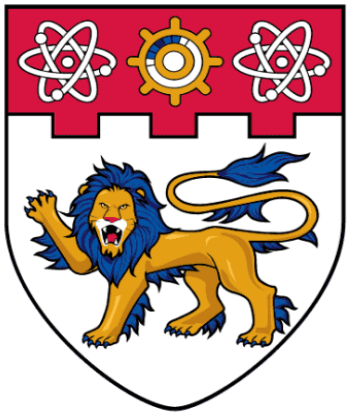


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

AI6121 Computer Vision

Direct Reading and Literature Review

Generative Modeling

Rohin Kumar Maheswaran
Email : rohinkum001@e.ntu.edu.sg
Matric No : G2303513K

Abstract

Generative Modeling is focused on creating models that can generate new data that is similar to existing data.These models are trained on a dataset and then used to produce data samples that have similar statistical properties to the training data. Generative models have a wide range of applications, including image generation, text generation, speech synthesis, and more. Generative modeling has been tested using Big-Bi-GAN models, although Autoregressive models have never been done at that time (2020).

In this directed reading and literature review, the topic of generative modeling will be introduced in the first chapter, followed by an exploration of both conventional and state-of-the-art approaches in generative modeling techniques. The research paper titled "**Generative Pretraining from Pixels**" will be reviewed and analyzed to establish its significance as one of the most groundbreaking papers in the modern generative modeling domain.

Contents

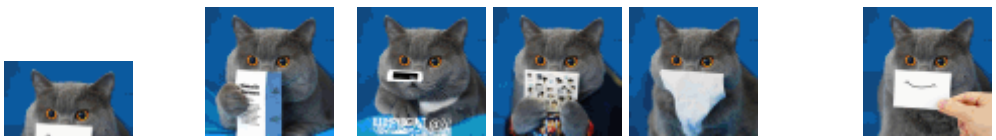
1. Introduction

1.1 Background :

Significant strides in Natural Language Processing have been achieved by employing pretraining transformer models with self-attention layers on extensive unlabeled text data, followed by fine-tuning for specific downstream tasks. Within the realm of auto-regressive language modeling, the process involves utilizing a sequence of n tokens to predict a masked token placed at the end of the sequence. Notably, OpenAI researchers have raised a fundamental question: Do more advanced generative models inherently possess more effective representations? This question is explored through an intriguing demonstration where a large model, sharing the same validation loss and pretraining tasks as a smaller model, reveals that the smaller counterpart exhibits superior representation transferability, particularly in the context of ImageNet classification.

In the research paper titled "Generative Pretraining from Pixels" by OpenAI, a pioneering approach is unveiled. It ingeniously involves partitioning images into two halves, reserving one part as the untouched original image. The challenge posed to the model is to precisely reconstruct what's missing from the other half, and remarkably, it accomplishes this task with meticulous attention to individual pixels. This methodology draws striking parallels with the operational principles of language models but is astutely adapted to the domain of pixel-level image comprehension. In essence, it effectively serves as a bridge between visual and linguistic cognition.

The noteworthy aspect of this paper is that while the pixel-by-pixel method itself isn't groundbreaking and has been utilized previously, the focus here is on exploring the extent to which generative models can be advanced through pretraining.



Input **Generated Images** **Original Image**

Figure 1. It displays both the input, the model-generated outputs, and the original image, which is positioned on the far right side. **[1]**

The images displayed in Figure **[1]** represent a stage before the final outcome, marking the pre-training phase. The true essence of this paper lies in its core objective: What happens

when we employ a large-scale pretraining process to generate high-quality images like these or to seamlessly complete images, and subsequently fine-tune the model for a classification task.

1.2 Drive and Purpose :

The drive and purpose of this paper is to investigate the potential of generative pre-training methods for unsupervised image representation learning. The authors aim to re-evaluate the effectiveness of unsupervised pre-training for image data, which has been successful in the field of Natural Language Processing (NLP) but has not been extensively studied for images. They propose to use generative image modeling to learn high-quality unsupervised image representations and compare them with recent self-supervised methods.

