UDC 004.31 **doi: 10.26906/SUNZ.2022.4.114**

O. Rudenko, O. Bilokin

National University «Yuri Kondratyuk Poltava Polytechnic», Poltava, Ukraine

APPLICATION OF GENERATIVE DIFFUSION MODELS IN DIGITAL IMAGE CREATION

Abstract. There's been a significant surge in the popularity of generative networks over the last year. With public releases of such advanced models as DALL-E, Stable Diffusions, or GPT-3, anyone with modest, run-of-the-mill hardware can dabble in machine learning [3]. Diffusion models are inspired by non-equilibrium thermodynamics. Diffusion models are a subcategory of likelihood-based models. They are known to offer reliably scalable, high-fidelity images while retaining a stationary training objective. These models generate samples by graduallyremoving noise from a signal, and their training objective can be expressed as a reweighted variationallower bound [2]. This class of models already holds the state-of-the-art [6] on CIFAR-10 [3], butstill lags behind GANs on difficult generation datasets like LSUN and ImageNet. Nichol and Dhariwal [4] found that these models improve reliably with increased compute, and can produce high-qualitysamples even on the difficult ImageNet 256×256 dataset using an upsampling stack. However, theFID of this model is still not competitive with BigGAN-deep [5], the current state-of-the-art on thisdataset. Even more, these models are capable of producing an infinite amount of unique, high-quality images, human-like speech, and realistic music, indistinguishable from human-made ones at the first glance. The popularity of generative models has grown rapidly. Likelihood-based models might provide better performance in comparison to GANs. Diffusion models are a promising new category of likelihood models. Disco Diffusion is a combination of CLIP and ImageNet models. It can generate digital art based on text prompts. Numerous applications are possible for this model, such as the creation of video, animation and image content. Several distinctions have to be considered when choosing Disco Diffusion over GAN.

Keywords: GAN, generative adversarial networks, artificial intelligence, non-equilibrium thermodynamics, diffusion models, digital art, ImageNet model, WordNet.

Introduction

There's been a significant surge in the popularity of generative networks over the last year.

With public releases of such advanced models as DALL-E, Stable Diffusions, or GPT-3, anyone with modest, run-of-the-mill hardware can dabble in machine learning [3]. Even more, these models are capable of producing an infinite amount of unique, high-quality images, human-like speech, and realistic music, indistinguishable from human-made ones at the first glance.

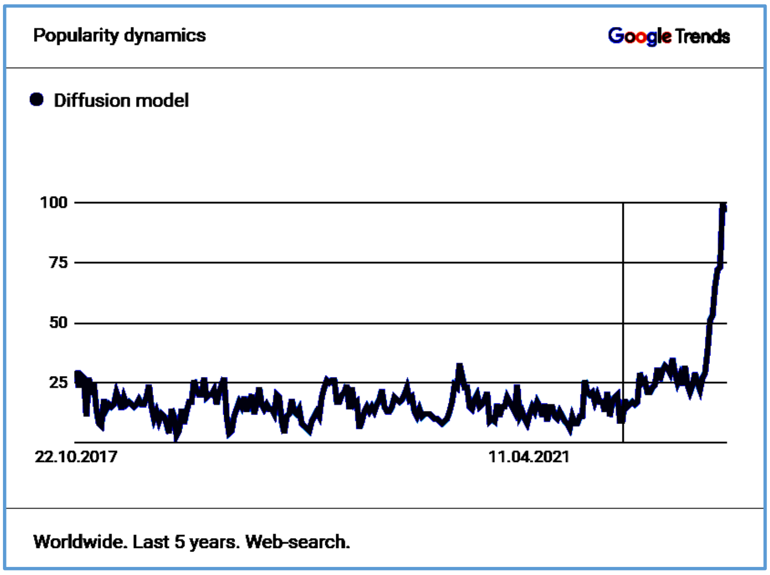


Fig. 1. Search trends over the last 5 years

No doubt, there is still much room for improvement beyondthe current state-of-the-art, and better generative models could have wide-ranging impacts on graphicdesign, games, music production, and countless other applications [1]. These models are created on the shoulders of their predecessors – GAN (Generative adversarial networks) and while those provide highly impressive results, the drawbacks of using GANs may prove to be too much when applied to other, yet unexplored domains. As an alternative, recent research on the usage of diffusion models was conducted to achieve the levels of quality that GANs possess, while evading the problems such as:

– lack of diversity, where GANs often perform worse than likelihood-based models;

– difficulties in scaling;

– difficulties in training, where GANs may often collapse if the selection of hyperparameters and regularizers proves to be of insufficient sophistication [6].

