

Deep Generative Models (AI314)

Master of Science in AI



Dr. Nibras Abdullah

*Assistant professor
Faculty of Informatics
Midocean University*



Introduction



Evaluation Metrics for Generative Models



Human Evaluation



Importance of Human Evaluation



Choosing the Right Metrics



Deep Generation Models

Deep generation models are a powerful class of neural networks used to generate new data similar to the data they were trained on. They have revolutionized fields like image and text generation.

Generative Models: A Comprehensive Overview

Generative models have emerged as a transformative force in the field of artificial intelligence, revolutionizing our ability to create new data that resembles real-world examples. From generating realistic images and music to writing compelling text and designing innovative products, generative models have proven their versatility and impact across various domains. This presentation provides a comprehensive overview of generative models, exploring their fundamental concepts, different types, evaluation metrics, and diverse applications.



Autoregressive Models

1 Sequential Prediction

Autoregressive models, such as recurrent neural networks (RNNs), generate data one element at a time, predicting the next element based on the previously generated elements. This sequential approach allows the model to capture complex dependencies and relationships within the data.

3 Limitations

Autoregressive models can be computationally expensive, especially for large datasets and complex sequences. They also struggle with generating long sequences effectively due to the accumulation of errors over time.

2 Examples

Popular autoregressive models include GPT-3 (text generation), WaveNet (audio synthesis), and PixelRNN (image generation). These models have achieved remarkable results in their respective domains, pushing the boundaries of generative modeling.

4 Applications

Autoregressive models find applications in natural language processing, music generation, speech synthesis, and image generation. Their ability to generate realistic sequences makes them invaluable tools in various creative and analytical tasks.

Variational Autoencoders (VAEs)

Encoding and Decoding

VAEs are a type of generative model that learn a latent representation of the data by compressing it into a lower-dimensional space. This compressed representation, known as the latent code, can be used to generate new data by decoding it back into the original space.

Variational Inference

VAEs use variational inference to approximate the distribution of the latent code. This involves optimizing a variational lower bound to minimize the difference between the true distribution and the approximate distribution.

Advantages

VAEs are known for their ability to generate diverse and realistic samples. They are also relatively efficient to train and can be used for various tasks, including image generation, anomaly detection, and data imputation.

Generative Adversarial Networks (GANs)

1

Generator

The generator network attempts to generate realistic data samples that resemble the real data distribution.

