



Computational Creativity Assignment: Image Synthesis and Inpainting with Diffusion Based Generative Models (GLIDE & CLIP)

Kumar Sindhurakshit , Id : 200854641

k.sindhurakshit@se21.qmul.ac.uk

School of EECS
Queen Mary University of London, UK

Abstract

Objective of this study is to research and develop diffusion based models for image synthesis. Generative adversarial networks (GANs) have been a research area of much focus in the last few years due to the quality of output they produce, however they do have some disadvantages like vanishing gradient, mode collapse and failure to converge in certain scenarios. Recent findings on diffusion based models have shown promising improvements in these aspects, a paper titled 'Diffusion Models Beat GANs on Image Synthesis' [\[1\]](#) by OpenAI researchers has shown that diffusion models can achieve image sample quality superior to the generative models, in related development Google AI introduced two connected approaches named Super-Resolution via Repeated Refinements (SR3) [\[4\]](#) and Cascaded Diffusion Models (CDM) [\[5\]](#) to improve the image synthesis quality for diffusion models last year. The deliverables of the project include a collab notebook which allows users to select different methods of diffusion described above for image synthesis and visualise image artifacts produced, it will be supplemented with report describing findings and scope of future work.

1. Introduction

In this project a photorealistic image to text generation and inpainting systems is implemented using diffusion based generative models. The system uses pretrained Open AI GLIDE (Guided Language to Image Diffusion for Generation and Editing) and CLIP (Contrastive Language–Image Pre-training).

To implement the project various approaches of text to image generation including Generative Adversarial Network (GAN), Variational Encoder (VAE) and diffusion based models were researched and analysed. The open AI GLIDE and CLIP were chosen because their enhanced technical capabilities including photo realism. This project is implemented in Google Colab notebook and designed to run in both GPU and non GPU environments.

As part of the project both system and results are analysed to understand current status and future improvements which are presented in results and future improvement sections of this report.

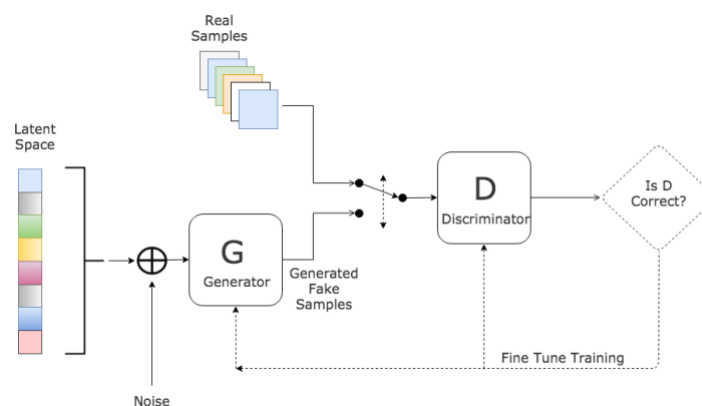
2. Background

One of the way people reflect their creativity is by visual representations, such as illustrations, paintings, and photographs. These artefacts can often be easily described in natural language, but generally require specialized skills and hours of labour to create. One of the research are of the computational creativity is to generate realistic visual representations from natural language to empower individual to create rich and diverse visual content with unprecedented ease. As of now there are four major approach to achieve generation of visual artefacts today.

1. Generative Adversarial Network (GAN) : A GAN consists of two models:

A discriminator D estimates the probability of a given sample coming from the real dataset. It works as a critic and is optimized to tell the fake samples from the real ones.

A generator G outputs synthetic samples given a noise variable input z (z brings in potential output diversity). It is trained to capture the real data distribution so that its generative samples can be as real as possible, or in other words, can trick the discriminator to offer a high probability.



Generative Adversarial Networks , Prof. Simon Colton , Computational Creativity (ECS7022P) Lecture 4, Part 1: Intro to GANs (source: <https://www.arxiv-vanity.com/papers/1801.04271/>)

These two models compete against each other during the training process: the generator G is trying hard to trick the discriminator, while the critic model D is trying hard not to be cheated. This interesting zero-sum game between two models motivates both to improve their functionalities.

