The Transformer Revolution Generative AI and Large Language Models Farhat Habib, PhD

Mphasis.ai

Bits & Scripts: Workshop on Computational Epigraphy, March 26 2024

Overview

- Generative models
- History
- Autoencoders
- Variational autoencoders
- Diffusion models
- Transformers
- LLMs



What are generative models?

- Aim to generate new data points similar to training set.
- Learns the distribution of data
- Can generate unseen instances by learning from dataset.
- Used in image, text, and audio synthesis.





Turn a linear model into a generative model

- Assume a distribution (e.g., Gaussian) for features in each class.
- Estimate parameters (mean, variance) of these distributions using linear model outputs.
- Data Generation:
 - Sample feature values from estimated distributions.
 - Generate new data points by sampling from these feature distributions



Timeline of generative AI models



https://www.desdevpro.com/blog/talk-rise-of-generative-ai

Autoencoders What is an autoencoder?

- A neural network trained using unsupervised learning
 - Trained to copy its input to its output
 - Learns an embedding h







What is an embedding?

- An embedding is a low-dimensional vector (e.g., PCA)
 - With fewer dimensions than the ambient space of which the manifold is a lowdimensional subset
- Embedding Algorithm
 - Maps any point in ambient space x to its embedding h

What does an autoencoder learn?

- Autoencoders are designed to be unable to copy perfectly
- Autoencoders learn salient features of the data
- Forced to prioritize which aspects of input should be copied
- When the decoder is linear and loss function is the mean squared error, an autoencoder learns to span the same subspace as PCA
- Autoencoders with nonlinear activation functions can learn more powerful nonlinear generalizations of PCA

X'1 Bottleneck X'2 х'з X'4 X'5 Hidden Output Input layer layer layer



Variational Autoencoders

- Training Objective:
 - Autoencoder: Minimize reconstruction error.
 - VAE: Minimize reconstruction error + Regularize latent space (KL divergence).
 - Kullback-Leibler divergence quantifies the difference between two probability distributions
 - Encodes inputs into a distribution over the latent space, characterized by mean and variance parameters
 - Can generate new, unseen data by sampling from latent space.



VAE generated images



Xianxu Hou, Linlin Shen, Ke Sun, Guoping Qiu, Deep Feature Consistent Variational Autoencoder, Neurocomputing

Generative Adversarial Networks

Generative Adversarial Networks

- GANs work by having two neural networks compete against each other: a generator network that creates the fake data, and a discriminator network that tries to distinguish the fake data from real data
- The generator's goal is to produce data that is indistinguishable from real data, while the discriminator's goal is to correctly identify the real and fake data.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. In Advances in neural information processing systems (pp. 2672–2680).





Generative Model

- How to make it generate different samples each time it is run?
 - input to model is noise
- Generative model as a neural network
 - computes $x = G(z|\theta)$
 - z typically has very high dimensionality (higher than x)



GANtraining

- Discriminator Training:
 - dataset and fake data generated by the generator.
- Generator Training:
 - weights.

• The discriminator is trained first within a training cycle. It receives both real data from the

• generator produces fake data, which is then passed to the discriminator. When training the generator, the goal is to maximize the mistake rate of the discriminator—essentially, the generator is rewarded if the discriminator classifies fake data as real. The generator's loss is calculated based on the discriminator's predictions on the fake data, with the aim of making these predictions incorrect. The gradient of this loss is then used to update the generator's



GANtraining

- limited varieties of data.
- Non-convergence: GANs can oscillate during training, making it difficult to achieve convergence.
- batch sizes, and network architectures.
- distributions.

Mode collapse: GANs can suffer from mode collapse, where the generator learns to produce

• Hyperparameter tuning: GANs require careful tuning of hyperparameters like learning rates,

• Data quality: High-quality and diverse training data is crucial for GANs to learn realistic data



GAN output



Progressive Growing of GANs for Improved Quality, Stability, and Variation, Progressive GAN, by NVIDIA, and Aalto University, 2018 ICLR

GAN (mode collapse)













Transformers

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2023). Attention Is All You Need. arXiv [Cs.CL]. Retrieved from http://arxiv.org/abs/1706.03762

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com Noam Shazeer* Google Brain noam@google.com Niki Parmar* Google Research nikip@google.com Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com Aidan N. Gomez^{* †} University of Toronto aidan@cs.toronto.edu Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin*[‡] illia.polosukhin@gmail.com



Attention

- "Attention" in neural networks is a mechanism that allows the model to focus on certain parts of the input data more than others,
- way for the model to allocate its 'attention' or focus to specific elements in the data it's processing, which helps it make better predictions or generate more coherent outputs.









