Incorporating adaptation in object recognition models captures temporal dynamics in neurophysiology and perception

Kasper Vinken[a,b,c,✉], **Xavier Boix**[a,b,d], **and Gabriel Kreiman**[a,b]

[a]Boston Children's Hospital, Harvard Medical School, Boston, MA 02115
[b]Center for Brains, Minds and Machines, Cambridge, MA 02139
[c]Laboratory for Neuro- and Psychophysiology, Department of Neurosciences, KU Leuven, 3000, Leuven, Belgium
[d]Department of Brain and Cognitive Sciences, MIT, Cambridge, MA 02139

The ability to rapidly adapt to the current sensory environment is a fundamental property of sensory systems. Such adaptation may depend on computations implemented at the neural circuit level, but also on intrinsic cellular mechanisms. We combined neural data, psychophysics, and computational models to evaluate whether an intrinsic neural fatigue mechanism in a feedforward hierarchical network is sufficient to explain the core properties of visual adaptation. We implemented activity-based response suppression in each unit of a convolutional deep neural network. The resulting units showed hallmark properties of repetition suppression, leading to a stimulus-specific and layer-dependent response attenuation for frequent input. Furthermore, the activation patterns could account for known perceptual aftereffects such as the tilt and face-gender aftereffect. The model was also able to capture the results of a visual categorisation experiment demonstrating enhanced object recognition when objects are embedded in a constant pattern of noisy inputs. The computations required for adaptation can be learned, as demonstrated by units in a recurrent neural network which learn to suppress themselves when trained on that same psychophysics visual recognition experiment. These results show that a learned intrinsic neural mechanism of fatigue can incorporate temporal context information accounting for key neurophysiological and perceptual properties and leading to efficient and enhanced processing of sensory inputs.

adaptation | object recognition | deep neural network | ventral stream | perceptual aftereffects | repetition suppression | neurophysiology | neuronal fatigue

Correspondence: ***kasper.vinken@childrens.harvard.edu***

Introduction

Information processing in biological vision is not just a constant function of current visual input, but is dynamically changed by that input. These changes can dramatically alter our visual experience, such as the illusory perception of upwards motion after watching flowing water in the so-called waterfall illusion (Addams, 1834). Likewise, in the brain neural responses change both during and after the presentation of a stimulus, resulting in complex temporal dynamics. The dependence of neural activity and perception on recent stimulation is generally called adaptation and is considered a fundamental property of our visual and other sensory systems (Whitmire and Stanley, 2016). Adaptation can improve metabolic efficiency by avoiding costly signal processing under conditions where the inputs are constant. In addition, adaptation matches the system's sensitivity to the prevailing conditions of the sensory environment, improving novelty detection (Clifford et al., 2007; Kohn, 2007). Rapid calibration of the system's operations is particularly relevant for understanding biological vision because it relates to the moment-to-moment changes of natural visual input (Whitmire and Stanley, 2016).

Our current best models of the primate ventral visual stream are based on a family of feedforward deep artificial neural networks (ANN) that assume a static stimulus-response relation. By incorporating computational principles that are directly inspired by the visual system (LeCun et al., 2015), these models have been shown to describe the responses in the ventral visual stream to brief stimulus presentations (Cadieu et al., 2014; Yamins et al., 2014; Güçlü and van Gerven, 2015; Kalfas et al., 2017, 2018) while capturing to some extent aspects of object recognition and perceived shape similarity in primates at the behavioral level (Yamins et al., 2014; Kubilius et al., 2016; Kalfas et al., 2018). These successes have shown that ANNs are a powerful tool for relating computational principles both to neural representations as well as perception and thus may provide a comprehensive framework for connecting the different levels at which adaptation phenomena have traditionally been described.

At the perceptual level, visual adaptation refers to fast but temporary changes during exposure to a stimulus, often leading to a reduced sensitivity for its features. For example, exposure to a high contrast grating reduces sensitivity for gratings similar in orientation and spatial frequency (Blakemore and Campbell, 1969). The lingering effects after removal of an adapter stimulus are referred to as aftereffects, as exemplified by the waterfall illusion resulting from adaptation to motion. Aftereffects have been originally described for a wide range of low-level stimulus properties such as motion, color, contrast, and orientation (Frisby and Stone, 2010). More recently however, they have also been reported for high-level level properties such as combinations of lower-level features, shape, or complex dimensions along which we classify faces (Webster and MacLeod, 2011; for an overview see Webster, 2015). This range of aftereffects for low to high-level properties suggest that the underlying mechanisms of visual adaptation operate at multiple levels of processing in the visual

system (Webster, 2015).

In visual cortex, neural adaptation effects have been reported in several different areas, from primary visual cortex (V1) to inferotemporal cortex (IT; Kohn, 2007; Vogels, 2016), and across species (Vinken et al., 2017). Mirroring the reduced perceptual sensitivity described in the previous paragraph, exposure to an adapter stimulus often reduces the neural response to a test stimulus that is similar to the adapter. The stimulus specificity of adaptation, that is the dependence on the similarity between adapter and test stimulus, is a hallmark property of adaptation (for an overview see Kohn, 2007). When a stimulus is repeated (i.e. maximum similarity), the resulting response reduction is called repetition suppression. The strength of repetition suppression has been shown to increase with the number of repetitions, even if there are intervening stimuli (Miller et al., 1991; Ulanovsky et al., 2003; Sawamura et al., 2006; Kaliukhovich and Vogels, 2014; Vinken et al., 2017), and decrease with the interstimulus interval (Sawamura et al., 2006).

Repetition suppression and aftereffects are usually considered to be manifestations of the same underlying mechanisms, but it remains uncertain what those mechanisms are. Adaptation effects could be implemented at the circuit level through suppression by inhibitory connections between neurons, but they might also emerge from intrinsic properties of single neurons or synapses (Whitmire and Stanley, 2016). At the cellular level, the responsiveness of cortical neurons can be controlled by intrinsic mechanisms that increase the membrane conductance. Indeed, contrast adaptation in cat visual cortex leads to a strong afterhyperpolarization of the membrane potential (Carandini and Ferster, 1997; Sanchez-Vives et al., 2000b). This afterhyperpolarization is caused by sodium-activated potassium currents that are triggered by the influx of sodium ions following high frequency firing (Sanchez-Vives et al., 2000a; Abolafia et al., 2011). Thus, in this scenario intrinsic properties of individual neurons control its responsiveness based on the strength of its previous activation, which is called firing-rate adaptation or fatigue.

Neural fatigue on its own cannot explain complex effects of adaptation such as stimulus selectivity, because fatigue should equally affect responses for all stimuli. However, several studies have suggested that adaptation effects cascade through the visual system (De Baene and Vogels, 2010; Dhruv and Carandini, 2014; Patterson et al., 2014). Thus, more complex effects could emerge when fatigue is considered in a hierarchical neural network, where simple suppressive effects could propagate through the circuit and dynamically change its state (Solomon and Kohn, 2014; Whitmire and Stanley, 2016). Here, we investigated whether core phenomena of visual adaptation can be explained by activity-based response suppression cascading through a feed-forward neural network. We started by implementing an exponentially decaying fatigue mechanism in the units of a pre-trained ANN (Krizhevsky et al., 2012) and asked whether it could account for the temporal dynamics of adaptation in neurophysiology and perception. Next, we ran a psychophysics experiment to test whether adaptation can improve object recognition by adapting to prevailing conditions. We show that an intrinsic adaptation mechanism can be learned by recurrent neural networks, when trained on the same object recognition task. Finally, we show that a circuit solution learned by a recurrent neural network is less robust than intrinsic neural fatigue.