2

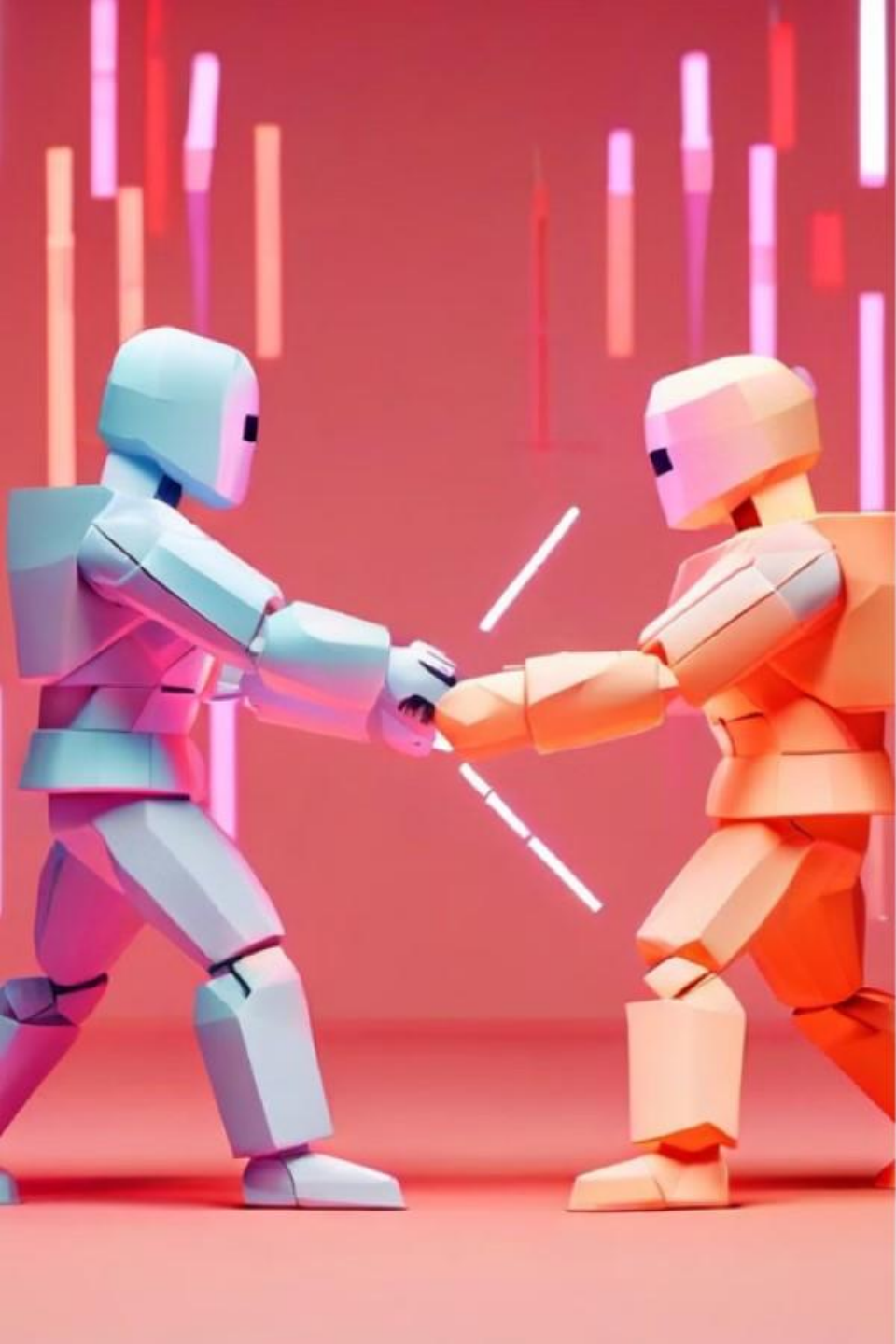
Discriminator

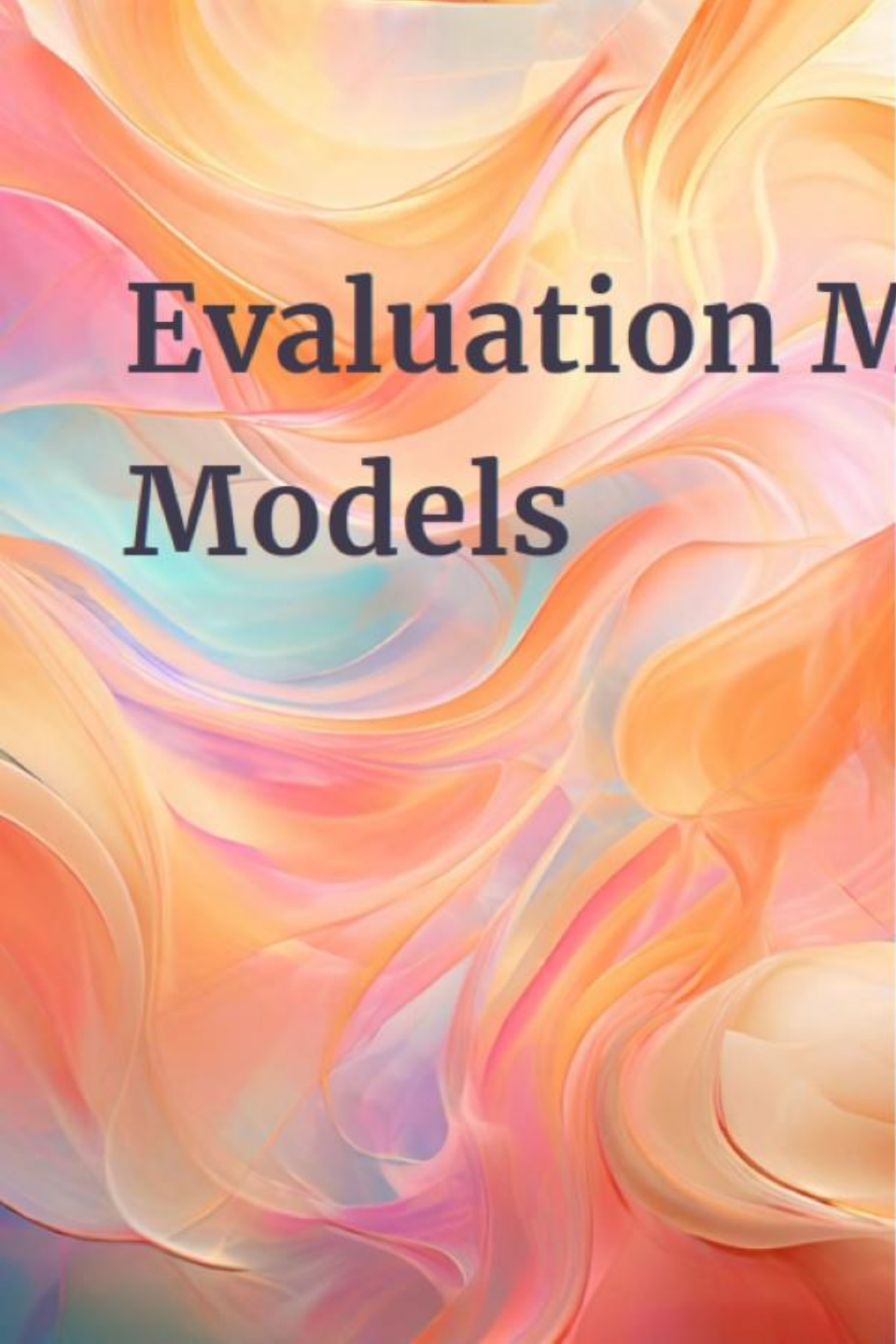
The discriminator network acts as a critic, evaluating the generated samples and determining their authenticity. It aims to distinguish between real and generated samples.

3

Adversarial Training

The generator and discriminator are trained in an adversarial manner, where the generator tries to fool the discriminator while the discriminator tries to improve its ability to detect fake samples. This constant competition leads to the generation of increasingly realistic and diverse data.





Evaluation Metrics for Generative Models

Generative models are powerful tools for creating new data that resembles real-world data. Evaluating the quality and performance of these models is crucial for understanding their capabilities and limitations. Several metrics have been developed to assess different aspects of generative models, providing insights into their ability to generate realistic and diverse outputs.

Evaluation Metrics for Generative Models

Metric	Description
Inception Score	Measures the quality and diversity of generated samples by evaluating the agreement between a pretrained classifier and the generated samples.
Fréchet Inception Distance (FID)	Calculates the distance between the feature distributions of real and generated samples using a pretrained inception network.
Likelihood	Evaluates the probability of the generated samples according to the learned model, indicating how well the model fits the data.
Human Evaluation	Involves subjective assessments by human experts, who provide feedback on the realism and quality of the generated samples.

Inception Score

1

Inception Score

Measures the quality and diversity of generated samples by evaluating the agreement between a pretrained classifier and the generated samples.

Inception Score

2

High Score

Indicates that the generated samples are both realistic and diverse, suggesting that the model has learned to capture the underlying distribution of the data.

Indicates that the generated samples are both diverse and belong to the correct classes according to the pretrained classifier.

Inception Score

3

Low Score

Suggests that the generated samples are either not realistic or not diverse, indicating that the model may not have learned the data distribution well.

Suggests that the generated samples are either not diverse or do not align well with the classifier's predictions.

Fréchet Inception Distance (FID)

FID

Calculates the distance between the feature distributions of real and generated samples using a pretrained inception network.

