CSDI: Conditional Score-based Diffusion Models for Probabilistic Time Series Imputation

Tashiro, V; Song, J; Song, T; Ermon, S

Stanford University

March 6, 2023

Presented by Scott Sun from Duke B&B

Introduction

The authors propose a new self-supervised learning approach to solve missing data problem in time series through imputation, which is called **Conditional Score-based Diffusion Model (CSDM)**.

A diffusion model is class of of deep generative model that generates samples by gradually converting (or say denoising) the original random noise into a plausible data sample. There have been lots of real-world applications built upon diffusion models (e.g. AI art).

Roadmap

- Background of DDPM
- Imputation with CSDI
- Training of CSDI (a self-supervised learning)
- Implementation for time series imputation
- Experimental results

Background: Denoising Diffusion Probabilistic Methods (DDPM by Ho, et. al.)

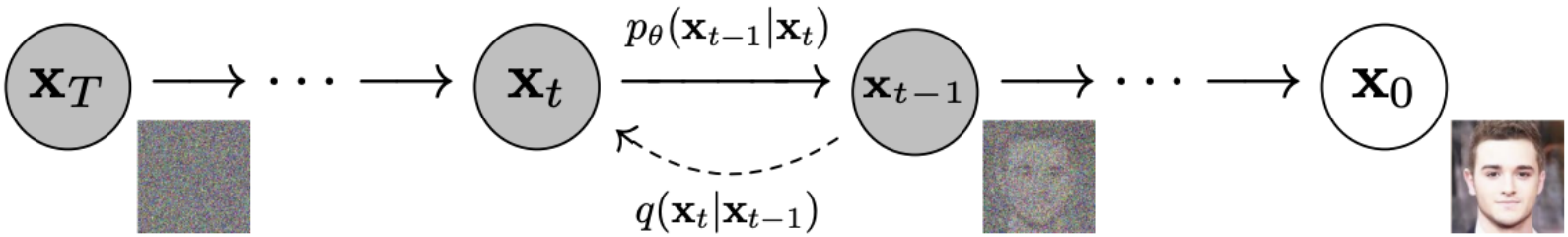

Figure: The directed graphical model considered in this work (Ho, et. al.)

Background: Denoising Diffusion Probabilistic Methods (DDPM by Ho, et. al.)

- Markov chain
- forward/diffusion process, $q(x_t|x_{t-1})$
- reverse/denoising process, $p_\theta(x_{t-1}|x_t)$

Background: Forward Process

By the Markov chain property,

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}), \text{ where } x_t|x_{t-1} \sim \mathcal{N}(\sqrt{1-\beta_t}x_{t-1}, \beta_t \boldsymbol{I})$$

Also, Ho et. al. proved that $\underbrace{x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon}_{\text{Reparameterization Trick}}$, where $\epsilon \sim \mathcal{N}(0, \boldsymbol{I})$,

and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_t = \prod_{i=1}^{t}(1-\beta_t)$

Background: Backward Process

By the Markov Chain property,

$$p_\theta(x_{0:T}) = p(x_T) \prod_{i=1}^{T} p_\theta(x_{t-1}|x_t), x_T \sim \mathcal{N}(0, \boldsymbol{I})$$

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_\theta(x_t, t)\boldsymbol{I})$$

To simply the math to some straightforward idea: μ_θ contains a trainable denoising function ϵ_θ, and σ_θ is a function of β_t's.

Optimization objective:

$$\min_\theta \mathcal{L}(\theta) := \min_\theta \mathbb{E}_{x_0 \sim q_0, \epsilon \sim \mathcal{N}(0, \boldsymbol{I}), t} ||\epsilon - \epsilon_\theta(x_t)||^2$$

In human language, we are trying to learn a denoising function that approximates the random noise we added to the data in the forward process.

