ROME is Forged in Adversity: RObust Distilled Datasets via InforMation BottlenEck

Zheng Zhou[1], Wenquan Feng[1], Qiaosheng Zhang[2,3], Shuchang Lyu[1,*], Qi Zhao[1], Guangliang Cheng[4]

[1]Beihang University, [2]Shanghai Artificial Intelligence Laboratory, [3]Shanghai Innovation Institute, [4]University of Liverpool (*Corresponding Author)

ICML International Conference On Machine Learning

Background & Motivation

✦What is Dataset Distillation?

- ❖ **Dataset distillation** compresses large datasets into compact synthetic subsets, significantly reducing training time and computation while maintaining model performance.
- ❖ Most dataset distillation methods are efficient but vulnerable to adversarial attacks, limiting their reliability in safety-critical areas like face recognition, autonomous driving, and object detection.

✦How to enhance the robustness of models?

- ❖ Adversarial robustness is a key research focus. A common way to improve it is adversarial training, but this method is costly and hard to apply in data-efficient settings like dataset distillation.

✦Existing challenges

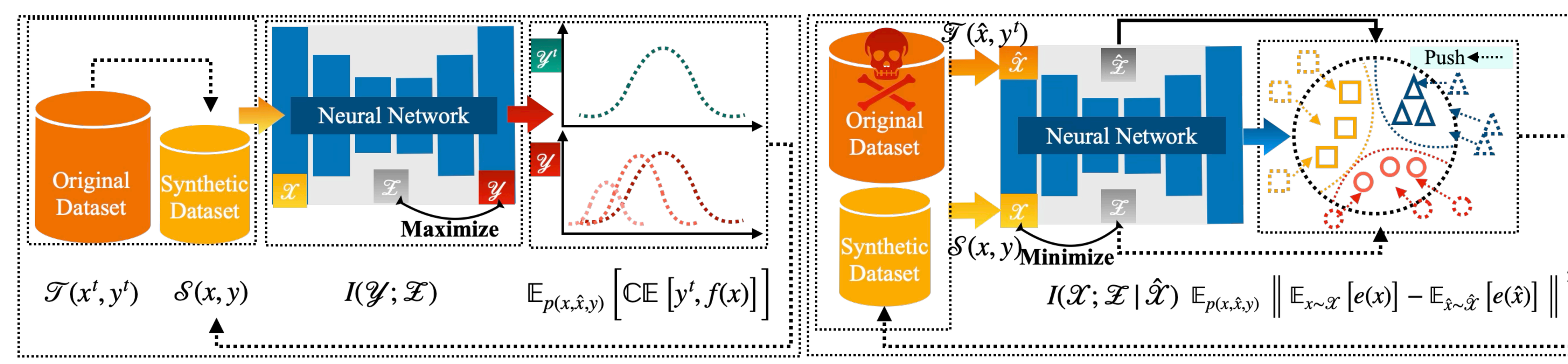
- ❖ **High retraining cost**, making the process computationally expensive.
- ❖ **Robustness–accuracy trade-off**, where improving adversarial robustness often reduces clean accuracy.

✦Contributions

- ❖ We propose ROME, which applies the information bottleneck to dataset distillation and incorporates adversarial perturbations to create robust distilled datasets.
- ❖ We present two training terms: a performance-aligned term that preserves accuracy and a robustness-aligned term that enhances adversarial robustness.
- ❖ We introduce I-RR, a refined metric for dataset distillation robustness. Experiments on CIFAR-10 and CIFAR-100 show our method outperforms others in both white-box and black-box attacks.

Method

✦Overview of ROME



(a) Performance-aligned Term (b) Robustness-aligned Term

Figure 1: The framework of ROME.

✦Formulating ROME via information bottleneck

$$\text{ROME} = I(\mathcal{Y};\mathcal{Z}) - \beta I(\mathcal{X};\mathcal{Z}\,|\,\hat{\mathcal{X}})$$
$$\geq \mathbb{E}_{p(x,\hat{x},y)p(z|x,\hat{x},y)}\left[\log q(y\,|\,z) - \beta \log \frac{p(z\,|\,x)}{q(z\,|\,\hat{x})}\right]$$

✦Performance-aligned term

$$\mathcal{L}_{\text{Perf\_Alig}} = \mathbb{E}_{p(x,\hat{x},y)p(z|x,\hat{x},y)}\left[\log q(y\,|\,z)\right]$$
$$= \mathbb{E}_{p(x,\hat{x},y)}\left[\mathbb{CE}\left[y^t, f(x)\right]\right]$$

✦Robustness-aligned term

$$\mathcal{L}_{\text{Rob\_Alig}} = \mathbb{E}_{p(x,\hat{x},y)p(z|x,\hat{x},y)}\left[\beta \log \frac{p(z\,|\,x)}{q(z\,|\,\hat{x})}\right]$$
$$= \mathbb{E}_{p(x,\hat{x},y)}\left\| \mathbb{E}_{x\sim\mathcal{X}}\left[e(x)\right] - \mathbb{E}_{\hat{x}\sim\hat{\mathcal{X}}}\left[e(\hat{x})\right] \right\|^2$$

✦Monte Carlo Approximation

$$\mathcal{L}_{\text{Perf\_Alig}} = \sum_{c=0}^{\mathcal{C}-1} \frac{1}{|\mathcal{X}_c|} \sum_{x\in\mathcal{X}_c} \mathbb{CE}\left[y_c^t, f(x)\right]$$

$$\mathcal{L}_{\text{Rob\_Alig}} = \sum_{c=0}^{\mathcal{C}-1} \left\| \frac{1}{|\mathcal{X}_c|} \sum_{x\in\mathcal{X}_c} e(x) - \frac{1}{|\hat{\mathcal{X}}_c|} \sum_{\hat{x}\in\hat{\mathcal{X}}_c} e(\hat{x}) \right\|^2$$

