DRM for Computer Vision Datasets

Originally published at https://www.unite.ai/drm-for-computer-vision-datasets/

History suggests that eventually the 'open' age of computer vision research, where reproducibility and favorable peer review are central to the development of a new initiative, must give way to a new era of IP protection – where closed mechanisms and walled platforms prevent competitors from undermining high dataset development costs, or from using a costly project as a mere stepping-stone to developing their own (perhaps superior) version.

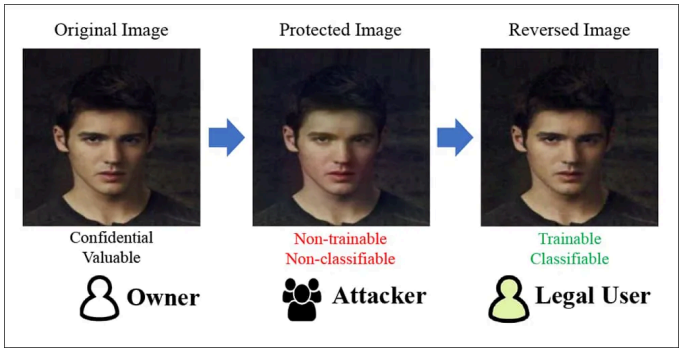
Currently the growing trend towards protectionism is mainly supported by fencing proprietary central frameworks behind API access, where users send sparse tokens or requests in, and where the transformational processes that make the framework's responses valuable are entirely hidden.

In other cases, the final model itself may be released, but without the central information that makes it valuable, such as the pre-trained weights that may have cost multiple millions to generate; or lacking a proprietary dataset, or exact details of how a subset was produced from a range of open datasets. In the case of OpenAI's transformative Natural Language model GPT-3, both protection measures are currently in use, leaving the model's imitators, such as GPT Neo, to cobble together an approximation of the product as best they can.

Copy-Protecting Image Datasets

However, interest is growing in methods by which a 'protected' machine learning framework could regain some level of portability, by ensuring that only authorized users (for instance, paid users) could profitably use the system in question. This usually involves encrypting the dataset in some programmatic way, so that it is read 'clean' by the AI framework at training time, but is compromised or in some way unusable in any other context.

Such a system has just been proposed by researchers at the University of Science and Technology of China at Anhui, and Fudan University at Shanghai. Titled *Invertible Image Dataset Protection*, the paper offers a pipeline that automatically adds adversarial example perturbation to an image dataset, so that it cannot be usefully used for training in the event of piracy, but where the protection is entirely filtered out by an authorized system containing a secret token.

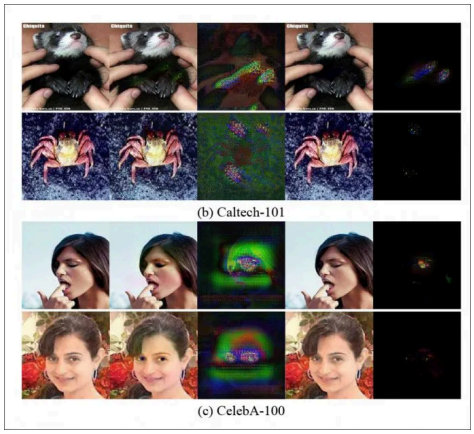


From the paper: a 'valuable' source image is rendered effectively untrainable with adversarial example techniques, with the perturbations removed systematically and entirely automatically for an 'authorized' user. Source: https://arxiv.org/pdf/2112.14420.pdf

The mechanism that enables the protection is called reversible adversarial example generator (RAEG), and effectively amounts to encryption on the actual *usability* of the images for classification purposes, using reversible data hiding (RDH). The authors state:

> *'The method first generates the adversarial image using existing AE methods, then embeds the adversarial perturbation into the adversarial image, and generates the stego image using RDH. Due to the characteristic of reversibility, the adversarial perturbation and the original image can be recovered.'*

The original images from the dataset are fed into a U-shaped invertible neural network (INN) in order to produce adversarially affected images that are crafted to deceive classification systems. This means that typical feature extraction will be undermined, making it difficult to classify traits such as gender, and other face-based features (though the architecture supports a range of domains, rather than just face-based material).



An inversion test of RAEG, where different sorts of attack are performed on the images prior to reconstruction. Attack methods include Gaussian Blur and JPEG artefacts.

Thus, if attempting to use the 'corrupted' or 'encrypted' dataset in a framework designed for GAN-based face generation, or for facial recognition purposes, the resulting model will be less effective than it would have been if it had been trained on unperturbed images.

Locking the Images

However, that's just a side-effect of the general applicability of popular perturbation methods. In fact, in the use case envisioned, the data is going to be crippled except in the case of authorized access to the target framework, since the central 'key' to the clean data is a secret token within the target architecture.

