAVG Federated Learning in Training Generative Image Models

Vassili Kovalev
Dept. Biomedical Image Analysis
United Institute of Informatics Problems
Surganova St., 6, 220012 Minsk, Belarus
vassili.kovalev@gmail.com

Dzmitry Karpenka
Dept. Biomedical Image Analysis
United Institute of Informatics Problems
Surganova St., 6, 220012 Minsk, Belarus
karpenko.dima.s11@gmail.com

***Abstract*—Training neural networks often requires very large amounts of image data. However, sharing images owned by different institutions can be problematic. One of the known solutions could be to train neural networks jointly by way of iterative sharing neural models, but not the restricted sets of training images. Such an approach is known as Federated Learning. In this paper, we present results of an experimental study of the efficiency of Federated Learning of generative neural networks of the DCGAN type. Specifically, the FedAVG approach has been investigated based on large collection of medical images, including chest x-ray images, axial slices of 3D computed tomography images, and Hematoxylin-Eosin stained histology images. The results of the FedAVG approach were found to be highly dependent on the homogeneity of the image datasets. Among the images being employed, the best potential for federated training was demonstrated by chest x-ray images, while the routine histology images were found to be unsuitable for FedAVG training. The 2D computed tomography image slices were situated somewhere in-between of these two image types and showed characteristically unstable behavior. The period of aggregation of training results on the federated server should be reasonably short and repeated after every 1-3 epochs performed on the local image datasets of federated clients.**

***Keywords*—federated learning, generative models, FedAVG**

I. Introduction

Training of modern neural networks typically requires a large amount of data. However, merging and sharing a sufficient amount of the original images can be problematic even with collaborating partners . This is due to factors such as data privacy, limitations applied by the national law, conditions of past contracts, ethical self-limitations, etc. In the context of medical applications, there could be several institutions in possession of the data necessary for the training of neural networks of common interest. However, these data may not be shared by the above reasons. Federated Learning (FL) methods [1] aim to train large neural network models together, without sharing image data between participating institutions. Instead, copies of the same model are trained on local image datasets of each client, and then, at the every training round, these copies are iteratively transferred to a server which aggregates them into a joint model and sends back.

Depending on the way the training data are generated and distributed, FL approaches are subdivided into two different categories. The first is primarily referred to as the horizontal one, whereas the other is called vertical [2]. In Horizontal FL participants share the same feature space but have different local samples. The goal is to train a global model that can generalize between samples from different clients. In the scenario of Vertical FL, clients share the same aligned samples but have different local features. The goal of Vertical FL is to train a global model that is capable of making predictions using the distributed features of shared samples. Thus, this paper is deals with Horizontal, Centralized, Non-Heterogeneous (same network architecture) Federated Learning.

There are several approaches exploring the FL setup including the FedProx [3], SCAFOLD [4], the FedProc [5], and some others. In this paper, we are considering a version of the FedAVG algorithm that averages weights of convolutional filters of the neural networks. Such an approach is known and commonly abbreviated as the FedAVG [1]. The family of the FedAVG algorithms are relatively simple but still under investigation in different applications. Unlike many other works, we are focusing on medical image generation rather than on the image classification.

Despite the basic idea of the FL method is transparent, implementation schemes of every particular algorithm could be different. For instance, in their work [1], McMahan with colleagues have introduced a version of FedAVG and performed a set of experiments that confirmed that the approach is robust to the unbalanced, non-independent and identically distributed datasets. Among the options under consideration, FedAVG appears to be one of the most flexible, with the potential for easy implementation, adjustment, and modification according to the specific problem the researchers are dealing with. Nevertheless, it remains not completely clear how much it is applicable to the generative neural network models and what results it would provide for different kinds of

medical images.

In this paper, we are using the weight-averaging approach to train Deep Convolutional Generative Adversarial Networks known as DC-GANs [6], [7]. The particular goal was to assess the impact of various factors on the quality of the FedAVG-based federated training such as the type of medical images, the amount of image data, the balance of different image classes owned by the clients participating in federated training, the output image sizes, and the way of integration of training and model aggregation steps. The relative simplicity of FedAVG algorithm and low communication cost it convenient to use it as a baseline for investigation of the technology of FL for widely used family of generative models of DC-GAN type (e.g., StyleGAN, CycleGAN, Conditional GAN, Wasserstain-GAN, Super-Resolution GAN) as compared to the extensively studied classification networks.

