Improving Structured Data ML Process with GANs

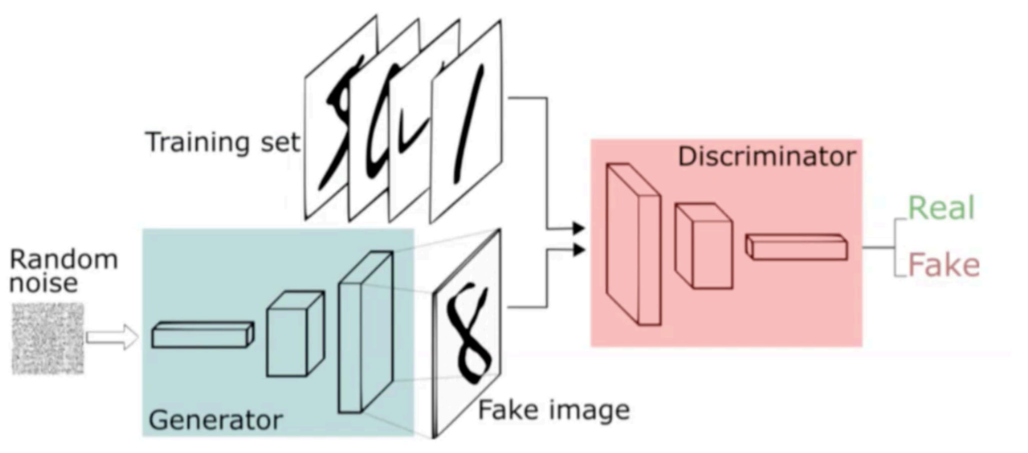
3/30/21 11:10 AM

Speaker: Srinivas Chilukuri (ZS Associates)
Date: 03/30/2021

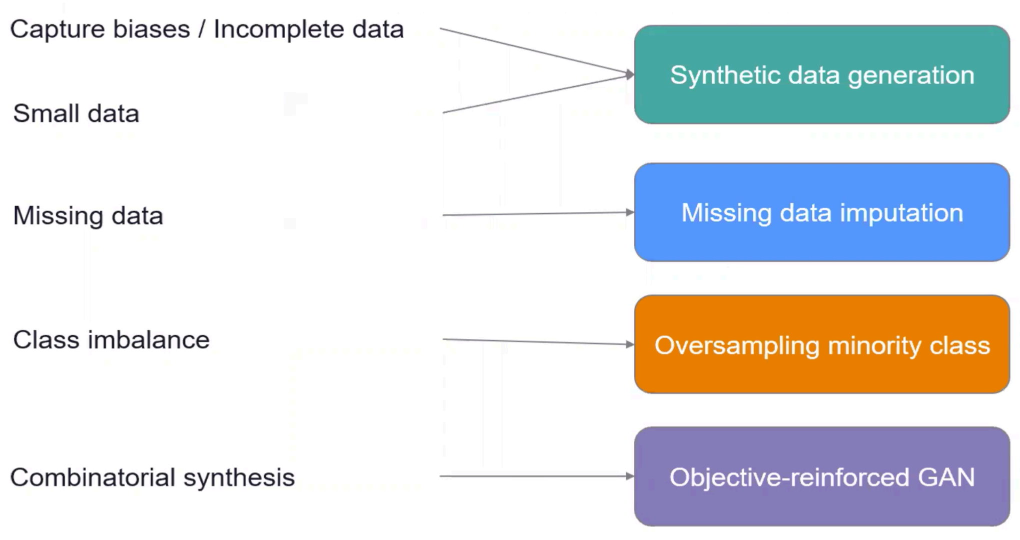
Structured datasets have unique challenges of their own

- Capture biases
- Missing data
- Small data
- Class imbalance

GAN overview for image generation



Applications of GANs


Capture biases / Incomplete data

Small data

Synthetic data generation

Missing data

Missing data imputation

Class imbalance

Oversampling minority class

Combinatorial synthesis

Objective-reinforced GAN

Synthetic generation of tabular data

- Generator create values according to schema of columns
- Discriminator produces log probability

