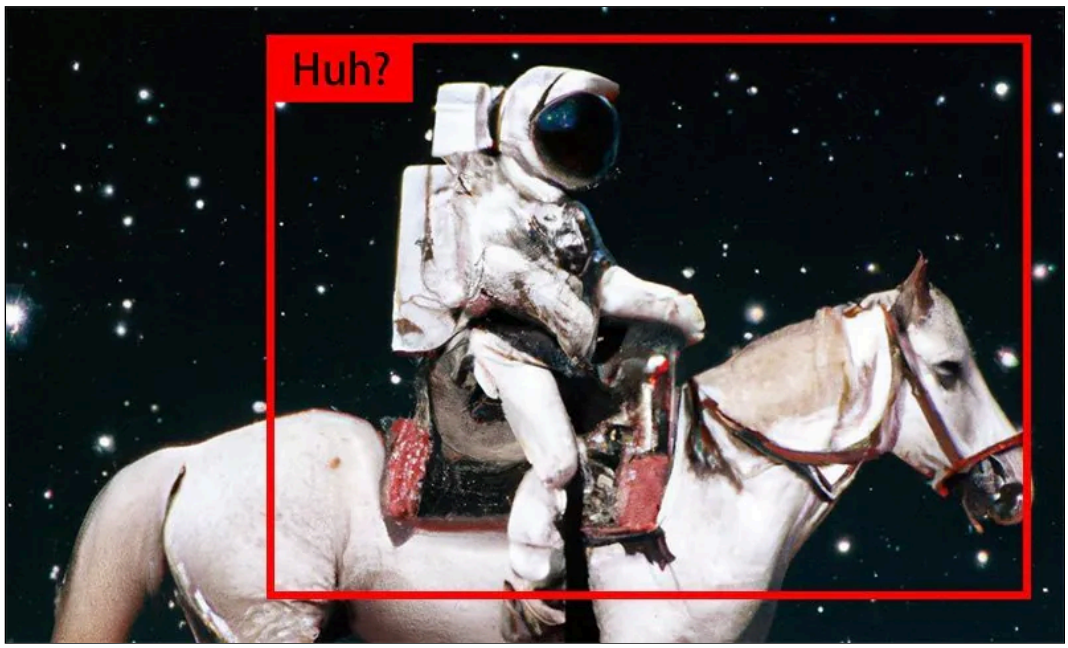
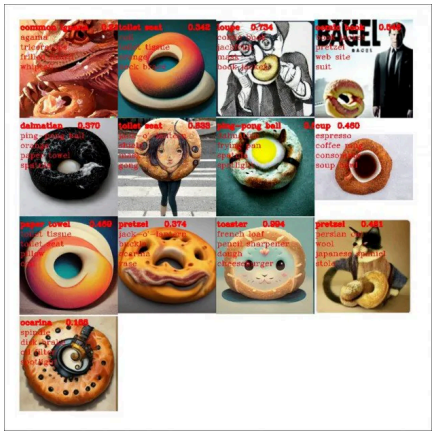
Deep Learning Models Might Struggle to Recognize AI-Generated Images

Originally published at https://www.unite.ai/deep-learning-models-might-struggle-to-recognize-ai-generated-images/

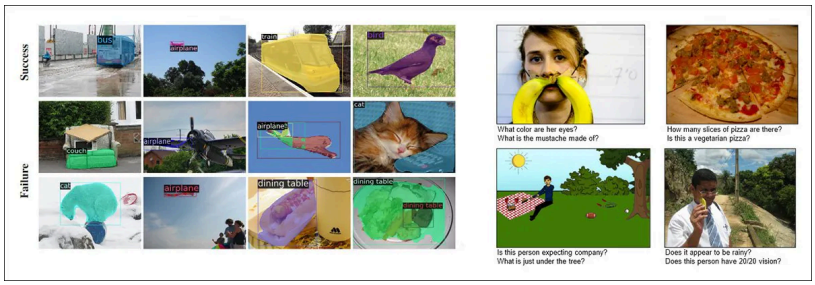


Findings from a new paper indicate that state-of-the-art AI is significantly less able to recognize and interpret AI-synthesized images than people, which may be of concern in a coming climate where machine learning models are increasingly trained on synthetic data, and where it won't necessarily be known if the data is 'real' or not.



Here we see the resnext101_32x8d_wsl prediction model struggling in the 'bagel' category. In the tests, a recognition failure was deemed to have occurred if the core target word (in this case 'bagel') was not featured in the top five predicted results. Source: https://arxiv.org/pdf/2208.10760.pdf

The new research tested two categories of computer vision-based recognition framework: object recognition, and visual question answering (VQA).