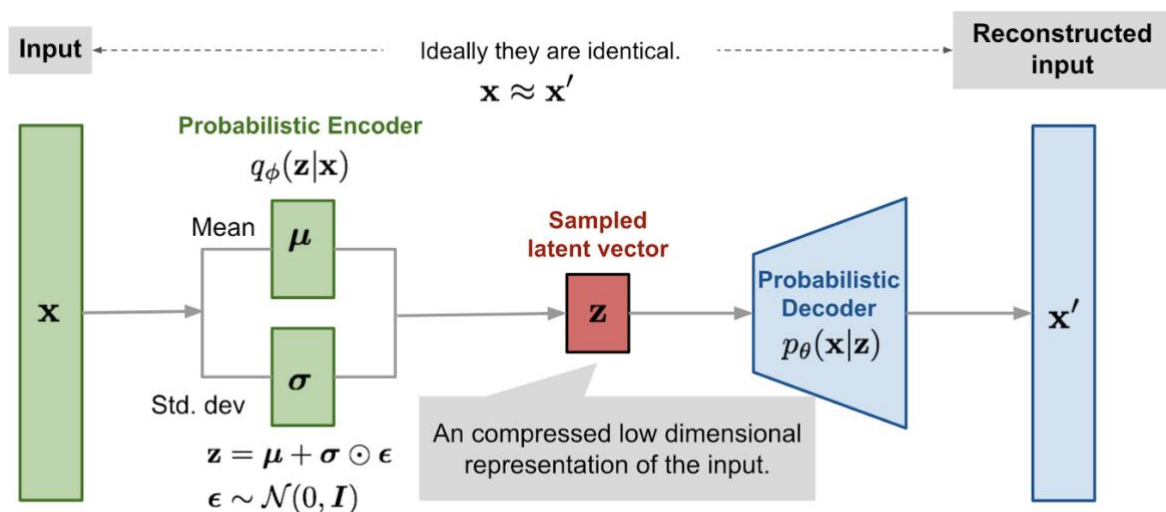
Problems in GANs

Although GAN has shown great success in the realistic image generation, the training is not easy; The process is known to be slow and unstable.

- Hard to achieve Nash equilibrium
- Low dimensional supports
- Vanishing gradient
- Mode collapse

Variational Encoder (VAE)

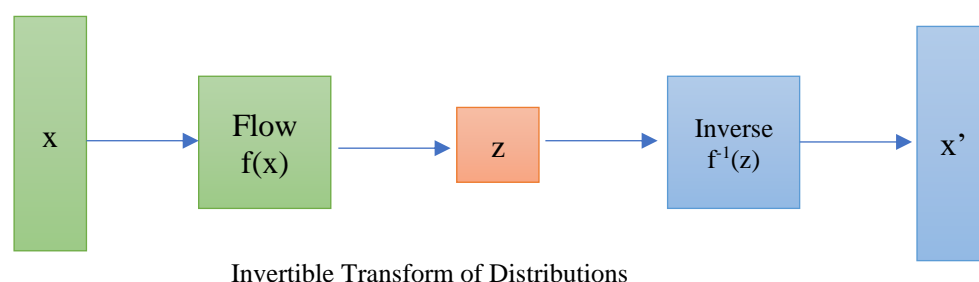
Traditional autoencoders learn to compress and reconstruct data but not really help with generating new data. This is where Variational Encoder (VAE) is very VAE learns the distribution of the data instead of just a compressed image, and by using the distribution, we can decode and generate new data. The encoder is trying to learn the parameters ϕ to compress data input x to a latent vector z , and the output encoding z is drawn from Gaussian density with parameters ϕ . As for the decoder, its input is encoding z , the output from the encoder. It parametrizes the reconstructed x' over parameters θ , and the output x' is drawn from the distribution of the data. Open AI



Source :<https://www.researchgate.net/profile/Ryan-Sander-2/publication/339323231/figure/fig2/AS:859731560263680@1581987385598/An-Illustration-of-a-Variational-Autoencoder-VAE-Note-that-our-data-is-not-shown.ppm>

Flow-based Generative Models

A flow-based generative model is constructed by a sequence of invertible transformations. Unlike other two, the model explicitly learns the data distribution $p(x)$ and therefore the loss function is simply the negative log-likelihood.



