DTU Civil Engineering
Department of Civil Engineering

Subject-Driven Generation Techniques for Stable Diffusion Model

Thesis subtitle

Master Thesis

Subject-Driven Generation Techniques for Stable Diffusion Model
Thesis subtitle

Master Thesis
June, 2023

By
Mario Lozano Cortés

Cover photo: Vibeke Hempler, 2012
Published by: DTU, Department of Civil Engineering, Brovej, Building 118, 2800 Kgs. Lyngby Denmark
www.byg.dtu.dk
ISSN: [0000-0000] (electronic version)
ISBN: [000-00-0000-000-0] (electronic version)

ISSN: [0000-0000] (printed version)
ISBN: [000-00-0000-000-0] (printed version)

Approval

This thesis has been prepared over six months at the Section for Indoor Climate, Department of Civil Engineering, at the Technical University of Denmark, DTU, in partial fulfilment for the degree Master of Science in Engineering, MSc Eng.

It is assumed that the reader has a basic knowledge in the areas of statistics.

Mario Lozano Cortés - s226536

.....

Signature

.....

Date

Abstract

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Acknowledgements

Mario Lozano Cortés, MSc Civil Engineering, DTU
Creator of this thesis template.

[Name], [Title], [affiliation]
[text]

[Name], [Title], [affiliation]
[text]

Contents

1 Introduction

2 State of the art

Text-to-image is an emerging field of deep learning where models can generate lifelike and highly detailed images from textual descriptions. The development of these models is a challenging task that requires the close integration of both computer vision and NLP approaches. The latest advancements in text-to-image models have led to the capability of producing high-quality images with rich semantic content that can now be used for tons of applications including video games and virtual reality, e-commerce, or education among others. Even though recent advancements have allowed the use for commercial applications, generative models remain a challenging and tough problem. This section aims to analyse the current state-of-the-art of text-to-image models

2.1 Historical review of text-to-image models

2.2 Diffusion probabilistic models

Throughout 2022, the capabilities and popularity of text-to-image models have exploded. The general public is aware of some models, such as DALL-E 2, Midjourney, or Stable Diffusion. Nonetheless, the vast majority of people are unaware of the technical prowess required in the field of Artificial Intelligence for these models to exist. This section aims to shed some light on the internal functioning and processes of these models from an academic perspective.

Diffusion probabilistic models are a class of latent variable models that introduce the ideas of nonequilibrium thermodynamics into data generation techniques by homogeneously adding noise into samples. Thus, they join the list of models that manage to generate high-quality images such as variational autoencoders (VAEs) or Generative adversarial networks (GANs). The latter models have been the reference of academic research in recent years and are the benchmark to be surpassed by diffusion models.

GANs were introduced in 2014 by researchers at the University of Montreal in the paper *Generative Adversarial Nets* [1]. The idea is to create generative models through an adversarial process in which two neural networks compete against each other. One of the networks will be generative while the other will be discriminative. Thus, the generative network will be in charge of capturing the distribution of the training dataset while the discriminative network must distinguish whether a sample comes from the generative network or the training data. The idea is that the generative network maximises the probability that the discriminative network makes errors.

Diffusion models, on the other hand, achieve high-quality image synthesis results in the paper *Denoising Diffusion Probabilistic Models* [2] by researchers from the University of California, Berkeley. These models are based on creating a Markov chain in which at each step they add Gaussian noise to an image in a diffusion process and then learn to undo it. In this way, a network is trained that is capable of reconstructing images from random noise. The differences between GANs and diffusion models are presented in figure 2.1.

Diving further into the workings of diffusion models, we define the **forward process** of the Markov chain. The first step is to take a sample of the target data distribution, which we will call X_0, and add Gaussian noise in T steps. The forward process is thus defined as a Markov chain in which the state of a sample at time n depends only on the state at time $n-1$. Therefore, one can denote the distribution of any sample conditioned on the initial state X_0.