Methods

Computational Model

We used the AlexNet architecture (Krizhevsky et al., 2012) (**Fig. 1A**), with weights pre-trained on the ImageNet dataset (Russakovsky et al., 2015) as a model for the ventral visual stream. We implemented an exponentially decaying fatigue mechanism (Bellec et al., 2018). For each unit in every convolutional and fully connected layer (except for the decoder) we assigned a suppression state s_t, which was updated at each time step t based on its previous state s_{t-1} and the previous response r_{t-1} (i.e. activations after ReLU):

$$s_t = \alpha s_{t-1} + (1-\alpha) r_{t-1} \quad \textbf{(1)}$$

where α is a constant in [0,1] determining the time scale of the decay (**Fig. 1B**). This suppression state is then subtracted from the encoding of the unit's current input x_t (given weights W and bias b) before applying the rectifier activation function σ, so that:

$$r_t = \sigma(b + W x_t - \beta s_t) \quad \textbf{(2)}$$

where β is a constant that scales the amount of suppression. These model updating rules result in an exponentially decaying response for constant input which recovers in case of no input (**Fig. 1B**), simulating a neural fatigue mechanism intrinsic to each individual neuron. By implementing this mechanism across discrete time steps in AlexNet, we essentially introduced a temporal dimension to the network (**Fig. 1C**). Throughout the paper, we use $\alpha = 0.96$ and $\beta = 0.7$ unless indicated otherwise.

Neurophysiology

We present neurophysiological data from two previously published studies in order to compare them with the neural adaptation effects of the proposed computational model: single cell ($N = 97$) recordings from inferior temporal (IT) cortex of one macaque monkey G (Vinken et al., 2018) and multi-unit recordings from primary visual cortex ($N = 55$) and latero-intermediate visual area ($N = 48$) of three rats (Vinken et al., 2017). For methodological details we refer to the original papers.

Psychophysics

Participants. A total of 12 volunteers (7 female, ages 19-50) participated in our doodle categorization experiments. All subjects gave informed consent and the studies were approved by the Institutional Review Board at Children's Hospital, Harvard Medical School.

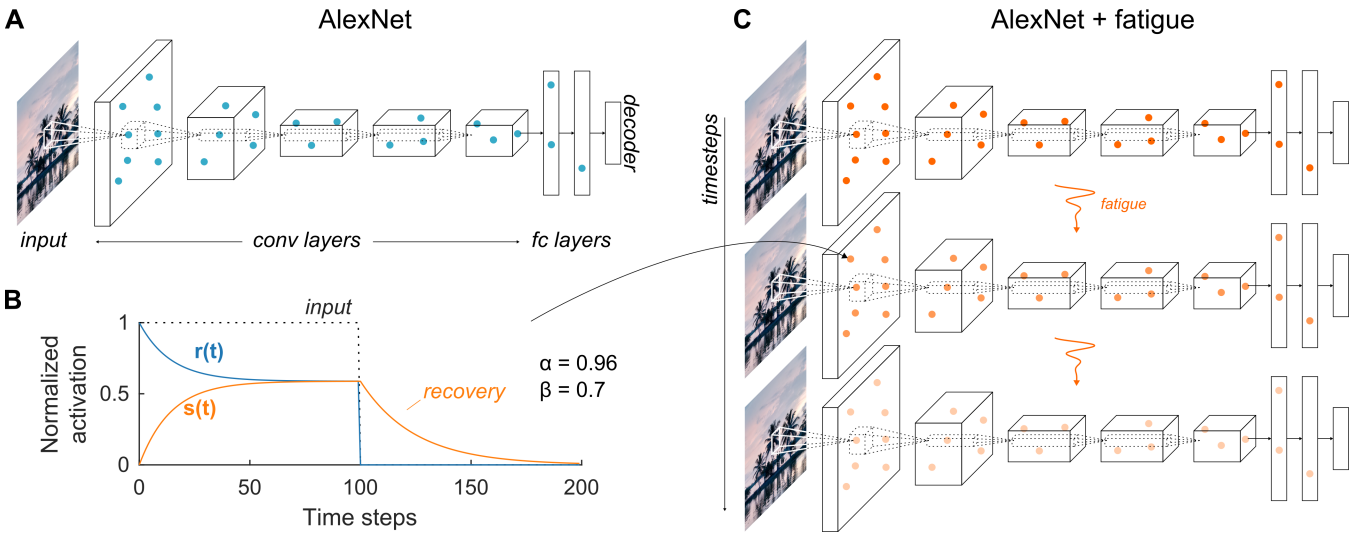

Fig. 1. Neural network architecture and expansion over time to include neural fatigue. **A** Architecture of a regular static deep convolutional neural network, in this case AlexNet (Krizhevsky et al., 2012). AlexNet contains five convolutional layers (conv1-5) and three fully connected layers (fc6, fc7, and the decoder fc8). The unit activations in each layer, and therefore the output of the network, are a fixed function of the input image. **B** Fatigue implemented by Equations 1 and 2 results in suppression over time for constant input (time steps 0-100) and recovery in the absence of input (time steps > 100). In this case $\alpha = 0.96$ and $\beta = 0.7$. **C** An expansion over time of the network in **A**, where the activation of each unit is a function of its inputs and its activation at the previous time step (Equations 1 and 2).

Stimuli. For the stimulus set we took hand drawn doodles of apples, cars, faces, fish, and flowers from the *Quick, Draw!* dataset (Google Creative Lab, 2019). We selected a total of 540 doodles (108 from each of the five categories) that were judged complete and identifiable. We lowered the contrast of each doodle image (28x28 pixels) to either 22 or 29% of the original contrast, before adding a Gaussian noise pattern (SD = 0.165 in normalized pixel values) of the same resolution. The higher contrast level (29%) was chosen so that the doodle was relatively visible as a control, was used in only one sixth of the trials, and was not included in the analyses.

Experimental protocol. Participants had to fixate a cross at the center of the screen in order to start a trial. Next, an adapter image was presented (for 0.5, 2, or 4 s), followed by a blank interval (of 50, 250, or 500 ms), a test image (for 500 ms), and finally a response prompt screen. The test images were noisy doodles described in the above paragraph. The adapter image could either be: an empty frame (defined by a white square filled with the background color), the same mosaic noise pattern as the one of the subsequent test image, or a randomly generated different noise pattern (**Fig. 4**). Participants were asked to keep looking at the fixation cross which remained visible throughout the entire trial until they were prompted to classify the test image using keyboard keys 1-5. All images were presented at 9 x 9° from a viewing distance of approximately 52 cm on a 19 inch CRT monitor (Sony Multiscan G520, 1024 × 1280 resolution), while we continuously tracked eye movements using a video-based eye tracker (EyeLink 1000, SR Research, Canada). Trials where the root-mean-square deviation of the eye-movements exceeded 1 degree of visual angle during adapter presentation were excluded from further analyses. The experiment was controlled by custom code written in MATLAB using Psychophysics Toolbox Version 3.0 (Brainard, 1997).