Domain specific tabular GANs

- E-commerce:
    - Generate low-dimensional dense representation of e-commerce orders
    - To generate plausible orders data for new products
- Insurance: (tackle data privacy issue)
    - Generate accessible insurance datasets for actuarial studies
    - Tweaked CTGAN for imbalance categorical levels
- Medical (medGAN)
    - Generate privacy preserving, accessible healthcare datasets
    - Use of recurrent Encoders/Decoders for handling sequences
    - Does not handle lab values, clinical notes, diagnostic images, …
- Relational datasets

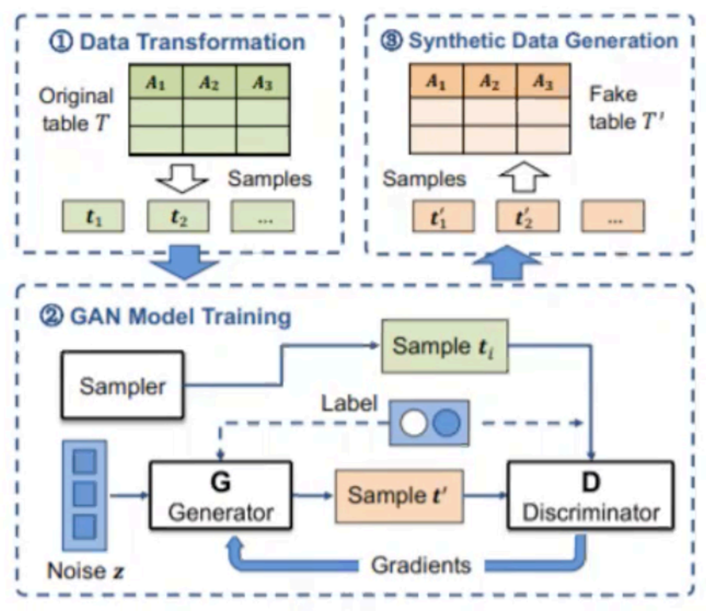
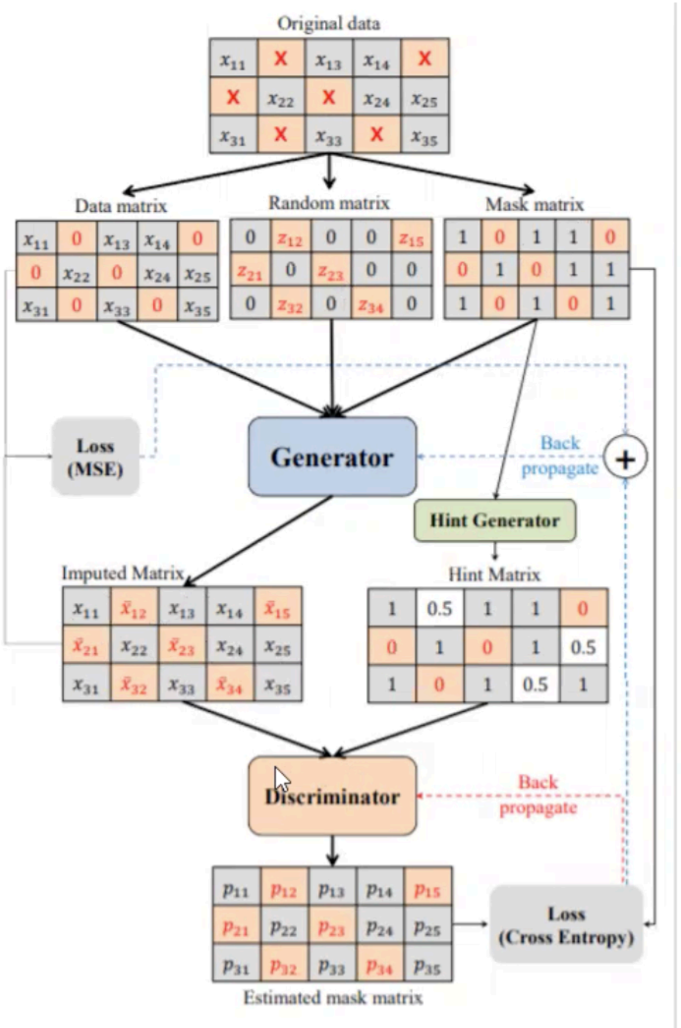


Figure 2: Overview of data synthesis using GAN. (1) It transforms each record in a relational table into a sample $t \in \mathbb{R}^d$. (2) It takes the samples as input to train a deep generative model G using the adversarial training framework in GAN. (3) It utilizes the trained G to generate a set of synthetic samples, which are then transformed back into fake records.