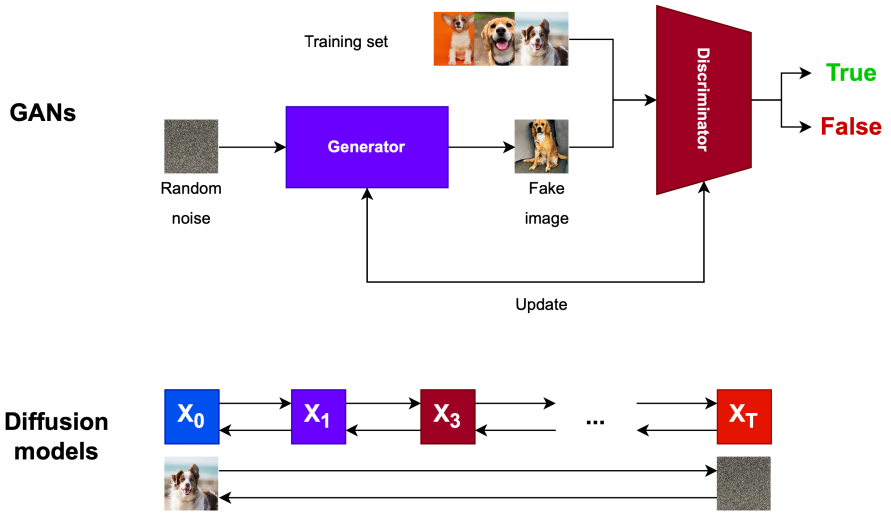

Figure 2.1: Overview of GANs and diffusion models

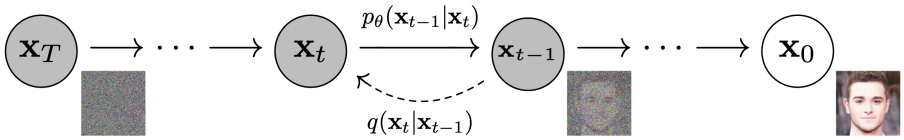

Figure 2.2: Markov chain of the diffusion process [2]

$$q\left(x_{1:T}|x_0\right) = \prod_{t=1}^{T} q\left(x_t|x_{t-1}\right)$$

In every step of the noising process Gaussian noise is added according to some variance schedule $\beta_1 ... \beta_t$, normally consider as hyperparameters. The restrictions applied to β_t are $\beta_1 < \beta_2 ... < \beta_t$ and $\beta_t \epsilon (0, 1)$. I stands for identity.

$$q\left(x_t|x_{t-1}\right) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t x_{t-1}}, \beta_t I\right)$$

As β_t grows in time and T approaches the limit ($T \to \infty$, $\beta_t \to 0$), the Gaussian mean will approach zero with identity covariance. In this way, the distribution will lose all the information contained in the original image. In practice, researchers use a T close to 1000 [2].

$$q\left(x_t|x_0\right) \approx \mathcal{N}\left(0, I\right)$$

Figure 2.2 shows the diffusion process described so far.

In summary, it is proven that the forward process destroys the structure of a data distribution step by step. The next challenge is to learn the **reverse diffusion process** in order to generate data that resembles the training distribution from pure Gaussian noise. As with the forward process, the reverse diffusion process can be expressed as a Markov chain where the probability of a sequence of samples can be expressed as the product of conditional probabilities.

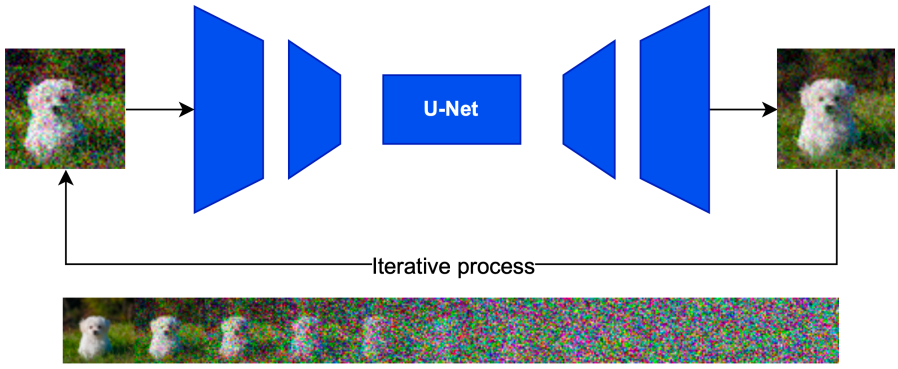

Figure 2.3: Reverse diffusion learning schema

$$p_\theta\left(x_{0:T}\right) := p\left(x_T\right) \prod_{t=1}^{T} p_\theta\left(x_{t-1} | x_t\right)$$

The reverse diffusion process involves a neural network to remove noise from an image in a stepwise manner. Thus, starting from pure Gaussian noise, noise is removed step by step to arrive at an image that resembles the training distribution. The reason that the process has to be done in a stepwise manner is that "*the estimation is more tractable than explicitly describing the full distribution*" as expressed in the publication *Deep Unsupervised Learning using Nonequilibrium Thermodynamics* [3].