Diffusion-based Generative Models

Diffusion models are inspired by non-equilibrium thermodynamics. They define a Markov chain of diffusion steps to slowly add random noise to data and then learn to reverse the diffusion process to construct desired data samples from the noise. Unlike VAE or flow models, diffusion models are learned with a fixed procedure and the latent variable has high dimensionality (same as the original data)

Diffusion

Diffusion is the process by which particles of one substance spread out through the particles of another substance. Diffusion is how smells spread out through the air and how concentrated liquids spread out when placed in water. Diffusion happens on its own when the particles spread out from an area of high concentration, where there are many of them, to areas of low concentration where there are fewer of them.

Diffusion models

Several diffusion-based generative models have been proposed with similar ideas underneath, including

1. Diffusion probabilistic models
2. Noise-conditioned score network
3. Denoising diffusion probabilistic models
4. Ablated Diffusion Model (ADM)
5. Ablated Diffusion Model (ADM -Guided) used in Open AI GLIDE
6. Cascaded Diffusion Models (CDM)

3. System Description

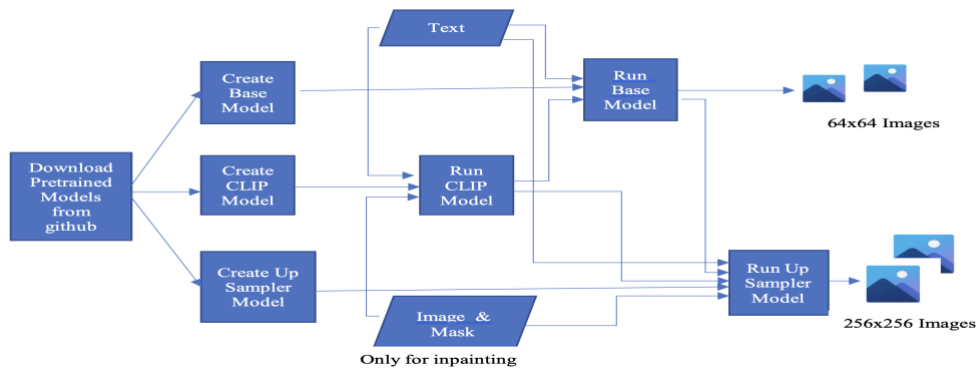
GLIDE Architecture

The GLIDE architecture is combination of three components an Ablated Diffusion Model (ADM) trained to generate a 64 x 64 image, a text model (transformer) that influences that image generation through a text prompt, and an up sampling model that takes our small 64 x 64 images to a more interpretable 256 x 256 pixels. The first two components interact with one another to guide the image generation process to accurately reflect the text prompt, and the latter is necessary to make the images we produce more easy to interpret.



System Architecture

Below diagram describes system design and architecture -



4. Discussion

Started computer creativity course with a bit of escapism if computers can be really creative with overwhelmed feeling complexity of technology, however as we progressed through the course, somehow it grabbed attention and imagination. This particular project of using diffusion based model has been further challenging as this being latest there is not much of literature and reference implementations available.

However started with research papers and OpenAI reference implementation, developed project code in structured way and added interactivity GUI elements with Jupyter widget library have led to creation of interesting system which is close to level 1 specifically the inpainting functions with image upload and mask control has been fun to play with. Some of the generated photo realistic images has been amazing (Please refer Appendix section for some interesting for some examples). On the other hand system times also generated some weird and disturbing images which are not sharable but this is very rare. In general images generated are appealing and entertaining. Achieved photo realism is surprisingly of high quality.

There is scope of significantly improving system both technically and functionally which is discussed in details section 5.

There is also technical issues observed when model is repeatedly called for inpainting masked areas sometimes simply coming from previous generation, i.e. system is memorising previous steps.

