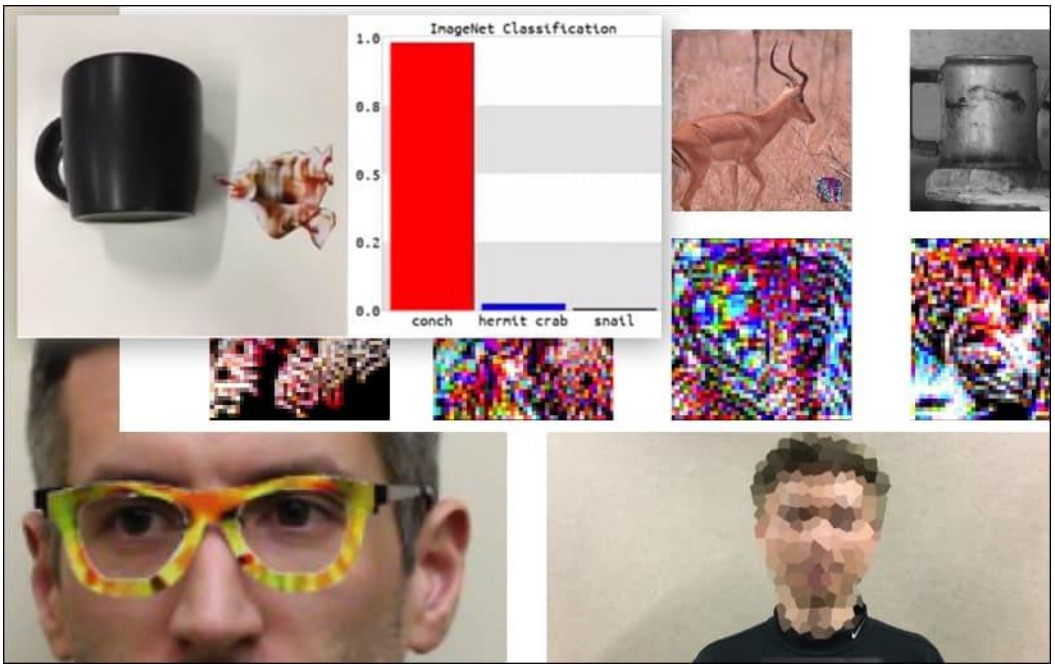
# Why Adversarial Image Attacks Are No Joke

*By* Martin Anderson



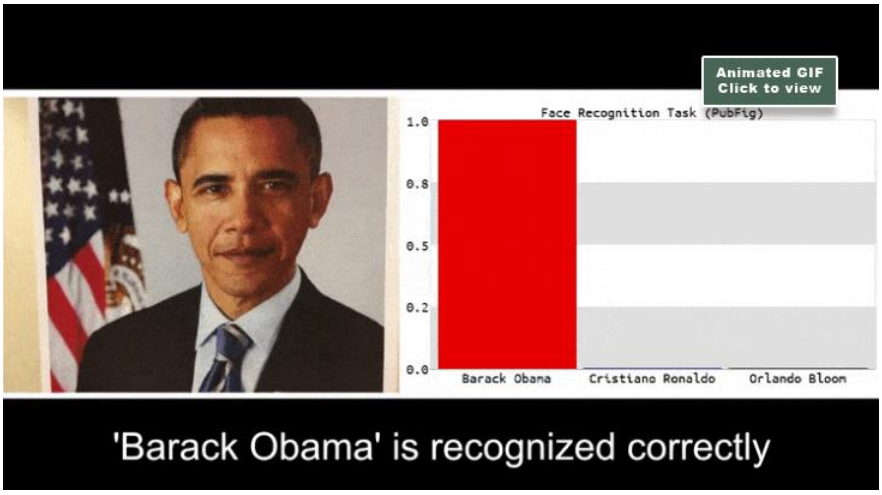

*First published* ***December 1st, 2021*** *at:*
*https://www.unite.ai/why-adversarial-image-attacks-are-no-joke/* | *Web-archived version*

***Attacking image recognition systems with carefully-crafted adversarial images has been considered an amusing but trivial proof-of-concept over the last five years. However, new research from Australia suggests that the casual use of highly popular image datasets for commercial AI projects could create an enduring new security problem.***

For a couple of years now, a group of academics at the University of Adelaide has been trying to explain something really important about the future of AI-based image recognition systems.