The neural network that the authors of DDPM propose aims to **predict the noise to subsequently eliminate it from the image**. This is equivalent to obtaining the mean of the distribution since the authors decide to fix the variance. The authors decide to use the **U-Net network** [4] for this purpose. U-Net consists of a bottleneck in the middle that ensures that the network removes irrelevant information and focuses on the important information. In addition, the network, between the encoder and the decoder, uses residual connections to improve efficiency. Finally, the authors of DDPM decide to employ self-attention at the 16 × 16 feature map resolution. Figure 2.3 shows a schema of the learning process.

Another question that arises when working with diffusion models is how conditional generations can be provided. This can be achieved through various techniques. One way is to feed a conditional variable into the training so that the model makes use of it in the generation to resemble a subset of the training distribution. However, guiding the generation process through a classifier is a more flexible technique that allows even more complex text descriptions than simple labels to be worked with. The idea is to take an already trained classifier and **guide the generation in the direction of the gradient of the classifier label**.

2.2.1 Improvements to diffusion probabilistic models

Although the results obtained by the *Denoising Diffusion Probabilistic Models* [2] paper are excellent and represent a great leap forward compared to the images that the generative models were capable of generating until then, researchers at OpenAI suggest some improvements that increase the quality of the results in their publication *Improved Denoising Diffusion Probabilistic Models* [5]. In it, the main improvements they propose to the model are **(i) the incorporation of learned variances and (ii) an improvement of the noise schedule**.

As discussed in section 2.2, the authors of the paper *Denoising Diffusion Probabilistic Models* [2] decided to fix the variance. However, the OpenAI researchers decide to learn the interpolation

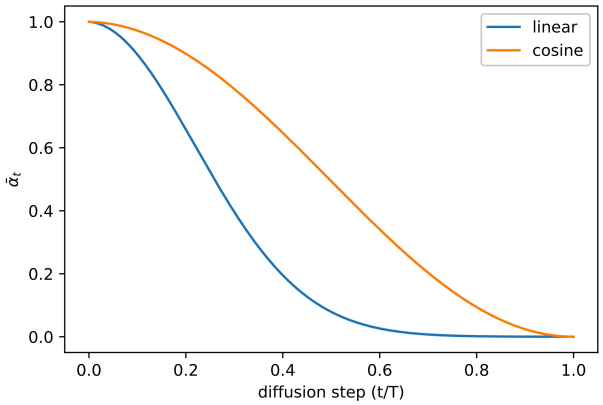

Figure 2.4: Cosine and linear schedules comparison

Figure 2.5: Cosine (bottom) and linear (top) schedules comparison on an image

of the variance between an upper and lower bound. This allows them to maintain the quality of the samples and improve the log-likelihood. Finally, they modify the loss to depend on the variance by a scaling factor λ set experimentally to 0.001.

On the other hand, the OpenAI authors present a new noise schedule designed to be linear in the central region and have little change at the beginning and end. It is defined through $\bar{\alpha}_t$, affecting the definition of the variances β_t as follows.

$$\beta_t = 1 - \frac{\bar{\alpha}_t}{\bar{\alpha}_{t-1}}$$

The proposed cosine noise schedule can be seen in figure 2.4. Whereas, figure 2.5 shows how each of the schedules adds noise to the image. The **linear schedule destroys the information faster and presents a sub-optimal behaviour** since the last steps are practically pure noise. Thus, the cosine schedule is superior as it allows a more controlled addition of noise.

However, the improvements do not stop there. The same OpenAI researchers in a later paper called *Diffusion Models Beat GANs on Image Synthesis* [6] demonstrate how a series of modifications to the architecture and the use of classifier guidance can produce images that are better than the state of the art at the time. The enhancement they make to the architecture are:

- **Increasing the depth while decreasing the width** to keep the size of the model relatively constant.
- Increased use of **attention heads and layers**
- Upsampling and downsampling the activations by means of the **BigGAN residual blocks** [7].
- Use of **adaptive group normalization** (AdaGN) layers, in which the concept of group

