bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Understanding Biological Visual Attention Using Convolutional Neural Networks

Grace W. Lindsay[a,b], Kenneth D. Miller[a,b]

[a] *Center for Theoretical Neuroscience, College of Physicians and Surgeons, Columbia University, New York, New York, USA*
[b] *Mortimer B. Zuckerman Mind Brain Behavior Institute, College of Physicians and Surgeons, Columbia University, New York, New York, USA*

Abstract

Covert visual attention has been shown repeatedly to enhance performance on tasks involving the features and spatial locations to which it is deployed. Many neural correlates of covert attention have been found, but given the complexity of the visual system, connecting these neural effects to performance changes is challenging. Here, we use a deep convolutional neural network as a large-scale model of the visual system to test the effects of applying attention-like neural changes. Particularly, we explore variants of the feature similarity gain model (FSGM) of attention—which relates a cell's tuning to its attentional modulation. We show that neural modulation of the type and magnitude observed experimentally can lead to performance changes of the type and magnitude observed experimentally. Furthermore, performance enhancements from attention occur for a diversity of tasks: high level object category detection and classification, low level orientation detection, and cross-modal color classification of an attended orientation. Utilizing the full observability of the model we also determine how activity should change to best enhance performance and how activity changes propagate through the network. Through this we find that, for attention applied at certain layers, modulating activity according to tuning performs as well as attentional modulations determined by backpropagation. At other layers, attention applied according to tuning does not successfully propagate through the network, and has a weaker impact on performance than attention determined by backpropagation. This thus highlights a discrepancy between neural tuning and function.

1. Introduction

Covert visual attention, applied according to spatial location or visual features, has been shown repeatedly to enhance performance on challenging visual tasks [10]. To explore the neural mechanisms behind this enhancement, neural responses to the same visual input are compared under different task conditions. Such experiments have identified numerous neural modulations associated with attention, including changes in firing rates, noise levels, and correlated activity [89, 14, 23, 56], however, the extent to which these changes are responsible for behavioral effects is debated. Therefore, theoretical work has been used to link sensory processing changes to performance changes. While offering helpful insights, much of this work is either based on small, hand-designed models [67, 77, 92, 11, 30, 98, 29] or lacks direct mechanistic interpretability [97, 8, 88]. Here, we utilize a large-scale model of the ventral visual stream to explore the extent to which neural changes like those observed in the biology can

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

lead to performance enhancements on realistic visual tasks. Specifically, we use a deep convolutional neural network trained to perform object classification to test variants of the feature similarity gain model of attention [90].

Deep convolutional neural networks (CNNs) are popular tools in the machine learning and computer vision communities for performing challenging visual tasks [73]. Their architecture—comprised of layers of convolutions, nonlinearities, and response pooling—was designed to mimic the retinotopic and hierarchical nature of the mammalian visual system [73]. Models of a similar form have been used in neuroscience to study the biological underpinnings of object recognition for decades [25, 74, 83]. Recently it has been shown that when these networks are trained to successfully perform object classification on real-world images, the intermediate representations learned are remarkably similar to those of the primate visual system [100, 38, 37]. Specifically, deep CNNs are state-of-the-art models for capturing the feedforward pass of the ventral visual stream [39, 35, 9]. Many different studies have now built on this fact to further compare the representations [91, 50, 43] and behavior [44, 26, 72, 75, 49] of CNNs to that of biological vision. A key finding has been the correspondence between different areas in the ventral stream and layers in the deep CNNs, with early convolutional layers able to capture the representation of V1 and deeper layers relating to V4 and IT [28, 22, 81]. Given that CNNs reach near-human performance on visual tasks and have architectural and representational similarities to the visual system, they are particularly well-positioned for exploring how neural correlates of attention can impact behavior.

We focus here on attention's ability to impact activity levels (rather than noise or correlations) as these findings are straightforward to implement in a CNN. Furthermore, by measuring the effects of firing rate manipulations alone, we make clear what behavioral enhancements can plausibly be attributable to them.

One popular framework to describe attention's effects on firing rates is the feature similarity gain model (FSGM). This model, introduced by Treue & Martinez-Trujillo, claims that a neuron's activity is multiplicatively scaled up (or down) according to how much it prefers (or doesn't prefer) the properties of the attended stimulus [90, 55]. Attention to a certain visual attribute, such as a specific orientation or color, is generally referred to as feature-based attention (FBA) and its effects are spatially global: that is, if a task performed at one location in the visual field activates attention to a particular feature, neurons that represent that feature across the visual field will be affected [102, 79]. Overall, this leads to a general shift in the representation of the neural population towards that of the attended stimulus [16, 34, 70]. Spatial attention implies that a particular portion of the visual field is being attended. According to the FSGM, spatial location is treated as an attribute like any other. Therefore, a neuron's modulation due to attention can be predicted by how well its preferred features and spatial receptive field align with the features and location of the attended stimulus. The effects of combined feature and spatial attention have been found to be additive [32].

While the FSGM does describe many findings, its components are not uncontroversial. For example, it is questioned whether attention impacts responses multiplicatively or additively [5, 2, 51, 59], and whether or not the activity of cells that do not prefer the attended stimulus is actually suppressed [6, 67]. Furthermore, only a handful of studies have looked directly at the relationship between attentional modulation and tuning [55, 78, 12, 95]. Another unsettled issue is where in the visual stream attention

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

effects can be seen. Many studies of attention focus on V4 and MT/MST [89], as these areas have reliable attentional effects. Some studies do find effects at earlier areas [65], though they tend to be weaker and occur later in the visual response [36]. Therefore, a leading hypothesis is that attention signals, coming from prefrontal areas [64, 62, 3, 41], target later visual areas, and the feedback connections that those areas send to earlier ones causes the weaker effects seen there later [7, 51].

In this study, we define the FSGM of attention mathematically and implement it in a deep CNN. By testing different variants of the model, applied at different layers in the network and for different tasks, we can determine the ability of these neural changes to change behavior. Given the complexity of these large nonlinear networks, the effects of something like FSGM are non-obvious. Because we have full access to all units in the model, we can see how neural changes at one area propagate through the network, causing changes at others. This provides a fuller picture of the relationship between neural and performance correlates of attention.

2. Methods

2.1. Network Model

This work uses a deep convolutional neural network (CNN) as a model of the ventral visual stream. Convolutional neural networks are feedforward artificial neural networks that consistent of a few basic operations repeated in sequence, key among them being the convolution. The specific CNN architecture used in the study comes from [84] (VGG-16D) and is shown in Figure 1A. A previous variant of this work used a smaller network [47].

Here, the activity values of the units in each convolutional layer are the result of applying a 2-D spatial convolution to the layer below, followed by positive rectification (rectified linear 'ReLu' nonlinearity):

$$x_{ij}^{lk} = [(W^{lk} \star X^{l-1})_{ij}]_+ \tag{1}$$

where W^{lk} is the k^{th} convolutional filter at the l^{th} layer. The application of each filter results in a 2-D feature map (the number of filters used varies across layers and is given in parenthesis in Figure 1A). x_{ij}^{lk} is the activity of the unit at the i, j^{th} spatial location in the kth feature map at the l^{th} layer. X^{l-1} is thus the activity of all units at the layer below the l^{th} layer. The input to the network is a 224 by 224 pixel RGB image, and thus the first convolution is applied to these pixel values. For the purposes of this study the convolutional layers are most relevant, and will be referred to according to their numbering in Figure 1A.

Max pooling layers reduce the size of the feature maps by taking the maximum activity value of units in a given feature map in non-overlapping 2x2 windows.

The final three layers of this network are each fully-connected to the layer below them, with the number of units per layer given in parenthesis in Figure 1A. Therefore, connections exist from all units from all feature maps in the last convolutional layer (layer 13) to all 4096 units of the next layer, and so on. This network was pre-trained [24] using backpropagation on the ImageNet classification task, which involves doing 1000-way object categorization (for details see [84]). The final layer of the network thus contains 1000 units upon which a softmax classifier is used to output a ranked list of category labels for a given image. Looking at the top-5 error rate (wherein an

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

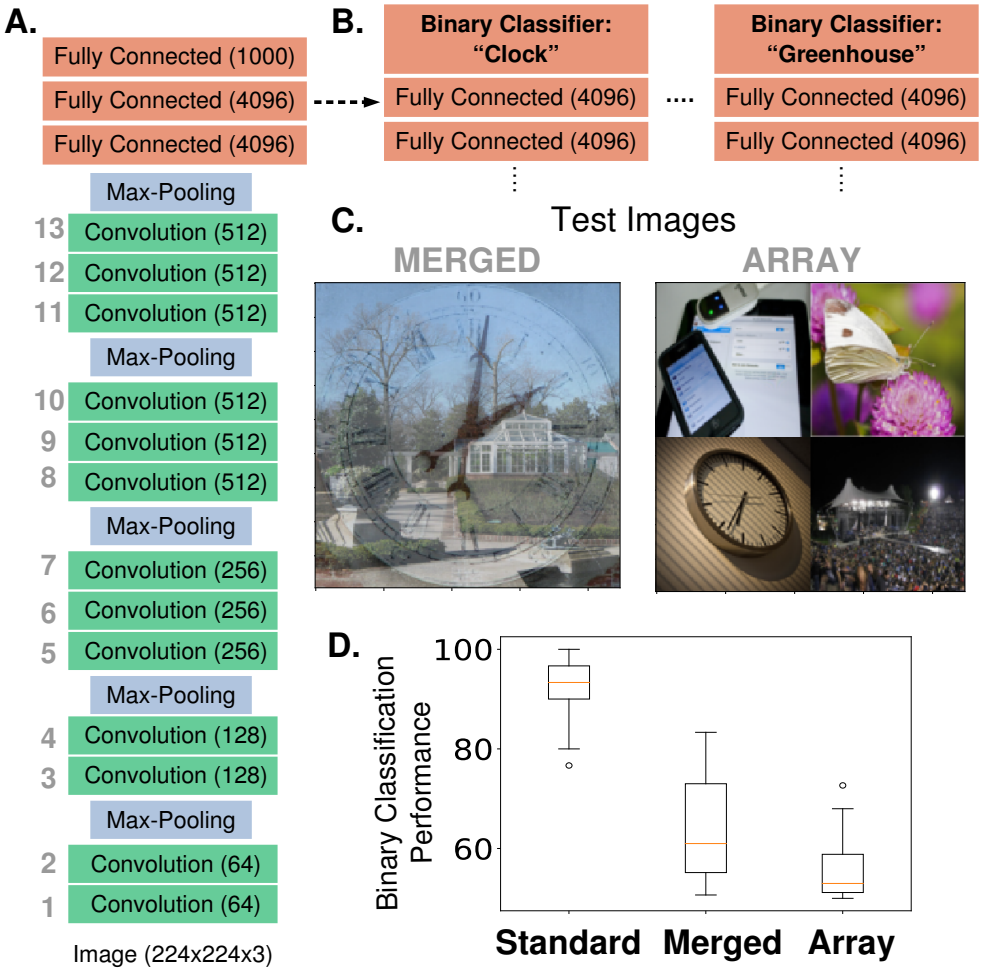

Figure 1: Network Architecture and Feature-Based Attention Task Setup. A.) The model used is a pre-trained deep neural network (VGG-16) that contains 13 convolutional layers (labeled in gray, number of feature maps given in parenthesis) and is pre-trained on the ImageNet dataset to do 1000-way object classification. All convolutional filters are 3x3. B.) Modified architecture for feature-based attention tasks. To perform our feature-based attention tasks, the final layer that was implementing 1000-way softmax classification is replaced by binary classifiers (logistic regression), one for each category tested (2 shown here). These binary classifiers are trained on standard ImageNet images. C.) Test images for feature-based attention tasks. Merged images (left) contain two transparently overlaid ImageNet images of different categories. Array images (right) contain four ImageNet images on a 2x2 grid. Both are 224 x 224 pixels. These images are fed into the network and the binary classifiers are used to label the presence or absence of the given category. D.) Performance of binary classifiers. Box plots describe values over 20 different object categories (median marked in red, box indicates lower to upper quartile values and whiskers extend to full range with outliers marked as dots). Standard images are regular ImageNet images not used in the binary classifier training set.

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

image is correctly labeled if the true category appears in the top five categories given by the network), this network achieves 92.7% accuracy.

2.2. Object Category Attention Tasks

The tasks we use to probe the effects of feature-based attention in this network involve determining if a given object category is present in an image or not, similar to tasks used in [86, 71, 40]. To have the network perform this specific task, we replaced the final layer in the network with a series of binary classifiers, one for each category tested (Figure 1B). We tested a total of 20 categories: paintbrush, wall clock, seashore, paddlewheel, padlock, garden spider, long-horned beetle, cabbage butterfly, toaster, greenhouse, bakery, stone wall, artichoke, modem, football helmet, stage, mortar, consomme, dough, bathtub. Binary classifiers were trained using ImageNet images taken from the 2014 validation set (and were therefore not used in the training of the original model). A total of 35 unique true positive images were used for training for each category, and each training batch was balanced with 35 true negative images taken from the remaining 19 categories. The results shown here come from using logistic regression as the binary classifier, though trends in performance are similar if support vector machines are used. Experimental results suggest that classifiers trained on unattended and isolated object images are appropriate for reading out attended objects in cluttered images [103].

Once these binary classifiers are trained, they are then used to classify more challenging test images. These test images are composed of multiple individual images (drawn from the 20 categories) and are of two types: "merged" and "array". Merged images are generated by transparently overlaying two images, each from a different category (specifically, pixel values from each are divided by two and then summed). Array images are composed of four separate images (all from different categories) that are scaled down to 112 by 112 pixels and placed on a two by two grid. The images that comprise these test images also come from the 2014 validation set, but are separate from those used to train the binary classifiers. See examples of each in Figure 1C. Test image sets are balanced (50% do contain the given category and 50% do not, 150 total test images per category). Both true positive and true negative rates are recorded and overall performance is the average of these rates.

To test the effects of spatial attention, only the "array" images are used. The task is to identify the category of the object at the attended location. Therefore, performance is measured using the original 1000-way classifier, with the category of the image in the attended quadrant as the true label (200 images were tested per quadrant).

2.3. Object Category Gradient Calculations

When neural networks are trained via backpropagation, gradients are calculated that indicate how a given weight in the network impacts the final classification. We use this same method to determine how a given unit's activity impacts the final classification. Specifically, we input a "merged" image (wherein one of the images belongs to the category of interest) to the network. We then use gradient calculations to determine the changes in activity that would move the 1000-way classifier toward classifying that image as belonging to the category of interest (i.e. rank that category highest). We average these activity changes over images and over all units in a feature map. This gives a single value per feature map:

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

$$g_c^{lk} = -\frac{1}{N_c}\sum_{n=1}^{N_c}\frac{1}{HW}\sum_{i=1,j=i}^{H,W}\frac{\partial E(n)}{\partial x_{ij}^{lk}(n)} \quad (2)$$

where H and W are the spatial dimensions of layer l and N_c is the total number of images from the category (here $N_C = 35$, and the merged images used were generated from the same images used to generate tuning curves, described below). $E(n)$ is the error of the classifier in response to image n, which is defined as the difference between the activity vector of the final layer (after the soft-max operation) and a one-hot vector, wherein the correct label is the only non-zero entry. Because we are interested in activity changes that would decrease the error value, we negate this term. The gradient value we end up with thus indicates how the feature map's activity would need to change to make the network more likely to classify an image as the desired category. Repeating this procedure for each category, we obtain a set of gradient values (one for each category, akin to a tuning curve), for each feature map: $\mathbf{g}^{lk}$. Note that, as these values result from applying the chain rule through layers of the network, they can be very small, especially for the earliest layers. For this study, the sign and relative magnitudes are of more interest than the absolute values.

2.4. Oriented Grating Attention Tasks

In addition to attending to object categories, we also test attention on simpler stimuli. In the orientation detection task, the network detects the presence of a given orientation in an image. Again, the final layer of the network is replaced by a series of binary classifiers, one for each of 9 orientations (0, 20, 40, 60, 80, 100, 120, 140, and 160 degrees. Gratings had a frequency of .025 cycles/pixel). The training sets for each were balanced (50% had only the given orientation and 50% had one of 8 other orientations) and composed of full field (224 by 224 pixel) oriented gratings of various colors (to increase the diversity of the training images, they were randomly degraded by setting blocks of pixels ranging uniformly from 0% to 70% of the image to 0 at random). Test images were each composed of two oriented gratings of different orientation and color (color options: red, blue, green, orange, purple). Each of these gratings were of size 112 by 112 pixels and placed randomly in a quadrant while the remaining two quadrants were black (Figure 6A). Again, the test sets were balanced and performance was measured as the average of the true positive and true negative rates (100 test images per orientation).

These same test images were used for a cross-modal attention task wherein the network had to classify the color of the grating that had the attended orientation. For this, the final layer of the network was replaced with a 5-way softmax color classifier. This color classifier was trained using the same full field oriented gratings used to train the binary classifiers (therefore, the network saw each color at all orientation values). The test sets contained images that all had the attended orientation as one of the two gratings (125 images per orientation). Performance was measured as the percent of trials wherein the color classifier correctly ranked the color of the attended grating highest (top-1 error).