Unsupervised pre-training has played a significant role in the resurgence of deep learning. In the mid-2000s, approaches such as the Deep Belief Network and Denoising Autoencoder were commonly used in neural networks for computer vision and speech recognition. It was believed that learning the data distribution would also lead to beneficial features for subsequent supervised modeling. However, advancements in activation functions, initialization techniques, and normalization strategies have reduced the need for pre-training to achieve strong results **[2]**.

Despite the success of unsupervised pre-training in NLP, its application to image data has not been extensively explored. The authors highlight recent advancements in self-supervised approaches for image modeling and the potential of generative image modeling to learn high-quality unsupervised image representations. They acknowledge the challenges of modeling images due to their higher dimensionality, noise, and redundancy compared to text. Therefore, the authors propose to re-examine generative pre-training methods for images and compare them with recent self-supervised methods

1.3 Types of Generative Models :

Generative Adversarial Networks (GANs): GANs are a type of generative model that consists of a generator network and a discriminator network. The generator network generates synthetic samples, while the discriminator network tries to distinguish between real and synthetic samples. The two networks are trained in a competitive manner, with the goal of the generator network generating samples that are indistinguishable from real data. **[7]**

Variational Autoencoders (VAEs): VAEs are another type of generative model that consists of an encoder network and a decoder network. The encoder network maps input data to a latent space, while the decoder network reconstructs the input data from the latent space. VAEs are trained to maximize the likelihood of the input data and learn a meaningful latent representation.

Autoregressive Models: Autoregressive models are generative models that model the joint probability distribution of the data by decomposing it into a product of conditional probabilities. These models generate samples by sequentially predicting each element of the data based on the previously generated elements. Examples of autoregressive models include PixelCNN and WaveNet.

Likelihood-based Training Objectives: Likelihood-based training objectives are used to train generative models by maximizing the likelihood of the observed data. This involves estimating the parameters of the generative model that maximize the probability of generating the observed data. Likelihood-based objectives include maximum likelihood estimation (MLE) and maximum a posteriori (MAP) estimation.

2. Architectures :

2.1 Previous Architecture :

This paper shares a lot of similarities with the PixelCNN model. PixelCNN also took on the task of this autoregressive pixel modeling with the images. But in PixelCNN they use convolutional neural network prior, in the neural network architecture that used to assign the probability mask to the new masked out pixel and it is not used in the scale of ImageGPT. There is a massive difference of using the CNN prior where we have this local kernel going across the image. They use this local prior to make prediction on this pixel.

The previous SOTA (State-Of-The-Art) in this simultaneous task of generative modeling and representation learning for taking representation out and fine tuning on the ImageNET classification is the Big-Bi-GAN model.

The Big-Bi-GAN model does not use the Big-GAN architecture to produce data but it also maps the data back into latent space with this encoder and then the latent **($\hat{z}$)** and **(z)** have to be a part of the discriminator that is telling its real or fake. It has to learn a good representation of going from latents into data. This representation is useful for image classification. Big GAN and Style GAN are two architectures that was previously (2020) used for GAN Modelling. It was pretty successful but not as successful as ImageNET and not on the same scale as ImageGPT.

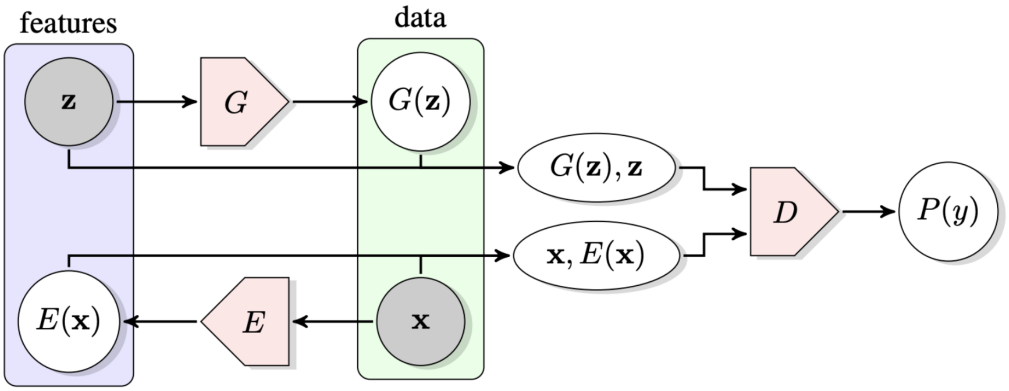


Figure 2. Structure of Bi - GAN architecture **[5]**

SOTA Architecture :

eDiff-I :

AI is a battleground for all tech conglomerates so the SOTA (State-Of-The-Art) is rapidly changing day to day. As per current standards, eDiff-I model by NVIDIA is considered as SOTA (2023 September).