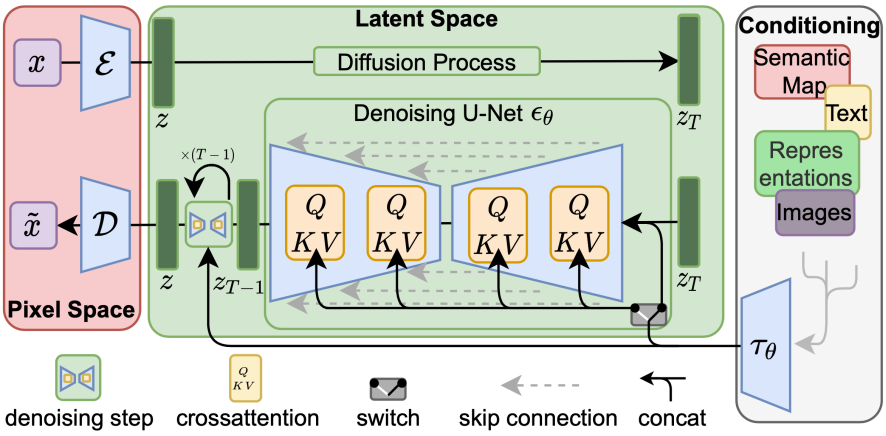

Figure 2.6: Latent diffusion architecture

normalization is expanded by adjusting the normalization parameters of each group separately according to the input data.

- **Classifier guidance**. Employing an additional classifier, the diffusion model is assisted in generating a certain class.

2.3 Latent diffusion models

Probabilistic diffusion models have enabled the generation of high-quality images with state-of-the-art results. However, they have a fundamental weakness that the successive iterations of improvements did not resolve. The fact that they operate in pixel space, dealing with additions and deletions of noise in a tensor of the same size as the input tensor, means that training these models requires enormous computational resources. Therefore, researchers from the Ludwig Maximilian University of Munich and Runway ML propose in the publication *High-Resolution Image Synthesis with Latent Diffusion Models* [8] to use **latent space instead of pixel space** to speed up the training and inference calculations of these models. The latent space is obtained from previously trained autoencoders, thus obtaining a representation of the images in a lower dimensional space that allows a balance to be reached between the quality of the details preserved and the reduction of the complexity obtained.

Thus, the operation of latent diffusion models can be summarised in the diagram present in figure 2.6. The first training step is to obtain a representation of the considered image in the latent space $\mathcal{Z}$ thanks to the encoder $\mathcal{E}$. Then, Gaussian noise is added to the diffusion process until $\mathcal{Z}_t$ is reached. For the inverse process, a U-Net network is used. However, the real strength of this approach lies in the ability to condition the generation. This is achieved thanks to a dedicated encoder τ_θ that maps the conditionings in the intermediate layers of the U-Net with cross-attention layers. Finally, the result of the latent space is returned to the pixel space thanks to the $\mathcal{D}$ decoder.

2.4 Stable diffusion

As detailed in section 2.1, text-to-image models have exploded in popularity and capabilities throughout 2022. One of the biggest drivers of this shift in public perception has been Stable Diffusion, an **open-source** model whose weights and architecture have been publicly released. As a consequence, many researchers and enthusiasts have put much effort into optimising and extending the project's capabilities. These efforts are led by the British generative AI startup Stability AI. As a result of the open-source philosophy, this model is capable of running on

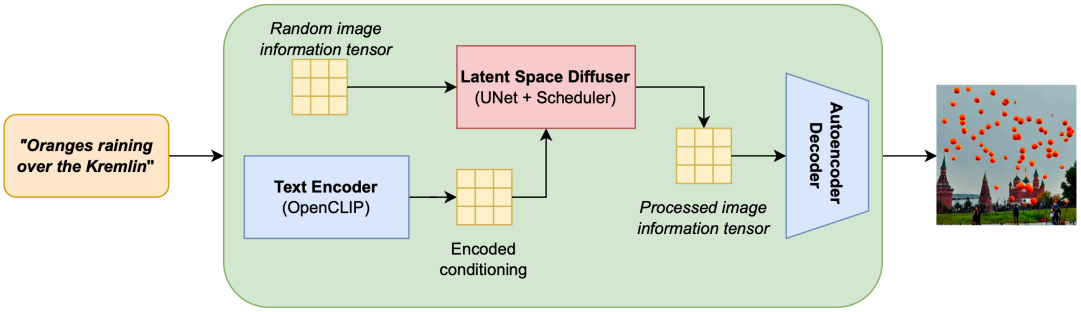

Figure 2.7: Stable Diffusion main components

consumer-available hardware. This fact allows the community to leverage its capabilities in a wide variety of cases.

