

Learn Python Generative AI

*Journey from autoencoders to
transformers to large language models*

Zonunfeli Ralte
Indrajit Kar



www.bpbonline.com

First Edition 2024

Copyright © BPB Publications, India

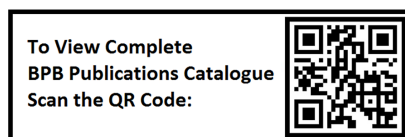
ISBN: 978-93-55518-972

All Rights Reserved. No part of this publication may be reproduced, distributed or transmitted in any form or by any means or stored in a database or retrieval system, without the prior written permission of the publisher with the exception to the program listings which may be entered, stored and executed in a computer system, but they can not be reproduced by the means of publication, photocopy, recording, or by any electronic and mechanical means.

LIMITS OF LIABILITY AND DISCLAIMER OF WARRANTY

The information contained in this book is true to correct and the best of author's and publisher's knowledge. The author has made every effort to ensure the accuracy of these publications, but publisher cannot be held responsible for any loss or damage arising from any information in this book.

All trademarks referred to in the book are acknowledged as properties of their respective owners but BPB Publications cannot guarantee the accuracy of this information.



Dedicated to

My beloved Parents:

R. Zohmingthanga and Sangthangseii

– Zonunfeli Ralte

My beloved Parents:

Avijit Kar and Puspa Kar

– Indrajit Kar

About the Authors

- **Zonunfeli Ralte**, a seasoned professional with a Master's in Business Administration and Economics, boasts 15 years of experience in Analytics, Finance, and AI. Currently, she is the CEO and Founder of RastrAI, while also serving as a Principal AI Consultant, developing GenAI applications for diverse industries. Zonunfeli has an impressive academic contribution with 6 IEEE research papers including **Large Language Models (LLM)**, Deep learning and computer vision, 3 of which received best paper awards. Her multifaceted expertise and leadership make her a notable figure in the AI community. Additionally, she has filed 1 patent in GenAI.
- **Indrajit Kar**, a master's graduate in Computational Biology from Bengaluru, also holds a Bachelor's in Science from the same institution with more than two decades of experience in AI and ML. He is an experienced intrapreneur, having built AI teams at Siemens, Accenture, IBM, and Infinite Data Systems. Presently, he is the AVP and Global Head of AI and ML leading AI research (ZAIR) and Data Practices. Indrajit has published 22 research papers across IEEE, Springers, Wiley Online Library, and CRC press, covering topics like LLM, Computer Vision, NLP, and more. He has 14 patents, including Generative AI. He is a mentor for startups and a recipient of multiple awards, including the 40 Under 40 Data Scientists award. He is also author of 2 AI books.

About the Reviewers

- ❖ **Utkarsh Mittal** is a Machine Learning manager at Gap Inc., a global retail company. He has more than ten years of practice experience in machine learning automation and is a leader of big AI-based database projects. He received his Master's in Industrial Engineering with a Supply Chain and Operations Research major from Oklahoma State University, USA. He is closely associated with research groups and editorial boards of high-profile International Journals and research organizations. He is passionate about solving complex business challenges and encouraging innovation through upcoming technologies. He is a Senior member of the IEEE Computer Society.
- ❖ **Arun Naudiyal** wears many hats, but they all share a common thread - a love for building, learning, and driving meaningful impact. As a Senior Product Engineer at a Big 4 Firm, his eight years of experience translate into a potent blend of expertise. He's the go-to for crafting and deploying Machine Learning pipelines on AWS, wielding frameworks like Kubernetes and Sagemaker with masterful hands. Arun's passion extends beyond code. He's an avid learner, constantly upskilling in MLOps and beyond. This thirst for knowledge led him to explore the fascinating world of Large Language Models. Under his guidance, teams have crafted projects that revolutionized document summarization for EdTech, unearthed customer sentiment like a treasure hunter, and even built an AI code assistant that boosts programmer productivity.

Acknowledgements

- **Zonunfeli Ralte:** To my family, especially my parents, sisters, and cousins, your unwavering encouragement and belief in my abilities have been the bedrock of my journey. Your support has been a guiding light, empowering me to pursue and accomplish this endeavor.

I also wish to express my heartfelt appreciation to the people of Mizoram, particularly the community of Ramthar Veng. Our rich culture and vibrant spirit have been a constant source of motivation and have deeply influenced my perspectives and writing.

A special acknowledgment goes to BPB Publications and my co-author Indrajit Kar for their patience and trust in my vision. Your flexibility in allowing the book to be published in multiple parts has been crucial in adequately covering the expansive and evolving field of AI.

Lastly, I thank my companies for providing an environment that fosters learning and growth. The opportunities to explore and develop GenAI applications have been fundamental in accumulating the knowledge shared in this book.

To all, your hidden and visible support has shaped this journey in countless ways, and for that, I am forever grateful.