Finally, for one analysis, a joint feature and spatial attention task was used. This task is almost identical to the setup of the orientation detection task, except that the searched-for orientation would only appear in one of the four quadrants. Therefore, performance could be measured when applying feature attention to the searched-for orientation, spatial attention to the quadrant in which it could appear, or both.

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

2.5. How Attention is Applied

This study aims to test variations of the feature similarity gain model of attention, wherein neural activity is modulated by attention according to how much the neuron prefers the attended stimulus. To replicate this in our model, we therefore must first determine the extent to which units in the network prefer different stimuli ("tuning values"). When attention is applied to a given category, for example, units' activities are modulated according to these values. We discuss below the options for how exactly to implement that modulation.

2.5.1. Tuning Values

To determine tuning to the 20 object categories used, we presented the network with images of each object category (the same images on which the binary classifiers were trained) and measured the relative activity levels.

Specifically, for the k^{th} feature map in the l^{th} layer, we define $r^{lk}(n)$ as the activity in response to image n, averaged over all units in the feature map (i.e., over the spatial dimensions). Averaging these values over all images in the training sets (N_c = 35 images per category, 20 categories. N=700) gives the mean activity of the feature map $\bar{r}^{lk}$:

$$\bar{r}^{lk} = \frac{1}{N}\sum_{n=1}^{N} r^{lk}(n) \tag{3}$$

Tuning values are defined for each object category, c as:

$$f_c^{lk} = \frac{\frac{1}{N_c}\sum_{n\in c} r^{lk}(n) - \bar{r}^{lk}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(r^{lk}(n) - \bar{r}^{lk})^2}} \tag{4}$$

That is, a feature map's tuning value for a given category is merely the average activity of that feature map in response to images of that category, with the mean activity under all image categories subtracted and standard deviation divided. These tuning values determine how the feature map is modulated when attention is applied to the category. Taking these values as a vector over all categories, $\mathbf{f}_{lk}$, gives a tuning curve for the feature map. We define the overall tuning quality of a feature map as its maximum absolute tuning value: $max(|\mathbf{f}_{lk}|)$. We also define the category with the highest tuning value as that feature map's most preferred, and the category with the lowest (most negative) value as the least or anti-preferred.

We apply the same procedure to generate tuning curves for orientation and for color by using the full field gratings used to train the orientation detection and color classification classifiers. The orientation tuning values were used when applying attention in these tasks. The color tuning curves were generated only to measure color tuning and its quality in the network.

When measuring how correlated tuning values are with gradient values, shuffled comparisons are used. To do this shuffling, correlation coefficients are calculated from pairing each feature map's tuning values with a random other feature map's gradient values.

2.5.2. Gradient Values

In addition to applying attention according to tuning, we also attempt to generate the "best possible" attentional modulation by utilizing gradient values. These gradient

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

values are calculated slightly differently from those described above (2.3), because they are meant to represent how feature map activity should change in order to increase overall task performance, rather than just increase the chance of classifying an image as a certain object or orientation.

The error functions used to calculate gradient values for the category and orientation detection tasks were for the binary classifiers associated with each object/orientation. A balanced set of test images was used. Therefore a feature map's gradient value for a given object/orientation is the averaged activity change that would increase binary classification performance for that object/orientation. Note that on images that the network already classifies correctly, gradients are zero. Therefore, the gradient values are driven by the errors: false negatives (classifying an image as not containing the category when it does) and false positives (classifying an image as containing the category when it does not). In our detection tasks, the former error is more prevalent than the latter, and thus is the dominant impact on the gradient values.

The same procedure was used to generate gradient values for the color classification task. Here, gradients were calculated using the 5-way color classifier: for a given orientation, the color of that orientation in the test image was used as the correct label, and gradients were calculated that would lead to the network correctly classifying the color. Averaging over many images of different colors gives one value per orientation that represents how a feature map's activity should change in order to make the network better at classifying the color of that orientation.

In both of the orientation tasks, the test images used for gradient calculations (50 images per orientation) differed from those used to assess performance. For the object detection task, images used for gradient calculations were merged images (45 per category) drawn from the same pool as, but different from, those used to test detection performance.

2.5.3. Spatial Attention

In the feature similarity gain model of attention, attention is applied according to how much a cell prefers the attended feature, and location is considered a feature like any other. In CNNs, each feature map results from applying the same filter at different spatial locations. Therefore, the 2-D position of a unit in a feature map represents more or less the spatial location to which that unit responds. Via the max-pooling layers, the size of each feature map shrinks deeper in the network, and each unit responds to a larger area of image space, but the "retinotopy" is still preserved. Thus, when we apply spatial attention to a given area of the image, we enhance the activity of units in that area of the feature maps (and, as we discuss below, possibly decrease the activity of units in other areas). In this study, spatial attention is tested using array images, and thus attention is applied to a given quadrant of the image.

2.5.4. Implementation Options

The values discussed above determine how strongly different feature maps or units should be modulated under different attentional conditions. We will now lay out the different implementation options for that modulation.

First, the modulation can be multiplicative or additive. That is, when attending to category c, the slope of the rectified linear units can be multiplied by a weighted function of the tuning value for category c:

$$x_{ij}^{lk} = (1 + \beta f_c^{lk})[(I_{lk}^{ij})]_+ \tag{5}$$

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

with I_{lk}^{ij} representing input to the unit coming from layer $l-1$. Alternatively, a weighted version of the tuning value can be added before the rectified linear unit:

$$x_{ij}^{lk} = [I_{ij}^{lk} + \mu_l \beta f_c^{lk}]_+ \tag{6}$$

Strength of attention is varied via the weighting parameter, β. For the additive effect, manipulations are multiplied by μ_l, the average activity level across all units of layer l in response to all images (for each of the 13 layers respectively: 20, 100, 150, 150, 240, 240, 150, 150, 80, 20, 20, 10, 1). When gradient values are used in place of tuning values, we normalize them by the maximum value at a layer, to be the same order of magnitude as the tuning values: $\mathbf{g}^l / max(|\mathbf{g}^l|)$.

Note that for feature-based attention all units in a feature map are modulated the same way, as feature attention has been found to be spatially global. In the case of spatial attention, object category tuning values are not used. Rather, the tuning value term is set to +1 if the i, j position of the unit is in the attended quadrant and to -1 otherwise. For feature attention tasks, β ranged from 0 to a maximum of 11.85 (object attention) and 0 to 4.8 (orientation attention). For spatial attention tasks, it ranged from 0 to 2.

Next, we chose whether attention only enhances units that prefer the attended feature/location, or also decreases activity of those that don't prefer it. For the latter, the tuning values are used as-is. For the former, the tuning values are positively-rectified: $[\mathbf{f}^{lk}]_+$.

Combining these two factors, there are four implementation options: additive positive-only, multiplicative positive-only, additive bidirectional, and multiplicative bidirectional.

The final option is the layer in the network at which attention is applied. We try attention at all convolutional layers individually and simultaneously (when applying simultaneously the strength range tested is a tenth of that when applying to a single layer).

Note that when gradient values were used, only results from using multiplicative bidirectional effects are reported (when tested on object category detection, multiplicative effects performed better than additive when using gradient values).

2.6. Signal Detection Calculations

For the joint spatial-feature attention task, we calculated criteria (c, "threshold") and sensitivity (d') using true (TP) and false (FP) positive rates as follows [52] :

$$c = -.5(\Phi^{-1}(TP) + \Phi^{-1}(FP)) \tag{7}$$

where Φ^{-1} is the inverse cumulative normal distribution function. c is a measure of the distance from a neutral threshold situated between the mean of the true negative and true positive distributions. Thus, a positive c indicates a stricter threshold (fewer inputs classified as positive) and a negative c indicates a more lenient threshold (more inputs classified as positive).

$$d' = \Phi^{-1}(TP) - \Phi^{-1}(FP) \tag{8}$$

This measures the distance between the means of the distributions for true negative and two positives. Thus, a larger d' indicates better sensitivity.

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

When necessary, a correction was applied wherein false positive rates of 0 were set to .01 and true positive rates of 1 were set to .99.

2.7. "Recording" Procedures

We examined the effects that applying attention at certain layers in the network (specifically 2, 6, 8, 10, and 12) has on activity of units at other layers. We do this for many different circumstances, using multiplicative bidirectional attention with $\beta = .5$ unless otherwise stated.

2.7.1. Unimodal Task Recording Setup

This recording setup is designed to mimic the analysis of [55]. Here, the images presented to the network are full-field oriented gratings of all orientation-color combinations. Feature map activity is measured as the spatially averaged activity of all units in a feature map in response to an image. Activity in response to a given orientation is further averaged over all colors. Each feature map's preferred (most positive tuning value) and anti-preferred (most negative tuning value) orientations are determined. Activity is recorded when attention is applied to the preferred or anti-preferred orientation and activity ratios are calculated. According to the FSGM, the ratio of activity when the preferred orientation is attended over when the anti-preferred is attended should be greater than one and the same regardless of whether the image is of the preferred or anti-preferred orientation. According to the feature matching (FM) model, the ratio of the activity when attending the presented orientation over attending an absent orientation should be greater than one and similar regardless of whether the orientation is preferred or not. We measure all of these ratios, and the fraction of total feature maps which show FM behavior, when attention is applied according to tuning values or gradient values.

As in [55], we also look at a measure of activity changes across all orientations. We calculate the ratio of activity when attention is applied to a given orientation (and the orientation is present in the image) over activity in response to the same image when no attention is applied. These ratios are then organized according to orientation preference: the most preferred is at location 0, then the average of next two most preferred at location 1, and so on with the average of the two least preferred orientations at location 4 (the reason for averaging of pairs is to match [55] as closely as possible). Fitting a line to these points gives a slope and intercept for each feature map. FSGM predicts a negative slope and an intercept greater than one.

We also calculate the same activity ratios described above when the images presented are standard (single image) ImageNet images from each of the 20 categories (activity is averaged over 5 images per category). Attention is applied according to object category tuning values or to gradient values for binary classification as described in 2.5.2.

2.7.2. Cross-modal Task Recording Setup

Cross-modal tasks involve attending to one modality (here, space or orientation) and reading out another (category or color, respectively). Specifically, in the first task, activity is recorded when spatial attention is applied to a given quadrant. Here, the activity for each feature map is averaged only over units in the quadrant that matches the attended quadrant. The images used are array images with 6 examples of each object category in the attended quadrant (for a total of 120 images). Activity ratios are

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

calculated as the activity when the recorded quadrant is attended over activity when no attention is applied. The average ratio for each category is organized according to category preference for each feature map and a line is fit to these points. The intercept (measured here as the true intercept minus one) and difference (slope multiplied by the number of categories minus one, 19) are calculated for each feature map. FSGM predicts a positive intercept and zero slope, because responses to all categories should be scaled equally by spatial attention.

The second cross-modal task setup involves measuring color encoding in different attention conditions. Here, images similar to those used in the orientation detection and color classification tasks are used. Specifically, images are generated that have two oriented gratings in two of the four quadrants. One is oriented at 160 degrees and the other nearly orthogonal at 80. All pairs of colors are generated for the two gratings (thus the two gratings may have the same color, which is a difference from the stimuli used in the orientation tasks). Activity is organized according to the color of the 160 degree grating (and averaged over the colors of the 80 degree grating), in order from most to least preferred color for each feature map. Lines were fit to these points in two cases: when attention was directed to 80 degrees and when it was directed to 160 degrees. We then asked if attention to 160 degrees led to better encoding of the color of the 160 degree stimulus compared to attention to 80 degrees. We considered a feature map to have better color encoding of the 160 degree grating if its mean increased (a stronger overall signal, measured as the activity value at the middle of the line) and if its slope became more negative (stronger differentiation between colors). Results are similar if only the latter condition is used. We measure the encoding changes for two separate populations of feature maps: those that prefer 160 degrees and those that anti-prefer it (most negative tuning value). Stimuli at 160 degrees were chosen as the focus of this analysis because across all layers there are roughly equal numbers of feature maps that prefer and anti-prefer it. Percent of feature maps that have better encoding were measured when attention was applied according to orientation tuning values or color classification gradient values.

In all cases, lines are fit using the least squares method, and any activity ratios with zero in the denominator were discarded.

2.8. Experimental Data

Model results were compared to previously published data coming from several studies. In [54], a category detection task was performed using stereogram stimuli (on object present trials, the object image was presented to one eye and a noise mask to another). The presentation of the visual stimuli was preceded by a verbal cue that indicated the object category that would later be queried (cued trials) or by meaningless noise (uncued trials). After visual stimulus presentation, subjects were asked if an object was present and, if so, if the object was from the cued category (categories were randomized for uncued trials). In Experiment 1, the object images were line drawings (one per category) and the stimuli were presented for 1.5 sec. In Experiment 2, the object images were grayscale photographs (multiple per category) and presented for 6 sec. True positives were counted as trials wherein a given object category was present and the subject correctly indicated its presence when queried. False positives were trials wherein no category was present and subjects indicated that the queried category was present.

In [53], a similar detection task is used. Here, subjects detect the presence of an

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

uppercase letter that is (on target present trials) presented rapidly and followed by a mask. Prior to the visual stimulus, a visual or audio cue indicated a target letter. After the visual stimulus, the subjects were required to indicate whether any letter was present. True positives were trials in which a letter was present and the subject indicated it (only uncued trials or validly cued trials—where the cued letter was the letter shown—were considered here). False positives were trials where no letter was present and the subject indicated that one was.

The task in [40] is also an object category detection task. Here, an array of several images was flashed on the screen with one image marked as the target. All images were color photographs of objects in natural scenes. In certain blocks, the subjects knew in advance which category they would later be queried about (cued trials). On other trials, the queried category was only revealed after the visual stimulus (uncued). True positives were trials in which the subject indicated the presence of the queried category when it did exist in the target image. False positives were trials in which the subject indicated the presence of the cued category when it was not in the target image. Data from trials using basic category levels with masks were used for this study.

Finally, we include one study using macaques wherein both neural and performance changes were measured [57]. In this task, subjects had to report a change in orientation that could occur in one of two stimuli. On cued trials, the change occurred in the cued stimulus in 80% of trials and the uncued stimulus in 20% of trials. On neutrally-cued trials, subjects were not given prior information about where the change was likely to occur (50% at each stimulus). Therefore performance could be compared under conditions of low (uncued stimuli), medium (neutrally cued stimuli), and high (cued stimuli) attention strength. Correct detection of an orientation change in a given stimulus (indicated by a saccade) is considered a true positive and a saccade to the stimulus prior to any orientation change is considered a false positive. True negatives are defined as correct detection of a change in the uncued stimulus (as this means the subject correctly did not perceive a change in the stimulus under consideration) and false negatives correspond to a lack of response to an orientation change.

In cases where the true and false positive rates were not published, they were obtained via personal communications with the authors.

3. Results

The ability to manipulate activities as well as measure performance on complicated visual tasks make CNNs a great testing ground for theories of attention. CNNs trained on visual object recognition learn representations that are similar to those of the ventral stream. The network used in this study was explored in [28], where it was shown that early convolutional layers of this CNN are best at predicting activity of voxels in V1, while late convolutional layers are best at predicting activity of voxels in the object-selective lateral occipital area (LO). In addition, CNN architecture makes comparison to biological vision straightforward. For example, the application of a given convolutional filter results in a feature map, which is a 2-D grid of artificial neurons that represent how well the bottom-up input aligns with the filter at each location. Therefore a "retinotopic" layout is built into the structure of the network, and the same visual features are represented across that retinotopy (akin to how cells that prefer different orientations exist at all locations across the V1 retinotopy). We

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

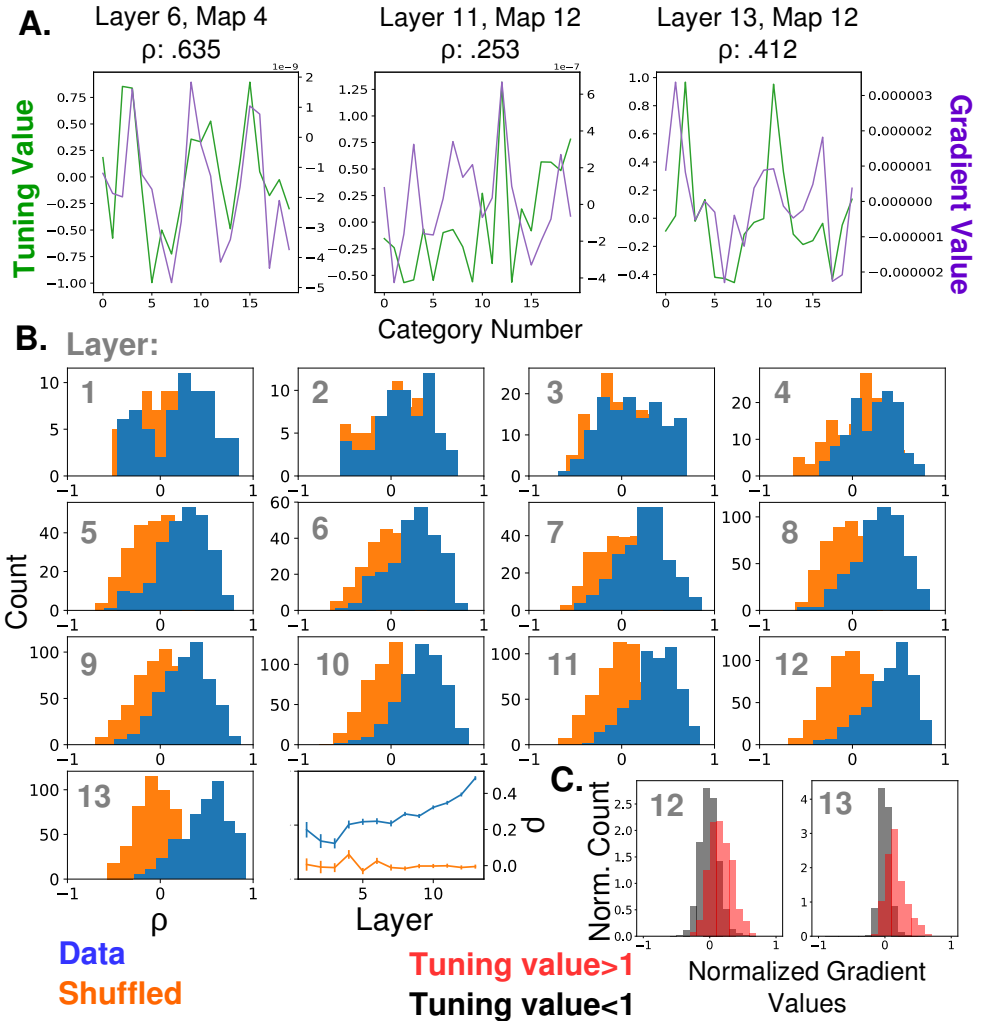

Figure 2: Relationship Between Feature Map Tuning and Gradients. A.) Example tuning values (green, left axis) and gradient values (purple, right axis) of three different feature maps from three different layers (identified in titles, layers as labeled in Fig 1A) over the 20 tested object categories. Correlation coefficients between tuning curves and gradient values given in titles. B.) Histograms of correlation coefficients across all feature maps at each layer (blue) along with shuffled comparisons (orange). Final subplot shows average correlation coeffecients across layers (errorbars +/- S.E.M.). C.) Distributions of gradient values when tuning is strong. In red, histogram of gradient values associated with tuning values larger than one, across all feature maps in layer 12 (left) and 13 (right). For comparison, histograms of gradient values associated with tuning values less than one are shown in black (counts are separately normalized for visibility, as the population in black is much larger than that in red).