Results

We introduced an adaptive mechanism into bottom-up computational models of vision by incorporating units that show neural fatigue (Methods). We show that the incorporation of neural fatigue is able to capture fundamental dynamic properties of adaptation both at the neurophysiological level as well as at the behavioral/perceptual level.

A neural network incorporating neural fatigue captures the attenuation in neurophysiological responses during repetition suppression

We first show that adding an intrinsic neural fatigue mechanism to each unit in an ANN replicates the hallmark properties of neural adaptation. The most prominent characteristic of neural adaptation is repetition suppression, or a response reduction when a stimulus is repeated. For example, in trials with presentation of two sequential stimuli, the response to the second stimulus is typically lower and this reduction is strongest when the second stimulus is identical to the first. This phenomenon is illustrated by the data in **Fig. 2A**, which shows a lower response in IT neurons for a face repetition (blue) compared to a face alternation (orange) (Vinken et al., 2018). We simulated the same experiment in our neural network by presenting the same trials assuming time step intervals of 50 ms ($\alpha = 0.96$, and $\beta = 0.7$ in Equations 1 and 2), which resulted in qualitatively similar results **Fig. 2B**. Stimulus-specific repetition suppression was more pronounced in IT neurons, which were recorded from the middle lateral face patch, compared to conv5 units of AlexNet, suggesting that there was more stimulus selectivity in the input population of the real neurons. The model units demonstrate the key features of adaptation at two time scales: (i) during presentation of any stimulus, including the first stimulus, there is a decrease in the response with time; (ii) the overall response to the second stimulus is smaller than

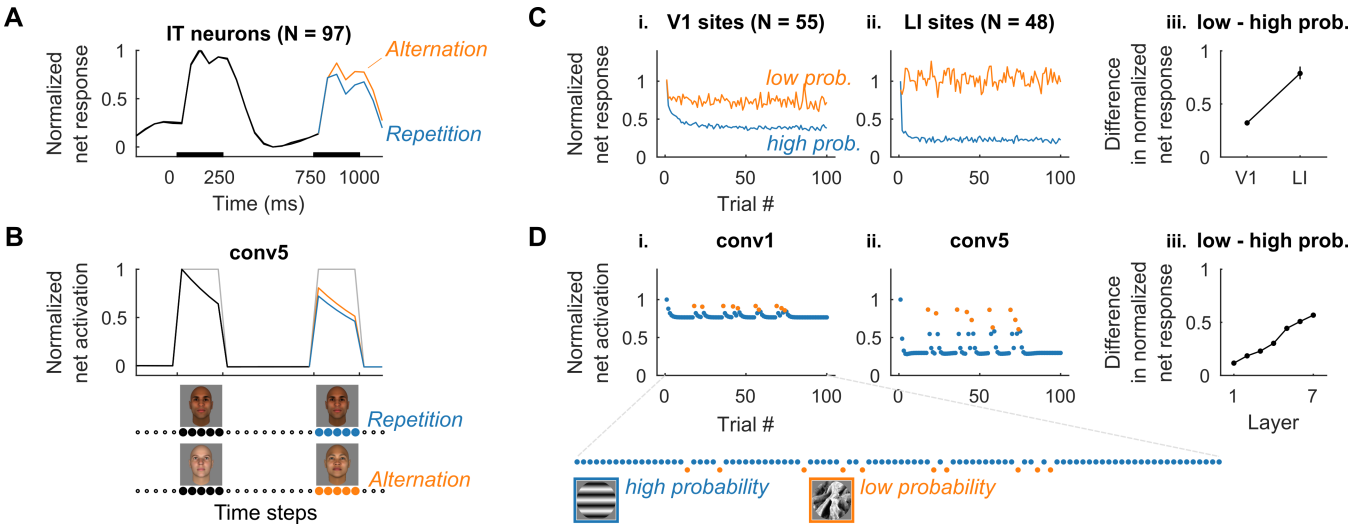

Fig. 2. Neural fatigue in an ANN captures temporal dynamics of adaptation at the neurophysiological level. A Neural responses in macaque inferior temporal cortex (IT) are suppressed more for a repeated stimulus (blue) than for a new stimulus (orange). The data are responses to faces recorded from $N = 97$ responsive neurons in the middle lateral face patch of one monkey (Vinken et al., 2018). Black bars indicate stimulus presentation. **B** A simulation of the same experiment as in **A** leads to similar stimulus-specific suppression in an ANN with neural fatigue. This plot shows the activity of $N = 6590$ responsive units in the conv5 layer of AlexNet. The x-axis units are time steps, mapping to bins of 50 ms in panel **A**. For comparison we show the response time course generated by the model without adaptation (grey). Below are example stimuli used in the actual and simulated experiment. **C** Accumulation of adaptation across multiple repeats leads to increased suppression for high probability stimuli (blue) compared to low probability stimuli (orange) in an oddball paradigm. The data are responses from multi-unit sites recorded during oddball sequences in rat **i.** primary (V1, $N = 55$) and **ii.** latero-intermediate (LI, $N = 48$) visual cortex (Vinken et al., 2017). Time courses were normalized by the response at the first trial. **iii.** Difference in response for the low and high probability stimulus increases from V1 to LI (error bars are 95% bootstrap confidence intervals calculated assuming no inter-animal difference). **B** A simulation of an oddball sequence from **C** leads to similar accumulation of suppression across stimulus presentations and stages of processing. Below are example stimuli of a high probability grating (blue) and a low probability texture (orange) used in the actual and simulated experiments. Note that stimulus type and probability were counterbalanced for each neural recording.

the overall response to the first stimulus; (iii) the response to the second stimulus is attenuated more when it is a repetition.

In addition to the two temporal scales illustrated in **Fig. 2A-B**, adaptation not only affects responses from one stimulus to the next, but also operates at longer time scales. For example, repetition suppression typically accumulates across multiple stimulus presentations and can survive intervening stimuli (Sawamura et al., 2006). To examine this longer time scale over multiple trials, we considered an *oddball paradigm* where two stimuli are presented with different probabilities in a sequence (**Fig. 2D**, bottom): the *standard* stimulus is shown with high probability (blue) and the *deviant* stimulus is shown with a low probability (orange). The sequence consisted of 100 stimulus presentations, each one shown for 300 ms and separated by 300 ms, with 90 standard stimuli and 10 deviant stimuli shown in random order. We illustrate the build-up in adaptation over time using data recorded from $N = 55$ neurons in the rat primary visual cortex (Vinken et al., 2017): the standard stimulus is far more likely to be repeated in the sequence, allowing adaptation to build up and cause a more decreased response for later trials. In contrast, the low probability stimulus does not show such a response reduction. There is evidence that adaptation effects increase in later stages of processing (Vinken et al., 2017; Kaliukhovich and Op de Beeck, 2018; Nieto-Diego and Malmierca, 2016). Indeed, in the same oddball paradigm, the difference in the response between the deviant stimulus and the standard stimulus was larger in the latero-medial (LI) visual cortex (**Fig. 2C**; Vinken et al., 2017).

The proposed model was able to qualitatively capture the response difference between standard and deviant stimuli **Fig. 2D**. The model also showed a partial release from adaptation after each presentation of a deviant, similar to the observaitons in rat visual cortex (see Supplemental Information of Vinken et al., 2017). Furthermore, there was a monotonic increase in the adaptation effect (difference in response between deviant and standard stimulus) from one layer to the next. These results demonstrate that neural fatigue propagating through a hierarchical system encodes stimulus probabilities with increasing sensitivity. Note that the L1 responses in rats also showed an *increased* response to the low probability deviant (normalized response above 1 in **Fig. 2C ii**; this effect was not captured by the model and is likely to require additional mechanisms, see Discussion).