- **Indrajit Kar:** I extend my deepest appreciation to my family, particularly my parents, wife, in-laws and children, whose steadfast encouragement and unwavering belief in my abilities have formed the cornerstone of my journey. Your support has illuminated my path, empowering me to pursue and fulfill this endeavor with confidence and dedication.

I must also express my profound gratitude to BPB Publications for their patience and trust in my vision. Their flexibility in allowing this book to be published in multiple segments has been pivotal in thoroughly addressing the broad and dynamic landscape of AI.

Furthermore, I am immensely thankful to my companies for creating an environment that nurtures learning and growth. The opportunities they have provided to delve into and develop GenAI applications have been instrumental in gathering the insights shared in this book.

To everyone involved, both in visible and unseen ways, your support has profoundly shaped this journey. For this, I am eternally grateful.

Preface

Learn Python Generative AI is an extensive and comprehensive guide that delves deep into the world of generative artificial intelligence. This book provides a thorough understanding of the various components and applications in this rapidly evolving field. It begins with a detailed analysis, laying a solid foundation for exploring generative models. The combination process of different generative models is discussed in depth, offering a roadmap to understand the complexities involved in integrating various AI models and techniques.

The early chapters emphasize the refinement of TransVAE, an advanced variational autoencoder, showcasing improvements in its encoder-decoder structure. This discussion sets the stage for a broader examination of the evolution of AI models, particularly focusing on the incorporation of the SWIN-Transformer in generative AI.

As the book progresses, it shifts focus to the practical applications of generative AI in diverse sectors. In-depth chapters explore its transformative potential in healthcare, including applications in hospital settings, dental, and radiology, underscoring the impact of AI in medical diagnostics and patient care. The role of GenAI in retail and finance is also thoroughly examined, with a special emphasis on corporate finance and insurance, demonstrating how AI can revolutionize customer engagement, risk assessment, and decision-making.

Each sector-specific chapter is enriched with real-world examples, challenges, and innovative solutions, offering a comprehensive view of how generative AI is reshaping various industries. The concluding chapters synthesize the key learnings from all topics, providing insights into the future trajectory of generative AI.

Chapter 1: Introducing Generative AI - The objective of this chapter is to provide a comprehensive understanding of generative models, including an overview of generative models, a comparison of discriminative vs generative models, an introduction to the types of discriminative and generative models, as well as their strengths and weaknesses. By the end of the content, readers should be able to differentiate between discriminative and generative models, understand the different types of each, and make informed decisions about which type of model is most appropriate for their needs.

Chapter 2: Designing Generative Adversarial Networks - In this chapter, the objective is to delve into the multifaceted landscape of GANs by comprehensively exploring various

types of GANs and their intricate architectures. By the end of this chapter, readers will be equipped with a solid understanding of the architecture, equations, and crucial design factors associated with different GAN variants. The chapter will dissect discriminator and generator losses, shed light on pivotal GAN types, including Vanilla GAN, Deep Convolutional GAN, Wasserstein GAN, Conditional GAN, CycleGAN, Progressive GAN, StyleGAN, and Pix2Pix, and address the major challenges encountered in designing effective GAN architectures. Through an in-depth analysis of each architecture, readers will gain the knowledge necessary to make informed decisions when selecting and designing GANs for various generative tasks.

Chapter 3: Training and Developing Generative Adversarial Networks - The objective of this book chapter is to provide readers with a comprehensive understanding of the process of training and tuning GANs, including the latest techniques and best practices for improving the stability and performance of GAN models.

Chapter 4: Architecting Auto Encoder for Generative AI - The primary goal of this chapter is to explore the fascinating world of autoencoders in the context of generative AI. We will delve into the inner workings of autoencoders, discussing their architectural variations, training strategies and their applications in generating diverse and high-quality outputs across various domains. Furthermore, we will examine advanced techniques that leverage autoencoders, such as Variational AutoEncoders (VAE) and Generative Adversarial Networks (GAN), which push the boundaries of generative AI even further.

Throughout this chapter, and the next, we will also discuss the key challenges associated with autoencoders for generative tasks, including issues like mode collapse, blurry outputs, and training instability. We will explore solutions and strategies to mitigate these challenges, providing practical insights and recommendations for building robust and effective generative models using autoencoders.

By the end of this chapter, readers will have gained a comprehensive understanding of autoencoders as a powerful tool in the realm of generative AI. They will have a solid grasp of the fundamental concepts, practical considerations, and cutting-edge advancements that can enable them to apply autoencoders effectively in their own projects and unlock the potential of generative models to create realistic and novel outputs.

Chapter 5: Building and Training Generative Autoencoders - The key objectives of this chapter are to provide the reader with a deep understanding of autoencoders and their applications. By the end of this chapter, readers will gain a comprehensive understanding of the concept of latent space and its significance in autoencoders, explore the concept of dual input autoencoders and their usefulness in handling missing values and multi-

modal data, and familiarize themselves with various loss functions commonly used in autoencoders and their role in training and reconstruction.

