UNIVERSITY OF MINNESOTA

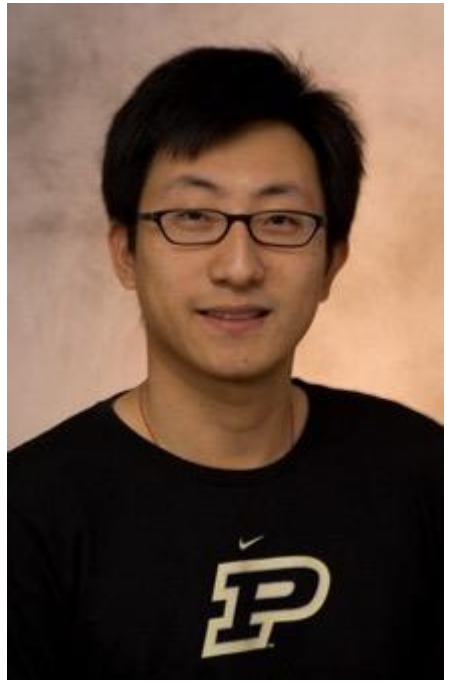
MissDiff: Training Diffusion Models on Tabular Data with Missing Values

Liyan Xie

Department of Industrial and Systems Engineering
University of Minnesota

Aknowledgements

Yidong Ouyang
UCLA



Guang Cheng
UCLA

Chongxuan Li
RUC

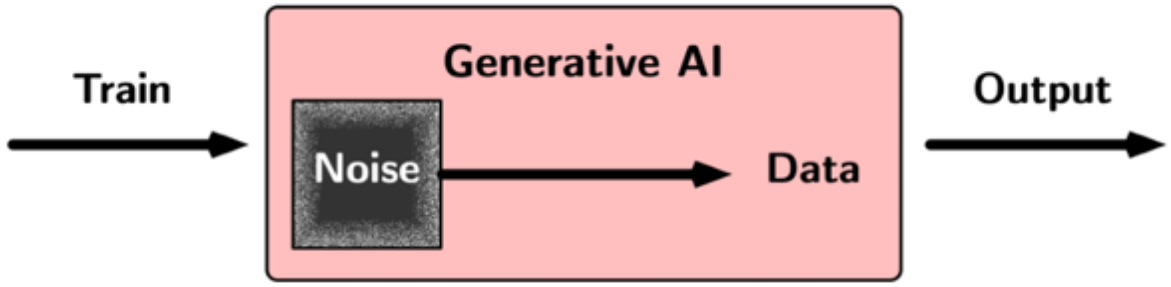
Reference:

Ouyang, X., Li, and Cheng. Missdiff: Training diffusion models on tabular data with missing values. arXiv preprint arXiv:2307.00467. (Partially presented at ICML 2023 Workshop on Structured Probabilistic Inference and Generative Modeling.)

Overview of Contribution

Incomplete data

Name	Gender	Age
Bill	Male	???
Miko	Female	15
Juan	???	45



Synthetic and **Complete** data

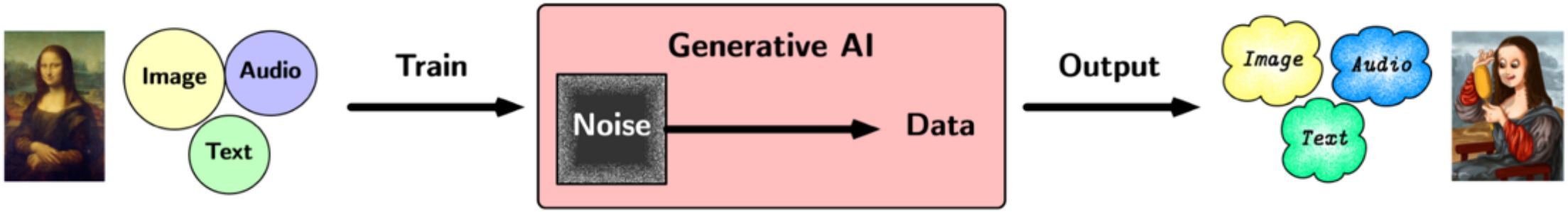
Name	Gender	Age
Bill	Male	25
Miko	Female	18
Mike	Male	45
Liyan	Female	32
Lisa	Female	20

- Vanilla Diffusion Model cannot directly learn from missing values.
- Propose **masked score matching** to learn a diffusion model on missing data.
- Provide some good properties of this objective and achieve good performance on both imputation tasks and generation tasks.

Outline

- Motivations and Preliminaries of Diffusion Models
- Problem Setup and Proposed Solution
- Some Theory
- Numerial Examples

Motivations: Synthetic Data Generation



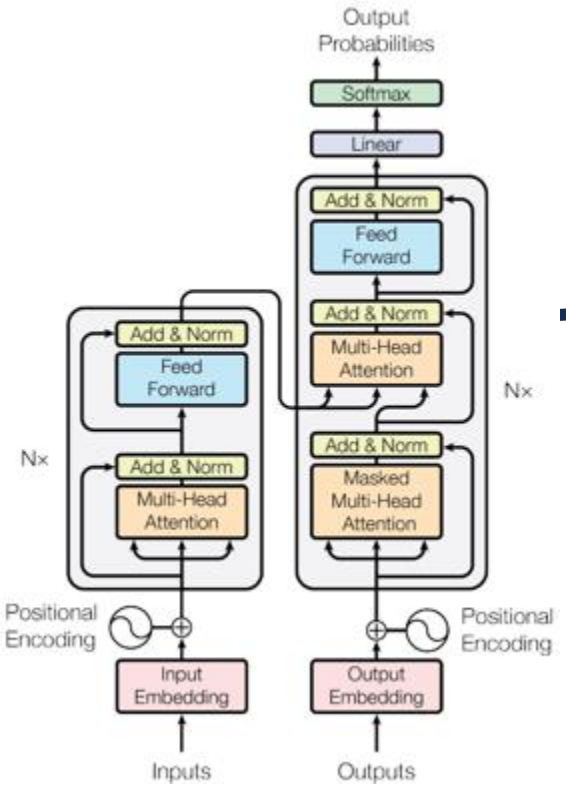
Given n i.i.d data sampled from some unknown distribution $p(x)$, generative models aim to learn the distribution $p(x)$ and generate samples from it.

