This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3040235, IEEE Access

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version October 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

A convolutional neural network reaches optimal sensitivity for detecting some, but not all, patterns

FABIAN H. REITH[1], BRIAN A. WANDELL[2]

[1]F. H. Reith is with Psychology Department, Stanford University, Stanford, CA 94305, USA (e-mail: Fabian.H.Reith@gmail.com)
[2]B. A. Wandell is with Psychology Department, Stanford University, Stanford, CA 94305, USA (e-mail: wandell@stanford.edu)

Corresponding author: Fabian H. Reith (e-mail: Fabian.H.Reith@gmail.com).

ABSTRACT We investigate the spatial contrast-sensitivity of modern convolutional neural networks (CNNs) and a linear support vector machine (SVM). To measure performance, we compare the CNN contrast sensitivity across a range of patterns with the contrast sensitivity of a Bayesian ideal observer (IO) with the signal-known-exactly and noise-known-statistically. A ResNet-18 reaches optimal performance for harmonic patterns, as well as several classes of real world signals including faces. For these stimuli the CNN substantially outperforms the SVM. We further analyze the case in which the signal might appear in one of multiple locations and found that CNN spatial sensitivity continues to match the IO. However, the CNN sensitivity is far below optimal at detecting certain complex texture patterns. These measurements show that CNNs spatial contrast-sensitivity differs markedly between spatial patterns. The variation in spatial contrast-sensitivity may be a significant factor, influencing the performance level of an imaging system designed to detect low contrast spatial patterns.

INDEX TERMS convolutional neural networks, deep learning, ideal observer, image systems, ResNet, signal detection, support vector machines

I. INTRODUCTION

DEEP convolutional neural networks (CNNs) - comprising a stack of convolutional layers connected by non-linearities and skip connections - have become an important computational tool. A network instance is defined by a large number of parameters that define the connection and computations. These parameters can be set by training, typically using back propagation, on a large number of examples. Much of the excitement about CNNs arises because a network trained on semantic categories, such as the texture of leather or a human face, generalizes well to new example images. The generalization accuracy far exceeds prior art and matches human accuracy on noise-free, undistorted images [1,2]. Furthermore, region proposal networks can locate the position of objects within these semantic categories anywhere in an image [3–5].

In addition to semantic classification, convolutional and related networks are being applied to image systems tasks including denoising, image reconstruction, super-resolution, pattern detection, part inspection and camera co-design [6–12]. When using a tool such as a CNN to analyze or design an imaging system, it is important to understand the limitations of the tool itself. An important limitation of an imaging system is its spatial contrast-sensitivity. Assessments of spatial sensitivity are a critical part of evaluating image systems performance. There are well-established methods for defining the spatial sensitivity of many critical components of imaging systems, such as lenses, pixel geometry, photon noise and electrical noise [13].

This paper introduces a metric to assess the spatial contrast-sensitivity limits of the CNN component of an imaging system: we compare the CNN performance to the performance of an ideal observer (equivalently, the Likelihood Ratio test described by the Neyman-Pearson Lemma). The ideal observer (IO) has a rigorous formal definition for the signal-known-exactly and noise-known-statistically case. We evaluate system spatial contrast-sensitivity by creating stimuli with known signals and statistically-known noise, and we compare the CNN performance on these stimuli with the IO performance. We also compare the performance of another important but simpler machine learning algorithm, the support vector machine (SVM).

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3040235, IEEE Access

Each CNN we evaluated has higher sensitivity to certain types of spatial patterns than others. The CNNs reach ideal sensitivity for some patterns, but in some cases the sensitivity is up to 5x lower than IO and sometimes even lower than SVM sensitivity. The approach we introduce and the experiments we describe should be helpful in assessing the CNN component of an imaging system for detection applications in vision science, astronomy, and medical imaging [14–16].

II. BACKGROUND

Convolutional neural networks have been particularly useful in fields where images are of central importance. Biomedical imaging has been a particularly active area - CNNs can be trained to identify low contrast targets that are revealed by a wide array of imaging modalities for diagnosis and monitoring [17,18]. CNN technology is also becoming important in Ophthalmology [19] and Vision Science [20,21].

There are many different uses for the CNNs. The best known application is semantic classification, in which images are classified based on their content. In a second application, direct comparisons are made between responses of trained networks and responses of neurons [21,22]. In a third application, investigators compare network and human perceptual performance, particularly with respect to stimulus generalization [20,23]. In a fourth application, most closely related to this paper, CNNs are used to detect spatial patterns in images. In this case, the CNN performance has been compared with the Hotelling observer [16,24].

Comparing system performance with respect to the absolute limits as determined by the physics of light and image formation is an important vision science approach. For example, in a classic study of the absolute sensitivity of vision, investigators found that the rod photodetectors are capable of responding to individual photons - performing at the absolute limits of light sensitivity [25,26]. As vision science evolved to measuring image contrast, rather than absolute light levels, investigators compared contrast sensitivity with ideal observers that were limited by photon noise and physiological optics [27–29].

Detecting a contrast pattern in noise has applications in many fields. The pattern of interest may vary considerably across applications, from tumors to textures to faces. To detect a pattern, a CNN learns internal representations of a particular class of spatial patterns. The CNN architecture may be closer to ideal observer performance for some patterns than others. Exploring performance over a range of spatial patterns can identify the strengths and limits of a particular CNN architecture.

Comparisons with ideal observer performance are useful for assessing and understanding CNN performance and guiding the direction of future investigations. For example, suppose a CNN performs at the theoretical limit when detecting certain types of patterns. In that case future research focus should not aim to improve CNN sensitivity but rather might aim to simplify computation and power consumption. Or one might aim to reach ideal performance using fewer training samples.

III. CONTRIBUTIONS

- We show that a modern CNN (ResNet) can be trained to detect certain spatial stimuli in the presence of Poisson noise (harmonics, faces, others) at an accuracy level that matches the sensitivity of an ideal observer.
- For other stimuli (certain textures) the asymptotic ResNet performance remains substantially lower than the ideal observer or even SVM performance; performance is best for stimuli with high spatial correlation.
- We show that the detection performance differs between CNN architectures, and the spatial sensitivity of a CNN can meaningfully impact the performance of an imaging system.

