Provably Improving Generalization of Few-shot Models with Synthetic Data

Lan-Cuong Nguyen, Quan Nguyen-Tri, Bang Tran Khanh,
Dung D.Le, Long Tran-Thanh, Khoat Than

Background and Motivation

Problems:

- Training with synthetic data faces ***performance degradation*** due to ***distribution gap*** between real and synthetic data.
- Recent methods narrowed the distribution gap but are ***heuristic-driven, lacking theoretical guarantees***.

Research questions:

- What properties can ***indicate the goodness*** of a synthetic dataset?
- How to ***generate*** a good synthetic dataset?
- How to efficiently ***train a predictor*** from a training set of both real and synthetic samples?
- How can the quality of a generator ***affect the generalization ability*** of the trained predictor?

Contributions

ICML
International Conference
On Machine Learning

1. ***Theory:*** Two novel generalization bounds shows that for good generalization, synthetic data must be both similar to real samples and diverse enough to ensure local robustness.
2. ***Methodology:*** A novel loss function and training paradigm, guided by theoretical bounds, to jointly optimize data partitioning and model training for minimizing generalization errors.
3. ***Empirical Validation:*** Our method consistently outperforms state-of-the-art few-shot image classification methods on multiple datasets when using synthetic data.

Definitions

S *and* G are real and synthetic datasets sampled from real and synthetic distribution,respectively. h is a model.

Model-based discrepancy:

$$\bar{d}_h(\boldsymbol{G}, \boldsymbol{S}) = \frac{1}{|\boldsymbol{G}|.|\boldsymbol{S}|} \sum_{\boldsymbol{u} \in \boldsymbol{G}, \boldsymbol{s} \in \boldsymbol{S}} \|\boldsymbol{h}(\boldsymbol{s}) - \boldsymbol{h}(\boldsymbol{u})\|$$

Local robustness in the area A:

$$\mathcal{R}_h(\boldsymbol{s}, \mathcal{A}|P) = \mathbb{E}_{\boldsymbol{z} \sim P}[\|\boldsymbol{h}(\boldsymbol{z}) - \boldsymbol{h}(\boldsymbol{s})\| : \boldsymbol{z} \in \mathcal{A}].$$

Theoretical Analysis

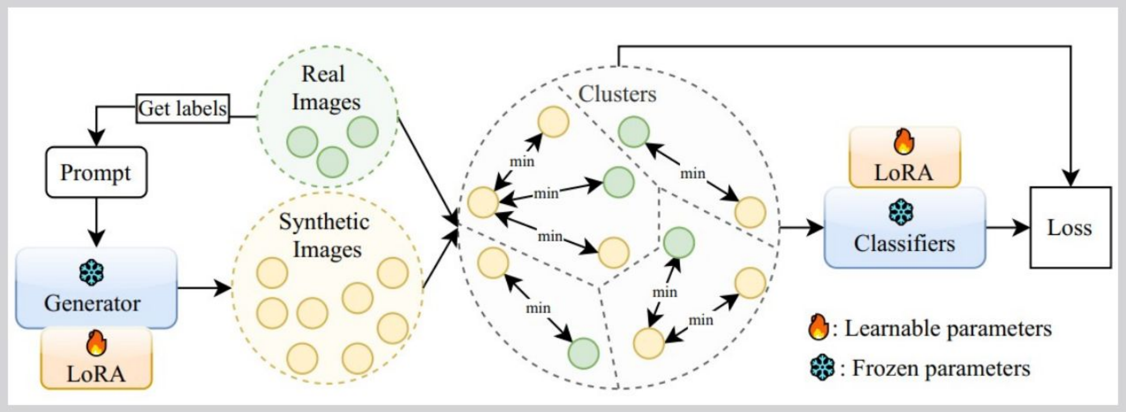
Generalization Bounds:

$$F(P_0, \boldsymbol{h}) \leq L_h \sum_{i \in \boldsymbol{T}_S} \frac{g_i}{g} \left[\bar{d}_h(\boldsymbol{G}_i, \boldsymbol{S}_i) + \mathcal{R}_h(\boldsymbol{G}_i, \mathcal{Z}_i \mid P_g) \right] + A$$

Asymptotic Cases:

$$F(P_0, \boldsymbol{h}) \leq L_h \sum_{i \in \boldsymbol{T}_S} \left[p_i^g \mathcal{R}_h(\boldsymbol{S}_i, \mathcal{Z}_i | P_g) + \frac{n_i}{n} \mathcal{R}_h(\boldsymbol{S}_i, \mathcal{Z}_i | P_0) \right] + A_1$$

Methodology



Overall algorithm pipeline

$$\mathcal{L} = \lambda F(\boldsymbol{S}, \boldsymbol{h}) + F(\boldsymbol{G}, \boldsymbol{h}) + \lambda_1 \sum_{i \in \boldsymbol{T}_S} \sum_{s \in \boldsymbol{S}_i, g \in \boldsymbol{G}_i} \frac{g_i}{g} \frac{1}{|\boldsymbol{G}_i||\boldsymbol{S}_i|} \|\boldsymbol{h}(s) - \boldsymbol{h}(g)\| + \lambda_2 \frac{1}{g} \sum_{i \in \boldsymbol{T}_S} \sum_{g_1, g_2 \in \boldsymbol{G}_i} \frac{1}{g_i} \|\boldsymbol{h}((g_1) - \boldsymbol{h}(g_2)\|$$

