

How biological attention mechanisms improve task performance in a large-scale visual system model

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

How does attentional modulation of neural activity enhance performance? Here we use a deep convolutional neural network as a large-scale model of the visual system to address this question. We model the feature similarity gain model of attention, in which attentional modulation is applied according to neural stimulus tuning. Using a variety of visual tasks, we show that neural modulations of the kind and magnitude observed experimentally lead to performance changes of the kind and magnitude observed experimentally. We find that, at earlier layers, attention applied according to tuning does not successfully propagate through the network, and has a weaker impact on performance than attention applied according to values computed for optimally modulating higher areas. This raises the question of whether biological attention might be applied at least in part to optimize function rather than strictly according to tuning. We suggest a simple experiment to distinguish these alternatives.

1. Introduction

1 Covert visual attention—applied according to spatial location or visual features—
2 has been shown repeatedly to enhance performance on challenging visual tasks [10].
3 To explore the neural mechanisms behind this enhancement, neural responses to the
4 same visual input are compared under different task conditions. Such experiments have
5 identified numerous neural modulations associated with attention, including changes
6 in firing rates, noise levels, and correlated activity [83, 14, 22, 52]. But how do these
7 neural activity changes impact performance? Previous theoretical studies have offered
8 helpful insights on how attention may work to enhance performance [62, 71, 86, 11, 27,
9 90, 26, 20, 4, 89, 8, 82, 88, 13]. However, much of this work is either based on small,
10 hand-designed models or lacks direct mechanistic interpretability. Here, we utilize a
11 large-scale model of the ventral visual stream to explore the extent to which neural
12 changes like those observed experimentally can lead to performance enhancements on
13 realistic visual tasks. Specifically, we use a deep convolutional neural network trained
14 to perform object classification to test effects of the feature similarity gain model of
15 attention [84].

16 Deep convolutional neural networks (CNNs) are popular tools in the machine learn-
17 ing and computer vision communities for performing challenging visual tasks [69].
18 Their architecture—comprised of layers of convolutions, nonlinearities, and response
19 pooling—was designed to mimic the retinotopic and hierarchical nature of the mam-
20 malian visual system [69]. Models of a similar form have been used to study the

21 biological underpinnings of object recognition for decades [24, 70, 78]. Recently it has
22 been shown that when these networks are trained to successfully perform object classi-
23 fication on real-world images, the intermediate representations learned are remarkably
24 similar to those of the primate visual system, making CNNs state-of-the-art models of
25 the ventral stream [92, 37, 36, 38, 34, 9, 85, 46, 42]. A key finding has been the corre-
26 spondence between different areas in the ventral stream and layers in the deep CNNs,
27 with early convolutional layers best able to capture the representation of V1 and mid-
28 dle and higher layers best able to capture V4 and IT, respectively [25, 21, 76]. Given
29 that CNNs reach near-human performance on visual tasks and have architectural and
30 representational similarities to the visual system, they are particularly well-positioned
31 for exploring how neural correlates of attention impact behavior.

32 One popular framework to describe attention’s effects on firing rates is the feature
33 similarity gain model (FSGM). This model, introduced by Treue & Martinez-Trujillo,
34 claims that a neuron’s activity is multiplicatively scaled up (or down) according to
35 how much it prefers (or doesn’t prefer) the properties of the attended stimulus [84,
36 51]. Attention to a certain visual attribute, such as a specific orientation or color,
37 is generally referred to as feature-based attention (FBA). FBA effects are spatially
38 global: if a task performed at one location in the visual field activates attention to
39 a particular feature, neurons that represent that feature across the visual field will
40 be affected [94, 73]. Overall, this leads to a general shift in the representation of the
41 neural population towards that of the attended stimulus [17, 33, 65]. Spatial attention
42 implies that a particular portion of the visual field is being attended. According to the
43 FSGM, spatial location is treated as an attribute like any other. Therefore, a neuron’s
44 modulation due to attention can be predicted by how well its preferred features and
45 spatial receptive field align with the features and location of the attended stimulus.
46 The effects of combined feature and spatial attention have been found to be additive
47 [29].

48 A debated issue in the attention literature is where in the visual stream attention
49 effects can be seen. Many studies of attention focus on V4 and MT/MST [83], as
50 these areas have reliable attentional effects. Some studies do find effects at earlier
51 areas [60], though they tend to be weaker and occur later in the visual response [35].
52 Therefore, a leading hypothesis is that attention signals, coming from prefrontal areas
53 [58, 57, 3, 40], target later visual areas, and the feedback connections that those areas
54 send to earlier ones cause the weaker effects seen there later [7, 47].

55 In this study, we define the FSGM of attention mathematically and implement
56 it in a deep CNN. By applying attention at different layers in the network and for
57 different tasks, we see how neural changes at one area propagate through the network
58 and change performance.

59 2. Results

60 The network used in this study—VGG-16, [79]—is shown in Figure 1A and ex-
61 plained in Methods 4.1. Briefly, at each convolutional layer, the application of a given
62 convolutional filter results in a feature map, which is a 2-D grid of artificial neurons
63 that represent how well the bottom-up input at each location aligns with the filter.
64 Therefore a ”retinotopic” layout is built into the structure of the network, and the
65 same visual features are represented across that retinotopy (akin to how cells that pre-
66 fer a given orientation exist at all locations across the V1 retinotopy). This network