A neural network incorporating neural fatigue reveals perceptual manifestations of adaptation

After demonstrating the core phenomena of neural adaptation during repetition suppression, we will now show that an ANN with neural fatigue can explain perceptual aftereffects of adaptation. We start with an aftereffect for the low-level stimulus property of orientation (Gibson and Radner, 1937; Frisby and Stone, 2010). The tilt aftereffect refers to a phenomenon where an observer perceives a vertically oriented grating to be slightly tilted in the direction *opposite* to the tilt direction of an adapter (see **Fig. 3A**). In other words, the decision boundary for perceptual orientation discrimination shifts towards the adapter (Frisby and Stone, 2010). On the other hand, no shift should occur when the adapter has the same orientation as the test stimulus.

To evaluate whether perceptual aftereffects can be described

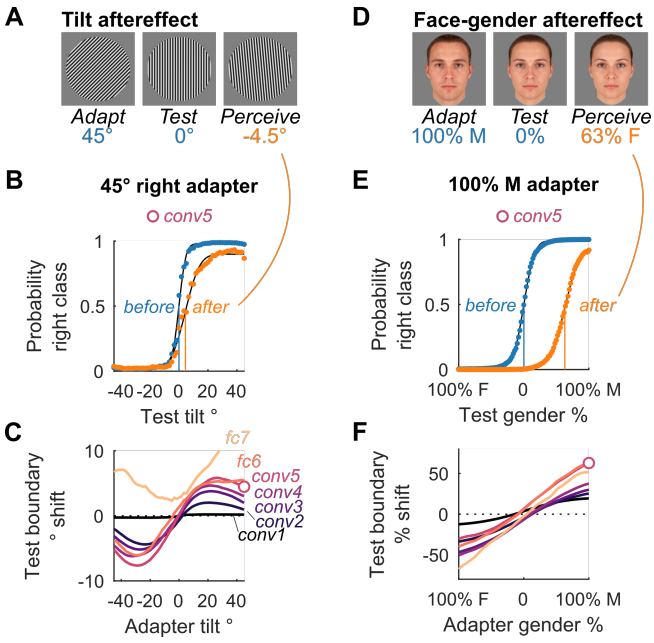

Fig. 3. An ANN incorporating neural fatigue demonstrates perceptual adaptation effects A,D Illustration of the tilt (**A**) and face-gender (**D**) aftereffect with the stimuli used in our simulated experiments. After exposure to an adapter (left), a test stimulus is presented (middle). The test stimulus is perceived differently as a result of a shift in the decision boundary toward the adapter. In (**A**), observers perceive the vertically oriented grating as tilted to the left. In (**D**), observers perceive the neutral face as more female like. The images for the illustration of perceived aftereffects (**A,D**) were picked based on the estimated conv5 boundary shift shown below. **B,E** Decision boundaries before (blue) versus after (orange) exposure to the adapter based on the conv5 layer of the model with neural fatigue. Markers show class probability estimates for each test stimulus, full lines indicate the corresponding psychometric functions, and vertical lines the classification boundaries. In (**B**), adaptation to a 45 degree grating leads to a shift in the decision boundary to positive orientations, hence perceiving test stimuli with more negative orientations. In (**E**), adaptation to a 100% male face leads to a shift in the decision boundary towards male faces, hence perceiving test stimuli as more female-like. **C,F** Decision boundary shifts for the test stimulus as a function of the adapter tilt/face-gender per layer. Round markers indicate the conv5 boundary shifts plotted in (**B,E**). Note that the classifier for the tilt aftereffect in fc7 is not robust to the effects of adaptation.

by an ANN with neural fatigue, we created a set of gratings that ranged parametrically from left to right oriented (-45° to 45° in 100 steps), and measured the category boundary for each layer of the model before and after adaptation. Specifically, these boundaries were estimated using a binary classifier (logistic regression, trained on the full stimulus set before adaptation) and fitting a psychometric function (Wichmann and Hill, 2001) on the class probability estimates given by that classifier. For a fair comparison, the before and after adaptation boundaries were always calculated using the same classifier in the same space, namely the principal component space obtained from the unadapted outputs to the full stimulus set. In **Fig. 3B** we show the psychometric curves fit on the conv5 class probability estimates before and after adaptation to a 45° right tilted grating. As predicted, the decision boundary shifted towards the tilt of the adapter. **Fig. 3C** shows that for all layers but fc7, adaptation to a tilted grating resulted in a boundary shift towards the adapter. Given that the test stimulus is vertical, adaptation to a vertically oriented grating had no effect. The effect of adaptation propagated and accumulated over the layers, and the shift in the test boundary was largest for conv5 and fc6. For conv1 there was also a shift which is too small to notice in **Fig. 3C**.

Aftereffects have also been described for high level stimulus properties, such as the gender of faces (Webster et al., 2004). Like the tilt aftereffect, adapting to a male or female face results in a face-gender boundary shift towards the adapter, whereas adapting to a gender-neutral face leads to no shift when the test stimulus is neutral. In other words, exposure to a male face will make a neutral face appear more female (**Fig. 3D**). To test this in the model proposed here, we created a set of face stimuli that morphed from average female to average male face in 100 steps (using Webmorph: DeBruine, 2019). As predicted, exposing the model to an adapter face shifted the face-gender boundary towards the adapter gender (**Fig. 3E, F**), whereas adaptation to a gender-neutral face had no effect. The aftereffect did not suddenly emerge in later layers, but slowly built up in an approximately monotonic fashion with increasing layers, consistent with the idea of adaptation cascading and accumulating across the stages of processing. In this framework, all stages of processing can contribute to aftereffects, based on the population of neurons that responds to both the adapter and test images.

Neural fatigue increases sensitivity to changes

One of the proposed computational roles of neural adaptation is to increase sensitivity to small but relevant changes in the sensory environment by adapting to the prevailing conditions (Clifford et al., 2007; Kohn, 2007). To test the hypothesis that adaptation can increase the ability to identify small but relevant changes in a sensory stream, we ran a psychophysics categorization task whereby participants were required to classify hand drawn doodles (**Fig. 4A**) hidden in a noise pattern. We asked whether adaptation to the same noise pattern would increase the ability to recognize the target object embedded in the noise pattern. We compared the behavioral results against two control conditions where we would expect to see no effect of adaptation: one where no adapter was used (i.e. an empty frame, **Fig. 4B**, left) and one where a different noise pattern was used (**Fig. 4B**, right) (Methods).

The task is not easy: whereas subjects can readily recognize the doodles in isolation, when they are embedded in noise and in the absence of any adapter, performance was 58% ($SD = 9$) where chance is 20%. Adapting to a noise pattern increased recognition performance by 9% (**Fig. 4C i**, $p = 0.016$, Wilcoxon signed rank test, $N = 12$ subjects). This increase in performance was contingent on the noise pattern presented during the test stimulus being the same as the noise pattern in the adapter. Performance in the same noise condition was 7% higher than in the different noise condition ($p = 0.009$, Wilcoxon signed rank test, $N = 12$ subjects).