II. Materials

The study was performed based on three different image datasets. The first radiological dataset consisted of 10,000 chest x-ray images of healthy subjects, whereas the second one included 10,000 axial slices of 3D Computer Tomography images of lung tuberculosis (TB) patients. In addition, the third dataset represents histological images and was composed of 100,000 high-resolution color microscopy images routinely used for breast cancer diagnosis. The choice of such diverse medical image types was motivated by the intention to better understand the role of specific types of image features used for federated training.

A. Chest X-ray images

A study group of chest x-ray images was created by a random sub-sampling of suitable subjects from the original repository. The age range was chosen to be 50 years spanning from 21 to 70 years. The availability of the large image repository allowed to create a study group that was well-balanced by both age and gender. For every year of life we selected exactly 100 male and 100 female subjects. Such sub-sampling resulted in the study group consisting of (100 male + 100 female) subjects * 50 years = 10,000 subjects. Thus, we believe that the use of such a balanced dataset of healthy subjects allows us to avoid additional factors caused by the natural variability of pathological changes in different patients.

Since the primary goal of this study was not the analysis of chest radiographs as such but comparison of different ways of medical image generation in the context of FL, the original x-ray images were preprocessed to avoid unnecessary large intensity range. This was done by adaptive intensity rescaling of the original images stored in medical DICOM format with 2 bytes per pixel down to the commonly used range of 0 – 225. Rescaling was done with the help of well-known technique of cutting the intensity range by histogram quantiles of 0.02 and 0.98. The characteristic large area of dark background was excluded using body masks which resulted from a preliminary image segmentation. Taking into account our close links with the manufacturers of x-ray machines, we have chosen to crop all the images by cutting out 25 % of bottom rows of the original chest scans and by 5% of pixels from the other three sides. Finally, all the images were resized down to 256x256 pixels proportionally. Example images are given in Fig. 1. Additional information related to the x-ray image properties in the context of classification tasks can be found in [8].

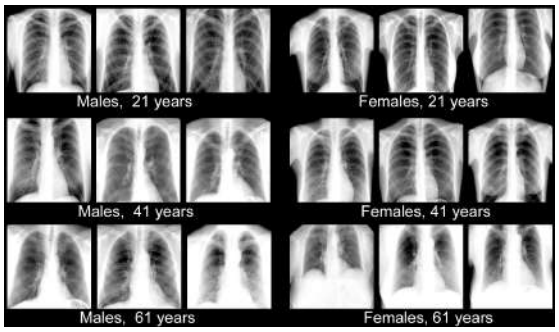

Figure 1: Examples of chest x-ray images.

B. Computed tomography images

For discovering possible dependencies of results on the medical image modality, we also used 2D slices of 3D CT images of tuberculosis patients. The original CT scans satisfy all the actual regulations, limitations, and the project agreements they are performed with. The original images were anonymized in due course before any steps of their computerized pre-processing and classifications. Thus, there are no ways to disclose, share, and disseminate any personal data. The approach and the sequence of steps of CT image data preparation are described below.

We started with a large CT image dataset containing as many as 10,714 3D CT scans. Then we excluded scans for which the information about Age and Gender of patients was not available. As a result, we end up with 8,463 CTs including 4,662 Males and 3,801 Females. The remaining 8,463 3D images were split into 2D axial slices. This resulted in 1,002,012 2D slice images of 512x512 pixels in size (574,309 in Male and 427,703 in Female image datasets). Finally, all the 2D images were exported to lossless PNG format with reduced 8 bit/pixel intensity resolution.

We started with a large CT image dataset containing as many as 10,714 3D CT scans. Then we excluded scans for which the information about Age and Gender of patients was not available. As a result, we end up with 8,463 CTs including 4,662 Males and 3,801 Females. The remaining 8,463 3D images were split into 2D axial

slices. This resulted in 1,002,012 2D slice images of 512x512 pixels in size (574,309 of Male and 427,703 of Female patients). Finally, they were exported to lossless PNG format with 8 bit/pixel intensity. In addition, all slice images were sub-divided into 3 conditional anatomical groups ("classes") to ease processes of balancing the semantic content of image datasets (see Fig. **??**).

Class c1: The upper part of Liver.

Class c2: The Heart class, which was represented by a middle heart section plus some limited amount of adjacent axial slices along with Z vertical axis.