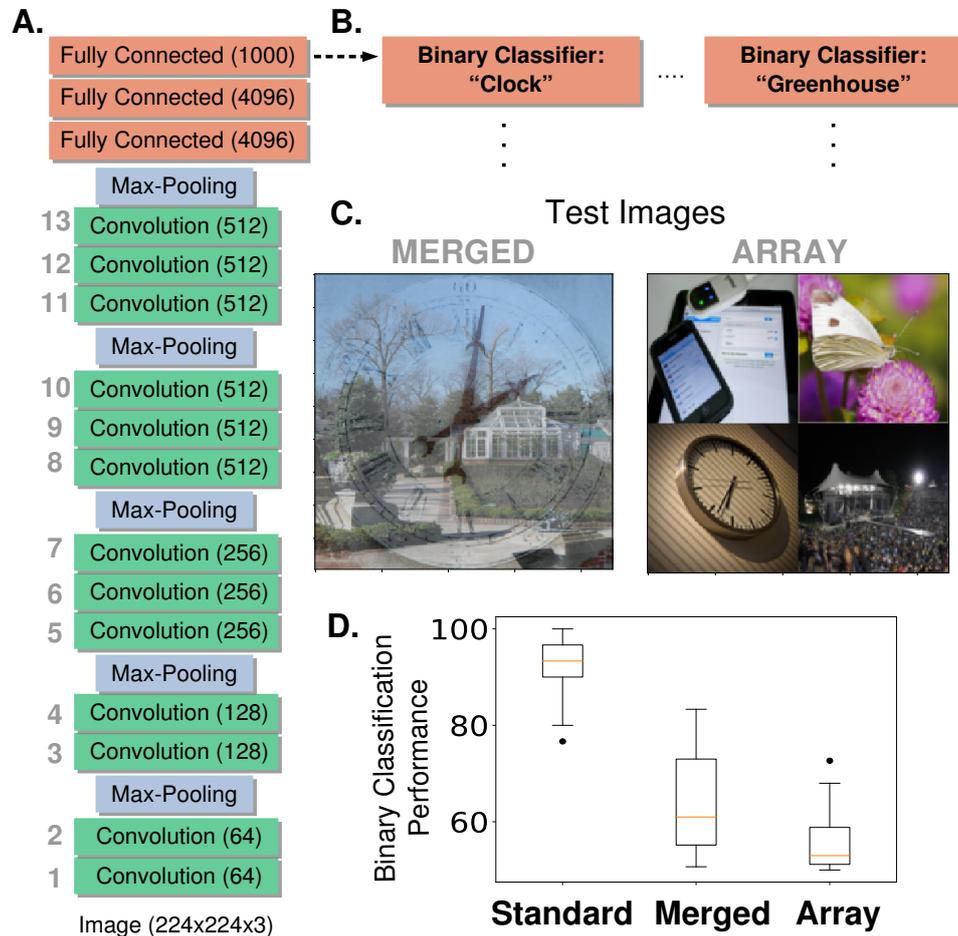


Figure 1: Network Architecture and Feature-Based Attention Task Setup. A.) The model used is a pre-trained deep neural network (VGG-16) that contains 13 convolutional layers (labeled in gray, number of feature maps given in parenthesis) and is trained on the ImageNet dataset to do 1000-way object classification. 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, 20 total). 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 the exception of outliers marked as dots). Standard images are regular ImageNet images not used in the binary classifier training set.

67 was explored in [25], where it was shown that early convolutional layers of this CNN
68 are best at predicting activity of voxels in V1, while late convolutional layers are best
69 at predicting activity of voxels in the object-selective lateral occipital area (LO).

70 *2.1. The Relationship between Tuning and Classification*

71 The feature similarity gain model of attention posits that neural activity is modu-
72 lated by attention in proportion to how strongly a neuron prefers the attended features,
73 as assessed by its tuning. However, the relationship between a neuron’s tuning and
74 its ability to influence downstream readouts remains a difficult one to investigate bio-
75 logically. We use our hierarchical model to explore this question. We do so by using
76 backpropagation to calculate “gradient values”, which we compare to tuning curves
77 (see Methods 4.3 and 4.5.1 for details). Gradient values indicate the ways in which fea-
78 ture map activities should change in order to make the network more likely to classify
79 an image as being of a certain object category. Tuning values represent the degree to
80 which the feature map responds preferentially to images of a given category. If there
81 is a correspondence between tuning and classification, a feature map that prefers a
82 given object category (that is, responds strongly to it) should also have a high positive
83 gradient value for that category. In Figure 2A we show gradient values and tuning
84 curves for three example feature maps. In Figure 2C, we show the average correlation
85 coefficients between tuning values and gradient values for all feature maps at each of
86 the 13 convolutional layers. As can be seen, tuning curves in all layers show higher
87 correlation with gradient values than expected by chance (as assayed by shuffled con-
88 trols), but this correlation is relatively low, increasing across layers from about .2 to .5.
89 Overall tuning quality also increases with layer depth (Figure 2B), but less strongly.

90 Even at the highest layers, there can be serious discrepancies between tuning and
91 gradient values. In Figure 2D, we show the gradient values of feature maps at the final
92 four convolutional layers, segregated according to tuning value. In red are gradient
93 values that correspond to tuning values greater than one (for example, category 12
94 for the feature map in the middle pane of Figure 2A). As these distributions show,
95 strong tuning values can be associated with weak or even negative gradient values.
96 Negative gradient values indicate that increasing the activity of that feature map
97 makes the network less likely to categorize the image as the given category. Therefore,
98 even feature maps that strongly prefer a category (and are only a few layers from the
99 classifier) still may not be involved in its classification, or even be inversely related to
100 it. This is aligned with a recent neural network ablation study that shows category
101 selectivity does not predict impact on classification [59].

102 *2.2. Feature-based Attention Improves Performance on Challenging Object Classifica-* 103 *tion Tasks*

104 To determine if manipulation according to tuning values can enhance performance,
105 we created challenging visual images composed of multiple objects for the network to
106 classify. These test images are of two types: merged (two object images transparently
107 overlaid, such as in [77]) or array (four object images arranged on a grid) (see Figure
108 1C examples). The task for the network is to detect the presence of a given object
109 category in these images. It does so using a series of binary classifiers trained on
110 standard images of these objects, which replace the last layer of the network (Figure
111 1B). The performance of these classifiers on the test images indicates that this is a
112 challenging task for the network (64.4% on merged images and 55.6% on array, Figure
113 1D. Chance is 50%), and thus a good opportunity to see the effects of attention.