It's something that would be difficult (and very expensive) to fix *right now*, and which would be unconscionably costly to remedy once the current trends in image recognition research have been fully developed into commercialized and industrialized deployments in 5-10 years' time.

Before we get into it, let's have a look at a flower being classified as President Barack Obama, from one of the six videos that the team has published on the project page:

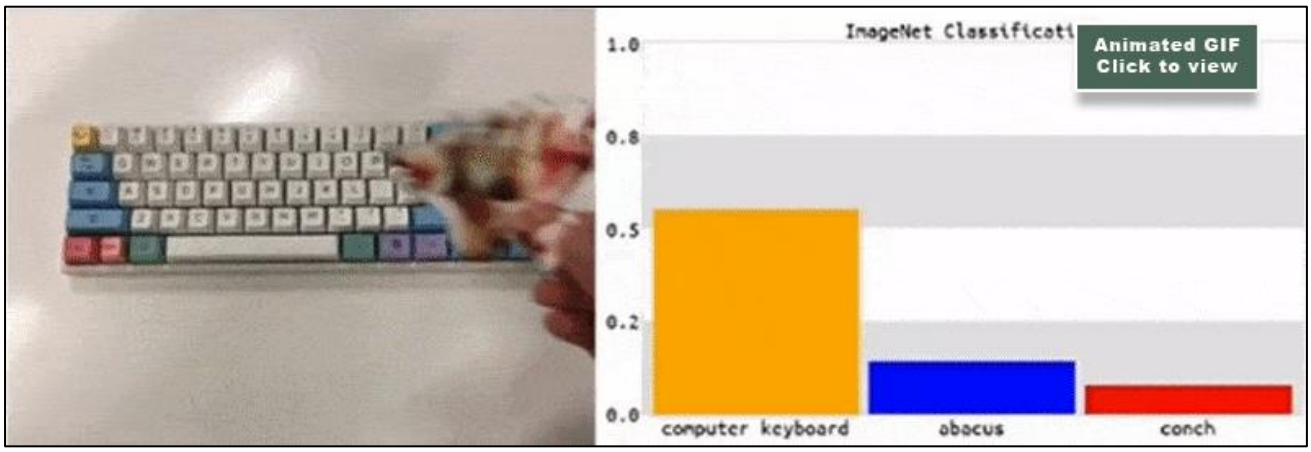



*Source: https://www.youtube.com/watch?v=Klepca1Ny3c*

In the above image, a facial recognition system that clearly knows how to recognize Barack Obama is fooled into 80% certainty that an anonymized man holding a crafted, printed adversarial image of a flower is also Barack Obama. The system doesn't even care that the 'fake face' is on the subject's chest, instead of on his shoulders.

Although it's impressive that the researchers have been able to accomplish this kind of identity capture by generating a coherent image (a flower) instead of just the usual random noise, it seems that goofy exploits like this crop up fairly regularly in security research on computer vision. For instance, those weirdly-patterned glasses that were able to fool face recognition back in 2016, or specially-crafted adversarial images that attempt to rewrite road signs.

If you're interested, the Convolutional Neural Network (CNN) model being attacked in the above example is VGGFace (VGG-16), trained on Columbia University's PubFig dataset. Other attack samples developed by the researchers used different resources in different combinations.