Stable Diffusion is a latent diffusion model that follows the architecture developed by the Computer Vision & Learning group of the Ludwig Maximilian University of Munich in the paper *High-Resolution Image Synthesis with Latent Diffusion Models* [8], which has already been explained in section 2.3. The proposed technique can also be adapted to other tasks such as inpainting, outpainting, generating image-to-image translations or increasing the resolution of an image, all tasks that Stable Diffusion can perform. A high-level diagram of the main components of the model can be seen in figure 2.7:

- **Text Encoder**: It creates an encoded representation of the text data's description. Its goal is to influence the diffusion process, ensuring that the resulting image corresponds to the given description. Stable Diffusion's first version utilizes CLIP [9], while its second version includes OpenClip [10]. In both cases, the text encoder is used in conjunction with an image encoder. CLIP and OpenClip strive to maximize the similarity between the two encodings, enabling the model to associate images with their respective descriptions.
- **Latent Space Diffuser**: It aims to utilize the diffusion process to eliminate noise from the image by manipulating the latent space information. As the process progresses, additional information is added to enhance the similarity between the image and the provided description. It is crucial to highlight that this operation takes place in latent space, resulting in improved efficiency and being a key advancement. Figure X provides a visual representation of the denoising process guided by the text encoder.
- **Autoencoder Decoder**: The final image is produced by utilizing the compressed information stored in the latent space. This step is carried out only once to construct the ultimate pixel image.

2.4.1 Training dataset

A significant challenge posed by models like Stable Diffusion is the choice of images they are trained on. This issue is not insignificant as image generation models require both textual descriptions of training images and a sufficient amount of variety to enable the model to comprehend how the world is constructed and thus be capable of reproducing it. However, the conventional datasets of the Machine Learning field (COCO, ImageNet, etc) fail to satisfy these requirements since they are not intended for this purpose. Researchers have discovered that the solution is the web, where a vast array of diverse images about the world can be found, many of which have HTML alt attribute tags.

Stable Diffusion has an advantage over some of its rivals, including *DALL-E 2*, in that it is an

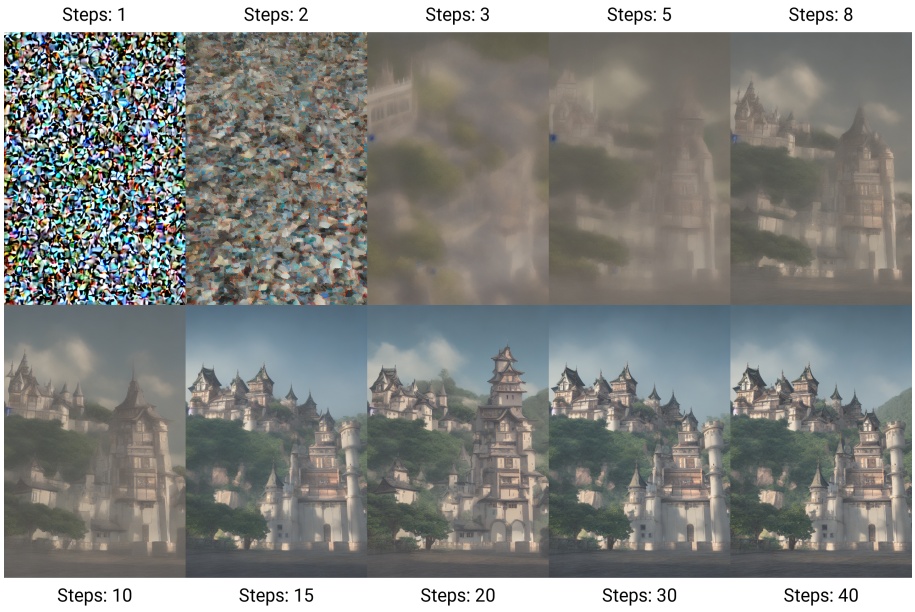

Figure 2.8: Diffusion steps [11]

open-source project, meaning that the dataset employed for training is well-known and accessible to everyone. Specifically, the dataset used by Stable Diffusion is "***LAION-5B**, a dataset of 5.85 billion CLIP-filtered image-text pairs, 14x larger than **LAION-400M**, previously the biggest openly accessible image-text dataset in the world*" [12]. In particular, Stable Diffusion presents several checkpoints on various LAION-5B assemblies. Some of these checkpoints in Stable Diffusion version 1 [13] are:

- **stable-diffusion-v1-1**: 256 x 256 images from a subset of 2.3 billion English-captioned images called **LAION-2B-EN**.
- **stable-diffusion-v1-2**: Resumed training on *stable-diffusion-v1-1* with 512x512 images from the subset **LAION-2B-EN**, containing a selection of improved aesthetics images compared to the others.
- **stable-diffusion-v1-3**: Resumed training on *stable-diffusion-v1-2* with the same subset of images but a 10% dropping of the text-conditioning.
- **stable-diffusion-v1-4**: Resumed training on *stable-diffusion-v1-2* with 512x512 images from the subset **LAION-Aesthetics v2 5+**, containing 600 million images from **LAION-2B-EN** with better aesthetics and low-resolution and watermarked images filtered out.
- **stable-diffusion-v1-5**: *stable-diffusion-v1-4* trained with more steps.

LAION-5B retrieves images from the internet that are not uniformly high in quality. Because these images are gathered automatically, they do not adhere to the same rigorous standards as other image datasets. As a result, the checkpoints for training Stable Diffusion use varying subsets of LAION-5B. Nonetheless, the fact that the images obtained are not accurately labelled as they are in standard vision supervised learning is actually an advantage. Consequently, Stable Diffusion is now included in the group of architectures, such as CLIP or DALL-E 2, that have proven the value of these vast datasets, even though they contain a significant amount of noise.

LAION-5B contains 5.85 billion image-text pairs divided into three subsets. **LAION2B-EN**, which

contains 2.32 billion English image-text pairs; **LAION2B-MULTI** with 2.26 billion image-text pairs from all other languages (Russian, French and German as top 3) and **LAION1B-NOLANG** of 1.27 billion samples where the language is not correctly defined.

LAION-5B Description
The attributes that can be found in LAION-5B are described in table 2.1.

Attribute	Description
id	Image identifier
URL	URL from where the image was obtained
Text string	Caption accompanying the image
Dimensions	Height and width of the image
Similarity	Cosine similarity between the text and image embeddings. CLIP-based models are employed to gauge the level of accuracy with which an image is described by a given textual description.
pwatermark	Probability that the image presents a watermark. The value is obtained by a custom model trained by LAION. Value between 0 and 1
punsafe	Probability that the image is NSFW. As some of the content acquired from the web may not be suitable for all audiences, LAION employs a custom model to assess its adequateness. Value between 0 and 1

Table 2.1: LAION-5B's attributes

Some statistics of the subsets computed by the LAION team can be found in table 2.2 [12].

Subset	Dimensions	NSFW	Watermark	Average text length
LAION2B-EN	- >256x256: 1324M - >512x512: 488M - >1024x1024: 76M	2.9%	6.1%	67
LAION2B-MULTI	- >256x256: 1299M - >512x512: 480M - >1024x1024: 57M	3.3%	5.6%	52
LAION1B-NOLANG	- >256x256: 1324M - >512x512: 488M - >1024x1024: 76M	3%	4%	46

Table 2.2: Statistics summary for LAION-5B

By analysing the data presented in tables 2.1 and 2.2, one can infer the rationale behind the various checkpoints employed in the Stable Diffusion model. The LAION-5B dataset, owing to its extensive diversity, can be partitioned into subsets that cater to various generation objectives. As a result, the model can be adapted to different resolutions or the quality of the generated images can be adjusted by filtering out low similarity image-to-description pairs, NSFW content,

or watermarked content.

It is noteworthy to mention that the primary characteristics of the entries in the dataset are produced by other pre-trained models. This highlights the significance of incorporating other models in the data collection process for large AI models, as they can assist in adding supplementary features to the dataset. A more detailed discussion of this fact can be found in section 2.4.1

LAION-5B Collection Methodology

The pipeline followed when creating the LAION-5B dataset involves: (i) obtaining Common Crawl data, (ii) filtering some web pages, (iii) downloading the image-text pairs, (iv) and filtering the content according to various characteristics.

Common Crawl is an organization dedicated to web crawling, data collection, and storage. It makes all the gathered data publicly available. In the October 2022 crawl, the total file size was 380 TBs, comprising 3.15 billion web pages. The dataset's key feature is that it contains HTML tag information about the images, including the ”alt” attribute, which provides an alternative description of the images. This attribute is widely used on the web, for example, to address page rendering issues, assist visually impaired individuals, or aid web content indexing by search engines. Therefore, it is a ubiquitous attribute on the web that is encouraged to improve page usability and ranking in web search engines.