Loss function

Algorithm 1 Fine-tuning few-shot models with synthetic data

Input: Real dataset $\boldsymbol{S}$, number g of synthesis samples, (conditional) Pretrained generator models $\mathcal{G}$

1: Initialize centroids $\boldsymbol{z}$ for every local area
2: Fine-tuning generator $\mathcal{G}$ by real dataset $\boldsymbol{S}$ with LoRA
3: Generate g synthetic images from generator $\mathcal{G}$
4: Use K-means clustering on both real and synthetic images to obtain partition $\Gamma(\mathcal{Z})$
5:
6: **for** each mini-batch A **do**
7: Assign datapoints to their nearest clusters
8: Train the model $\boldsymbol{h}$ using the loss function $\mathcal{L}$ on the combined dataset $\boldsymbol{S}_A \cup \boldsymbol{G}_A$ that includes both real data and synthetic data. ▷ Refer to equation 7.
9: **end for**

Experiments Results

Method	R	S	IN	CAL	DTD	EuSAT	AirC	Pets	Cars	SUN	Food	FLO	Avg
CLIP (zero-shot)			70.2	96.1	46.1	38.1	23.8	91.0	63.1	72.2	85.1	71.8	64.1
Real-finetune	✓		73.4	96.8	73.9	93.5	59.3	94.0	87.5	77.1	87.6	98.7	84.2
IsSynth	✓	✓	73.9	97.4	75.1	93.9	64.8	92.1	88.5	**77.7**	86.0	99.0	84.8
DISEF	✓	✓	73.8	97.0	74.3	94.0	64.3	92.6	87.9	77.6	86.2	99.0	84.7
DataDream$_{cls}$	✓	✓	73.8	97.6	73.1	93.8	68.3	94.5	91.2	77.5	87.5	99.4	85.7
DataDream$_{dset}$	✓	✓	**74.1**	96.9	74.1	93.4	72.3	**94.8**	92.4	77.5	87.6	**99.4**	86.3
Ours (lightweight)	✓	✓	73.7	**97.9**	**75.5**	94.2	71.5	94.5	90.2	77.6	90.0	99.0	86.4
Ours (full)	✓	✓	73.8	97.3	74.5	**94.7**	**74.3**	94.6	**93.1**	**77.7**	**90.4**	99.3	**87.0**

Main experiment results of 16-shot fine-tuning settings

Ablation Studies

ICML
International Conference
On Machine Learning

Table 2. Ablation of the loss function components.

Discre.	Rob.	EuroSAT	DTD	AirC	Cars
		93.5	74.1	72.5	92.6
	✓	94.6	74.4	73.1	**93.1**
✓		94.3	74.3	**74.8**	93.0
✓	✓	**94.7**	**74.5**	74.3	**93.1**

Table 3. Methods performance on CLIP-Resnet50.

Methods	AirC	Cars	Food	CAL
Real fine-tune	61.57	78.86	63.52	93.29
IsSynth	70.94	90.82	68.77	94.54
DISEF	65.99	79.18	70.10	94.34
DataDream$_{cls}$	79.21	92.99	66.70	94.37
DataDream$_{dset}$	81.46	93.30	66.63	**94.62**
Ours	**82.67**	**93.71**	**70.35**	94.17

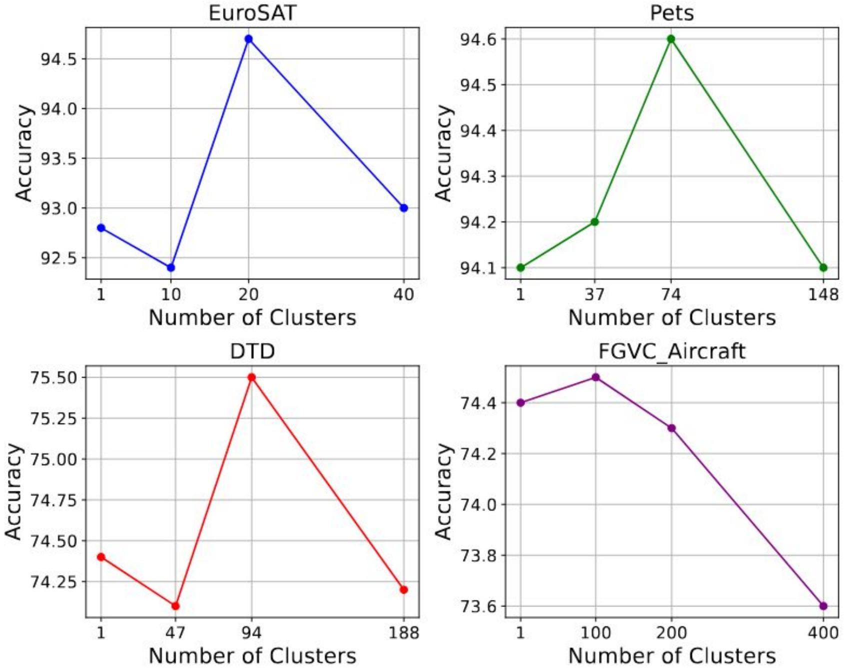


Figure 3. Results with increasing number of clusters on 4 datasets

THANK YOU!

Paper QR: