

**VILNIUS UNIVERSITY**

**ŠIAULIAI ACADEMY**

BACHELOR PROGRAMME SOFTWARE ENGINEERING

**Artificial Intelligence**

**Report on**

**“Investigation of Neural Network Model” task**

Student: Anna Kutova

Lecturer: Dr. Gintautas Daunys

Šiauliai, 2023

*Table of Contents*

[*Table of Contents* 2](#_Toc154423088)

[1. Introduction 3](#_Toc154423089)

[ **Model Description** 3](#_Toc154423090)

[2. Description of Neural Network Architecture: 3](#_Toc154423091)

[ **Residual Networks (ResNets):** 3](#_Toc154423092)

[ **Identity Block:** 4](#_Toc154423093)

[ **Convolutional Block:** 4](#_Toc154423094)

[ **ResNet-50 Model:** 5](#_Toc154423095)

[ **Summary:** 5](#_Toc154423096)

[3. Training 6](#_Toc154423097)

[ **Training Data** 6](#_Toc154423098)

[ **Training Procedure** 6](#_Toc154423099)

[ **Available Checkpoints** 6](#_Toc154423100)

[4. Evaluation Results 7](#_Toc154423101)

[ **DDIM Sampling Steps** 7](#_Toc154423102)

[ **Classifier-Free Guidance Scales** 7](#_Toc154423103)

[5. Model Execution 8](#_Toc154423104)

[**General Options** inside the program: 9](#_Toc154423105)

[6. Conclusion 10](#_Toc154423106)

1. **Introduction**

* **Model Description**

The [Stable Diffusion v2-1 model](https://huggingface.co/stabilityai/stable-diffusion-2-1) is a text-to-image generation model based on Latent Diffusion Models. It is fine-tuned from stable-diffusion-2 with additional steps on the same dataset. The model uses a fixed, pretrained text encoder (OpenCLIP-ViT/H) for encoding text prompts.

1. **Description of Neural Network Architecture:**

The Stable Diffusion v2-1 model employs a text-to-image generation architecture based on Latent Diffusion Models. Fine-tuned from stable-diffusion-2, it incorporates additional steps on the same dataset. The model's core architecture involves a fixed, pretrained text encoder (OpenCLIP-ViT/H) for encoding text prompts.

* **Residual Networks (ResNets):**

To understand the neural network architecture, it's crucial to delve into the concept of Residual Networks (ResNets). ResNets were introduced by Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun to address challenges in training very deep neural networks. The primary issue, vanishing gradients, occurs when gradients become too small during training, impeding effective weight updates.

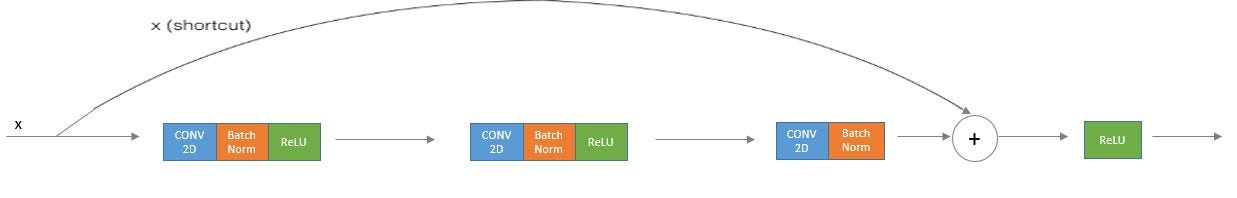
ResNets tackle this problem by introducing skip or shortcut connections, allowing the input signal to bypass certain layers and connect directly to subsequent layers. This mitigates the vanishing gradient problem and facilitates the training of much deeper networks.

A diagram of a flowchart

Description automatically generated

* **Identity Block:**

The identity block is a fundamental component of ResNets and is used when the input and output dimensions are the same. It consists of a "shortcut path" and a "main path."



ResNets Identity block

The shortcut path allows the direct flow of information without any alteration.