We next evaluated the model's performance in this same task. We first fine-tuned the pre-trained fully connected layers of AlexNet to classify high contrast (i.e. 40% as opposed to 22% in the experiment) doodles on a noisy background, using 10,000 independent training images for each of the 5 categories for 5 epochs. In the experiment the model demonstrated the same effects as the human participants, showing increased performance for the same noise condition com-

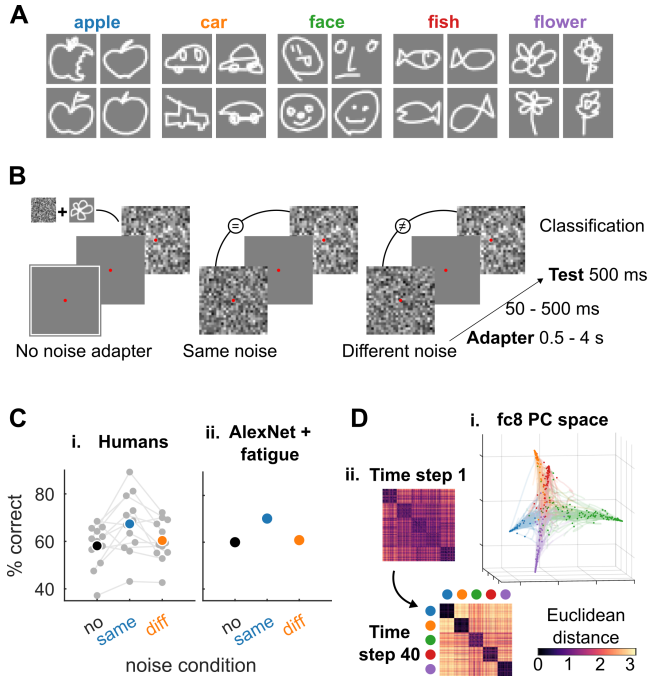

Fig. 4. Neural fatigue increases sensitivity to changes by adapting to previous input. A Representative examples for each of the five doodle categories from the total set of 540 selected images (Google Creative Lab, 2019). **B** Schematic illustration of the conditions used in the doodle experiment. In each trial participants or the model had to classify a hand drawn doodle hidden in noise (test), after adapting to the same (middle), a different (right), or no (left) noise pattern. **C** Both **i.** human participants and **ii.** the proposed computational model showed an increase in performance after adapting to the same noise pattern. Gray circles and lines denote individual participants ($N = 12$). The colored circles show average performance. Chance = 20%. Note that for this figure we decreased the suppression scaling constant to $\beta = 0.1$ to get comparable adapter effects (for all other figures it was set to $\beta = 0.7$). **D** Neural fatigue enhanced the representation of the signal (doodle) by adapting to the prevailing sensory conditions (noise pattern). **i.** Representation of the dynamic evolution of the representation of the images embedded in noise. The 3 axes correspond to the first 3 principal components of the fc8 layer representation of all the test images. Each dot represents a separate doodle+noise image, the color corresponds to the category (as shown by the text in part A). The transparent lines denote the trajectory joining the initial and final representation of each doodle+noise image. Adaptation to the same noise pattern moves the doodle representations in fc8 principal component space to separable clusters. **ii.** Dissimilarity matrix for all pairs of images. Entry (i,j) shows the Euclidian distance between image i and image j based on the fc8 features at time step 1 (top) or time step 40 (bottom). The distance is represented by the color of each point in the matrix (see scale on bottom right). Images are sorted based on their categories (colored circles refer to category labels in **A**). Adaptation leads to an increase in between category distances and a decrease in within category distances as shown by the pairwise distance matrices.

pared to the no adapter condition or different noise condition (**Fig. 4C ii.**). Neural fatigue enhanced the representation of the signal (doodle) by adapting to the current sensory conditions. **Fig. 4D i.** shows the dynamic evolution of the representation of each doodle+noise image in a space determined by the first 3 principal components of the fc8 feature representation. The separation between images in the feature space was quantified by computing the dissimilarity matrix for all possible pairs of images (**Fig. 4D ii.**. Adaptation led to increased differentiation of the between category comparisons (off diagonal squares) and increased similarity of images within each category (diagonal squares) from the initial conditions (top) to the final time step (bottom).

Adaptation can be learned in a network trained to maximize recognition performance

The structure and parameters of the model considered thus far were hard coded by introducing a neural fatigue mechanism. The proof-of-principle model allowed us to demonstrate that an intrinsic neural mechanism is sufficient to capture core neurophysiological and perceptual adaptation effects in a feedforward neural network. Next we asked whether it is possible to learn adaptation in a neural network without imposing a fatigue mechanism. In a purely feed-forward network without any temporal dynamics, it is not possible to learn adaptation. Therefore, we considered the recurrent neural network schematically illustrated in **Fig. 5A**, consisting of three convolutional layers and one fully connected layer (before the decoder). We started by considering a simplified version of the experiment presented in the previous section with two time steps (without the intervening blank delay), with the adapter shown at step one and the test image at step two (**Fig. 5B**). We used the same doodle images hidden in noise from (**Fig. 4**), with 100,000 doodles in each of the five categories for training (5 epochs), and 1000 doodles in each category for testing.

As a baseline, we first evaluated the performance of the same network without recurrent connections (**Fig. 5B i.** black bar); this network reached a performance of 74.7%. Upon including the recurrent connections, the network achieved a performance of 97.5% when trained on same noise trials (**Fig. 5B i.** blue bar). In contrast, when the network was trained on different noise trials, performance was similar to the one obtained without recurrent connectivity (**Fig. 5B i.** orange bar). To investigate how the recurrent neural network solved the task, we compared the on-diagonal recurrent weights (i.e., the weights connecting a unit onto itself across time steps), with off-diagonal recurrent weights (i.e., the weights connecting different units horizontally within a layer). In the first convolutional layer (conv1), on-diagonal weights tended to be strongly negative compared to off-diagonal weights (**Fig. 5B ii.**). Thus, the units in layer conv1 learned to suppress themselves based on the previous time step, a hallmark signature of intrinsic adaptation. In contrast, when the recurrent neural network was trained on different noise trials, the noise pattern information at time step one was not useful and we did not observe the negative diagonal weights.

This self-suppression solution only works when there is no gap between the adapter and test images, because the units have no memory capacity by themselves, unlike units with a slowly decaying hidden state. However, when we trained a recurrent neural network to solve the task when there is a gap (**Fig. 5C**), the network could still learn to use its recurrent loops to "remember" the adapter noise pattern. To compare this circuit solution with that of an exponentially decaying decaying hidden state, we built the same network without recurrent connections, but with the hidden state dictated by Equations 1 and 2. In addition to training the feedforward weights on the task, we now also trained parameters α and β per layer. The value of α determines how fast the hidden state updates, ranging from no update ($\alpha = 1$) to completely re-

