Review of generative models

Ruixi Wang

College of Information Science and Engineering, Northeastern University, ShenYang, Liaoning Province, 110819, China

20205397@stu.neu.edu.cn

Abstract. The advancement of generative AI models has been remarkable since 2022. Several visually appealing generative AI models have been introduced to the public, including those for text and image generation. Despite being generated by large-scale neural networks and deep learning algorithms through extensive training, generative models are capable of achieving average or above-average quality and creativity in many fields, such as painting and literature. This paper will examine some of the AI models currently available, delve into their underlying principles and histories, and provide insight into what the future may hold. With the advancement of technology, we can expect to see even more innovative and creative applications in the future.

Keywords: Generative models, artificial intelligence, NLP, image generation, deep learning.

1. Introduction

Generative models have their origins in the development of the Boltzmann machine, which dates back to 1983[1]. Boltzmann machines encountered challenges in training and were not widely adopted at the time. Nevertheless, they laid the foundation for subsequent generative models. In 2006, Geoffrey Hinton successfully reduced the dimensionality of data using Deep Belief Networks (DBN) based on Restricted Boltzmann Machines (RBM), thereby solving the problem of difficult training[2]. In 2013, Kingma and Welling used Variational Autoencoder (VAE) to learn the latent representation of high-dimensional data. The encoder maps the data to the latent representation space, and the decoder maps the latent representation back to the original data space[3]. This algorithm is also capable of generating new data in the potential space. In 2014, Goodfellow I used Generative Adversarial Networks (GAN) to make two neural networks play games with each other to learn the generation model of data distribution[4]. A GAN consists of a generator and a discriminator. The generator takes random noise as input to produce samples that are similar to the training data, while the discriminator determines whether a sample is from the generator or real data. By playing a game between these two neural networks, GAN can generate diverse and high-quality samples. In 2015, Ashish Vaswani proposed a novel, simple network architecture based solely on attention mechanisms, dispensing with recurrence and convolutions entirely[5]. The Transformer model has demonstrated excellent performance in natural language processing tasks, including text classification, sentiment analysis, and question-answering. Today, many generative models have been developed and improved upon the previous models. For instance, the GPT model was originally based on a one-way Transformer decoder architecture.

This paper provides a comprehensive summary of the current state and future development direction of generative model research, as well as an overview of current research hotspots and challenges, in order to enrich the literature review in the field of generative models.

2. Popular generative model

This section will provide an introduction to popular generative models, covering training strategies, evaluation methods, and performance indicators. Additionally, it will explore the principles, benefits, and challenges of these models, as well as their performance in various application domains. Specifically, the models covered in this section are GPT, Diffusion Models, and Dall-E, with one natural language model and two image-generating models.

2.1. GPT

The first GPT model was released in 2018, and it explored a semi-supervised approach for language understanding tasks using a combination of unsupervised pre-training and supervised fine-tuning. GPT-1 uses a Transformer based architecture for sequence modeling. In this way, models can learn intrinsic structure and semantic representation. GPT-1 adopts a Transformer-based architecture for sequence modeling, allowing the model to learn the intrinsic structure and semantic representation. By doing so, it acquires significant world knowledge and the ability to process long-range dependencies, which can be effectively transferred to solving discriminative tasks such as natural language inference, question answering, sentence similarity, and classification [6].

Compared to GPT-1, GPT-2 has a model size and pre-training data that are nearly 10 times larger, more parameters, a deeper network structure, and the ability to handle more complex language tasks, resulting in better language learning knowledge. Additionally, GPT-2 has a more varied set of pre-training tasks, which improves its performance in different scenarios. GPT-2 also introduces "style transfer" technology, enabling the control of the tone of the generated text and making it more personalized and richer [7].

GPT-3 has a larger scale, which enables it to handle more complex natural language tasks with greater expressive and reasoning abilities. The significant number of parameters in GPT-3 can provide ample prior knowledge for learning from a small number of samples, allowing the model to learn from just a few examples. This type of learning is referred to as "few-shot learning." GPT-3 can identify common structures between tasks and transfer previously learned knowledge to new tasks seamlessly[8].

GPT-4, the most recent iteration of the GPT model, distinguishes itself from its predecessor, GPT-3, by being a multimodal input model. GPT-4 can accept not only text input but also image input, allowing it to handle any task involving vision or language. It can be generated from input text and images, such as generating image descriptions, image questions and answers, or image to code. GPT-4 used an alternating dense and locally banded Sparse attention pattern, similar to Sparse Transformer. This allows the model to maintain high performance while reducing computation and memory costs. Despite its capabilities, GPT-4 has similar limitations to earlier GPT models, it is not fully reliable[9].

