Human versus machine: comparing visual object recognition systems on a level playing field.

Nicolas Pinto, Najib J. Majaj, Youssef Barhomi, Ethan A. Solomon, David D. Cox & James J. DiCarlo

McGovern Inst/Dept of Brain & Cog Sci, MIT
Rowland Institute, Harvard

It is received wisdom that biological visual systems easily outmatch current artificial systems at complex visual tasks like object recognition. But have the appropriate comparisons been made? Because artificial systems are improving every day, they may surpass human performance some day. We must understand our progress toward reaching that day, because that success is one of several necessary requirements for "understanding" visual object recognition. How large (or small) is the difference in performance between current state-of-the-art object recognition systems and the primate visual system?

In practice, the performance comparison of any two object recognition systems requires a focus on the computational crux of the problem and sets of images that engage it. Although it is widely believed that tolerance ("invariance") to identity-preserving image variation (e.g. variation in object position, scale, pose, illumination) is critical, systematic comparisons of state-of-the-art artificial visual representations almost always rely on "natural" image databases that can fail to probe the ability of a recognition system to solve the invariance problem [Pinto et al.]. Thus, to understand how well current state-of-the-art visual representations perform relative to each other, relative to low-level neuronal representations (e.g. retinal-like and V1-like), and relative to high-level representations (e.g. human performance), we tested all of these representations on a common set of visual object recognition tasks that directly engage the invariance problem.

Specifically, we used a synthetic testing approach that allows direct engagement of the invariance problem, as well as knowledge and control of all the key parameters that make object recognition challenging. We successfully re-implemented a variety of state-of-the-art visual representations, and we confirmed the high published performance of all of these state-of-the-art representations on large, complex "natural'' image benchmarks. Surprisingly, we found that most of these representations were weak on our simple synthetic tests of invariant recognition, and only high-level biologically-inspired representations showed performance gains above the neuroscience "null" representation (V1-like).

While in aggregate, we found that the performance of these state-of-the-art representations pales in comparison to human performance, humans and computers seem to fail in different and potentially enlightening ways when faced with the problem of invariance. We also show how our synthetic testing approach can more deeply illuminate the strengths and weaknesses of different visual representations and thus guide progress on invariant object recognition.