* **Convolutional Block:**

When input and output dimensions differ, a convolutional layer is added to the shortcut path, creating a convolutional block.

A diagram of a person with a long curved line

Description automatically generated with medium confidence

This arrangement ensures compatibility between input and output dimensions.

* **ResNet-50 Model:**

With these essential building blocks, you are equipped to construct an extensively deep ResNet. The comprehensive architecture of this neural network is intricately detailed in the following figure. In the diagram, "ID BLOCK" represents the "Identity block," and "ID BLOCK x3" indicates the stacking of three identity blocks in succession.

A diagram of a block diagram

Description automatically generated

The architecture of a full ResNet model with 50 layers involves stacking multiple identity and convolutional blocks. Each stage comprises specific configurations, including convolutional filters, sizes, and strides. The model incorporates stages with varying filter sizes (64, 128, 256, 512) and achieves down-sampling through convolutional blocks.

The final architecture includes an average pooling layer followed by a fully connected layer with softmax activation for classification. The ResNet-50 model is constructed using the Keras framework, with details provided for each stage and block.

* **Summary:**

In summary, the Stable Diffusion v2-1 model's architecture is rooted in Latent Diffusion Models, while its underlying ResNet-inspired structure addresses challenges associated with training very deep neural networks. The incorporation of identity and convolutional blocks, as well as specific configurations in each stage, contributes to the model's capacity for high-resolution image synthesis. The Keras implementation of a ResNet-50 model serves as a practical example, emphasizing the significance of skip-connections in overcoming the vanishing gradient problem and enabling the training of deep networks.

1. **Training**

* **Training Data**

The model developers utilized the LAION-5B dataset and its subsets for training. The training data underwent further refinement through LAION's NSFW detector, with a "p\_unsafe" score of 0.1 (conservative).

* **Training Procedure**

Stable Diffusion v2, a latent diffusion model, combines an autoencoder with a diffusion model trained in the autoencoder's latent space. The training process involves the following steps:

* Images are encoded using an encoder, converting them into latent representations. The autoencoder uses a relative downsampling factor of 8, mapping images of shape H x W x 3 to latents of shape H/f x W/f x 4.
* Text prompts are encoded through the OpenCLIP-ViT/H text-encoder. The output of the text encoder is fed into the UNet backbone of the latent diffusion model via cross-attention.
* The loss involves a reconstruction objective between the noise added to the latent and the prediction made by the UNet, including the v-objective (refer to <https://arxiv.org/abs/2202.00512>).
* **Available Checkpoints**