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

utilize these properties to test variants of the feature similarity gain model (FSGM) on a diverse set of visual tasks that are challenging for the network. We also take advantage of the full observability of this network model to compare the FSGM to "optimal" attentional manipulation, as determined by backpropagation calculations.

3.1. The Relationship between Tuning and Classification

The feature similarity gain model of attention posits that neural activity is modulated by attention in proportion to how strongly a neuron prefers the attended features, as assessed by its tuning. However, the relationship between a neuron's tuning and its ability to influence downstream readouts remains a difficult one to investigate biologically. We use our hierarchical model to explore this question directly. We do so by calculating gradient values, which we compare to tuning curves (see Methods Sections 2.3 and 2.5.1 for details). These gradient values indicate the way in which activity of a feature map should change in order to make the network more likely to classify an image as being of a certain object category. If there is a correspondence between tuning and classification, a feature map that prefers a given object category (that is, responds strongly to it compared to other categories) should also have a high positive gradient value for that category. In Figure 2A we show gradient values and tuning curves for three example feature maps. In Figure 2B, we show the distribution of correlation coefficients between tuning values and gradient values for all feature maps at each of the 13 convolutional layers. As can be seen in the final subplot, on average, tuning curves show higher than expected correlation with gradient values at all layers (compared to shuffled controls). Furthermore, this correlation increases with later layers. While the correlation between tuning and gradient values suggests that a feature map's response is indicative of its functional role, the correspondence is not perfect. In Figure 2C, we show the gradient values of feature maps at layers 12 and 13, segregated according to tuning value. In red are gradient values that correspond to tuning values greater than one (for example, category 12 for the feature map in the middle pane of Figure 2A). As these distributions show, strong tuning values can be associated with weak or even negative gradient values. Negative gradient values indicate that increasing the activity of that feature map makes the network less likely to categorize the image as the given category. Therefore, even feature maps that strongly prefer a category (and are only a few layers from the classifier) still may not be involved in its classification, or even be inversely related to it.

3.2. Feature-based Attention Improves Performance on Challenging Object Classification Tasks

To determine if manipulation according to tuning values can enhance performance, we created challenging visual images composed of multiple objects for the network to classify. These test images are of two types: merged (two object images transparently overlaid, such as in [82]) or array (four object images arranged on a grid) (see Figure 1C for an example of each). The task for the network is to detect the presence or absence of a given object category in these images. It does so using a series of binary classifiers trained on standard images of these objects, which replace the last layer of the network (Figure 1B). The performance of these classifiers on the test images indicates that this is a challenging task for the network (Figure 1D), and thus a good opportunity to see the effects of attention. Without attention, the average performance of the binary classifiers across all categories is 64.4% on merged images and 55.6%

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

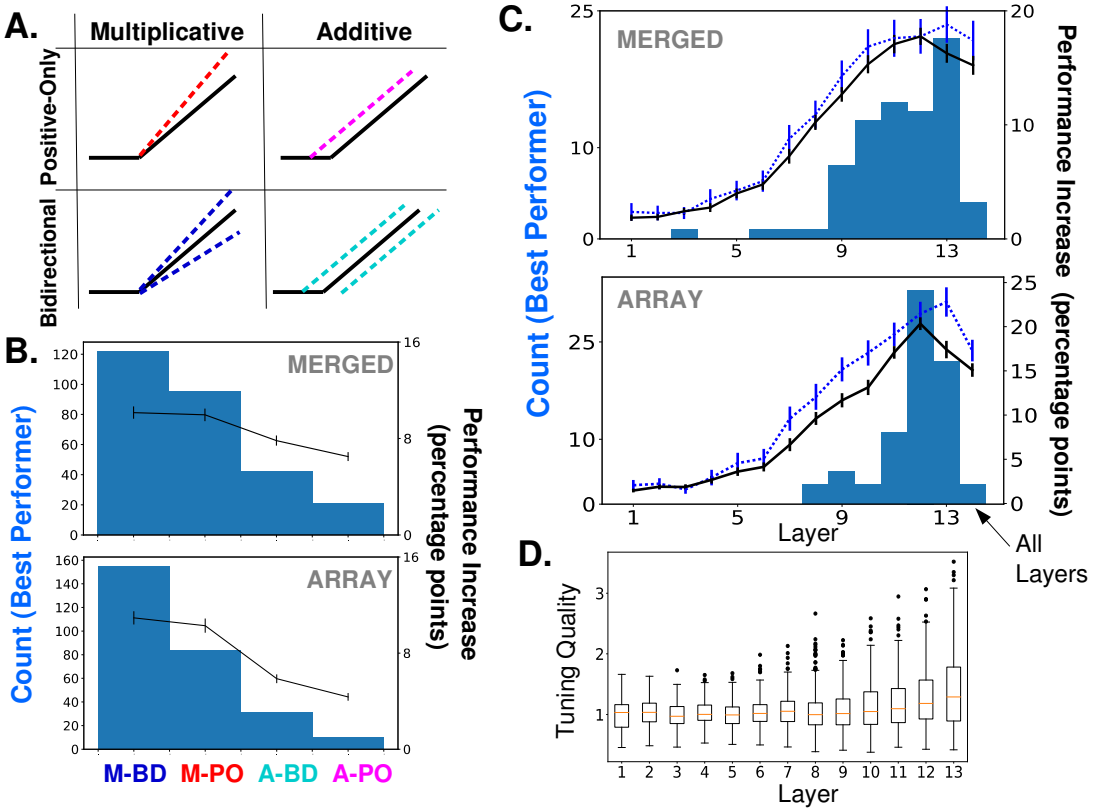

Figure 3: Effects of Applying Feature-Based Attention on Object Category Tasks. A.) Schematics of how attention can modulate the activity function. Feature-based attention modulates feature maps according to their tuning values but this modulation can scale the activity multiplicatively or additively, and can either only enhance feature maps that prefer the attended category (positive-only) or also decrease the activity of feature maps that do not prefer it (bidirectional). B.) Considering the combination of attention applied to a given category at a given layer as an instance (20 categories * 14 layer options = 280 instances), histograms (left axis) show how often the given option is the best performing, for merged (top) and array (bottom) images. Average increase in binary classification performance for each option also shown (right axis, averaged across all instances, errorbars +/- S.E.M.) C.)Comparison of performance effects of layer options. Considering each instance as the combination of attention applied to a given category using a given implementation option (20 categories * 4 implementation options = 80 instances), histograms show how often applying attention to the given layer is the best performing, for merged (top) and array (bottom) images. The final column corresponds to attention applied to all layers simultaneously with the same strength (strengths tested are one-tenth of those when strength applied to individual layers). Average increase in binary classification performance for each layer also shown in black (right axis, errorbars +/- S.E.M.). Average performance increase for MBD option only shown in blue. In all cases, best performing strength from the range tested is used for each instance. D.) Tuning quality across layers. Tuning quality is defined per feature map as the maximum absolute tuning value of that feature map. Box plots show distribution across feature maps for each layer.

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

on array (compared to a chance performance of 50%, as the test sets contained the attended category 50% of the time).

We implement feature-based attention in this network by modulating the activity of feature maps according to how strongly they prefer the attended object category (see Methods 2.5.1). While tuning values determine the relative strength and direction of the modulation, there are still options regarding how to implement those changes. We test additive effects (wherein attention alters the activity of a feature map by the same amount regardless of its activity level) and multiplicative effects (attention changes the slope of the activity function). We also consider the situation where attention only increases the activity of feature maps that prefer the attended category (i.e., have a positive tuning value), or when attention also decreases the activity of feature maps that do not prefer the attended category. Taken together this leads to four implementation options: additive positive-only, multiplicative positive-only, additive bidirectional, and multiplicative bidirectional (see Figure 3A for depictions of each, and Methods 2.5.4 for details). A final option is the choice of convolutional layer at which these manipulations are applied.

To determine which of these attention mechanisms is best, attention is applied to each object category and the performance of the binary classifier associated with that category is compared with and without the different activity manipulations. The results of this are shown in Figure 3B and C (the best performing strength, including 0 if necessary, is assumed for each category. See Methods for details).

As Figure 3B shows, multiplicative bi-directional effects are best able to enhance performance, measured in terms of the number of times that the multiplicative bidirectional option beats out the other three options when compared for the same category and layer (blue histogram). The second best option is multiplicative positive-only, then additive bidirectional, and additive positive-only. This ordering is the same when looking at the average increase in performance (black line), however, the differences between multiplicative bi-directional and multiplicative positive-only performance are not significant. Furthermore, these trends are identical regardless of whether tested on merged (top) or array (bottom) images, though the differences are starker for array images.

Figure 3C shows a similar analysis but across layers at which attention is applied. Again, the trends are the same for merged and array images and show a clear increase in performance as attention is applied at later layers in the network (numbering is as in 1A). Across all implementation options, attention at layer 12 best increases average performance (black lines). However this is driven by the additive implementations. We show the average performance increase with layer for multiplicative bi-directional effects alone (blue dotted line). For this the final layer is best, leading to an 18.8% percentage point increase in binary classification on the merged image task and 22.8% increase on the array task.

The trends in performance track trends in tuning quality shown in 3D. That is, layers with better object category tuning lead to better performance when attention is applied at them. They also track the correlation between tuning values and gradient values, as that correlation increases with later layers.

Overall, the best performing options for implementing attention—multiplicative bidirectional effects applied at later layers—are in line with what has been observed biologically and described by the feature similarity gain model [90, 56].

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

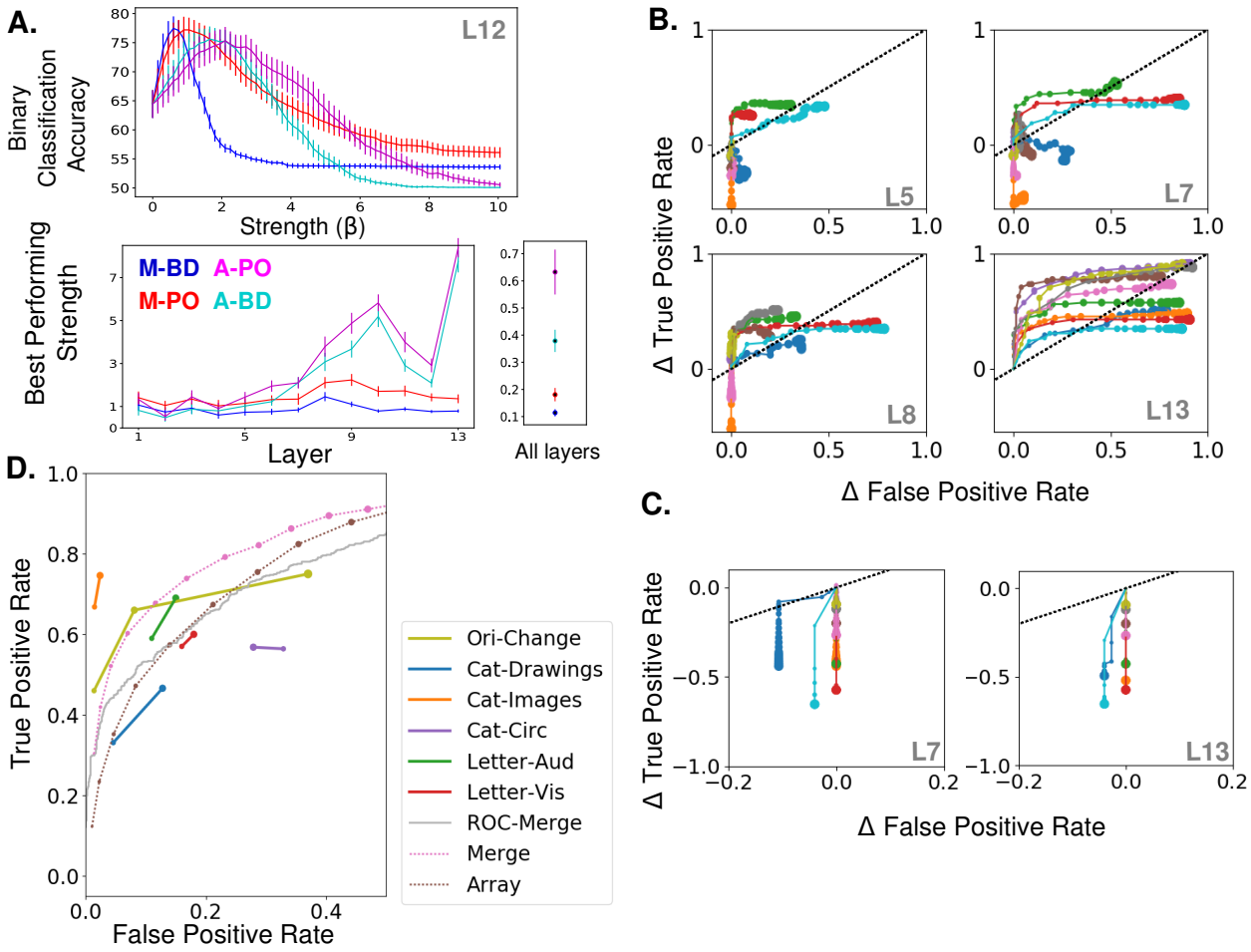

Figure 4: Effects of Varying Attention Strength in Feature-Based Attention Tasks. A.) Effect of strength on different implementation options. On the top, performance averaged over categories (errorbars +/- S.E.M.) shown as a function of the strength parameter, β, for each implementation option. Attention is applied to layer 12 and on merged images. The location of the peak for each category individually is the best performing strength for that category. On the bottom, the best performing strength averaged across categories (errorbars +/- S.E.M.) at each layer for each implementation option. When applied at all layers simultaneously, the range of attention strength tested was smaller. Color scheme as in Figure 1A. B.) and C.) multiplicative bidirectional attention is used, on merged images. B.) Effect of strength increase in true- and false-positive rate space for each of four layers (layer indicated in bottom right of each panel). Each line represents performance changes that arise from applying attention to a different category (only 10 categories shown for visibility), with each increase in dot size representing a .15 increase in strength. Baseline (no attention) values are subtracted for each category such that all start at (0,0) and the layer attention is applied to is indicated in gray. The black dotted line represents equal changes in true and false positive rates. C.) Effect of strength increase in true- and false-positive rate space when tuning values are negated. Same as B, but with sign of attention effects switched (only attention at layer 7 and 13 shown). D.) Comparisons from experimental data. The true and false positive rates from four previously published studies are shown for conditions of increasing attentional strength (solid lines). True and false positive rates are shown for merged and array images (dotted lines, averaged over categories) when attention is applied with increasing strengths (starting at 0, each increasing dot size equals .15 increase in β) at layer 13 (multiplicative bidirectional effects). Receiver operator curve for merged images shown in gray. Cat-Drawings=[54], Exp. 1; Cat-Images=[54],Exp. 2; Objects=[40], Letter-Aud.=[53], Exp. 1; Letter-Vis.=[53], Exp. 2. Ori-Change=[57]. See Methods for details of experiments.