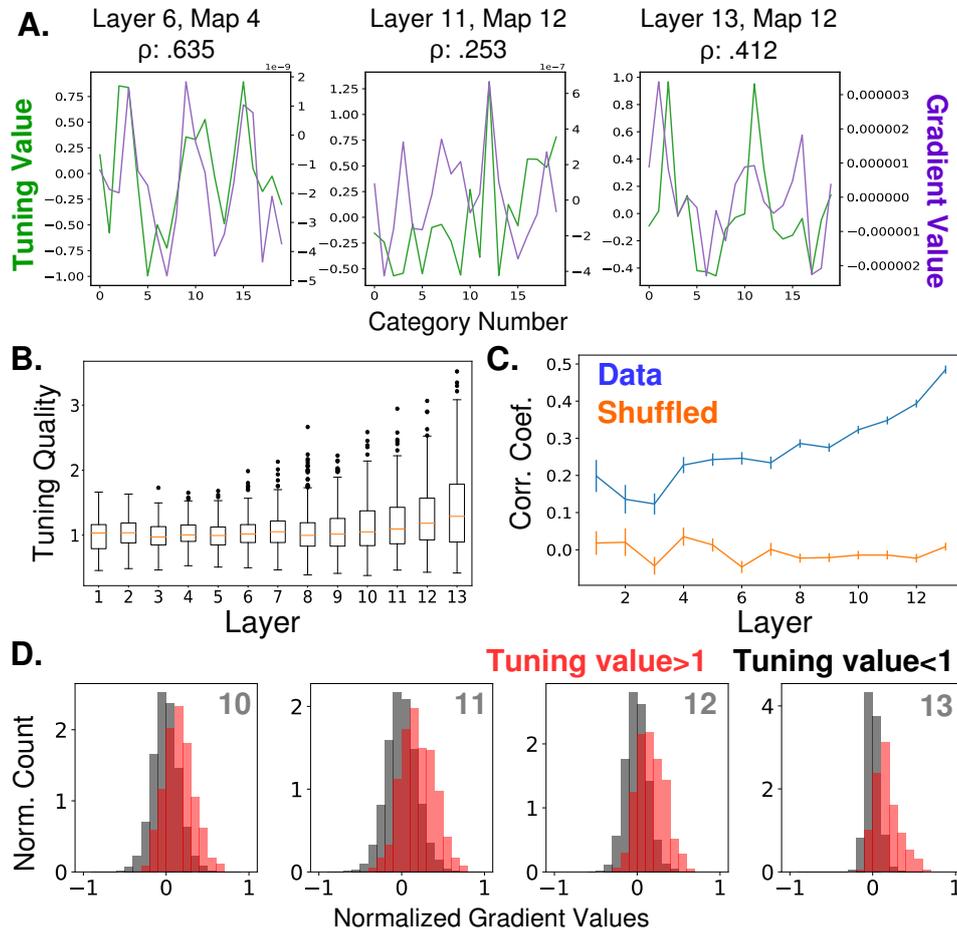


Figure 2: Relationship Between Feature Map Tuning and Gradient Values. A.) Example tuning values (green, left axis) and gradient values (purple, right axis) of three different feature maps from three different layers (identified in titles, layers as labeled in Figure 1A) over the 20 tested object categories. Tuning values indicate how the response to a category differs from the mean response; gradient values indicate how activity should change in order to classify input as from the category. Correlation coefficients between tuning curves and gradient values given in titles. B.) 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. Average tuning quality for shuffled data: $.372 \pm .097$ (this value does not vary significantly across layers) C.) Correlation coefficients between tuning curves and gradient value curves averaged over feature maps and plotted across layers (errorbars \pm S.E.M., data values in blue and shuffled controls in orange). D.) 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 layers 10, 11, 12, and 13. 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).

114 We implement feature-based attention in this network by modulating the activity
115 of units in each feature map according to how strongly the feature map prefers the
116 attended object category (see Methods 4.5.1 and 4.5). A schematic of this is shown
117 in Figure 3A. The slope of the activation function of units in a given feature map is
118 scaled according to the tuning value of that feature map for the attended category
119 (positive tuning values increase the slope while negative tuning values decrease it).
120 Thus the impact of attention on activity is multiplicative and bi-directional.

121 The effects of attention are measured when attention is applied in this way at
122 each layer individually, or all layers simultaneously (Figure 3B; solid lines). For both
123 image types (merged and array), attention enhances performance and there is a clear
124 increase in performance enhancement as attention is applied at later layers in the
125 network (numbering is as in Figure 1A). In particular, attention applied at the final
126 convolutional layer performs best, leading to an 18.8% percentage point increase in
127 binary classification on the merged images task and 22.8% increase on the array images
128 task. Thus, FSGM-like effects can have large beneficial impacts on performance.

129 Attention applied at all layers simultaneously does not lead to better performance
130 than attention applied at any individual layer. The reasons for this will be addressed
131 later.

132 Some components of the FSGM are debated, e.g. whether attention impacts re-
133 sponses multiplicatively or additively [5, 2, 47, 55], and whether the activity of cells
134 that do not prefer the attended stimulus is actually suppressed [6, 62]. Comparisons
135 of different variants of the FSGM can be seen in Supplementary Figure 8. In general,
136 multiplicative and bidirectional effects work best.

137 We also measure performance when attention is applied using gradient values rather
138 than tuning values (these gradient values are calculated to maximize performance
139 on the binary classification task, rather than classify the image as a given category;
140 therefore technically they differ from those shown in Figure 2, however in practice
141 they are strongly correlated. See Methods 4.3 and 4.5.2 for details). Attention applied
142 using gradient values shows the same layer-wise trend as when using tuning values.
143 It also reaches the same performance enhancement peak when attention is applied at
144 the final layers. The major difference, however, comes when attention is applied at
145 middle layers of the network. Here, attention applied according to gradient values
146 outperforms that of tuning values.

