# Multimodal Models and Methods for Biomedical Tasks

Chaychuk Mikhail
*Faculty of Computer Science*
*National Research University Higher School of Economics*
Moscow, Russia
mvchaychuk@edu.hse.ru

***Abstract*—The development of artificial neural networks for healthcare applications has recently become one of the most promising research directions within the field of artificial intelligence. These networks have the potential to significantly improve the quality and accessibility of healthcare services in the future. However, their further development requires the creation of effective AI tools for generating artificial training data. This article proposes a method for developing such a tool to generate a dataset of colonoscopy images, which is being developed as part of participation in the ImageCLEF 2024 competition. We suggest using a multimodal text-to-image model that is based on a state-of-the-art architecture and has been modified using advanced techniques for fine-tuning pre-trained models.**

***Keywords*—Artificial neural networks, machine learning, multimodal architectures, text-to-image models, LoRA, fine-tuning, CLEF, ImageCLEF 2024, ImageCLEFmedical.**

## I. Introduction

Recently, the advancement of neural networks has reached an extremely high level, and they are being widely employed in a variety of fields. The medical sector is no exception, with artificial intelligence being utilized to analyze patient data and identify possible diseases, as well as make accurate diagnoses. Frequently, a pivotal component of the diagnostic process involves the neural network analysis of medical images, such as radiography,computed tomography (CT) scans, ultrasound scans, etc.

A significant challenge in this domain is the necessity for a substantial amount of labeled data for training models. However, obtaining such data is associated with certain difficulties. Firstly, patients' medical records are confidential. Secondly, for some diseases, there may not exist a sufficient number of images at their different stages and conditions (this is especially true for rare and novel diseases). Finally, it is essential to engage qualified medical professionals in the process of labeling the data.

One of the proposed solutions to this problem is the creation of artificial training datasets through the use of cutting-edge generative neural networks. These models are trained on large volumes of images and can generate pictures with virtually any desired content, based on a text description.

However, for the creation of medical datasets, such models cannot be utilized in their original form, as the necessary images are highly specific and their quality must meet increased standards. Therefore, it is necessary to adapt existing solutions to each distinct task by utilizing real data. Due to the differences in quality and quantity of data available, along with significant differences in the types and formats of medical images, each of these tasks presents unique challenges.

Within the scope of this project, participation in the international competition ImageCLEFmed 2024 [1] is anticipated. The aim of the competition is to develop a model that can generate colonoscopy images. Throughout the research process, it is intended to examine existing state-of-the-art architectures, select the most suitable one, conduct experiments to refine this architecture for the task and to adjust model hyperparameters. The outcome is expected to be a multimodal model that generates high-quality colonoscopy images based on textual descriptions, capable of competing with the solutions of other participants in the competition.

In the present article, an analysis of relevant sources is initially conducted (Section II), followed by a description of the proposed methodology for solving the problem (Section III) and planned results (Section IV). In the concluding Section V, a summary of the work and certain outcomes are provided.

## II. Literature review

An integral part of text-to-image models is the textual encoder, which extracts semantic information from the prompt. Modern works utilize transformer-like LLMs (Large Language Models) for this purpose. The concept of the transformer was initially introduced in the paper [2]. In that article, the entire model was built around the attention mechanism, originally introduced in the paper [3]. This paper sparked a true revolution in the field of NLP (Natural Language Processing) and laid the groundwork for the emergence of models that further advanced the technology. The papers [4] and [5] introduced models GPT and BERT based on the transformer's decoder architecture. These models, trained on vast volumes of text data, enable solving of a wide array of text processing tasks due to their extensive knowledge about the general structure of language embedded within them. It is also noteworthy that the cross-attention mechanism, introduced in [2], is an essential component in many text-to-image architectures, bridging together the textual and generative models.