"Hallucinations" may be caused by models relying too much on patterns and biases in the data, ignoring the true meaning of the input, or by the lack of quality and diversity of data resulting in models that cannot cover all situations and knowledge, or by the complexity and opacity of the model making it difficult to understand and control. To reduce hallucinations, more diverse and high-quality data can be used to train models to reduce data bias and noise. Various evaluation metrics can be used to measure the degree of "Hallucinations" of the model, such as factual consistency, semantic relevance, logical coherence, etc. Post-processing techniques can be used to correct or filter the "Hallucinations" output of the model, such as knowledge retrieval, counter-sample, dialogue strategy, etc. In addition, interpretive techniques can be used to analyze the causes of the model's "Hallucinations", such as attention visualization, gradient analysis, contrast learning [10].

2.2. Diffusion models

Image synthesis is one of the computer vision fields with the most spectacular recent development, Diffusion models have achieved record performance in many applications, including image synthesis and video generation [11].

Diffusion models are inspired by non-equilibrium thermodynamics; they define a Markov chain of diffusion steps to slowly add random noise to the data and then learn to reverse the diffusion process to construct the desired data sample from the noise. Unlike VAE or stream models, diffusion models are learned through a fixed process and latent variables have a high dimension (the same as the original data) [12,13].

In 2022, Robin Rombach proposed a Latent Diffusion model. The Latent Diffusion Model is a novel generative model that can be used to synthesize high-resolution images of natural scenes. It combines the diffusion process and autoregressive modeling approaches, leveraging the reversible diffusion process to efficiently sample and perform inference. The model also incorporates the autoregressive model to capture local structural information, thus allowing it to model both global and local features of an image [11].

The Latent Diffusion Model exhibits robust controllability and interpretability. The noise level and diffusion process can be effectively manipulated to achieve variations in image sharpness and detail. By leveraging the sampling and reverse diffusion processes of the hidden variables, the model can generate high-resolution images with both high naturalness and artistic merit [11].

Stable diffusion is a text-to-image model based on Latent Diffusion Models (LDMs). Specifically, it benefits from Stability AI's computing resources and LAION's data resources [14]. In fact, the diffusion model can process image, text, audio, and other types of data, and the generated samples have the advantages of high quality and diversity.

The disadvantage of Diffusion Models is that the training time is long, a lot of computing and data resources are needed, and the parameter setting of the model is complicated. But now the computer's power is enough to produce better pictures.

In conclusion, Diffusion Model is a very promising generation model. It is expected to play a role in the future of AI and image generation.

2.3. Dall-E

Since the pioneering work of Reed, the Generative adversarial text to image synthesis has been an active area. Dall-E is a text-to-image synthesis model[15].

The DALL-E 2 uses a Transformer architecture similar to that of GPT-3 but modified to fit the image generation task. In particular, the authors divide the image into small blocks and encode the pixel values of those blocks into discrete tokens, which are then used as inputs to the model. DALL-E 2 is trained using a large amount of data containing text descriptions and associated images. The data comes from the Internet and includes various types of images and text information [16]. The relationship between the text semantics in DALL-E 2 and its relative visual images is learned by another OpenAI model, CLIP (Contrastive Language-Image Pre-training). CLIP is trained on hundreds of millions of images and their associated text, learning how a given piece of text relates to an image. The DALL-E 2 uses an improved GLIDE model that uses projected CLIP text embedding in two ways. The first is to add them to GLIDE's existing time-step embed, and the second is to create four additional context tags that connect to the output sequence of the GLIDE text encoder. The diffusion model acts as a prior of DALL-E 2 in order to map from the text encoding of the image title to the image encoding of the corresponding image [17].

In general, the CLIP text encoder maps the image description to the presentation space. The diffused priors are then mapped from the CLIP text encoding to the corresponding CLIP image encoding. Finally, a modified version of the GLIDE generation model maps from the representation space to the image space by reverse diffusion, generating one of many possible images [17].

Because of the source of its data, DALL-E 2 may generate images with biased or discriminatory content. Although you can provide some text, DALL-E 2 may produce some text that is inconsistent with the description [16,18].

In conclusion, DALL-E 2 is capable of generating many types of images, including images that do not exist in reality, images that have certain properties (such as certain colors, shapes, or styles), and images that satisfy abstract text descriptions [16,18].

3. Discussion

The generative model still leaves much to be desired. Despite recent advances in deep learning, some of the most advanced models still don't produce realistic enough output in all situations. In addition, generative models still face challenges when handling large amounts of data, so more research is needed to improve their efficiency and scalability. In fact, GPT4 has a relatively high accuracy level for question-answering tasks, especially in the field of text tasks. Fortunately, at the recent 2023 GTC conference, Nvidia will support the development of generative AI with even more powerful chips.

The world is undergoing a significant transformation with the advent of generative AI. By 2022, AI-generated images will have a major impact on the painting industry, resulting in the phasing out of low-quality paintings from the market and raising the threshold of the industry. This creates new opportunities for individuals who can utilize these generative AI models and produce high-quality paintings that exceed those of the average painter. Additionally, the original artist may choose to modify or enhance the AI-generated image. Some artists enter completely unrelated and abstract labels into the model to get a completely unknown image, and use the composition to inspire their own paintings. However, some artists may feel that their copyright is being infringed upon during training, which has led to the launch of an anti-AI-generated image campaign on various painting websites in February 2023. Creative industries like painting have been slower to adopt generative AI, and it may take some time before artists fully embrace this technology. The issue of how to use AI-generated content while protecting the copyright of these works as training sets is an important consideration for future generative models.