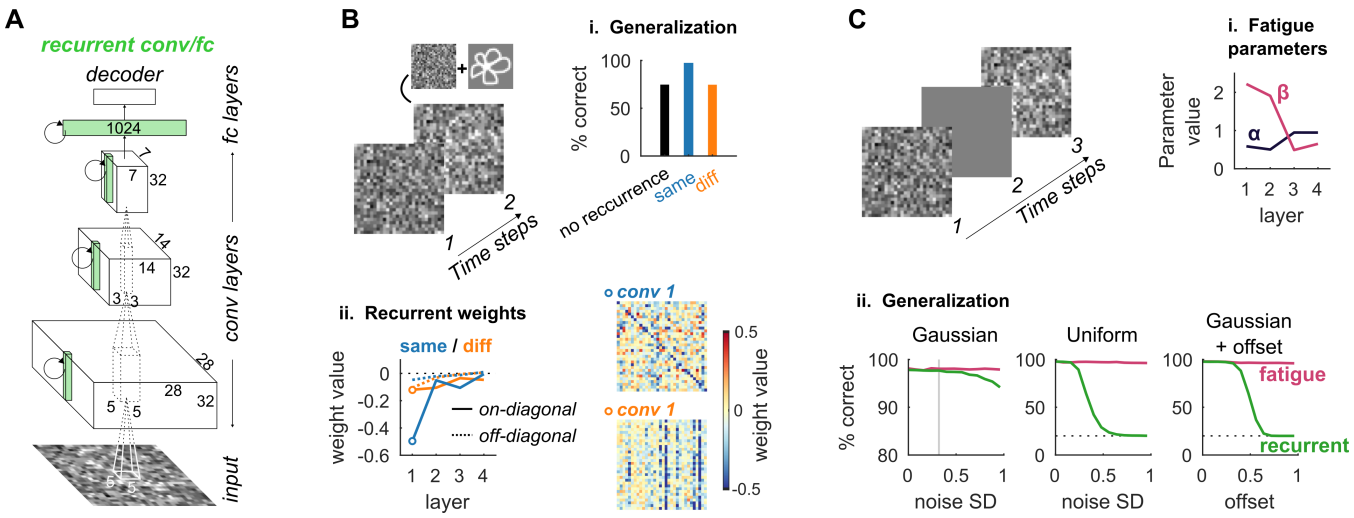

Fig. 5. A recurrent neural network can learn to use the adapter noise pattern information. **A** Recurrent convolutional neural network, with fully connected recurrent weights for the fully connected layer and convolutional recurrent kernels of size 1 x 1 x 32 and stride 1 for the three convolutional layers. **B i** Generalization performance for a network trained on the same noise condition, without recurrence (black bar), or a network with recurrent connections trained on same noise trials (blue) or different noise trials (orange). Chance = 20% **ii.** (left) Average recurrent connection weights for layers 1 through 4 separately showing on-diagonal weights (solid lines), off-diagonal weights (dotted lines) when the network is trained with the same noise condition (blue) or different noise condition (orange). Right: all recurrent weights for the conv1 layer in the same noise condition (top) and different noise condition (bottom). Entry (i,j) indicates the connection weight between unit i and unit j, see color scale on right. **C** We trained a recurrent network on same noise trails with a gap between adapter and test images and compare with a network of the same architecture but without recurrent weights and with units with a decaying hidden state built in (according to Equations 1 and 2). **i.** Here, in contrast with previous figures, the parameters α and β were learned separately for each layer. This figure shows the fitted parameters for each layer. **ii.** Categorization performance (chance = 20%, indicated by horizontal dotted lines) when different amounts of noise (x-axis) was added during testing for the network implementing neural fatigue (magenta) or the recurrent network (green). (left) Gaussian noise. (middle) Uniformly distributed noise. (right) Gaussian noise (SD = 0.32) with an added offset value. The vertical dotted line shows the amount of noise used during training (only in the panel on the left which was the type of noise used for training).

newing at each time step ($\alpha = 0$). The value of β determines whether the hidden state is used for activation-based fatigue ($\beta > 0$), enhancement ($\beta < 0$) or nothing at all ($\beta = 0$). Training these parameters led to fatigue, which was strongest in the first two layers (higher β and lower α (**Fig. 5C i.**). Both the neural fatigue mechanism and the recurrent network generalized well to different trials with the same Gaussian noise distribution as the training set (**Fig. 5C ii.**, left). However, the recurrent network failed to generalize to higher standard deviations of uniformly distributed noise, or Gaussian noise with an offset (**Fig. 5C ii.**, middle and right). This suggests that in contrast with an intrinsic neural fatigue mechanism, the circuit implementation learned by the recurrent neural network is not robust to deviations from the particular training conditions.

Discussion

The visual system continuously adapts to previous stimulation. The dynamics of incorporating such temporal contextual information could be a consequence of the interactions in the whole circuit, but they could also arise from intrinsic biophysical mechanisms in each individual neuron. The basic building blocks in most of the existing neural network models of the visual system are simplified linear summation units plus a nonlinear activation function such as the rectifying linear unit (ReLU). Deep convolutional networks are based on hierarchical cascades of such units, and do not incorporate the temporal dynamics of neuronal responses. Previous attempts at incorporating a time dimension have relied on adding recurrent weights to a network (Tang et al., 2018; Kar et al., 2019).

A particularly notable aspect of how temporal context is incorporated in visual processing is the effect of adaptation, which is evident in the attenuated neuronal responses to repeated stimuli and is also evident in perceptual aftereffects and other visual illusions. In this study, we investigated whether the paradigmatic neurophysiological and perceptual signatures of adaptation effects can be explained by an intrinsic activation-dependent neural fatigue mechanism inspired by afterhyperpolarization of the membrane potential (Sanchez-Vives et al., 2000b). We showed that an ANN which implements neural fatigue for each unit can explain classical perceptual aftereffects of adaptation, such as the tilt and face-gender aftereffects (Gibson and Radner, 1937; Webster et al., 2004). In addition, the units in this network exhibited stimulus-specific repetition suppression (Kohn, 2007; Vogels, 2016), which recovers over time but also builds up across repeats despite intervening stimuli (Ulanovsky et al., 2003), and builds up across stages of processing (Vinken et al., 2017; Kaliukhovich and Op de Beeck, 2018; Nieto-Diego and Malmierca, 2016). Thus, activation-based fatigue in deep neural network units leads to neural adaptation effects and aftereffects.

Next, we used psychophysics to demonstrate perceptual effects of adaptation on object recognition under varying conditions. As predicted by the proposed model, adapting to a noise pattern increased object recognition performance when the target object was hidden in a temporally constant noise pattern. Finally, when we trained a recurrent neural network to perform the same task, the neurons learned to suppress themselves using recurrent weights, showing that the basic

activation-based suppression of neural fatigue could be readily learnt such that adaptation can emerge from the goal of maximizing recognition performance. However, when there was a gap between the adapter and test images, the recurrent neural network relied on a circuit solution that was less robust to different noise conditions than the fatigue based solution.

Functional benefits

In the proposed computational model, the response strength of units for a given stimulus sequence was inversely proportional to the probability of each stimulus. Such automatic encoding of stimulus probability has been observed across species in somatosensory (Musall et al., 2014, 2015), auditory (Ulanovsky et al., 2003; Fishman and Steinschneider, 2012; Farley et al., 2010), and visual cortices (Kaliukhovich and Vogels, 2014; Hamm and Yuste, 2016; Vinken et al., 2017) and has been proposed to enhance novelty detection and separate behaviorally relevant information from the prevailing input (Ulanovsky et al., 2003; Musall et al., 2015; Vinken et al., 2017). Because of this property, stimulus-specific repetition suppression has been interpreted as a manifestation of a reduced prediction error within the predictive coding framework (Friston, 2005; Summerfield et al., 2008). This framework stresses the role of top-down modulations by internally generated perceptual expectations rather than neural fatigue. However, repetition suppression can be dissociated in time from expectation effects (Todorovic and de Lange, 2012) and is not modulated by perceptual expectations induced by repetition probability in macaque IT neurons (Kaliukhovich and Vogels, 2014; Vinken et al., 2018). Therefore, we argue that basic repetition suppression does not rely on top-down circuitry and can be explained by feedforward effects of intrinsic neural mechanisms. On the other hand, in rat higher level visual cortex we did observe a response enhancement for rare stimuli in addition to repetition suppression, which might require additional mechanisms at the circuit level (Vinken et al., 2017). Note that this enhancement not observed in macaque IT (Kaliukhovich and Vogels, 2014).