The readers will also learn about potential issues during training, such as overfitting, vanishing gradients, and noisy data, along with strategies to mitigate them, discover optimization techniques specific to autoencoders for effective model training and performance enhancement, as well as understand the differences between autoencoders and variational autoencoders and their respective benefits.

Lastly, the reader will acquire the knowledge and skills to leverage autoencoders in practical scenarios for data representation, generation, and anomaly detection.

Chapter 6: Designing Generative Variation Auto Encoder - By the end of this chapter, the reader will be able to understand the fundamental differences between VAEs and traditional AEs. We will also explore the network architecture of VAEs, including the encoder and decoder networks, and their role in learning latent representations. The reader will also gain insight into the mathematical principles underlying VAEs, including the reparameterization trick and the ELBO objective function.

The chapter will then move to advanced techniques in VAEs, such as employing different prior distributions, utilizing various forms of the encoder network, and handling missing or incomplete data. We will also discover methods for interpreting the latent space of a VAE and visualizing its representations, explore the generative capabilities of VAEs by generating novel samples using the decoder network, and lastly, acquire the necessary knowledge and skills to apply VAEs in practical applications, including image generation, natural language processing, and anomaly detection.

By achieving these key objectives, readers will develop a comprehensive understanding of VAEs and be able to leverage their power and flexibility in various domains, ultimately enhancing their ability to learn and generate meaningful representations from complex data.

Chapter 7: Building Variational Autoencoders for Generative AI - By the end of this chapter, the reader will have explored various architectural choices, including convolutional or Non convolution networks, to handle complex dependencies in VAEs. We will also investigate the impact of KL divergence and different prior distributions on the generative process of VAEs, and develop strategies to effectively handle missing or incomplete data within the VAE framework. The reader will also understand the role of loss functions and address potential issues during training to ensure stable convergence, as well as optimize VAE performance and generative capabilities for diverse data modalities.

By achieving these key objectives, readers will develop a comprehensive understanding of VAEs and be able to leverage their power and flexibility in various domains, ultimately enhancing their ability to learn and generate meaningful representations from complex data.

Chapter 8: Fundamental of Designing New Age Generative Vision Transformer - By the end of this chapter, readers will have a solid understanding of transformers, their underlying principles, and their various applications in natural language processing and computer vision. They will also have the necessary knowledge to build, train, and fine-tune transformer models for their own use cases. The readers will gain a comprehensive introduction to transformers as a class of neural networks. This includes explaining their significance in revolutionizing natural language processing and their current applications in computer vision. Then, we will explore fundamental transformer concepts, delve into the basic principles and key components of transformers, such as self-attention mechanisms and the transformer architecture. This chapter will cover generative transformers and highlight the main differences between regular transformers and those designed for generative tasks. Apart from this, the reader will also be able to analyze different types of attention, such as self-attention, cross-attention, and multi-headed attention, and elucidate their specific applications in image processing.

Lastly, we will explore transformer math and positional encoding.

Chapter 9: Implementing Generative Vision Transformer - In this chapter, our primary objective is to explore and understand the fundamental distinctions between Generative Transformers and conventional Transformers, highlighting their key differences and applications within the realm of image generation. We will then delve into VAE models and their application to the STL dataset, emphasizing their capability to capture latent features and generate images. Building upon this foundation, our objective further extends to the conversion of a VAE model into a Generative Transformer model, showcasing the integration of these two powerful architectures to enhance image synthesis.

Throughout the chapter, we will compare Generative Transformers and Transformers. We will thoroughly dissect the distinctions between Generative Transformers and traditional Transformers in terms of architecture, training methodologies, and their respective strengths and weaknesses. We'll construct VAEs for the STL dataset, then transition to Generative Transformer models, adapting VAE components to fit Transformer's self-attention and positional encodings. Our comprehensive evaluation will compare image quality, diversity, and speed against traditional models. We'll also explore real-world applications, demonstrating the model's capability to produce diverse, contextually coherent images. Ultimately, this chapter aims to deepen understanding of Generative

Transformers versus traditional models, guide in VAE construction, and reveal the innovative transition to Generative Transformer architecture.

Chapter 10: Architectural Refactoring for Generative Modeling - In this chapter, our primary objective is to explore the combination process, and delve into the process of synergistically combining an encoder-decoder architecture with a transformer model for enhanced generative modeling in computer vision. We will investigate how to enhance the transformer model by introducing modifications and optimizations, contributing to improved performance and suitability for specific tasks, and provide an in-depth exploration of the SWIN transformer implementation, including detailing its architecture, components, and distinctions from other transformer variants.

Moreover, this chapter will introduce readers to advanced concepts encompassing combining hyper parameter tuning and model refactoring and aims to equip readers with a comprehensive understanding of the entire process, encompassing motivations for combining architectures, technical implementation details, and an appreciation of the intricacies of the SWIN transformer model.

Through this holistic approach, readers will gain both theoretical insights and practical skills, setting the stage for innovative generative modeling using combined encoder-decoder-transformer architectures.