5. Conclusions and Future Work

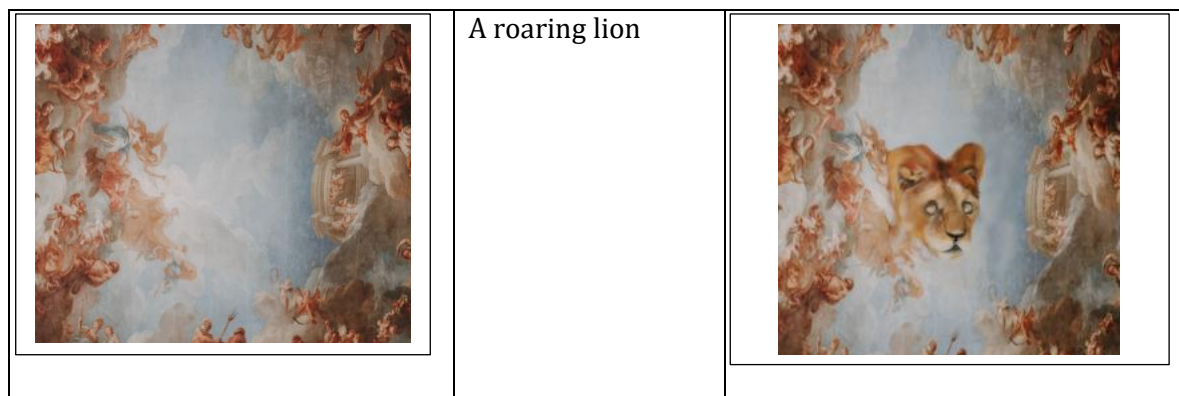
High Level Issues

- No fitness function for self-evaluation
- Limited Autonomy
- This is filtered version of GLIDE with limited dataset output will be much better with complete data set.
- Inpainting is not very effective for small mask window size.

Weak computational creativity

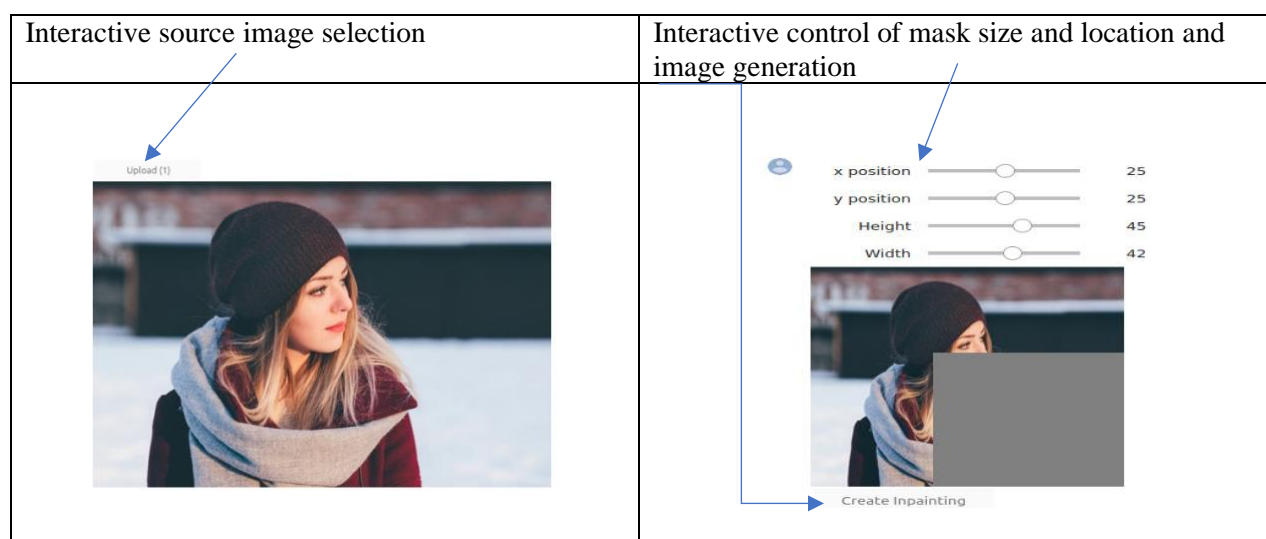
System inpainting capability with interactive control of source image along with mask location and size for inpainting can be categorised as weak computational creativity as it is able to produce some awe-inspiring artefacts of real value and beauty with some human intelligence applied like location of the mask. Knowing the fact that system use random pixel from the source image for inpainting help creating better artefacts.