Besides increasing the salience of novel stimuli, neural adaptation may serve to increase coding efficiency by normalizing responses for the current sensory conditions (Kohn, 2007). Neurons have a limited dynamic range with respect to the feature they encode and because they are noisy there are only a limited number of response levels. The idea is that adaptation maximizes information a neurons can carry by re-centering tuning around the prevailing conditions and thus preventing response saturation and increasing sensitivity (Webster et al., 2005). ANNs on the other hand usually do not suffer from these constraints, because they rely on the non-saturating rectifying linear unit (ReLU) activation function and are inherently not noisy.

Finally, neurons use significant amounts of energy to generate action potentials and this constrains the amount of neural activity that can be used for a neural representation (Laughlin, 2001; Lennie, 2003). By reducing neural responsiveness for redundant information, adaptation has the advantage of increasing metabolic efficiency of the neural code.

Other mechanisms of adaptation

Several mechanisms have been proposed to be involved in adaptation. At the circuit level, responses may be suppressed by inhibitory connections. Indeed, postsynaptic inhibition has been shown to contribute to adaptation in rat auditory cortex in the first 50-100 ms after a stimulus, but it does not explain adaptation at slower time scales (Wehr and Zador, 2005). At the level of the synapse, repetitive stimulation could cause short-term depression, weakening synaptic strength by decreasing neurotransmitter release through several molecular mechanisms (for an overview see Fioravante and Regehr, 2011). For example, depression at thalamocortical synapses could contribute to adaptation in sensory cortex (Chung et al., 2002). Still, compared to cortical slices, short-term synaptic depression appears to be less pronounced *in vivo*, where it may be decreased by ongoing network activity (Reig et al., 2006).

Time scales of adaptation

We modeled adaptation as an exponential process, after Bellec et al. (2018), therefore limiting adaptation effects to one particular time scale. In reality, adaptation operates over a range of time scales from milliseconds to minutes (Kohn, 2007) and it has been proposed that the mechanism can be better approximated by a scale-invariant power-law (Wark et al., 2007; Drew and Abbott, 2006). However, power-law adaptation can be approximated over a finite time interval using a sum of exponential functions (Drew and Abbott, 2006), so in principle we could extend our model with additional exponential processes.

Conclusion

By simulating a cellular mechanism in a deep neural network we were able to connect systems to cellular neuroscience in one comprehensive model. Our results demonstrate that response fatigue cascading through a feedforward hierarchical network was sufficient for explaining the hallmarks of visual adaptation. This implies that intrinsic neural mechanisms may contribute substantially to the dynamics of sensory processing and perception in a temporal context.

ACKNOWLEDGEMENTS

This work was supported by Research Foundation Flanders, Belgium (fellowship of K.V.), by NIH grant R01EY026025 and by the Center for Brains, Minds and Machines, funded by NSF Science and Technology Centers Award CCF-1231216. We thank Ricardo Henriques for this bioRxiv template.

Bibliography

Abolafia, J. M., Vergara, R., Arnold, M. M., Reig, R., and Sanchez-Vives, M. V. 2011. "Cortical Auditory Adaptation in the Awake Rat and the Role of Potassium Currents." *Cerebral Cortex* 21:977–990.

Addams, R. 1834. "An account of a peculiar optical phenomenon seen after having looked at a moving body." *London and Edinburgh Philosophical Magazine and Journal ofScience* 5:373–374.

Bellec, G., Salaj, D., Subramoney, A., Legenstein, R., and Maass, W. 2018. "Long short-term memory and learning-to-learn in networks of spiking neurons." *Conference on Neural Information Processing Systems* .

Blakemore, C. and Campbell, F. W. 1969. "On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal images." *Journal of Physiology* 203:237–260.

Brainard, D. H. 1997. "The Psychophysics Toolbox." *Spatial Vision* 10:433–436.

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., and DiCarlo, J. J. 2014. "Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition." *PLoS Computational Biology* 10.

Carandini, M. and Ferster, D. 1997. "A Tonic Hyperpolarization Underlying Contrast Adaptation in Cat Visual Cortex." *Science* 276:949–952.

Chung, S., Li, X., and Nelson, S. B. 2002. "Short-term depression at thalamocortical synapses contributes to rapid adaptation of cortical sensory responses in vivo." *Neuron* 34:437–446.

Clifford, C. W. G., Webster, M. A., Stanley, G. B., Stocker, A. A., Kohn, A., Sharpee, T. O., and Schwartz, O. 2007. "Visual adaptation: Neural, psychological and computational aspects." *Vision Research* 47:3125–3131.

De Baene, W. and Vogels, R. 2010. "Effects of adaptation on the stimulus selectivity of macaque inferior temporal spiking activity and local field potentials." *Cerebral Cortex* 20:2145–65.

DeBruine, L. M. 2019. "Webmorph (Version v0.0.0.9001)."

Dhruv, N. T. and Carandini, M. 2014. "Cascaded Effects of Spatial Adaptation in the Early Visual System." *Neuron* 81:529–535.

Drew, P. J. and Abbott, L. F. 2006. "Models and Properties of Power-Law Adaptation in Neural Systems." *Journal of Neurophysiology* 96:826–833.

Farley, B. J., Quirk, M. C., Doherty, J. J., and Christian, E. P. 2010. "Stimulus-specific adaptation in auditory cortex is an NMDA-independent process distinct from the sensory novelty encoded by the mismatch negativity." *The Journal of neuroscience : the official journal of the Society for Neuroscience* 30:16475–84.

Fioravante, D. and Regehr, W. G. 2011. "Short-term forms of presynaptic plasticity." *Current Opinion in Neurobiology* 21:269–274.

Fishman, Y. I. and Steinschneider, M. 2012. "Searching for the Mismatch Negativity in Primary Auditory Cortex of the Awake Monkey: Deviance Detection or Stimulus Specific Adaptation?" *Journal of Neuroscience* 32:15747–15758.

Frisby, J. and Stone, J. 2010. "Seeing Aftereffects: The Psychologist's Microelectrode." In *Seeing: The computational approach to biological vision*, pp. 75–110. Cambridge, MA: MIT Press, 2 edition.

Friston, K. 2005. "A theory of cortical responses." *Philosophical transactions of the Royal Society of London B* 360:815–36.

Gibson, J. J. and Radner, M. 1937. "Adaptation, after-effect and contrast in the perception of tilted lines. I. Quantitative studies." *Journal of Experimental Psychology* 20:453–467.

Google Creative Lab. 2019. "The Quick, Draw! Dataset." *GitHub repository https://github.com/googlecreativelab/quickdraw-dataset* .

Güçlü, U. and van Gerven, M. A. J. 2015. "Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream." *Journal of Neuroscience* 35:10005–10014.