147 *2.3. Attention Strength and the Tradeoff between Increasing True and False Positives*

148 In the previous section, we examined the best possible effects of attention by choos-
149 ing the strength for each layer and category that optimized performance. Here, we
150 look at how performance changes as we vary the overall strength (β) of attention.

151 In Figure 4A we break the binary classification performance into true and false
152 positive rates. Here, each colored line indicates a different category and increasing dot
153 size represents increasing strength of attention. Ideally, true positives would increase
154 without an equivalent increase (and possibly with a decrease) in false positive rates.
155 If they increase in tandem, attention does not have a net beneficial effect. Looking at
156 the effects of applying attention at different layers, we can see that attention at lower
157 layers is less effective at moving the performance in this space and that movement is in
158 somewhat random directions, although there is an average increase in performance with
159 moderate attentional strength. With attention applied at later layers, true positive
160 rates are more likely to increase for moderate attentional strengths, while substantial

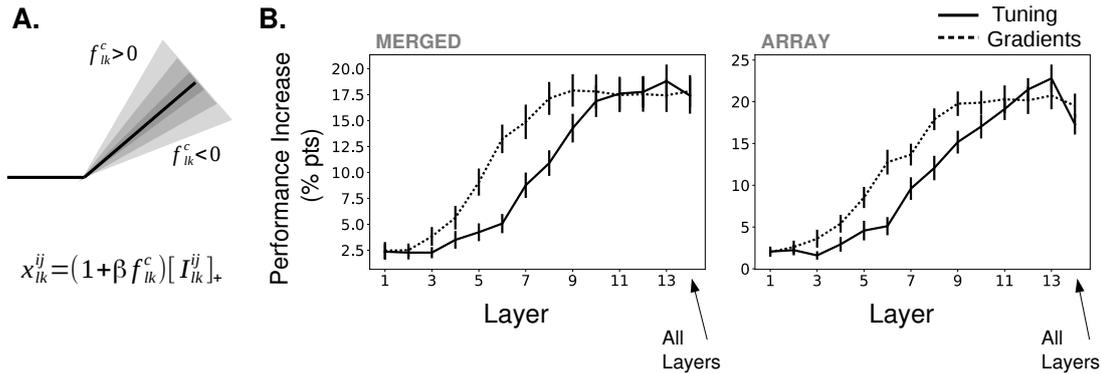


Figure 3: Effects of Applying Feature-Based Attention on Object Category Tasks. A.) Schematic of how attention modulates the activity function. All units in a feature map are modulated the same way. The slope of the activation function is altered based on the tuning (or gradient) value, f_{lk}^c , of a given feature map (here, the k^{th} feature map in the l^{th} layer) for the attended category, c , along with an overall strength parameter β . I_{lk}^{ij} is the input to this unit from the previous layer. For more information, see Methods 4.5. B.) Average increase in binary classification performance as a function of layer attention is applied at (solid line represents using tuning values, dashed line using gradient values, errorbars +/- S.E.M.). 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). In all cases, best performing strength from the range tested is used for each instance. Performance shown separately for merged (left) and array (right) images. Gradients perform significantly ($p < .05$, $N = 20$) better than tuning at layers 5-8 ($p = 4.6e-3, 2.6e-5, 6.5e-3, 4.4e-3$) for merged images and 5-9 ($p = 3.1e-2, 2.3e-4, 4.2e-2, 6.1e-3, 3.1e-2$) for array images.

161 false positive rate increases occur only with higher strengths. Thus, when attention
 162 is applied with modest strength at layer 13, most categories see a substantial increase
 163 in true positives with only modest increases in false positives. As strength continues
 164 to increase however, false positives increase substantially and eventually lead to a net
 165 decrease in overall classifier performance (representing as crossing the dotted line in
 166 Figure 4A).

167 Applying attention according to negated tuning values leads to a decrease in true
 168 and false positive values with increasing attention strength, which decreases overall
 169 performance (Supplementary Figure 9A). This verifies that the effects of attention are
 170 not from non-specific changes in activity.

171 Experimentally, when switching from no or neutral attention, neurons in MT
 172 showed an average increase in activity of 7% when attending their preferred motion
 173 direction (and similar decrease when attending the non-preferred) [51]. In our model,
 174 when $\beta = .75$ (roughly the value at which performance peaks at later layers; Figure 9),
 175 given the magnitude of the tuning values (average magnitude: .38), attention scales
 176 activity by an average of 28.5%. This value refers to how much activity is modulated
 177 in comparison to the $\beta = 0$ condition, which is probably more comparable to passive
 178 or anesthetized viewing, as task engagement has been shown to scale neural responses
 179 generally [64]. This complicates the relationship between modulation strength in our
 180 model and the values reported in the data.

181 To allow for a more direct comparison, in Figure 4B, we collected the true and
 182 false positive rates obtained experimentally during different object detection tasks
 183 (explained in Methods 4.9), and plotted them in comparison to the model results
 184 when attention is applied at layer 13 using tuning values (pink line) or gradient value
 185 (brown line) (results are similar). Five experiments (second through sixth studies) are