For example in the below inpainting as there is natural similarity between painting overall colour and natural colour of lion and knowledge that while the empty space in the centre of painting could be a good place for inpainting lead to generation of below artefact –

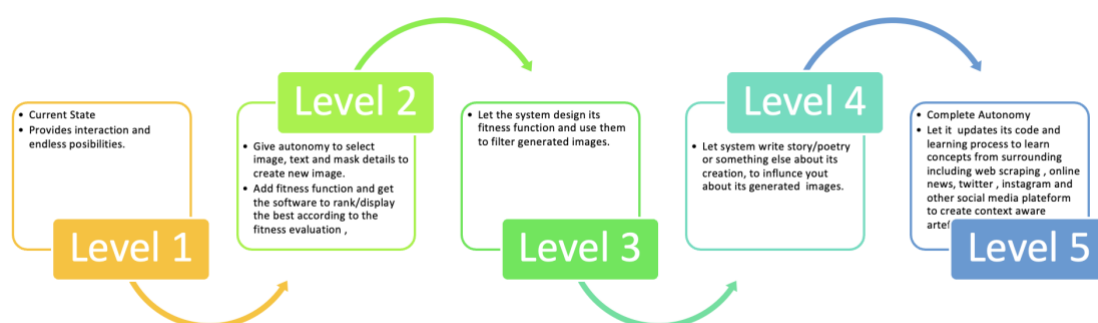


Level 1 - Generative System

This system can be best described as Level 1 system and interactive GUI for inpainting allows to play with a number of source images and changes in applied mask location , mask size along with different guiding text increases the fun and variety. With this it also becomes obvious that space of input/parameters is very big with huge possibility space making it level 1 generative system.



Scope of Future Work (Functional)











Future Scope of Work (Technical)


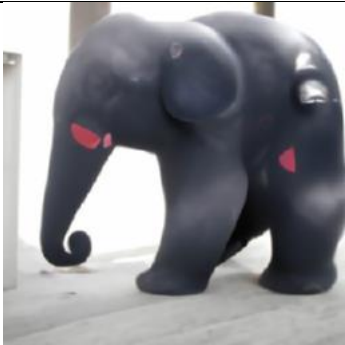


1. Currents system is based on pretrained model , next step is to develop its training model.
2. GLIDE is based on filtered dataset for training , train it with extended data set.
3. For inpainting system selects random pixels from source images , altering algorithm to also consider colour space of targeted concepts may lead to better artefacts or complete latent space , this requires more R&D and experimentation.
4. Add Fitness function for self-evaluation and ranking of generated images.
5. Add web scrapping to learn and generate context aware artefacts

6. Results

In-Painting

S.N.	Source Image	Guiding Text	In painted Image
1.	 <p>A painting from unsplash.com</p>	A male roaring lion	
2.	 <p>QMUL Great Hall</p>	Painting of Minerva the goddess of wisdom	
		A beautiful chandelier	
		Dolphin jumping out of the water	

Text to Image with Guided CLIP

S. N.	Text	Generated Image
1.	A beautiful girl walking on the beach in red gown	
2	An elephant dancing on the street	
3	A village on the bank of river with lots of trees	
4	A painting of a fox in the style of starry night	

7. References

1. [Diffusion Models Beat GANs on Image Synthesis](#) , Prafulla Dhariwal, Alex Nichol, Cornell University
2. [What are diffusion Models.](#) Lil Long
3. [Multimodal Image Synthesis and Editing: A Survey](#) , angneng Zhan, Yingchen Yu, Rongliang Wu, Jiahui Zhang, Shijian Lu
4. [Image Super-Resolution via Iterative Refinement](#) , Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, Mohammad Norouzi
5. [Cascaded Diffusion Models for High Fidelity Image Generation](#) , Google AI
6. [Comparative Study on Generative Adversarial Networks](#), Saifuddin Hitawala
7. [Understanding VQ-VAE \(DALL-E Explained1\).](#) [Charlie Snell](#)