Available online www.ejaet.com

European Journal of Advances in Engineering and Technology, 2023, 10(12):21-23

Research Article ISSN: 2394 - 658X

Understanding Generative AI Architectures: A comprehensive Overview and Classification

Sai Kalyana Pranitha Buddiga

Boston, USA
pranitha.bsk3@gmail.com

ABSTRACT

Generative Artificial Intelligence (Generative AI) encompasses a diverse set of architectures and models that enable machines to create new data samples that resemble real-world examples. This white paper provides a comprehensive overview of Generative AI architectures, categorizing them into distinct classes based on their underlying principles, structures, and applications. By exploring the various Generative AI architectures, readers will gain insights into the state-of-the-art techniques and their potential applications across different domains.

Key words: Generative AI, Deep Learning, AI, Data Generation, Autoencoders, Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs)

1. INTRODUCTION

The emergence of Generative AI has transformed the landscape of artificial intelligence, empowering machines to create authentic and innovative data samples. This introduction offers insights into Generative AI architectures, their pivotal role in modern machine learning exploration, and their profound impact on the trajectory of AI-driven innovations. By delving into the core concepts and classifications of Generative AI architectures, readers will be equipped with the knowledge needed to understand the breadth and depth of this rapidly evolving field [1].

2. CLASSIFICATION OF GENERATIVE AI ARCHITECTURES

This involves categorizing various models and frameworks based on their fundamental principles, structures, and applications. One prominent classification scheme distinguishes between traditional models and more modern approaches. Traditional models include autoencoders, which aim to learn compact representations of input data by compressing it into a latent space and then reconstructing it. Gaussian Mixture Models (GMMs) are another traditional technique used for modeling complex data distributions by representing them as a mixture of Gaussian components. Hidden Markov Models (HMMs) are commonly employed for sequential data generation tasks, such as speech recognition and natural language processing.

In contrast, modern Generative AI Architectures encompass cutting-edge techniques that have revolutionized the field of artificial intelligence. This includes Generative Adversarial Networks (GANs), which consist of two neural networks, a generator, and a discriminator, trained adversarial to generate realistic data samples. Variational Autoencoders (VAEs) leverage probabilistic inference to learn a latent space representation of input data, enabling the generation of novel samples. Autoregressive Models, such as PixelCNN and WaveNet, generate data by modeling the conditional probability distribution of each data point given previous data points. By categorizing Generative AI Architectures into traditional and modern classes, researchers can gain a deeper understanding of the diverse range of techniques available for data generation and synthesis [2-4].

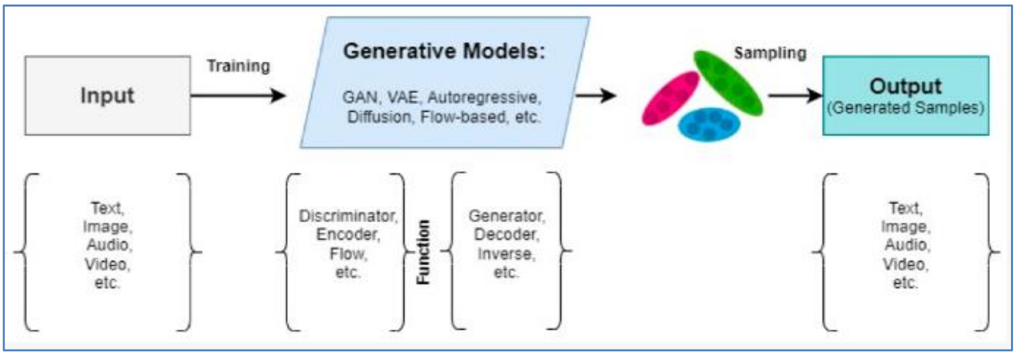

Figure 1: GenAI Models

3. OVERVIEW OF GENERATIVE AI ARCHITECTURE

Generative Artificial Intelligence systems encompass various models and frameworks designed to produce synthetic data that resembles authentic data found in the real world. Different architectures are employed in GenAI based on the specific function and use case, yet they commonly share certain fundamental elements and guiding concepts:

3.1. Generator Network

At the core of GenAI systems is a generator network tasked with creating artificial data samples. This network starts with an input of random noise or a latent space representation and processes it into an output that is akin to authentic training data. The specific neural network architecture within the generator can range from dense networks to CNNs or RNNs, according to the required task.

3.2. Discriminator Network

In architectures like GANs, there is a discriminator network that works to tell apart real data from synthetic creations. Trained simultaneously with the generator, the discriminator learns to identify what's real and what's generated. The competition between the generator trying to fool the discriminator, and the discriminator trying to remain accurate, drives the improvement of the quality of the generated data [4].