On the left, inference successes and failures from an object recognition system; on the right, VQA tasks designed to probe AI understanding of scenes and images in a significant way. Sources: https://arxiv.org/pdf/2105.05312.pdf and https://arxiv.org/pdf/1505.00468.pdf

Out of ten state-of-the-art models tested on curated datasets generated by image synthesis frameworks DALL-E 2 and Midjourney, the best-performing model was able to achieve only 60% and 80% top-5 accuracy across the two types of test, whereas ImageNet, trained on non-synthetic, real-world data, can respectively achieve 91% and 99% in the same categories, while human performance is typically notably higher.

Addressing issues around distribution shift (aka 'Model Drift', where prediction models experience diminished predictive capacity when moved from training data to 'real' data), the paper states:

> *'Humans are able to recognize the generated images and answer questions on them easily. We conclude that a) deep models struggle to understand the generated content, and may do better after fine-tuning, and b) there is a large distribution shift between the generated images and the real photographs. The distribution shift appears to be category-dependent.'*

Given the volume of synthetic images already flooding the internet in the wake of last week's sensational open-sourcing of the powerful Stable Diffusion latent diffusion synthesis model, the possibility naturally arises that as 'fake' images flood into industry-standard datasets such as Common Crawl, variations in accuracy over the years could be significantly affected by 'unreal' images.

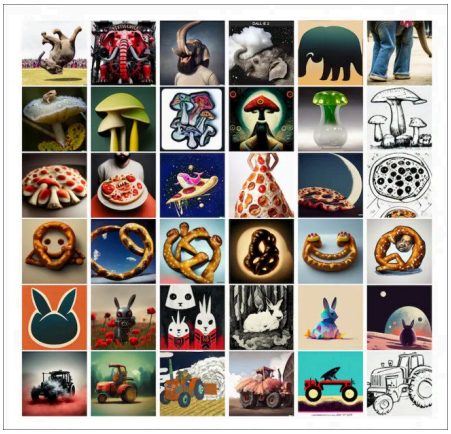
Though synthetic data has been heralded as the potential savior of the data-starved computer vision research sector, which often lacks resources and budgets for hyperscale curation, the new torrent of Stable Diffusion images (along with the general rise in synthetic images since the advent and commercialization of DALL-E 2) are unlikely to all come with handy labels, annotations and hashtags distinguishing them as 'fake' at the point that greedy machine vision systems scrape them from the internet.

The speed of development in open source image synthesis frameworks has notably outpaced our ability to categorize images from these systems, leading to growing interest in 'fake image' detection systems, similar to deepfake detection systems, but tasked with evaluating whole images rather than sections of faces.

The new paper is titled *How good are deep models in understanding the generated images*?, and comes from Ali Borji of San Francisco machine learning startup Quintic AI.

Data

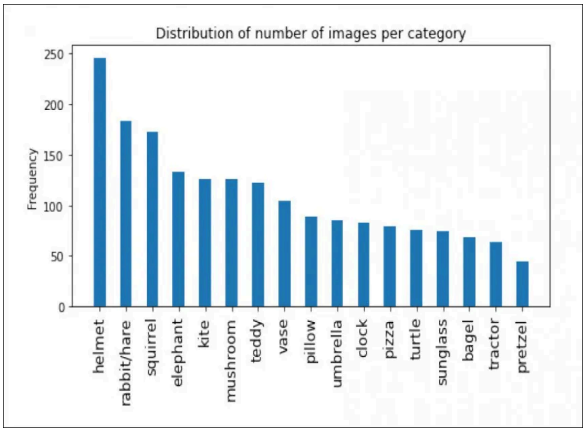
The study predates the Stable Diffusion release, and the experiments use data generated by DALL-E 2 and Midjourney across 17 categories, including *elephant*, *mushroom*, *pizza*, *pretzel*, *tractor* and *rabbit*.



Examples of the images from which the tested recognition and VQA systems were challenged to identify the most important key concept.

Images were obtained via web searches and through Twitter, and, in accordance with DALL-E 2's policies (at least, at the time), did not include any images featuring human faces. Only good quality images, recognizable by humans, were chosen.

Two sets of images were curated, one each for the object recognition and VQA tasks.

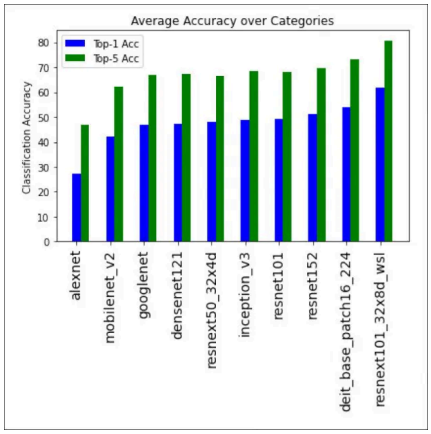


The number of images present in each tested category for object recognition.

Testing Object Recognition

For the object recognition tests, ten models, all trained on ImageNet, were tested: AlexNet, ResNet152, MobileNetV2, DenseNet, ResNext, GoogleNet, ResNet101, Inception_V3, Deit, and ResNext_WSL.

