УДК 811.111:004.8

AI SYSTEM FOR CREATING REALISTIC IMAGES

student Melnichuk A.V.
student Oboznaya A.A.
scientific supervisor – senior lecturer Vanik I.Y.
Belarusian National University of Technology
Minsk, Belarus

Currently, in our modern world, neural networks are beginning to gain momentum. They are used where we don't even expect: in speech recognition and synthesis, navigation systems, and even in industrial robots.

Neural networks, also known as artificial neural networks (ANNs) or simulated neural networks (SNNs), are a subset of machine learning and are at the heart of deep learning algorithms. Their name and structure are inspired by the human brain, mimicking the way that biological neurons signal to one another [1].

The aim of this paper is to discuss the DALL-E 2 neural network. DALL-E 2 is an AI system that can create realistic images and art from a description in natural language. At the beginning of 2021, OpenAI company released an AI system called DALL-E that could generate realistic images from the description of the scene or object. The generator’s name was a frankenword coined after combining the artist Salvador Dali and the robot WALL-E from the Pixar movie of the same name. Within days, it had taken the world of computer vision and artificial intelligence by storm [2].

How does DALL-E 2 work? DALL-E 2’s goal is to train two models. The first is Prior, it is trained to take text labels and create CLIP image embeddings. The second is the Decoder, which takes the CLIP image embeddings and produces a learned image. After training, the workflow of inference looks like

this. The entered caption is transformed into a CLIP text embedding using a neural network [3].

Next, Prior reduces the dimensionality of the text embedding using Principal Component Analysis or PCA. Image embedding is created using the text embedding. In the decoder step, a diffusion model is used to transform the image embedding into the image. The image is upscaled from 64×64 to 256×256 and then finally to 1024×1024 using a Convolutional Neural Network [3].

Here's a quick rundown of the DALL-E 2 text-to-image generation process. A text encoder takes the text prompt and generates text embeddings. These text embeddings serve as the input for a model called the Prior which generates the corresponding image embeddings.

Finally, an image decoder model generates an actual image from the embeddings. Sounds straightforward, but how does each of these steps actually work? The text and image embeddings used by DALL-E 2 come from another network created by OpenAI called CLIP [3].

References

1. What are neural networks? [Electronic resource]. – Mode of access: https://www.ibm.com/topics/neural-networks. – Data of access: 25.03.2023.

2. What is DALL-E 2 and how does it work? [Electronic resource]. – Mode of access: https://openai.com/product/dall-e-2. – Data of access: 25.03.2023.

3. How does DALL-E 2 work? [Electronic resource]. – Mode of access: https://medium.com/augmented-startups/how-does-dall-e-2-work-e6d492a2667f. – Data of access: 25.03.2023.