# Detecting Backdoor Samples in Contrastive Language Image Pretraining

Hanxun Huang[1] Sarah Erfani[1], Yige Li[2], Xingjun Ma[3], James Bailey[1]

[1]The University of Melbourne [2]Singapore Management University [3]Fudan University

Paper Website Code

## Overview

- CLIP is highly vulnerable to **poisoning backdoor attacks**, where an adversary can compromise the model by poisoning as little as **0.01%** of the training data.
- Given the web-scale nature of CLIP's training data, **acquiring 0.01%** of the dataset is feasible—an attacker could achieve this by purchasing expiring domains for as little as **$10 USD**.

## Contributions

- We present a systematic study on the detectability of poisoning backdoor attacks on CLIP, and show that existing detection methods designed for supervised learning can fail on CLIP.
- We reveal one major weakness of CLIP backdoor samples related to the **sparsity of their representation local neighbourhood**, which facilitates **highly accurate** and **efficient detection** using efficient local density-based detectors. With these detectors, one can clean up a million-scale poisoned dataset (e.g., CC3M) within 15 minutes using 4 Nvidia-A100 GPUs.
- Our experiments in the clean setting reveal that there exist **unintentional (natural) backdoors** in the CC3M dataset, which has been injected into a popular open-source model released by OpenCLIP.

## Local Outlier Detection

The Simplified Local Outlier Factor (SLOF) provides a simplified version of Local Outlier Factor (LOF) using the distance to the $k$-th nearest neighbour, defined as the following:

$$\mathrm{SLOF}_k(\boldsymbol{q}) \triangleq \frac{1}{k} \sum_{\boldsymbol{o} \in \mathrm{NN}_k(\boldsymbol{q})} \frac{k\text{-dist}(\boldsymbol{q})}{k\text{-dist}(\boldsymbol{o})}.$$

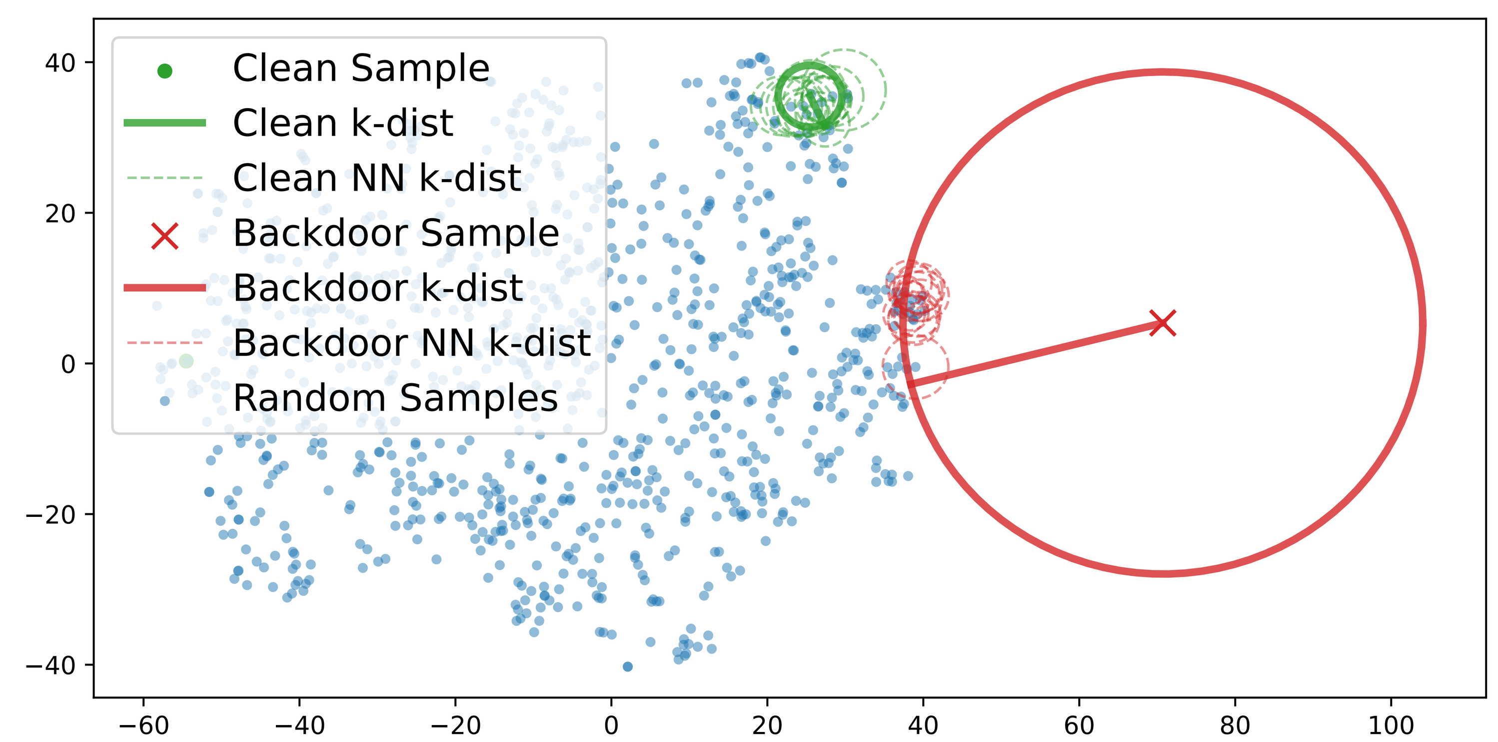
Local Intrinsic Dimensionality (LID) describes the rate of growth in the number of data objects encountered as the distance from the reference sample increases:

