Efficiency of object recognition networks on an absolute scale

Richard F. Murray (rfm@yorku.ca)
Devin H. Kehoe (dhkehoe@gmail.com)

Department of Psychology and Centre for Vision Research, York University
4700 Keele Street, LAS 0003J, Toronto, Ontario, Canada, M3J 1P3

Abstract:

Deep neural networks have made rapid advances in object recognition, but progress has mostly been made through experimentation, with little guidance from normative theories. Here we use ideal observer theory and associated methods to compare current network performance to theoretical limits on performance. We measure network performance and ideal observer performance on a modified ImageNet task, where model observers view samples from a limited number of object categories, in several levels of external white Gaussian noise. We find that although current networks achieve 90% performance or better on the standard ImageNet task, the ideal observer performs vastly better on the more limited task we consider here. The networks' "calculation efficiency", a measure of the extent to which they use all available information to perform a task, is on the order of 10^{-5}, an exceedingly small value. We consider reasons why efficiency may be so low, and outline further uses of ideal obsevers and noise methods to understand network performance.

Keywords: object recognition; neural networks; ideal observers; noise masking; efficiency

Introduction

Deep neural networks (DNNs) have made rapid advances in many areas of computer vision, including object recognition. This progress has been made largely through experimentation and intuition, with little guidance from normative theory. Here we show that ideal observers and noise masking methods from visual psychophysics are useful theoretical tools for understanding networks trained for object recognition.

An ideal observer is a theoretical observer that achieves the best possible performance on a particular task (Geisler, 2011). On many tasks, perfect performance is unachievable, even in principle. A classic example is identifying faint signals in visual noise, where stimulus fluctuations mean that every strategy will sometimes lead to errors. Comparing human performance on such tasks to an ideal observer's performance can be highly informative. For example, people are better at some tasks than others, and ideal observer analyses show that some of these performance differences are not due to inefficiencies in visual processing, but to some tasks being intrinsically easier than others in non-obvious ways.

Noise masking methods have also been useful for characterizing performance on tasks where underlying processes are still poorly understood, such as identifying complex patterns (Pelli & Farell, 1999). One simple but powerful method is to measure contrast energy thresholds in a task, as a function of the variance of white noise added to the stimulus; this is a "noise masking function". If the observer is contrast-invariant (often approximately true for humans), then noise masking functions are straight lines, and the line's parameters have natural interpretations. The x-intercept is "equivalent input noise", a measure of the observer's internal noise. The slope is a measure of "calculation efficiency", the extent to which the observer uses all relevant stimulus information when choosing a response.

Many DNNs for object identification are well-suited for noise masking methods. They are contrast-invariant, as the stimulus is normalized before processing, so their noise masking functions are linear. Furthermore, they have no internal noise, so calculation efficiency is the main parameter of interest.

Experiment

We measured performance in noise for several well-known networks trained on the ImageNet dataset (Deng et al., 2009). We chose ImageNet object identification because it is a mature research area where several architectures achieve human-level accuracy, so it is an interesting test case for examining theoretical limits of performance.

We modified the standard ImageNet object identification task to make it amenable to ideal observer analysis. We randomly chose n=10 categories from the 1000 available. In each category we randomly chose m=50 sample images. The stimulus on each trial was a randomly chosen image (out of nm=500 total) in pixelwise additive white Gaussian noise of contrast variance σ^2=0.01, 0.02, 0.03, and 0.04. The model observer identified which of the *n* categories the stimulus belonged to.

Results

Figure 1 shows proportion correct for three DNNs and the ideal observer, as a function of signal contrast, in Gaussian white noise with σ=0.05. The ideal observer is dramatically better than the DNNs, even though the networks achieve up to 90% correct performance on the standard ImageNet task. The average signal contrast threshold (defined as the contrast required for 55% correct responses, estimated from a cumulative normal fit) for DNNs is 0.22, whereas for the ideal observer it is 0.0016, a more than hundredfold difference.