Missing data imputation

- Through masking data points in original dataset



Original data

x_{11}	X	x_{13}	x_{14}	X
X	x_{22}	X	x_{24}	x_{25}
x_{31}	X	x_{33}	X	x_{35}

Data matrix

x_{11}	0	x_{13}	x_{14}	0
0	x_{22}	0	x_{24}	x_{25}
x_{31}	0	x_{33}	0	x_{35}

Random matrix

0	z_{12}	0	0	z_{15}
z_{21}	0	z_{23}	0	0
0	z_{32}	0	z_{34}	0

Mask matrix

1	0	1	1	0
0	1	0	1	1
1	0	1	0	1

Imputed Matrix

x_{11}	$\bar{x}_{12}$	x_{13}	x_{14}	$\bar{x}_{15}$
$\bar{x}_{21}$	x_{22}	$\bar{x}_{23}$	x_{24}	x_{25}
x_{31}	$\bar{x}_{32}$	x_{33}	$\bar{x}_{34}$	x_{35}

Hint Matrix

1	0.5	1	1	0
0	1	0	1	0.5
1	0	1	0.5	1

Estimated mask matrix

p_{11}	p_{12}	p_{13}	p_{14}	p_{15}
p_{21}	p_{22}	p_{23}	p_{24}	p_{25}
p_{31}	p_{32}	p_{33}	p_{34}	p_{35}

Generate data based on an objective

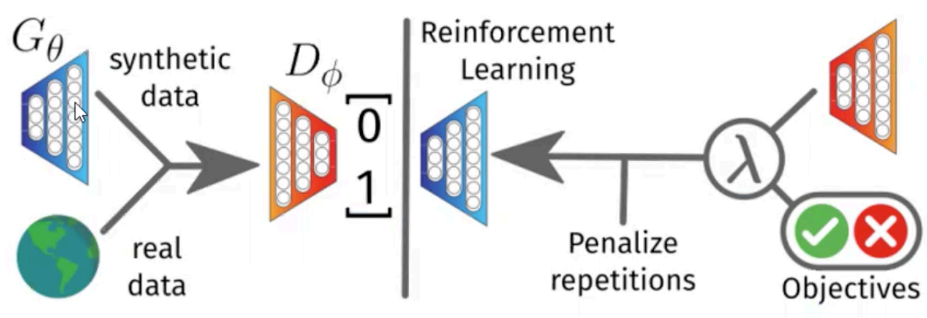


Figure 1: Schema for ORGAN. *Left*: D is trained as a classifier receiving as input a mix of real data and generated data by G. *Right*: G is trained by RL where the reward is a combination of D and the objectives, and is passed back to the policy function via Monte Carlo sampling. We penalize non-unique sequences.

Future directions

1. *How to evaluate tabular GANs?*

   Sample similarity
    - Basic statistics
    - Column correlations
    - Mirror column associations
    - PCA variance correlation

   Machine learning efficacy
    - Evaluate how downstream ML model performance varies with real data, synthetic data, real data + synthetic

   Privacy evaluations
    - Check how many samples are replicated from the training (real) dataset

   Human evaluations
    - On the lines of Turing test - to see whether humans can distinguish between real and synthetic data
    - In domain specific implementations, this will help with understanding which patterns GANs are able to learn

2. *Open challenges*
    - Data hungry - Adaptive Discriminator Augmentation
    - Mode collapse for imbalanced categorical data
    - Domain specific data still requires significant work e.g. HER data is still not realistic