Imputation with Diffusion Model (conditionally)

The goal of probabilistic imputation is to estimate the true conditional data distribution with a model distribution $p_\theta(x_0^{ta}|x_0^{co})$. Presumably, this method is valid for MAR.

$$p_\theta(x_{0:T}^{ta}|x_0^{co}) = p(x_T^{ta}) \prod_{i=1}^{T} p_\theta(x_{t-1}^{ta}|x_t^{ta}, x_0^{co}), x_T^{ta} \sim \mathcal{N}(0, \boldsymbol{I})$$

$$p_\theta(x_{t-1}^{ta}|x_t^{ta}, x_0^{co}) = \mathcal{N}(x_{t-1}^{ta}; \mu_\theta(x_t^{ta}, t|x_0^{co}), \sigma_\theta(x_t^{ta}, t|x_0^{co})\boldsymbol{I})$$

CSDI: A Visual Representation

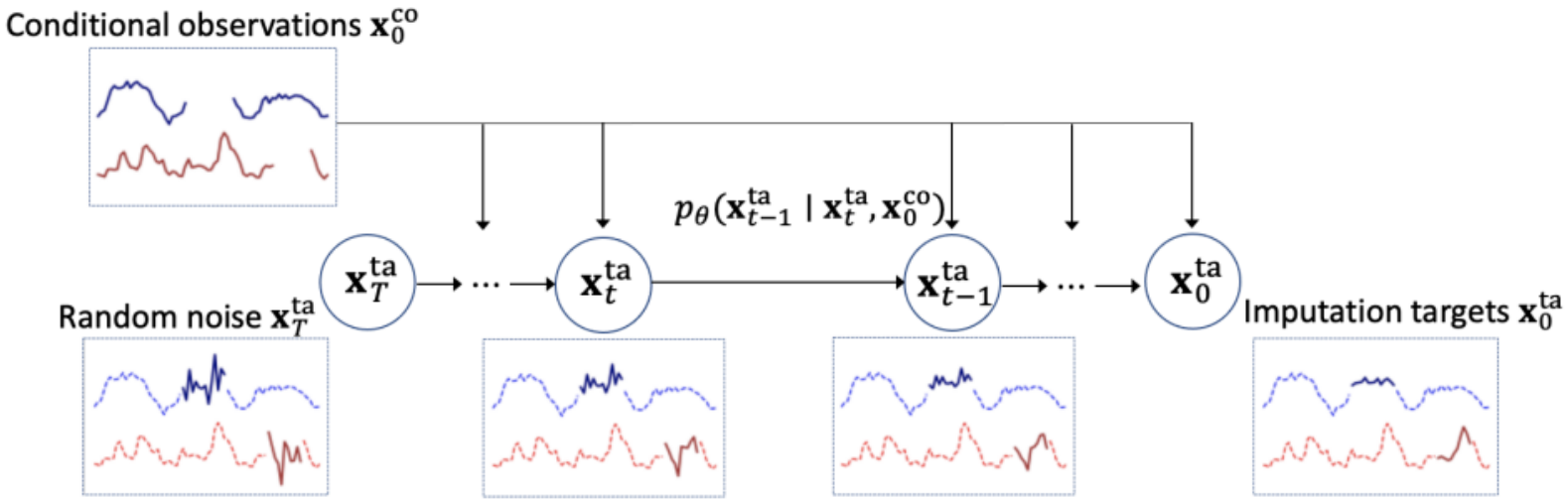

Figure: The procedure of time series imputation with CSDI (Tashiro et. al.)