*A keyboard is re-classed as a conch, in a WideResNet50 model on ImageNet. The researchers have also ensured that the model has no bias towards conches. See the full video for extended and additional demonstrations at https://www.youtube.com/watch?v=dhTTj*

## Image Recognition as an Emerging Attack Vector

The many impressive attacks that the researchers outline and illustrate are not criticisms of individual datasets or specific machine learning architectures that use them. Neither can they be easily defended against by switching datasets or models, retraining models, or any of the other 'simple' remedies that cause ML practitioners to scoff at sporadic demonstrations of this kind of trickery.

Rather, the Adelaide team's exploits exemplify a *central weakness* in the entire current architecture of image recognition AI development; a weakness which could be set to expose many future image recognition systems to facile manipulation by attackers, and to put any subsequent defensive measures on the back foot.

Imagine the latest adversarial attack images (such as the flower above) being added as 'zero-day exploits' to security systems of the future, just as current anti-malware and antivirus frameworks update their virus definitions every day.

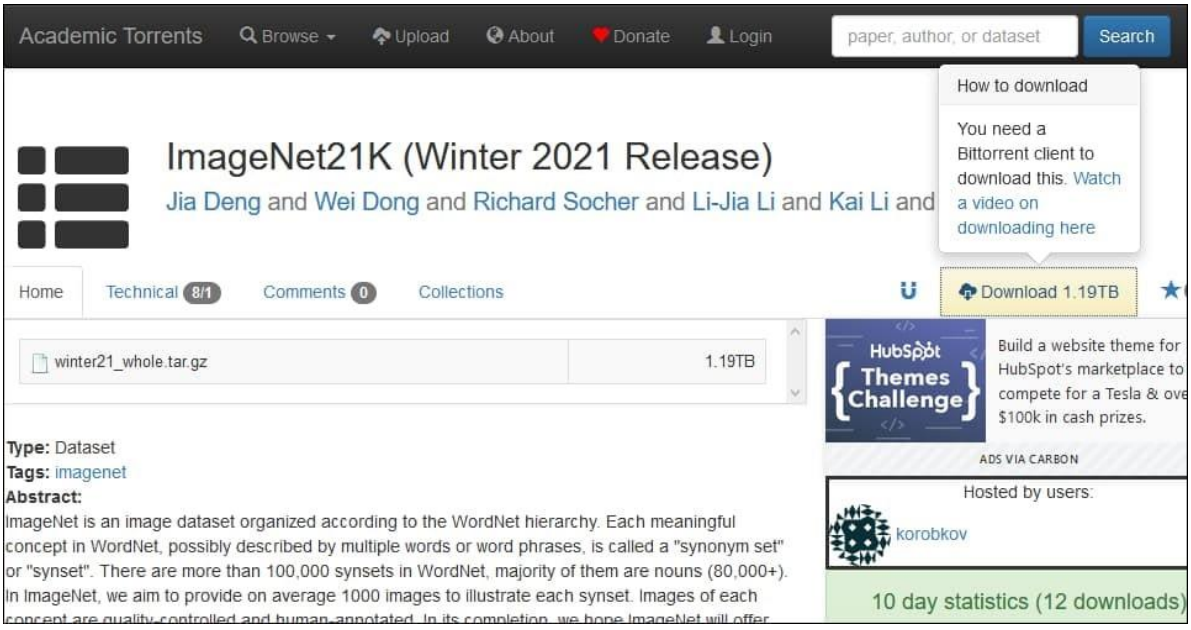
The potential for novel adversarial image attacks would be inexhaustible, because the foundation architecture of the system didn't anticipate downstream problems, as occurred with the internet, the Millennium Bug and the leaning Tower of Pisa.

In what way, then, are we setting the scene for this?

## Getting the Data for an Attack

Adversarial images such as the 'flower' example above are generated by having access to the image datasets that trained the computer models. You don't need 'privileged' access to training data (or model architectures), since the most popular datasets (and many trained models) are widely available in a robust and constantly-updating torrent scene.

For instance, the venerable Goliath of Computer Vision datasets, ImageNet, is available to Torrent in all its many iterations, bypassing its customary restrictions, and making available crucial secondary elements, such as validation sets.