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

3.3. Strength of Attention Influences True and False Positive Tradeoff

As mentioned above, strength is a relevant variable when implementing attention. Specifically, the strength parameter, which we call β, scales the tuning values to determine how strongly attention modulates activities (in the case of additive effects, this value is further multiplied by the average activity level of the layer before being added to the response). We tested a range of β values and the analysis in Figure 3 assumes the best-performing β for each combination of category, layer, and implementation option. Here, we look at how performance changes as the strength varies.

Figure 4A (top) plots the increase in binary classification performance (averaged across all categories) as a function of strength for the four different implementation options, when attention is applied at layer 12 for merged images (results similar for array images). From this we can see that not only is the multiplicative bidirectional manipulation the best performing, it also reaches its peak at a lower strength than the other options.

On the bottom of Figure 4A, we show the best performing strength (calculated for each category individually and averaged) across layers, and when applied to all layers simultaneously. It is clear from this analysis that multiplicative bidirectional effects consistently require lower strength to reach maximum performance than other options. Furthermore, the fact that the best performing strengths occur below the peak strength tested ($\beta = 11.85$ for individual layers and $\beta = 1.19$ for all layers simultaneously) indicates that any performance limitations are not due to a lack of strength. The best performing strength for additive attention at layer 13 is surprisingly high. To understand why this may be, it is important to remember that, when using additive attention, the attention value added to each unit's response is the product of the relevant tuning value, β, and the average activity level of the layer. This is necessary because average activity levels vary by 2 orders of magnitude across layers. The variability of activity across feature maps, however, is much higher at layer 13 compared to layers 1 through 12. This makes the mean activity level used to calculate attention effects less reliable, which may contribute to why higher β values are needed.

Performance can change in different ways with attention. In Figure 4B we break the binary classification performance down into true and false positive rates. Here, each colored line indicates a different category and increasing dot size indicates increasing strength of attention (multiplicative bidirectional effects used). True and false positive rates in the absence of attention have been subtracted such that all categories start at (0,0). Ideally, true positives would increase without an equivalent increase (and possibly with a decrease) in false positive rates. If they increase in tandem (i.e., follow the black dotted lines) then attention would not have a net beneficial effect on performance.

Looking at the effects of applying attention at different layers (layer labeled in gray), we can see that attention at lower layers is less effective at moving the performance in this space, and that movement is in somewhat random directions. As attention is applied at later layers, true positive rates are more likely to increase and the increase in false positive rates is delayed. Thus, when attention is applied with modest strength at layer 13, most categories see a substantial increase in true positives with only modest increases in false positives. As strength continues to increase however, false positives increase substantially and eventually lead to a net decrease in overall classifier performance (i.e., cross the black dotted line). Without attention the false negative rate is $69.7 \pm 21.8\%$ and decreases to $19.9 \pm 10\%$ using the best perform-

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

ing strength for each category. Without attention the false positive rate is $1.4 \pm 3.1\%$ and increases to $13.7 \pm 7.7\%$ using the best performing strength for each category.

To confirm that these behavioral enhancements result from the targeted effects of attention, rather than a non-specific effect of activity manipulation, we apply multiplicative bi-directional attention using negated tuning values. Because tuning values sum to zero over all feature maps and categories, using negated tuning values doesn't change the overall level of positive and negative modulation applied to the network. Applying attention this way, however, leads to unambiguously different results. Figure 4C shows these results, plotted in the same format as Figure 4B, for attention at layers 7 and 13. Using negated tuning values leads to a decrease in true and false positive values with increasing attention strength. Thus, attention appears to function as a knob that can turn true and false positives up or down in an intuitive way.

It would be useful to know how the magnitude of neural activity changes in our model compare to those used by the brain. Experimentally, the strength of attention can be manipulated by controlling the presence and/or validity of cues [57], switching attention from the non-preferred to preferred stimulus can have large effects on firing rate (111% increase in MT [45]). Before the presentation of a target array, cells in IT showed a 40% increase in firing when the to-be-detected object was preferred versus non-preferred [12]. Of most direct relevance to this study, however, is the modulation strength when switching from no or neutral attention to specific feature-based attention, rather than switching attention from a non-preferred to a preferred stimulus. In [55], neurons in MT showed an average increase in activity of 7% when attending their preferred motion direction (and similar decrease when attending the non-preferred) versus a neutral attention condition.

In our model, when $\beta = .75$ (roughly the value at which performance with multiplicative bidirectional effects peaks at later layers), given the magnitude of the tuning values (average magnitude: .38), attention scales activity by an average of 28.5%. This value refers to how much activity is modulated in comparison to a the $\beta = 0$ condition. This $\beta = 0$ condition is probably more comparable to passive or anesthetized viewing, as task engagement has been shown to scale neural responses generally [69]. This complicates the relationship between modulation strength in our model and the values reported in the data.

To allow for a more direct comparison, in Figure 4D, we have collected the true and false positive rates obtained experimentally during different object detection tasks (explained in detail in Methods), and plotted them in comparison to the model results. The first five studies plotted in Figure 4D come from human studies. In all of these studies, uncued trials are those in which no information about the upcoming visual stimulus is given, and therefore attention strength is assumed to be low. In cued trials, the to-be-detected category is cued before the presentation of a challenging visual stimulus, allowing attention to be applied to that object or category. The tasks range from detecting simple, stereotyped stimuli (e.g. letters) to highly-varied photographic instances of a given category. Not all changes in performance were statistically significant, but we plot them here to show general trends.

The majority of these experiments show a concurrent increase in both true and false positive rates as attention strength is increased. The rates in the uncued conditions (smaller dots) are generally higher than the rates produced by the $\beta = 0$ condition in our model, which suggests that neutrally cued conditions do indeed correspond to a value of $\beta > 0$. We can determine the average β value for the neutral and cued

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

conditions by projecting the data values onto the nearest point on the model line (each dot on the model line corresponds to an increase in β of .15). Specifically, we project the values from the four datasets whose experiments are most similar to our merged image task (Cat-Drawings, Cat-Images, Letter-Aud, and Letter-Vis) onto the model line generated from using the merged images. Through this, we find that the average β value for the neutral conditions is .39 and for the attended conditions .53. Because attention scales activity by $1 + \beta f_c^{lk}$ (where f_c^{lk} is the tuning value and the average tuning value magnitude is .38), these changes correspond to a $\approx$5% change in activity. Thus, the size of observed performance changes is broadly consistent with the size of observed neural changes.

Among the experiments used, the one labeled "Cat-Images" is an outlier, as it has much higher true positive and lower true negative rates than the model can achieve simultaneously. This experimental setup is the one most similar to the merged images used in the model (subjects are cued to attend a given category and grayscale category images are presented with a concurrent noise mask), however, the images were presented for 6 seconds. This presumably allows for several rounds of feedback processing, which our purely feedforward model cannot capture. Notably though, true and false positive rate still increase with attention in this ask.

Another exception is the experiment labeled as "Cat-Circ", which has a larger overall false positive rate and shows a decrease in false positives with stronger attention. In this study, a single target image is presented in a circular array of distractor images, and the subject may be cued ahead of time as to which object category will need to be detected in that target image. The higher false positive rates in this experiment may be attributable to the fact that the distractors were numerous and were pixelated versions of real images. Attention's ability to decrease false positives, however, suggests a different mechanism than the one modeled here. The reason for this difference is not clear. However, in this experiment, the cued trials were presented in blocks wherein the same category was to be detected in each trial, whereas for the uncued trials, the to-be-detected category changed trialwise. The block structure for the attended trials may have allowed for a beneficial downstream adaptation to the effects of attention, which reined in the false positive rate.

The last dataset included in the plot (Ori-Change) differs from the others in several ways. First, it comes from a macaque study that also measured neural activity changes, which allows for a direct exploration of the relationship between neural and performance effects. The task structure is different as well: subjects had to detect an orientation change in one of two stimuli. For cued trials, the change occurs at the cued stimulus on 80% of trials. Attention strength could thus be low (for the uncued stimuli on cued trials), medium (for both stimuli on neutrally-cued trials), or high (for the cued stimuli on cued trials). While this task includes a spatial attention component, it is still useful as a test of feature-based attention effects. Previous work has demonstrated that, during a change detection task, feature-based attention is deployed to the pre-change features of a stimulus [15, 58]. Therefore, because the pre-change stimuli are of differing orientations, the cueing paradigm used here controls the strength of attention to orientation as well. So, while this task differs somewhat from the one performed by the model, it can still offer broad insight into how the magnitude of neural changes relates to the magnitude of performance changes.

We plot the true positive (correct change detection) and false positive (premature response) rates as a function of strength as the yellow line in 4D. Like the other

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

studies, this study shows a concurrent increase in both true and false positive rates with increasing attention strength. According to recordings from V4 taken during this task, average firing rates increase by 3.6% between low and medium levels of attention. To achieve the performance change observed between these two levels the model requires a roughly 12% activity change. This gap may indicate the role of other biologically observed effects of attention (e.g., on Fano Factor and correlations) in performance enhancement, or the smaller effect in the data may be due to the averaging of both positive and negative changes (because the stimuli were optimized for a subset of the recorded neurons, positive changes would be expected on average). Firing rates increased by 4.1% between medium and high attention strength conditions. For the model to achieve the observed changes in true positive rates alone between these levels requires a roughly 6% activity change. However, the data shows a very large increase in false positives between these two attention strengths, which would require a roughly 20% activity change in the model. This high rate of false positives points to a possible effect of attention downstream of sensory processing.

Finally, we show in this plot the change in true and false positive rates when the threshold of the final layer binary classifier is varied (a receiver operating characteristic analysis. No attention was applied during this analysis). The gray line in Figure 4D shows this analysis for merged images. Comparing this to the effect of varying attention strength (pink line), it is clear that varying the strength of attention applied at the final convolutional layer has more favorable performance effects than altering the classifier threshold. This points to the role of attentional modulation in sensory areas, rather than targeting only downstream "readout" areas.

Overall, the findings from these studies suggest that much of the change in true and false positive rates observed experimentally could be attributed to moderately-sized changes in neural activity in sensory processing areas. However, it is clear that the details of the experimental setup are relevant, both for the absolute performance metrics and how they change with attention [67].

An analysis of performance changes in the context of signal detection theory (sensitivity and criteria) will come later.

3.4. Spatial Attention Increases Object Categorization Performance

In addition to feature-based attention, we also test the effects of spatial attention in this network. For this, we use our array images, and the task of the network is to correctly classify the object category in the attended quadrant of the image. Therefore, the original final layer of the network which performs 1000-way object categorization is used (Figure 5A). The same implementation and layer options were tested and compared to 1000-way classification performance without attention (see Methods 2.5.4). However, tuning values were not used; rather, because the spatial layout of activity is largely conserved in CNNs, an artificial neuron was assumed to "prefer" a given quadrant of the image if that unit was in the corresponding quadrant of the feature map.

In Figure 5B, the performance (classification was considered correct if the true category label appeared in the top five categories outputted by the network, but trends are the same for top-1 error) is shown as a function of attention strength for each of the four options. The layer at which attention is applied is indicated by the line color. Because tuning values are not used for the application of spatial attention, the β value can be interpreted directly as the amount of activity modulation due to attention

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

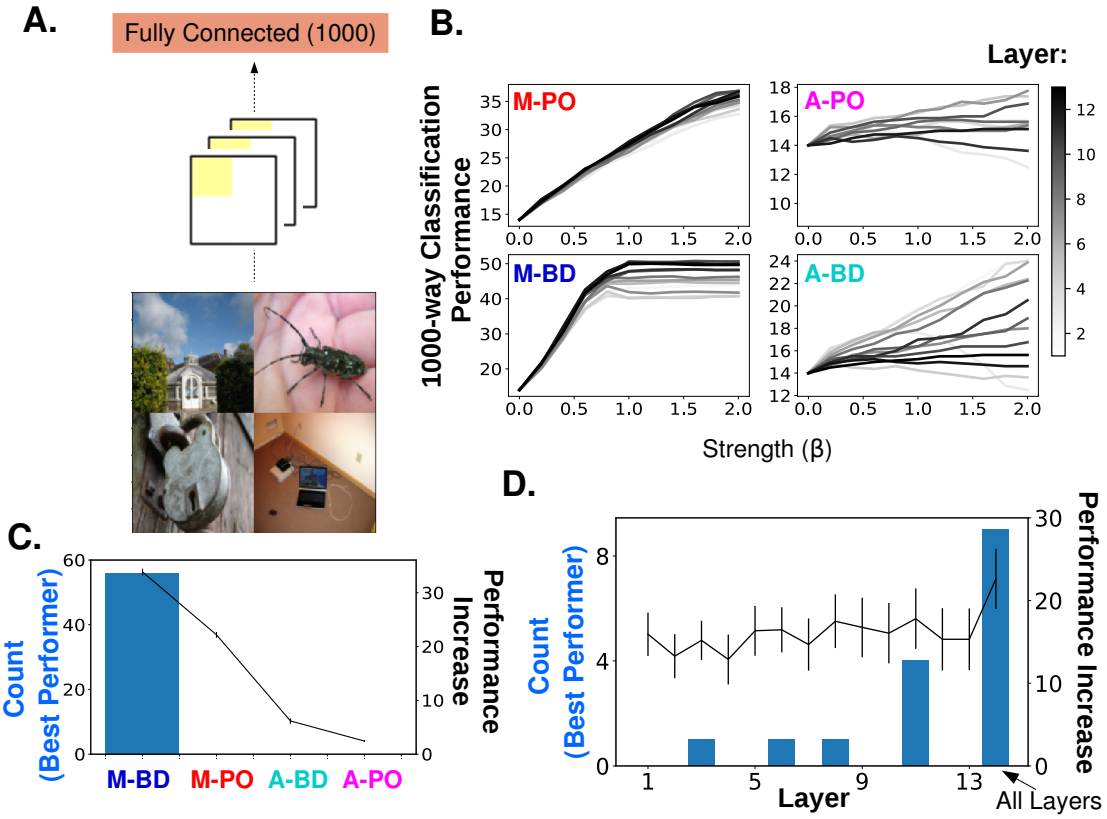

Figure 5: Spatial Attention Task and Results. A.) Array images were used to test spatial attention. Performance was measured as the ability of the original 1000-way classifier to identify the category in the attended quadrant (measured as top-5 error). Attention was applied according to the spatial layout of the feature maps (for example, when attending to the upper left quadrant of the image, units in the upper left quadrant of the feature maps are enhanced). B.) 1000-way classification performance as a function of attention strength, when applied at different layers (indicated by line darkness) and for each of the four attention options. C.) Comparison of performance effects of attention options (using best performing strength). Histograms (left axis) show how often the given option is the best performing (over 4 quadrants * 14 layer options = 56 instances). Average increase in 1000-way classification performance for each option also shown (right axis, errorbars +/- S.E.M.). D.) Histograms (over 4 quadrants * 4 implementation options = 16 instances) show how often the applying attention to the given layer is the best performing. The final column corresponds to attention applied to all layers simultaneously (strength at one-tenth that of strength applied to individual layers). Average increase in 1000-way classification performance for each layer also shown (right axis, errorbars +/- S.E.M.).

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(recall that for multiplicative effects rates are multiplied by $1 + \beta$).

Using experimentally-observed performance changes to relate our model to data (as we did in Figure 4D) is more challenging for the spatial attention case because the specific tasks used are more varied. Using the performance on trials with a neutral spatial cue as a baseline, we report the impact of spatial attention as the factor by which performance increases on trials with valid spatial cues. Experimentally, spatial attention scales performance by ≈19% on a color recognition task [27], ≈16% on an orientation categorization task [19], ≈10% on an orientation classification task [76] and a gap detection task [63], and ≈3.3% on a red line detection task [87]. Spatial attention effects range in magnitude but have been shown to increase neural activity by ≈20% in several studies [60, 17] when calculated for attend-in versus attend-out conditions. If we assume that attend-in and attend-out conditions scale activity in opposite directions (bi-directional effects) but with equal magnitude from a baseline [57], then spatially cued trials should have a roughly 10% change in activity compared to neutral trials. As mentioned above, the $\beta = 0$ condition in our model is not necessarily comparable to a neutrally-cued condition experimentally, so it is unclear what performance level in our model should be used as a baseline. However, going from $\beta = 0$ to $\beta = .1$ enhances performance from 14% correct to an average (across attention at each layer) of 17.4% correct. This is a 24.2% increase in accuracy stemming from a 22% change in activity on attend-in versus attend-out conditions. Again, these simple calculations suggest that the experimentally-observed magnitude of neural modulations could indeed lead to the observed magnitude of behavioral changes.

It is also of note that performance in the case of multiplicative bidirectional effects plateaus around $\beta = 1$, yet for multiplicative positive-only effects it continues to climb. This suggests that the suppressing of the three non-attended quadrants is a strong driver of the performance changes when using multiplicative bidirectional effects, as this suppression is complete at $\beta = 1$ (i.e., activity is 100% silenced at that value). While it is not believed that spatial attention leads to complete silencing of cells representing unattended locations, these results highlight the potential importance of scaling such activity downward.