Chapter 11: Major Technical Roadblocks in Generative AI and Way Forward - The designated sections of this chapter aim to unravel the challenges and innovative solutions in the fields of data representation, retrieval, and cross-modal understanding. Obstacles and technical hurdles delve into the multifaceted challenges faced in various domains, such as generative AI and computer vision.

Text and image embeddings provide insights into the pivotal role of embeddings in transforming textual and visual data into condensed, meaningful vectors. It examines how embeddings facilitate the understanding of semantic relationships and contextual nuances within language and images. The objective is to showcase how embeddings bridge the gap between raw data and AI models, contributing to better comprehension, representation, and manipulation of diverse data types.

Vector databases delves into the construction and application of databases where items are represented as vectors. The section emphasizes efficient retrieval through indexing, particularly similarity searches. It aims to elucidate the construction of structures that enable quick and accurate querying of semantically related items, illustrating their significance in real-world applications.

Image-to-image search utilizing the liberated pinecone vector databases explores the practical implementation of vector databases for image search tasks. It sheds light on the liberation of these databases for open exploration and outlines how they power efficient image retrieval mechanisms. This section aims to demonstrate how vector databases can revolutionize image search, transforming the way users discover visually similar content across a spectrum of applications.

Chapter 12: Overview and Application of Generative AI Models - In this chapter, we embark on a journey through the dynamic landscape of technology's role in various industries, without delving into complex code or algorithms. Imagine a world where cutting-edge innovations like LLM and Gen AI are not just buzzwords but integral tools reshaping healthcare, retail, finance, and insurance.

The story begins in healthcare, where LLM streamlines compliance, analyzes intricate medical documents, and guides professionals through complex regulatory mazes. Meanwhile, Gen AI steps in to provide personalized medical advice, automate appointment scheduling, and deliver vital information to patients and healthcare providers, ensuring the highest quality of care. Transitioning to the retail sector, LLM ensures contractual accuracy, compliance, and vendor agreement efficiency. Gen AI transforms the customer experience, captivating shoppers with personalized recommendations and dynamic marketing strategies, creating a retail environment tailored to each individual. In the financial realm, LLM takes center stage, enhancing risk assessment, detecting fraud, and analyzing contracts with unparalleled precision. Simultaneously, Gen AI optimizes customer service through AI-powered chatbots and virtual assistants, providing real-time and context-aware responses to financial inquiries.

Finally, in the insurance sector, LLM drives claims efficiency, fraud detection, and regulatory compliance. Gen AI revolutionizes insurance by reshaping underwriting processes, crafting personalized policy offerings, and elevating customer interactions.

Chapter 13: Key Learnings - The objective of this chapter is to synthesize and distill the core teachings and insights from chapters one through twelve. It aims to provide readers with a comprehensive summary, highlighting the key concepts, important takeaways, and significant learnings obtained from each preceding chapter. By consolidating this knowledge, the chapter seeks to offer a holistic understanding of the subject matter, reinforcing key ideas, and preparing readers for further exploration or application of the discussed principles. Ultimately, the objective is to enhance comprehension, retention, and practical application of the cumulative wisdom acquired throughout the previous chapters.

Code Bundle and Coloured Images

Please follow the link to download the *Code Bundle* and the *Coloured Images* of the book:

<https://rebrand.ly/7cicq52>

The code bundle for the book is also hosted on GitHub at

<https://github.com/bpbpublications/Learn-Python-Generative-AI>.

In case there's an update to the code, it will be updated on the existing GitHub repository.

We have code bundles from our rich catalogue of books and videos available at **<https://github.com/bpbpublications>**. Check them out!

Errata

We take immense pride in our work at BPB Publications and follow best practices to ensure the accuracy of our content to provide with an indulging reading experience to our subscribers. Our readers are our mirrors, and we use their inputs to reflect and improve upon human errors, if any, that may have occurred during the publishing processes involved. To let us maintain the quality and help us reach out to any readers who might be having difficulties due to any unforeseen errors, please write to us at :

errata@bpbonline.com

Your support, suggestions and feedbacks are highly appreciated by the BPB Publications' Family.

Did you know that BPB offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at www.bpbonline.com and as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at :

business@bpbonline.com for more details.

At **www.bpbonline.com**, you can also read a collection of free technical articles, sign up for a range of free newsletters, and receive exclusive discounts and offers on BPB books and eBooks.

Piracy

If you come across any illegal copies of our works in any form on the internet, we would be grateful if you would provide us with the location address or website name. Please contact us at **business@bpbonline.com** with a link to the material.

If you are interested in becoming an author

If there is a topic that you have expertise in, and you are interested in either writing or contributing to a book, please visit **www.bpbonline.com**. We have worked with thousands of developers and tech professionals, just like you, to help them share their insights with the global tech community. You can make a general application, apply for a specific hot topic that we are recruiting an author for, or submit your own idea.

Reviews

Please leave a review. Once you have read and used this book, why not leave a review on the site that you purchased it from? Potential readers can then see and use your unbiased opinion to make purchase decisions. We at BPB can understand what you think about our products, and our authors can see your feedback on their book. Thank you!