Fréchet Inception Distance (FID)

Low FID

Indicates that the generated samples are similar to real samples in terms of their feature distributions, suggesting that the model has learned to generate realistic data.

Indicates that the generated samples are closer to the real data distribution, suggesting higher realism.

Fréchet Inception Distance (FID)

High FID

Suggests that the generated samples are different from real samples in terms of their feature distributions, indicating that the model may not have learned the data distribution well.

Implies a larger discrepancy between the generated and real data distributions, indicating lower realism.

Likelihood

Likelihood

Evaluates the probability of the generated samples according to the learned model, indicating how well the model fits the data.

High Likelihood

Indicates that the generated samples are likely to have been generated by the model, suggesting that the model has learned the data distribution well.

Low Likelihood

Suggests that the generated samples are unlikely to have been generated by the model, indicating that the model may not have learned the data distribution well.



Human Evaluation

Metric

Description

Human Evaluation

Involves subjective assessments by human experts, who provide feedback on the realism and quality of the generated samples.



Human Evaluation

1

Subjective Assessments

Involves subjective assessments by human experts, who provide feedback on the realism and quality of the generated samples.

2

Expert Feedback

Human experts can provide valuable insights into the strengths and weaknesses of the generated samples, which can be used to improve the model.

3

Realism and Quality

Human experts can assess the realism and quality of the generated samples based on their experience and knowledge of the real world.

Subjective Assessments

1

Realism

Human experts assess the degree to which the generated samples resemble real-world data.

2

Quality

Human experts evaluate the overall quality of the generated samples, considering factors such as clarity, detail, and aesthetic appeal.

3

Diversity

Human experts assess the variety and uniqueness of the generated samples, ensuring that the model is not generating repetitive or predictable outputs.

Importance of Human Evaluation



Subjective Feedback

Human evaluation provides valuable subjective feedback that cannot be captured by objective metrics.



User Experience

Human evaluation helps to assess the user experience of interacting with the generated samples, ensuring that they are enjoyable and engaging.



Quality Control

Human evaluation acts as a quality control mechanism, ensuring that the generated samples meet specific standards and expectations.





Combining Metrics

Combining different evaluation metrics provides a more comprehensive understanding of the performance of generative models. By considering both objective and subjective assessments, researchers and developers can gain valuable insights into the strengths and weaknesses of their models, leading to improvements in their design and training.

Evaluation Metrics for Generative Models

Generative models are powerful tools for creating new data that resembles real-world data. Evaluating the quality and performance of these models is crucial to ensure they generate realistic and meaningful outputs. This presentation explores various evaluation metrics commonly used for generative models.

Metrics for Text Generation



BLEU Score

Measures the similarity between the generated text and a set of reference translations, indicating how well the model captures the meaning and style of the original text.

ROUGE Score

Evaluates the overlap between the generated text and a set of reference summaries, indicating how well the model captures the key information from the original text.

Metrics for Text Generation



Perplexity

Measures the uncertainty of the model's predictions, indicating how well the model fits the data and how likely it is to generate fluent and coherent text.

Human Evaluation

Involves subjective assessments by human experts, who provide feedback on the fluency, coherence, and quality of the generated text.

Metrics for Audio Generation

1

Perceptual Evaluation of Speech Quality (PESQ)

Measures the perceived quality of generated speech by comparing it to a reference speech signal, indicating how well the model captures the naturalness and intelligibility of human speech.

