УДК 004.8

Creation of a dataset of vector digits for training LLMs for vector graphics generation

Timofeenko B.A. (ITMO University)

Scientific research supervisor — Filchenkov A.A., PhD
(ITMO University)

Introduction. Large Language Models (LLMs), such as GPT-4 [1], GPT 3.5[2] or Llama 2, are becoming a cutting-edge frontier of AI advancements. Besides their natural language processing capabilities, such models can generate source code in different programming languages. In my previous research, I found out that these models are also capable of generating simple vector graphics in form of SVG code with the trick of using DAN master prompt[3]. But can we push the limits and train a LLM on a custom dataset for better performance in this task? In this research, I collect such a dataset and explore the possibility of using it in the training process of LLMs.

Main part. During this research, I was aiming to create a MNIST-like dataset, but for vector graphics. Currently, vector graphic datasets are almost absolutely absent, and those that do exist, aim at very specific tasks and don't provide good data quality in general. The conduction process of the dataset consisted of scraping the web for different fonts, including the ones that imitate handwritten symbols, with a custom script. After collecting the fonts, they were processed with another custom script to extract glyphs of digits and convert them to SVG format. A file structure was proposed for better organization of the dataset.

Выводы. A high-quality dataset of SVG vector digits was created. In continuation of this research, the dataset will be used to train Large Language Models, and their performance will be evaluated within the task of generating SVG digits.

List of used sources:

1. OpenAI team, "ChatGPT: Optimizing Language Models for Dialogue", OpenAI blog, 2022
2. OpenAI team, "Language Models are Few-Shot Learners", Advances in Neural Information Processing Systems, pages 1877—1901, 2020
3. AfSch001, "DAN 2.0" // "ChatGPT" community on Reddit, 2022