For more information about BPB, please visit **www.bpbonline.com**.

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

<https://discord.bpbonline.com>



Table of Contents

1. Introducing Generative AI	1
Introduction	1
Structure	2
Objectives	2
Overview of generative models.....	2
Discriminative vs. generative models.....	4
Types of discriminative and generative models.....	6
Strengths and weaknesses	10
<i>Class imbalance scenario</i>	15
<i>Generative modeling framework</i>	16
<i>Sample Space</i>	18
<i>Probability density function</i>	19
<i>Maximum likelihood</i>	26
<i>KL divergence</i>	26
<i>GMM code using TensorFlow probability</i>	31
Conclusion	33
2. Designing Generative Adversarial Networks	35
Introduction	35
Structure	36
Objectives	36
Generative Adversarial Networks	36
Types of GANs available.....	37
Architecture of a GAN.....	38
<i>Equation</i>	38
<i>Discriminator loss</i>	39
<i>Generator loss</i>	40
Vanilla GAN.....	40

<i>Outline crucial factors in GAN architecture design</i>	41
<i>Major challenges in designing GANs architecture</i>	41
Architecture of Deep Convolutional GANs.....	42
Architecture of Wasserstein GANs	43
Architecture of Conditional GANs.....	44
Architecture of CycleGANs.....	45
Architecture of progressive GANs	46
Architecture of StyleGANs.....	47
Architecture of Pix2Pix.....	48
Conclusion	49
Multiple choice questions.....	49
<i>Answers</i>	51
3. Training and Developing Generative Adversarial Networks	53
Introduction	53
Structure	54
Objectives	54
Generative Adversarial Training	54
Generating MNIST data: Basic GAN implementation.....	55
Issues during training a GANs	62
<i>Mode collapse</i>	62
<i>Vanishing gradients</i>	64
<i>Oscillation</i>	66
<i>Unstability</i>	68
<i>Evaluation</i>	69
Case study: Common practical implementation of GANs for augmentation and balancing classes	70
Conclusion	74
4. Architecting Auto Encoder for Generative AI.....	77
Introduction	77
Structure	78
Objectives	78

Auto Encoders	78
<i>Regularization</i>	81
<i>Creating a bottleneck</i>	81
Key distinctions with autoencoders	82
<i>Autoencoders</i>	82
<i>GANs</i>	82
Importance of regularization in auto encoders	83
Cifar10	86
Anomaly detection using auto encoder	91
Autoencoders with convolutional layers	97
<i>Architecture</i>	97
<i>Capturing spatial information</i>	97
<i>CNN versus ANN Autoencoders</i>	98
Conclusion	99
5. Building and Training Generative Autoencoders.....	101
Introduction	101
Structure	102
Objectives	102
Latent space	102
Difference between GANs latent space and AE latent space	107
Key distinctions with autoencoders latent space	108
Adding color to a grayscale image using autoencoders	111
Coding advanced auto encoders	115
<i>Multi modal auto encoders</i>	115
Loss in autoencoders	122
<i>Mean squared error loss</i>	122
<i>Binary cross-entropy loss</i>	123
<i>Categorical cross-entropy loss</i>	123
<i>Kullback-leibler divergence loss</i>	123
<i>Huber loss</i>	124
Challenges in training auto encoders and mitigation	124

AE vs. VAE	126
Conclusion	127
6. Designing Generative Variation Auto Encoder.....	129
Introduction	129
Structure	130
Objectives	130
Story of VAE.....	131
VAE vs AE	132
<i>Math behind the latent space.....</i>	<i>133</i>
<i>Deterministic Autoencoder</i>	<i>134</i>
<i>Stochastic Variational Autoencoder.....</i>	<i>135</i>
Key distinctions with autoencoder latent space	137
<i>Can the VAE Latent space be stochastic as well as deterministic.....</i>	<i>137</i>
<i>Dirichlet distribution</i>	<i>138</i>
Importance of the latent space when designing a VAE.....	140
Vanilla VAE architecture	142
<i>The ELBO.....</i>	<i>143</i>
<i>The reparameterization trick</i>	<i>144</i>
Challenges in Vanilla VAE	149
Types of VAE.....	150
Conclusion	152
7. Building Variational Autoencoders for Generative AI.....	155
Introduction	155
Structure	156
Objectives	157
Key focus areas in VAE research.....	157
Building a VAE with Dirichlet distribution: Non-CNN Approach	158
Building a VAE with Dirichlet distribution: CNN Approach.....	162
<i>Difference between two networks</i>	<i>166</i>
VAE with Non Dirichlet distribution	171
KL divergence.....	175

Common loss function sin VAE	177
Common issues and possible solutions while training VAE	178
Missing data handling during generation.....	180
Optimization techniques.....	181
Conclusion	182
8. Fundamental of Designing New Age Generative Vision Transformer	183
Introduction	183
Structure	183
Objectives	184
The evolution.....	184
<i>The birth of transformers.....</i>	<i>186</i>
<i>Overview of transformer architectures.....</i>	<i>186</i>
<i>Applications in NLP</i>	<i>188</i>
<i>Generative transformers and language modeling</i>	<i>189</i>
<i>Transformer in computer vision.....</i>	<i>189</i>
Difference between VAE, GANs, and Transformers.....	190
<i>Transformers.....</i>	<i>190</i>
<i>Generative Adversarial Networks</i>	<i>191</i>
<i>Variational autoencoders.....</i>	<i>191</i>
<i>Differences and applications.....</i>	<i>192</i>
Vision Transformer.....	193
Understanding self-attention	194
NLP vs vision.....	195
<i>NLP transformer</i>	<i>196</i>
<i>Self-attention mechanism.....</i>	<i>196</i>
<i>Feed-forward neural networks</i>	<i>197</i>
<i>Vision transformer.....</i>	<i>197</i>
<i>Patch embeddings.....</i>	<i>197</i>
<i>Positional embeddings</i>	<i>198</i>
<i>Transformer encoder.....</i>	<i>199</i>
Architectural attention	199

<i>Dot product attention</i>	199
<i>Scaled dot product attention</i>	199
<i>Additive attention</i>	200
<i>Multi-head attention</i>	200
<i>Cross attention</i>	200
<i>Compute attention scores</i>	201
<i>Compute cross-attention output</i>	201
When to use which architectural attention	201
Functional attention.....	203
<i>Hard attention</i>	203
<i>Equation: Sampling-based hard attention</i>	203
<i>Soft attention</i>	203
<i>Equation: Soft attention</i>	204
<i>Global attention</i>	204
<i>Equation: Global attention</i>	204
<i>Local attention</i>	204
<i>Equation: local attention</i>	204
When to use which functional attention.....	205
<i>Hard attention</i>	205
<i>Soft attention</i>	206
<i>Global attention</i>	206
<i>Local attention</i>	206
Conclusion	207
9. Implementing Generative Vision Transformer.....	209
Introduction	209
Structure	210
Objectives	210
STL dataset.....	211
<i>Key features of the STL-10 dataset</i>	211
Developing a VAE model on STL dataset.....	212
Implementation of VAE architecture with TensorFlow	213

<i>Outputs</i>	216
Pytorch.....	217
Transition from VAE to Generative Transformer Model: Keras Vit Library	223
Implementing a ViT model from scratch.....	225
<i>Outputs</i>	227
Implementing a ViT model pre trained with ViT model.....	229
<i>Outputs</i>	233
Training Pretrained ViT vs ViT scratch	234
<i>Pretrained Vision Transformer</i>	234
<i>Advantages</i>	234
<i>Disadvantages</i>	234
<i>Training a ViT model from scratch</i>	234
<i>Advantages</i>	235
<i>Disadvantages</i>	235
Examining the loss curve	235
Optimization of ViT models	236
Conclusion	237
10. Architectural Refactoring for Generative Modeling	239
Introduction	239
Structure	240
Objectives	240
STL dataset.....	240
Exploring the combination process: Outline.....	241
Refactoring TransVAE and improving	242
<i>Cyclic Learning Rate Schedule</i>	242
<i>LearningRateScheduler</i>	245
<i>EarlyStopping</i>	245
<i>Weight decay: L2 regularization</i>	245
Improved Encoder Decoder	251
SWIN-Transformer.....	252
Implementation of SWIN Transformer: VAE.....	254

Improving the models	260
Conclusion	263
11. Major Technical Roadblocks in Generative AI and Way Forward	265
Introduction	265
Structure	266
Objectives	266
Challenges and hurdles in Generative AI	267
<i>NLP based generative models</i>	268
Large language models and image-based foundation models	270
Embedding in language models	272
Embedding in image.....	274
Generative AI and embeddings	275
Vector data bases and image embeddings.....	276
<i>Vector databases</i>	277
<i>Image embeddings</i>	277
Building an image search using pinecone and vector database.....	278
Conclusion	288
12. Overview and Application of Generative AI Models.....	289
Introduction	289
Structure	290
Objectives	290
GenAI in hospital	291
GenAI in dental	292
GenAI in radiology	293
GenAI in retail	295
GenAI in finance.....	296
GenAI in corporate finance.....	298
GenAI in insurance	299
Conclusion	301

13. Key Learnings	303
Introduction	303
Structure	303
Objectives	303
Key learning from all the chapters	304
<i>Chapter 1: Introducing Generative AI</i>	<i>304</i>
<i>Chapter 2: Designing Generative Adversarial Networks</i>	<i>305</i>
<i>Chapter 3: Training and Developing Generative Adversarial Networks.....</i>	<i>305</i>
<i>Chapter 4: Architecting Auto Encoder for Generative AI</i>	<i>306</i>
<i>Chapter 5: Building and Training Generative Autoencoders</i>	<i>307</i>
<i>Chapter 6: Designing Generative VAE</i>	<i>308</i>
<i>Chapter 7: Building Variational AutoEncoders for Generative AI.....</i>	<i>308</i>
<i>Chapter 8: Designing New Age Generative Vision Transformer for</i> <i>Generative Learning</i>	<i>309</i>
<i>Chapter 9: Implementing Generative Vision Transformers</i>	<i>310</i>
<i>Chapter 10: Architectural Refactoring Combining Encoder-decoder and</i> <i>Transformers for Generative Modeling.....</i>	<i>311</i>
<i>Chapter 11: Major Technical Roadblocks in Generative AI</i>	<i>312</i>
<i>Chapter 12: Overview of Applications of Generative AI Models.....</i>	<i>313</i>
Index	315-324

CHAPTER 1

Introducing Generative AI

Introduction

In this chapter, you will learn about the evolution of generative AI and how it has progressed over the years. It also highlights the approaches previously used for generative models, and how these have changed with the emergence of deep learning and vast amounts of data. Some of the latest techniques, such as **Generative Adversarial Networks (GANs)** and **Variational Autoencoders (VAEs)**, and their applications in generating high-quality images, audio, and text are also discussed.

In addition, you can learn about the difference between discriminative and generative models and how generative models aim to generate new data that follows the original data distribution. An introduction to generative models and an overview of the various generative models available are also provided.

Finally, the chapter discusses the strengths and weaknesses of generative models and highlights that there is still much room for further innovation and improvement in generative AI. Overall, the chapter provides an excellent introduction to the evolution of generative AI and the different techniques used in the field.

Structure

In this chapter, we will learn about the following topics:

- Overview of generative models
- Discriminative vs generative modes
- Types of discriminative and generative models
- Strengths and weaknesses

Objectives

The objective of this chapter is to provide a comprehensive understanding of generative models, including an overview of generative models, a comparison of discriminative vs generative models, an introduction to the types of discriminative and generative models, as well as their strengths and weaknesses. By the end of the content, readers should be able to differentiate between discriminative and generative models, understand the different types of each, and make informed decisions about which type of model is most appropriate for their needs.

Overview of generative models

Generative AI refers to a type of artificial intelligence that can generate new data or content, such as images, videos, or text, with similar characteristics to the training data it was given. Generative AI has progressed rapidly over the years, and much of this progress has been driven by advances in deep learning.

One of the earliest examples of generative AI was the autoencoder, developed in the 1980s. Autoencoders are neural networks that can learn to compress and reconstruct data, and they can also be used to generate new data by sampling from the known compressed representation. However, autoencoders have limitations regarding the types of data they can develop and the quality of the generated output.

In the 1990s, Boltzmann machines were developed, which are neural networks that can model the joint probability distribution of a set of input variables. Boltzmann Machines can be used for generative modeling by sampling from the learned distribution, but they are challenging to train and scale to large datasets.

More recently, deep learning has enabled significant progress in generative AI, particularly with the development of GANs and VAEs. GANs were first introduced in 2014. They consist of two neural networks: a generator network that generates new data and a discriminator network that distinguishes between generated and real data. The generator is trained to produce indistinguishable data from real data, while the discriminator is trained to correctly classify the data as real or fake. Through this adversarial training process, GANs

can generate high-quality data in a variety of domains, including images, videos, and music.

VAEs were also introduced in 2014, they are similar to autoencoders, with the addition of a probabilistic encoder that learns a distribution over the compressed representation. VAEs can generate new data by sampling from the learned distribution, and they have been used for generative modeling in various domains, including images and text.

More recent advancements in generative AI have focused on improving the quality and diversity of generated data, such as using attention mechanisms and self-attention, as well as exploring new domains for generative modeling, such as 3D object generation and interactive storytelling. Generative AI has also made significant progress in **natural language processing (NLP)**, where language models such as OpenAI's **Generative Pre-trained Transformer (GPT)** series have achieved remarkable performance in tasks such as language generation, language understanding, and even question answering. These models use a generative approach to learn the underlying structure and patterns of human language, allowing them to generate coherent and fluent sentences almost indistinguishable from those written by humans.

Moreover, generative AI has also been used in creative domains such as art, music, and fashion, enabling new forms of artistic expression and creativity. For example, DeepDream, a generative model developed by Google, has been used to create surreal and psychedelic images by altering the features of an input image. Similarly, the Magenta project by Google has developed generative models for music creation that can generate original compositions in various styles and genres. In recent years, there have been several new generative models that have emerged, which have shown impressive results in generating realistic and diverse outputs. Two such models are the **Stable Diffusion (SDE)** and DALL-E.

Stable Diffusion (SDE) is a recently proposed generative model that builds upon the idea of continuous-time stochastic processes. The model is based on the diffusion process, a stochastic process that describes the movement of particles in a fluid or gas. The SDE model uses a **Markov Chain Monte Carlo (MCMC)** approach to learn a stochastic differential equation that describes the dynamics of the diffusion process. This allows the model to understand the complex correlations between the inputs and generate high-quality samples that exhibit a high degree of diversity.