- **Utility**: Generative models can be used to enhance the performance of the downstream model (classification/regression tasks) by generating a large amount of data for training.
- **Privacy**: Generative models can avoid using sensitive data from users by learning the distribution and generating "fake" but similar data.
- **Cost**: Generative models can reduce the cost of collecting and labeling data.

Emergence of Deep Generative AI

Billions even trillions of parameters

Model

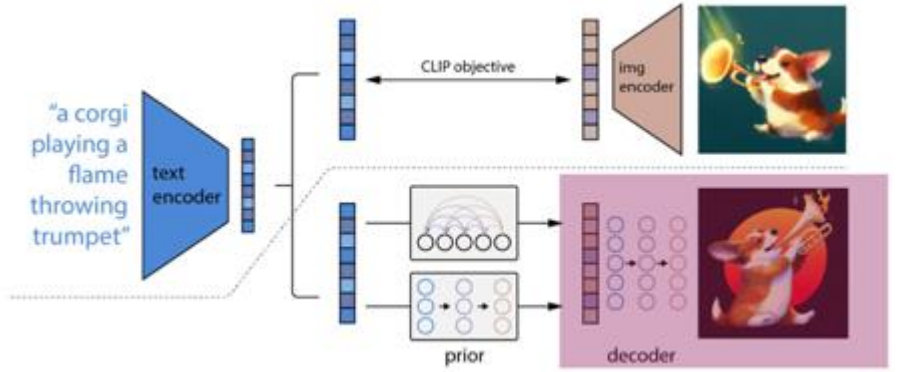


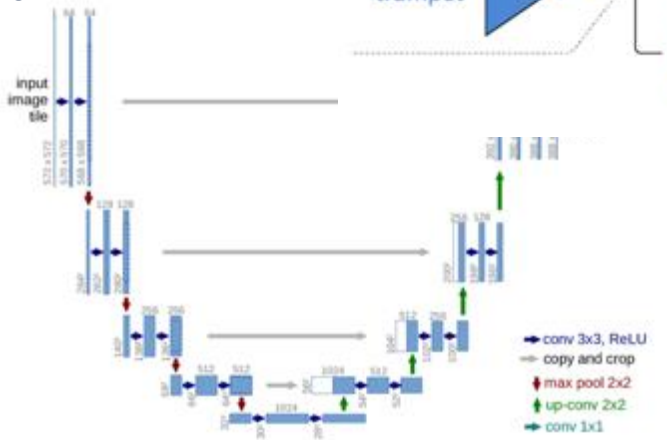
ChatGPT

Stable D

DALL-E 3

Transformer





Deep Generative AI

1998 2012 2015 2018 2021 2022 2023

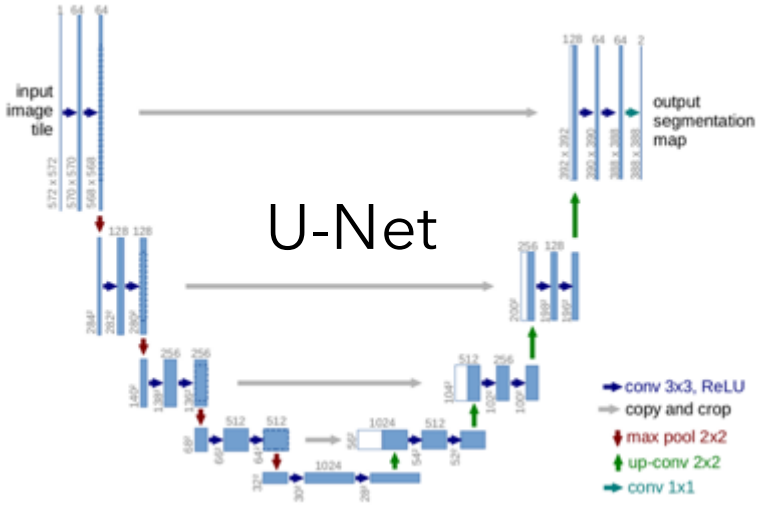
-- Thanks to blogs by Rockwell Anyoha, Toloka Team and Rick Merritt

A Revolution – Diffusion Model

- Sequential transformation

Noise

High-D



> 890M parameters
(Stable Diffusion)

Backbone of Stable Diffusion, Sora, etc.