The architecture of the eDiff-I model consists of a modified U-net architecture **[3]**. It incorporates global conditioning by adding projected pooled CLIP text embeddings, CLIP image embeddings, and time step embeddings as inputs. Cross-attention blocks are used to perform cross-attention between image embeddings and conditioning embeddings. The keys in the cross-attention layers are the concatenation of pre-pooled CLIP text embeddings, T5 embeddings, and pooled CLIP image embeddings. A learnable null embedding is also included for cases where conditioning embeddings are not needed.

For the super-resolution models, a modified version of the Efficient U-net architecture is used **[4]**. The SR1024 model, which generates images at 1024x1024 resolution, utilizes the block structure of the Efficient U-net architecture. Self-attention layers are removed to improve efficiency during inference, and only cross-attention layers are retained. During training, the SR1024 model is trained using random patches of size 256x256 and applied to 1024x1024 resolution during inference.

In this paper, the authors propose a model called eDiff-I that can generate high-quality images from text descriptions. The model uses a technique called diffusion, which involves gradually refining a noisy image to generate a final image.

To improve the image generation process, the authors introduce an ensemble of expert denoisers. These are specialized models that focus on different aspects of the image generation process. By using multiple expert denoisers, the model can capture different stages of image synthesis, such as capturing the text prompt in the early stages and focusing on visual details in the later stages.

The model also takes advantage of text embeddings, which are representations of the text descriptions. These embeddings help the model understand the meaning of the text and generate images that align with the given descriptions. The authors experiment with different types of text embeddings, such as CLIP text and image embeddings, to improve the quality of the generated images.

Additionally, the paper introduces a feature called "paint-with-words." This feature allows users to specify the spatial layout of objects in the generated images by selecting and scribbling phrases on the image. This gives users more control over the generated outputs and allows for more creative expression.

Overall, the eDiff-I model achieves state-of-the-art results in generating images from text prompts. It improves the image generation process by using an ensemble of expert denoisers, leveraging text embeddings, and providing user control through the "paint-with-words" feature.

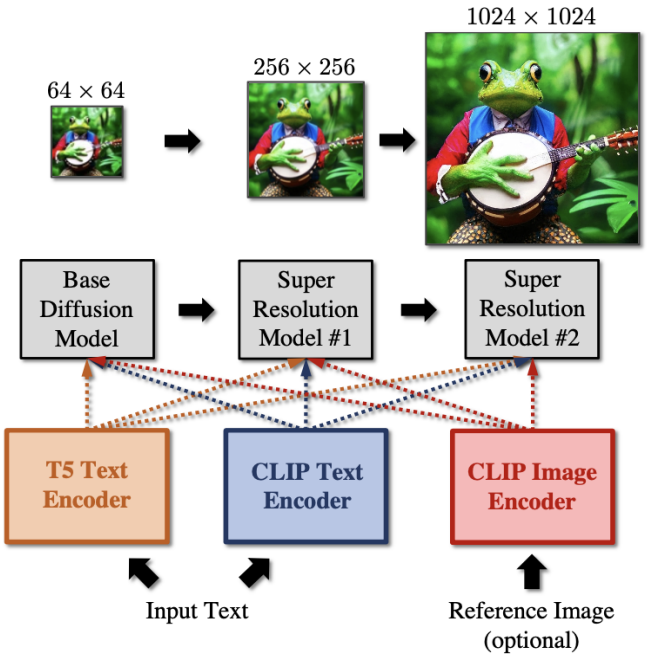


Figure 3. eDiff-I consists of a base diffusion model that gener- ates images in 64→64 resolution. This is followed by two super- resolution diffusion models that upsample the images to 256→256 and 1024→1024 resolution, respectively, referred to as SR256 and SR1024 through the paper. All models are conditioned on text through both T5 and CLIP text embeddings. eDiff-I also allows the user to optionally provide an additional CLIP image embed- ding. This can enable detailed stylistic control over the output **[4]**