2

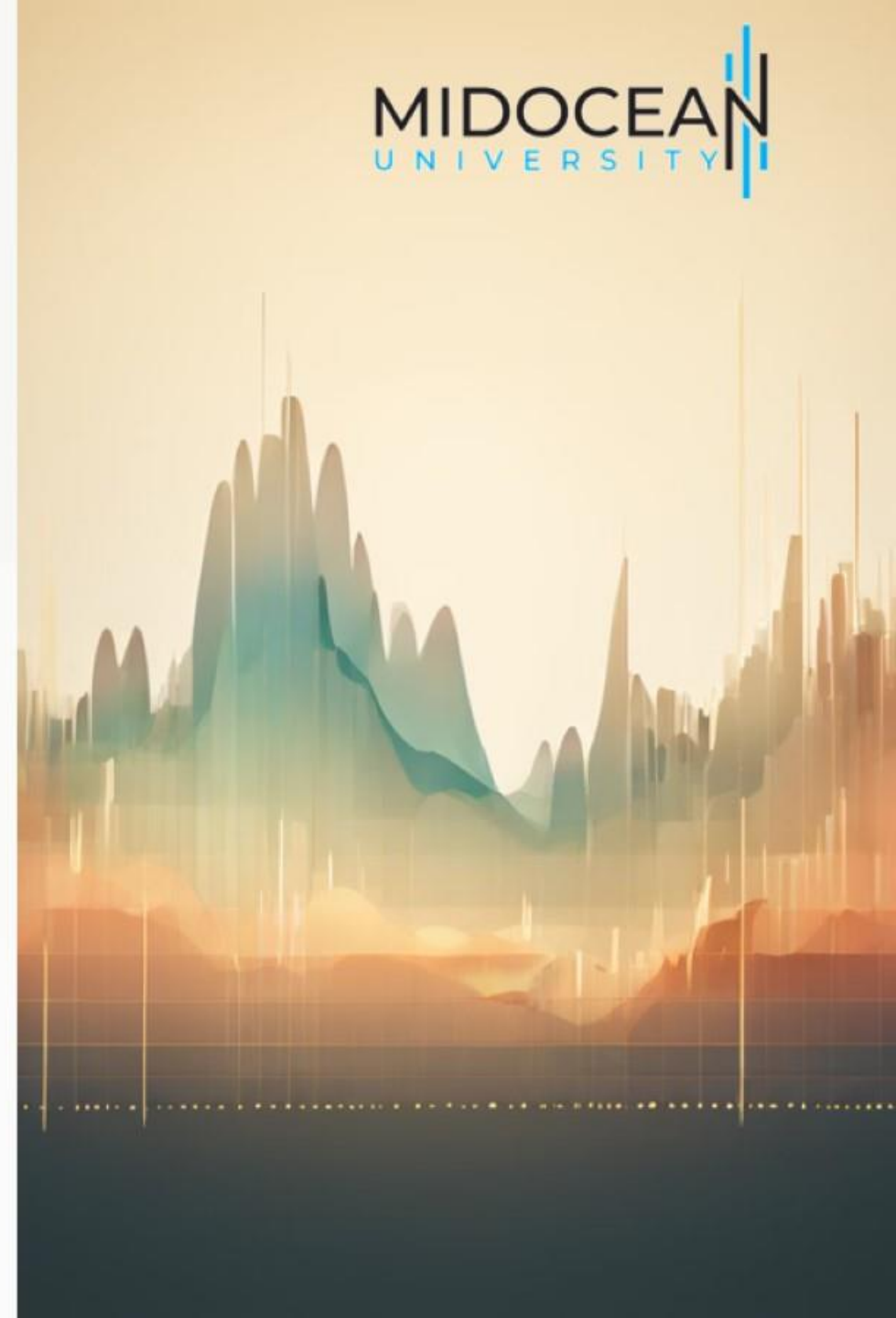
Short-Time Objective Intelligibility (STOI)

Evaluates the intelligibility of generated speech by measuring the amount of information that can be understood from the speech signal, indicating how well the model captures the meaning of the speech.

3

Human Evaluation

Involves subjective assessments by human listeners, who provide feedback on the naturalness, intelligibility, and overall quality of the generated audio.



Evaluation Metrics for Deep Generation Models

1

Perplexity

Measures how well a model predicts the next element in a sequence. Lower perplexity indicates better prediction accuracy.

2

BLEU Score

Evaluates the quality of machine translation by comparing the generated translation to reference translations. A higher BLEU score means a more fluent and accurate translation.