Some of the classes in the tested systems were more granular than others, necessitating the application of averaged approaches. For instance, ImageNet contains three classes retaining to 'clocks', and it was necessary to define some kind of arbitrational metric, where the inclusion of any 'clock' of any type in the top five obtained labels for any image was regarded as a success in that instance.



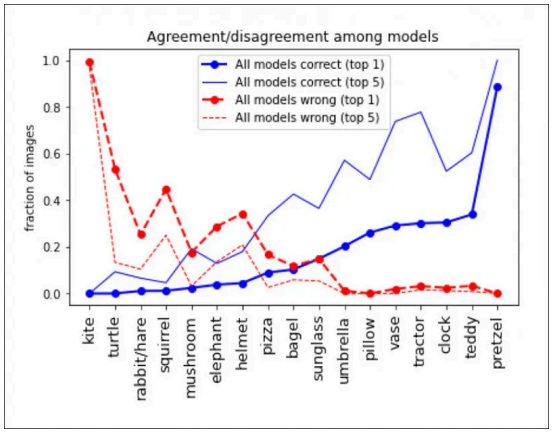
Per-model performance across 17 categories.

The best-performing model in this round was resnext101_32x8d_ws, achieving near 60% for top-1 (i.e., the times where its preferred prediction out of five guesses was the correct concept embodied in the image), and 80% for top-five (i.e. the desired concept was at least listed somewhere in the model's five guesses about the picture).

The author suggests that this model's good performance is due to the fact that it was trained for the weakly-supervised prediction of hashtags in social media platforms. However, these leading results, the author notes, are notably below what ImageNet is able to achieve on real data, i.e. 91% and 99%. He suggests that this is due to a major disparity between the distribution of ImageNet images (which are also scraped from the web) and generated images.

The five most difficult categories for the system, in order of difficulty, were *kite*, *turtle*, *squirrel*, *sunglasses* and *helmet*. The paper notes that the *kite* class is often confused with *balloon*, *parachute* and *umbrella*, though these distinctions are trivially easy for human observers to individuate.

Certain categories, including *kite* and *turtle*, caused universal failure across all models, while others (notably *pretzel* and *tractor*) resulted in almost universal success across the tested models.



Polarizing categories: some of the target categories chosen either foxed all the models, or else were fairly easy for all the models to identify.

The authors postulate that these findings indicate that all object recognition models may share similar strengths and weaknesses.

Testing Visual Question Answering

Next, the author tested VQA models on open-ended and free-form VQA, with binary questions (i.e. questions to which the answer can only be 'yes' or 'no'). The paper notes that recent state-of-the-art VQA models are able to achieve 95% accuracy on the VQA-v2 dataset.

For this stage of testing, the author curated 50 images and formulated 241 questions around them, 132 of which had positive answers, and 109 negative. The average question length was 5.12 words.

This round used the OFA model, a task-agnostic and modality-agnostic framework to test task comprehensiveness, and was recently the leading scorer in the VQA-v2 test-std set. OFA scored 77.27% accuracy on the generated images, compared to its own 94.7% score in the VQA-v2 test-std set.

Example questions and results from the VQA section of the tests. 'GT" is 'Ground Truth', i.e., the correct answer.

The paper's author suggests that part of the reason may be that the generated images contain semantic concepts absent from the VQA-v2 dataset, and that the questions written for the VQA tests may be more challenging the general standard of VQA-v2 questions, though he believes that the former reason is more likely.

LSD in the Data Stream?

Opinion

The new proliferation of AI-synthesized imagery, which can present instant conjunctions and abstractions of core concepts that do not exist in nature, and which would be prohibitively time-consuming to produce via conventional methods, could present a particular problem for weakly supervised data-gathering systems, which may not be able to fail gracefully – largely because they were not designed to handle high volume, unlabeled synthetic data.

In such cases, there may be a risk that these systems will corral a percentage of 'bizarre' synthetic images into incorrect classes simply because the images feature distinct objects which do not really belong together.

'Astronaut riding a horse' has perhaps become the most emblematic visual for the new generation of image synthesis systems – but these 'unreal' relationships could enter real detection systems unless care is taken. Source: https://twitter.com/openai/status/1511714545529614338?lang=en

Unless this can be prevented at the preprocessing stage prior to training, such automated pipelines could lead to improbable or even grotesque associations being trained into machine learning systems, degrading their effectiveness, and risking to pass high-level associations into downstream systems and sub-classes and categories.

Alternatively, disjointed synthetic images could have a 'chilling effect' on the accuracy of later systems, in the eventuality that new or amended architectures should emerge which attempt to account for *ad hoc* synthetic imagery, and cast too wide a net.

In either case, synthetic imagery in the post Stable Diffusion age could prove to be a headache for the computer vision research sector whose efforts made these strange creations and capabilities possible – not least because it imperils the sector's hope that the gathering and curation of data can eventually be far more automated than it currently is, and far less expensive and time-consuming.

First published 1st September 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