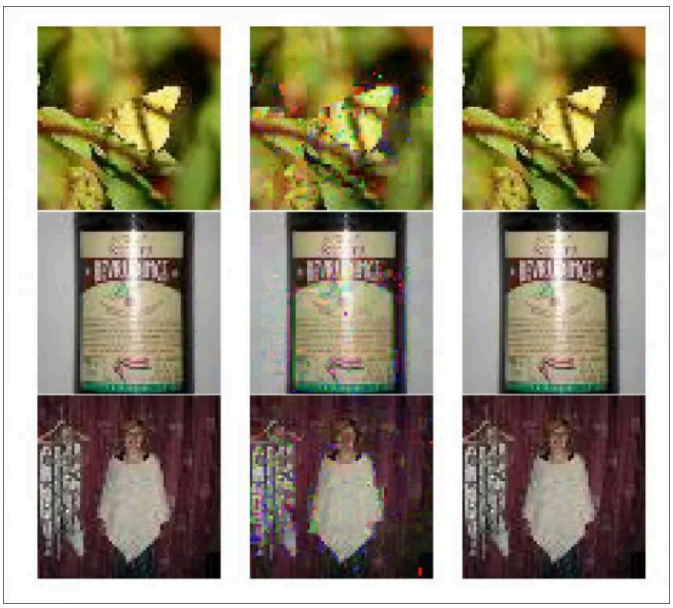
Hobbling Computer Vision Datasets Against Unauthorized Use

Originally published at https://www.unite.ai/hobbling-computer-vision-datasets-against-unauthorized-use/

Researchers from China have developed a method to copyright-protect image datasets used for computer vision training, by effectively 'watermarking' the images in the data, and then decrypting the 'clean' images via a cloud-based platform for authorized users only.

Tests on the system show that training a machine learning model on the copyright-protected images causes a catastrophic drop in model accuracy. Testing the system on two popular open source image datasets, the researchers found it was possible to drop accuracies from 86.21% and 74.00% for the clean datasets down to 38.23% and 16.20% when attempting to train models on the non-decrypted data.

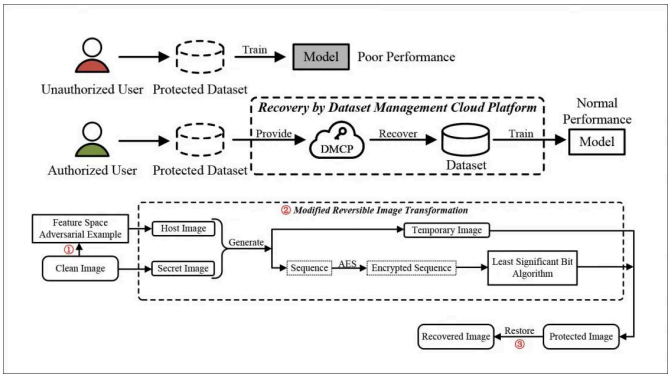


From the paper – examples, left to right, of clean, protected (i.e. perturbed) and recovered images. Source: https://arxiv.org/pdf/2109.07921.pdf

This potentially allows wide public distribution of high-quality, expensive datasets, and (presumably), even semi-crippled 'demo' training of the datasets in order to demonstrate approximate functionality.

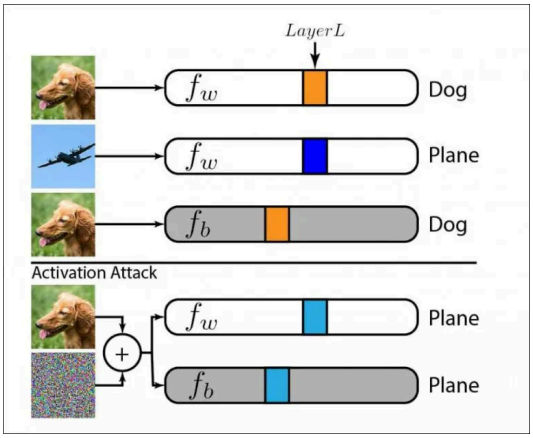
Cloud-Based Dataset Authentication

The paper comes from researchers at two departments at the Nanjing University of Aeronautics and Astronautics, and envisages the routine use of a Dataset Management Cloud Platform (DMCP), a remote authentication framework that would provide the same kind of telemetry-based pre-launch validation as has become common in burdensome local installations such as Adobe Creative Suite.



The flow and framework for the proposed method.

The protected image is generated through feature space perturbations, an adversarial attack method developed at North Carolina's Duke University in 2019.