$$\widehat{\mathrm{LID}}^*_{F_{\boldsymbol{q}}} = \left( -\frac{1}{k} \sum_{\boldsymbol{o} \in \mathrm{NN}_k(\boldsymbol{q})} \log \frac{\mathrm{dist}(\boldsymbol{q}, \boldsymbol{o})}{k\text{-dist}(\boldsymbol{q})} \right)^{-1}.$$

Dimensionality-Aware Outlier Detection (DAO) is a criterion that extends LOF and SLOF using theory in dimensionality characteristics:

$$\mathrm{DAO}_k(\boldsymbol{q}) \triangleq \frac{1}{k} \sum_{\boldsymbol{o} \in \mathrm{NN}_k(\boldsymbol{q})} \left( \frac{k\text{-dist}(\boldsymbol{q})}{k\text{-dist}(\boldsymbol{o})} \right)^{\widehat{\mathrm{LID}}^*_{F_{\boldsymbol{o}}}}.$$

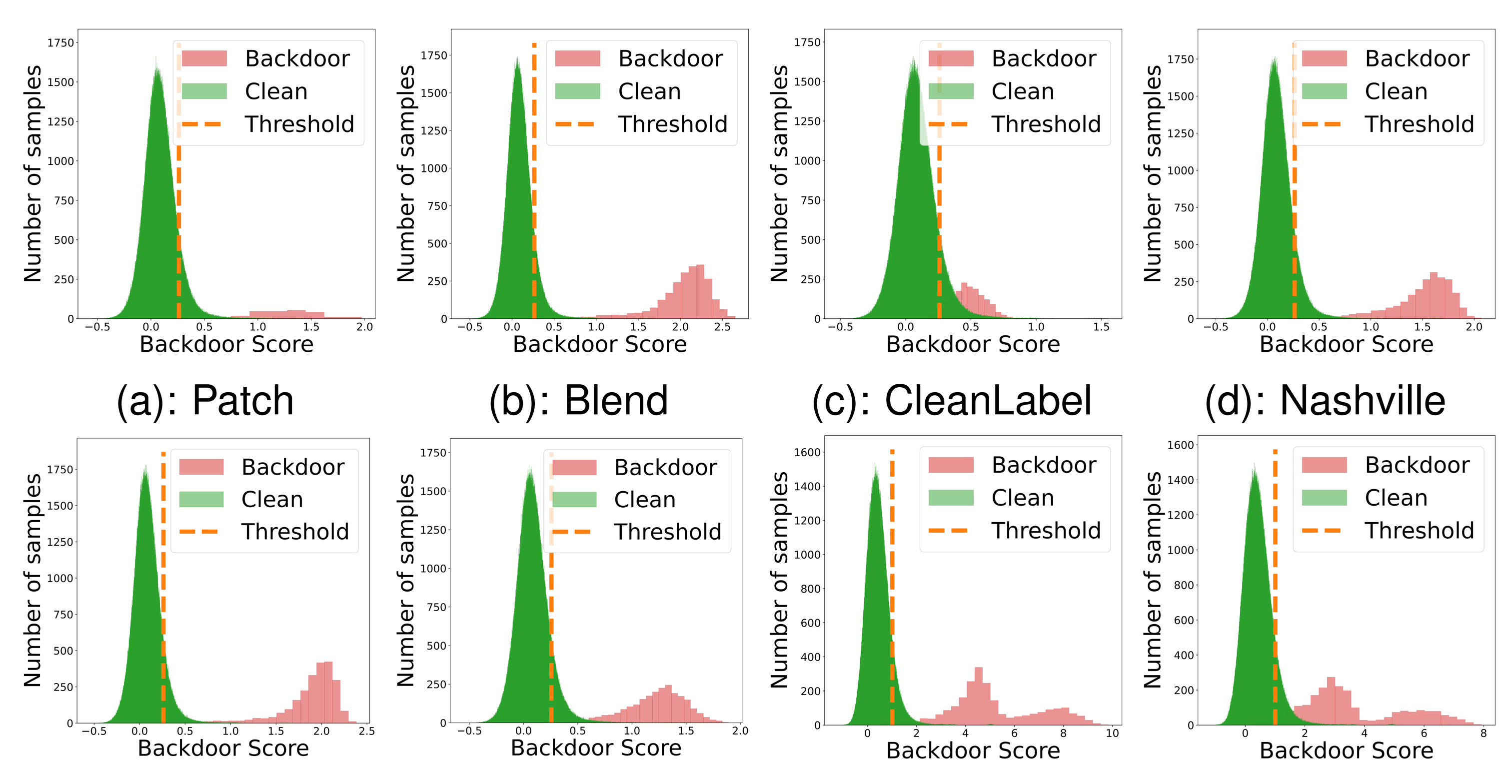
## Sparsity in the Local Neighbourhood




The CLIP learned presentations are projected into a 2-D space using t-SNE. $k$ is set to 16.

- The red cross is a backdoor data point.
- The green dot is a clean data point.
- The blue dot is a randomly sampled data point.
- The circle with the solid line is the region containing all $k$ nearest neighbours.
- The circle with a dashed line is the region containing $k$ nearest neighbours for the $k$-th neighbours.