Training of CSDI

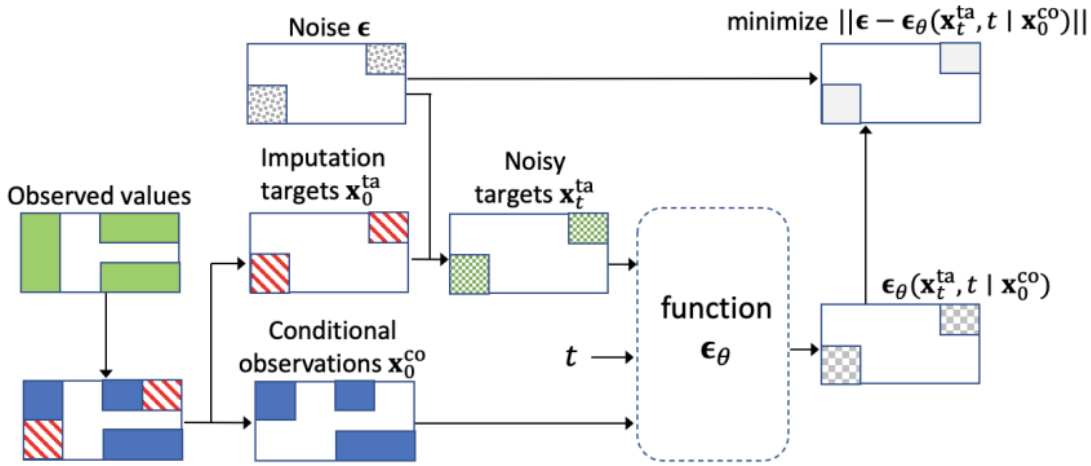

Figure: The self-supervised training procedure of CSDI (Tashiro et. al.)

On the middle left rectangle, the green and white areas represent observed and missing values, respectively. The observed values are into imputation targets and conditional covariates through a specified strategy so that we can train ϵ_θ.

Training of CSDI

	x_0^{ta}	x_0^{co}
training	a subset of observed values (red)	the remaining observed values (blue)
sampling/imputation	all missing values	all observed values (green)

Table: Imputation targets x_0^{ta} and conditional observations x_0^{co} for CSDI at training and sampling

Three training strategies are discussed here, including 1. random strategy, 2. historical strategy, 3. mix strategy, and 4. test pattern strategy.

Implementation of CSDI for imputation

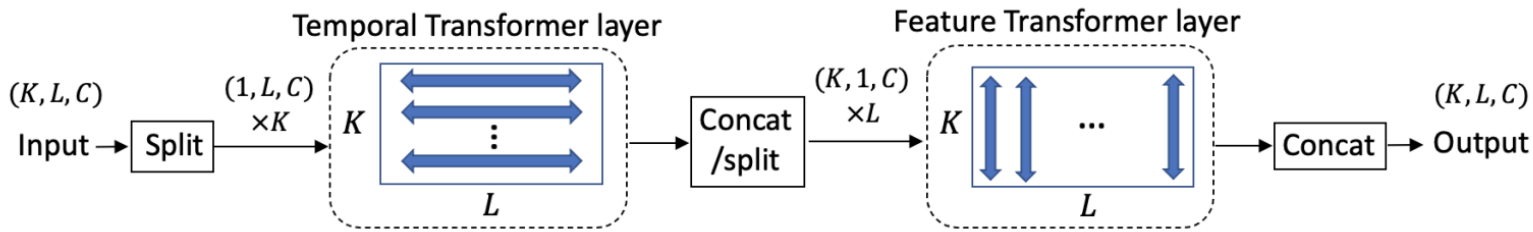

Figure: The architecture of 2D attention. Given a tensor with K features, L length, and C channels, the temporal Transformer layer takes tensors with $(1, L, C)$ shape as inputs and learns temporal dependency. The feature Transformer layer takes tensors with $(K, 1, C)$ shape as inputs and learns feature dependency. The output shape of each layer is the same as the input shape.

Experiment: Evaluation metric

Continuous Ranked Probability Score (CRPS) (Matheson & Winkler, 1976) is a measure of error for probabilistic forecast, which can be thought as a generalization of *MAE* for deterministic forecast.

$$CRPS(\hat{F}, x_{ta}) = \int_x \Big(\hat{F}(x) - H(x - x_{ta})\Big)^2 dx$$

However, the paper derives it in another form in terms of quantile loss

$$\begin{aligned} CRPS(\hat{F}^{-1}, x) &= \int_0^1 2\lambda_\alpha(\hat{F}^{-1}(\alpha), x) d\alpha \\ &\approx \frac{1}{19} \sum_{i=1}^{19} 2\lambda_{0.05i}(\hat{F}^{-1}(0.05i), x) \end{aligned}$$