IV. METHODS

A. IMAGE SIMULATION

Test and training images were created using a simulation of a simple camera with diffraction-limited optics and a sensor with Poisson noise. The sensor images were calculated using the open-source and freely available software, ISETCam[1] [30–32]. Unless stated otherwise, the stimuli were simulated as being presented on a uniform background with a mean level of about 300 photons per pixel per capture. This level is typical of many imaging applications.

CNN sensitivity was analyzed using an input-referred measurement: We calculated stimulus detection accuracy for a range of logarithmically spaced performance levels, sweeping out a performance versus contrast curve. We then estimate the contrast level needed to obtain 75% correct detection in an present-absent discrimination. For most spatial patterns contrast was defined as the peak stimulus intensity minus minimum intensity divided by twice the mean intensity. In some cases, the contrast was defined by the standard deviation of the spatial pattern. The source code for creating the stimuli and for training and evaluating the networks can be downloaded from GitHub[2] .

1) Harmonics and Textures

The inputs to the CNN were simulated image sensor data. The simulations calculated a camera's sensor response from a planar scene defined by its spatial-spectral radiance (e.g., a harmonic pattern at some contrast, frequency, phase and orientation). The scene has a horizontal field of view of 10 deg, sampled at 512 rows and columns, and 31 wavelengths (400-700 nm with 10 nm spacing). We modeled the imaging lens as diffraction limited (f/# = 4) with a focal distance of 3.9 mm. The monochrome sensor was ideal (no electronic noise) with a pixel size of 2.8 microns, approximately equal to the full-width half maximum of the diffraction limited lens (2.4 microns). In this configuration the 10 deg scene spans 238

[1] https://github.com/iset/isetcam

[2] https://github.com/FabianRei/optimal_networks

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3040235, IEEE Access

x 238 sensor pixels and the Nyquist sampling frequency for the sensor is approximately 119 cycles/image. We changed the sensor to 256x256 pixels for cellular automata in order to match the distinct 256x256 pixel values of the cellular automaton itself. The sensor image data include only Poisson noise, which is the classic description of photon absorptions in an electronic device [33].

2) Face Stimuli

Face images were taken from the MIT-CBCL database[3] [34]. We converted these images to a contrast image (mean of zero) and added each to a uniform gray background. The face contrast was measured by its standard deviation, and set to 0.7071, which matches the mean and standard deviation of a harmonic pattern with a contrast of one. We simulated presenting this monochrome image on a display monitor in which each pixel emits an equal photon spectral radiance. The scene radiance was adjusted so that the mean number of photons captured by each pixel was close to 300.

3) Cellular Automaton Textures

We generated complex textures using a cellular automaton method [35]. We scale the scene resolution to 256 x 256, the resolution of the automaton we create, and slightly increase the lens field of view. This way, each pixel within the scene reaches exactly one pixel of the simulated sensor. For the textures the mean and standard deviation of the images were adjusted as we did for the face stimuli (scene radiance standard deviation of 0.7071; mean scene radiance set to create an average of 300 photons per pixel).

B. IDEAL OBSERVER

The neural network was compared to an ideal observer with signal-known-exactly and background-known-statistically. The number of electrons at each position is given by a Poisson distribution [36], whose rate parameter λ is equal to the intensity of the signal at each position in the image:

$$P(N) = \frac{exp(-\lambda)\lambda^N}{N!} \tag{1}$$

The ideal observer chooses the more likely signal based on a maximum likelihood calculation. For a candidate signal in noise, θ, measured independently at each pixel, the likelihood is the product of the Poisson density scaled by the a priori likelihood of the signal:

$$L(\theta) = P(\theta)\prod_{i=1}^{p}(P(N_i|\theta) \tag{2}$$

For computational simplicity it is usual to calculate the log likelihood:

$$LL(\theta) = \log(P(\theta)) + \sum_{i=1}^{p}\log(P(N_i|\theta)) \tag{3}$$

[3]http://cbcl.mit.edu/software-datasets/heisele/facerecognition-database.html

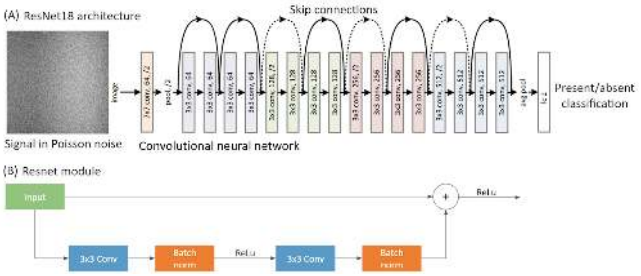

FIGURE 1. The ResNet CNN architecture [4]. (A) The input is processed through 18 stages; many of these stages include connections that transmit the input to a later stage through a skip connection (dashed implies resizing). The final stage is a fully connected layer that provides a classification decision – signal present or absent. The text in each of the stages describes its key properties: N x N conv is the kernel size; the next integer is the number of kernels; if present /N describes the spatial sub-sampling (stride). In our implementation the last fully connected layer only has 2 output classes (signal vs. noise) though in many applications this layer can be quite large. (B) The key concept of the network can also be described as comprising 8 modules. Each module performs a standard set of operations: convolution, batch normalization, half wave rectification (ReLu), convolution, batch norm, skip connection sum, and ReLu.

No training is necessary to implement the ideal observer. When there are N different signals, the system selects the most likely of these given the data. This algorithm performs optimally given the available information [28].

C. SUPPORT VECTOR MACHINE

Support vector machines (SVMs) were introduced by [37] under the slightly different name 'Support-Vector Networks'. The widely used linear SVM uses training data to learn a support vector such that the value of the inner product between this vector and a data sample decides the classification (e.g., signal vs. noise). A linear SVM separating two classes implicitly defines a hyperplane separating the two classes. To solve nonlinear classification tasks it is possible to use a nonlinear kernel, which is an extension of the dot product, as described by [38].