Feature space perturbations perform an 'Activation Attack' where the features of one image are pushed towards the feature space of an adversarial image. In this case, the attack is forcing a machine learning recognition system to classify a dog as a plane.
Source: https://openaccess.thecvf.com

Next, the unmodified image is embedded into the distorted image via block pairing and block transformation, as proposed in the 2016 paper *Reversible Data Hiding in Encrypted Images by Reversible Image Transformation*.

The sequence containing the block pairing information is then embedded into a temporary interstitial image using AES encryption, the key to which will later be retrieved from the DMCP at authentication time. The Least Significant Bit steganographic algorithm is then used to embed the key. The authors refer to this process as Modified Reversible Image Transformation (mRIT).

The mRIT routine is essentially reversed at decryption time, with the 'clean' image restored for use in training sessions.

Testing

The researchers tested the system on the ResNet-18 architecture with two datasets: the 2009 work CIFAR-10, which contains 6000 images across 10 classes; and Stanford's TinyImageNet, a subset of the data for the ImageNet classification challenge which contains a training dataset of 100,000 images, along with a validation dataset of 10,000 images and a test set of 10,000 images.

The ResNet model was trained from zero on three configurations: the clean, protected and decrypted dataset. Both datasets used the Adam optimizer with an initial learning rate of 0.01, a batch size of 128 and a training epoch of 80.

Model	Dataset		Training Accuracy	Test Accuracy
ResNet-18	CIFAR-10	Clean	96.44%	86.21%
		Protected	100.00%	38.23%
		Reversed	95.21%	85.86%
	TinyImageNet	Clean	99.98%	74.00%
		Protected	100.00%	16.20%
		Reversed	99.96%	73.20%

Training and test accuracy results from tests on the encryption system. Minor losses are observable in training statistics for the reversed (i.e. decrypted) images.

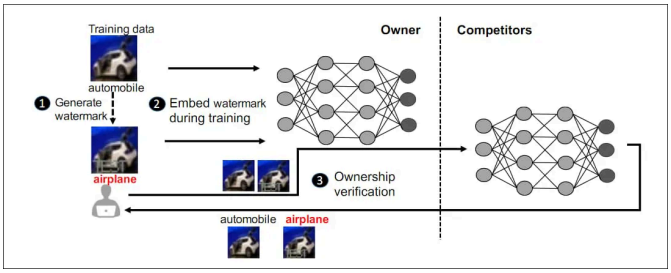
Though the paper concludes that ' the performance of the model on recovered dataset is not affected', the results do show minor losses for accuracy on recovered data vs. original data, from 86.21% to 85.86% for CIFAR-10, and 74.00% to 73.20% on TinyImageNet.

However, given the way that even minor seeding changes (as well as GPU hardware) can affect training performance, this seems to be a minimal and effective trade-off for IP-protection against accuracy.

Model Protection Landscape

Prior work has concentrated primarily on IP-protecting actual machine learning models, on the assumption that training data itself is more difficult to protect: a 2018 research effort from Japan offered a method to embed watermarks in deep neural networks; earlier work from 2017 offered a similar approach.

A 2018 initiative from IBM made perhaps the deepest and most committed investigation into the potential of watermarking for neural network models. This approach differed from the new research, in that it sought to embed non-reversible watermarks into training data and then use filters inside the neural network to 'discount' the perturbations in the data.



IBM's scheme for a neural network to 'ignore' watermarks hinged on protecting the parts of the architecture that were designed to recognize and discard the watermarked sections of the data. Source: https://gzs715.github.io/pubs/WATERMARK_ASIACCS18.pdf

Piracy Vector

Though the pursuit of IP-protecting dataset encryption frameworks might seem like an edge case in the context of a machine learning culture that is still dependent on open source review and the sharing of information among the global research community, ongoing interest in privacy-preserving identity protection algorithms seem likely to periodically produce systems that may be of interest to corporations looking to protect specific data rather than PII.

The new research does not add random perturbations to the image data, but rather crafted, forced shifts in the feature space. Therefore the current slew of watermark-removal and image enhancement computer vision projects could potentially 'restore' the images to a human-perceived higher quality without actually removing the feature perturbations that cause misclassification.

In many applications of computer vision, specifically those involving labeling and entity recognition, such illegitimately restored images would likely still cause misclassification. However, in cases where image transformations are the core objective (such as face generation or deepfake applications), algorithmically-restored images could likely still be useful in the development of functional algorithms.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More