Figure 5C and D summarize the performance enhancements that result from different options (assuming the best performing strengths, as in Figure 3B and C). Unlike feature-based attention, spatial attention is relatively insensitive to the layer at which it is applied, but is strongly enhanced by using multiplicative bidirectional effects compared to others. This discrepancy makes sense when we consider that spatial attention tasks are cross-modal—that is, they involve attending to one dimension (space) and reading out another (object category)—whereas the object detection tasks used above are unimodal—the same dimension (object category) is attended to and read out. In a cross-modal task it is not valuable just to amplify the attended attribute, but rather to amplify the information carried by the attended attribute. Assuming the absolute difference in rates across cells is relevant for encoding object identity, multiplicative effects amplify these informative differences and can thus aid in object classification in the attended quadrant. In a system with noise, attention's benefits would depend on the extent to which it simultaneously enhanced the non-informative noise. Experimentally, attention leads to a decrease in mean-normalized variance in firing across trials [14].

Another difference between feature-based and spatial attention is the effect of applying attention at all layers simultaneously. When applying attention at all layers,

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

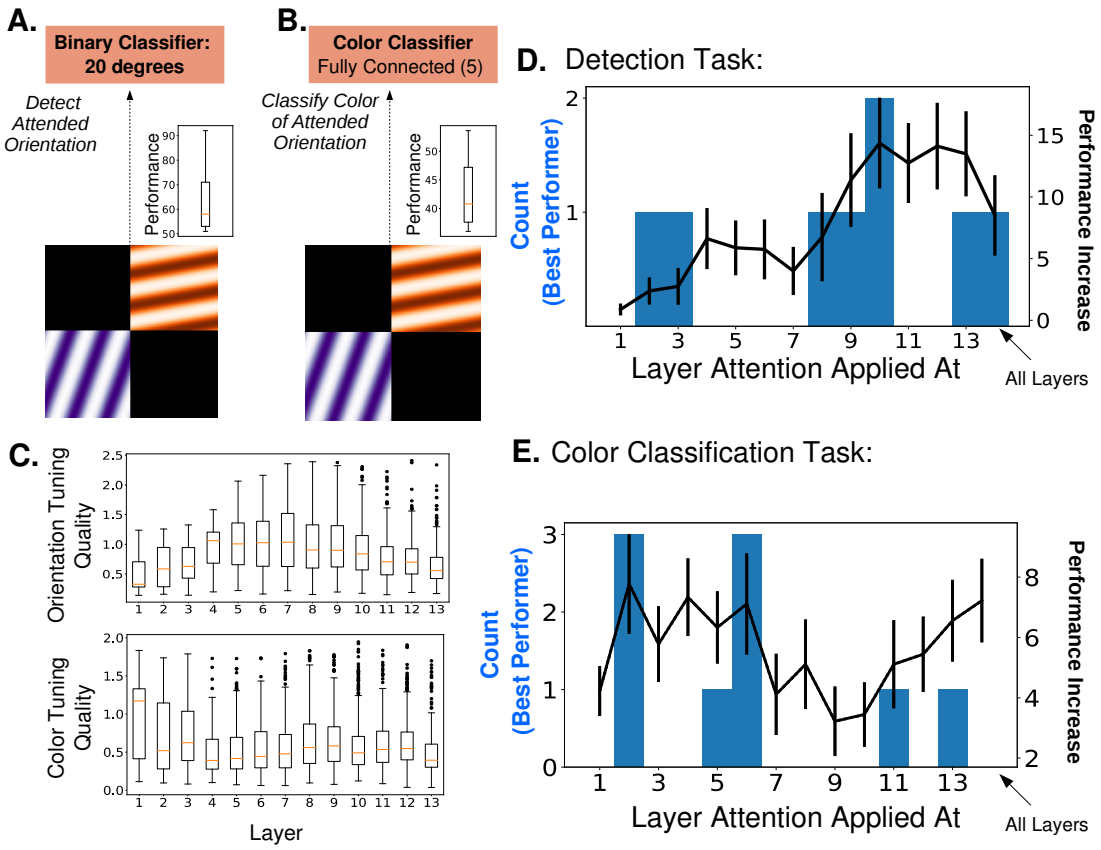

Figure 6: Attention Tasks and Results Using Oriented Gratings. A.) Orientation detection task. Like with the object category detection tasks, separate binary classifiers trained to detect each of 8 different orientations replaced the final layer of the network. Test images included 2 oriented gratings of different color and orientation located at two of 4 quadrants. Insets show performance over 9 orientations without attention B.) Color classification task. The final layer of the network is replaced by a single 5-way color classifier. The same test images are used as in the detection task and performance is measured as the ability of the classifier to identify the color of the attended orientation. Inset shows performance over 9 orientations without attention (chance is 25%) C.) Orientation tuning quality (top) and color tuning quality (bottom) as a function of layer. D.) Comparison of performance on detection task when attention (determined by orientation tuning values) is applied at different layers. Histogram of best performing layers in blue, average increase in binary classification performance in black. E.) Comparison of performance on color classification task when attention (determined by orientation tuning values) is applied at different layers. Histogram of best performing layers in blue, average increase in 5-way classification performance in black. Errorbars are +/- S.E.M.

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the β values tested are one-tenth that of when attention is applied at individual layers. Despite this weakened strength, applying attention at all layers leads to better performance in the spatial attention task than applying it to any layer individually. In the feature-based attention task, this is not the case (Figure 3C). This difference is explored more directly later.

3.5. Feature-based Attention Enhances Performance on Orientation Detection and Color Classification Tasks

Some of the results presented above, particularly those related to the layer at which attention is applied, may be influenced by the fact that we are using an object categorization task. To see if results are comparable using simpler stimuli, we created an orientation detection task (Figure 6A), wherein binary classifiers trained on full field oriented gratings are tested using images that contain two gratings of different orientation and color. The performance of these binary classifiers without attention is above chance (distribution across orientations shown in inset of Figure 6A). The performance of the binary classifier associated with vertical orientation (0 degrees) was abnormally high (92% correct without attention, other orientations average 60.25%) and this orientation was excluded from further analysis for the detection task.

Attention is applied according to orientation tuning values of the feature maps (tuning quality by layer is shown in Figure 6C) and tested across layers (using multiplicative bidirectional effects). We find that the trend in this task is similar to that of the object task: applying attention at later layers leads to larger performance increases (14.4% percentage point increase at layer 10). This is despite the fact that orientation tuning quality peaks in the middle layers.

We also explore a cross-modal attention task that is in line with the style of certain attention experiments in neuroscience and psychology [78, 66, 96]. Specifically, the task for the network is to readout the color of the stimulus in the image with the attended orientation (Figure 6B, mean 5-way classification performance without attention: 42.89%). Thus, attention is applied according to orientation tuning values, but the final layer of the network is a 5-way color classifier. This is akin to studies where the task of the subject is, for example, to report a speed change in random dots that are moving in the attended direction. Interestingly, in this case attention applied at earlier layers (specifically layers 2-6, best performance increase is 7.8 percentage points at layer 2) performs best. Color tuning quality is stronger at earlier layers as well (layers 1-3 particularly).

The β values that lead to peak performance in the detection task at later layers ranges from .5 to 1. Given that β scales the tuning values and the average tuning value magnitude at later layers is .32, the average modulation strength (compared to the $\beta = 0$ condition) is 16%-32%. For the color classification task the successful modulation at earlier layers ranges from 13-28%. Therefore the two different tasks require similar modulations.

3.6. Gradient Values Offer Performance Comparison

Previously, we used gradient values to determine if object category tuning values were related to classification behavior. Here, we use a similar procedure to obtain gradient values that tell us how feature map activity should change in order to make the network better at the tasks of orientation detection and color classification (see

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

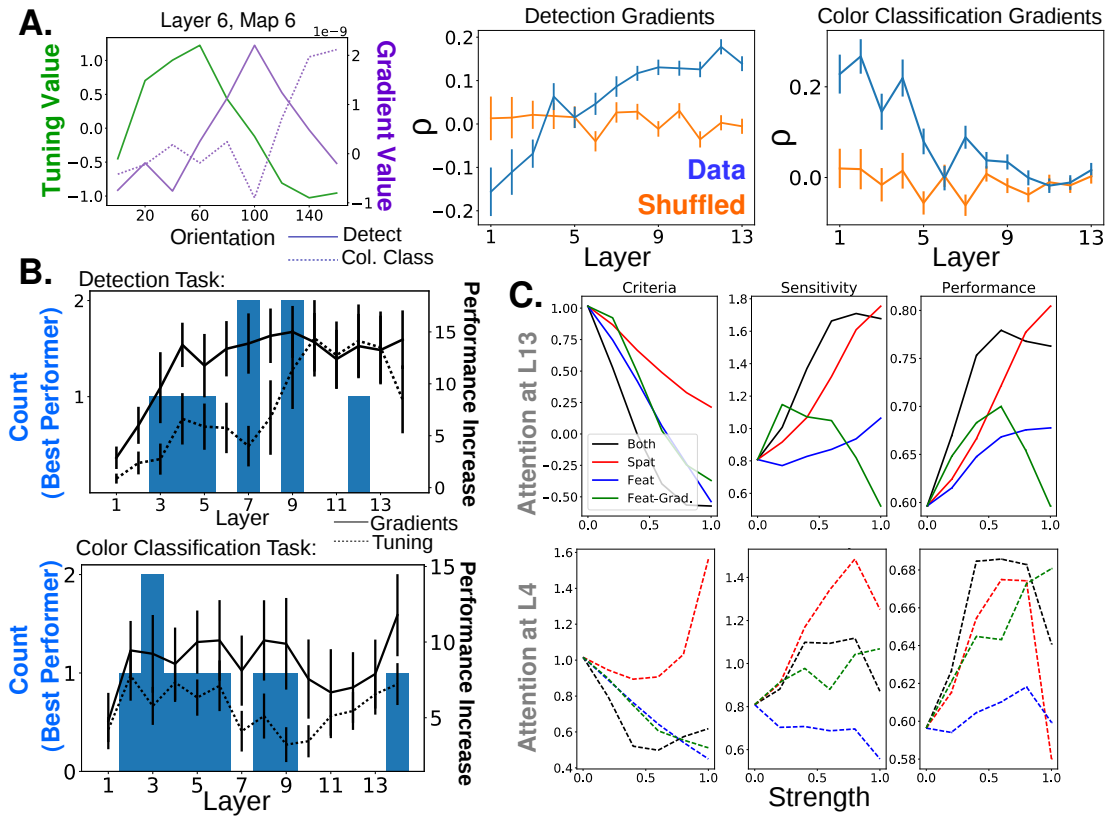

Figure 7: Comparison of Orientation Task Gradient Values to Tuning Values. A.) Correlation of gradient values with tuning values for the detection and color classification tasks. On the left, an example feature maps orientation tuning curve (green) and curves generated from detection gradient values (solid purple) and color classification gradient values (dashed purple). Correlation coefficients with tuning curve are -.196 and -.613, respectively. Average correlation coefficient values between tuning curves and detection gradient curves (middle) and color classification gradient curves (right) across layers (blue). Shuffled correlation values in orange. Errorbars are +/- S.E.M. B.) Comparison of performance on detection task when attention is determined by detection gradient values and applied at different layers (top). Comparison of performance on color classification task when attention is by determined by color classification gradient values and applied at different layers (bottom). Histograms of best performing layers in blue, average increase in binary or 5-way classification performance in black. Errorbars are +/- S.E.M. In both, performance increase when attention is determined by tuning values is shown for comparison (dashed lines). Only multiplicative bidirectional effects are used. C.) Change in signal detection values when attention is applied in different ways (spatial, feature according to tuning, both spatial and feature according to tuning, and feature according to gradient values) for the task of detecting a given orientation at a given quadrant. Top row is when attention is applied at layer 13 and bottom when applied at layer 4 (multiplicative bidirectional effects).

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Methods 2.5.2). We then use these values in place of the orientation tuning values when applying attention, and compare the performances.

In Figure 7A, we first show the extent to which these gradient values correlate with the tuning values. On the left, an example feature map's tuning curve (green) along with curves generated from gradient values for the orientation detection task (solid purple) and color classification task (dashed purple). The middle and right panels show the average correlation coefficients between tuning curves and the respective gradient values across layers. Correlation with orientation detection gradients peaks at later layers, while correlation with color classification gradients peaks at early layers. In Figure 7B, the solid lines and histograms document the performance using gradient values. For comparison, the dashed lines give the performance improvement from using the tuning values. In the orientation detection task, gradient values perform better than tuning values at earlier layers, but the performance difference vanishes at later layers (where the tuning values and gradient values are most correlated). Thus, tuning values can actually reach the same performance level as the gradient values suggesting that, while they are not identical to the values determined by the gradient calculations, they are still well-suited for increasing detection performance. The performance for color classification using gradient values has the reverse pattern. It is most similar to the performance using tuning values at earlier layers (where the two are more correlated), and the performance gap is larger at middle layers. At all layers, the mean performance using gradient values is larger than that using tuning values.

The results of applying this procedure to the object category detection task are discussed later (Figure 8E).

3.7. Feature-based Attention Primarily Influences Criteria and Spatial Attention Primarily Influences Sensitivity

Signal detection theory is frequently used to characterize the effects of attention on performance [94]. Here, we use a joint feature-spatial attention task to explore effects of attention in the model. The task uses the same 2-grating stimuli described above. The same binary orientation classifiers are used and the task of the model is to determine if a given orientation is in a given quadrant. Performance is then measured when attention is applied according to orientation, space, or both (effects are combined additively), and two key signal detection measurements are computed. Criteria is a measure of how lenient the threshold that's used to mark an input as a positive is. Sensitivity is a measure of how separate the two populations of true positive and negatives are.

Figure 7C shows how these values, along with the overall binary classification performance, vary with the strength and type of attention applied at two example layers. Performance is best when both spatial and feature-based attention are applied simultaneously. The ways in which these two types of attention affect performance can be teased apart by looking at their effects when applied separately. Criteria decreases more when feature-based attention is applied alone than when spatial is. Sensitivity increases more for spatial attention alone than feature-based attention alone. These general trends hold regardless of the layer at which attention is applied, though when applied at layer 4, feature-based attention alone actually decreases sensitivity.

Applying feature-based attention according to orientation detection gradient values has a very similar effect on criteria as applying it with tuning values. The effect

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

on sensitivity however, is slightly different, as the gradient values are better able to increase sensitivity. Therefore, attending to feature using gradient values leads to slightly better overall performance than when using tuning values in this example.

Various impacts of attention on sensitivity and criteria have been found experimentally. Task difficulty (an assumed proxy for attentional strength) was shown to increase both sensitivity and criteria [85]. Spatial attention increases sensitivity and (less reliably) decreases criteria [31, 20]. A study that looked explicitly at the different effects of spatial and category-based attention [86] found that, in line with our results, spatial attention increases sensitivity more than category-based attention (most visible in their Experiment 3c, which uses natural images) and that the effects of the two are additive.

The diversity of results in the literature (including discrepancies with our model) may be attributed to different task types and to the fact that attention is known to impact neural activity in various ways beyond pure sensory areas [42]. This idea is borne out by a study that aimed to isolate the neural changes associated with sensitivity and criteria changes [52]. In this study, the authors designed behavioral tasks that encouraged changes in sensitivity or criteria exclusively: high sensitivity was encouraged by associating a given stimulus location with higher overall reward, while high criteria was encouraged by rewarding correct rejects more than hits (and vice versa for low sensitivity/criteria). Differences in V4 neural activity were observed between trials using high versus low sensitivity stimuli. No differences were observed between trials using high versus low criteria stimuli. This indicates that areas outside of the ventral stream (or at least outside V4) are capable of impacting criteria. Importantly, it does not mean that changes in V4 don't impact criteria, but merely that those changes can be countered by downstream processes. Indeed, to create sessions wherein sensitivity was varied without any change in criteria, the authors had to increase the relative correct reject reward (i.e., increase the criteria) at locations of high absolute reward, presumably to counter the decrease in criteria that appeared naturally as a result of attention-induced neural changes in V4 (similarly, they had to decrease the correct reject reward at low reward locations). Our model demonstrates clearly how such effects from sensory areas alone can impact detection performance, which, in turn highlights the role downstream areas play in determining the final behavioral outcome.

3.8. Recordings Show How Feature Similarity Gain Effects Propagate

To explore how attention applied at one location in the network impacts activity later on, we apply attention at various layers and "record" activity at others (Figure 8A). In particular, we record activity of feature maps at all layers while applying multiplicative bidirectional attention at layers 2, 6, 8, 10, and 12 individually. The results of these recordings show both which features of the activity changes are correlated with performance enhancements as well as how FSGM effects at one area can lead to very different effects at another.

