Matching-based Data Valuation for Generative Model

Jiaxi Yang[1,2], Wenlong Deng[2], Benlin Liu[3], Yangsibo Huang[4], Xiaoxiao Li[2]

[1]University of Electronic Science and Technology of China, [2]University of British Columbia, [3]University of Washington, [4]Princeton University

Try Our Demo

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

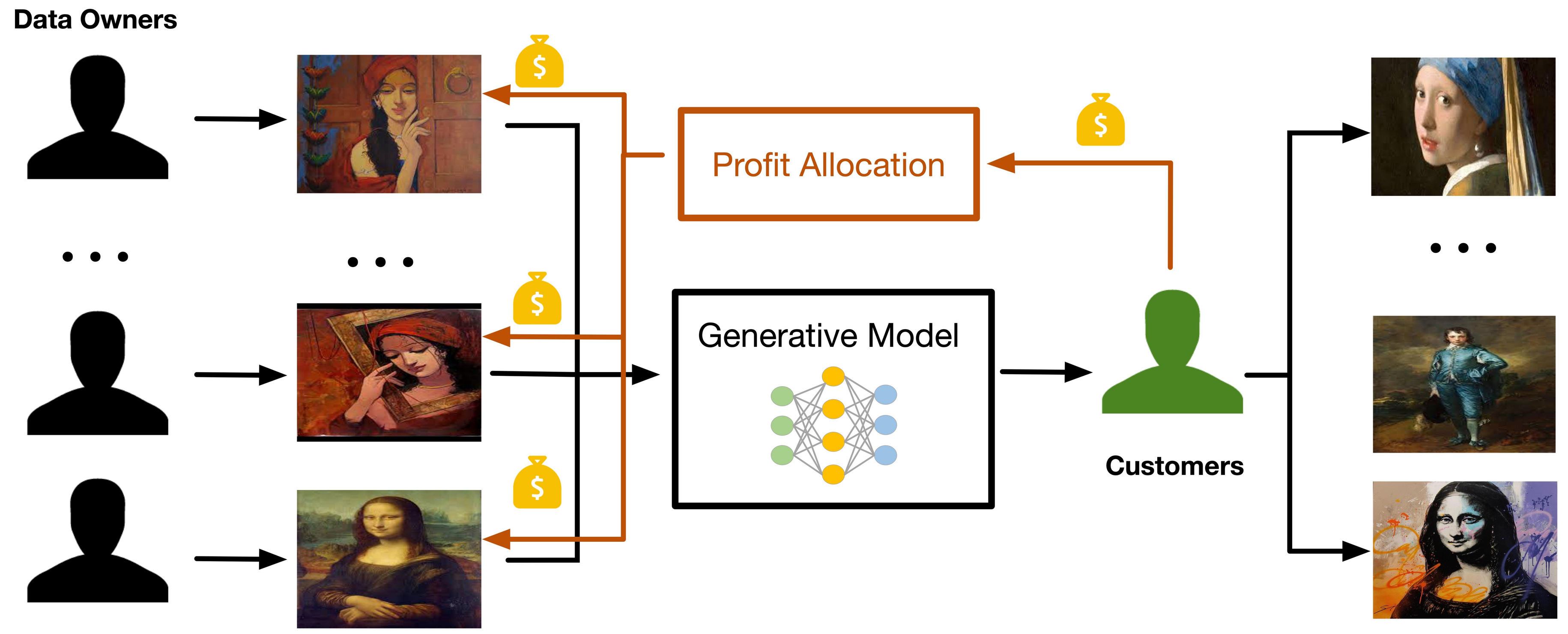
Motivation

Background:

- Training process of generative model requires loads of training data.
- Training data (e.g. art paintings) is considered both private property and subject to copyright protection by its owners.

Goals:

- Evaluating the value of training data for generated images, thus the **credits** could be distributed to data owners fairly.

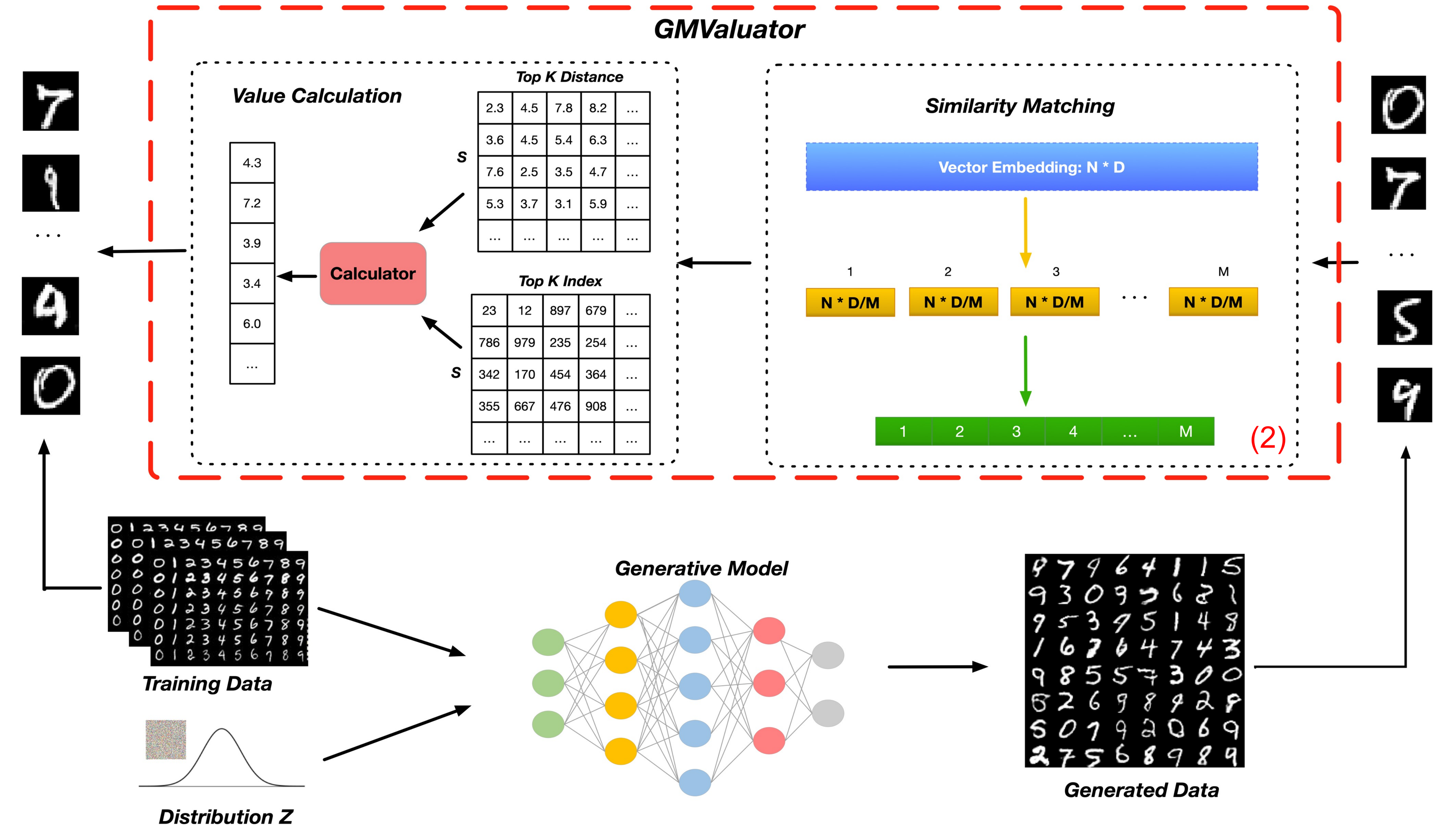


Challenges

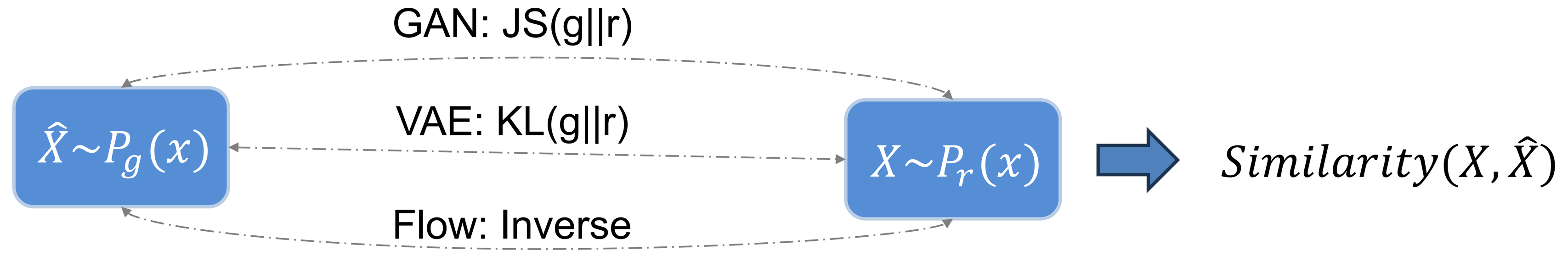
- ***Lack of metrics to determine data valuation (1)***
  - ✓ Generative model aims to capture the underlying distribution of data, which have **no explicit labels** or ground-truth **for valuation**.
  - ✓ Evaluation metrics (e.g. FID, IS) are not objective to determine the value of generated data.
- ***Expensive Computational Cost. (2)***
  - ✓ **Generative** models have **complex** architectures with a **large** number of parameters, which requires substantial computational costs.
  - ✓ **Existing methods** (Shapley Value, Banzhaf et.) is impractical for generative model due to the **need** for **retraining** and high computational cost.

Methodology

GMValuator Architecture:



(1) *Similarity & Valuation:*



We calculate similarity between training and generated data to determine data valuation as generative model optimize the distribution similarity.

(2) *Image Similarity Matching:*