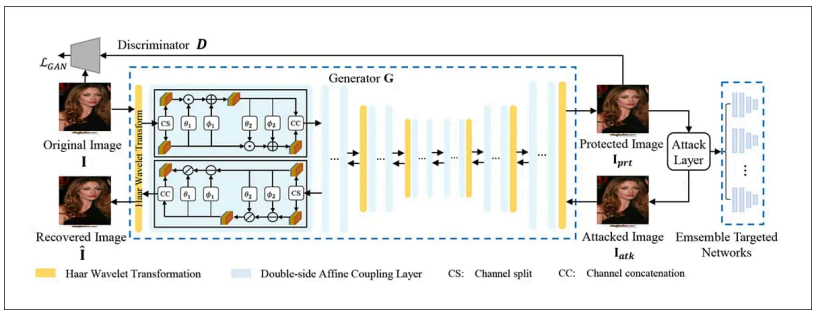
This encryption does come with a price; the researchers characterize the loss of original image quality as 'slight distortion', and state *'[The] proposed method can almost perfectly restore the original image, while the previous methods can only restore a blurry version.'*

The previous methods in question are from the November 2018 paper *Unauthorized AI cannot Recognize Me: Reversible Adversarial Example*, a collaboration between two Chinese universities and the RIKEN Center for Advanced Intelligence Project (AIP); and *Reversible Adversarial Attack based on Reversible Image Transformation*, a 2019 paper also from the Chinese academic research sector.

The researchers of the new paper claim to have made notable improvements in the usability of restored images, in comparison to these prior approaches, observing that the first approach is too sensitive to intermediary interference, and too easy to circumvent, while the second causes excessive degradation of the original images at (authorized) training time, undermining the applicability of the system.

Architecture, Data, and Tests

The new system consists of a generator, an attack layer that applies perturbation, pre-trained target classifiers, and a discriminator element.



The architecture of RAEG. Left-middle, we see the secret token 'I_{prt}', which will allow de-perturbation of the image at training time, by identifying the perturbed feat images and discounting them.

Below are the results of a test comparison with the two prior approaches, using three datasets: CelebA-100; Caltech-101; and Mini-ImageNet.

Table 1. Comparison of accuracy on pretrained ResNet-50/DenseNet-121 using RAE[28], RIT[29] and RAEG.

Dataset	Method	No Attack	JPEG QF=90	JPEG QF=50	Resize	Gaussian Noise	Medium Filter	ComDefend	DIPDefend
Caltech-101	RAEG	0.479 / 0.495	0.587 / 0.570	0.752 / **0.721**	**0.615 / 0.701**	**0.607 / 0.732**	**0.504 / 0.672**	**0.632 / 0.709**	**0.587 / 0.704**
	RAE-IGSM	**0.076 / 0.063**	**0.376 / 0.394**	0.890 / 0.873	0.755 / 0.712	0.816 / 0.822	0.823 / 0.799	0.748 / 0.752	0.843 / 0.835
	RAE-CW	0.656 / 0.673	0.955 / 0.958	0.961 / 0.967	0.950 / 0.961	0.957 / 0.942	0.880 / 0.868	0.937 / 0.921	0.891 / 0.903
	RIT-IGSM	0.679 / 0.850	0.697 / 0.854	**0.693** / 0.854	0.677 / 0.872	0.729 / 0.888	0.693 / 0.829	0.779 / 0.891	0.747 / 0.893
CelebA-100	RAEG	0.060 / 0.093	0.070 / 0.090	**0.073 / 0.083**	0.103 / 0.127	**0.093 / 0.113**	**0.070 / 0.083**	**0.080 / 0.093**	**0.080 / 0.089**
	RAE-IGSM	**0.003 / 0.001**	**0.003 / 0.003**	0.275 / 0.267	**0.053 / 0.066**	0.113 / 0.125	0.343 / 0.359	0.257 / 0.298	0.387 / 0.365
	RAE-CW	0.493 / 0.452	0.762 / 0.733	0.796 / 0.762	0.773 / 0.745	0.752 / 0.736	0.657 / 0.662	0.710 / 0.694	0.656 / 0.677
	RIT-IGSM	0.114 / 0.442	0.380 / 0.428	0.137 / 0.421	0.127 / 0.391	0.174 / 0.515	0.204 / 0.522	0.224 / 0.539	0.197 / 0.572
Mini-ImageNet	RAEG	**0.068 / 0.092**	**0.082 / 0.113**	**0.171 / 0.183**	**0.117 / 0.148**	**0.129 / 0.118**	**0.087 / 0.124**	**0.099 / 0.124**	**0.101 / 0.183**
	RAE-IGSM	0.192 / 0.225	0.488 / 0.473	0.819 / 0.835	0.687 / 0.691	0.770 / 0.796	0.790 / 0.788	0.796 / 0.802	0.823 / 0.819
	RAE-CW	0.434 / 0.451	0.929 / 0.917	0.923 / 0.940	0.902 / 0.925	0.924 / 0.912	0.859 / 0.863	0.900 / 0.878	0.870 / 0.856
	RIT-IGSM	0.572 / 0.805	0.586 / 0.791	0.623 / 0.791	0.609 / 0.813	0.751 / 0.879	0.655 / 0.799	0.755 / 0.860	0.751 / 0.880

The three datasets were trained as target classification networks, with a batch size of 32, on a NVIDIA RTX 3090 over the course of a week, for 50 epochs.

The authors claim that RAEG is the first work to offer an invertible neural network that can actively generate adversarial examples.

First published 4th January 2022.

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