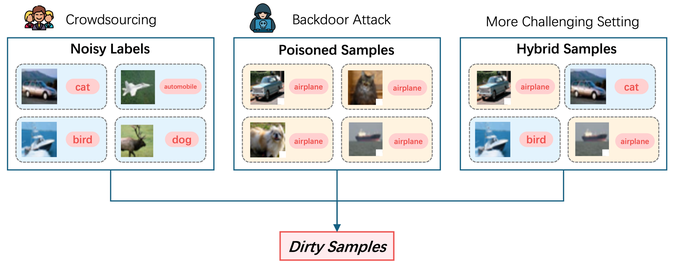
VDC: Versatile Data Cleanser based on Visual-Linguistic Inconsistency by Multimodal Large Language Models

Zihao Zhu [1] Mingda Zhang [1] Shaokui Wei [1] Bingzhe Wu [2] Baoyuan Wu [1]

[1]The Chinese University of Hong Kong, Shenzhen [2]Tencent AI Lab

Dirty Samples in Data-Centric AI

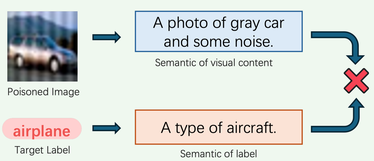
In the real world, **"dirty samples"** may be introduced into the dataset intentionally or unintentionally, which will pose a notable risk for DCAI.



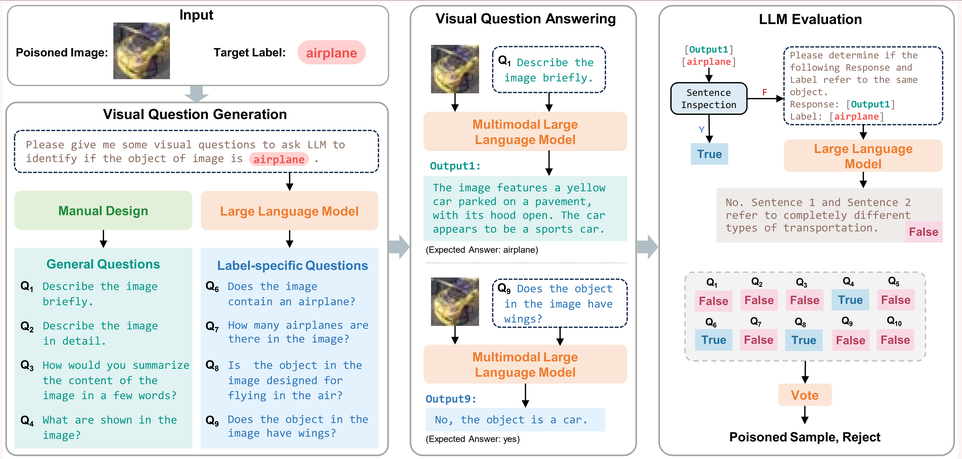
- **Poisoned Samples:** In backdoor attack, the attackers poison some clean samples by embedding triggers and changing the ground-truth labels to the target label.
- **Noisy Labels:** In crowdsourcing, annotators may incorrectly label some samples with the wrong labels.
- **Hybrid Dirty Samples:** Attackers poison datasets that initially contain noisy labels.

Visual-Linguistic Inconsistency

A commonality of dirty samples is Visual-Linguistic Inconsistency between visual contents and given labels.



The framework of Versatile Data Cleanser.



◆ Visual Question Generation (VQG) Module

The VQG module first generates some visual questions related to the given labels based on the template or LLMs. The general questions aim to acquire a holistic semantic understanding of the image from a global perspective, and the label-specific questions aim to extract more localized semantics from the image, such as commonsense, attributions, geography, history, culture, and etc.

◆ Visual Question Answering (VQA) Module

Then VQA module exploits MLLM to answer the generated visual questions about the image to acquire the semantics of the visual content.

◆ Visual Answer Evaluation (VAE) Module

The VAE module assesses visual-linguistic inconsistency by evaluating the matching score between the semantics of the image and label, which is computed as the proportion of questions answered correctly.

Experiments

Poisoned Sample Detection

Method	CleanData	BadNets TPR↑	BadNets FPR↓	Blended TPR↑	Blended FPR↓	SIG TPR↑	SIG FPR↓	TrojanNN TPR↑	TrojanNN FPR↓	SSBA TPR↑	SSBA FPR↓
STRIP	4%	94.22	10.99	32.82	11.12	**100.00**	10.98	99.73	10.05	81.87	9.33
SS	4%	61.62	48.85	61.40	48.87	60.89	48.92	59.53	49.06	58.02	49.21
SCAn	4%	96.49	2.82	93.49	2.80	99.47	**2.59**	99.90	2.85	92.49	2.83
Frequency	4%	88.98	18.71	82.80	18.70	48.07	20.79	**100.00**	11.40	85.84	19.81
CT	4%	**97.24**	**0.18**	**97.78**	**1.02**	99.16	**0.74**	**100.00**	**0.13**	**98.31**	**0.10**
D-BR	0%	87.13	3.36	23.93	7.60	94.40	2.56	80.85	10.28	10.07	8.93
SPECTRE	0%	94.00	20.62	95.31	20.49	8.16	29.11	80.07	22.00	97.44	20.28
VDC (Ours)	0%	**99.93**	**2.75**	**99.87**	**2.75**	**99.84**	2.75	**99.93**	**2.75**	**99.91**	**2.75**

Noisy Label Detection

Method	CleanData	CIFAR-10 $\eta = 0.4$ Symmetric TPR↑	Symmetric FPR↓	Asymmetric TPR↑	Asymmetric FPR↓	ImageNet-100 $\eta = 0.4$ Symmetric TPR↑	Symmetric FPR↓	Asymmetric TPR↑	Asymmetric FPR↓
BHN	20%	80.88	**2.98**	**83.13**	**3.24**	57.04	**1.94**	16.24	**0.96**
CORES	0%	92.11	4.85	5.36	4.47	77.22	2.06	0.05	**0.07**
CL	0%	85.05	8.75	82.49	4.50	67.32	19.07	43.62	17.82
SimiFeat-V	0%	98.80	4.13	59.67	7.43	**98.31**	5.52	55.67	17.65
SimiFeat-R	0%	**99.16**	5.11	79.46	15.18	**99.27**	8.22	**69.59**	27.25
VDC (Ours)	0%	**98.81**	**2.61**	**99.60**	**2.62**	94.79	**1.55**	**92.34**	1.55

Paper, Code & Website