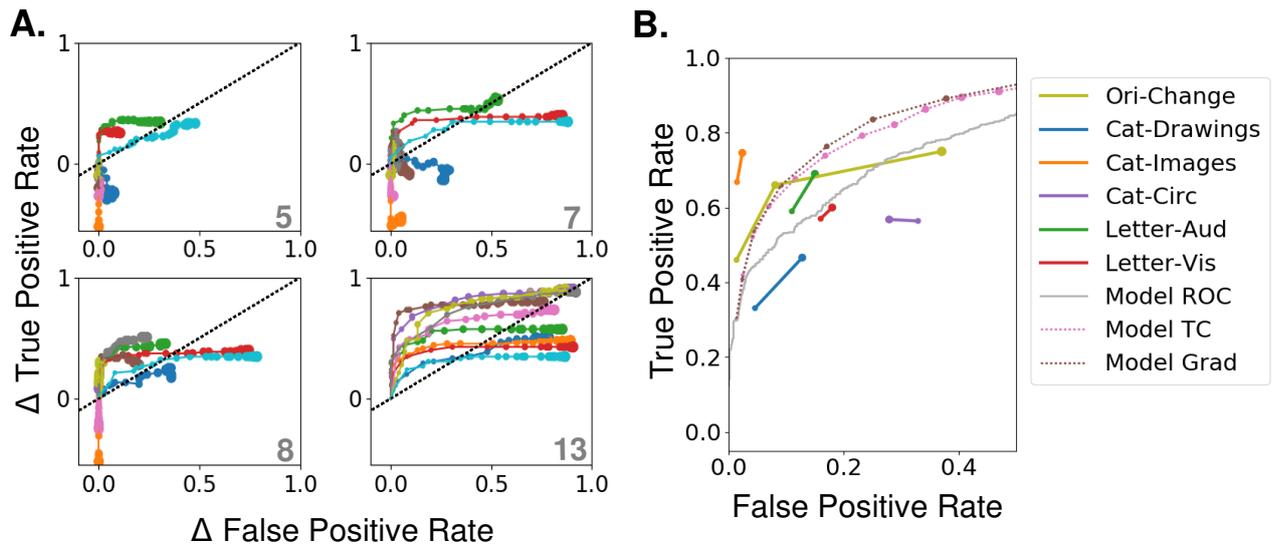


Figure 4: Effects of Varying Attention Strength A.) Effect of increasing attention strength (β) in true and false positive rate space for attention applied at each of four layers (layer indicated in bottom right of each panel, attention applied using tuning values). Each line represents performance for an individual category (only 10 categories shown for visibility), with each increase in dot size representing a .15 increase in β . Baseline (no attention) values are subtracted for each category such that all start at (0,0). The black dotted line represents equal changes in true and false positive rates. B.) Comparisons from experimental data. The true and false positive rates from six experiments in four previously published studies are shown for conditions of increasing attentional strength (solid lines). Cat-Drawings=[50], Exp. 1; Cat-Images=[50],Exp. 2; Objects=[39], Letter-Aud.=[49], Exp. 1; Letter-Vis.=[49], Exp. 2. Ori-Change=[53]. See Methods 4.9 for details of experiments. Dotted lines show model results for merged images, averaged over all 20 categories, when attention is applied using either tuning (TC) or gradient (Grad) values at layer 13. Model results are shown for attention applied with increasing strengths (starting at 0, with each increasing dot size representing a .15 increase in β). Receiver operating curve (ROC) for the model using merged images, which corresponds to the effect of changing the threshold in the final, readout layer, is shown in gray.

186 human studies. In all of these, uncued trials are those in which no information about
187 the upcoming visual stimulus is given, and therefore attention strength is assumed to
188 be low. In cued trials, the to-be-detected category is cued before the presentation of a
189 challenging visual stimulus, allowing attention to be applied to that object or category.

190 The majority of these experiments show a concurrent increase in both true and
191 false positive rates as attention strength is increased (with the exception of Cat-Circ,
192 which has a larger initial false positive rate and shows a decrease in false positives with
193 stronger attention). The rates in the uncued conditions (smaller dots) are generally
194 higher than the rates produced by the $\beta = 0$ condition in our model, consistent with
195 neutrally cued conditions corresponding to $\beta > 0$. We find (see Methods 4.9), that the
196 average corresponding β value for the neutral conditions is .37 and for the attended
197 conditions .51. Because attention scales activity by $1 + \beta f_c^{lk}$ (where f_c^{lk} is the tuning
198 value), these changes correspond to a $\approx 5\%$ change in activity. Thus, according to our
199 model, the size of observed performance changes is broadly consistent with the size of
200 observed neural changes.

201 The first dataset included in the plot (Ori-Change; yellow line in Figure 4B) comes
202 from a macaque change detection study (see Methods 4.9 for details). Because the
203 attention cue was only 80% valid, attention strength could be of three levels: low (for
204 the uncued stimuli on cued trials), medium (for both stimuli on neutrally-cued trials),
205 or high (for the cued stimuli on cued trials). Like the other studies, this study shows a
206 concurrent increase in both true positive (correct change detection) and false positive
207 (premature response) rates with increasing attention strength. However, for the model
208 to achieve the performance changes observed between low and medium attention a
209 roughly 12% activity change is needed, but average V4 firing rates recorded during
210 this task show an increase of only 3.6%. This discrepancy may suggest that changes
211 in correlations [14] or firing rate changes in areas aside from V4 also make important
212 contributions to observed performance changes.

213 Finally, we show the change in true and false positive rates when the threshold of
214 the final layer binary classifier is varied (a "receiver operating characteristic" analy-
215 sis, Figure 4B, gray line; no attention was applied during this analysis). Comparing
216 this to the pink line, it is clear that varying the strength of attention applied at the
217 final convolutional layer has more favorable performance effects than altering the clas-
218 sifier threshold (which corresponds to an additive effect of attention at the classifier
219 layer). This points to the limitations that could come from attention targeting only
220 downstream readout areas.

221 Overall, the model roughly matches experiments in the amount of neural modula-
222 tion needed to create the observed changes in true and false positive rates. However,
223 it is clear that the details of the experimental setup are relevant, and changes aside
224 from firing rate and/or outside the ventral stream also likely play a role [62].