Activity was recorded in response to multiple different stimuli and attentional conditions. In Figure 8B we explore whether applying feature attention according to the FSGM at one layer continues to have FSGM-like effects at later layers. To do this we use an analysis taken from [55]. Specifically, full field oriented gratings were shown to the network that were either of the preferred (most positive tuning value) or anti-preferred (most negative tuning value) orientation for a given feature map. Attention

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

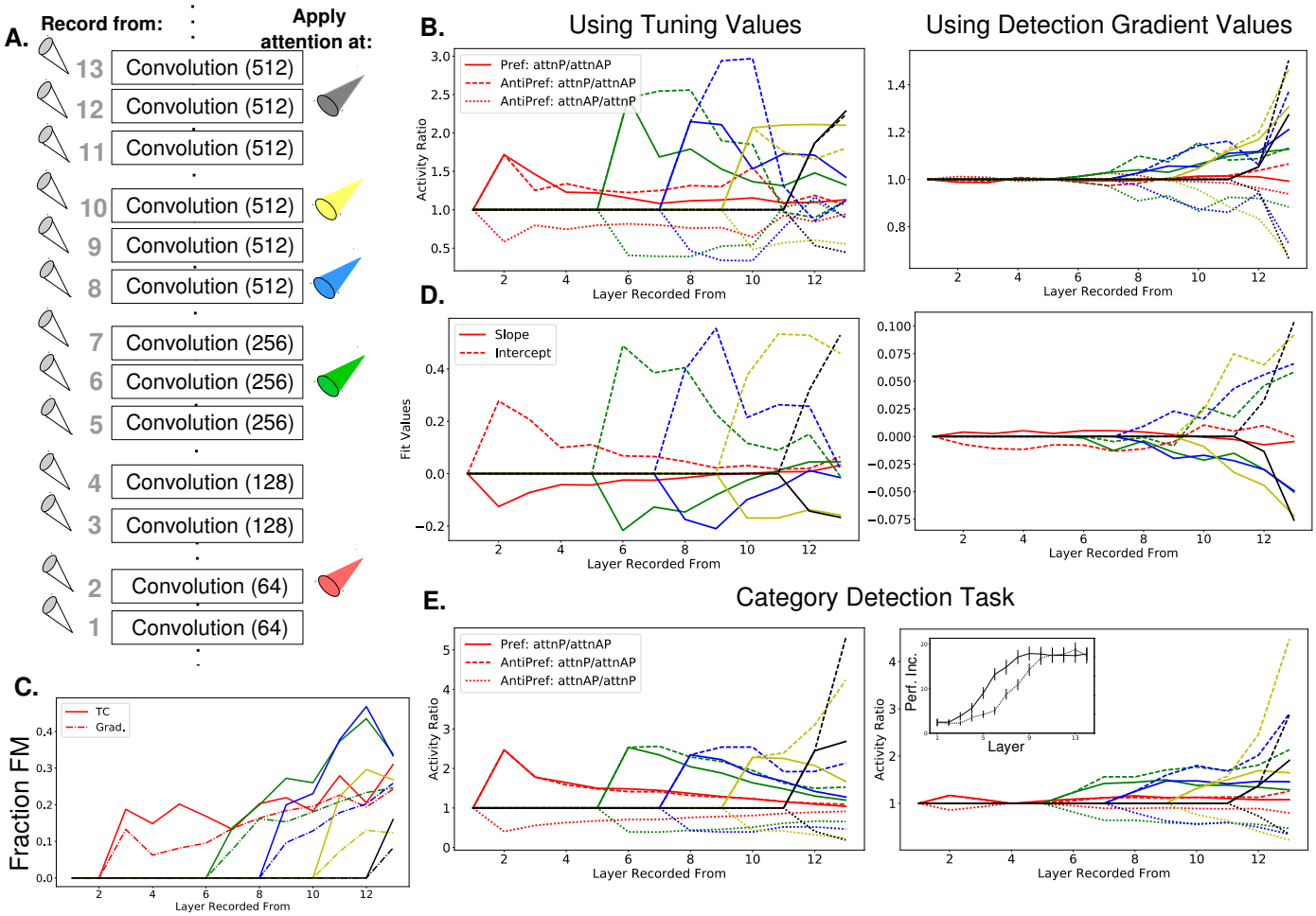

Figure 8: How Activity Changes from Attention Propagate for Unimodal Tasks. A.) Recording setup. The spatially averaged activity of feature maps at each layer was recorded (left) while attention was applied at layers 2, 6, 8, 10, and 12 individually. Activity was in response to a full field orientated grating for (B), (C), and (D) or full field standard ImageNet images for (E). Attention was always multiplicative and bidirectional. B.) Activity ratios for different attention conditions as a function of recorded layer when attention is applied at different layers (given by color as in (A)). Line style indicates whether the stimulus presented is preferred (solid line) or anti-preferred (dashed and dotted lines), and whether the ratio is calculated as activity when the preferred is attended divided by when the anti-preferred is attended (solid and dashed) or the reverse (dotted). Values are medians over all feature maps. Orientation tuning values (left) or orientation detection gradient values (right) are used for applying attention. C.) The fraction of feature maps that display feature matching (FM) behavior, defined as activity ratios greater than one for Pref:AttnP/AttnAP and AntiPref:AttnAP/AttnP) when attention is applied according to orientation tuning curve values (solid) or detection gradient values (dashed). D.) Dividing activity when a given orientation is present and attended by activity when no attention is applied gives a set of activity ratios. Ordering these ratios from most to least preferred orientation and fitting a line to them gives the slope and intercept values plotted here (intercept values are plotted in terms of how they differ from 1, so positive values are an intercept greater than 1). Values are medians across all feature maps at each layer with attention applied at layers indicated in (A). E.) Same as in (B) but using object category images, tuning values, and detection gradient values. The inset on the right shows mean performance detection over all 20 categories when attention is applied at diffferent layers using category detection gradient values (solid line, performance using tuning values shown as dotted line for comparison. Errorbars S.E.M.)

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

was also applied either to the preferred or anti-preferred orientation. According to the FSGM, the ratio of activity when the preferred orientation is attended divided by activity when the anti-preferred orientation is attended should be larger than one regardless of whether the orientation of the stimulus is preferred or not (indeed, the ratio should be constant for any stimulus). An alternative model, the feature matching (FM) model, suggests that the effect of attention is to amplify the activity of a neuron whenever the stimulus in its receptive field matches the attended stimulus. In this case, the ratio of activity when the preferred stimulus is attended over when the anti-preferred is attended would only be greater than one when the stimulus is the preferred orientation. If the stimulus is the anti-preferred orientation, the inverse of the that ratio would be greater than one.

In Figure 8B, we plot the median value of these ratios across all feature maps at a layer when attention is applied at different layers, indicated by color. When attention is applied directly at a layer according to its tuning values (left), FSGM effects are seen by default. As these activity changes propagate through the network, however, the FSGM effects wear off. Thus, when attention is applied at an early layer, it does not create strong changes in the final convolutional layer and thus cannot strongly impact the classifier. This explains the finding (Figure 6D) that attention works best for the detection task when applied at later layers, as the only way for strong FSGM effects to exist at the final layers is to apply attention near the final layers.

The notion that strong FSGM-like effects at the final layer are desirable for increasing classification performance is further supported by findings using the gradient values. In Figure 8B(right), we show the same analysis, but while applying attention according to orientation detection gradient values rather than tuning values. The effects at the layer at which attention is applied do not look strongly like FSGM, however FSGM properties evolve as the activity changes propagate through the network, leading to clear FSGM-like effects at the final layer.

These results are recapitulated in Figure 8D using a broader analysis also from [55]. Here, the activity of a feature map is calculated when attention is applied to the orientation in the stimulus and divided by the activity in response to the same orientation when no attention is applied. These ratios are organized according to orientation preference (most to least) and a line is fit to them. According to the FSGM of attention, this ratio should be greater than one for more preferred orientations and less than one for less preferred, creating a line with an intercept greater than one and negative slope. As expected, applying attention according to tuning values causes similar changes at the layer at which it is applied in this model (intercept values are plotted in terms of how they differ from one. Comparable average values from [55] are intercept: .06 and slope: 0.0166). Again, these effects wear off as the activity changes propagate through the network. Also gradient values ultimately lead to this kind of change at the final layer (right panel).

While Figure 8B and D show FSGM-like effects according to median values across all feature maps, some individual feature maps may show different behavior. In Figure 8C, we calculate the fraction of feature maps at a given layer that show feature matching behavior (defined as having activity ratios greater than one when the stimulus orientation matches the attended orientation for both preferred and anti-preferred orientations). As early as one layer post-attention feature maps start showing feature matching behavior, and the fraction grows as activity changes propagate. Interestingly, applying attention according to gradient values also causes an increase in the

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

fraction of feature maps with FM behavior, even as the median values become more FSGM-like. The attention literature contains conflicting findings regarding the feature similarity gain model versus the feature matching model [66, 78]. This may result from the fact that FSGM effects can turn into FM effects as they propagate through the network. In particular, this mechanism can explain the observations that feature matching behavior is observed more in FEF than V4 [104] and that match information is more easily readout from perirhinal cortex than IT [68].

We explore the propagation of these effects for category-based attention as well. In Figure 8E, we perform the same analysis as 8B, but with attention applied according to object category tuning values and stimuli that are full-field standard ImageNet images. We also calculate gradient values that would increase performance on category detection tasks (the same procedure used to calculate orientation detection gradients). The binary classification performance increase that results from applying attention according to these values is shown in Figure 8E(right, inset, solid line) in comparison to that when applying according to tuning values (dashed line). Like with orientation detection gradient values, applying attention according to these values propagates through the network to result in FSGM-like effects at the final layer. Also as with the orientation findings, the size of the FSGM effects that reach the final layer track with how well applying attention increases performance; for example, applying attention at layer 2 (red lines) does not lead to strong FSGM effects at the final layer and does not strongly increase performance.

3.9. Attention Alters Encoding Properties in Cross-Modal Tasks

The above recordings looked at how encoding of the attended dimension changed with attention. In cross-modal tasks, such as the spatial attention task and color classification task, the encoding that is relevant for performance is the that of the read-out dimension. We therefore measured how that encoding changes with attention at different layers as well.

For the spatial attention task, we measured category encoding by fitting a line to a set of activity ratios (see Figure 9A, left). Those activity ratios represent the activity of a quadrant when a given object category was in it and the quadrant was attended divided by activity when the same category was in the quadrant and no attention was applied. Arranging these from most to least preferred category for each feature map and fitting a line to them gives two values per feature map: the intercept (the ratio for the most preferred category, measured in terms of its difference from one) and the difference (the ratio for the most preferred minus the ratio for the least preferred, akin to the slope). A purely multiplicative effect leads to a positive intercept value and zero difference. This effect is clearly observed at the layers at which attention is applied in Figure 9A(right). It also continues with only a small amount of decay as the activity changes propagate through the network. By the final layer, the median intercept is still positive. The median difference becomes negative, indicating that preferred categories are enhanced more than non-preferred. The values at the final layer are fairly similar regardless of the layer at which attention was applied. This is in line with the fact that performance with multiplicative spatial attention is only moderately affected by the layer at which is attention is applied (Figure 5B).

We also looked at how color encoding changes when attention is applied to orientation. Here, we use 2-grating stimuli like those in Figure 6B to ask if encoding of the color of the grating with a given orientation increases when attention is applied

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

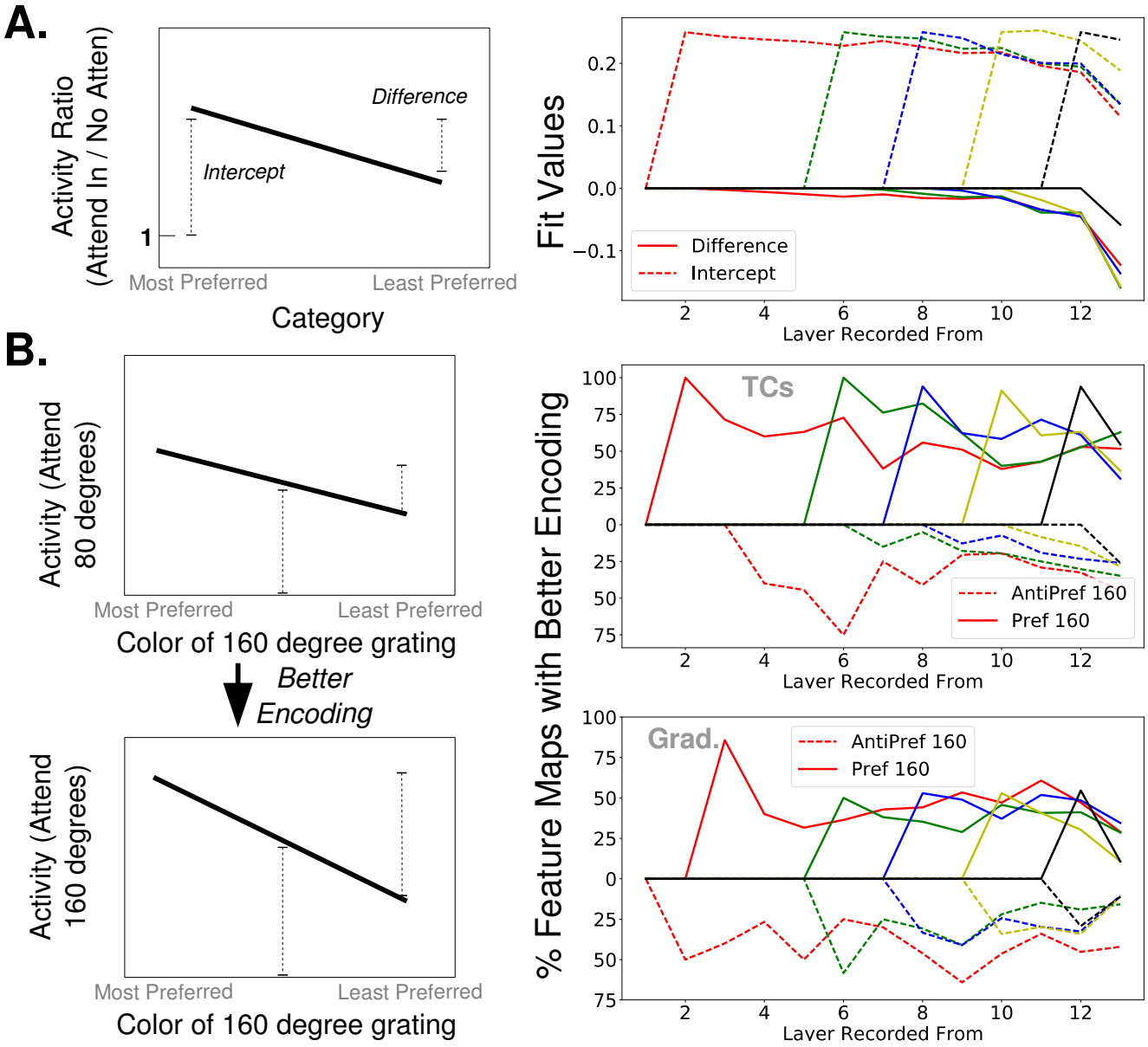

Figure 9: How Activity Changes from Attention Propagate for Cross-modal Tasks. A.) For each feature map, activity averaged over the attended quadrant when attention is applied to it is divided by activity when attention is not applied. Arranging these activity ratios from when the most to least preferred category is present in the quadrant and fitting a line to them results in the intercept and difference values as diagrammed on the left. Specifically, the intercept is the ratio for the most preferred category minus 1 and the difference is the ratio for the most preferred category minus the ratio for the least preferred. On the right, the median fit values across all feature maps are shown for each layer when attention is applied at layers indicated in 8A. B.) Orientated grating stimuli like those in 6B were designed with one grating at 140 degrees and the other at 60. Encoding of the color of the 140 degree grating is measured by fitting a line to the activity (spatially averaged over entire feature map) evoked by when each color is presented in the 140 degree grating (averaged over all colors presented in the 60 degree grating), ordered from most to least preferred. If the intercept (at the middle of this line) and difference increase when attention is applied to 140 degrees compared to attention at 60 degrees, the feature map has better encoding. On the right, the percent of feature maps with better encoding, segregated according to those that prefer 140 degrees (solid line) and those that anti-prefer (least prefer) 140 degrees (dashed lines, presented on a mirrored y-axis for visibility). Attention applied according to orientation tuning values (top) or color classification gradients (bottom).

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

to that orientation versus when it is applied to the orientation of the other grating (160 and 80 degree gratings were used). Arranging activity levels from most to least preferred color, we consider the encoding better if both the overall activity level is higher and the slope is more negative (see Figure 9B, left). We then measure the percent of feature maps that have better encoding of 160 degrees when attending 160 degrees versus attending 80 degrees. Looking at those feature maps that most prefer 160 degrees (sold lines, Figure 9B, right), nearly all feature maps enhance their color encoding at the layer at which attention was applied. However this percent decreases as the activity changes propagate through the network. On the other hand, for feature maps that anti-(or least) prefer 160 degrees, none have better encoding at the layer at which attention was applied, but the percent increases as activity changes propagate through the layers. Essentially, the burden of better encoding becomes evenly spread across feature maps regardless of preferred orientation.

This helps understand why, when applying attention according to tuning values, color classification performance is high at early layers, falls off at mid layers, and then recovers at final layers (Figure 6E, bottom). This is due to the different effects attention at these layers have on the final layer. When attention is applied at early layers, fewer final layer feature maps that prefer the attended orientation have better encoding, but many that don't prefer it do. When applied at late layers, a high percent of final layer feature maps that prefer the attended orientation have better encoding, even if those that don't prefer it do not. When attention is applied at middle layers, the effect on final layer feature maps that prefer the orientation has decayed, but the effect on those that don't prefer it hasn't increased much yet. Therefore performance is worse.