However, much work has been done to achieve GAN-like sample quality with likelihood-based models and, while these models capture more diversity and are typically easier to scale and train than GANs, they still fall short in terms of visual sample quality. Furthermore, except for VAEs, sampling from these models is slower than GANs in terms of wall-clock time [4, 10].

Diffusion models are inspired by non-equilibrium thermodynamics. Diffusion models are a subcategory of likelihood-based models. They are known to offer reliably scalable, high-fidelity images while retaining a stationary training objective. These models generate samples by graduallyremoving noise from a signal, and their training objective can be expressed as a reweighted variationallower bound [2]. This class of models already holds the state-of-the-art [6] on CIFAR-10 [3], butstill lags behind GANs on difficult generation datasets like LSUN and ImageNet. Nichol and Dhariwal [4] found that these models improve reliably with increased compute, and can produce high-qualitysamples even on the difficult ImageNet 256×256 dataset using an upsampling stack. However, theFID of this model is still not competitive with BigGAN-deep [5], the current state-of-the-art on thisdataset.

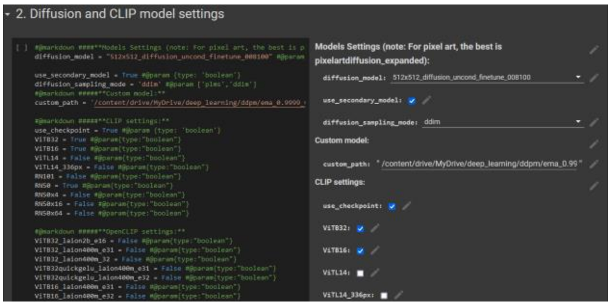
The main part of the article

This particular model, Disco Diffusion, is an amalgamation of OpenAI's ImageNet model combined with CLIP. Such conjunction allows us to connect text-based prompts with pictures, creating a fully-functional text-to-image model that can generate digital art from just a few sentences.

What is CLIP? CLIP is a transformer model created by Open AI to match text prompts with their respective image. The reasoning for the creation of CLIP is simple: If we can create models that accurately describe the text and we can create models that accurately describe images, then we are capable of joining the two to easily map image to text by combining similar descriptions [5].

And what is ImageNet? It is an image dataset organized according to the WordNet hierarchy, a foundation for the advancement of deep learning research and self-supervised (without the usage of human-labeled data) computer vision technologies. Each meaningful concept in WordNet, possibly described by multiple words or word phrases, is called a "synonym set" or "synset". There are more than 100,000 synsets in WordNet; the majority of them are nouns (80,000+). The goal of this service is to provide on average 1000 images to illustrate each synset. Images of each concept are quality-controlled and human-annotated. In its completion, ImageNet offers tens of millions of cleanly labeled and sorted images for most of the concepts in the WordNet hierarchy, creating a monumental basis for the training of image processing models. [7, 8].

Disco Diffusion is an open-source project, as is its larger counterpart – Stable Diffusions project. These models give permission under MIT license to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute and sublicense.



2. Diffusion and CLIP model settings

```
[ ] #@markdown ####**Models Settings (note: For pixel art, the best is p
    diffusion_model = "512x512_diffusion_uncond_finetune_008100" #@param

    use_secondary_model = True #@param {type: 'boolean'}
    diffusion_sampling_mode = 'ddim' #@param ['plms','ddim']
    #@markdown #####**Custom model:**
    custom_path = '/content/drive/MyDrive/deep_learning/ddpm/ema_0.9999_

    #@markdown #####**CLIP settings:**
    use_checkpoint = True #@param {type: 'boolean'}
    ViTB32 = True #@param{type:"boolean"}
    ViTB16 = True #@param{type:"boolean"}
    ViTL14 = False #@param{type:"boolean"}
    ViTL14_336px = False #@param{type:"boolean"}
    RN101 = False #@param{type:"boolean"}
    RN50 = True #@param{type:"boolean"}
    RN50x4 = False #@param{type:"boolean"}
    RN50x16 = False #@param{type:"boolean"}
    RN50x64 = False #@param{type:"boolean"}

    #@markdown #####**OpenCLIP settings:**
    ViTB32_laion2b_e16 = False #@param{type:"boolean"}
    ViTB32_laion400m_e31 = False #@param{type:"boolean"}
    ViTB32_laion400m_32 = False #@param{type:"boolean"}
    ViTB32quickgelu_laion400m_e31 = False #@param{type:"boolean"}
    ViTB32quickgelu_laion400m_e32 = False #@param{type:"boolean"}
    ViTB16_laion400m_e31 = False #@param{type:"boolean"}
    ViTB16_laion400m_e32 = False #@param{type:"boolean"}
```