We use the linear support vector classifier implementation by the Python library Scikit-learn [39], based on the libsvm implementation [40]. The SVM classifier optimizes for the hinge loss [41], which finds the maximum margin classification. The SVM is optimized via a SMO-type decomposition method proposed in [42]. We set the maximum iterations performed to 1000, unless the convergence tolerance criterion of 0.001 is reached [40].

D. CONVOLUTIONAL NEURAL NETWORK

We used a ResNet network architecture because of its high quality [4]. The ResNet comprises multiple modules that each perform a convolution, batch normalization, and nonlinear operation (rectified linear unit). The network also includes skip connections.

If not declared differently, the architecture was a ResNet-18 [4], which has a good trade-off between speed and accuracy for our experiments. We use the PyTorch [43] implementation with a few minor adjustments. We changed the first convolution layer to account for the fact that the sensor

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3040235, IEEE Access

data are monochrome. We also replaced the average pooling layer through the PyTorch implementation of an adaptive average pooling layer [44] to allow the network to be more flexible to variations in image size. The network weights were randomly initialized by the default PyTorch initialization method, a method known as He Initialization [45]. This algorithm specifically addresses rectifier nonlinearities. The last layer of the ResNet-18, a fully connected layer, is replaced by a smaller layer to accommodate the very small output dimension (binary choice).

The data consists of one scene per class that has random Poisson noise. There is no inherent limit to the epoch size and this parameter can be set arbitrarily. We used 10,000 samples to define one epoch. The batch size was 32, and we used Adam [46] as the gradient-based optimization function.

The outputs of the neural network are normalized into a probability distribution via the softmax function. These processed outputs are then used to calculate the loss function. For this, we use cross-entropy loss where y_c is the ground truth and $\hat{y}_c$ is the model output for class c.

$$L(y, \hat{y}) = -\sum_{c=1}^{M} y_c \log(\hat{y}_c) \tag{4}$$

The initial learning rate is 1e-3 and after 10 epochs, the learning rate is decreased to 1e-4. After another 10 epochs, the CNN is trained with a learning rate of 1e-5. The network's performance is tested on 5,000 data samples. Seeds are used to ensure the same random initialization of ResNet-18 on all experiments. Training data are generated with the same, specified, seeds to initiate the random number generator.

The ResNet-18 is trained using a parallel algorithm to permit the server to use all available GPUs. While each neural network runs on one specific GPU, each GPU runs multiple training and testing experiments in parallel. On the server used for training, there are six Nvidia GK210 graphics processors. On one GPU, training ResNet-18 with 300,000 data samples, generated in real-time, takes 1:18 hours.

E. NETWORK PERFORMANCE

1) Metrics

The detection experiments are two-class classification problems. We vary the size of the signal contrast, position, or orientation and measure classification performance by the hit (true positive) false alarm rates of the IO, ResNet-18, and SVM. The network training was carried out for each stimulus at each contrast level. We estimate the discriminability between the two classes using d-prime (d') [47,48] from these two rates. Specifically, we calculate the z-scores (inverse of the standard normal cumulative distribution) for these rates and subtract the false alarm z-score from hit rate z-score:

$$d' = Z(hit\,rate) - Z(false\,alarm\,rate) \tag{5}$$

We manage extreme hit or false alarm rates (zero errors) by a small adjustment to the hit and false alarm rates [49]:

$$hit\,rate = \frac{0.5 + \sum hits}{1 + \sum hits + \sum misses} \tag{6}$$

$$false\,alarm\,rate = \frac{0.5 + \sum false\,alarms}{1 + \sum false\,alarms + \sum correct\,rejections} \tag{7}$$

Without this modification, a hit rate of 100% would result in a d' of infinity, given the false alarm rate is not at 100% as well. The modified equations for false alarm and hit rates provides a finite and only slightly biased underestimate of the true d' [50].

Given the mean number of signal photons in a detection task with only Poisson noise, d' can also be calculated by a formula that only requires the mean photon absorptions of both classes. In this formula, the sum of scaled Poisson random variables is approximated with normal density [51]:

$$d' = \frac{\sum_{i=1}^{n} (\beta_i - \alpha_i) \ln(\beta_i / a_i)}{[0.5 \sum_{i=1}^{n} (\alpha_i + \beta_i) \ln^2(\beta_i / \alpha_i)]^{1/2}}. \tag{8}$$

Our results show that the IO d', calculated via hit and false alarm rate, matches the theoretical d'. We also calculate the sensitivity of a discriminator.

In most analyses, we calculate how d' increases as the stimulus parameters - contrast, position shift, or angle - change. This produces a curve relating performance (d') to the stimulus parameter. In many analyses we summarize network sensitivity using an input-referred measure. Specifically, we calculate the contrast level, spatial shift or orientation angle needed to achieve d' = 1.5. The contrast, phase shift or angle metric is calculated by linearly interpolating the performance curve.

In certain select cases, we repeated the training experiments five times with different random seeds. This results in different training and test data, different neural network random states, as well as varied random states of the SVM. In these cases, we report the mean and standard deviation of the sensitivity measure. The size of the main effects we describe are many multiples of the estimated standard deviations.

2) Size of Training Data

ResNet-18 performance improves as training set size increases (Figure A1). The ResNet-18 reaches asymptote - in this case the maximum theoretical performance level - when the training set reaches 100,000 to 300,000 samples. The SVM performance reaches asymptote at a much smaller training set size, prior to the initial portion of the graph (about 10,000 training samples).