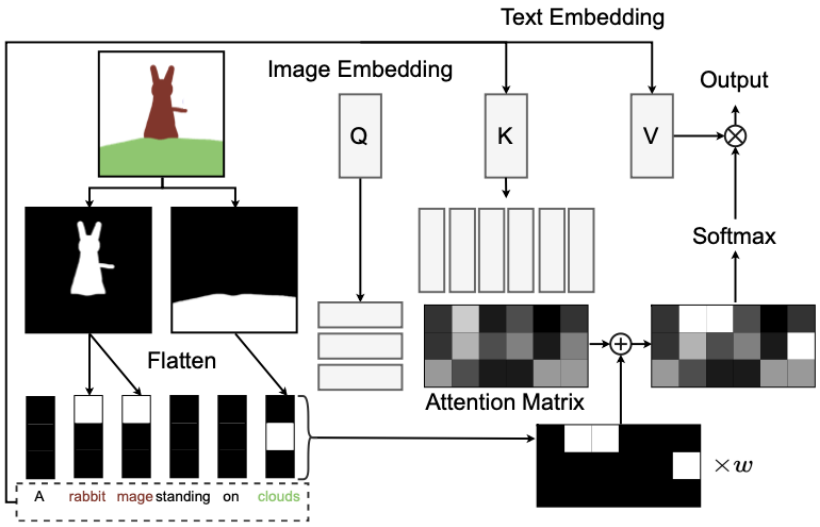


Figure 4. Illustration of the proposed *paint-with-words* method. The user can control the location of objects by selecting phrases (here "rabbit mage" and "clouds"), and painting them on the canvas. The user-specified masks increase the value of corresponding entries of the attention matrix in the cross-attention layers. **[4]**

3. Overview : Generative Pretraining from pixels

In this research paper, the authors tackle the challenge of processing images by first transforming them into a format suitable for their model. Initially, an image from the ImageNet dataset is unrolled, resulting in an incredibly large matrix of pixels, [(244)^2 pixels X 3] (representing the three color channels). However, this size is impractical, so they take a strategic approach.

To make it more manageable, they begin by downsizing the image to either 32 by 32 or 64 by 64 dimensions, and then proceed to unroll it. In essence, they traverse the pixels from left to right, creating a sequential representation. This process is inspired by the model's original design, which was primarily tailored for handling text sequences.

To further simplify matters, they reduce the image's three color channels to a single channel, creating their unique representation where each index signifies a specific color. Eventually, this transformation results in a compact (32)^2 representation of the image. At this point, they have two choices for subsequent processing:

1) Autoregressive
2) BERT

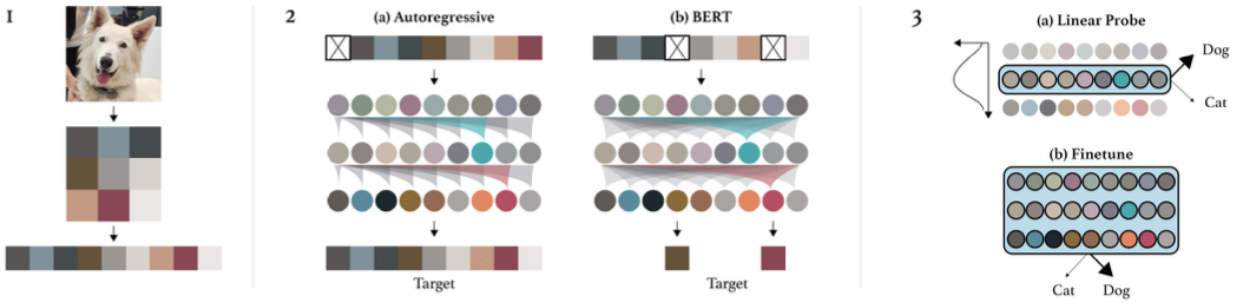


Figure 5. This image represents flow of autoregressive and BERT models for pixel generation **[1]**

1) Autoregressive :

The main idea is to predict the next pixel in a sequence, using a one-way (unidirectional) modeling approach inspired by GPT-2-style pretraining.