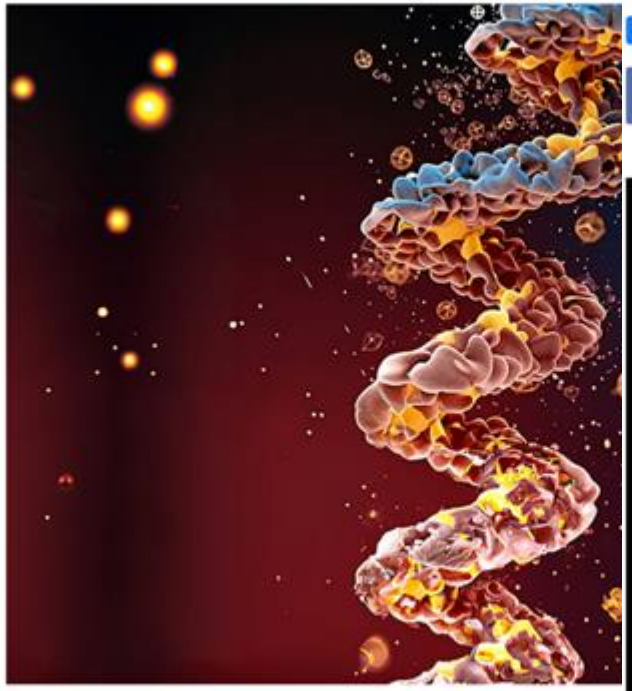
(Sohl-Dickstein et al., 2015)
(Song and Ermon, 2019)
(Ho et al., 2020)

Diffusion Model Is A Game Changer

Generative AI imagines new pr

"FrameDiff" is a computational tool that uses g
protein structures, with the aim of accelerating
improving gene therapy.

Rachel Gordon | MIT CSAIL
July 12, 2023



Biology is a wondrous yet delicate tapestry. At the heart is
encodes proteins, responsible for orchestrating the many
within the human body. However, our body is akin to a fine
losing its harmony. After all, we're faced with an ever-chan
pathogens, viruses, diseases, and cancer.

Diffusion models are
turbocharging reinfo
learning systems

By **Ben Dickson** - March 4, 2024

Like 75

Facebook Twitter Reddit

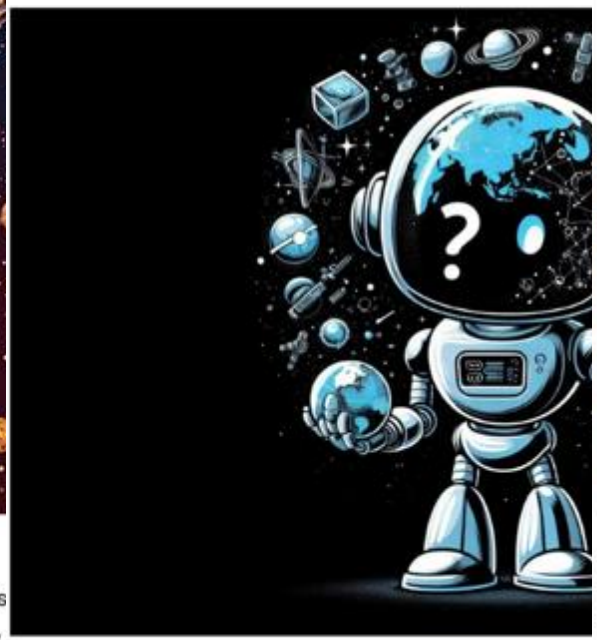


Image generated with Bing Image Creator

This article is part of our coverage of the late

ARTIFICIAL INTELLIGENCE

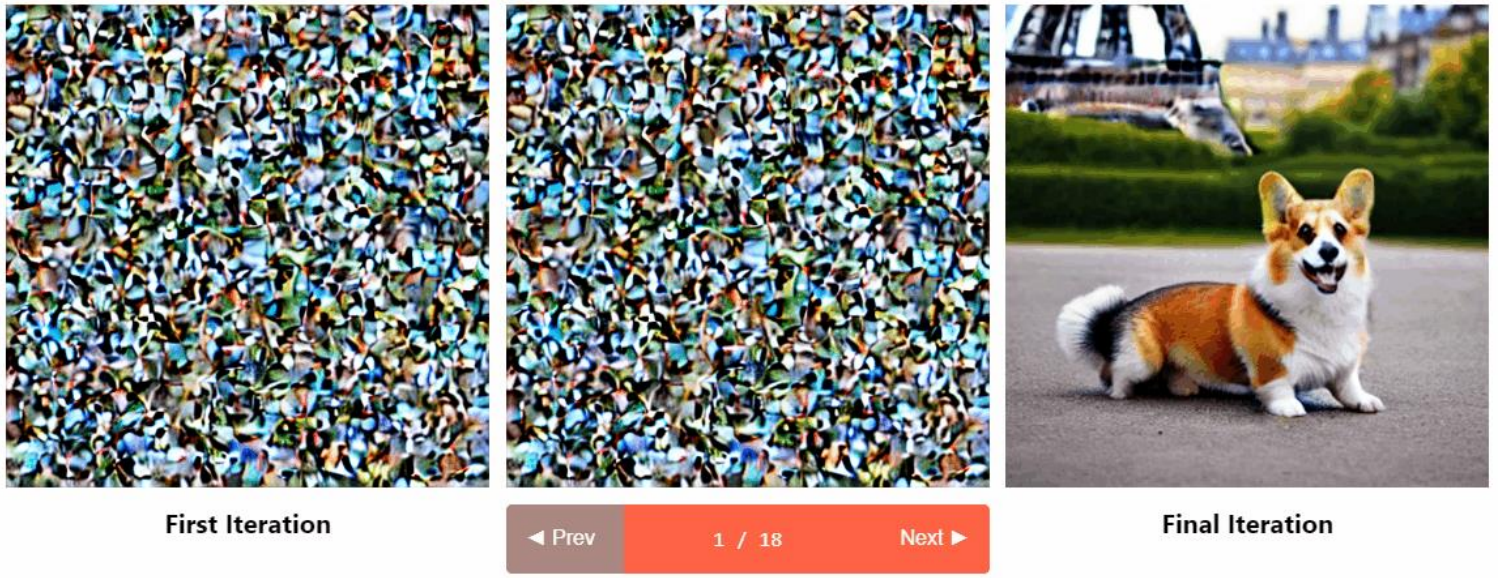
AniPortrait: Audio-Driven Synthesis of Photorealistic Portrait Animation

Published 3 days ago on May 3, 2024
By **Kunal Kejriwal**

Over the years, the creation of realistic and expressive portraits animations from static images and audio has found a range of applications including gaming, digital media, virtual reality, and a lot more. Despite its potential application, it is still difficult for developers to create frameworks capable of generating high-quality animations that maintain temporal consistency and are visually captivating. A major cause for the complexity is the need for intricate coordination of lip movements, head positions, and facial expressions to craft a visually compelling effect.