225 *2.4. Feature-based Attention Enhances Performance on Orientation Detection Task*

226 Some of the results presented above, particularly those related to the layer at which
227 attention is applied, may be influenced by the fact that we are using an object catego-
228 rization task. To see if results are comparable using the simpler stimuli frequently used
229 in macaque studies, we created an orientation detection task (Figure 5A). Here, binary
230 classifiers trained on full-field oriented gratings are tested using images that contain
231 two gratings of different orientation and color. The performance of these binary clas-
232 sifiers without attention is above chance (distribution across orientations shown in

233 inset of Figure 5A). The performance of the binary classifier associated with vertical
234 orientation (0 degrees) was abnormally high (92% correct without attention, other ori-
235 entations average 60.25%) and this orientation was excluded from further performance
236 analysis.

237 Attention is applied according to orientation tuning values of the feature maps
238 (tuning quality by layer is shown in Figure 5B) and tested across layers. We find
239 (Figure 5D, solid line) that the trend in this task is similar to that of the object
240 task: applying attention at later layers leads to larger performance increases (14.4%
241 percentage point increase at layer 10). This is despite the fact that orientation tuning
242 quality peaks in the middle layers.

243 We also calculate the gradient values for this orientation detection task. While
244 overall the correlations between gradient values and tuning values are lower (and even
245 negative for early layers), the average correlation still increases with layer (Figure
246 5C), as with the category detection task. Importantly, while this trend in correlation
247 exists in both detection tasks tested here, it is not a universal feature of the network
248 or an artifact of how these values are calculated. Indeed, an opposite pattern in
249 the correlation between orientation tuning and gradient values is shown when using
250 attention to orientation to classify the color of a stimulus with the attended orientation
251 (Supplementary Figure 10A, Methods 4.4 and 4.5.2).

252 The results of applying attention according to gradient values is shown in Figure
253 5D (dashed line). Here again, using gradient value creates similar trends as using
254 tuning values, with gradient values performing better in the middle layers.

255 *2.5. Feature-based Attention Primarily Influences Criteria and Spatial Attention Pri-* 256 *marily Influences Sensitivity*

257 Signal detection theory is frequently used to characterize the effects of attention
258 on performance [88]. Here, we use a joint feature-spatial attention task to explore
259 effects of attention in the model. The task uses the same two-grating stimuli described
260 above. The same binary orientation classifiers are used and the task of the model is
261 to determine if a given orientation is present in a given quadrant. Performance is
262 then measured when attention is applied according to orientation, quadrant, or both
263 (effects are combined additively, for more, see Methods 4.5). Two key signal detection
264 measurements are computed: criteria is a measure of the threshold that's used to mark
265 an input as positive, with a higher criteria leading to fewer positives; and sensitivity
266 is a measure of the separation between the populations of true positive and negatives,
267 with higher sensitivity indicating a greater separation.

268 Figure 5E shows how criteria decreases more when feature-based attention is ap-
269 plied alone than when spatial is. Intuitively, feature-based attention shifts the repre-
270 sentations of all stimuli in the direction of the attended category, implicitly lowering
271 the detection threshold. Sensitivity increases more for spatial attention alone than
272 feature-based attention alone, indicating that spatial attention amplifies differences in
273 the representation of whatever features are present. These general trends hold regard-
274 less of the layer at which attention is applied. Changes in true and false positive rates
275 for this task can be seen explicitly in Supplementary Figure 10B.

276 Experimentally—in line with our results—spatial attention was found to increase
277 sensitivity and (less reliably) decrease criteria [28, 19], and feature attention is known
278 to decrease criteria, with minimal effects on sensitivity [68, 1]. A study that looked
279 explicitly at the different effects of spatial and category-based attention [81] found