Hamm, J. P. and Yuste, R. 2016. "Somatostatin Interneurons Control a Key Component of Mismatch Negativity in Mouse Visual Cortex." *Cell Reports* 16:597–604.

Kalfas, I., Kumar, S., and Vogels, R. 2017. "Shape Selectivity of Middle Superior Temporal Sulcus Body Patch Neurons." *Eneuro* 4:ENEURO.0113–17.2017.

Kalfas, I., Vinken, K., and Vogels, R. 2018. "Representations of regular and irregular shapes by deep Convolutional Neural Networks, monkey inferotemporal neurons and human judgments." *PLOS Computational Biology* 14:e1006557.

Kaliukhovich, D. A. and Op de Beeck, H. 2018. "Hierarchical stimulus processing in rodent primary and lateral visual cortex as assessed through neuronal selectivity and repetition suppression." *Journal of Neurophysiology* 120:926–941.

Kaliukhovich, D. A. and Vogels, R. 2014. "Neurons in Macaque Inferior Temporal Cortex Show No Surprise Response to Deviants in Visual Oddball Sequences." *Journal of Neuroscience* 34:12801–12815.

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., and DiCarlo, J. J. 2019. "Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior." *Nature Neuroscience* .

Kohn, A. 2007. "Visual Adaptation: Physiology, Mechanisms, and Functional Benefits." *Journal of Neurophysiology* 10461:3155–3164.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. 2012. "ImageNet Classification with Deep Convolutional Neural Networks." *Advances In Neural Information Processing Systems* pp. 1–9.

Kubilius, J., Bracci, S., and Op de Beeck, H. P. 2016. "Deep Neural Networks as a Computational Model for Human Shape Sensitivity." *PLoS Computational Biology* 12:1–26.

Laughlin, S. 2001. "Energy as a constraint on the coding and processing of sensory information." *Current Opinion in Neurobiology* 11:475–480.

LeCun, Y., Bengio, Y., and Hinton, G. 2015. "Deep learning." *Nature* 521:436–444.

Lennie, P. 2003. "The Cost of Cortical Computation." *Current Biology* 13:493–497.

Miller, E. K., Li, L., and Desimone, R. 1991. "A neural mechanism for working and recognition memory in inferior temporal cortex." *Science* 254:1377–1379.

Musall, S., Haiss, F., Weber, B., and von der Behrens, W. 2015. "Deviant Processing in the Primary Somatosensory Cortex." *Cerebral Cortex* p. bhv283.

Musall, S., von der Behrens, W., Mayrhofer, J. M., Weber, B., Helmchen, F., and Haiss, F. 2014. "Tactile frequency discrimination is enhanced by circumventing neocortical adaptation." *Nature neuroscience* 17:1567–73.

Nieto-Diego, J. and Malmierca, M. S. 2016. "Topographic Distribution of Stimulus-Specific Adaptation across Auditory Cortical Fields in the Anesthetized Rat." *PLOS Biology* 14:e1002397.

Patterson, C., Wissig, S., and Kohn, A. 2014. "Adaptation Disrupts Motion Integration in the Primate Dorsal Stream." *Neuron* 81:674–686.

Reig, R., Gallego, R., Nowak, L. G., and Sanchez-Vives, M. V. 2006. "Impact of cortical network activity on short-term synaptic depression." *Cerebral Cortex* 16:688–695.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. 2015. "ImageNet Large Scale Visual Recognition Challenge." *International Journal of Computer Vision* 115:211–252.

Sanchez-Vives, M. V., Nowak, L. G., and McCormick, D. A. 2000a. "Cellular mechanisms of long-lasting adaptation in visual cortical neurons in vitro." *The Journal of neuroscience : the official journal of the Society for Neuroscience* 20:4286–4299.

Sanchez-Vives, M. V., Nowak, L. G., and McCormick, D. A. 2000b. "Membrane mechanisms underlying contrast adaptation in cat area 17 in vivo." *Journal of Neuroscience* 20:4267–4285.

Sawamura, H., Orban, G. A., and Vogels, R. 2006. "Selectivity of neuronal adaptation does not match response selectivity: a single-cell study of the FMRI adaptation paradigm." *Neuron* 49:307–318.

Solomon, S. G. and Kohn, A. 2014. "Moving Sensory Adaptation beyond Suppressive Effects in Single Neurons." *Current Biology* 24:R1012–R1022.

Summerfield, C., Trittschuh, E. H., Monti, J. M., Mesulam, M. M., and Egner, T. 2008. "Neural repetition suppression reflects fulfilled perceptual expectations." *Nature Neuroscience* 11:1004–1006.

Tang, H., Schrimpf, M., Lotter, W., Moerman, C., Paredes, A., Ortega Caro, J., Hardesty, W., Cox, D., and Kreiman, G. 2018. "Recurrent computations for visual pattern completion." *Proceedings of the National Academy of Sciences* p. 201719397.

Todorovic, A. and de Lange, F. P. 2012. "Repetition suppression and expectation suppression are dissociable in time in early auditory evoked fields." *The Journal of Neuroscience* 32:13389–13395.

Ulanovsky, N., Las, L., and Nelken, I. 2003. "Processing of low-probability sounds by cortical neurons." *Nature neuroscience* 6:391–8.

Vinken, K., Op de Beeck, H. P., and Vogels, R. 2018. "Face Repetition Probability Does Not Affect Repetition Suppression in Macaque Inferotemporal Cortex." *The Journal of Neuroscience* 38:7492–7504.

Vinken, K., Vogels, R., and Op de Beeck, H. 2017. "Recent Visual Experience Shapes Visual Processing in Rats through Stimulus-Specific Adaptation and Response Enhancement." *Current Biology* 27:914–919.

Vogels, R. 2016. "Sources of adaptation of inferior temporal cortical responses." *Cortex* 80:185–195.

Wark, B., Lundstrom, B. N., and Fairhall, A. 2007. "Sensory adaptation." *Current Opinion in Neurobiology* 17:423–429.

Webster, M. A. 2015. "Visual Adaptation." *Annual Review of Vision Science* 1:547–567.

Webster, M. A., Kaping, D., Mizokami, Y., and Duhamel, P. 2004. "Adaptation to natural facial categories." *Nature* 428:557–561.

Webster, M. A. and MacLeod, D. I. A. 2011. "Visual adaptation and face perception." *Philosophical Transactions of the Royal Society B: Biological Sciences* 366:1702–1725.

Webster, M. A., Werner, J. S., and Field, D. J. 2005. "Adaptation and the Phenomenology of Perception." In *Fitting the Mind to the WorldAdaptation and After-Effects in High-Level Vision*, chapter 10, pp. 241–278. Oxford University Press.

Wehr, M. and Zador, A. M. 2005. "Synaptic mechanisms of forward suppression in rat auditory cortex." *Neuron* 47:437–445.

Whitmire, C. and Stanley, G. 2016. "Rapid Sensory Adaptation Redux: A Circuit Perspective." *Neuron* 92:298–315.

Wichmann, F. A. and Hill, N. J. 2001. "The psychometric function: I. Fitting, sampling, and goodness of fit." *Perception & Psychophysics* 63:1293–1313.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. 2014. "Performance-optimized hierarchical models predict neural responses in higher visual cortex." *Proceedings of the National Academy of Sciences* 111:8619–8624.