VECTOR INSTITUTE

Vector's Diffusion Models Bootcamp Educational Resources

August 2024

Welcome to Vector's Diffusion Models Bootcamp educational resources! Explore what diffusion models are, their wide-ranging applications in healthcare, finance, and the environment, and learn about different algorithms. These essential readings will prepare you for the upcoming bootcamp, equipping you to apply diffusion models effectively in practice. At the end of this document, you will find an Extra Read section which offers supplementary materials for those eager to delve deeper into specific applications, advanced algorithms, and theoretical foundations of diffusion models.

Table of Contents

Introduction to Generative Models

A generative model is a machine learning model designed to understand the fundamental patterns or distributions within data so that it can create new data that resembles the original. There are various types of generative models: Bayesian networks, Diffusion Models, Generative Adversarial Networks (GANs), Variational Encoders (VAEs), etc.

- [video] Cornell CS 6785: Deep Generative Models. Lecture 1 https://www.youtube.com/watch?v=IZgvgLy1wyg
- [blog] Diffusion Models vs. GANs vs. VAEs: Comparison of Deep Generative Models https://towardsai.net/p/machine-learning/diffusion-models-vs-gans-vs-vaes-comparison-of-deep-generative-models

Generative models offer significant advantages beyond data generation. They enhance data augmentation by creating additional data in areas where it is scarce or costly, such as medical imaging. They are excellent at detecting anomalies by identifying outliers, which is vital in sectors like finance for spotting fraudulent transactions. Their versatility allows them to be used in various learning scenarios and to generate personalized content based on user preferences, improving experiences in entertainment. In design, generative models drive innovation by proposing novel ideas. They also reduce costs by automating content creation, making processes more efficient in industries like manufacturing and entertainment.

- [paper/book] Synthetic data generation: State of the art in healthcare domain. Provides comparison as well. https://www.sciencedirect.com/science/article/pii/S1574013723000138
- [paper/book] A computationally intelligent agent for detecting fake news using generative adversarial networks https://www.sciencedirect.com/science/article/pii/B9780128186992000044

Introduction to Diffusion Models

Diffusion models (DMs) are a class of generative models that create data by simulating a diffusion process. The process involves two main phases: the forward diffusion process and the reverse diffusion process. In the forward diffusion process, noise is gradually added to the data, transforming it into a simple distribution, usually

Gaussian noise, over several steps. Each step slightly corrupts the data, making it progressively noisier. The reverse diffusion process involves learning how to reverse this corruption. A neural network is trained to denoise the data step-by-step, effectively learning to transform noise back into the original data distribution. During training, the model learns the parameters needed to reverse the diffusion process by minimizing a loss function that measures the discrepancy between the generated data and the original data at each step. To generate new data, the model starts with a sample of noise and applies the learned reverse diffusion process, step-by-step, to produce data samples resembling the original training data.

- [blog] How diffusion models work: the math from scratch https://theaisummer.com/diffusion-models/
- [paper/book] A Comprehensive Survey On Generative Diffusion Models For Structured Data https://arxiv.org/pdf/2306.04139

DM for Tabular Data

DMs for tabular problems are being actively studied in the machine learning community. By leveraging DM's iterative noise-adding and noise-removing processes, these models can learn the underlying distributions and complex relationships within the data. For data generation, diffusion models create realistic synthetic datasets that preserve the properties of the original data, aiding in tasks such as data augmentation and privacy-preserving data sharing. For imputation, they effectively handle missing values by predicting and filling in gaps based on learned patterns, improving data quality and usability. Below we provide a list of algorithms that will be covered during the bootcamp:

- [paper/book] TabSyn: https://arxiv.org/pdf/2310.09656
- [paper/book] TabDDPM: https://proceedings.mlr.press/v202/kotelnikov23a/kotelnikov23a.pdf
- [paper/book] ClavaDDPM: Multi-relational Data Synthesis with Cluster-guided Diffusion Models https://arxiv.org/pdf/2405.17724

Evaluation Metrics

The quality of the synthetic data generated by the DMs can be evaluated using a variety of metrics. Below we discuss the most commonly used metrics.

- [paper/book] Mixed-Type Tabular Data Synthesis With Score-Based Diffusion In Latent Space https://arxiv.org/pdf/2310.09656

Low-order statistics: column-wise density estimation and pair-wise column correlation, which estimate the density of every single column and the correlation between every column pair.

- *Kolmogorov-Smirnov Test (KST)*: given two continuous distributions, KST quantifies the distance between the two distributions using the upper bound of discrepancy between two corresponding cumulative distribution functions.
- *Total Variation Distance (TVD)*: computes the frequency of each category value and expresses it as a probability. Then reports the average difference between the probabilities.
- *Pearson Correlation Coefficient*: measures whether two continuous distributions are linearly correlated.