Class c3: The shoulders which include the upper part of lungs and their close neighboring sections above and below them by Z axis.

Figure 2: Examples of 2D slices of computed tomography images.

It should be noted that we were not able to consider other distinct anatomical sections such as the ones situated at the Neck and Kidney levels. This is because the project the image data came from, was focused on the lung tuberculosis and therefore patients were scanned only within the regions of lungs plus few additional upper and bottom safety image slices.

C. Histology images

Histological images playing the role of a "Gold standard" in diagnosis of oncological diseases world-wide. A study group of Hematoxilin-Eosin stained histological images was sampled from the dataset of whole-slide images used in an international challenge of breast cancer diagnosis (see "Minsk Team" in [9]). A total of 100,000 RGB image pieces (image tiles) of 256x256 pixels in size were sampled and pre-processed from the large original whole-slide images. The resultant image dataset included 50,000 images representing the norm and 50,000 images of cancerous tissue. Some distinctive Illustrative examples that help to imagine the image variability are given in [9]. Few illustrative examples are given in Fig. 3

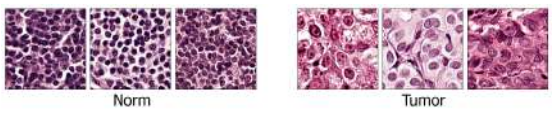

Figure 3: Examples of histological images.

III. Methods

The primary objective of the present study was to conduct a series of computational experiments to evaluate the efficacy and productivity of FL collaborative processes on training medical image generation models, as well as to identify key factors playing the role. The experimental conditions are summarized and presented in an itemized style below.

- We used Deep Convolutional Generative Adversarial Networks because they remain very popular despite the emergence of several other image generators such as the diffusion models which produce higher-quality image samples and more easy to scale and control [10], [11].
- Training of generative models was executed with different number of training epochs before performing the aggregation step on the FL server. The aggregation was carried out after every 1, 2, and 5 epochs.
- In each experiment we trained generative network during the total of 100 rounds of federated training.
- The Influence of specific medical image modality, i.e., the image type was studied by way of running same set of experiments. Each experiment followed by comparison of the FL results including the federated training efficiency as well as the quality of generated images.
- In all the occasions we used the Fréchet Inception Distance (FID) [12] to evaluate the quality of the generated images. This metric combines two different and controversial properties. The FID score expresses quantitatively how similar the generated images to their parental image dataset and how variable the generated image dataset is. The smaller FID value, the more similar generated images to the original ones.
- During the federated training, the FID distance computed between the generated images produced by each copy of the local generative model as compared to the local training image dataset of each client. It should be noted that the resultant FID value is computed using Inception v3 neural net which is trained by itself each time it is called.
- For the research purposes, the FID score was also calculated for an aggregated model on the server relative to a common fetched dataset, which is the union of local image datasets.
- Dependence of federated training convergence on the image size was evaluated using relatively small image resolution of 64x64 and 128x128 pixels. This is because of commonly known substantive problems of GANs associated with generation of high-resolution images. An additional, purely technical reason was to accelerate the massive computations necessary for distributed training of image generators.
- Considering the very high computation expenses caused by a grid-like set of experiments, the size of generated images and corresponding number of image generation experiments was limited to 64x64 and 128x128. All the experiments were repeated for each image dataset and each resolution.
- Our FL setup has 2 clients that send all the weights of the local models to the server every round. Once the weights of local models are sent to the server, they are aggregated by weights averaging. Then the new copy of aggregated model is sent back to the clients to continue training on local data.
- The number of image data items was the same in each client and they do not overlap.
- Experiments were performed on a dedicated server equipped with 4 GPU of NVIDIA V-100 type with 16Gb of video RAM each.

IV. Results

The results of the computational experiments are summarized in figures 4 and 5. Fig. 4 reports results of joint federated training of generative model aimed at creating images of 64x64 pixels whereas Fig 5 represent the training of the model for generating artificial images of 128x128 pixels in size. It is easy to see that the structure of both figures is identical. The rows of both figures correspond to 3 different types of medical images examined in this study including chest x-ray, computed tomography, and histological images.

The figure columns illustrate dynamics of convergence of federated training processes with respect to 3 different ways of aggregation of particular clients' results by the federated server. Namely, 3 figure columns depict results obtained when aggregation (averaging) of neural net weights performed after 1, 3, and 5 training epochs accomplished by the clients on their private image data.