✦Training Objective Term

$$\mathcal{L}_{\text{TOTAL}} = (1-\alpha)\mathcal{L}_{\text{Perf\_Alig}} + \alpha\mathcal{L}_{\text{Rob\_Alig}}$$

Results

✦Experimental Results

Dataset	Method	Targeted Attack				Untargeted Attack			
		RR	CREI	I-RR	I-CREI	RR	CREI	I-RR	I-CREI
CIFAR-10	Full-size	20.42%	24.98%	67.24%	48.39%	28.33%	25.12%	28.82%	25.36%
	DC 2020	30.79%	29.35%	88.51%	58.21%	31.87%	26.70%	56.02%	38.78%
	DSA 2021	45.22%	36.43%	86.81%	57.22%	36.53%	27.75%	53.66%	36.32%
	MTT 2022	36.00%	32.26%	83.95%	56.24%	33.30%	26.26%	48.34%	33.77%
	DM 2023	46.01%	36.01%	85.76%	55.89%	34.50%	28.32%	56.19%	39.16%
	IDM 2023	32.35%	27.75%	87.07%	55.11%	33.03%	28.46%	53.43%	38.66%
	BACON 2024	36.83%	33.05%	84.37%	56.82%	32.87%	27.20%	50.49%	36.01%
	ROME	**81.36%** (35.35 ↑)	**55.28%** (18.85 ↑)	**97.44%** (8.93 ↑)	**63.32%** (5.11 ↑)	**49.86%** (13.33 ↑)	**35.05%** (6.59 ↑)	**67.01%** (10.82 ↑)	**43.62%** (4.46 ↑)
CIFAR-100	Full-size	6.77%	18.18%	65.50%	47.55%	19.91%	18.60%	20.08%	18.69%
	DC 2020	33.11%	30.31%	77.14%	52.32%	28.74%	22.40%	32.33%	24.19%
	DSA 2021	43.97%	35.01%	72.97%	49.51%	28.53%	20.40%	33.29%	22.77%
	MTT 2022	36.06%	31.16%	74.54%	50.40%	26.07%	19.65%	31.10%	22.17%
	DM 2023	39.32%	31.32%	71.29%	47.30%	26.72%	19.78%	29.74%	21.28%
	IDM 2023	34.44%	27.16%	74.57%	47.23%	26.28%	20.36%	30.83%	22.63%
	BACON 2024	31.81%	29.78%	69.96%	48.86%	25.26%	19.30%	27.42%	20.38%
	ROME	**103.09%** (59.12 ↑)	**66.18%** (31.17 ↑)	**100.65%** (23.51 ↑)	**64.96%** (12.64 ↑)	**44.10%** (15.36 ↑)	**28.29%** (5.89 ↑)	**46.24%** (12.95 ↑)	**29.36%** (5.17 ↑)

Table 1: Robustness of models trained on distilled datasets under white-box attacks.

Method	Targeted Attack		Untargeted Attack	
	Transfer	Query	Transfer	Query
DC	85.84%	88.71%	83.97%	43.81%
DSA	94.09%	94.95%	92.31%	54.60%
MTT	91.40%	92.76%	89.02%	48.71%
DM	92.22%	93.86%	90.36%	57.53%
IDM	92.17%	94.37%	89.22%	63.23%
BACON	92.46%	94.67%	89.25%	63.26%
ROME	**99.90%** (5.81 ↑)	**99.79%** (4.84 ↑)	**98.44%** (6.13 ↑)	**78.46%** (15.2 ↑)

Table 2: Robustness of models trained on distilled datasets under black-box attacks.

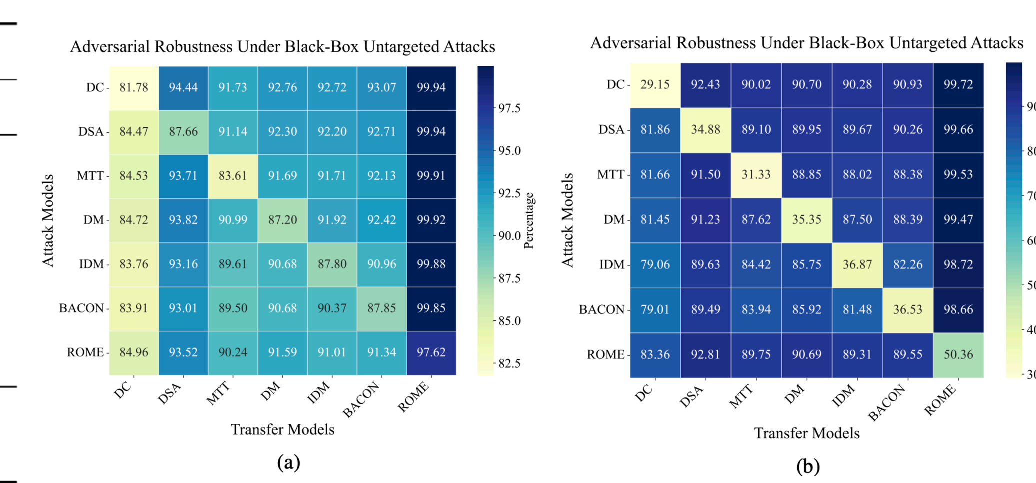


Figure 2: Robustness heatmaps of models trained on distilled datasets against black-box attacks.

More Information

✦Scan the QR codes for more information

Code · Project Page · Contact us