Diffusion Model Generates Samples

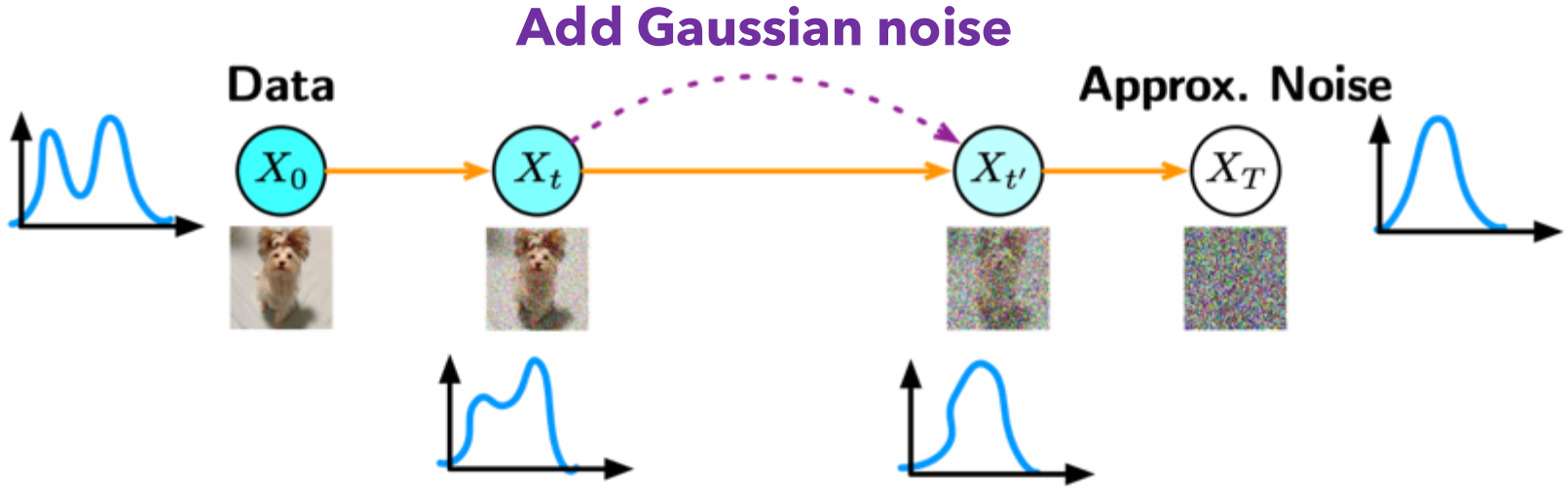
- Generate samples from noise



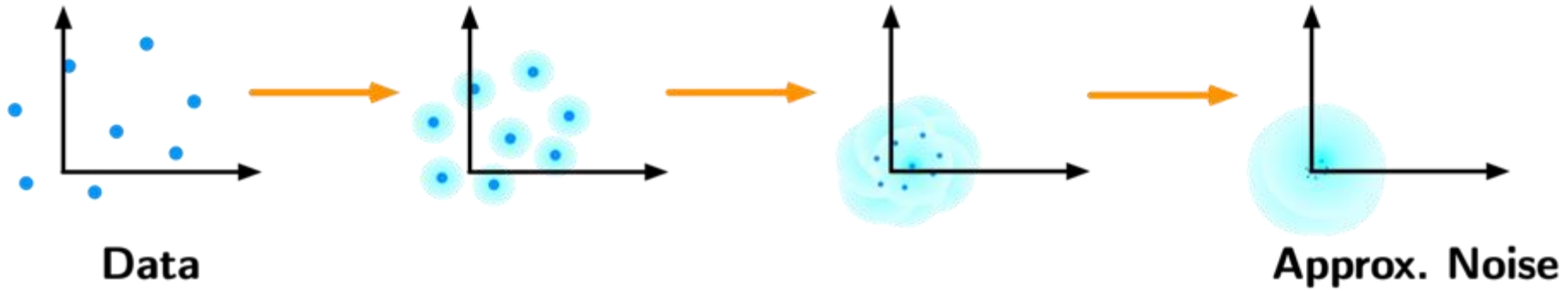
-- Credit: online

Forward Process – Noise Corruption

- Noise corruption process $\mathrm{d}X_t = -\frac{1}{2}X_t\mathrm{d}t + \mathrm{d}W_t$

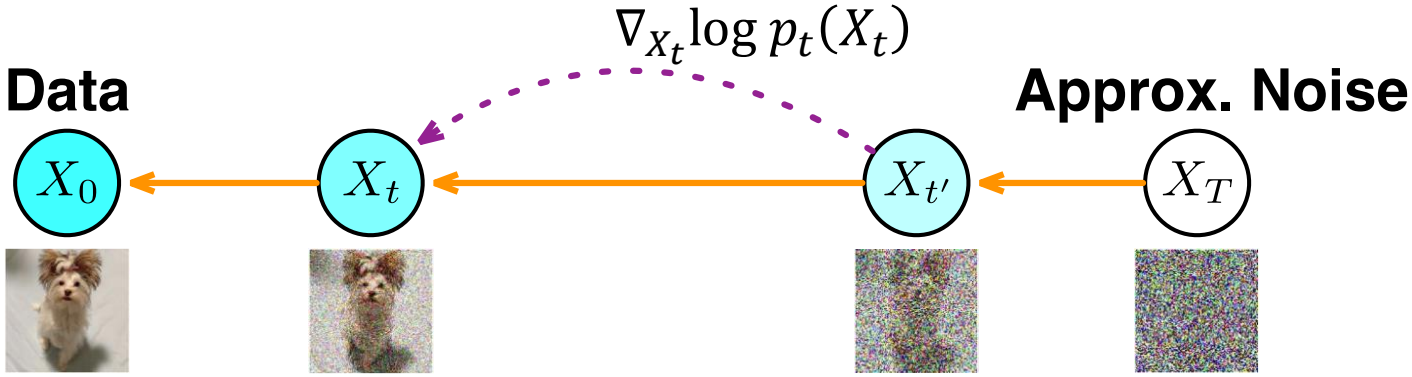


- The noise corruption



Backward Process – Sample Generation

- Time reversal in distribution



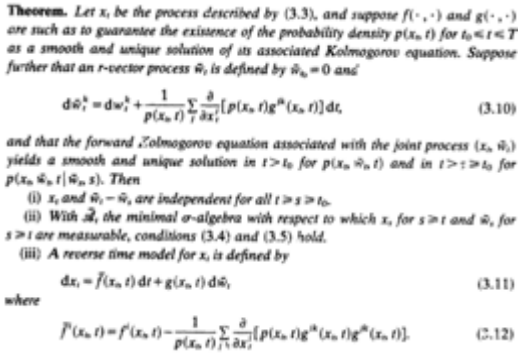
- The math (Anderson, 1982; Haussmann and Pardoux, 1986)

Forward $$dX_t = -\frac{1}{2}\beta_t X_t dt + \sqrt{\beta_t} dW_t$$

Brownian

Backward $$dX_t = \frac{1}{2}\beta_t X_t dt + \beta_t \nabla_{X_t} \log p_t(X_t) dt + \sqrt{\beta_t} dW_t$$

Score Function



Theorem. *Let x_t be the process described by (3.3), and suppose $f(\cdot,\cdot)$ and $g(\cdot,\cdot)$ are such as to guarantee the existence of the probability density $p(x_t, t)$ for $t_0 \leqslant t \leqslant T$ as a smooth and unique solution of its associated Kolmogorov equation. Suppose further that an r-vector process $\bar{w}_t$ is defined by $\bar{w}_{t_0} = 0$ and*

$$d\bar{w}_t^k = dw_t^k + \frac{1}{p(x_t, t)} \sum_i \frac{\partial}{\partial x_t^i} [p(x_t, t) g^{ik}(x_t, t)]\, dt, \tag{3.10}$$

and that the forward Kolmogorov equation associated with the joint process $(x_t, \bar{w}_t)$ yields a smooth and unique solution in $t > t_0$ for $p(x_t, \bar{w}_t, t)$ and in $t > \tau \geqslant t_0$ for $p(x_t, \bar{w}_t, t \mid \bar{w}_\tau, \tau)$. Then

(i) x_t and $\bar{w}_t - \bar{w}_\tau$ are independent for all $t \geqslant \tau \geqslant t_0$.

(ii) With $\bar{\mathcal{B}}_t$ the minimal σ-algebra with respect to which x_τ for $\tau \leqslant t$ and $\bar{w}_\tau$ for $\tau \geqslant t$ are measurable, conditions (3.4) and (3.5) hold.

(iii) A reverse time model for x_t is defined by

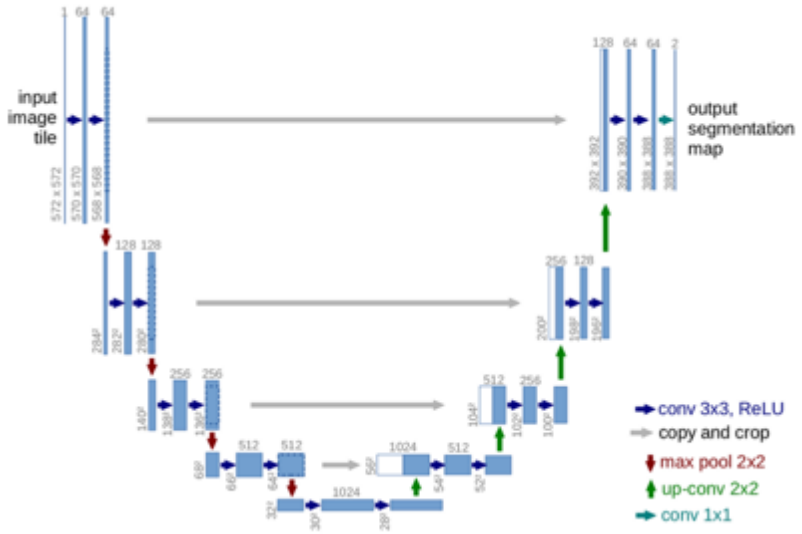
$$dx_t = \bar{f}(x_t, t)\, dt + g(x_t, t)\, d\bar{w}_t \tag{3.11}$$

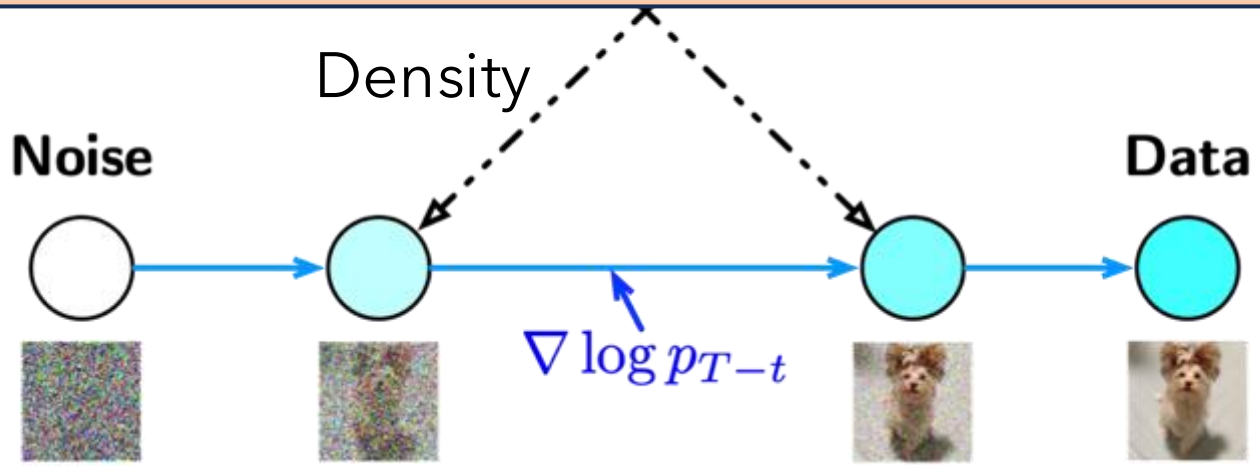
where

$$\bar{f}^i(x_t, t) = f^i(x_t, t) - \frac{1}{p(x_t, t)} \sum_j \frac{\partial}{\partial x_t^j} [p(x_t, t) g^{ik}(x_t, t) g^{jk}(x_t, t)]. \tag{3.12}$$

Forward and Backward Coupling

Training $\int_0^T \mathbb{E}_{x_t}\left[\|\nabla \log p_t(x_t) - s(x_t, t)\|_2^2\right] \mathrm{d}t$





Problem Setup

Tabular data with Missing Values

Name	Gender	Age
Bill	Male	???
Miko	Female	15
Juan	???	45

- Underlying complete d-dimensional data $\mathbf{x} = (x_1, \ldots, x_d) \in \mathcal{X}$
- Binary mask $\mathbf{m}$

$$m_i = \begin{cases} 1 & \text{if } x_i \text{ is observed,} \\ 0 & \text{if } x_i \text{ is missing.} \end{cases}$$

- Observed data:

$$S^{\text{obs}} = \{\mathbf{x}_i^{\text{obs}}\}_{i=1}^{\tilde{n}} \text{ with } \mathbf{x}^{\text{obs}} = \mathbf{x} \odot \mathbf{m} + \text{na} \odot (\mathbf{1} - \mathbf{m})$$

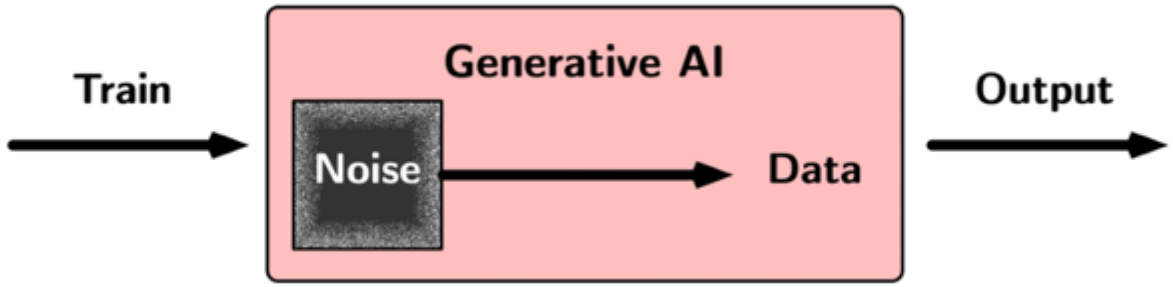
- Some common missing mechanisms
  - Missing Completely At Random (MCAR): $\mathbf{m}$ is independent with the complete data $\mathbf{x}$
  - Missing At Random (MAR): $\mathbf{m}$ only depends on the observed value $\mathbf{x}^{\text{obs}}$
  - Not Missing At Random (NMAR): $\mathbf{m}$ depends on the observed value $\mathbf{x}^{\text{obs}}$ and missing value

Problem Setup

Research objective

Incomplete data

Name	Gender	Age
Bill	Male	???
Miko	Female	15
Juan	???	45



Synthetic and **Complete** data

Name	Gender	Age
Bill	Male	25
Miko	Female	18
Mike	Male	45
Liyan	Female	32
Lisa	Female	20

- Learn a diffusion model, i.e., learn the score network $s_\theta(\mathrm{x}_t, t)$, from training data with missing values directly.
- Usage: the learned diffusion model can be used to generate more synthetic (and complete) data, as well as doing imputation task for missing data.

Related Work #1 (on Generation task)