Finally, a normalized average *CRPS* is used to summarize the overall performance

$$\frac{\sum_{k,l} CRPS(F_{k,l}^{-1}, x_{k,l})}{\sum_{k,l} |x_{k,l}|}$$

Experiment: Result

Table 2: Comparing CRPS for probabilistic imputation baselines and CSDI (lower is better). We report the mean and the standard error of CRPS for five trials.

	healthcare			air quality
	10% missing	50% missing	90% missing	
Multitask GP [31]	0.489(0.005)	0.581(0.003)	0.942(0.010)	0.301(0.003)
GP-VAE [10]	0.574(0.003)	0.774(0.004)	0.998(0.001)	0.397(0.009)
V-RIN [32]	0.808(0.008)	0.831(0.005)	0.922(0.003)	0.526(0.025)
unconditional	0.360(0.007)	0.458(0.008)	0.671(0.007)	0.135(0.001)
CSDI (proposed)	**0.238(0.001)**	**0.330(0.002)**	**0.522(0.002)**	**0.108(0.001)**

The paper refers that GP-VAE once was the state-of-the-art

Experiment: Result

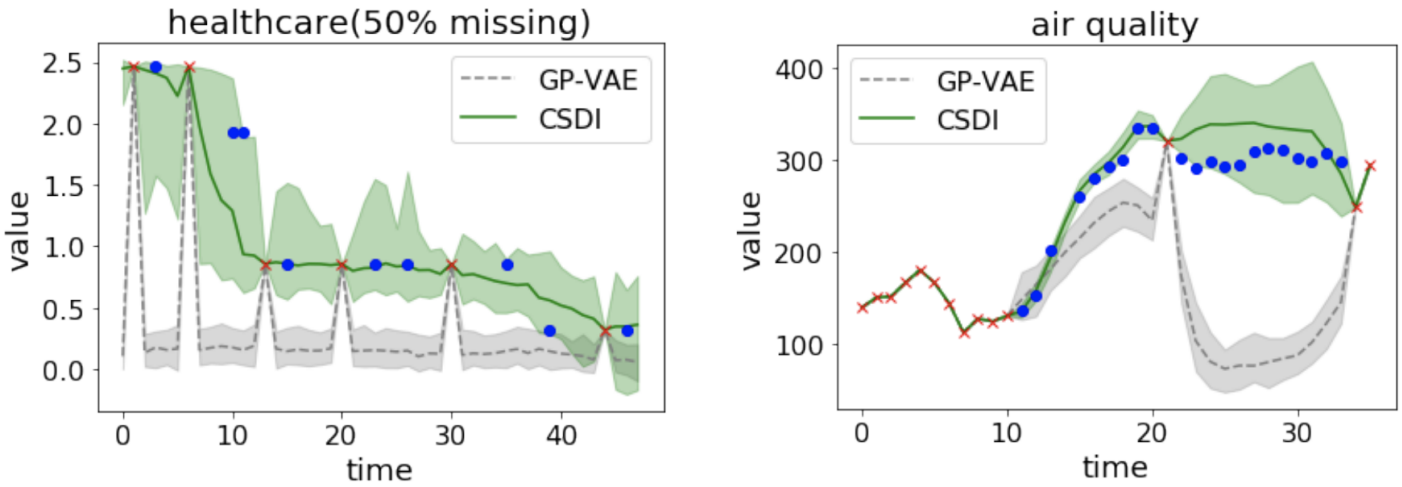

Figure 4: Examples of probabilistic time series imputation for the healthcare dataset with 50% missing (left) and the air quality dataset (right). The red crosses show the observed values and the blue circles show the ground-truth imputation targets. For each method, median values of imputations are shown as the line and 5% and 95% quantiles are shown as the shade.

The paper refers that GP-VAE once was the state-of-the-art

Future study

- Some basic stat theory about Variational Inference
- VAE
- More about diffusion model