*Source: https://academictorrents.com*

If you have the data, you can (as the Adelaide researchers observe) effectively 'reverse-engineer' any popular dataset, such as CityScapes, or CIFAR.

In the case of PubFig, the dataset which enabled the 'Obama Flower' in the earlier example, Columbia University has addressed a growing trend in copyright issues around image dataset redistribution by instructing researchers how to *reproduce* the dataset via curated links, rather than making the compilation directly available, observing *'This seems to be the way other large web-based databases seem to be evolving'*.

In most cases, that's not necessary: Kaggle estimates that the ten most popular image datasets in computer vision are: CIFAR-10 and CIFAR-100 (both directly downloadable); CALTECH-101 and 256 (both available, and both currently available as torrents); MNIST (officially available, also on torrents); ImageNet (see above); Pascal VOC (available, also on torrents); MS COCO (available, and on torrents); Sports-1M (available); and YouTube-8M (available).

This availability is also representative of the wider range of available computer vision image datasets, since obscurity is death in a 'publish or perish' open source development culture.

In any case, the scarcity of manageable new datasets, the high cost of image-set development, the reliance on 'old favorites', and the tendency to simply adapt older datasets all exacerbate the problem outlined in the new Adelaide paper.

## Typical Criticisms of Adversarial Image Attack Methods

The most frequent and persistent criticism of machine learning engineers against the effectiveness of the latest adversarial image attack technique is that the attack is *specific to a particular dataset, a particular model, or both*; that it is not 'generalizable' to other systems; and, consequently, represents only a trivial threat.

The second-most frequent complaint is that the adversarial image attack is *'white box'*, meaning that you would need direct access to the training environment or data. This is indeed an unlikely scenario, in most cases – for instance, if you wanted to exploit the training process for the facial recognition systems of London's Metropolitan Police, you'd have to hack your way into NEC, either with a console or an axe.

## The Long-Term 'DNA' of Popular Computer Vision Datasets

Regarding the first criticism, we should consider not only that a mere handful of computer vision datasets dominate the industry by sector year-on-year (i.e. ImageNet for multiple types of object, CityScapes for driving scenes, and FFHQ for facial recognition); but also that, as simple annotated image data, they are 'platform agnostic' and highly transferable.

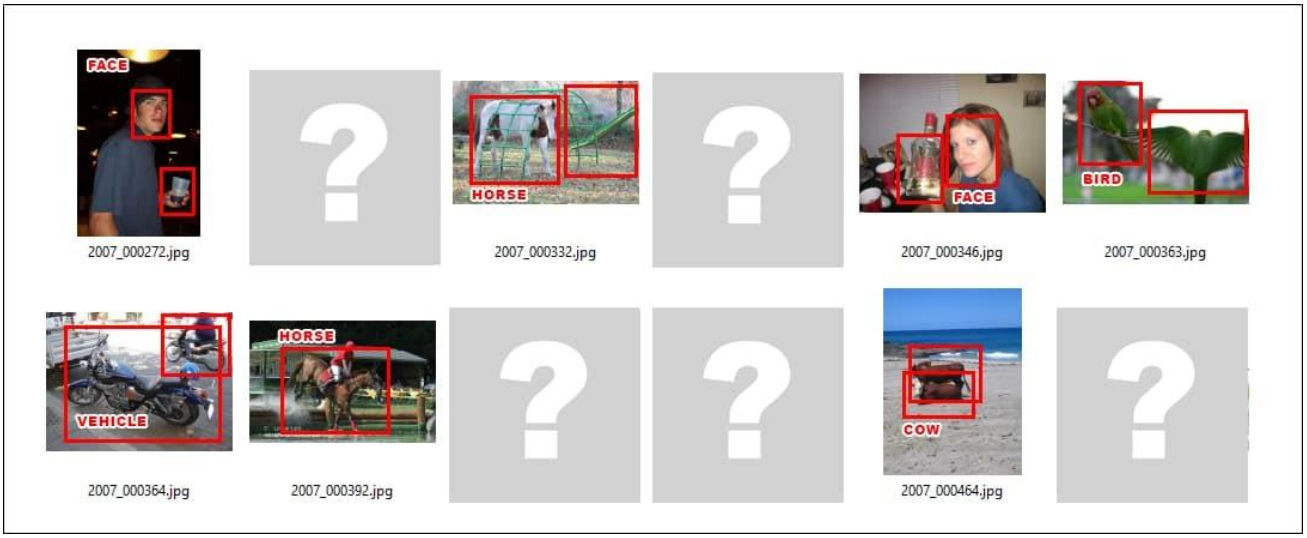
Depending on its capabilities, any computer vision training architecture will find *some* features of objects and classes in the ImageNet dataset. Some architectures may find more features than others, or make more useful connections than others, but *all* should find at least the highest-level features:


