Watertight Data Splits:

A ciFAIRer benchmark and guide to robust deduplication

Christin Whitton | Mentors: Nathan DeBardeleben, Vanessa Job | HPC-DES

The Problem

DATA LEAKAGE

When information is repeated between training and test sets, a machine learning (ML) model has the opportunity for memorization. This can result in overly optimistic outcomes. Data leakage of all kinds is problematic—we focus on duplicates and near duplicates.

CIFAR and ciFAIR data

CIFAR-10: Labeled subsets of the 80 million tiny images dataset – consists of 10 classes, each with 5000 train images and 1000 test images.
CIFAR-100: Consists of 100 classes, each with 500 train images and 100 test images.

ciFAIR-10: Bartz and Denzier found 3.25% of the test set contained duplicate and near duplicates in common with the training dataset in CIFAR-10—this is a modified dataset with all duplicate test images replaced by new images.
ciFAIR-100: 10% duplicate images were found—this is a modified dataset with all duplicate test images replaced by new images.

METHOD: Bartz and Denzier trained a lightweight CNN architecture on the training set and then extracted L^2 normalized features from the average pooling layer of the trained network for both training and testing images.

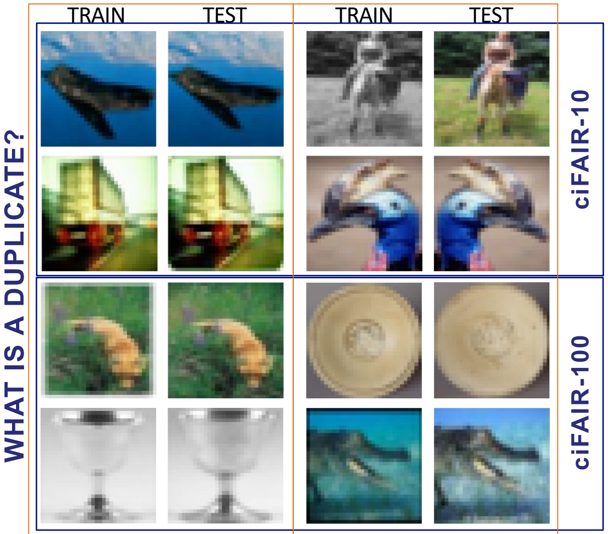
The pairs in the top box were found inside the ciFAIR-10 dataset along with **192 other duplicate test images, an additional 68%** over their original 286 duplicate images. The bottom box contains train/test pairs found inside the ciFAIR-100 dataset, along with **32 other test images, an additional 4%** over their original 891 duplicate images.

ciFAIRer Data

AN UPDATED ciFAIR DATASET, WITH ADDITIONAL DUPLICATES REPLACED

Coming soon:

- **ciFAIRer dataset available online**
- **Comparison of CNN architectures using CIFAR, ciFAIR, and ciFAIRer datasets.**



These images are not pixel-identical but are visually similar enough that the human visual system and (more importantly) our duplicate detection metrics can identify them as duplicates.

These images are considered near-duplicates. They are not identical but contain roughly the same information. These include images with color variation, rotations, introduced noise, scaling and cropping.

FOUND IN:

Duplicate Detection: Metrics

STRUCTURAL SIMILARITY INDEX (SSIM): Uses local statistics (mean, variance, covariance) to assess similarity based on structure, luminance and contrast.

PEAK SIGNAL-TO-NOISE RATIO (PSNR): Measures how well a processed signal represents the original signal.

PERCEPTUAL HASH (pHash): The Hamming distance between two pHashes—the count of differing bits.

EARTH MOVERS DISTANCE (EMD): Measures the amount of work to transform one image into the other.

FACEBOOK AI SIMILARITY SEARCH (FAISS): Builds index vectors and performs a nearest-neighbor search.

NORMALIZED CROSS CORRELATION (NCC): Measures similarity by computing the cross-correlation of two normalized image patches.

KULLBACK-LEIBLER DIVERGENCE (KL): Measures how one probability distribution diverges from a reference distribution.

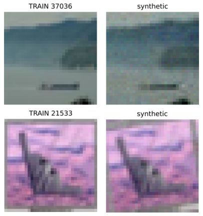
ANDERSON-DARLING TEST: A statistical test to determine if two datasets come from the same distribution

> **"Data leakage is indeed a widespread problem and has led to severe reproducibility failures."**
> - Kapoor and Narayanan, 2022

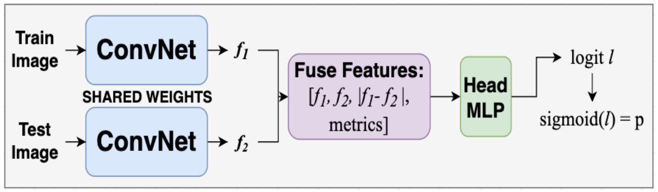
Siamese Network

Each metric has its own strengths and weaknesses – no metric predicts duplicates or near duplicates with full accuracy. These metrics were used in the initial labeling phase of the CIFAR/ciFAIR data, but our goal is robust deduplication with minimal human involvement.

A Siamese Network is designed to determine the probability of two inputs being the same (in this case, a train image and a test image). There are only 482 known duplicates in the CIFAR-10 dataset, so we created 4,000 synthetic duplicates for training only.

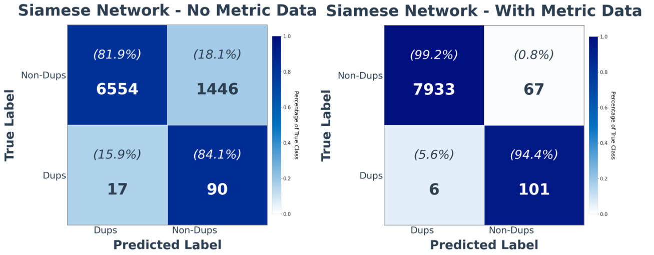


Synthetic duplicates



Siamese Network Architecture

RESULTS WHEN TRAINED ON CIFAR-10



WITH GRAYSCALE IMAGE DATA ONLY

WITH DUPLICATE DETECTION METRICS INCLUDED

Managed by Triad National Security, LLC, for the U.S. Department of Energy's NNSA.
ROSY ID: 1b8b5605 LA-UR-25-28265