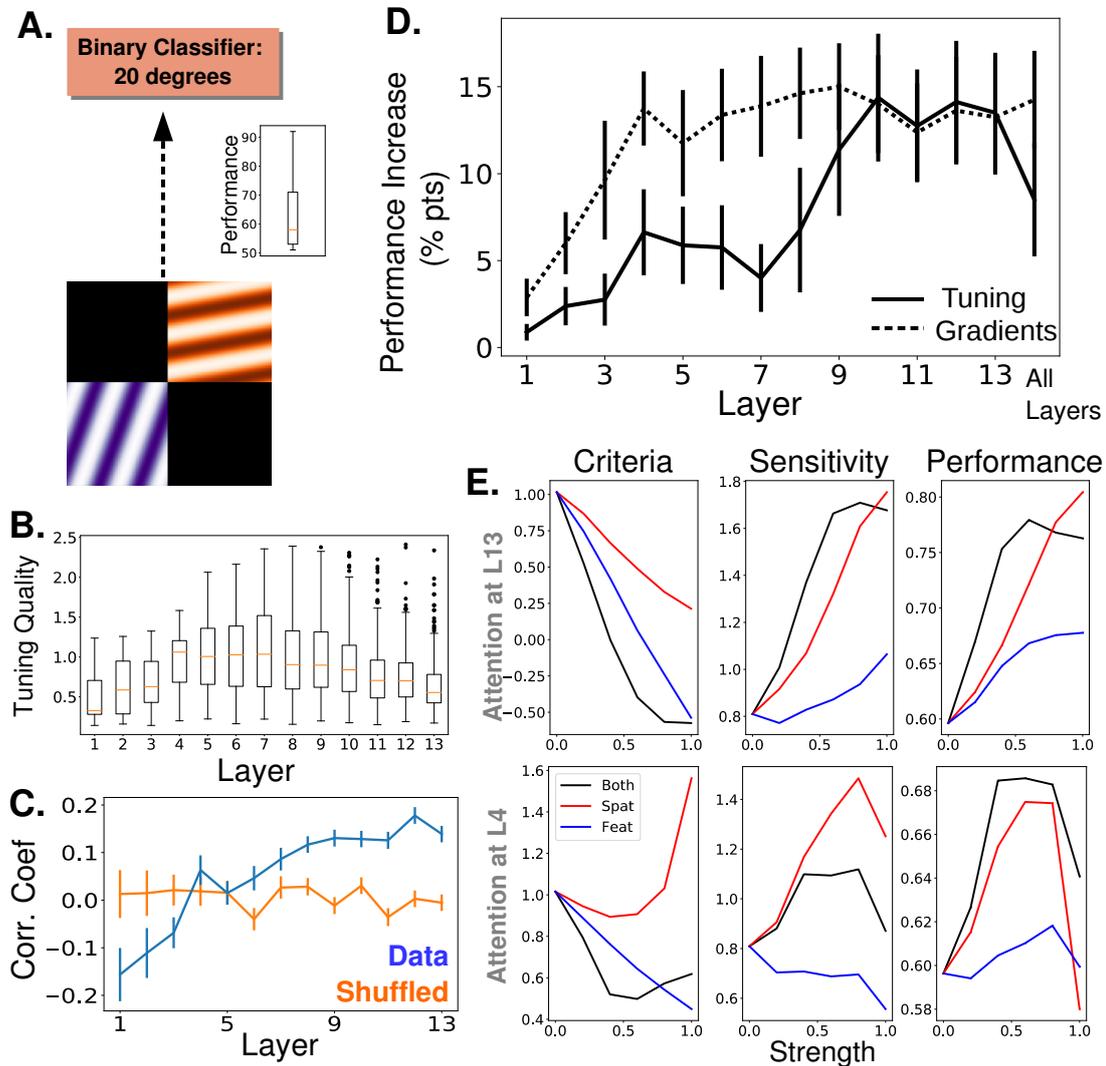


Figure 5: Attention Task and Results Using Oriented Gratings. A.) Orientation detection task. Like with the object category detection tasks, separate binary classifiers trained to detect each of 9 different orientations replaced the final layer of the network. Test images included 2 oriented gratings of different color and orientation located at 2 of 4 quadrants. Inset shows performance over 9 orientations without attention B.) Orientation tuning quality as a function of layer. C.) Average correlation coefficient between orientation tuning curves and gradient curves across layers (blue). Shuffled correlation values in orange. Errorbars are \pm S.E.M. D.) Comparison of performance on orientation detection task when attention is determined by tuning values (solid line) or gradient values (dashed line) and applied at different layers. As in Figure 3B, final column is performance when attention is applied at all layers, and best performing strength is used in all cases. Errorbars are \pm S.E.M. Gradients perform significantly ($p = 1.9e - 2$) better than tuning at layer 7. E.) Change in signal detection values and performance (percent correct) when attention is applied in different ways—spatial, feature (according to tuning), and both spatial and feature—for the task of detecting a given orientation in a given quadrant. Top row is when attention is applied at layer 13 and bottom when applied at layer 4.

280 that spatial attention increases sensitivity more than category-based attention (most
281 visible in their Experiment 3c, which uses natural images), and the effects of the two
282 are additive.

283 However, attention and priming are known to impact neural activity beyond pure
284 sensory areas [41, 16]. This idea is borne out by a study that aimed to isolate the
285 neural changes associated with sensitivity and criteria changes [48]. In this study, the
286 authors designed behavioral tasks that encouraged changes in behavioral sensitivity
287 or criteria exclusively: high sensitivity was encouraged by associating a given stimulus
288 location with higher overall reward, while high criteria was encouraged by rewarding
289 correct rejects more than hits (and vice versa for low sensitivity/criteria). Differences
290 in V4 neural activity were observed between trials using high versus low sensitivity
291 stimuli. No differences were observed between trials using high versus low criteria
292 stimuli. This indicates that areas outside of the ventral stream (or at least outside
293 V4) are capable of impacting criteria [80]. Importantly, it does not mean that changes
294 in V4 don't impact criteria, but merely that those changes can be countered by the
295 impact of changes in other areas. Indeed, to create sessions wherein sensitivity was
296 varied without any change in criteria, the authors had to increase the relative correct
297 reject reward (i.e., increase the criteria) at locations of high absolute reward, which
298 may have been needed to counter a decrease in criteria induced by attention-related
299 changes in V4 (similarly, they had to decrease the correct reject reward at low reward
300 locations). Our model demonstrates clearly how such effects from sensory areas alone
301 can impact detection performance, which, in turn highlights the role downstream areas
302 may play in determining the final behavioral outcome.

303 *2.6. Recordings Show How Feature Similarity Gain Effects Propagate*

304 To explore how attention applied at one location in the network impacts activity
305 later on, we apply attention at various layers and "record" activity at others (Figure
306 6A, in response to full field oriented gratings). In particular, we record activity of fea-
307 ture maps at all layers while applying attention at layers 2, 6, 8, 10, or 12 individually.

308 To understand the activity changes occurring at each layer, we use an analysis from
309 [51] that was designed to test for FSGM-like effects and is explained in Figure 6B.
310 Here, the activity of a feature map in response to a given orientation when attention is
311 applied is divided by the activity in response to the same orientation without attention.
312 These ratios are organized according to the feature map's orientation preference (most
313 to least) and a line is fit to them. According to the FSGM of attention, this ratio
314 should be greater than one for more preferred orientations and less than one for less
315 preferred, creating a line with an intercept greater than one and negative slope.

316 In Figure 6C, we plot the median value of the slopes and intercepts across all
317 feature maps at a layer, when attention is applied at different layers (indicated by
318 color). When attention is applied directly at a layer according to its tuning values
319 (left), FSGM effects are seen by default (intercept values are plotted in terms of how
320 they differ from one; comparable average values from [51] are intercept: .06 and slope:
321 0.0166, but we are using $\beta = 0$ for the no-attention condition in the model which,
322 as mentioned earlier, is not necessarily the best analogue for no-attention conditions
323 experimentally. Therefore we use these measures to show qualitative effects). As these
324 activity changes propagate through the network, however, the FSGM effects wear off,
325 suggesting that activating units tuned for a stimulus at one layer does not necessarily
326 activate cells tuned for that stimulus at the next. This misalignment between tuning