The image data sets used in this study are different in the degree of their heterogeneity. The most homogeneous dataset among them is the dataset of chest x-ray images of 10,000 healthy subjects. Typically, for non-specialists majority of them appear nearly the same except for subtle visual features associated with age, gender, and (sometimes) general body constitution. More details can be found in [13], [14]. As a result, from the 1st row of Fig 4 representing training of tiny 64x64 x-ray images we see that the federated training progresses reliable and consistent for all 3 ways of aggregation of particular training results obtained by different clients. However, when it turns out to larger image samples (see the 1st row of Fig 5), the plot curves expressing the consistency of results produced by trained generative models measured with the help of FID scores, become more noisy. In case the aggregation was done after every 5 epochs (last plot of the 1st row of Fig 5), the training is not converged at all.

In general, the training trajectory of computed tomography images (see the 2nd rows of figures 4 and 5) is somewhat similar to the x-ray images discussed above. However, in all the occasions the training becomes much more unstable. This is evident from the bahaviour of curves characteristic for both training processes in clients (blue and yellow curves) as well as for the results of aggregation in the server (green curve).

It is easy to note even from the small set of histological images shown in Fig. 3 that they are highly variable morphologically. Usually, there is no repetition of the same cell patterns presented in the very large whole-slide microscopy image taken from the biopsy sample. This case, the FedAVG approach behaves similarly for both image sizes (see the 3rd rows of figures 4 and 5). Namely, in the majority of training loops, the model that aggregates results obtained by two clients (note the green line) is always situated above the blue and orange lines that correspond to the two independent clients. The aggregated results are notably worse than the ones obtained when each client trains generative model locally. This practically means that the Fréchet distance between the real and generated images is large and the use of FL technology leads to high dissimilarity of real and generated images. Thus, under given specific conditions, it is not worth considering Federated training.

V. Conclusions

Results obtained in this study allow to draw the following conclusions.

1. It was found that the potential utility of the horizontal FedAVG FL approach depends strongly on the natural homogeneity of the image datasets involved in the federated training.

2. Among the examined images, the best potential for federated training has been demonstrated by chest x-ray images whereas the typical Hematoxylin-Eosin stained images were rendered as not suitable for FedAVG training. The 2D computed tomography image slices are situated somewhere between the two aforementioned image types with some unstable behavior during the training.

3. The period of aggregation of training results on the federated server should be reasonably short and repeated after every 1-3 epochs performed on the local image datasets of clients.

VI. An outline of future research directions

Finally, let us make a sketch of possible future research for two different directions. The firs direction is related to the near future of developments in the field of FL technologies while the second one is more general and associated with the potential synergy of joining basic ideas and technologies of *General-AI* and Large Language Models (LLMs) [16].

A. Vertical Federated Learning in medical image analysis

Based on the distributed way of data, FL can be primarily categorized into three scenarios [15]:

- Horizontal FL, which is dealing with image data with similar distribution among the clients (e.g., chest x-rays of Norm and Pathology, MRI tomography images acquired by different MRI scanners with the strengths of magnetic field of 1.5 and 3.0 Tesla, etc).
- Vertical FL in which clients share the same samples (e.g., patients) but have different local features.
- Federated Transfer Learning [17] in which clients share both common samples and parts of the feature spaces.

In our view, the main way of further development of FL technologies in computerized medical diagnosis and treatment is to jointly use multi-sort image, signal, and laboratory data and proceed with them by the algorithms of Federated Transfer Learning technology. In this context, the Federated Transfer technology could be used without the need to share aforementioned private data to fulfill the local law and privacy concerns applied in different places (e.g., hospitals, regions, countries).