Models for text-to-image generation have become increasingly popular in recent years. In early architectures, such as DALL-E [6], the concept of GANs, introduced in the paper [7],

was employed for generation. However, the most contemporary models, like DALL-E 2 [8], Imagen [9], StableDiffusion [10], and Kandinsky [11], utilize the principle of diffusion - the restoration of images from random noise. Diffusion was considered the most promising method for image generation. Yet, in the latest works, which surpass diffusion models in quality, generators based on transformers are used. The Muse model [12], for example, uses a T5-XXL [13] text encoder and VQGAN [14] image generator, which incorporates the strengths of both GANs and transformers. The CM3Leon model [15] utilizes only a modified transformer decoder for generation, and it is claimed to produce higher-quality results than previous models.

Large pre-trained models are an integral part of most modern solutions, as they significantly enhance the overall quality of models trained on small data volumes for specialized tasks. However, fine-tuning such models in their entirety can be extremely resource-intensive. Parameter-Efficient Fine-Tuning represents a set of approaches that address this issue by altering only a small number of additional parameters during the training process. The LORA article [16] introduces the method of low-rank adapters, which allows approximating the change in gradients of a large model through the product of two low-rank matrices. This approach allows the use of the weights of the original model in their frozen form, while changing only a small number of additional parameters, substantially reducing training costs and increasing its speed. Furthermore, the QLoRA article [17] proposes the idea of additional model quantization - reducing data storage costs by lowering the precision of the numerical representation of weights. Specifically for the fine-tuning of large textual transformer-based models, paper [18] suggests the concept of adding small "adapter" layers to the model. These layers enable the modification of the model's weights in the desired direction while requiring relatively few resources.

To address the current challenge of ImageCLEF 2024 [1], it may be beneficial to examine the papers of the ImageCLEF 2023 participants who solved the inverse problem – a textual response to a question regarding a colonoscopy image. The papers [19] and [20] outline the models that demonstrated the highest quality in solving this task. Although these models cannot be directly applied to the current challenge, the experience accumulated by their authors in working with a similar dataset, and their approach to the selection of textual encoder model may prove useful in the research process.

## III. Methodology

Multimodal neural networks are models capable of processing, combining, and translating between various types of data, such as images, sound, and text. This process becomes possible through the integration of multiple neural networks, each specialized in working with its own type of data, into a single entity by using various tools and architectural solutions.

Modern multimodal models for image generation based on text, also known as text-to-image models, typically have a similar architecture that is also planned to be used in this work. This architecture consists of the following components.

Firstly, a textual encoder is required, which will extract features from the incoming prompt and translate them into an embedding vector. Recently, it has become standard to use LLMs based on transformers, such as BERT [5] and GPT [4], for this purpose. These models can be used as they are or further fine-tuned to better suit a specific task.

Secondly, a generative model is needed, which will be directly responsible for creating the image itself. This could be a generative adversarial network or diffusion, which is typically used in most of the latest architectures.

These models need to be integrated with each other. For this purpose, the cross-attention mechanism is most often used, which allows adding the prompt embedding into the generation process and directing the creation of the image towards a specific result.

In state-of-the-art architectures, various additional features are often incorporated into the models described above, further enhancing the ability to precisely control and optimize the final image. For example, Muse paper [12] authors have incorporated a super-resolution module into the processing pipeline, allowing for the production of higher-resolution outputs without the need to work at that resolution during the training phase, which significantly reduces computational costs.

Within the framework of this research, it will be necessary to select the most suitable encoder model, guided by state-of-the-art articles as well as previous years' works of ImageCLEF [1] participants, who addressed similar tasks. It is also planned to use LoRa [16] or some other techniques for fine-tuning the selected model. However, the main focus of the research will be on the process of selecting a generative model. It is proposed to test several different open-source models (such as StableDiffusion [10] and Kandinsky [11]) and choose the most suitable one, both in terms of the result and in terms of ease of training and tuning. Open-source models were chosen for the reason that the dataset provided within the competition consists of a relatively small number of pictures (about 2,000). This quantity is insufficient for training a model of proper quality from scratch, so it is necessary to utilize a pre-trained model. Additionally, it will be essential to try various fine-tuning methods and select optimal hyperparameters, which is likely to require conducting a significant number of experiments. If necessary, the possibility of adding additional modifications and architectural elements to improve the quality of the generated images may also be considered.