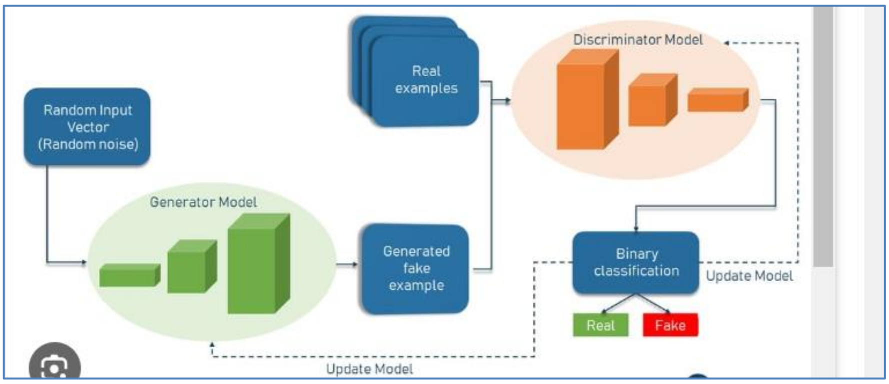

Figure 2: GAN Architecture

3.3. Latent Space Representation

An essential feature of GenAI is the encoding of data within a latent space, which is a compressed representation capturing the essence of the data. This encoding allows the generator to make new, significant outputs. For instance, VAEs use latent space to guide the generation process and allow for a highly controlled and interpretable synthesis of data.

3.4.1 Training Technique

Training these systems generally involves refining a loss function that measures how well the generated data compares to the real thing. In the context of GANs, the generator aims to decrease the discriminator's accuracy, while the discriminator works to increase its ability to separate real from fake [4]. Other models like autoencoders focus on minimizing the difference between the input and the recreated outputs.

3.5.1 Variants and Improvement:

As GenAI technology has evolved, various offshoots and improvements have emerged to overcome hurdles. These include conditional GANs, which generate data based on certain conditions, and GANs that enhance image resolution gradually. Innovations such as attention mechanisms and more sophisticated loss functions have been introduced to GenAI structures to improve their efficiency and stability.

Overall, GenAI architectures involve a tapestry of models and methods dedicated to crafting synthetic data that is convincingly realistic, with wide-ranging implications in AI and creative industries [5]. They utilize neural networks, latent space concepts, and advanced training strategies to excel in tasks like image generation, text creation, and even music production, leading to breakthroughs in the scope.

4. CONCLUSION

As we conclude our exploration of Generative AI architectures, we reflect on the rich landscape of techniques and applications that have emerged in this field. Generative AI architectures offer a powerful toolkit for creating new and realistic data samples, with implications for creativity, innovation, and problem-solving across diverse domains. By providing a comprehensive overview and classification of Generative AI architectures, this white paper aims to inspire further research, experimentation, and application of these transformative technologies.

REFERENCES

[1]. Z. Epstein, A. Hertzmann, L. Herman, R. Mahari, M. R. Frank, M. Groh, H. Schroeder, A. Smith, M. Akten, J. Fjeld, H. Farid, N. Leach, A. Pentland, and O. Russakovsky, "Art and the science of generative AI: A deeper dive," arXiv:2306.04141 [cs.AI], June 7, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2306.04141.

[2]. A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta and A. A. Bharath, "Generative Adversarial Networks: An Overview," in IEEE Signal Processing Magazine, vol. 35, no. 1, pp. 53-65, Jan. 2018, doi: 10.1109/MSP.2017.2765202.

[3]. Y. Hong, U. Hwang, J. Yoo, and S. Yoon, "How Generative Adversarial Networks and Their Variants Work: An Overview of GAN," arXiv:1711.05914 [cs.LG], submitted on Nov. 16, 2017, revised Jul. 27, 2018 (this version, v7), latest version Feb. 28, 2019 (v10). [Online]. Available: https://doi.org/10.48550/arXiv.1711.05914.

[4]. S. Bengesi, H. El-Sayed, M. K. Sarker, Y. Houkpati, J. Irungu, and T. Oladunni, "Advancements in Generative AI: A Comprehensive Review of GANs, GPT, Autoencoders, Diffusion Model, and Transformers," arXiv:2311.10242 [cs.LG], Nov. 17, 2023, revised Nov. 21, 2023 (this version, v2). [Online]. Available: https://doi.org/10.48550/arXiv.2311.10242.

[5]. R. Gozalo-Brizuela and E. C. Garrido-Merchán, "A survey of Generative AI Applications," arXiv:2306.02781 [cs.LG], June 5, 2023, revised June 14, 2023 (this version, v2). [Online]. Available: https://doi.org/10.48550/arXiv.2306.02781.