## Detecting Backdoor Samples in CLIP



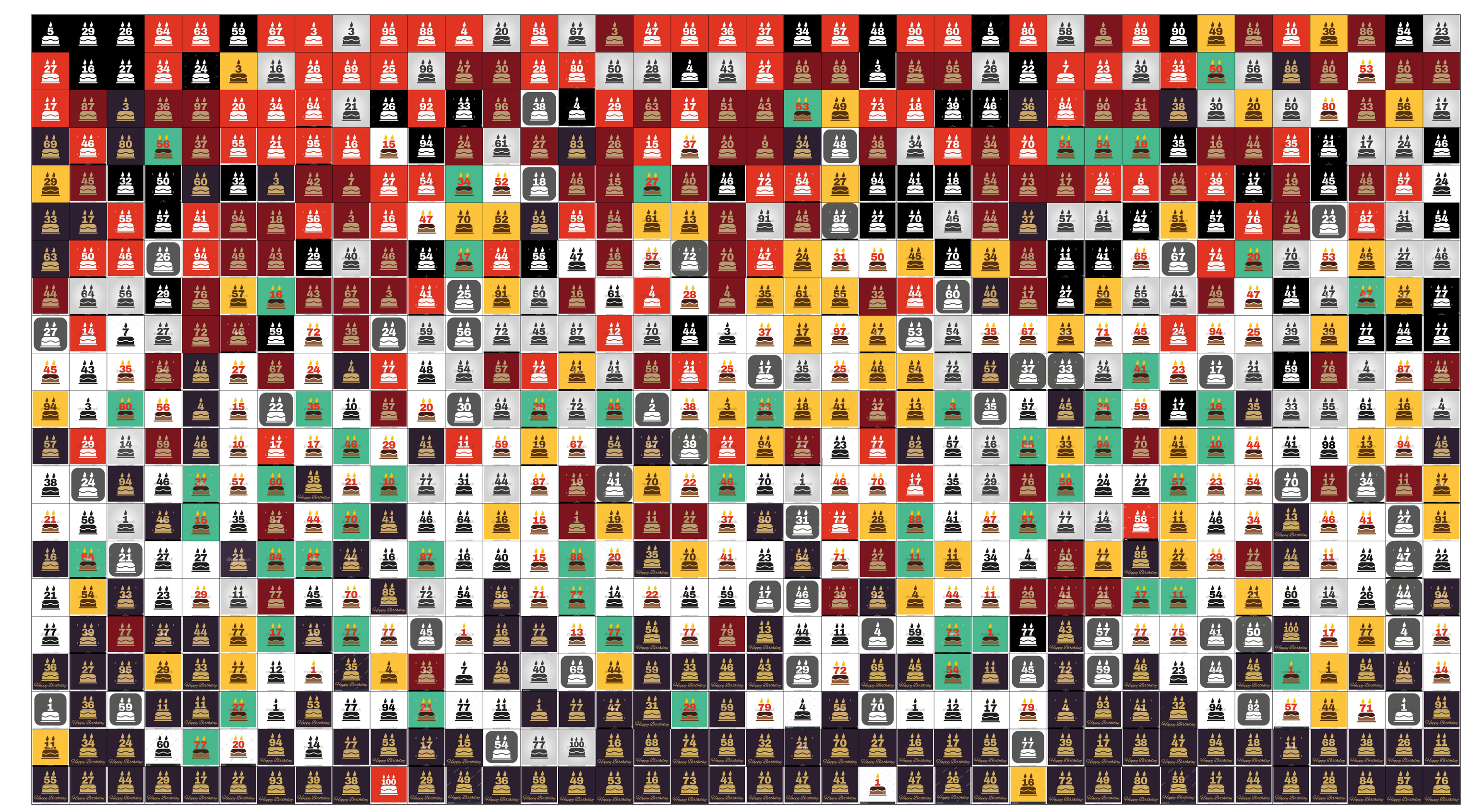

(a): Patch (b): Blend (c): CleanLabel (d): Nashville

(e): SIG (f): WaNet (g): MT-s (h): MT-M

The distributions of the DAO detection score on poisoned CC3M using ResNet-50 as the vision encoder.

## Case Study on Conceptual Captions 3 Million

The unintentional (natural) backdoor samples found in CC3M.



Caption: The birthday cake with candles in the form of number icon.

- These images appear 798 times in the dataset, accounting for approximately 0.03% of the CC3M dataset.
- These images share similar content and the same caption.
- We recovered the backdoor trigger from these samples and achieving high ASR.


The recovered trigger pattern of the birthday cake image on our pre-trained CLIP (left) and a model (mid) released by OpenCLIP that uses ResNet-50 as the vision encoder. An example of the unavailable images (right). The top 1,000 samples with the highest backdoor scores (below).