Self supervised learning

- This formulation has been used in the BERT, RoBERTa and ALBERT papers.





• In this formulation, words in a text are randomly masked and the task is to predict them.

A quick [MASK] fox jumps over the [MASK] dog A quick brown fox jumps over the lazy dog

Training Compute optimal models

- training larger models with less data. This contrasts with earlier beliefs emphasizing increasing model size.
- Optimal Training Regime: Chinchilla findings propose an optimal training regime that balances the scale of the dataset and the size of the model, demonstrating that the factors in harmony, rather than focusing solely on model size.

• Data Efficiency Over Model Size: Chinchilla research indicates that, for a given compute budget, training models with more data and slightly fewer parameters is more effective than

efficiency of language model training can be significantly improved by adjusting these



Open LLM Leaderboard

T A	Model	Average 🚹 🔺
•	davidkim205/Rhea-72b-v0.5 🖹	81.22
•	<u>SF-Foundation/Ein-72B-v0.11</u>	80.81
•	<u>SF-Foundation/Ein-72B-v0.13</u>	80.79
•	SF-Foundation/Ein-72B-v0.12	80.72
•	<u>abacusai/Smaug-72B-v0.1</u>	80.48
•	ibivibiv/alpaca-dragon-72b-v1 🐚	79.3
	moreh/MoMo-72B-lora-1.8.7-DPO 📑	78.55
•	cloudyu/TomGrc FusionNet 34Bx2 MoE v0.1 DP0 f16 📑	77.91
•	saltlux/luxia-21.4b-alignment-v1.0 🖹	77.74
•	cloudyu/TomGrc FusionNet 34Bx2 MoE v0.1 full linear DPO 📄	77.52

Diffusion models

- Diffusion models are a class of generative models that learn data distributions by iteratively corrupting data and reverse the process.
- Work by gradually adding Gaussian noise to data, and then training a neural network to reverse the diffusion process and reconstruct the original data.
- Capable of generating high-quality samples across various data types and domains, including images, audio, and video

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. arXiv [Cs.LG].



Training diffusion models

- Forward diffusion process: Training starts by corrupting the data with Gaussian noise over multiple diffusion steps, gradually destroying the data into pure noise.
- Denoising objective: At each diffusion step, the model is trained to predict the noise that was added, effectively learning to denoise the data.
- Optimize parameters: The model parameters are optimized to minimize the difference between the predicted noise and the actual noise added during the forward diffusion process.



Inferencing from Diffusion Models

- Reverse diffusion process: During inference, the model starts from pure noise and iteratively denoises the data by predicting and subtracting the added noise at each step.
- Sampling from distribution: At each step, the model samples from the learned distribution to generate the denoised data for the next step.
- Conditional guidance: Diffusion models can be conditioned on additional inputs, like text
 prompts or class labels, to guide the generation process towards desired outputs





DALLE-2



vibrant portrait painting of Salvador Dalí with a robotic half face



an espresso machine that makes coffee from human souls, artstation

panda mad scientist mixing sparkling chemicals, artstation



Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv [Cs.CV]. Retrieved from http://arxiv.org/abs/2204.06125

a shiba inu wearing a beret and black turtleneck

a close up of a handpalm with leaves growing from it

100



a corgi's head depicted as an explosion of a nebula



CLIP (Contrastive Language-Image Pre-training)

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. arXiv preprint arXiv:2103.00020.

• Multimodal Learning: CLIP is trained to understand and associate images with textual descriptions, making it capable of understanding a wide range of visual concepts expressed in natural language.

• Contrastive Pre-training: It uses a contrastive learning approach to pre-train on a large dataset of images and text pairs, learning to predict the correct pairing among a set of incorrect ones.

• The contrastive pre-training approach works by teaching the model to distinguish between matching and non-matching pairs of data across two different modalities (e.g., images and text).

• The model is presented with pairs of images and text captions. For each image, there is one matching caption that describes the image, and several non-matching captions. The model's task is to predict which caption correctly matches the image among the set of possible captions.





Figure 2: A high-level overview of unCLIP. Above the dotted line, we depict the CLIP training process, through which we learn a joint representation space for text and images. Below the dotted line, we depict our text-to-image generation process: a CLIP text embedding is first fed to an autoregressive or diffusion prior to produce an image embedding, and then this embedding is used to condition a diffusion decoder which produces a final image. Note that the CLIP model is frozen during training of the prior and decoder.





Figure 17: unCLIP samples show low levels of detail for some complex scenes.

(a) A high quality photo of a dog playing in a green field next to a lake.

(b) A high quality photo of Times Square.

Vision Transformer



An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, <u>Alexey Dosovitskiy</u>, <u>Lucas</u> Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby



Future...

• Tap into video data



Thank you!