- To fast calculate the similarity, we first utilize *Product Quantization* (PQ) to compress the training images.
- Given a generated image, top k most similar training images, which are regarded as contributors, are recalled by Wasserstein distance.

Value Calculation:

- For the data valuation for generative model, we propose *GMValuator*, which is the first method to address this problem.
- For one generated data sample, the utility function of each training sample can be denoted as:

$$\mathcal{V}(x_i,\hat{x}_j)=\begin{cases}\frac{\exp(-d_i)}{\sum_{i=1}^{K}\exp(-d_i)} & x_i\in\mathcal{P}\\ 0 & x_i\notin\mathcal{P}\end{cases}$$

$$s.t.\quad d_1,d_2,...,d_k>0$$

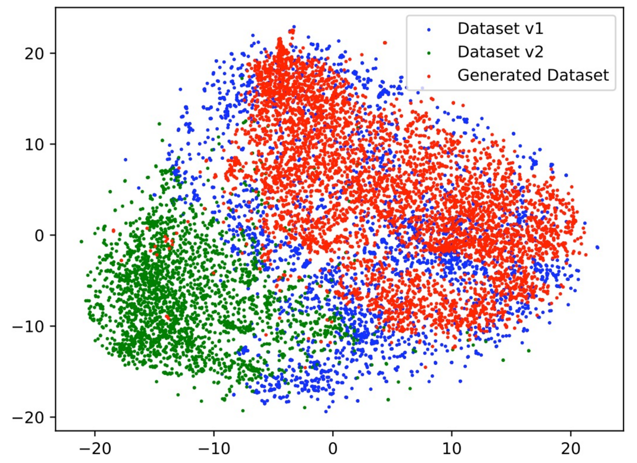
- With the fixed generated dataset, the value of a training data sample can be calculated by summing up all its utility values for each training data sample, which can be presented as follows:

$$\phi(x_i,\hat{X})=\sum_{j}^{m}\gamma_{i,j}\mathcal{V}(x_i,\hat{x}_j), i\in X, j\in\hat{X}$$