Models Settings (note: For pixel art, the best is pixelartdiffusion_expanded):

diffusion_model: 512x512_diffusion_uncond_finetune_008100

use_secondary_model: ☑

diffusion_sampling_mode: ddim

Custom model:

custom_path: " /content/drive/MyDrive/deep_learning/ddpm/ema_0.99 "

CLIP settings:

use_checkpoint: ☑

ViTB32: ☑

ViTB16: ☑

ViTL14: □

ViTL14_336px: □

Fig. 2. Model configuration

The model allows the user to control the output via diffusion model presets, sampling modes, numerous CLIP, and OpenCLIP parameter presets, batch and configurations depending on the required art style and available hardware.

Upon receiving the CLIP configurations, the model then may be modified further with different values for image sizes, diffusion steps, rescaling and resolutions overriding previously selected presets. Then, additional parameters may be specified for the creation of video content, such as the number of steps, guidance scale, number of skipped steps, animation settings, and input files.

At last, the model receives a text prompt from the user. For demonstrational purposes, we will create a unique image based on a text prompt using a modified copy of the model with the following input:

```
"A gorgeous paintingofan isolated
lighthouse,shiningitslightacross a
tumultuousseaofredin the style of
bygregrutkowskiandthomaskinkade.",
"yellowcolorscheme".
```

Once the model is started, the generation of the prompted image begins. The model defines a random seed and iteratively creates.

As you can see (Fig. 3), the model starts by generating a raster of monochrome noise, where only vague shapes of the main landscape objects can be discerned.

Fig. 3. Initial Image, the first iteration of the target image

As the cycles advance, more and more details are generated by the model to create a distinctive picture.

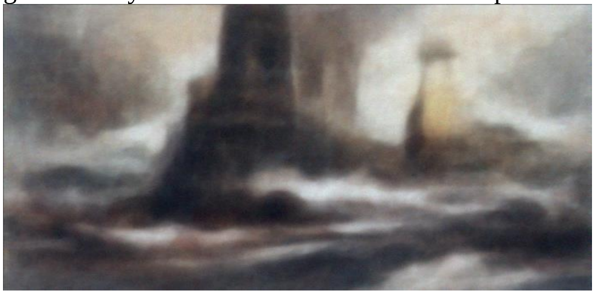


Fig. 4. Third iteration

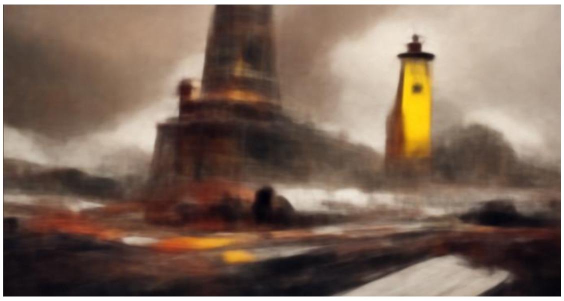


Fig. 5. Fifth iteration

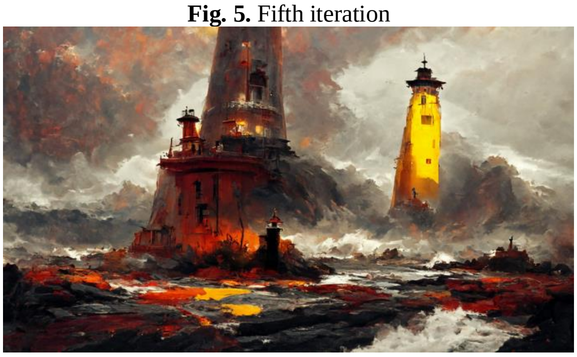


Fig. 6. Final Image

Conclusions

We have explored the possibilities, advantages, and downsides of using Disco Diffusion – a diffusion model made to generate videos, animations, illustrations, and artworks based on a plain text input.