The latest version of GPT-4 is poised to bring about a significant transformation in the way we interact with computers, offering faster, more accurate, and smarter search results. Previously, becoming a programmer required a foundational understanding of programming languages and concepts. However, with GPT-4, anyone can become a programmer and create using the open-source code provided by the system. GPT-4 has revolutionized the field of writing, replacing many tedious and repetitive tasks that were previously considered necessary. For instance, clerks who create PowerPoint presentations can now rely on GPT-4-based copilots to generate high-quality PPTs based on their requirements, freeing them from creating mundane and uninteresting content.

In early testing, GPT-4 is able to use the tools with very minimal instruction and no demonstrations and then make appropriate use of the output (note how the second search result contains potentially conflicting information, and GPT-4 is still able to infer the right answer). For example, GPT-4 knows when to use a calculator and can use them efficiently. There may be future results from training on some other tools. At the 2022 Mathematical Olympiad, GPT-4 demonstrated remarkable problem-solving abilities by correctly proving a highly creative problem. GPT4 can also act as your personal butler, helping you automate your planning [19].

Occupations were assessed according to their correspondence to GPT ability, combined with human expertise and GPT-4 classification. The findings suggest that about 80% of the US workforce could have at least 10% of their work tasks affected by the introduction of GPT, while about 19% of workers could have at least 50% of their tasks affected. The effect spans all wage levels, with higher-paying jobs likely to be more at risk. It is important to note that this effect is not limited to industries with higher productivity growth recently [20]. In summary, these models may have significant economic, social, and policy implications.

4. Conclusion

Generative models have come a long way since the development of the Boltzmann machine in 1983. With advancements in deep learning and the emergence of models like GPT, Diffusion Models, and Dall-E, the field of generative AI has made tremendous progress in recent years. These models have demonstrated their potential in various applications, such as natural language processing, image synthesis, and video generation.

However, there are still challenges to be addressed in order to make these models more reliable, efficient, and scalable. Issues such as hallucinations, data bias, and the protection of copyrights for original content creators need to be tackled to ensure the ethical and responsible use of generative AI technologies.

As these models continue to improve and evolve, they will undoubtedly transform numerous industries, from the creative arts to programming and beyond. The potential economic, social, and policy implications of these advancements are vast, and it is essential for researchers, practitioners, and policymakers to work together to harness the power of generative AI while addressing its challenges and mitigating potential risks.

In conclusion, the future of generative AI is promising, with endless possibilities for innovation and growth. By understanding the state of current research, exploring the principles, benefits, and challenges of popular generative models, and discussing their potential impact on various domains, this paper provides a comprehensive overview of the exciting developments in the field of generative models and paves the way for future research and applications.

References

[1] Hinton, G. E., & Sejnowski, T. J. (1983). Optimal perceptual inference. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 448-453.

[2] Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. Science, 313(5786), 504-507.

[3] Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

[4] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems, 2672-2680.

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

[6] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

[7] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog, 1(8), 9.

[8] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Sutskever, I. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

[9] OpenAI. (2023). GPT-4 Technical Report. arXiv preprint arXiv:2303.08774.

[10] Ji, Z., Lee, N., Frieske, R., Chen, L., & Wang, W. Y. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1-38.

[11] Rombach, R., Blattmann, A., Lorenz, D., Kingma, D. P., & Welling, M. (2022). High-resolution image synthesis with latent diffusion models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10684-10695.

[12] Yang, L., Zhang, Z., Song, Y., & Wu, F. (2022). Diffusion models: A comprehensive survey of methods and applications. arXiv preprint arXiv:2209.00796.

[13] Weng, L. (2021, July 11). What are Diffusion Models? Lil'Log. https://lilianweng.github.io/posts/2021-07-11-diffusion-models/.

[14] Sean. (2022, November 25). Stable Diffusion principle interpretation. Zhihu column.

https://zhuanlan.zhihu.com/p/583124756.

[15] Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). Generative adversarial text to image synthesis. International conference on machine learning, 1060-1069.

[16] OpenAI. (2021). DALL-E. OpenAI. https://openai.com/research/dall-e.

[17] O'Connor, R. (2022, April 21). Here's how the DALL-E 2 works. Zhihu column. https://zhuanlan.zhihu.com/p/502389739.

[18] OpenAI. (2021). DALL-E. OpenAI. https://openai.com/research/dall-e.

[19] Bubeck, S., Chandrasekaran, V., Eldan, R., Klivans, A., Raghunathan, A., & Zeevi, A. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712.

[20] Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). GPTs are GPTs: An early look at the labor market impact potential of large language models. arXiv preprint arXiv:2303.10130.