3

METEOR

A metric similar to BLEU but also considers word stemming, synonymy, and paraphrasing.

4

ROUGE Score

Measures the quality of text summarization by comparing the generated summary with reference summaries.



Perplexity

1

Text Generation

Predicts the next word in a sentence based on the preceding words. A lower perplexity indicates a more coherent and grammatically correct sentence.

2

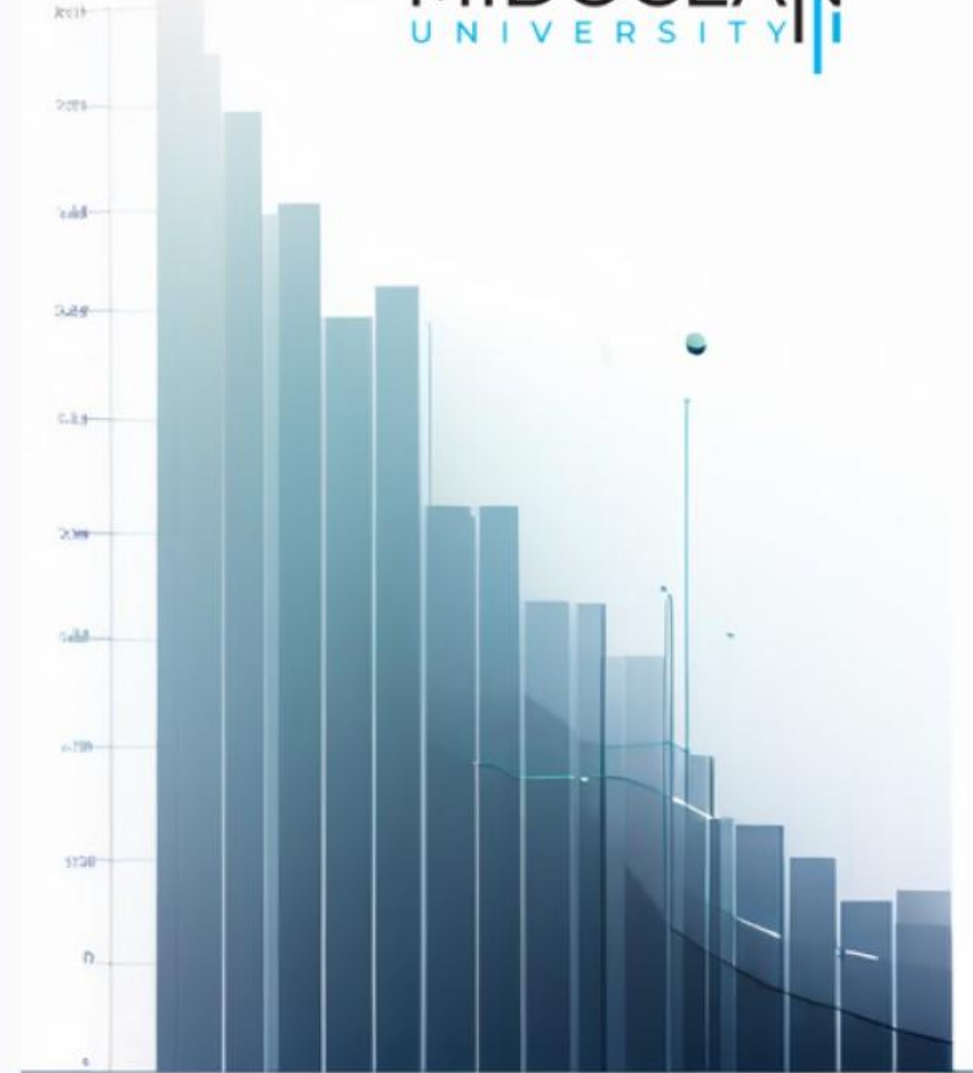
Speech Recognition

Evaluates how well a model transcribes spoken language. Lower perplexity means the model accurately recognizes the spoken words.