- 512-base-ema.ckpt

- 768-v-ema.ckpt

- 512-depth-ema.ckpt

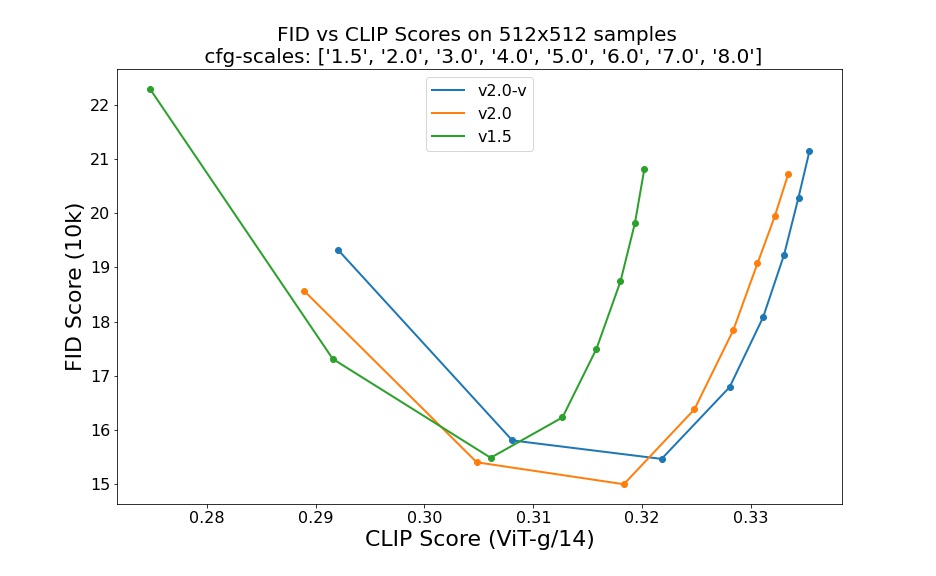
- 512-inpainting-ema.ckpt

- x4-upscaling-ema.ckpt

1. **Evaluation Results**

* **DDIM Sampling Steps**

Evaluation includes different classifier-free guidance scales with 50 DDIM sampling steps. Performance improvements are measured relative to checkpoints.



* **Classifier-Free Guidance Scales**

Evaluation results with guidance scales such as 1.5, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0 are provided. The evaluations were conducted at 512x512 resolution.

1. **Model Execution**

I found this model in Apple store <https://apps.apple.com/ua/app/diffusers/id1666309574?l=uk&mt=12>, so the installation was quite easy.

Here are examples of programs’ window and an example of generated picture:

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a clown

Description automatically generated

**General Options** inside the program:

* Diffusers launches with **a set of 5 models** that can be downloaded from the Hugging Face Hub:

A close up of a screen

Description automatically generated

A grey and white text

Description automatically generated with medium confidence

**Stable Diffusion 1.4**

This is the original Stable Diffusion model that changed the landscape of Al image generation.

**Stable Diffusion 1.5**

Same architecture as 1.4, but trained on additional images with a focus on aesthetics.

**Stable Diffusion 2**

Improved model, heavily retrained on millions of additional images.

**Stable Diffusion 2.1**

The last reference in the Stable Diffusion family. Works great with negative prompts.

**OFA small vO**

This is a special so-called distilled model, half the size of the others. It runs faster and requires less RAM, try it out if you find generation slow

* **Promts**

A screenshot of a computer

Description automatically generated

Prompt is the description of what is wanted, and negative prompt is what isn’t wanted on the image.

* **Guidance Scale**

A screen shot of a computer

Description automatically generated

Indicates how much the image should resemble the prompt.

Low values produce more varied results, while excessively high ones may result in image artifacts such as posterization.

Values between 7 and 10 are usually good choices, but they affect differently to different models.

* **Step count**

A close up of a blue screen

Description automatically generated

How many times to go through the diffusion process.

Quality increases the more steps is choosen, but marginal improvements get increasingly smaller.

* **Seed**

A close up of a computer screen

Description automatically generated

The seed is a numerical value that enables the reproduction of a previous image generation. Choosing -1 will result in the selection of a random seed.

* **Advanced**

A screenshot of a computer

Description automatically generated

This section provides options for experimenting with different optimization settings. While Diffusers attempts to automatically select the optimal configuration for the system, it may not always be the most efficient.

1. **Conclusion**

In summary, exploring the Stable Diffusion v2-1 model has been captivating. Rooted in Latent Diffusion Models, it's a robust text-to-image tool fine-tuned from stable-diffusion-2. The model's efficiency is emphasized with a fixed, pretrained text encoder (OpenCLIP-ViT/H) for prompts.

Unveiling the architecture, especially Residual Networks (ResNets), shows how the model handles deep network challenges. The design of identity and convolutional blocks in ResNet-50 contributes to image synthesis.

Training with LAION-5B and subsets shows a methodical approach, refined by LAION's NSFW detector. The model's sophistication lies in combining an autoencoder with a diffusion model, incorporating factors like downsampling and cross-attention.

Evaluation results, through DDIM steps and guidance scales, offer a comprehensive view. Checkpoints enhance adaptability.

Accessibility via the Apple store to the Diffusers app, makes it user-friendly. The simple installation, with pre-trained models and options, makes it a practical choice.

However, generated images may not always meet high-quality standards and frequently exhibit an unsettling appearance. As such, continual updates are expected to refine and improve the model's image synthesis capabilities.