The idea that both feature maps that prefer and anti-prefer the attended orientation should enhance their color encoding is borne out by the gradient results. When attention is applied according to gradient values (Figure 9B, bottom), the percent of feature maps with better encoding is roughly equal for both those that prefer and anti-prefer the attended orientation. Experimentally, MT neurons have been found to better encode the direction of motion of a stimulus of the attended color as compared to a simultaneously presented stimulus of a different color [96]. Importantly, this effect of attention was *not* stronger when the preferred color was attended (indeed, there was a slight negative correlation between color preference and attention effect strength). This is not predicted by the FSGM directly, but as our model indicates, could result from FSGM-like effects at earlier areas, such as V1.

3.10. Applying Feature-based Attention at Multiple Layers Counteracts Effects

It is conceivable that feature-based attention applied at a lower layer could be as (or more) effective in modulating the activity of feature maps at a later layer as applying attention at that layer directly. In particular, for a given filter at layer l that prefers the attended category, bidirectional attention applied at layer $l-1$ could decrease the activity of units that have negative weights to the filter and increase the activity of units that have positive weights to the filter (note that in a more biologically-realistic model, the negatively weighted components would come indirectly from di-synaptic feedforward inhibition or surround interactions, as feedforward connections are largely excitatory). For example, if for a given unit in response to a given image the sum of its positively-weighted inputs is a, and the sum of its negatively-weighted inputs is b, without any attention, net input is $a-b$. If attention at $l-1$ scales positively-

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

weighted inputs up by 20% and negatively-weighted inputs down by 20%, the total input is now $1.2a - .8b$. These would lead to a greater net activity level than attention at l itself, which would just scale the net input by 1.2: $1.2(a - b)$. Therefore, given the same strength, applying attention at layer $l - 1$ could be a more effective way to modulate activity than applying it at layer l directly. However this assumes a very close alignment between the preferences of the feature maps at $l - 1$ and the weighting of the inputs into l.

We investigate this alignment by applying attention to object categories at various layers and recording at others (stimuli are standard ImageNet images of the attended category). The ratio of activity when attention is applied at a lower layer is divided by that when no attention is applied. Feature maps are then divided according to whether they prefer the attended category (have a tuning value greater than zero) or don't prefer it (tuning value less than zero). The strength value used is $\beta = .5$, therefore if attention at lower layers is more effective, we should see activity ratios greater than 1.5 for feature maps that prefer the attended category. The histograms in Figure 10A (left) show that the majority of feature maps that prefer the attended category (red) have ratios less than 1.5, regardless of the layer of attention or recording. In many cases, these feature maps even have ratios less than one, indicating that attention at a lower layer decreases the activity of feature maps that prefer the attended category. The misalignment between lower and later layers is starker the larger the distance between the attended and recorded layers. For example, when looking at layer 12, attention applied at layer 2 appears to increase and decrease feature map activity equally, without respect to category preference.

This helps to understand why feature-based attention applied at multiple layers simultaneously is not particularly effective at enhancing detection performance (Figure 3C). Specifically, if attention at a lower layer decreases the activity of feature maps that prefer the attended category at a later layer, it is actively counteracting the effects of attention applied at that layer. In Figure 10A, the effects of applying attention simultaneously at all layers is shown in black (using the same analysis of Figure 8B. The results from that figure are also replicated in paler colors for comparison). Attention is applied at each layer at one-tenth the strength ($\beta = .05$) as when it is applied to an individual layer. It is clear these effects are not accumulating effectively, as the activity ratios at the final layer (after passing through 13 layers of $\beta = .05$) are weaker than effects applied at layer 12 with $\beta = .5$.

Spatial attention, on the other hand, does lead to an effective accumulation of effects when applied at multiple layers. Figure 10B(left) uses the same analysis as Figure 9A, and shows the effect of applying spatial attention at all layers (with $\beta = .025$) in black. The effect on the intercept at the tenth layer is equal whether applying attention at all layers or only at layer 10 with $\beta = .25$. The difference parameter, however, is more negative when attention is applied at all layers than when attention is applied at layer 10. This demonstrates something that spatial attention can achieve at a given layer only when it is applied at a lower one: amplify preferred categories more than non-preferred. When all activity for all images is scaled multiplicatively at $l - 1$, some feature maps at layer l may see only a small increase when the image is of their non-preferred categories, due to the scaling up of their negatively-weighted inputs. In the cases where this effect is so strong that attention causes a decrease in activity in response to non-preferred category images (i.e., activity ratio less than one) while still causing an increase for preferred, attention would have the effect of

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

sharpening the tuning curve. Tuning curve sharpening as a result of spatial attention is generally not found experimentally [59, 90].

Activity ratios plotted in Figure 10B(right) are calculated as the activity recorded from a given quadrant when attention was applied to that quadrant over when no attention was applied. They are organized according to whether the feature map prefers or does not prefer the category present in the quadrant. By looking at different attended and recorded layers, we can see that spatial attention at lower layers can indeed lead to a higher scaling of feature maps that prefer the presented category, and that feature maps that do not prefer the presented category can have their activity decreased due to attention (especially when the gap between attended and recorded layers is larger).

4. Discussion

In this work, we utilized a deep convolutional neural network (CNN) as a model of the visual system to probe the relationship between neural activity and performance. Specifically, we provide a formal mathematical definition of the feature similarity gain model (FSGM) of attention, the basic tenets of which have been described in several experimental studies. This formalization allows us to investigate the FSGM's ability to enhance a CNN's performance on challenging visual tasks. Through this, we show that neural activity changes matching the type and magnitude of those observed experimentally can indeed lead to performance changes of the kind and magnitude observed experimentally. Furthermore, these results hold for a variety of tasks, from high level category detection to spatial tasks to color classification. The benefit of these particular activity changes for performance can be analyzed more formally in a signal detection or Bayesian framework [94, 21, 4, 67, 13], however such analysis is outside the scope of this work.

A finding from our model is that the layer at which attention is applied can have a large impact on performance. For detection tasks in particular, attention at early layers does little to enhance performance while attention at later layers such as 9-13 is most effective. According to [28], these layers correspond most to areas V4 and LO. Such areas are known and studied for reliably showing attentional effects, whereas earlier areas such as V1 are generally not [51]. In a study involving detection of objects in natural scenes, the strength of category-specific preparatory activity in object selective cortex was correlated with performance, whereas such preparatory activity in V1 was anti-correlated with performance [70]. This is in line with our finding that feature-based attention effects at earlier areas can counter the beneficial effects of that attention at later areas.

While CNNs have representations that are similar to the ventral stream, they lack many biological details including recurrent connections, dynamics, cell types, and noisy responses. Preliminary work has shown that these elements can be incorporated into a CNN structure, and attention can enhance performance in this more biologically-realistic architecture [48]. Furthermore, while the current work does not include neural noise independent of the stimulus, the images used do introduce variable responses. Take for example, the merged images, wherein a given image from one category is overlaid with an image from another. This can be thought of as highly structured noise added to the first image (rather than, for example, pixel-wise Gaussian noise). Such noise in the signal direction is known to be particularly challenging to overcome [1].

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

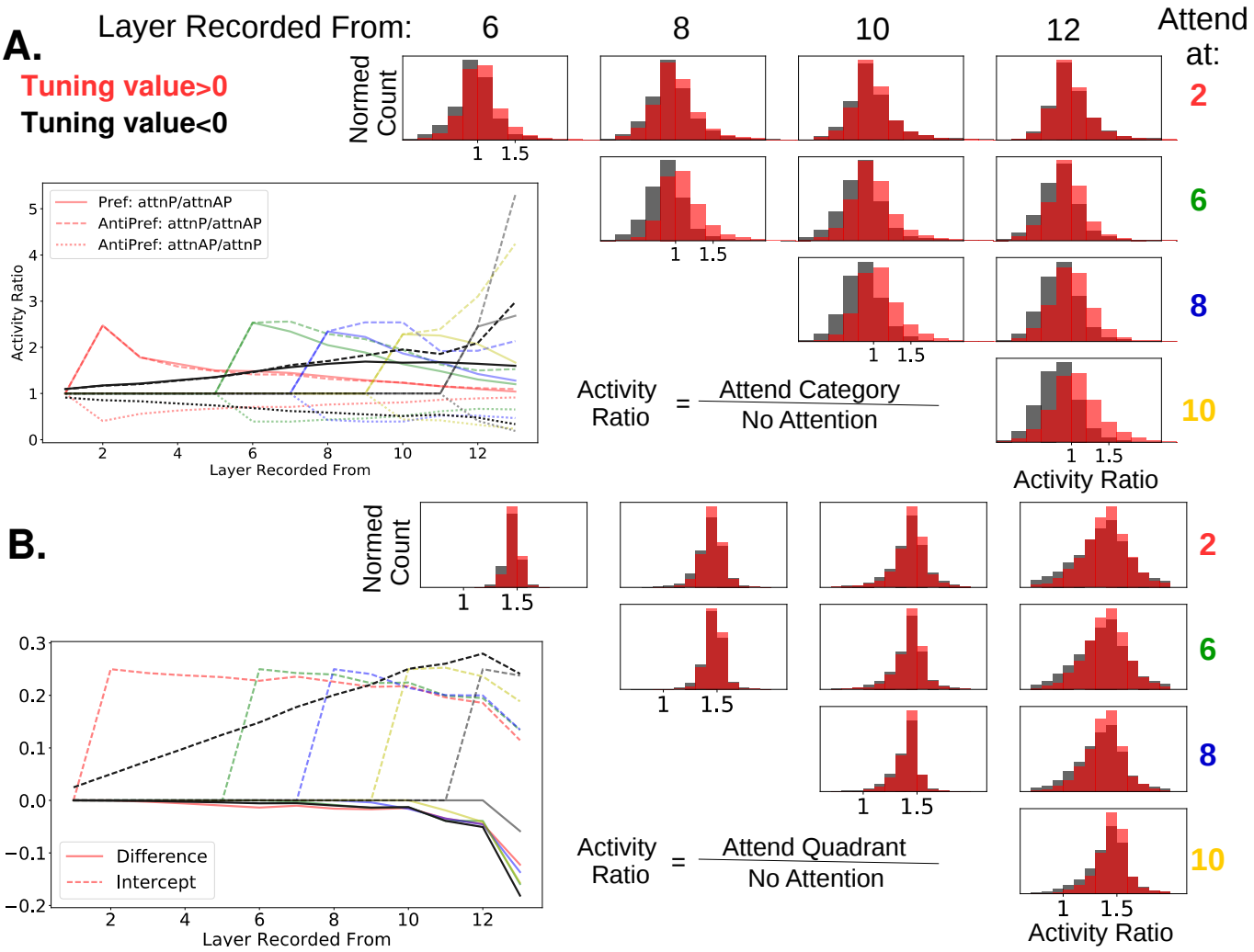

Figure 10: Differences When Applying Attention at All Layers for Feature and Spatial Attention. A.) Feature attention is not enhanced by being applied at multiple layers simultaneously. On the left, activity ratios as described in 8E are reproduced in lighter colors. Black lines show ratios when attention is applied at all layers ($\beta = .05$). On the right activity ratios are shown for when attention is applied at various layers individually and activity is recorded from later layers. In all cases, the category attended was the same as the one present in the input image. Histograms are of ratios of feature map activity when attention is applied to the category divided by activity when no attention is applied, dividing according to whether the feature map prefers (red) or does not prefer (black) the attended category. B.) Attention at multiple layers aides spatial attention. On the left, fit values for lines as described in 9A are shown in paler colors. Black lines are when attention is applied at all layers simultaneously ($\beta = .025$). On the right, histograms of activity ratios are given. Here the activity ratio is activity when attention is applied to the recorded quadrant over when no attention is applied. Feature maps are divided are according whether they prefer (red) or do not prefer (black) the category present in the quadrant.

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Another biological detail that this model lacks is "skip connections," when one layer feeds into both the layer directly above and layers above that. This is seen frequently in the brain, for example, in connections from V2 to V4 or V4 to parietal areas [93]. Our results show that the effects on attention at the final convolutional layer are important for performance changes, suggesting that synaptic distance from the classifier is a relevant feature—one that is less straight forward to determine in a network with skip connections. It may be the case though that thinking about visual areas in terms of their synaptic distance from decision-making areas such as prefrontal cortex [33] may be more useful for the study of attention than in terms of their distance from the retina. Finally, a major challenge for understanding the biological implementation of selective attention is determining how the attention signal is carried by feedback connections. Feature-based attention in particular appears to require targeted cell-by-cell modulation, which if implemented directly by top-down inputs, would require an unrealistic amount of fine tuning. A mechanism wherein feedback targeting is coarse, but the effects of it are refined by local processing is more plausible. It may be useful to take inspiration from the machine learning literature on attention and learning for hypotheses on how the brain does this [99, 46].

While they lack certain biological details, a benefit of using CNNs as a model is the ability to backpropagate error signals and understand causal relationships. Here we use this to calculate gradient values that estimate how attention should modulate activity, and compare these to the tuning values that the FSGM uses. The fact that these values are correlated and can lead to similar performance changes at task-specific layers (including similar changes in true and false positive rates, not shown) raises a question about the nature of biological attention: are neurons really targeted according to their tuning, or does the brain use something like gradient values? In [12] the correlation coefficient between an index of tuning and an index of attentional modulation was .52 for a population of V4 neurons, suggesting factors other than selectivity influence attention. Furthermore, many attention studies, including that one, use only preferred and non-preferred stimuli and therefore don't include a thorough investigation of the relationship between tuning and attentional modulation. [55] use multiple stimuli to provide support for the FSGM, however the interpretation is limited by the fact that they only report population averages. Furthermore, those population averages are closer to the average values in our model when attention is applied according to gradient values, rather than tuning values (Figure 8D). [78] investigated the relationship between tuning strength and the strength of attentional modulation on a cell-by-cell basis. While they did find a correlation (particularly for binocular disparity tuning), it wasn't very strong, which leaves room for the possibility that tuning is not the primary factor that determines attentional modulation.

Another finding from comparing gradient values with tuning values (and doing "recordings") is that tuning does not always predict how effectively one unit in the network will impact downstream units or the classifier. In particular, applying attention according to gradient values leads to changes that are hard to interpret when looked at through the lens of tuning, especially at earlier layers (Figure 8). However these changes eventually lead to large and impactful changes at later layers. Because experimenters can easily control the image, defining a cell's function in terms of how it responds to stimuli makes practical sense. A recent study looking at the relationship between tuning and choice probabilities suggests that tuning is not always an indication of a causal role in classification [101]. Studies that activate specific neurons in

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

one area and measure changes in another area or in behavioral output will likely be of significant value for determining function. Thus far, coarse stimulation protocols have found a relationship between the tuning of neural populations and their impact on perception [61, 18, 80]. Ultimately though, targeted stimulation protocols and a more fine-grained understanding of inter-area connections will be needed.

In this study, we used a diversity of attention tasks to see if the same mechanism could enhance performance universally. While we do find support for the feature similarity gain model's broad applicability, it is likely the case that the effects of attention in the brain are influenced substantially by the specifics of the task. Naturally, unimodal detection tasks have different challenges than cross-modal readout tasks (such as detecting a motion change in dots of a certain color). Generally, studies probing the neural mechanisms of attention care largely about the stimulus that is being attended, and less so about the information the animal needs from that stimulus to do the task. The task, then, is merely a way to get the subject to attend. However, as we see in our results, the best attention strategy is dependent on the task. Performance on our category detection task is only somewhat influenced by the choice of activity modulation (additive vs. multiplicative, etc), however, performance on the category classification task depends strongly on the use of multiplicative spatial attention. This task-dependency is even more stark in the orientation tasks, where the pattern of performance for attention at different layers is different for the detection and color classification tasks, even though the attention applied is identical. The effects of attention on firing rates, noise, and correlations may be more similar across studies if more similar tasks were used.

5. Acknowledgements

We are very grateful to the authors who so readily shared details of their behavioral data upon request: J. Patrick Mayo, Gary Lupyan, and Mika Koivisto. We further thank J. Patrick Mayo for helpful comments on the manuscript. GWL was supported by a Google PhD Fellowship and NIH (T32 NS064929). The authors declare no competing financial interests.

6. References

[1] Bruno B Averbeck, Peter E Latham, and Alexandre Pouget. Neural correlations, population coding and computation. *Nature reviews. Neuroscience*, 7(5):358, 2006.

[2] Jalal K Baruni, Brian Lau, and C Daniel Salzman. Reward expectation differentially modulates attentional behavior and activity in visual area v4. *Nature neuroscience*, 18(11):1656, 2015.

[3] Narcisse P Bichot, Matthew T Heard, Ellen M DeGennaro, and Robert Desimone. A source for feature-based attention in the prefrontal cortex. *Neuron*, 88 (4):832–844, 2015.

[4] Ali Borji and Laurent Itti. Optimal attentional modulation of a neural population. *Frontiers in computational neuroscience*, 8, 2014.

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[5] Geoffrey M Boynton. A framework for describing the effects of attention on visual responses. *Vision research*, 49(10):1129–1143, 2009.

[6] David A Bridwell and Ramesh Srinivasan. Distinct attention networks for feature enhancement and suppression in vision. *Psychological science*, 23(10):1151–1158, 2012.

[7] Elizabeth A Buffalo, Pascal Fries, Rogier Landman, Hualou Liang, and Robert Desimone. A backward progression of attentional effects in the ventral stream. *Proceedings of the National Academy of Sciences*, 107(1):361–365, 2010.