References

1. Prafulla Dhariwal, Alex Nichol – Diffusion Models Beat GANs on Image Synthesis URL: https://arxiv.org/pdf/2105.05233.pdf
2. Sakib Shahriar - GAN Computers Generate Arts? A Survey on Visual Arts, Music, and Literary Text Generation using Generative Adversarial Network, URL: https://arxiv.org/ftp/arxiv/papers/2108/2108.03857.pdf
3. Ali Razavi, Aaron van den Oord, Oriol Vinyals – Generating Diverse High-Fidelity Images with VQ-VAE-2, URL: https://arxiv.org/pdf/1906.00446
4. Rewon Child Very Deep VAEs Generalize Autoregressive Models and Can Outperform Them on Images, URL: https://arxiv.org/pdf/2011.10650
5. Matthew Cateer – CLIP Prompt Engineering for Generative Art, URL: https://matthewmcateer.me/blog/clip-prompt-engineering/
6. Andrew Brock, Theodore Lim, J.M. Ritchie, Nick Weston - Neural Photo Editing with Introspective Adversarial Networks, URL: https://arxiv.org/pdf/1609.07093
7. Open AI Image GPT, URL: https://openai.com/blog/image-gpt/
8. ImageNet: About, URL: https://www.image-net.org/about.php
9. Google Trends, URL: https://trends.google.com/
10. Golovko G. V., Nikiforova K. M. Information systems use at Poltava national technical Yuri Kondratyuk University. *Control, navigation and communication systems*. 2018. Vol. 3. P. 103-105.

Received (Надійшла) 05.09.2022
Accepted for publication (Прийнята до друку) 09.11.2022

Використання генеративних дифузійних моделей машинного навчання у створенні зображень

О. Руденко, О. Білокінь

Анотація. За останній рік стався значний сплеск популярності генеративних мереж. Завдяки публічним випускам таких просунутих моделей, як DALL-E, Stable Diffusions або GPT-3, кожен із скромним, звичайним апаратним забезпеченням може спробувати машинне навчання [3]. Моделі дифузії натхненні нерівноважною термодинамікою. Дифузійні моделі є підкатегорією моделей на основі ймовірності. Відомо, що вони пропонують надійно масштабовані високоточні зображення, зберігаючи при цьому нерухомий тренувальний об'єкт. Ці моделі генерують вибірки шляхом поступового видалення шуму із сигналу, а їх мета навчання може бути виражена як перезважена варіаційна нижня межа [2]. Цей клас моделей уже відповідає найсучаснішому [6] на CIFAR-10 [3], але все ще відстає від GAN щодо складних наборів даних, таких як LSUN і ImageNet. Нікол і Дхарівал [4] виявили, що ці моделі надійно вдосконалюються зі збільшенням обсягу обчислень і можуть створювати високоякісні зразки навіть на складному наборі даних ImageNet 256 × 256 за допомогою стека підвищення дискретизації. Проте FID цієї моделі все ще неконкурентоспроможний із BigGAN-deep [5], поточним сучасним сучасним набором даних. Більше того, ці моделі здатні створювати нескінченну кількість унікальних високоякісних зображень, людської мови та реалістичної музики, яку на перший погляд неможливо відрізнити від рукотворної. Популярність генеративних моделей швидко зростає. Ймовірнісні моделі можуть забезпечити кращу продуктивність у порівнянні з GAN. Дифузійні моделі є новою перспективною категорією ймовірнісних моделей. DiscoDiffusion – це комбінація моделей CLIP та ImageNet. Вона може генерувати цифрові картини на основі текстових підказок. Для цієї моделі можливі численні застосування, такі як створення відео, анімації та графічного контенту. При виборі DiscoDiffusion замість GAN слід враховувати певні відмінності.

Ключові слова: GAN, generative ad versarial net works, штучний інтелект, дифузійні моделі, цифрове мистецтво, ImageNetmodel, WordNet.