2) **BERT :**

They introduce perturbations to select pixels and task the model with reconstructing these altered portions. This involves bidirectional attention, unlike BERT where predicting pixels in a single forward pass is more complex. This situation illustrates a tradeoff between BERT and autoregressive models.

Once this process is completed, they employ two approaches for assessment:

1) Linear Probing
2) Fine Tuning

1) **Linear Probing :**

Linear probing is a method where we forward propagate to one of the layers and create classes and train on a linear classifier on the classes alone and classify it.

2) **Fine Tuning :**

You can create classes on top of the existing architecture and train on the small dataset. This is called fine tuning.

Dataset :

1) CIFAR-10
2) CIFAR - 100
3) STL -10

The pretraining is done on the imageNET and they transferlearn or train on these small datasets.

Linear probe accuracy :

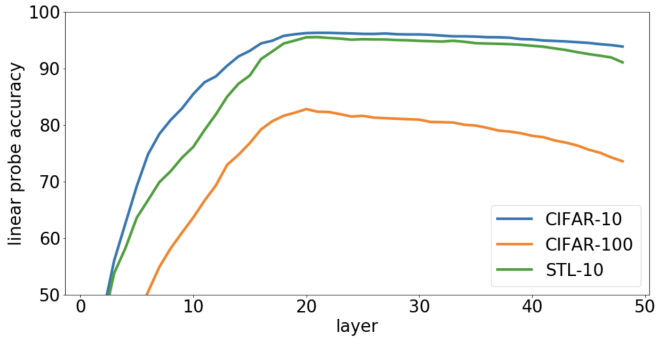


Figure 6. Representation quality depends on the layer from which we extract features. In contrast with supervised models, the best representations for these generative models lie in the middle of the network. We plot this unimodal dependence on depth by showing linear probes for iGPT-L on CIFAR-10, CIFAR-100, and STL-10. **[1]**

The linear probe accuracy achieves a range of 95% to 96% in this model, while the state-of-the-art (SOTA) benchmark in 2020 stands at 99%. Although it falls short of the 99% mark, this performance is still considered commendable. In typical classification models, the final layer often yields the best results, but in this case, the intermediate layers emerge as the key contributors to superior representation. Notably, the quality of linear probing diminishes as one progresses to higher layers.

To generate consistent pixel predictions, it becomes necessary to incorporate a sense of global image information, which is crucial for accurate classification. The neural network's initial layer, reminiscent of Convolutional Neural Networks (CNNs), focuses on low-level feature transformation. Conversely, the ultimate layer is primarily concerned with the precise prediction of individual pixels. The underlying hypothesis posits that the middle layers encode and transform global information, subsequently shaping the outcomes of the final layers. This elucidates why the most effective representations are found within the intermediate layers.

Findings :

They have a range of model sizes, including iGPT-L (60 layers), iGPT-M (48 layers), and iGPT-S (32 layers), all in line with the GPT-2 scale. The y-axis in their analysis represents linear probe accuracy, indicating the layer where the best performance is achieved after probing and recording accuracy.

A noticeable pattern emerges: as the validation loss decreases, linear probe accuracy consistently improves. This connection mirrors what's seen in text models, where lower perplexity in language models results in higher-quality representations for downstream tasks. In this context, achieving a lower validation loss corresponds to better performance in image modeling tasks. It's important to note that, despite reaching the same validation loss, larger models outperform their smaller counterparts in these tasks.

They have primarily focused on supervised transfer learning for fine-tuning. However, they suggest that further fine-tuning efforts could lead to an impressive 99% accuracy on CIFAR-10, which would be on par with the best models from 2020.

It's worth mentioning that iGPT-L employs less data augmentation compared to Gpipe and follows a supervised transfer learning approach.