*ImageNet data, with the minimum viable number of correct identifications – 'high level' features.*

It's those 'high-level' features that distinguish and 'fingerprint' a dataset, and which are the reliable 'hooks' on which to hang a long-term adversarial image attack methodology that can straddle different systems, and grow in tandem with the 'old' dataset as the latter is perpetuated in new research and products.

A more sophisticated architecture will produce more accurate and granular identifications, features and classes:

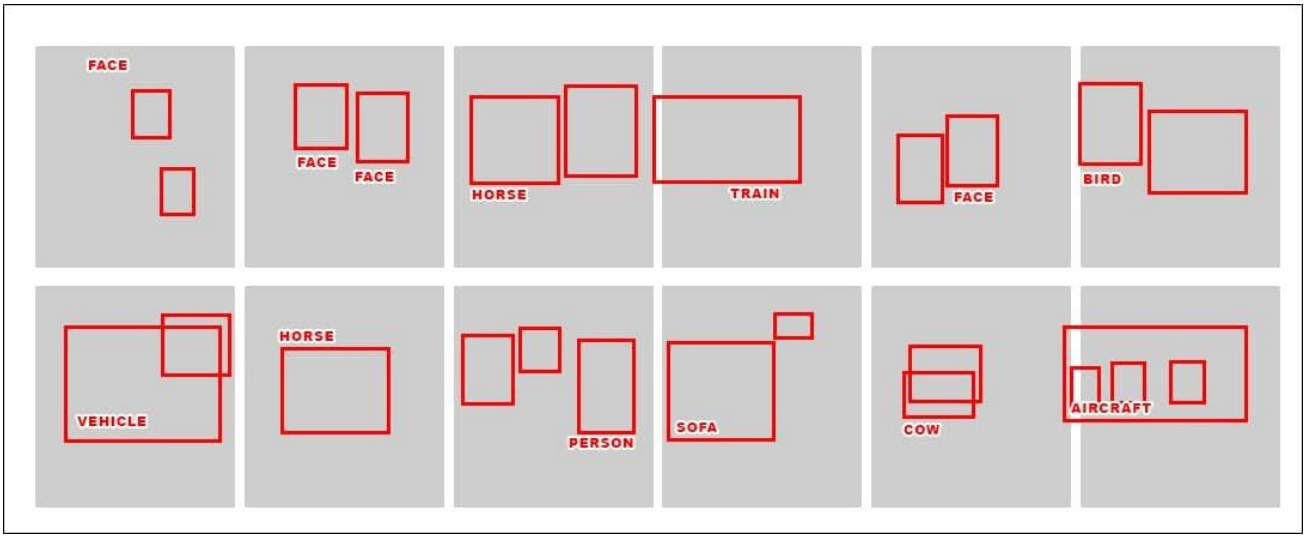

However, the more an adversarial attack generator relies on these *lower* features (i.e. 'Young Caucasian Male' instead of 'Face'), the less effective it will be in cross-over or later architectures that use *different versions* of the original dataset – such as a sub-set or filtered set, where many of the original images from the full dataset are not present:




## Adversarial Attacks on 'Zeroed', Pre-Trained Models

What about cases where you just download a pre-trained model that was originally trained on a highly popular dataset, and give it completely new data?