ResNet-18 performance is significantly better than that of the SVM after 10,000 training samples, continuing to rise up to the IO level at approximately 100,000 training samples. Based on these experiments, we used a training set size of 10,000 samples for the SVM and 300,000 samples for ResNet-18.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3040235, IEEE Access

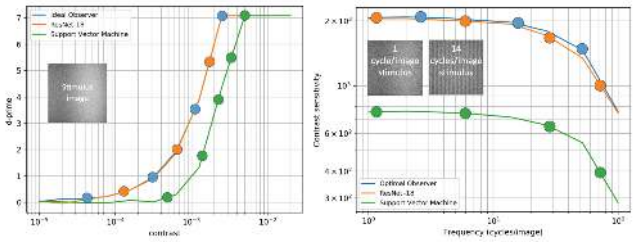

FIGURE 2. Comparison of IO, ResNet and SVM detection performance for a harmonic presented in Poisson noise. (A) Performance (d') increases as a function of contrast. The blue and orange points are representative points along the two detection curves. The IO and ResNet performance are very similar, so the underlying blue and orange curves overlap. (B) Contrast-sensitivity functions of the IO, ResNet-18 and SVM for spatial frequencies up to the sensor Nyquist frequency. Contrast sensitivity is the inverse of the contrast needed to achieve discrimination performance of d' = 1.5. Higher contrast sensitivity means performance is reached with lower contrast.

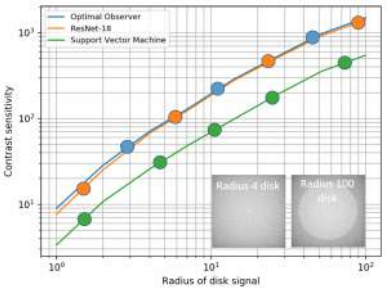

FIGURE 3. Contrast sensitivity to disks for IO, ResNet-18 and SVM. Detection performance for disks with sizes from radius 1 to radius 100. Disk contrast sensitivity is shown for a performance level of d' = 1.5. Other details as in Figure 2.

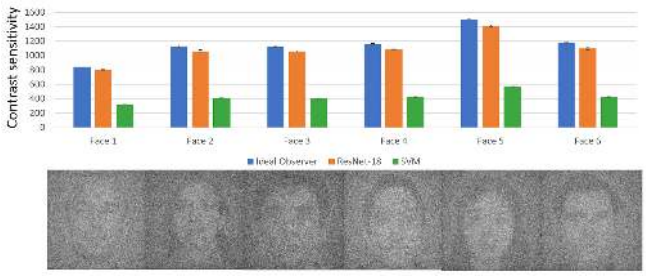

FIGURE 4. Contrast sensitivity to faces for IO, ResNet-18, and SVM. The first five graphs show detection of a single face. The sixth graph shows sensitivity to a collage comprising nine faces. Contrast sensitivity is shown for a performance level of d' = 1.5. Error bars are +/- 1 SD (N=5). Other details as in Figure 2.

V. RESULTS

First, we consider the detection of harmonics. We measure discriminability as a function of spatial frequency, position (phase shift) and orientation. Second, we measure signal detection based on signal size (disks of various sizes). Third, we consider a collection of biological images (faces). Fourth, we measure texture signals that are not compact in space or spatial frequency (white noise, cellular automata). Fifth, we analyze the detection performance for targets in which the signal may be present in one of multiple positions.

A. HARMONICS

1) Contrast

For all networks detection sensitivity (d') of a harmonic in Poisson noise increases with contrast. The ResNet-18 can be trained to achieve a performance that closely matches the ideal observer's performance and the SVM performance is about half a log (4x) unit less sensitive (Figure 2A).

We repeated these calculations for a range of harmonic spatial frequencies, extending to the Nyquist limit of the sensor (Figure 2B). The input-referred contrast sensitivity (1 over the contrast for d'=1.5) matched the performance of the ideal observer closely, being only only slightly lower than the IO, by an average of 2.86% (0.013 log10 units). The SVM contrast sensitivity was an average of 63.39% lower (0.44 log10 units) compared to IO.

B. DISKS

Detection sensitivity grows systematically with disk radius, approximately as the square root of the disk area (Figure 3). Deviations from this rule are present for small disks which are blurred by the optics and very large disks that span nearly the whole sensor. The ResNet-18 again approximates IO performance for all disk sizes tested, and the SVM is about half a log unit lower.

C. FACES

Disks and harmonic signals are very simple patterns compared to many natural objects. Hence, we decided to measure contrast sensitivity for an important and complex object, the human face (Figure 4).

The ResNet-18 contrast sensitivity is similar but slightly lower than the IO sensitivity. ResNet-18 contrast sensitivity is on average 5.87% lower than the IO sensitivity. This difference is slightly larger than the sensitivity difference using the harmonics. The SVM performance is about 1/3rd the sensitivity of the IO and ResNet-18 network.

D. TEXTURES

In addition to test stimuli (harmonics, disks) and natural objects (faces), there are applications in which the target is a texture pattern (see Discussion). We used cellular automata to generate an organized list of texture patterns [35,52]. We focused on rules which converge to a structured repetitive pattern (class 2) and rules in which the texture pattern remains random (class 3). We generated textures using four different class 2 rules and four different class 3 rules.

1) Class 2 Cellular Automata

Class 2 automata converge to a repetitive texture pattern. We suspect that CNNs might learn filters to identify repetitive patterns. To measure detection performance, we used experiments for four class 2 automata (rules 3, 57, 76 and 78). The contrast sensitivity for these patterns is slightly higher for the IO than ResNet-18, and substantially higher than SVM

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3040235, IEEE Access

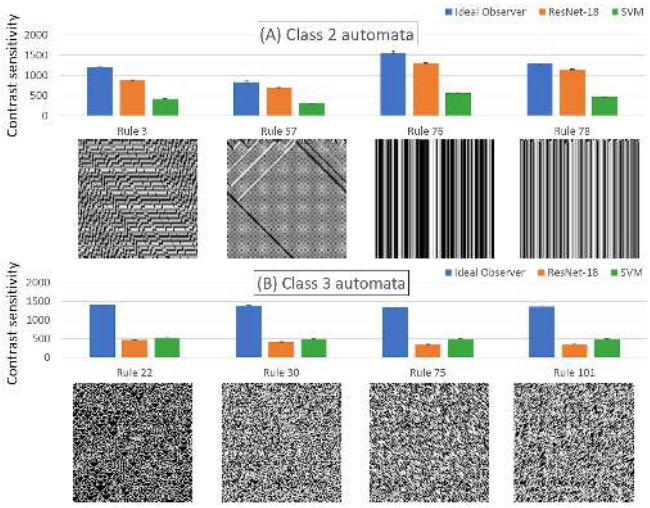

FIGURE 5. Contrast sensitivity for class 2 cellular automata (A). ResNet-18 contrast sensitivity is lower than IO; the sensitivity of SVM is around one third of IO sensitivity. The highest sensitivity is for rule 76, followed by rule 78. One-dimensional patterns are easiest to detect. Contrast sensitivity for class 3 cellular automata (B). IO contrast sensitivity is three-fold higher than either ResNet-18 or SVM. In several cases, SVM contrast sensitivity to these patterns exceeds that of ResNet-18. Unlike the class 2 cellular automata, these textures are dense and not space-invariant. The variance of the contrast sensitivity is smaller than the measured difference in contrast sensitivity. Error bars are +/- 1 SD (N=5).