3

Code Generation

Measures the accuracy of a model in predicting the next line of code based on the existing code.



BLEU Score

Machine Translation

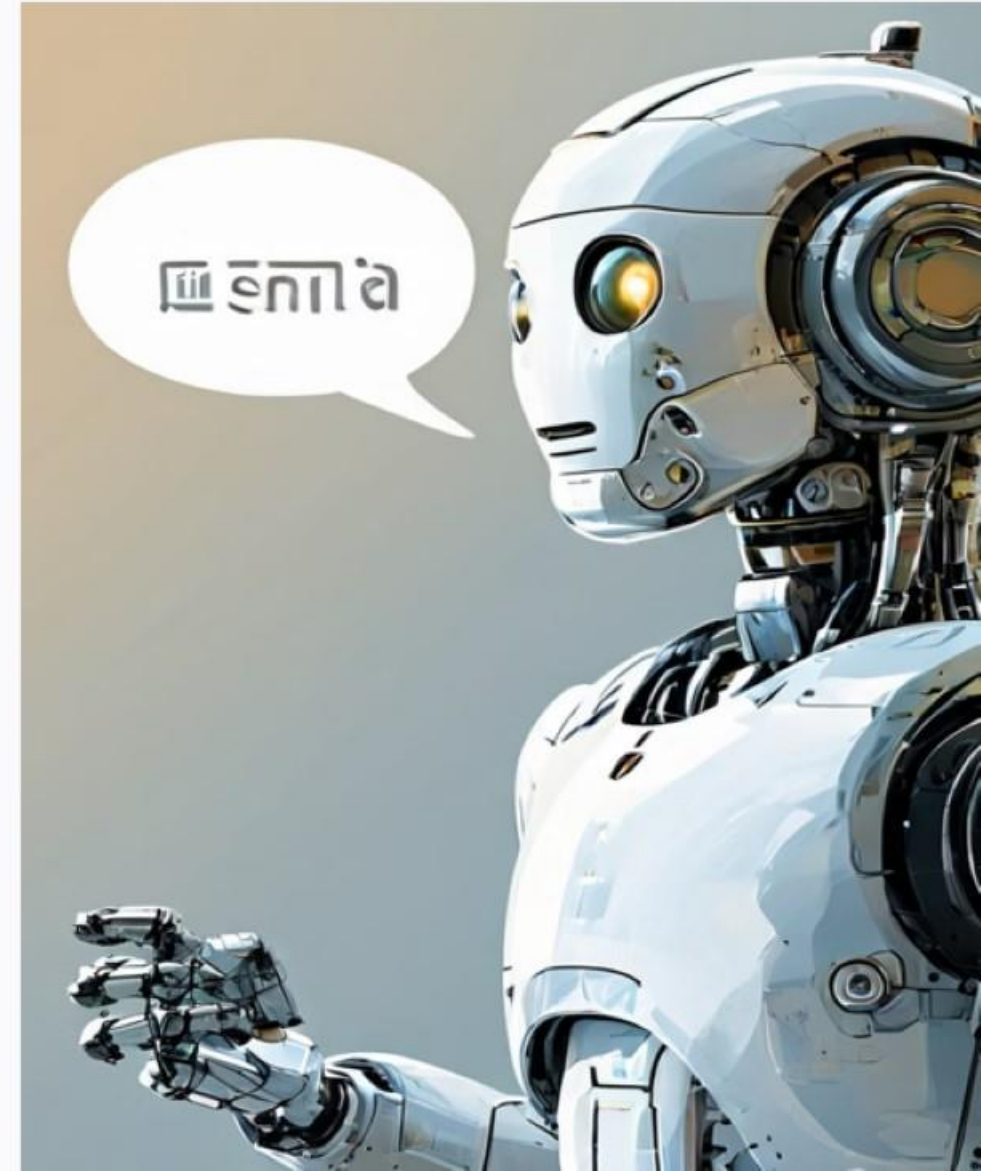
Evaluates the quality of a machine translation by comparing the generated translation to human-translated reference sentences.