After the Common Crawl data is accessible, images that have information in the ”alt” attribute are chosen. Once both images and descriptions are available, a language detection model is employed on the descriptions, and the data is then divided into three subsets: LAION2B-EN, LAION2B-MULTI, and LAION1B-NOLANG, as mentioned earlier. It is noteworthy that in order to incorporate data into LAION1B-NOLANG, a confidence threshold is determined based on the prediction of the language detection model, and if it is insufficient, it is included in this subset.

The next step is to clean the dataset of poor-quality images and descriptions. For this purpose, images, and descriptions with less than 5KB data, 5 words and 0.28 cosine similarity (in LAION2B-EN) are removed. **The cosine similarity is computed thanks to Open AI's CLIP model, which computes the embedding of images and text.**

It is important to notice the importance of the CLIP contrastive model in understanding how the Stable Diffusion training dataset was created. As explained above, CLIP is able to associate images and text. The way in which it achieves this is very clever as it can solve the classic problem of labels in Deep Learning. Thus, CLIP is a pioneer in bringing together language models and vision models by making supervision in natural language. And therein lies the key to LAION-5B: unlike other datasets that require a specialized team to create carefully curated tags, this dataset relies on natural language descriptions provided by internet users. This allows for much faster scalability. It is worth noting that CLIP is not only essential in the creation of the dataset, but it also plays a vital role in the Stable Diffusion model, as previously explained in section 2.4.

The final stage of the pipeline involves incorporating additional attributes that help categorize the image in a useful way, beyond just its similarity to text. One such attribute is the probability that the image contains NSFW content, which is determined using a custom model. Another attribute is the probability that the image has a watermark, which is determined using a separate model designed for that purpose.

Summing up, the creation of LAION-5B relies on multiple AI models that help gather reliable content from the internet and guarantee the accuracy of image descriptions. This marks a significant shift in the way we collect data for training models, where the emphasis is on scaling

the dataset rather than carefully generating accurate labels. Instead, models like CLIP enable the use of natural language descriptions that accompany web images for data collection.

2.5 Subject-driven generation techniques

3 Examples of figures, tables, equations and listings

Bibliography

[1] Ian Goodfellow et al. “Generative adversarial networks”. In: *Communications of the ACM* 63.11 (2020), pp. 139–144.

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising diffusion probabilistic models”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 6840–6851.

[3] Jascha Sohl-Dickstein et al. “Deep unsupervised learning using nonequilibrium thermodynamics”. In: *International Conference on Machine Learning*. PMLR. 2015, pp. 2256–2265.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer. 2015, pp. 234–241.

[5] Alexander Quinn Nichol and Prafulla Dhariwal. “Improved denoising diffusion probabilistic models”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 8162–8171.

[6] Prafulla Dhariwal and Alexander Nichol. “Diffusion models beat gans on image synthesis”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 8780–8794.

[7] Andrew Brock, Jeff Donahue, and Karen Simonyan. “Large scale GAN training for high fidelity natural image synthesis”. In: *arXiv preprint arXiv:1809.11096* (2018).

[8] Robin Rombach et al. “High-resolution image synthesis with latent diffusion models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10684–10695.

[9] Alec Radford et al. “Learning transferable visual models from natural language supervision”. In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763.

[10] Mehdi Cherti et al. “Reproducible scaling laws for contrastive language-image learning”. In: *arXiv preprint arXiv:2212.07143* (2022).

[11] Wikipedia. *Stable Diffusion, Wikipedia, The Free Encyclopedia*. Online; accessed 12-April-2023. 2023. URL: https://en.wikipedia.org/wiki/Stable_Diffusion.

[12] Christoph Schuhmann et al. “Laion-5b: An open large-scale dataset for training next generation image-text models”. In: *arXiv preprint arXiv:2210.08402* (2022).

[13] Robin Rombach and Patrick Esser. *Hugging Face - Stable Diffusion*. 2023. URL: https://huggingface.co/CompVis/stable-diffusion (visited on 03/02/2023).

A Title

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Technical
University of
Denmark

Brovej, Building 118
2800 Kgs. Lyngby
Tlf. 4525 1700

www.byg.dtu.dk