Experiments

RA1: The generated data should exhibit a higher degree of similarity to the data points used to train the generator than the ones not used for training, despite originating from the same distribution. (Figure 1)

RA2: The data points (used for training) are expected to have a higher value than data points that are not used for training. (Table 1)



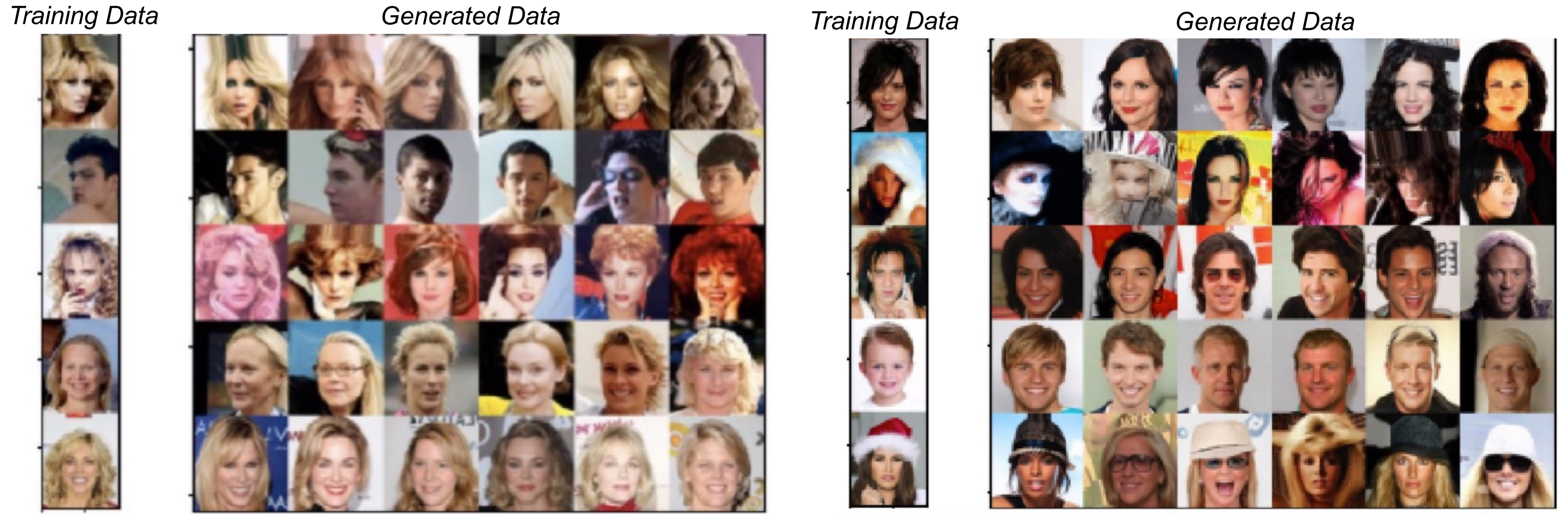
RA 1

H_0: $\phi(D_i,S,\mu_i)>\phi(D_j,S,\mu_i)$
H_1: $\phi(X_i,S,\mu_i)\le\phi(X_j,S,\mu_i), i\in X_{v1}, j\in X_{v2}$

	BigGAN	Classifier-free Guidance Diffusion
Average value (v1)	1.632352	0.369565
Average value (v2)	0.319654	0.030434
P-value	6.937027×10^{-68}	8.053195×10^{-55}
T-statistic	17.924512	15.947860
Significance level	0.01	0.01
Result	p-value less than 0.01, reject H_0, value of v2 less than v1 averagely	p-value less than 0.01, reject H_0, value of v2 less than v1 averagely

RA 2

RA3: Training data points that contribute more to the generated data (high values) are expected to exhibit greater semantic alignment to the fixed-generated dataset.



High Value ***Low Value***