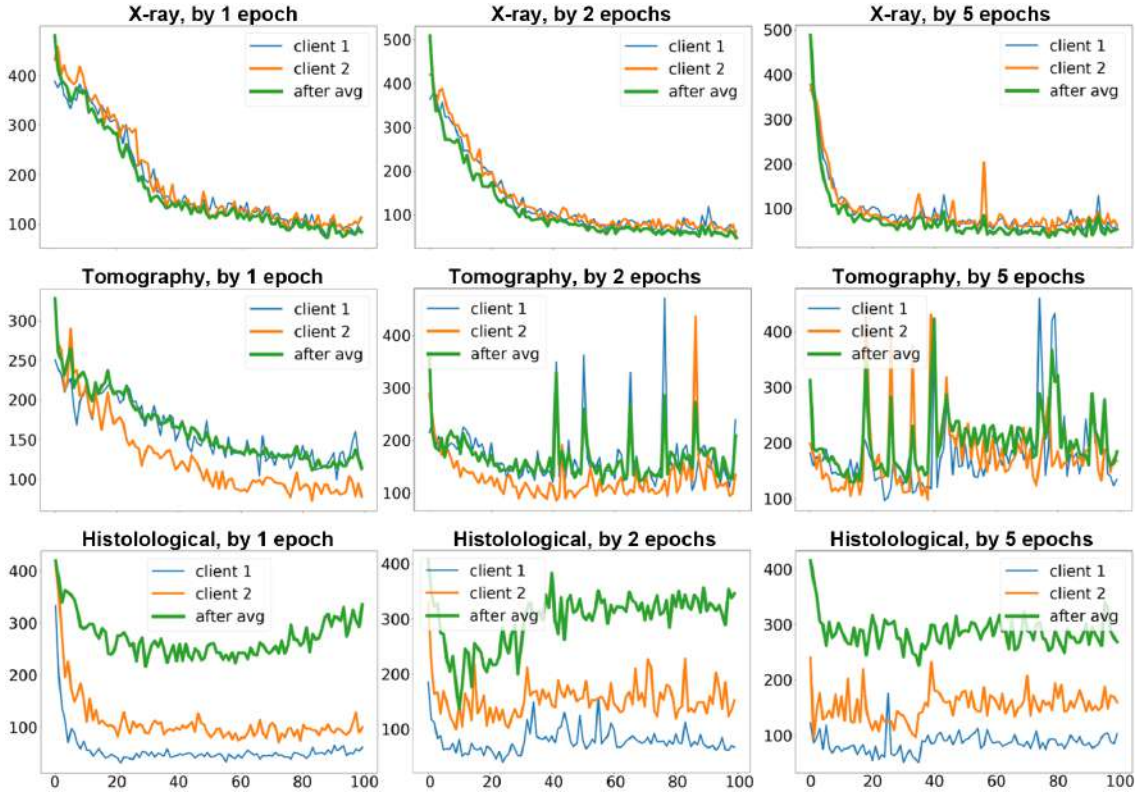

Figure 4: Results of training generative model for images of 64x64 pixels in size (explanations in text).