"Impute-then-generate" Framework

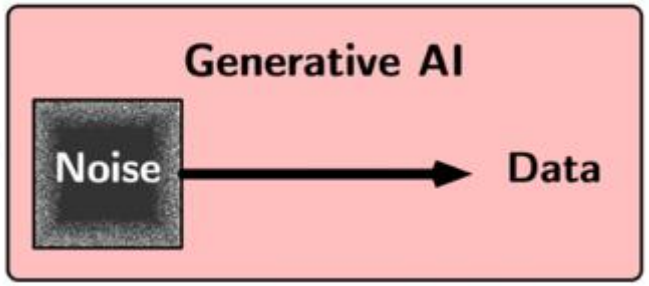
Incomplete data

Name	Gender	Age
Bill	Male	???
Miko	Female	15
Juan	???	45

Impute

Name	Gender	Age
Bill	Male	25
Miko	Female	15
Juan	Male	45

Train



- Maximum likelihood framework
  - Use $f_\varphi(\mathbf{x}^{\text{obs}})$ to impute, where $f_\varphi(\mathbf{x}^{\text{obs}})$ is trained by minimizing the regression loss $\mathbb{E}_{(\mathbf{x}^{\text{obs}},\mathbf{x}^{\text{miss}})\sim p_0(\mathbf{x})}\|f_\varphi(\mathbf{x}^{\text{obs}}) - \mathbf{x}^{\text{miss}}\|^2$.
  - Train the generative model by $\max_\phi \log p_\phi(\mathbf{x}^{\text{obs}}, \mathbf{x}^{\text{miss}} := f_\varphi(\mathbf{x}^{\text{obs}}))$.
- Limitation: optimal single/multiple imputation may fail to capture the data variability.

Bertsimas, Dimitris, Arthur Delarue, and Jean Pauphilet. "Prediction with missing data." *stat* 1050 (2021): 7.
Niels Bruun Ipsen, Pierre-Alexandre Mattei, and Jes Frellsen. How to deal with missing data in supervised deep learning? ICLR 2022.

Related Work #2 (on Imputation task)

Existing imputation methods cannot be used for generating new samples

Name	Gender	Age
Bill	Male	???
Miko	Female	15
Juan	???	45

Impute →

Name	Gender	Age
Bill	Male	25
Miko	Female	15
Juan	Male	45

- When using the conditional diffusion model for generation, there exist a **mismatch** between training and inference since no information can be conditioned for generation tasks.
- Variational Auto-encoder usually uses a student t distribution with location, scale, and degrees of freedom outputted by the decoder, which has **limited representation** power for the real distribution.

Tashiro, Y., Song, J., Song, Y., & Ermon, S. CSDI: Conditional Score-based Diffusion Models for Probabilistic Time Series Imputation. NeurIPS 2021.
Mattei, P., & Frellsen, J. MIWAE: Deep Generative Modelling and Imputation of Incomplete Data Sets. ICML 2019.

Proposed Method

Denoising score matching on incomplete data:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} J_{DSM}(\boldsymbol{\theta})$$

$$:= \frac{T}{2}\mathbb{E}_t\left\{\lambda(t)\mathbb{E}_{\mathbf{x}^{\text{obs}}(0)}\mathbb{E}_{\mathbf{x}^{\text{obs}}(t)|\mathbf{x}^{\text{obs}}(0)}\left[\left\|\left(\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}^{\text{obs}}(t),t) - \nabla_{\mathbf{x}^{\text{obs}}(t)}\log p(\mathbf{x}^{\text{obs}}(t)\mid\mathbf{x}^{\text{obs}}(0))\right)\odot\mathbf{m}\right\|_2^2\right]\right\}$$

- where $\lambda(t)$ is a positive weighting function
- recall the binary mask $\mathbf{m}$ indicates the observed entries $m_i = \begin{cases} 1 & \text{if } x_i \text{ is observed,} \\ 0 & \text{if } x_i \text{ is missing.} \end{cases}$

Algorithm 1 *MissDiff*: Denoising Score Matching on Data with Missing Values

Require: Diffusion process hyperparameter β_t, σ_t, denote $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t}\alpha_s$.

1: **repeat**
2: Sample $\mathbf{x}_0^{\text{obs}}$ according to the data distribution and missing mechanism;
3: Infer mask $\mathbf{m} = \mathbb{1}[\mathbf{x}_0^{\text{obs}} \neq \text{na}]$;
4: $t \sim \text{Uniform}(\{1,\ldots,T\})$;
5: $\epsilon_t \sim \mathcal{N}(\mathbf{0},\mathbf{I})$;
6: Take gradient descent step on

$$\nabla_\theta \left\|(\epsilon_t - \mathbf{s}_{\boldsymbol{\theta}}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0^{\text{obs}} + \sqrt{1-\bar{\alpha}_t}\epsilon_t, t))\odot\mathbf{m}\right\|^2.$$

7: **until** converged.

Detaild Algorithms

For Imputation

Algorithm 2 *MissDiff* for Imputation

Require: Observed data $\mathbf{x}_0^{\text{obs}}$, Diffusion model $\mathbf{s}_{\boldsymbol{\theta}}$, hyperparameter β_t, σ_t, denote $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$.

Sample $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbb{I})$;
Infer mask $\mathbf{m} = \mathbb{I}[\mathbf{x}_0^{\text{obs}} \neq \text{na}]$;
$t = T$;
while $t \neq 0$ **do**
  Sample $\epsilon_t^{\text{obs}} \sim \mathcal{N}(\mathbf{0}, \mathbb{I})$ if $t > 1$, else $\epsilon_t^{\text{obs}} = \mathbf{0}$;
  $\mathbf{x}_{t-1}^{\text{obs}} = \sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0^{\text{obs}} + (1 - \bar{\alpha}_{t-1})\,\epsilon_t^{\text{obs}}$
  Sample $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbb{I})$ if $t > 1$, else $\epsilon_t = \mathbf{0}$;
  $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)) + \sigma_t\epsilon_t$;
  $\mathbf{x}_{t-1} = \mathbf{m} \odot \mathbf{x}_{t-1}^{\text{obs}} + (\mathbf{1} - \mathbf{m}) \odot \mathbf{x}_{t-1}$
  $t = t - 1$;
end while
return $\mathbf{x}_0$.

For Generation

Algorithm 3 *MissDiff* for Generation

Require: Diffusion model $\mathbf{s}_{\boldsymbol{\theta}}$, hyperparameter β_t, σ_t, denote $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$.