The model has already been trained on (for instance) ImageNet, and all that's left are the weights, which may have taken weeks or months to train, and are now ready to help you identify similar objects to those that existed in the original (now absent) data.




*With the original data removed from the training architecture, what's left is the 'predisposition' of the model to classify objects in the way that it originally learned to do, which will essentially cause many of the original 'signatures' to reform and become vulnerable once again to the same old Adversarial Image Attack methods.*

Those weights are valuable. Without the data *or* the weights, you essentially have an empty architecture with no data. You're going to have to train it from scratch, at great expense of time and computing resources, just like the original authors did (probably on more powerful hardware and with a higher budget than you have available).

The trouble is that the weights are already pretty well-formed and resilient. Though they will adapt somewhat in training, they're going to behave similarly on your new data as they did on the original data, producing signature features that an adversarial attack system can key back in on.

In the long term, this too preserves the 'DNA' of computer vision datasets that are twelve or more years old, and may have passed through a notable evolution from open source efforts through to commercialized deployments – even where the original training data was completely jettisoned at the start of the project. Some of these commercial deployments may not occur for years yet.

## No White Box Needed

Regarding the second common criticism of adversarial image attack systems, the authors of the new paper have found that their ability to deceive recognition systems with crafted images of flowers is highly transferable across a number of architectures.

Whilst observing that their 'Universal NaTuralistic adversarial paTches' (TnT) method is the first to use recognizable images (rather than random perturbation noise) to fool image recognition systems, the authors also state:

> *'[TnTs] are effective against multiple state-of-the-art classifiers ranging from widely used WideResNet50 in the Large-Scale Visual Recognition task of ImageNet dataset to VGG-face models in the face recognition task of PubFig dataset in both targeted and untargeted attacks.*
>
> *'TnTs can possess: i) the naturalism achievable [with] triggers used in Trojan attack methods; and ii)the generalization and transferability of adversarial examples to other networks.*
>
> *'This raises safety and security concerns regarding already deployed DNNs as well as future DNN deployments where attackers can use inconspicuous natural-looking object patches to misguide neural network systems without tampering with the model and risking discovery.'*

The authors suggest that conventional countermeasures, such as degrading the Clean Acc. of a network, could theoretically provide some defense against TnT patches, but that *'TnTs still can successfully bypass this SOTA provable defense methods with most of the defending systems achieving 0% Robustness'.*

Possible other solutions include federated learning, where the provenance of contributing images is protected, and new approaches that could directly 'encrypt' data at training time, such as one recently suggested by the Nanjing University of Aeronautics and Astronautics.

Even in those cases, it would be important to train on genuinely *new* image data – by now the images and associated annotations in the small cadre of the most popular CV datasets are so embedded in development cycles around the world as to resemble software more than data; software that often hasn't been notably updated in years.

## Conclusion

Adversarial image attacks are being made possible not only by open source machine learning practices, but also by a corporate AI development culture that is motivated to reuse well-established computer vision datasets for several reasons: they've already proved effective; they're far cheaper than 'starting from scratch';

and they're maintained and updated by vanguard minds and organizations across academia and industry, at levels of funding and staffing that would be difficult for a single company to replicate.

Additionally, in many cases where the data is not original (unlike CityScapes), the images were gathered prior to recent controversies around privacy and data-gathering practices, leaving these older datasets in a kind of semi-legal purgatory that may look comfortingly like a 'safe harbor', from a company's point of view.

TnT Attacks! Universal Naturalistic Adversarial Patches Against Deep Neural Network Systems *is co-authored by Bao Gia Doan, Minhui Xue, Ehsan Abbasnejad, Damith C. Ranasinghe from the University of Adelaide, together with Shiqing Ma from the Department of Computer Science at Rutgers University.*