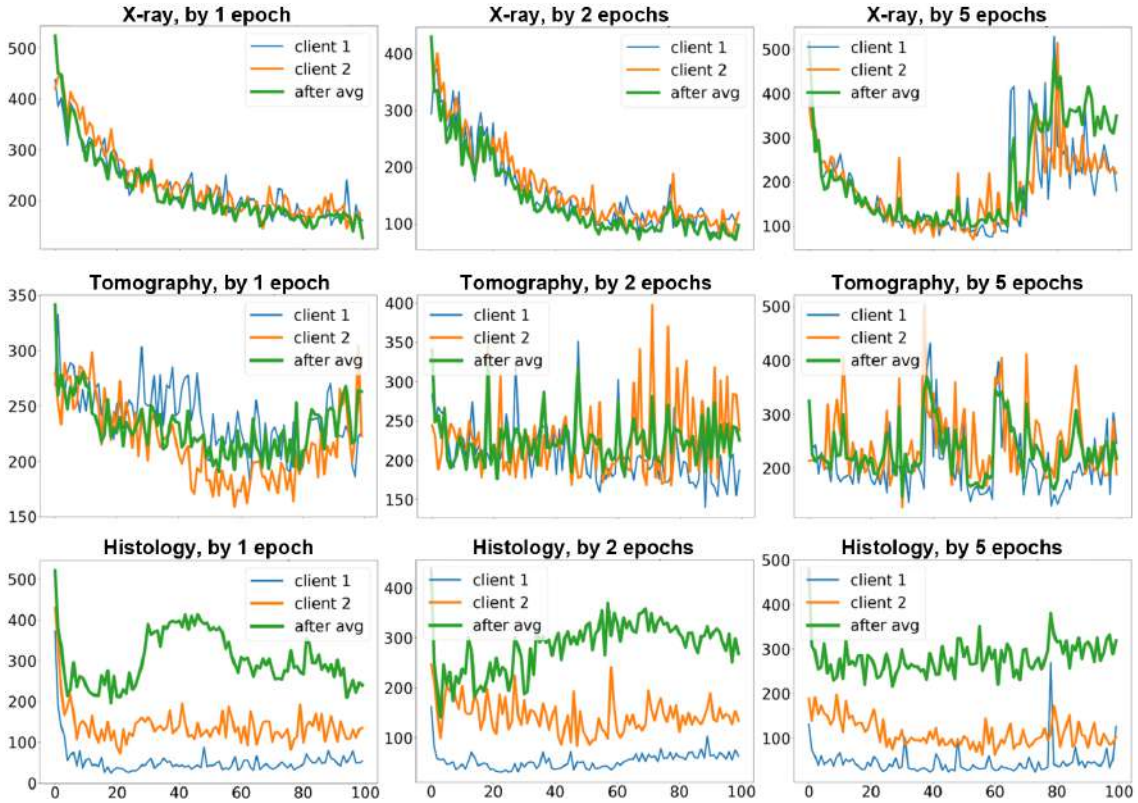

Figure 5: Results of training generative model for images of 128x128 pixels in size (explanations in text).

B. Relationships with classical General-AI Methodology

The process of integration of classical AI concepts such as semantic networks, logical inference, frame-based systems, and various connectionist approaches (i.e., methods of cognitive modeling that utilizes large networks of simple computational units) with modern LLM technologies offers promising ways to address existing limitations in reasoning, knowledge grounding, and interpretability. A short list of promising approaches to bridge these two is given below. There is a large body of relevant references that can be explored separately according to the specific area of interests.

- Structured Knowledge Integration by way of using semantic networks as Knowledge Bases in explicit relationships (e.g., "Paris is capital of France"). Generation of commonsense knowledge by fine-tuning LLMs on semantic graphs.
- Enforcement of frame-based representations including generation of structured output using frame templates (e.g., "disease: symptoms, treatments") to guide LLMs in generating consistent outputs by converting unstructured LLM outputs into structured formats.
- Improving explainability and debugging by tracing inference paths and map LLM outputs to paths in a semantic network to explain decisions (e.g., "The model inferred X because of relationships Y and Z") and using tools like AllenNSP to visualize reasoning steps. Using logical inconsistency detection to identify hallucination patterns in LLMs. For instance, detecting claims that violate knowledge graphs.
- Using Hybrid Models that combine symbolic AI (rule-based systems) with LLMs to improve reasoning and decision-making.
- Employing few-shot and zero-shot for learning LLMs' ability to generalize from minimal examples, aligning with General AI's goal of adaptability.
- Promoting cross-disciplinary applications by way of using LLMs in fields like neuroscience and robotics to emulate human-like learning and adaptability.

Thus, classical AI techniques can mitigate LLM weaknesses (e.g., hallucination, poor reasoning) by injecting structured knowledge, enabling hybrid neuro-symbolic architectures, and improving interpretability. This synergy could lead to systems that are not only powerful in language tasks but also capable of broader, more generalized intelligence.

Acknowledgments

This work was partly supported by the Belarus Fundamental Research Fund (projects F23UZB-109 and F23KUB-005).

References

[1] B. McMahan, E. Moore, D. Ramage, et al. "Communication-efficient learning of deep networks from decentralized data". 20th International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 20-22 April 2017, pp. 1273-1282.

[2] Q. Yang, Y. Liu, T. Chen, Y. Tong. "Federated Machine Learning: Concept and Applications", ACM Transactions on Intelligent Systems and Technology, vol 10, No 2, 2019, pp 1-19.

[3] T. Li, A. Kumar Sahu, M. Zaheer, et. al. "Federated optimization in heterogeneous networks", Proceedings of Machine learning and systems, vol. 2, 2020, pp. 429-450.

[4] S. P. Karimireddy, S. Kale, M. Mohri, et al. SCAFFOLD: Stochastic Controlled Averaging for Federated Learning, 37th International Conference on Machine Learning (PMLR), 2020, vol. 119, pp. 5132-5143.

[5] X. Mu, Y. Shen, K. Cheng et. al. "FedProc: Prototypical contrastive federated learning on non-IID data", Future Generation Computer Systems, 2023, vol. 143. pp. 93-104.