(Figure 5A).

Slightly worse performance is achieved for the other two automata. At rule 3, IO has a contrast sensitivity of 1213.31, while ResNet-18 reaches 861.14 and SVM achieves 438.02. IO contrast sensitivity for the rule 57 automaton is the lowest. Here, IO reaches 824.34, ResNet-18 achieves 688.76 and SVM reaches 298.32. Compared to IO, ResNet-18 performance drops by an average of 18.18%, while SVM performance drops by an average of 63.74%.

2) Class 3 Cellular Automata

Class 3 automata have a complex irregular spatial pattern (Figure 5B). We examine four class 3 automata (rules 22, 30, 75 and 101). The CNN sensitivity to these patterns is far from the sensitivity of the IO, dropping to the level of the SVM performance.

3) Block Randomization

In addition to the cellular automata, we produced texture patterns by randomizing the pixel positions in an existing image. We performed a series of experiments by block-wise scrambling the pixels in a one-cycle per image harmonic (Figure 6).

The IO performance is indifferent to the scrambling, as expected from the computational formula (Equation 1). Similarly, the SVM adjusts its critical vector and learns to detect the pattern with reordered pixels. The ResNet-18 sensitivity is substantially reduced by scrambling. The scrambled texture pattern does not repeat regularly across the image, and like the cellular automata in class 3, the ResNet-18 sensitivity is below the IO.

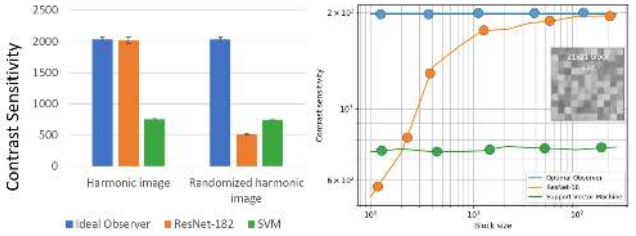

FIGURE 6. Performance for spatial randomization of frequency one harmonic signal. Panel (a) shows 70% reduction in ResNet-18 contrast sensitivity (compared to IO) for randomization of all pixel locations of harmonic signal (1x1 block). The contrast sensitivity is 20% lower than the SVM. The bar heights represent the mean of five runs with training data, test data and different random number seeds. Error bars are +/- 1 SD (N=5). Panel (b) displays contrast sensitivity for various block sizes.

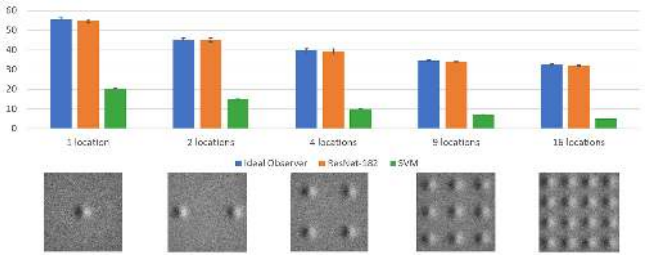

FIGURE 7. Detection performance of frequency one harmonic with Gabor for one or multiple locations. The signal examples show the signal in all its potential locations. In the signal case, the signal can be seen in exactly one location. The bar heights represent the mean of five runs with training data, test data and different random number seeds. Error bars are +/- 1 SD (N=5).

E. MULTIPLE TARGET POSITIONS

The ability to detect and localize a signal anywhere in a scene is one of the most important contributions of CNN technology [53]. We compare the CNN sensitivity with the ideal observer sensitivity to a simple stimulus (a Gabor patch) that might be presented at one of multiple possible locations (Figure 7). When there are N different locations, the ideal observer selects the most likely of these locations, or no signal, given the image data.

Introducing position uncertainty reduces the sensitivity of the ResNet, SVM and the IO. Although sensitivity declines, the ResNet-18 continues to match the IO performance. Both methods are about half as sensitive when the target can appear in 16 locations rather than one location. The SVM sensitivity declines by a larger fraction, becoming about one-fourth as sensitive as the number of possible positions increases to 16 from one.

F. COMPARISON TO VGG-16 AND ALEXNET

We compared the performance of ResNet-18 to two other CNNs, VGG-16 [54] and AlexNet [3]. We measured contrast sensitivity using a harmonic with one cycle per image (Figure A2, cf. Figure 2). The VGG-16 and AlexNet sensitivities

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3040235, IEEE Access

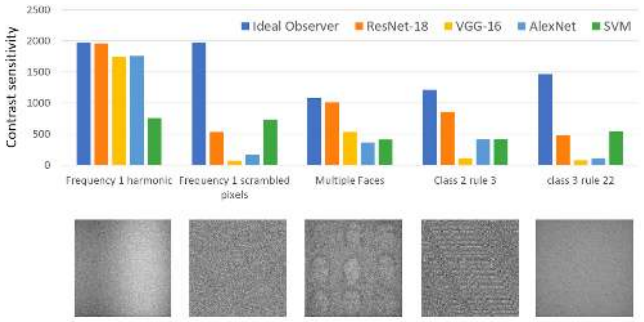

FIGURE 8. Detection sensitivity of IO, ResNet-18, VGG-16, AlexNet and SVM to six different stimuli that were also used in the main text. See the text for details.

are close to the IO for higher contrasts. But in both cases the IO performs above chance ($d' > 0$) at contrasts where the two networks are still at $d' = 0$. Even at the higher contrasts AlexNet sensitivity is slightly lower than IO sensitivity.