Sample $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbb{I})$;
$t = T$;
while $t \neq 0$ **do**
  Sample $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbb{I})$ if $t > 1$, else $\epsilon_t = \mathbf{0}$;
  $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)) + \sigma_t\epsilon_t$;
  $t = t - 1$;
end while
return $\mathbf{x}_0$.

Some Theoretical Guarantee

Theoretical Results

- Denoising Score Matching on missing data can learn the oracle score

Theorem (oracle score)

Denote $\boldsymbol{\rho}(\mathbf{x}) = [\rho_1, \ldots, \rho_d] = \mathbb{E}_{p(\mathbf{m}|\mathbf{x})}[\mathbf{1} - \mathbf{m}]$ *as the missing probability of each entry. Define* $\rho_{max} := \max_{i=1,\ldots,d} \sup_{\mathbf{x}} \rho_i(\mathbf{x})$ *as the supreme of missing rates and assume* $\rho_{max} < 1$. *Let* $\boldsymbol{\theta}^*$ *be the solution to our training objective. Then we have*

$$\mathbf{s}_{\boldsymbol{\theta}^*}(\mathbf{x}(t), t) = \nabla_{\mathbf{x}(t)} \log p_t(\mathbf{x}(t)).$$

Theoretical Results

- Denoising Score Matching objectives on missing data is an upper bound for the negative likelihood of the generative model on observed data $\mathbf{x}^{\mathbf{obs}}$

Theorem

Under the same condition of the last Theorem and mild regularity conditions, we have

$$-\mathbb{E}_{p(\mathbf{x}^{obs})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x})\right] \leq \frac{1}{1-\rho_{max}} J_{\mathrm{DSM}}\left(\boldsymbol{\theta}\right) + C_1,$$

where C_1 is a constant independent of $\boldsymbol{\theta}$.

➢For the special case of no missing, i.e., $\rho_{\max} = 0$, our result degenerates to previous results $-\mathbb{E}_{p(\mathbf{x})}[\log p_{\boldsymbol{\theta}}(\mathbf{x})] \leq J_{\mathrm{DSM}}(\boldsymbol{\theta}) + C_1$

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. NeurIPS 2021.

Numerical Experiments

For Imputation tasks

- Evaluation Criterion: Root Mean Squared Error (RMSE) between the predicted value with the oracle missing value.

Table: Evaluation on imputation tasks. The standard deviations of five independent trials are shown in the parenthesis. The *lower* the RMSE, the *better* the performance.

Method	Census	Breast	Wine	Concrete	Libras	diabetes
Mean /Mode	0.120(0.003)	0.263(0.009)	0.076(0.003)	0.217(0.007)	0.099(0.001)	0.222(0.003)
MICE(linear)	0.101(0.002)	0.154(0.011)	0.065(0.003)	0.153(0.006)	0.034(0.001)	0.263(0.002)
MissForest	0.112(0.004)	0.163(0.014)	0.060(0.002)	0.173(0.005)	0.024(0.001)	0.216(0.003)
GAIN	0.123(0.057)	0.165(0.006)	0.072(0.004)	0.203(0.007)	0.089(0.006)	0.202(0.003)
MIWAE	0.113(0.042)	0.1874(0.079)	0.074 (0.005)	0.195(0.006)	0.083(0.003)	0.194(0.081)
CSDI_T	0.099(0.003)	0.153(0.003)	0.065(0.004)	**0.131(0.008)**	**0.011(0.001)**	0.197(0.001)
MissDiff	**0.089(0.006)**	**0.136(0.002)**	**0.053(0.001)**	0.161(0.001)	0.0787(0.002)	**0.051(0.004)**

UCI datasets with up to 20k training samples and 20 columns.

Numerical Experiments

For Generation tasks

- Evaluation Criterion: Training various models, including Decision Tree, AdaBoost, Logistic/Linear Regression, MLP classifier/regressor, RandomForest, and XGBoost, on synthetic data, and test them with real data.

Table: *Utility* (RMSE) evaluation of *MissDiff* on MIMIC4ED dataset. The *lower* the RMSE, the *better* the performance.

	MissDiff	*Diff-mean*	*STaSy-mean*	CSDI_T
Row Missing	**1.826**	2.166	1.894	1.853
Column Missing	**1.834**	2.011	1.935	1.874
Independent Missing	**1.852**	2.483	1.972	1.879

MIMIC4ED is an electronic health record in the emergency department with more than 350k training samples and 73 columns.

Summary

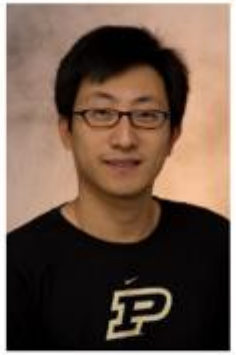
- A diffusion-based unified framework for imputation and complete sample generation by learning from data with missing values
- Theoretical justifications on recovering the oracle score function and upper bounding the negative likelihood of the observed data under mild assumptions
- The proposed method outperforms existing state-of-the-art methods in most real datasets on both imputation tasks and generation tasks
- Future directions:
  - ❖ Beyond tabular data: time series data, multi-modal data…
  - ❖ More involved theoretical study on the finite-sample estimation error of the score function, and the resulting convergence rate of the estimated distributions

Thank you!

xlyustc@gmail.com

https://liyanxie.github.io/

Yidong Ouyang
UCLA



Guang Cheng
UCLA

Chongxuan Li
RUC

Reference:

Ouyang, X., Li, and Cheng. Missdiff: Training diffusion models on tabular data with missing values. arXiv preprint arXiv:2307.00467. (Partially presented at ICML 2023 Workshop on Structured Probabilistic Inference and Generative Modeling.)