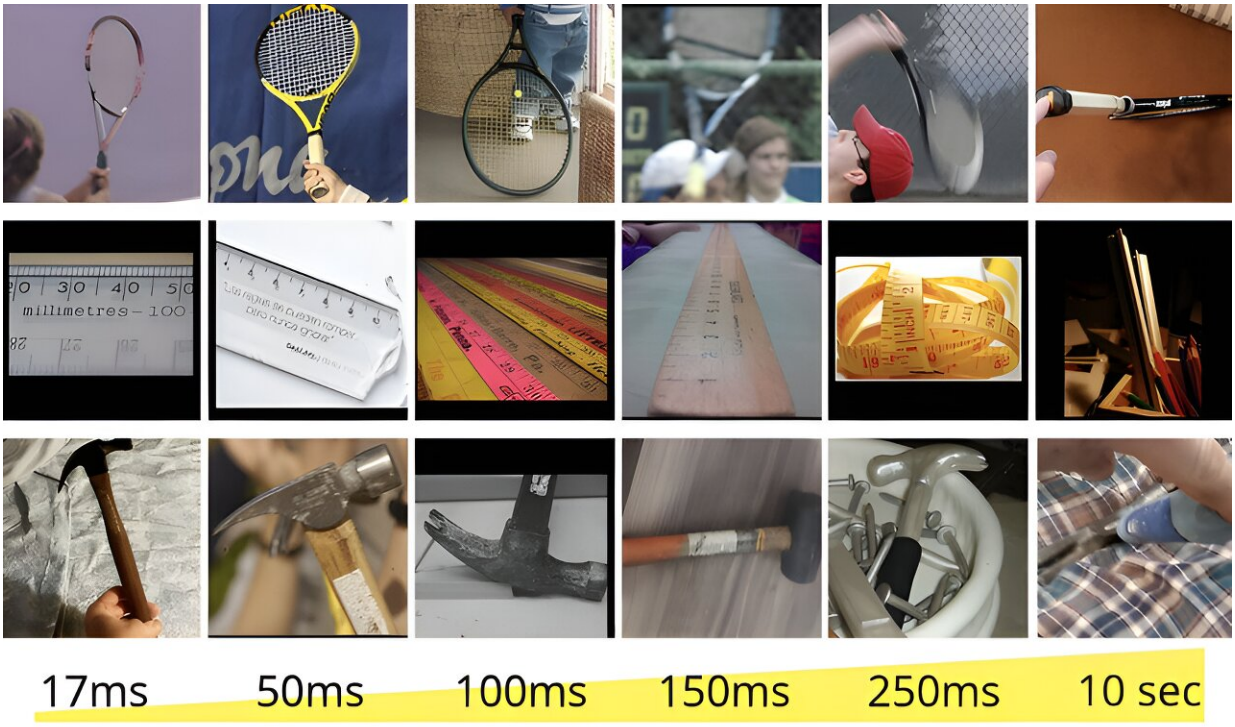
Image recognition accuracy: An unseen challenge confounding today's AI

December 15 2023, by Rachel Gordon



MVT, minimum viewing time, is a dataset difficulty metric measuring the minimum presentation time required for an image to be recognized. Researchers hope this metric will be used to evaluate models' performance and biological plausibility and guide the creation of new, more difficult datasets, leading to new computer vision techniques that perform better in real life. Credit: Images courtesy of the researchers/MIT CSAIL.

Imagine you are scrolling through the photos on your phone and you come across an image that at first you can't recognize. It looks like maybe something fuzzy on the couch; could it be a pillow or a coat? After a couple of seconds it clicks—of course! That ball of fluff is your friend's cat, Mocha. While some of your photos could be understood in an instant, why was this cat photo much more difficult?

MIT Computer Science and Artificial Intelligence Laboratory (CSAIL) researchers were surprised to find that despite the critical importance of understanding visual data in pivotal areas ranging from health care to transportation to household devices, the notion of an image's recognition difficulty for humans has been almost entirely ignored.

One of the major drivers of progress in deep learning-based AI has been datasets, yet we know little about how data drives progress in large-scale deep learning beyond that bigger is better.

In real-world applications that require understanding visual data, humans outperform object recognition models despite the fact that models perform well on current datasets, including those explicitly designed to challenge machines with debiased images or distribution shifts.

This problem persists, in part, because we have no guidance on the absolute difficulty of an image or dataset. Without controlling for the difficulty of images used for evaluation, it's hard to objectively assess progress toward human-level performance, to cover the range of human abilities, and to increase the challenge posed by a dataset.

To fill in this knowledge gap, David Mayo, an MIT Ph.D. student in electrical engineering and computer science and a CSAIL affiliate, delved into the deep world of image datasets, exploring why certain images are more difficult for humans and machines to recognize than others.

"Some images inherently take longer to recognize, and it's essential to understand the brain's activity during this process and its relation to machine learning models. Perhaps there are complex neural circuits or unique mechanisms missing in our current models, visible only when tested with challenging visual stimuli. This exploration is crucial for comprehending and enhancing machine vision models," says Mayo, a lead author of a new paper on the work.

This led to the development of a new metric, the "minimum viewing time" (MVT), which quantifies the difficulty of recognizing an image based on how long a person needs to view it before making a correct identification.

Using a subset of ImageNet, a popular dataset in machine learning, and ObjectNet, a dataset designed to test object recognition robustness, the team showed images to participants for varying durations from as short as 17 milliseconds to as long as 10 seconds and asked them to choose the correct object from a set of 50 options.