[6] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza et. al. "Generative Adversarial Nets", Advances in Neural Information Processing Systems, Curran Associates, Inc., vol. 27, 2014, pp. 1-9.

[7] S. Kozlovski, V. Kovalev. "Generation of Artificial Biomedical Image Datasets for Training Deep Learning Models", Pattern Recognition and Information Processing (PRIP-2019), Minsk, Belarus, BSUIR, 21-23 May, 2019, pp. 278-281.

[8] V. Kovalev, A. Radzhabov, E. Snezhko. "Automatic detection of pathological changes in chest X-Ray screening images using deep learning methods", Chapter 8: Diagnostic Biomedical Signal and Image Processing Applications, Elsevier, London, 2023, pp. 155-178.

[9] B.E. Bejnordi, M. Veta, P.J. van Diest, et al. "Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women with Breast Cancer", Journal of the American Medical Association (JAMA), vol. 318, No 22, pp. 2199-2210, 2017.

[10] Y. Song, S. Ermon. "Generative Modeling by Estimating Gradients of the Data Distribution", 33rd Conference on Neural Information Processing Systems, Vancouver, Canada, Dec. 8–14, 2019, pp. 115–119.

[11] J. Ho, A. Jain, P. Abbeel. (2020) "Denoising Diffusion Probabilistic Models", Advances in Neural Information Processing Systems vol. 33, 2020, pp. 6840–6851.

[12] 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", NIPS'17: Proc. of the 31st International Conference on Neural Information Processing Systems, 2017, pp. 6629–6640.

[13] V.A. Kovalev. Assessing the Security of Personal Data in Large Scale chest X-Ray Image Screening, Int. Conf. on Pattern Recognition and Information Processing, 17-19 Oct, 2023, Minsk, Belarus, pp. 328-331.

[14] V. Kovalev, V. Liauchuk, A. Kalinovsky, A. Shukelovich. "A comparison of conventional and Deep Learning methods of image classification on a database of chest radiographs", International Journal of Computer Assisted Radiology and Surgery, vol. 12, Suppl. 1, 2017, pp. S139-S140.

[15] M. Ye, W. Shen, B. Du, E. Snezhko, V. Kovalev, P.C. Yuen. "Vertical Federated Learning for Effectiveness, Security, Applicability: A Survey", ACM Computing Surveys, 2025, pp. 1-23.

[16] P. Kumar. "Large language models (LLMs): survey, technical frameworks, and future challenges", Springer Nature Link, Artificial Intelligence Review, 2024, vol. 57, No. 260, pp. 1-51.

[17] Y. Chen, X. Qin, J. Wang, C. Yu, W. Gao. "Fedhealth: A federated transfer learning framework for wearable healthcare". IEEE Intelligent Systems, vol. 35, No 4, 2020, pp. 83–93.

ОЦЕНКА ЭФФЕКТИВНОСТИ ТЕХНОЛОГИИ FEDAVG ПРИ ФЕДЕРАТИВНОМ ОБУЧЕНИИ ГЕНЕРАТИВНЫХ НЕЙРОННЫХ СЕТЕЙ

Ковалев В. А., Карпенко Д. С.

Цель данной статьи – представить результаты экспериментального исследования эффективности федеративного обучения генеративных нейронных сетей типа DC-GAN. В качестве основы федеративного обучения был выбран подход FedAVG, который был исследован на больших наборах медицинских изображений, включающем рентгеновские снимки грудной клетки, аксиальные срезы трехмерных компьютерных томограмм, а также гистологические изображения, окрашенные гематоксилином-эозином. Было установлено, что результаты подхода FedAVG сильно зависят от однородности наборов изображений. Среди рассматриваемых изображений наилучший потенциал для федеративного обучения продемонстрировали рентгеновские снимки грудной клетки, в то время как типичные гистологические снимки, окрашенные гематоксилиномэозином, оказались непригодными для обучения методом FedAVG. В этом отношении, 2D слои компьютернотомографических изображений оказались где-то между указанными двумя классами. При этом процесс обучения генеративной нейронной сети на томографических изображениях отличался значительной нестабильностью при переходе от эпохи к эпохе. Установлено, что период агрегирования частных результатов обучения на стороне федеративного сервера должен быть достаточно коротким, порядка 1 раз в течении 1-3 эпох тренировки участниками федеративного обучения их копий нейронных сетей на локальных наборах изображений.

Received 01.04.2025