Next we explored network sensitivity to variations of the harmonic signal, the multiple faces signal, and two different cellular automata (Figure 8). All network hyperparameters were the same as for ResNet-18 with one exception: for the network solution to converge it was necessary to decrease the VGG-16 and AlexNet initial learning rates to 1e-5 rather than using the ResNet initial learning rate of 1e-3. It might be possible to find further improvements in VGG-16 and AlexNet performance by modifying other hyperparameters.

Like the ResNet, randomizing image pixel positions (1x1 blocks) causes VGG-16 and AlexNet sensitivity to drop significantly. The sensitivity of these two networks is also substantially lower for detecting multiple faces signal and the two types of automata signals. In several cases the SVM outperforms the VGG-16 and AlexNet CNNs.

VI. DISCUSSION

A. CNN SPATIAL SENSITIVITY

For many spatial patterns (harmonics, disks, faces), a ResNet-18 can be trained to detect a contrast pattern at the same sensitivity level as an ideal observer. The main requirement to achieve this optimal performance is a large number of training samples (more than 1e+5 samples).

The ResNet-18 spatial sensitivity for certain textures (class 3 cellular automata and block-scrambled images) is 2.5x lower than the ideal observer and comparable to the SVM. All of the networks have reduced sensitivity to these patterns. Network sensitivity to the repetitive textures, such as class 2 automata is higher than sensitivity to random class 3 automata textures.

There is also a substantial spatial sensitivity difference for the two other architectures (VGG-16 and AlexNet). For these architectures both class 2 and class 3 textures are detected poorly compared to ideal, and sensitivity to faces is only half that of the IO (Figure 8). It is worth noting that individual networks have their distinct spatial sensitivity profiles.

B. UNCERTAIN SIGNAL POSITION

An important value of CNNs is their ability to detect patterns even when the pattern's position is uncertain. We performed an initial analysis of the ResNet's ability to detect signals present at one of multiple positions and found that the CNN matches IO contrast sensitivity. The experiments examining the sensitivity when position is uncertain could be significantly expanded to include variations in the stimulus pattern, size and a systematic analysis of position bias. The methods in this paper - input-referred contrast measures and a comparison with the ideal observer - can provide meaningful numerical assessments for such evaluations.

C. ARCHITECTURE

We compared ResNet-18 detection sensitivity to other well-known CNN architectures, VGG-16 and AlexNet. Sensitivity to harmonics is comparable, but sensitivity to more complex signals (e.g., faces) is substantially lower for AlexNet and VGG-16; in several cases these networks are less sensitive than a linear SVM.

ResNet-18 contrast sensitivity is lower than the IO sensitivity when the stimuli comprise fine textures that do not repeat regularly across the image. Block randomization and cellular automata examples fit this pattern. Why limits performance on the class 3 cellular automata and block-scrambled images? The central difference between cellular automata of class 2 vs. class 3 and block-scrambled images is the image complexity. The repeating patterns of class 2 automata can be summarized by a shift-invariant representation compared to the non-repetitive class 3 and block-scrambled patterns. The stimulus structure's complexity matches the architecture of the CNN, which has a small number of weights compared to fully connected NNs. The number of weights needed by the IO to distinguish the signal from noise using is (256 x 256 x 2, row x col x classes). The initial stage of the ResNet-18 uses only 7x7x64 weights for low-level feature extraction which is just 2% of the IO weights. The total number of ResNet-18 weights is vastly larger (1.1e+7).

D. APPLICATIONS

There are signal detection applications whose signals resemble class 3 cellular automata (MRI k-space, which is the Fourier Transform of the image, skin rashes or retinal bleeding). The conventional CNN architectures limit performance for such signals. The tests in this paper can discriminate between different CNN architectures to determine which may be most effective for specific classes of signals. An advantage of our testing procedure is that it does not require large amounts of labeled data which is especially helpful for signal types that are only observed in certain clinical cases.

In addition to the benchmarking of existing CNN architectures, we hope that our tools will furthermore be helpful in the design of new, innovative CNNs that allow improved performance on non-standard signal types.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3040235, IEEE Access

E. CONCLUSION

We present a way to assess CNN performance by measuring performance with respect to a fundamental image science tool, the ideal observer. This approach quantifies how well a CNN architecture learns to detect signals of varying shapes and abstraction. As in other branches of image systems engineering, we hope that characterizing CNN architectures will help designers find the right deep learning algorithms for specific tasks. Just as we characterize the impact on spatial resolution of the lens, pixel sampling array, and electrical noise, we can characterize the impact of a CNN detection network.

ResNet, along with many other CNN architectures, was designed for semantic categorization and commonly tested with the categories in ImageNet. Compared to fully connected neural networks and transformer networks, the CNN architecture processes low-level features via convolutional filters which reduces the number of weights required. We evaluated several CNNs for signal detection sensitivity. For many signals we find that even in the presence of pixel-wise Poisson noise the ResNet CNN has the same sensitivity as an ideal observer. We conclude that current CNN architectures are able to detect signal types, such as the ones found in ImageNet, at near optimal levels.

Image systems may be designed to detect a wide range of spatial targets in applications spanning medical imaging and industrial inspection: from localized tumors and moles to a widespread rash. Some of these targets are not similar to the images in ImageNet. We find that ResNet's sensitivity to certain types of textures is substantially lower than ideal.

The high sensitivity of a CNN for identifying certain targets, but not others, should be a part of decision-making in image system design. The experiments in this paper are a start towards developing this technology. It would be useful to develop consensus methods that assess the spatial sensitivity profile of a CNN with respect to the target objects for each application.