After over 200,000 image presentation trials, the team found that existing test sets, including ObjectNet, appeared skewed toward easier, shorter MVT images, with the vast majority of benchmark performance derived from images that are easy for humans.

The project identified interesting trends in model performance—particularly in relation to scaling. Larger models showed considerable improvement on simpler images but made less progress on more challenging images. The CLIP models, which incorporate both language and vision, stood out as they moved in the direction of more human-like recognition.

"Traditionally, object recognition datasets have been skewed towards less-complex images, a practice that has led to an inflation in model

performance metrics, not truly reflective of a model's robustness or its ability to tackle complex visual tasks. Our research reveals that harder images pose a more acute challenge, causing a distribution shift that is often not accounted for in standard evaluations," says Mayo.

"We released image sets tagged by difficulty along with tools to automatically compute MVT, enabling MVT to be added to existing benchmarks and extended to various applications. These include measuring test set difficulty before deploying real-world systems, discovering neural correlates of image difficulty, and advancing object recognition techniques to close the gap between benchmark and real-world performance."

"One of my biggest takeaways is that we now have another dimension to evaluate models on. We want models that are able to recognize any image even if—perhaps especially if—it's hard for a human to recognize. We're the first to quantify what this would mean. Our results show that not only is this not the case with today's state of the art, but also that our current evaluation methods don't have the ability to tell us when it is the case because standard datasets are so skewed toward easy images," says Jesse Cummings, an MIT graduate student in electrical engineering and computer science and co-first author with Mayo on the paper.

From ObjectNet to MVT

A few years ago, the team behind this project identified a significant challenge in the field of machine learning: Models were struggling with out-of-distribution images or images that were not well-represented in the training data. Enter ObjectNet, a dataset comprised of images collected from real-life settings.

The dataset helped illuminate the performance gap between machine

learning models and human recognition abilities by eliminating spurious correlations present in other benchmarks—for example, between an object and its background. ObjectNet illuminated the gap between the performance of machine vision models on datasets and in real-world applications, encouraging use for many researchers and developers—which subsequently improved model performance.

Fast forward to the present, and the team has taken their research a step further with MVT. Unlike traditional methods that focus on absolute performance, this new approach assesses how models perform by contrasting their responses to the easiest and hardest images.

The study further explored how image difficulty could be explained and tested for similarity to human visual processing. Using metrics like c-score, prediction depth, and adversarial robustness, the team found that harder images are processed differently by networks. "While there are observable trends, such as easier images being more prototypical, a comprehensive semantic explanation of image difficulty continues to elude the scientific community," says Mayo.

In the realm of health care, for example, the pertinence of understanding visual complexity becomes even more pronounced. The ability of AI models to interpret medical images, such as X-rays, is subject to the diversity and difficulty distribution of the images. The researchers advocate for a meticulous analysis of difficulty distribution tailored for professionals, ensuring AI systems are evaluated based on expert standards rather than layperson interpretations.

Mayo and Cummings are currently looking at neurological underpinnings of visual recognition as well, probing into whether the brain exhibits differential activity when processing easy versus challenging images. The study aims to unravel whether complex images recruit additional brain areas not typically associated with visual processing, hopefully

helping demystify how our brains accurately and efficiently decode the visual world.

Toward human-level performance

Looking ahead, the researchers are not only focused on exploring ways to enhance AI's predictive capabilities regarding image difficulty. The team is working on identifying correlations with viewing-time difficulty in order to generate harder or easier versions of images.

Despite the study's significant strides, the researchers acknowledge limitations, particularly in terms of the separation of object recognition from visual search tasks. The current methodology concentrates on recognizing objects, leaving out the complexities introduced by cluttered images.

"This comprehensive approach addresses the long-standing challenge of objectively assessing progress towards human-level performance in object recognition and opens new avenues for understanding and advancing the field," says Mayo.

"With the potential to adapt the Minimum Viewing Time difficulty metric for a variety of visual tasks, this work paves the way for more robust, human-like performance in object recognition, ensuring that models are truly put to the test and are ready for the complexities of real-world visual understanding."

"This is a fascinating study of how human perception can be used to identify weaknesses in the ways AI vision models are typically benchmarked, which overestimate AI performance by concentrating on easy images," says Alan L. Yuille, Bloomberg Distinguished Professor of Cognitive Science and Computer Science at Johns Hopkins University, who was not involved in the paper.

"This will help develop more realistic benchmarks leading not only to improvements to AI but also make fairer comparisons between AI and human perception."

"It's widely claimed that computer vision systems now outperform humans, and on some benchmark datasets, that's true," says Anthropic technical staff member Simon Kornblith Ph.D. '17, who was also not involved in this work.

"However, a lot of the difficulty in those benchmarks comes from the obscurity of what's in the images; the average person just doesn't know enough to classify different breeds of dogs. This work instead focuses on images that people can only get right if given enough time. These images are generally much harder for computer vision systems, but the best systems are only a bit worse than humans."

More information: Paper: How hard are computer vision datasets? Calibrating dataset difficulty to viewing time

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology

Citation: Image recognition accuracy: An unseen challenge confounding today's AI (2023, December 15) retrieved 21 May 2024 from https://techxplore.com/news/2023-12-image-recognition-accuracy-unseen-confounding.html