Text Summarization

BLEU can be used to assess how well a model summarizes a document by comparing the generated summary to human-written reference summaries.

METEOR

Focus	BLEU Score	METEOR
Precision & Recall	High Emphasis	Balances Precision & Recall
Word Stemming & Synonymy	Not Considered	Considers Stemming & Synonymy
Paraphrasing	Not Considered	Considers Paraphrasing



ROUGE

1

Reference Summaries

ROUGE compares the generated summary with human-written reference summaries.

2

Overlap Calculation

ROUGE calculates the overlap between the generated summary and the reference summaries.

3

ROUGE Scores

Different ROUGE scores are generated based on various aspects of overlap, such as unigram, bigram, and longest common subsequence.

CIDEr

Image Captioning

CIDEr is a metric specifically designed for evaluating image captioning models.

N-gram Co-occurrence Statistics

CIDEr considers the co-occurrence of n-grams (sequences of words) between the generated caption and human-written reference captions.

Higher Scores

Higher CIDEr scores indicate that the generated caption is more similar to the human-written reference captions.



Distinct-n



Vocabulary Richness

Distinct-n measures the vocabulary richness of the generated text.



Text Diversity

A higher Distinct-n score indicates a more diverse vocabulary and a wider range of words used in the generated text.



Avoid Repetition

Distinct-n helps ensure that the generated text is not repetitive and avoids using the same words over and over again.

Conclusion and Key Takeaways

Deep generation models have become a powerful tool for generating creative content. Understanding and utilizing appropriate evaluation metrics is crucial for measuring the quality of generated outputs.

Choosing the Right Metrics

The choice of evaluation metrics depends on the specific task and the type of generative model being evaluated. It is often beneficial to use a combination of metrics to provide a comprehensive assessment of the model's performance.



Future Directions

The field of generative models is constantly evolving, with new metrics and evaluation methods being developed. Future research will focus on developing more sophisticated and reliable metrics that can capture the nuances of generative models, leading to more accurate and insightful assessments of their performance.



Challenges in Generative Model Evaluation

Subjectivity

Evaluating the quality of generated data is inherently subjective, as different individuals may have varying opinions on what constitutes "good" or "realistic." This makes it difficult to establish objective and universally accepted evaluation criteria.

Mode Collapse

Some generative models can fall into the trap of mode collapse, where they generate only a limited range of samples, leading to a lack of diversity and realism in the generated data. This makes it challenging to assess the model's performance accurately.

Challenges in Generative Model Evaluation

Data Distribution Shift

The distribution of generated data can differ from the real data distribution, making it difficult to directly compare the two. This requires careful consideration of the evaluation metrics and the intended use of the generated samples.

Lack of Ground Truth

In many cases, there is no clear ground truth or reference data for evaluating the generated samples. This makes it challenging to assess the model's performance accurately and reliably.

Applications of Generative Models



Image Generation

Generative models have made significant strides in creating photorealistic images, enabling applications such as artistic expression, image editing, and data augmentation.



Text Generation

Generative models have revolutionized natural language processing, enabling applications such as chatbot development, language translation, and creative writing.



Music Composition

Generative models can create original musical pieces in various styles and genres, offering new creative tools for musicians and composers.



3D Model Generation

Generative models can generate complex 3D models, facilitating applications such as game development, virtual reality, and product design.

Conclusion and Future Directions

Generative models have come a long way in recent years, demonstrating their potential to revolutionize various industries. From creating realistic and diverse data to facilitating new creative possibilities, these models are rapidly changing the landscape of artificial intelligence. As research continues to progress, we can expect even more innovative applications and breakthroughs in the field of generative modeling.



Conclusion and Future Directions

Future directions for generative models include improving their ability to generate higher-quality and more diverse data, addressing challenges in model evaluation, and exploring novel applications in domains such as drug discovery, materials science, and personalized medicine. With continued research and development, generative models hold the promise of revolutionizing our understanding of the world and unlocking new possibilities for innovation.



THANK YOU