High-order statistics: α-precision and β-recall scores that measure the overall fidelity and diversity of synthetic data

- [paper/book] How Faithful is your Synthetic Data? Sample-level Metrics for Evaluating and Auditing Generative Models https://arxiv.org/pdf/2102.08921

Machine Learning Efficiency: testing accuracy on real data when trained on synthetically generated tabular datasets

Privacy Protection: evaluate if the synthetic data is randomly sampled according to the distribution density rather than copied from the training data via Distance to Closest Records (DCR). If there is a privacy issue (e.g. if the synthetic set is directly copied from the training set), then the DCR scores for the training set should be closer to 0 than those for the testing set.

- Data Synthesis based on Generative Adversarial Networks https://arxiv.org/pdf/1806.03384

DM for Time Series Data

In addition to tabular problems, DMs have been used for time series forecasting, imputation and generation tasks. For forecasting, DMs can predict future values by learning from historical data trends. For generation, DMs can synthesize realistic time series data and for imputation they can fill in missing values by leveraging the learned

temporal structure, ensuring the imputed data remains consistent with observed patterns.

- [paper/book] A Survey on Diffusion Models for Time Series and Spatio-Temporal Data https://arxiv.org/pdf/2404.18886

Below we provide a list of algorithms that will be covered during the bootcamp:

- [paper/book] CSDI: Conditional Score-based Diffusion Models for Probabilistic Time Series Imputation https://openreview.net/pdf?id=VzuIzbRDrum
- [paper/book] Predict, Refine, Synthesize: Self-Guiding Diffusion Models for Probabilistic Time Series Forecasting https://arxiv.org/pdf/2307.11494

Evaluation Metrics

The quality of the forecasted and imputed data by DMs can be assessed using the following metrics:

- [paper/book] Predict, Refine, Synthesize: Self-Guiding Diffusion Models for Probabilistic Time Series Forecasting https://arxiv.org/pdf/2307.11494

Continuous ranked probability score (CRPS): a statistical metric that compares distributional predictions to ground-truth values.

- [blog] Continuous ranked probability score https://www.lokad.com/continuous-ranked-probability-score/

Linear Predictive Score (LPS): test CRPS of a linear (ridge) regression model trained on synthetic samples. The ridge regression model is a simple, standard model available in standard machine learning libraries (e.g., scikit-learn) that can effectively gauge the predictive quality of synthetic samples.

DM Use-cases

DMs are extensively applied across diverse fields, including healthcare, finance and transportation.

Health care

- [paper/book] A Transformer-based Diffusion Probabilistic Model for Heart Rate and Blood Pressure Forecasting in Intensive Care Unit https://arxiv.org/abs/2301.06625
- [paper/book] Synthetic Health-related Longitudinal Data with Mixed-type Variables Generated using Diffusion Models https://openreview.net/pdf/65b6780afb6f4970774b18552e37529381d88d01.pdf
- [paper/book] Density-Aware Temporal Attentive Step-wise Diffusion Model For Medical Time Series Imputation https://dl.acm.org/doi/abs/10.1145/3583780.3614840
- [paper/book] Improving Diffusion Models for ECG Imputation with an Augmented Template Prior https://arxiv.org/pdf/2310.15742

Finance

- [paper/book] Diffusion Variational Autoencoder for Tackling Stochasticity in Multi-Step Regression Stock Price Prediction https://arxiv.org/abs/2309.00073

Transportation

- [paper/book] SpecSTG: A Fast Spectral Diffusion Framework for Probabilistic Spatio-Temporal Traffic Forecasting https://arxiv.org/pdf/2401.08119
- [paper/book] Using a Diffusion Model for Pedestrian Trajectory Prediction in Semi-Open Autonomous Driving Environments https://ieeexplore.ieee.org/document/10489838

Extra Read

This section contains extra reads—additional resources to explore if you're interested in delving deeper into the topic. These readings provide supplementary insights and perspectives that complement the core material.

Visual Data

- [paper/book] Your Diffusion Model is Secretly a Zero-Shot Classifier https://diffusion-classifier.github.io/static/docs/DiffusionClassifier.pdf
- [blog] Diffusers library https://huggingface.co/docs/diffusers/en/index
- [paper/book] Tutorial on Diffusion Models for Imaging and Vision https://arxiv.org/pdf/2403.18103

Tabular Data

- [blog] Differentially Private (tabular) Generative Models Papers with Code https://github.com/ganevgv/dp-generative-models
- [blog] Synthcity: A library for generating and evaluating synthetic tabular data. https://github.com/vanderschaarlab/synthcity
- [paper/book] Quantifying and Mitigating Privacy Risks for Tabular Generative Models https://arxiv.org/pdf/2403.07842
- [paper/book] Systematic Assessment of Tabular Data Synthesis Algorithms: https://arxiv.org/pdf/2402.06806

Time Series Data

- [blog] Diffusion Model for Time Series and Spatio-Temporal Data https://github.com/yyysjz1997/Awesome-TimeSeries-SpatioTemporal-Diffusion-Model
- [paper/book] Diffusion Models for Time Series Applications: A Survey https://arxiv.org/pdf/2305.00624
- [paper/book] Imputation-based Time-Series Anomaly Detection with Conditional Weight-Incremental Diffusion Models https://dl.acm.org/doi/pdf/10.1145/3580305.3599391