Translation of text to images

Teodoras Šaulys
Informatikos institutas
Matematikos if informatikos fakultetas
Vilnius, Lietuva
teodoras.saulys@mif.stud.vu.lt

Povilas Slivinskas
Informatikos institutas
Matematikos if informatikos fakultetas
Vilnius, Lietuva
povilas.slivinskas@mif.stud.vu.lt

***Abstract*—In this paper we analyze the performance of fine-tuned text-to-image models, specifically the StableDiffusion and DALL-E, using a Pokemon dataset (lambdalabs/pokemon-blip-captions). The primary focus of this team project is to thoroughly analyze the performance of these models and assess their ability to be fine-tuned effectively with the same dataset and parameters, particularly in generating images based on textual descriptions. By utilizing pre-trained models on the selected dataset, we compare the results and delve into the field of fine-tuning text-to-image models.**

***Index Terms*—5-6 Raktiniai žodžiai**

I. Introduction

Text-to-image synthesis has gained significant attention in recent years due to its potential applications in various domains such as computer vision, creative arts, and multimedia content generation. The ability to generate realistic images from textual descriptions opens up new possibilities for content creation, storytelling, and even assisting individuals with limited artistic skills to visualize their ideas. In this paper, we delve into the realm of fine-tuning text-to-image models, focusing on the StableDiffusion and <insert model>, using the Pokemon dataset (lambdalabs/pokemon-blip-captions). Our objective is to explore the effectiveness of fine-tuning these models with the same dataset and parameters, aiming to enhance their capability to generate visually coherent and accurate images. Through a comprehensive analysis of the performance and results, we aim to provide insights into the potential of fine-tuning text-to-image models and contribute to the advancement of this exciting field.

II. Methods

A. Fine-tuning

Fine-tuning text-to-image models involves the process of adapting pre-trained models, originally trained on a large dataset, to generate visually coherent and accurate images based on textual descriptions. By fine-tuning these models with a specific dataset and parameters, we aim to enhance their ability to generate images that align with the given textual input. Fine-tuning leverages the pre-existing knowledge and learned features of the models, allowing them to specialize and generate more contextually relevant images in response to textual prompts. This process facilitates the transfer of knowledge from the pre-trained models to the target task, enabling improved performance and generating visually compelling images that effectively capture the essence of the provided textual descriptions.

1) Fine-tuning parameters: For fine-tuning both the StableDiffusion and <insert model> text-to-image models, we employed a consistent set of parameters. The following list provides a brief explanation of each parameter:

- Dataset name: lambdalabs/pokemon-blip-captions - a dataset containing used for training that contains Pokemon-related image-caption pairs.
- Use_ema: true - Enables the use of Exponential Moving Average (EMA) during training, which can stabilize the model's performance.
- Resolution: 512 - sets the resolution of the generated images to 512x512 pixels, ensuring a consistent output size.
- Center_crop: true - Applies center cropping to the input images, focusing on the central region for training.
- Random_flip: true - Randomly flips the images horizontally, introducing additional diversity during training.
- Train_batch_size: 4 - Specifies the batch size for training, determining the number of samples processed in each training iteration.
- Gradient_accumulation_steps: 4 - accumulates gradients over multiple steps, effectively increasing the effective batch size and - allowing for more stable training.
- Gradient_checkpointing: true - saving memory by trading off computational efficiency for memory consumption during backpropagation.
- Max_train_steps: 100 - limits the maximum number of training steps to 100, controlling the duration of the training process.
- Learning_rate: 1e-05 - sets the initial learning, governing the step size for adjusting the model's parameters during training.
- Max_grad_norm: 1 - Specifies the maximum gradient norm value, which is used for gradient clipping to prevent exploding gradients.
- LR_scheduler: Constant - sets the learning rate scheduler to "constant," indicating a fixed learning rate throughout the training process.

By utilizing these parameters consistently for both models, we aim to ensure a fair comparison and evaluate the effectiveness of fine-tuning in generating images based on textual

descriptions from the chosen Pokemon dataset.

B. Low-Rank Adaptation of Large Language Models (LoRA)

$$\mathcal{L}(\mathbf{X}, \mathbf{Y}) = \frac{1}{w \cdot h} \sum_{i=1}^{h} \sum_{j=1}^{w} (X_{i,j} - Y_{i,j})^2 \tag{1}$$

Taikyta nuostolių funkcija (1).

$$\begin{aligned} y &= f(x) \\ f &= f_1(f_2(x)) \end{aligned} \tag{2}$$

Taikytas modelis (2).

III. Duomenys

Aprašote naudojamus duomenis.

A. Equations

B. Paveikslėliai ir lentelės

Paveikslėlį cituojame “ 2 pav.”.
Paveikslėlį cituojame “ I lentelė”.

TABLE I – Lentelės aprašas

a	b	c	d

pav 1 – Paveikslėlio aprašas.

IV. Results

Cituojame šaltinį [2], [1].

References

[1] Karan Desai et al. *RedCaps: web-curated image-text data created by the people, for the people*. 2021. arXiv: 2111.11431 [cs.CV].

[2] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. In: *nature* 521.7553 (2015), pp. 436–444.

pav 2 – Paveikslėlio aprašas.