APPENDIX

ACCURACY BASED ON SIZE OF TRAINING SET

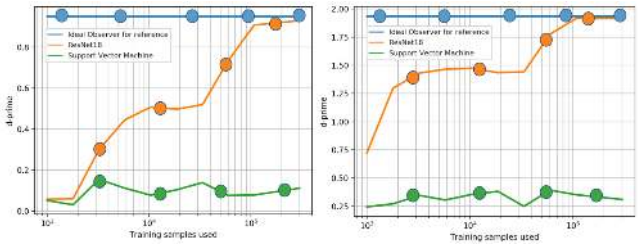

FIGURE A1. Increase in detectability (d') of a harmonic image in Poisson noise as a function of the number of training samples (horizontal axis). The two panels show performance for stimuli at two different contrasts: 3.2e-4 (left) and 6.3e-4 (right). Irrespective of training set size, ResNet-18 was trained for 9375 iterations with a batch size of 32. The ideal observer requires no training. The SVM reaches asymptotic performance before 1e+4 training samples. The ResNet performance increases until approximately 3e+5 training samples. As in the main text, colored disks are superimposed on the lines at every other measurement point, which is helpful when the ResNet and IO curves superimpose.

COMPARISON OF NETWORK CONTRAST-DEPENDENT ACCURACY

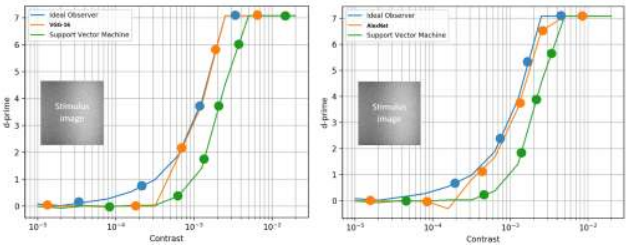

FIGURE A2. Detection performance of VGG-16 (left) and AlexNet (right) for a harmonic stimulus of frequency one. VGG-16 (A) approximates the IO for higher contrasts but does not reach IO performance at lower contrasts. AlexNet (B) is able to discriminate a low contrast signal slightly better than VGG-16, but never reaches IO performance for high contrast harmonic curves.

ACKNOWLEDGMENT

We thank David Donoho, Zhenyi Liu, Zheng Lyu, Laura Leal-Taixé, and Daniel Yamins for useful discussions.

REFERENCES

[1] Dodge S, Karam L. A Study and Comparison of Human and Deep Learning Recognition Performance Under Visual Distortions. arXiv [cs.CV]. 2017. Available from: http://arxiv.org/abs/1705.02498

[2] Geirhos R, Janssen DHJ, Schütt HH, Rauber J. Comparing deep neural networks against humans: object recognition when the signal gets weaker. arXiv preprint arXiv. arxiv.org; 2017; Available from: https://arxiv.org/abs/1706.06969

[3] Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, editors. Advances in Neural Information Processing Systems 25. Curran Associates, Inc.; 2012. p. 1097–105.

[4] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. arXiv [cs.CV]. 2015. Available from: http://arxiv.org/abs/1512.03385

[5] Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. arXiv [cs.CV]. 2015. Available from: http://arxiv.org/abs/1506.01497

[6] McCann MT, Jin KH, Unser M. Convolutional Neural Networks for Inverse Problems in Imaging: A Review. IEEE Signal Process Mag. 2017;34:85–95.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3040235, IEEE Access

[7] Ledig C, Theis L, Huszar F, Caballero J, Cunningham A, Acosta A, et al. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. arXiv [cs.CV]. 2016. Available from: http://arxiv.org/abs/1609.04802

[8] Karras T, Aila T, Laine S, Lehtinen J. Progressive Growing of GANs for Improved Quality, Stability, and Variation. arXiv [cs.NE]. 2017. Available from: http://arxiv.org/abs/1710.10196

[9] Jain V, Seung S. Natural Image Denoising with Convolutional Networks. In: Koller D, Schuurmans D, Bengio Y, Bottou L, editors. Advances in Neural Information Processing Systems 21. Curran Associates, Inc.; 2009. p. 769–76.

[10] Jackson AS, Bulat A, Argyriou V, Tzimiropoulos G. Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. Proceedings of the IEEE International Conference on Computer Vision. openaccess.thecvf.com; 2017. p. 1031–9.

[11] Schlemper J, Caballero J, Hajnal JV, Price A, Rueckert D. A Deep Cascade of Convolutional Neural Networks for MR Image Reconstruction. Information Processing in Medical Imaging. Springer International Publishing; 2017. p. 647–58.

[12] Liu Z, Shen M, Zhang J, Liu S, Blasinski H, Lian T, et al. A system for generating complex physically accurate sensor images for automotive applications. arXiv [cs.CV]. 2019. Available from: http://arxiv.org/abs/1902.04258

[13] Holst GC. CCD arrays, cameras, and displays. Citeseer; 1998.

[14] Wandell BA. Foundations of vision. Sunderland, MA: Sinauer Associates; 1995.

[15] Starck J, Murtagh F. Astronomical image and signal processing: looking at noise, information and scale. IEEE Signal Process Mag. 2001;18:30–40.

[16] Zhou W, Li H, Anastasio MA. Approximating the Ideal Observer and Hotelling Observer for binary signal detection tasks by use of supervised learning methods. IEEE Trans Med Imaging. 2019; Available from: http://dx.doi.org/10.1109/TMI.2019.2911211

[17] Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, et al. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? IEEE Trans Med Imaging. 2016;35: 1299–1312.

[18] Chen Y-C, Hong DJ-K, Wu C-W, Mupparapu M. The Use of Deep Convolutional Neural Networks in Biomedical Imaging: A Review. J Orofac Sci. 2019;11: 3.

[19] Sengupta S, Singh A, Leopold HA, Gulati T, Lakshminarayanan V. Ophthalmic diagnosis using deep learning with fundus images - A critical review. Artif Intell Med. 2020;102: 101758.

[20] Geirhos R, Rubisch P, Michaelis C, Bethge M, Wichmann FA, Brendel W. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv [cs.CV]. 2018. Available: http://arxiv.org/abs/1811.12231

[21] Yamins DLK, DiCarlo JJ. Using goal-driven deep learning models to understand sensory cortex. Nat Neurosci. 2016;19: 356–365.