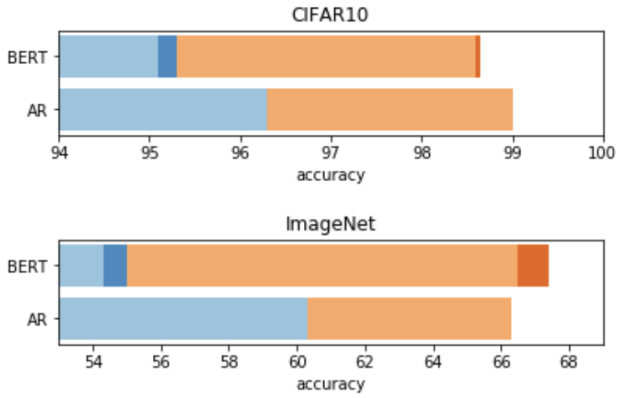


Figure 7. CIFAR - 10 is pretrained with imageNET and ImageNET itself is pretrained like a larger collection from internet. All the pre training is done without labeling. The blue part is what you can get after linear probe and orange is what you get after fine tuning. Ie On top of linear probe. Fine tuning is always done at the end. They have tried fine tuning inbetween but the results are best when they take the layers from the last. The important thing is coming up with higher level representation and then once you fine tune. You are probably able to push that representation through the end by your training signals. **[1]**

Challenges :

The paper delves into a significant challenge posed by this technique. Self-attention introduces a quadratic complexity when applying query and key matrices simultaneously. When considering the full input resolution (full RGB color) of an ImageNet image, it becomes practically impossible to fit even a single layer of this into memory. To circumvent this issue, the authors employ context reduction and employ downsampling along with k-means color clustering, effectively sidestepping this bottleneck.

Furthermore, the authors acknowledge that their experiments have shed light on areas where their approach can be enhanced. They acknowledge that their method does not match the performance of supervised methods, indicating that there is still ample room for improvement in unsupervised image representation learning.

Future Work :

The authors of the paper outline several avenues for future research. They suggest that exploring alternative architectures, such as employing convolutional neural networks (CNNs) instead of transformers, holds promise as a fruitful research direction. **[6]**

Furthermore, the authors propose investigating the influence of diverse pre-training objectives and training strategies on the quality of acquired representations. They posit that combining multiple objectives or integrating domain-specific knowledge has the potential to enhance the performance of unsupervised image representation learning. **[6]**

In addition, the authors stress the importance of conducting comprehensive evaluations of unsupervised methods on extensive datasets and examining their adaptability to various tasks and domains. They also underscore the significance of studying the transferability of representations obtained through unsupervised pre-training to downstream applications. **[6]**

In conclusion, the authors underscore the imperative for ongoing research and development in unsupervised image representation learning to advance state-of-the-art capabilities and address the evolving challenges in this field.

Conclusion :

The paper concludes that generative image modeling shows promise for learning unsupervised image representations. The authors demonstrate that their method of predicting pixels achieves state-of-the-art results on low-resolution datasets and is competitive with other self-supervised methods on high-resolution settings.

However, the authors acknowledge that there are still challenges to be addressed in unsupervised image representation learning. Their method does not perform as well as

supervised methods, indicating that there is room for improvement in terms of the quality of learned representations. To address these challenges, the authors suggest several directions for future work. One direction is to explore different architectures, such as using convolutional neural networks (CNNs) instead of transformers. This could potentially improve the performance of unsupervised image representation learning.

Another area for future research is to investigate the impact of different pre-training objectives and training strategies on the quality of learned representations. The authors suggest that combining multiple objectives or incorporating domain-specific knowledge could lead to better unsupervised image representation learning.

Furthermore, the authors emphasize the need for more comprehensive evaluation of unsupervised methods on large-scale datasets. They also highlight the importance of studying the transferability of representations learned from unsupervised pre-training to downstream tasks.

In conclusion, while generative image modeling shows promise for unsupervised image representation learning, there are still challenges to be addressed. Further research is needed to improve the quality of learned representations and to explore different architectures, pre-training objectives, and training strategies.

7. References

[1] Image GPT by OpenAI : https://openai.com/research/image-gpt
[2] Paine, T. L., Khorrami, P., Han, W., and Huang, T. S. An analysis of unsupervised pre-training in light of recent advances. arXiv preprint arXiv:1412.6597, 2014.
[3] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In *Proc. NeurIPS*, 2021.
[4] EDiff-I arXiv:2211.01324 [cs.CV]
[5] Adversarial Feature Learning : arXiv:1605.09782 [cs.LG]
[6] Rives, A., Goyal, S., Meier, J., Zitnick, C. L., & Susskind, J. (2019). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. bioRxiv.
[7] Generative Adversarial Networks: arXiv:1406.2661 [stat.ML]