Bridging the Gap: Enhancing the Utility of Synthetic Data via Post-Processing Techniques

Andrea Lampis, Eugenio Lomurno, Matteo Matteucci

POLITECNICO MILANO 1863 — AIRLAB — ARTIFICIAL INTELLIGENCE AND ROBOTICS LAB

BMVC 2023

Background

Deep **generative models** have become so powerful that they are able to produce high-quality samples that are almost indistinguishable from real ones.
While most work focuses on improving this quality, this research focuses on maximising their **utility** in the context of training a **downstream** machine learning model.

Contributions

- Two improved post-processing techniques, namely **Dynamic Sample Filtering** and **Dynamic Dataset Recycle**, and a novel method called **Expansion Trick.**
- The **GaFi pipeline**, which consists of a set of **post-processing techniques** suitable for any generative model to **maximise the CAS** achieved with its generated data.
- Empirical CAS results that approach the upper bound of real accuracy performance, setting a new **state of the art** in the generation of synthetic data for classification tasks.

Pipeline Steps

1) Generative Model Training: the initial step of the pipeline entails training a generative model and saving its **checkpoints** after **every epoch** for subsequent use.

2) Checkpoint Optimisation: the checkpoints are sampled and evaluated with respect to the **Classification Accuracy Score (CAS)** rather than to IS or FID metrics. The sampling hyperparameters are kept fixed and **Dataset Recycle** technique is used with N=10, i.e. the synthetic dataset is renewed every 10 epochs. During this step, the checkpoint with the best CAS is selected.

3) Stddev Optimisation: the optimal sampling standard deviation is tuned to find the best input **noise distribution**, which corresponds to the application of the **Expansion Trick**.

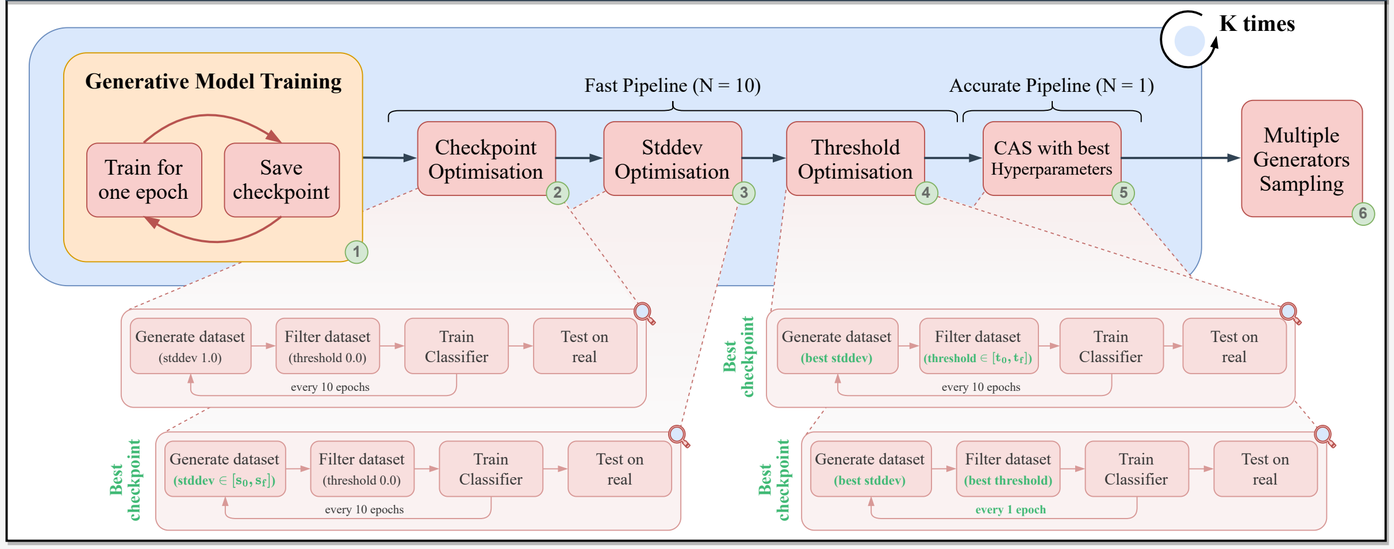
4) Threshold Optimisation: the optimal filtering threshold is found through the **Dynamic Sample Filtering** technique.

5) CAS with best hyperparameters: with the optimal hyperparameters selected so far, the **Dynamic Dataset Recycle** technique is reapplied with N=1. The single optimal generative model is found in this step.

	Checkpoint	Standard Deviation	Filtering Threshold	CAS
Fashion-MNIST	112	2.00	0.0	**94.03%**
CIFAR-10	460	1.60	0.3	**92.60%**
CIFAR-100	443	1.70	0.1	**68.92%**
CINIC-10	490	1.25	0.0	**84.37%**
DermaMNIST	80	1.30	0.4	**73.66%**

6) Multiple Generators Sampling: multiple generative models are built following the previous steps and then used to sample the data and train a single optimal classifier. This is achieved by repeating all the previous steps of the pipeline K=6 times.

Pipeline Architecture



Results

#Generators		Fashion-MNIST	CIFAR-10	CIFAR-100	CINIC-10	DermaMNIST
	Real Data	96.01%	94.98%	75.64%	89.05%	77.25%
	Baseline	88.70%	87.11%	57.74%	75.58%	67.48%
1	Dat *et al.*	-	88.25%	62.22%	-	-
	GaFi (ours)	94.03%	92.60% (+4.35%)	68.92% (+6.70%)	84.37%	73.66%
2	Dat *et al.*	-	89.68%	64.33%	-	-
	GaFi (ours)	93.98%	92.74% (+3.06%)	70.22% (+5.89%)	85.42%	75.06%
4	Dat *et al.*	-	90.68%	67.22%	-	-
	GaFi (ours)	93.99%	93.02% (+2.34%)	71.75% (+4.53%)	85.62%	74.71%
6	Dat *et al.*	-	91.14%	67.56%	-	-
	GaFi (ours)	93.98%	93.20% (+2.06%)	71.95% (+4.39%)	85.72%	75.21%