Figure 2 shows noise masking functions for the same DNNs and ideal observer. As expected, the plots are linear with zero x-intercept. An observer's calculation efficiency is given by the ratio of the ideal observer's slope to the observer's slope, and here these ratios are 2.8 x 10^{-5} (AlexNet), 7.6 x 10^{-5} (VGG), and 5.9 x 10^{-5} (ResNet). These values show that the DNNs use only a tiny amount of the available stimulus information.

Discussion

DNNs can achieve high efficiency when trained on a task with a small number of stimuli (Reith & Wandell, 2020). Here we find that in tasks with more alternatives, their efficiency can be startlingly low. This may be partly because the ideal observer is perfectly tuned to the stimuli in the task at hand, and it exploits all differences between the stimuli to find the optimal response. In this sense, it is by definition an over-fitting algorithm that is not expected to generalize well to new samples. A useful development in ideal observer theory would be a measure of the extent to which an observer uses all category-relevant information to perform a task.

An important goal for future work will be to compare efficiency of human observers and DNNs. Proportion correct for human observers depends on their calculation efficiency, but also on internal noise. It would be informative to measure noise masking functions for human observers in this modified ImageNet task, in order to compare human and DNN calculation efficiency, since unlike human observers, DNN performance is not limited by internal noise.

It will also be informative to degrade the ideal observer's performance using factors that affect DNNs and human observers, such as uncertainty about signal position and contrast. Such experiments may establish the limiting factors on performance of DNNs and human observers, and provide a deeper understanding of similarities and differences between biological and computer vision.

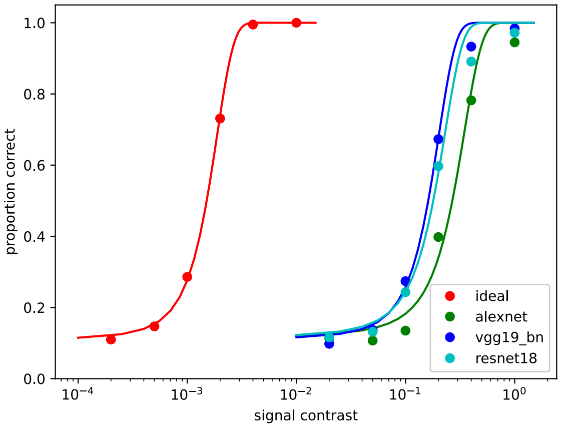

Figure 1: Psychometric functions for DNNs and ideal observer in Gaussian white noise (σ=0.05)

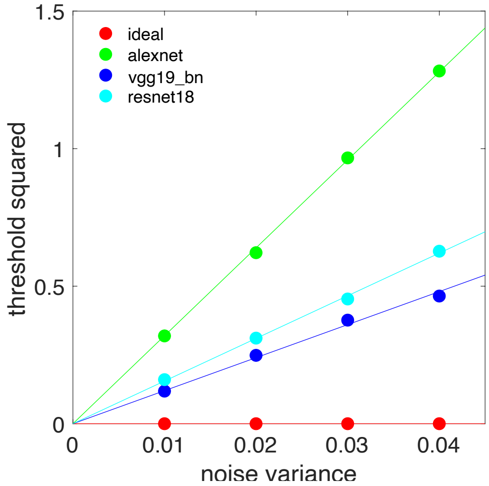

Figure 2: Noise masking functions for DNNs and ideal observer

References

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Li, F.-F. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*.

Geisler, W. S. (2011). Contributions of ideal observer theory to vision research *Vision Research, 51,* 771-781.

Pelli, D. G, & Farell, B. (1999). Why use noise? *Journal of the Optical Society of America A, 16,* 647-653.

Reith, F. H., & Wandell, B. A. (2020). A convolutional neural network reaches optimal sensitivity for detecting some, but not all, patterns. *IEEE Access, 8,* 213522-213530.