The experiments are planned to be conducted on the HSE University cHARISMa supercomputer cluster.

## IV. Expected Results

The expected result of this work is a multimodal text-to-image model that generates colonoscopy images that correspond to a given textual description (prompt). This model is expected to be based on one of the most recent and effective architectures, showing sufficiently high quality on test data

as measured by the FID metric (and potentially other metrics that will be identified in the course of the study). Moreover, it should possess the capability to compete in terms of quality with the solutions offered by other participants in the ImageCLEF 2024 [1] competition. Additionally, as a culmination of the efforts within the participation in the ImageCLEF [1] forum, there is also an intention to prepare and publish a scholarly article describing the results of participation in the competition and the developed model.

## V. Conclusion

This article describes the proposed process of working on the task within the framework of participation in the ImageCLEFmed 2024 [1] international competition. In the competition, participants are challenged with the task of designing a neural network capable of generating high-quality colonoscopy images from textual descriptions. The aim is for the generated images to be sufficiently similar to real-world data so that, theoretically, they could be used as a training set for medical neural networks

To address this challenge, we propose creating a multimodal text-to-image model based on one of the contemporary state-of-the-art architectures using fine-tuning.

A brief analysis of the most popular modern image generation methods, as well as multimodal text-to-image architectures and fine-tuning techniques for large models, is presented. Assumptions are made regarding how the described methods could be applied in this work.

It is expected that the final solution will be able to demonstrate high quality and compete with the methods employed by other participants in the competition. Subsequently, the developed model could potentially serve as the basis for solving a multitude of similar tasks from the same or different fields.

## VI. Acknowledgments

This research was supported in part through computational resources of HPC facilities at HSE University.

Word Count: 1800

## References

[1] Imageclef 2024. [Online]. Available: https://www.imageclef.org/2024

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023.

[3] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2016.

[4] A. Radford and K. Narasimhan, "Improving language understanding by generative pre-training," 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:49313245

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[6] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," 2021.

[7] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

[8] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," 2022.

[9] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 36 479–36 494.

[10] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2022.

[11] A. Razzhigaev, A. Shakhmatov, A. Maltseva, V. Arkhipkin, I. Pavlov, I. Ryabov, A. Kuts, A. Panchenko, A. Kuznetsov, and D. Dimitrov, "Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion," 2023.

[12] H. Chang, H. Zhang, J. Barber, A. Maschinot, J. Lezama, L. Jiang, M.-H. Yang, K. Murphy, W. T. Freeman, M. Rubinstein, Y. Li, and D. Krishnan, "Muse: Text-to-image generation via masked generative transformers," 2023.

[13] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," 2023.

[14] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," 2021.

[15] L. Yu, B. Shi, R. Pasunuru, B. Muller, O. Golovneva, T. Wang, A. Babu, B. Tang, B. Karrer, S. Sheynin, C. Ross, A. Polyak, R. Howes, V. Sharma, P. Xu, H. Tamoyan, O. Ashual, U. Singer, S.-W. Li, S. Zhang, R. James, G. Ghosh, Y. Taigman, M. Fazel-Zarandi, A. Celikyilmaz, L. Zettlemoyer, and A. Aghajanyan, "Scaling autoregressive multi-modal models: Pretraining and instruction tuning," 2023.

[16] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 2021.

[17] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," 2023.

[18] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and

S. Gelly, “Parameter-efficient transfer learning for nlp,” 2019.

[19] P. Cieplicka, J. Klos, M. Morawski, and J. Opala, “Language-based colonoscopy image analysis with pretrained neural networks,” 2023. [Online]. Available: https://ceur-ws.org/Vol-3497/paper-120.pdf

[20] T. M. Thai, A. T. Vo, H. K. Tieu, L. N. Bui, and T. T. Nguyen, “Uit-saviors at medvqa-gi 2023: Improving multimodal learning with image enhancement for gastrointestinal visual question answering,” 2023. [Online]. Available: https://ceur-ws.org/Vol-3497/paper-129.pdf