[22] Tanaka H, Nayebi A, Maheswaranathan N, McIntosh L, Baccus S, Ganguli S. From deep learning to mechanistic understanding in neuroscience: the structure of retinal prediction. In: Wallach H, Larochelle H, Beygelzimer A, d' Alché-Buc F, Fox E, Garnett R, editors. Advances in Neural Information Processing Systems 32. Curran Associates, Inc.; 2019. pp. 8537–8547.

[23] Geirhos R, Janssen D, Schütt H, Bethge M, Wichmann F. Of Human Observers and Deep Neural Networks: A Detailed Psychophysical Comparison. Journal of Vision. 2017. p. 806. doi:10.1167/17.10.806

[24] Yao J, Barrett HH. Predicting human performance by a channelized Hotelling observer model. Mathematical Methods in Medical Imaging. International Society for Optics and Photonics; 1992. pp. 161–168.

[25] Hecht S, Shlaer S, Pirenne MH. ENERGY AT THE THRESHOLD OF VISION. Science. 1941;93: 585–587.

[26] Sakitt B. Counting every quantum. J Physiol. 1972;223: 131–150.

[27] Banks MS, Geisler WS, Bennett PJ. The physical limits of grating visibility. Vision Res. 1987;27: 1915–1924.

[28] Geisler WS. Contributions of ideal observer theory to vision research. Vision Res. 2011;51: 771–781.

[29] Cottaris NP, Jiang H, Ding X, Wandell BA, Brainard DH. A computational-observer model of spatial contrast-sensitivity: Effects of wave-front-based optics, cone-mosaic structure, and inference engine. J Vis. 2019;19: 8.

[30] Farrell JE, Xiao F, Catrysse PB, Wandell BA. A simulation tool for evaluating digital camera image quality. Image Quality and System Performance. International Society for Optics and Photonics; 2003. pp. 124–132.

[31] Farrell JE, Catrysse PB, Wandell BA. Digital camera simulation. Appl Opt. 2012;51: A80–90.

[32] Farrell JE, Wandell BA, editors. Image Systems Simulation. Handbook of Digital Imaging. Chichester, UK: John Wiley & Sons, Ltd; 2015. pp. 1–28.

[33] Schottky W. Über spontane Stromschwankungen in verschiedenen Elektrizitätsleitern. Ann Phys. 1918;362: 541–567.

[34] Weyrauch B, Heisele B, Huang J, Blanz V. Component-Based Face Recognition with 3D Morphable Models. 2004 Conference on Computer Vision and Pattern Recognition Workshop. 2004. pp. 85–85.

[35] Wolfram S. Statistical mechanics of cellular automata. Rev Mod Phys. 1983;55: 601–644.

[36] Snyder DL, Miller MI. Random point processes. j. Willey & Sons, New York. 1975.

[37] Cortes C, Vapnik V. Support-Vector Networks. Mach Learn. 1995;20: 273–297.

[38] Aizerman MA, Braverman EA, Rozonoer L. Theoretical foundations of the potential function method in pattern recognition learning. Automation and Remote Control,. 1964. pp. 821–837.

[39] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12: 2825–2830.

[40] Chang C-C, Lin C-J. LIBSVM: A Library for Support Vector Machines. ACM Trans Intell Syst Technol. 2011;2: 27:1–27:27.

[41] Rosasco L, Vito ED, Caponnetto A, Piana M. Are loss functions all the same? Neural Comput. 2004. Available: https://www.mitpressjournals.org/doi/abs/10.1162/089976604773135104

[42] Fan R-E, Chen P-H, Lin C-J. Working Set Selection Using Second Order Information for Training Support Vector Machines. J Mach Learn Res. 2005;6: 1889–1918.

[43] Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, et al. Automatic differentiation in PyTorch. 2017 [cited 15 Jan 2019]. Available: https://openreview.net/forum?id=BJJsrmfCZ

[44] Lin M, Chen Q, Yan S. Network In Network. arXiv [cs.NE]. 2013. Available: http://arxiv.org/abs/1312.4400

[45] He K, Zhang X, Ren S, Sun J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. arXiv [cs.CV]. 2015. Available: http://arxiv.org/abs/1502.01852

[46] Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv [cs.LG]. 2014. Available: http://arxiv.org/abs/1412.6980

[47] Stanislaw H, Todorov N. Calculation of signal detection theory measures. Behav Res Methods Instrum Comput. 1999;31: 137–149.

[48] Green DM, Swets JA. Signal Detection Theory and Psychophysics. Peninsula Pub.; 1988.

[49] Knoke D, Burke PJ, Burke PJ. Log-Linear Models. SAGE; 1980.

[50] Hautus MJ. Corrections for extreme proportions and their biasing effects on estimated values ofd'. Behav Res Methods Instrum Comput. 1995;27: 46–51.

[51] Geisler WS. Physical limits of acuity and hyperacuity. J Opt Soc Am A. 1984;1: 775–782.

[52] Wolfram S. A New Kind of Science (English Edition). Kindle. Wolfram Media, Inc.; 2016.

[53] Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R, editors. Advances in Neural Information Processing Systems 28. Curran Associates, Inc.; 2015. pp. 91–99.

[54] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv [cs.CV]. 2014. Available: http://arxiv.org/abs/1409.1556

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3040235, IEEE Access

FABIAN H. REITH graduated from the Technical University of Munich in 2019 and is now a PhD student at Humboldt University of Berlin. In 2019 he was a visiting research scholar at Stanford University. Reith's research centers on deep learning, vision science, spanning topics from signal processing, medical data analysis and personalized medicine.

BRIAN A. WANDELL is the first Isaac and Madeline Stein Family Professor. He joined the Stanford Psychology faculty in 1979 and is a member, by courtesy, of Electrical Engineering, Ophthalmology, and the Graduate School of Education. He is Director of Stanford's Center for Cognitive and Neurobiological Imaging and Deputy Director of Stanford's Neurosciences Institute. Wandell's research centers on vision science, spanning topics from visual disorders, reading development in children, to digital imaging devices and algorithms for both magnetic resonance imaging and digital imaging.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/