DALL-E is another recently proposed generative model that OpenAI developed. DALL-E is a transformer-based model that can generate high-quality images from textual descriptions. The model uses a conditioning mechanism that takes in a textual description as input and generates an image that corresponds to the description. DALL-E is trained on a massive dataset of text-image pairs, which allows it to learn the complex relationships between text and images.

One of the advantages of these new generative models is that they are capable of generating high-quality and diverse outputs that are difficult to distinguish from real data. This has

important implications for a range of applications, such as image and video synthesis, text-to-image generation, and natural language processing.

Discriminative vs. generative models

Discriminative modeling involves directly learning the decision boundary between classes, which allows for the direct classification of new examples. For example, in a binary classification problem, a discriminative model learns to predict whether an input belongs to class A or class B. Discriminative models do not attempt to model the underlying distribution of the data, but rather focus on learning the boundary between classes. Common discriminative models include logistic regression, support vector machines, and neural networks.

On the other hand, generative modeling involves modeling the underlying distribution of the data and using that model to generate new examples like the training data. Generative models can also be used for classification by computing the probability of a new example belonging to each class and choosing the class with the highest probability. Common generative models include Naive Bayes, Gaussian mixture models, and Hidden Markov models.

One advantage of generative models is that they can generate new data points, which can be useful in scenarios where the amount of training data is limited. However, generative models may be more computationally expensive than discriminative models, as they require modeling the entire data distribution. In addition, generative models may not perform as well as discriminative models in situations where the decision boundary between classes is complex.

Both discriminative and generative modeling is important in today's deep learning era.

Discriminative models, such as convolutional neural networks and recurrent neural networks, are commonly used in tasks such as image classification, object detection, and natural language processing. These models are highly effective at learning complex decision boundaries between classes and can achieve state-of-the-art performance on many tasks.

Generative models, such as variational autoencoders and generative adversarial networks, have become increasingly popular recently. These models can generate new data points that are similar to the training data, which can be useful in scenarios where the amount of training data is limited. Generative models are also being used in applications such as image and video synthesis, text generation, and data augmentation.

Overall, both discriminative and generative models have essential roles in the deep learning era, and the choice between them depends on the specific task and available resources. As deep learning continues to advance, it is likely that both types of models will continue to play essential roles in different applications. In the following figure, we can clearly see the difference between discriminative and generative models and how they illustrate the

decision boundary. Understanding these concepts is crucial for anyone looking to work with machine learning models:

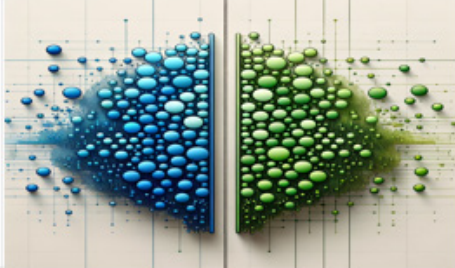
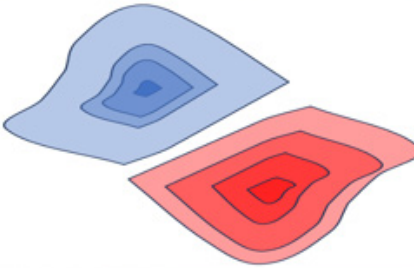
	Discriminative model	Generative model
Goal	Directly estimate $P(y x)$	Estimate $P(x y)$ to then deduce $P(y x)$
What's learned	Decision boundary	Probability distributions of the data
Illustration		
Examples	Regressions, SVMs	GDA, Naive Bayes

Figure 1.1: Difference between discriminative and generative models

Discriminative models learn the boundary between classes directly, while generative models learn the joint probability distribution of the input and output variables. Let's delve deeper into these topics and explore the various types of discriminative and generative models, as well as their strengths and weaknesses. Let us understand the significance of these modeling in today's Deep learning era. Discriminative models, such as convolutional neural networks and recurrent neural networks, are commonly used in tasks such as image classification, object detection, and natural language processing. These models are highly effective at learning complex decision boundaries between classes and can achieve state-of-the-art performance on many tasks.

Generative models, such as variational autoencoders and generative adversarial networks, have become increasingly popular recently. These models can generate new data points similar to the training data, which can be useful in scenarios where the amount of training data is limited. Generative models are also being used in applications such as image and video synthesis, text generation, and data augmentation.

Overall, both discriminative and generative models have essential roles in the deep learning era, and the choice between them depends on the specific task and available resources. As deep learning continues to advance, it is likely that both types of models will continue to play essential roles in many different applications.

Let us clarify a common misconception about **convolutional neural networks (CNNs)** and **recurrent neural networks (RNNs)**. The question many ask, are they generative models? CNNs are primarily used for discriminative modeling tasks such as image classification and object detection. They learn to extract features from the input data and use those features to make predictions about the class of the input.