[8] Claus Bundesen. A theory of visual attention. *Psychological review*, 97(4):523, 1990.

[9] Santiago A Cadena, George H Denfield, Edgar Y Walker, Leon A Gatys, Andreas S Tolias, Matthias Bethge, and Alexander S Ecker. Deep convolutional models improve predictions of macaque v1 responses to natural images. *bioRxiv*, page 201764, 2017.

[10] Marisa Carrasco. Visual attention: The past 25 years. *Vision research*, 51(13): 1484–1525, 2011.

[11] Kyle R Cave. The featuregate model of visual selection. *Psychological research*, 62(2):182–194, 1999.

[12] Leonardo Chelazzi, John Duncan, Earl K Miller, and Robert Desimone. Responses of neurons in inferior temporal cortex during memory-guided visual search. *Journal of neurophysiology*, 80(6):2918–2940, 1998.

[13] Sharat Chikkerur, Thomas Serre, Cheston Tan, and Tomaso Poggio. What and where: A bayesian inference theory of attention. *Vision research*, 50(22):2233–2247, 2010.

[14] Marlene R Cohen and John HR Maunsell. Attention improves performance primarily by reducing interneuronal correlations. *Nature neuroscience*, 12(12): 1594–1600, 2009.

[15] Marlene R Cohen and John HR Maunsell. Using neuronal populations to study the mechanisms underlying spatial and feature attention. *Neuron*, 70(6):1192–1204, 2011.

[16] Tolga Çukur, Shinji Nishimoto, Alexander G Huth, and Jack L Gallant. Attention during natural vision warps semantic representation across the human brain. *Nature neuroscience*, 16(6):763–770, 2013.

[17] Mohammad Reza Daliri, Vladislav Kozyrev, and Stefan Treue. Attention enhances stimulus representations in macaque visual cortex without affecting their signal-to-noise level. *Scientific reports*, 6, 2016.

[18] Gregory C DeAngelis, Bruce G Cumming, and William T Newsome. Cortical area mt and the perception of stereoscopic depth. *Nature*, 394(6694):677, 1998.

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[19] Rachel N Denison, William T Adler, Marisa Carrasco, and Wei Ji Ma. Humans flexibly incorporate attention-dependent uncertainty into perceptual decisions and confidence. *bioRxiv*, page 175075, 2017.

[20] Cathryn J Downing. Expectancy and visual-spatial attention: effects on perceptual quality. *Journal of Experimental Psychology: Human perception and performance*, 14(2):188, 1988.

[21] Miguel P Eckstein, Matthew F Peterson, Binh T Pham, and Jason A Droll. Statistical decision theory to relate neurons to behavior in the study of covert visual attention. *Vision research*, 49(10):1097–1128, 2009.

[22] Michael Eickenberg, Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152:184–194, 2017.

[23] Pascal Fries, John H Reynolds, Alan E Rorie, and Robert Desimone. Modulation of oscillatory neuronal synchronization by selective visual attention. *Science*, 291 (5508):1560–1563, 2001.

[24] Davi Frossard. *VGG in TensorFlow*. Accessed: 2017-03-01.

[25] Kunihiko Fukushima. Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural networks*, 1(2):119–130, 1988.

[26] Robert Geirhos, David HJ Janssen, Heiko H Schütt, Jonas Rauber, Matthias Bethge, and Felix A Wichmann. Comparing deep neural networks against humans: object recognition when the signal gets weaker. *arXiv preprint arXiv:1706.06969*, 2017.

[27] Ivan C Griffin and Anna C Nobre. Orienting attention to locations in internal representations. *Journal of cognitive neuroscience*, 15(8):1176–1194, 2003.

[28] Umut Güçlü and Marcel AJ van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

[29] FH Hamker. The role of feedback connections in task-driven visual search. In *Connectionist models in cognitive neuroscience*, pages 252–261. Springer, 1999.

[30] Fred H Hamker and James Worcester. Object detection in natural scenes by feedback. In *International Workshop on Biologically Motivated Computer Vision*, pages 398–407. Springer, 2002.

[31] Harold L Hawkins, Steven A Hillyard, Steven J Luck, Mustapha Mouloua, Cathryn J Downing, and Donald P Woodward. Visual attention modulates signal detectability. *Journal of Experimental Psychology: Human Perception and Performance*, 16(4):802, 1990.

[32] Benjamin Y Hayden and Jack L Gallant. Combined effects of spatial and feature-based attention on responses of v4 neurons. *Vision research*, 49(10):1182–1187, 2009.

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[33] Hauke R Heekeren, Sean Marrett, Peter A Bandettini, and Leslie G Ungerleider. A general mechanism for perceptual decision-making in the human brain. *Nature*, 431(7010):859–862, 2004.

[34] Daniel Kaiser, Nikolaas N Oosterhof, and Marius V Peelen. The neural dynamics of attentional selection in natural scenes. *Journal of neuroscience*, 36(41):10522–10528, 2016.

[35] Kohitij Kar, Jonas Kubilius, Elias Issa, Kailyn Schmidt, and James DiCarlo. Evidence that feedback is required for object identity inferences computed by the ventral stream. COSYNE, 2017.

[36] Sabine Kastner and Mark A Pinsk. Visual attention as a multilevel selection process. *Cognitive, Affective, & Behavioral Neuroscience*, 4(4):483–500, 2004.

[37] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915, 2014.

[38] Seyed-Mahdi Khaligh-Razavi, Linda Henriksson, Kendrick Kay, and Nikolaus Kriegeskorte. Fixed versus mixed rsa: Explaining visual representations by fixed and mixed feature sets from shallow and deep computational models. *Journal of Mathematical Psychology*, 76:184–197, 2017.

[39] Saeed Reza Kheradpisheh, Masoud Ghodrati, Mohammad Ganjtabesh, and Timothée Masquelier. Deep networks can resemble human feed-forward vision in invariant object recognition. *Scientific reports*, 6:32672, 2016.

[40] Mika Koivisto and Ella Kahila. Top-down preparation modulates visual categorization but not subjective awareness of objects presented in natural backgrounds. *Vision Research*, 133:73–80, 2017.

[41] Simon Kornblith and Doris Y Tsao. How thoughts arise from sights: inferotemporal and prefrontal contributions to vision. *Current Opinion in Neurobiology*, 46:208–218, 2017.

[42] Richard J Krauzlis, Lee P Lovejoy, and Alexandre Zénon. Superior colliculus and visual spatial attention. *Annual review of neuroscience*, 36:165–182, 2013.

[43] Jonas Kubilius, Stefania Bracci, and Hans P Op de Beeck. Deep neural networks as a computational model for human shape sensitivity. *PLoS computational biology*, 12(4):e1004896, 2016.

[44] Brenden M Lake, Wojciech Zaremba, Rob Fergus, and Todd M Gureckis. Deep neural networks predict category typicality ratings for images. In *CogSci*, 2015.

[45] Joonyeol Lee and John HR Maunsell. Attentional modulation of mt neurons with single or multiple stimuli in their receptive fields. *Journal of Neuroscience*, 30(8):3058–3066, 2010.

[46] Timothy P Lillicrap, Daniel Cownden, Douglas B Tweed, and Colin J Akerman. Random synaptic feedback weights support error backpropagation for deep learning. *Nature communications*, 7, 2016.

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[47] Grace W Lindsay. Feature-based attention in convolutional neural networks. *arXiv preprint arXiv:1511.06408*, 2015.

[48] Grace W Lindsay, Dan B Rubin, and Kenneth D Miller. The stabilized supralinear network replicates neural and performance correlates of attention. COSYNE, 2017.

[49] Drew Linsley, Sven Eberhardt, Tarun Sharma, Pankaj Gupta, and Thomas Serre. What are the visual features underlying human versus machine vision? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2706–2714, 2017.

[50] Bradley C Love, Olivia Guest, Piotr Slomka, Victor M Navarro, and Edward Wasserman. Deep networks as models of human and animal categorization. In *CogSci*, 2017.

[51] Steven J Luck, Leonardo Chelazzi, Steven A Hillyard, and Robert Desimone. Neural mechanisms of spatial selective attention in areas v1, v2, and v4 of macaque visual cortex. *Journal of neurophysiology*, 77(1):24–42, 1997.

[52] Thomas Zhihao Luo and John HR Maunsell. Neuronal modulations in visual cortex are associated with only one of multiple components of attention. *Neuron*, 86(5):1182–1188, 2015.

[53] Gary Lupyan and Michael J Spivey. Making the invisible visible: Verbal but not visual cues enhance visual detection. *PLoS One*, 5(7):e11452, 2010.

[54] Gary Lupyan and Emily J Ward. Language can boost otherwise unseen objects into visual awareness. *Proceedings of the National Academy of Sciences*, 110(35): 14196–14201, 2013.

[55] Julio C Martinez-Trujillo and Stefan Treue. Feature-based attention increases the selectivity of population responses in primate visual cortex. *Current Biology*, 14(9):744–751, 2004.

[56] John HR Maunsell and Erik P Cook. The role of attention in visual processing. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 357(1424):1063–1072, 2002.

[57] J Patrick Mayo and John HR Maunsell. Graded neuronal modulations related to visual spatial attention. *Journal of Neuroscience*, 36(19):5353–5361, 2016.

[58] J Patrick Mayo, Marlene R Cohen, and John HR Maunsell. A refined neuronal population measure of visual attention. *PloS one*, 10(8):e0136570, 2015.

[59] Carrie J McAdams and John HR Maunsell. Effects of attention on orientation-tuning functions of single neurons in macaque cortical area v4. *Journal of Neuroscience*, 19(1):431–441, 1999.

[60] Jude F Mitchell, Kristy A Sundberg, and John H Reynolds. Differential attention-dependent response modulation across cell classes in macaque visual area v4. *Neuron*, 55(1):131–141, 2007.

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[61] Sebastian Moeller, Trinity Crapse, Le Chang, and Doris Y Tsao. The effect of face patch microstimulation on perception of faces and objects. *Nature Neuroscience*, 20(5):743–752, 2017.

[62] Ilya E Monosov, David L Sheinberg, and Kirk G Thompson. The effects of prefrontal cortex inactivation on object responses of single neurons in the inferotemporal cortex during visual search. *Journal of Neuroscience*, 31(44):15956–15961, 2011.

[63] Barbara Montagna, Franco Pestilli, and Marisa Carrasco. Attention trades off spatial acuity. *Vision research*, 49(7):735–745, 2009.

[64] Tirin Moore and Katherine M Armstrong. Selective gating of visual signals by microstimulation of frontal cortex. *Nature*, 421(6921):370, 2003.

[65] Sancho I Moro, Michiel Tolboom, Paul S Khayat, and Pieter R Roelfsema. Neuronal activity in the visual cortex reveals the temporal order of cognitive operations. *Journal of Neuroscience*, 30(48):16293–16303, 2010.

[66] Brad C Motter. Neural correlates of feature selective memory and pop-out in extrastriate area v4. *Journal of Neuroscience*, 14(4):2190–2199, 1994.

[67] Vidhya Navalpakkam and Laurent Itti. Search goal tunes visual features optimally. *Neuron*, 53(4):605–617, 2007.

[68] Marino Pagan, Luke S Urban, Margot P Wohl, and Nicole C Rust. Signals in inferotemporal and perirhinal cortex suggest an untangling of visual target information. *Nature neuroscience*, 16(8):1132–1139, 2013.

[69] William K Page and Charles J Duffy. Cortical neuronal responses to optic flow are shaped by visual strategies for steering. *Cerebral cortex*, 18(4):727–739, 2007.

[70] Marius V Peelen and Sabine Kastner. A neural basis for real-world visual search in human occipitotemporal cortex. *Proceedings of the National Academy of Sciences*, 108(29):12125–12130, 2011.

[71] Marius V Peelen, Li Fei-Fei, and Sabine Kastner. Neural mechanisms of rapid natural scene categorization in human visual cortex. *Nature*, 460(7251):94, 2009.

[72] Joshua C Peterson, Joshua T Abbott, and Thomas L Griffiths. Adapting deep network features to capture psychological representations. *arXiv preprint arXiv:1608.02164*, 2016.

[73] Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation*, 2017.

[74] Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11), 1999.

[75] Samuel Ritter, David GT Barrett, Adam Santoro, and Matt M Botvinick. Cognitive psychology for deep neural networks: A shape bias case study. *arXiv preprint arXiv:1706.08606*, 2017.

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[76] Mariel Roberts, Rachel Cymerman, R Theodore Smith, Lynne Kiorpes, and Marisa Carrasco. Covert spatial attention is functionally intact in amblyopic human adultsroberts et al. *Journal of vision*, 16(15):30–30, 2016.

[77] Edmund T Rolls and Gustavo Deco. Attention in natural scenes: neurophysiological and computational bases. *Neural networks*, 19(9):1383–1394, 2006.

[78] Douglas A Ruff and Richard T Born. Feature attention for binocular disparity in primate area mt depends on tuning strength. *Journal of neurophysiology*, 113 (5):1545–1555, 2015.

[79] Melissa Saenz, Giedrius T Buracas, and Geoffrey M Boynton. Global effects of feature-based attention in human visual cortex. *Nature neuroscience*, 5(7):631, 2002.

[80] C Daniel Salzman, Kenneth H Britten, and William T Newsome. Cortical microstimulation influences perceptual judgements of motion direction. *Nature*, 346(6280):174–177, 1990.

[81] K Seeliger, M Fritsche, U Güçlü, S Schoenmakers, J-M Schoffelen, SE Bosch, and MAJ van Gerven. Cnn-based encoding and decoding of visual object recognition in space and time. *bioRxiv*, page 118091, 2017.

[82] John T Serences, Jens Schwarzbach, Susan M Courtney, Xavier Golay, and Steven Yantis. Control of object-based attention in human cortex. *Cerebral Cortex*, 14(12):1346–1357, 2004.

[83] Thomas Serre, Lior Wolf, Stanley Bileschi, Maximilian Riesenhuber, and Tomaso Poggio. Robust object recognition with cortex-like mechanisms. *IEEE transactions on pattern analysis and machine intelligence*, 29(3):411–426, 2007.

[84] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[85] Hedva Spitzer, Robert Desimone, Jeffrey Moran, et al. Increased attention enhances both behavioral and neuronal performance. *Science*, 240(4850):338–340, 1988.

[86] Timo Stein and Marius V Peelen. Object detection in natural scenes: Independent effects of spatial and category-based attention. *Attention, Perception, & Psychophysics*, 79(3):738–752, 2017.

[87] Jan Theeuwes, Arthur F Kramer, and Paul Atchley. Attentional effects on preattentive vision: spatial precues affect the detection of simple features. *Journal of Experimental Psychology: Human Perception and Performance*, 25(2):341, 1999.

[88] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

[89] Stefan Treue. Neural correlates of attention in primate visual cortex. *Trends in neurosciences*, 24(5):295–300, 2001.

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[90] Stefan Treue and Julio C Martinez Trujillo. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, 399(6736):575, 1999.

[91] Bryan P Tripp. Similarities and differences between stimulus tuning in the inferotemporal visual cortex and convolutional networks. In *Neural Networks (IJCNN), 2017 International Joint Conference on*, pages 3551–3560. IEEE, 2017.

[92] John K Tsotsos, Scan M Culhane, Winky Yan Kei Wai, Yuzhong Lai, Neal Davis, and Fernando Nuflo. Modeling visual attention via selective tuning. *Artificial intelligence*, 78(1-2):507–545, 1995.

[93] Leslie G Ungerleider, Thelma W Galkin, Robert Desimone, and Ricardo Gattass. Cortical connections of area v4 in the macaque. *Cerebral Cortex*, 18(3):477–499, 2007.

[94] Preeti Verghese. Visual search and attention: A signal detection theory approach. *Neuron*, 31(4):523–535, 2001.

[95] Bram-Ernst Verhoef and John HR Maunsell. Attention-related changes in correlated neuronal activity arise from normalization mechanisms. *Nature Neuroscience*, 20(7):969–977, 2017.

[96] Aurel Wannig, Valia Rodríguez, and Winrich A Freiwald. Attention to surfaces modulates motion processing in extrastriate area mt. *Neuron*, 54(4):639–651, 2007.

[97] Louise Whiteley and Maneesh Sahani. Attention in a bayesian framework. *Frontiers in human neuroscience*, 6, 2012.

[98] Jeremy M Wolfe. Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review*, 1(2):202–238, 1994.

[99] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015.

[100] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

[101] Adam Zaidel, Gregory C DeAngelis, and Dora E Angelaki. Decoupled choice-driven and stimulus-related activity in parietal neurons may be misrepresented by choice probabilities. *Nature Communications*, 8, 2017.

[102] Weiwei Zhang and Steven J Luck. Feature-based attention modulates feedforward visual processing. *Nature neuroscience*, 12(1):24–25, 2009.

[103] Ying Zhang, Ethan M Meyers, Narcisse P Bichot, Thomas Serre, Tomaso A Poggio, and Robert Desimone. Object decoding with attention in inferior temporal cortex. *Proceedings of the National Academy of Sciences*, 108(21):8850–8855, 2011.

bioRxiv preprint doi: https://doi.org/10.1101/233338; this version posted December 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[104] Huihui Zhou and Robert Desimone. Feature-based attention in the frontal eye field and area v4 during visual search. *Neuron*, 70(6):1205–1217, 2011.