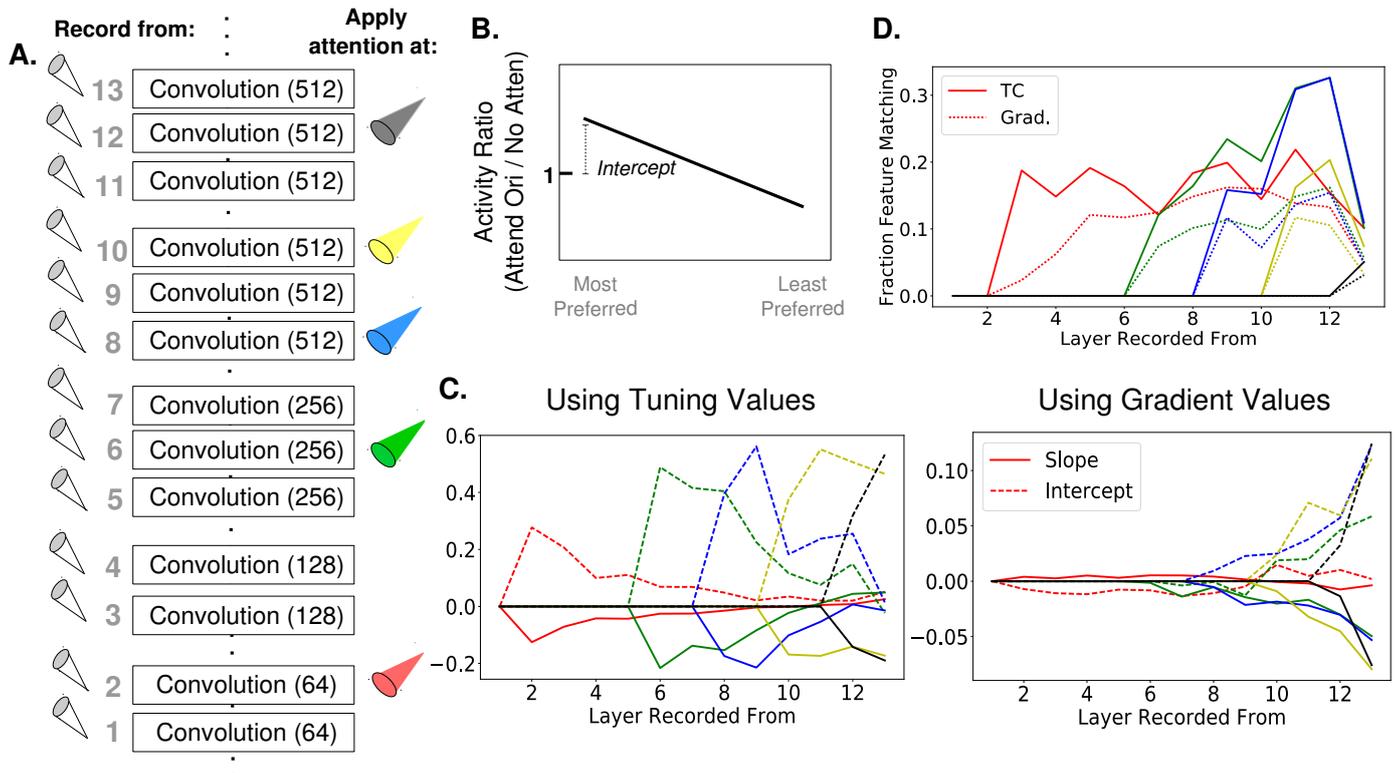


Figure 6: How Attention-Induced Activity Changes Propagate through the Network. A.) Recording setup. The spatially averaged activity of feature maps at each layer was recorded (left) while attention was applied at layers 2, 6, 8, 10, or 12 individually. Activity was in response to a full field oriented grating. B.) Schematic of metric used to test for the feature similarity gain model. Activity when a given orientation is present and attended is divided by the activity when no attention is applied, giving 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 in (C). Intercept values are plotted in terms of how they differ from 1, so positive values are an intercept greater than 1. (FSGM predicts negative slope and positive intercept) C.) The median slope (solid line) and intercept (dashed line) values as described in (B) plotted for each layer when attention is applied to the layer indicated by the line color as labeled in (A). On the left, attention applied according to tuning values and on the right, attention applied according to gradient values. D.) Fraction of feature maps displaying feature matching behavior at each layer when attention is applied at the layer indicated by line color. Shown for attention applied according to tuning (solid lines) and gradient values (dashed line).

327 at one layer and the next explains why attention applied at all layers simultaneously
328 isn't more effective (Figure 3B). In fact, applying attention to a category at one layer
329 can actually have effects that counteract attention at a later layer (see Supplementary
330 Figure 11).

331 In Figure 6C (right), we show the same analysis, but while applying attention
332 according to gradient values. The effects at the layer at which attention is applied do
333 not look strongly like FSGM, however FSGM properties evolve as the activity changes
334 propagate through the network, leading to clear FSGM-like effects at the final layer.
335 Finding FSGM-like behavior in neural data could thus be a result of FSGM effects at
336 that area or non-FSGM effects at an earlier area (here, attention applied according to
337 gradients which, especially at earlier layers, are not aligned with tuning).

338 An alternative model of the neural effects of attention—the feature matching (FM)
339 model—suggests that the effect of attention is to amplify the activity of a neuron when-
340 ever the stimulus in its receptive field matches the attended stimulus. In Figure 6D,
341 we calculate the fraction of feature maps at a given layer that show feature match-
342 ing behavior (defined as having activity ratios greater than one when the stimulus
343 orientation matches the attended orientation for both preferred and anti-preferred ori-
344 entations). As early as one layer post-attention, some feature maps start showing
345 feature matching behavior. The fact that the attention literature contains conflicting
346 findings regarding the feature similarity gain model versus the feature matching model
347 [61, 72] may result from this finding that FSGM effects can turn into FM effects as
348 they propagate through the network. In particular, this mechanism can explain the
349 observations that feature matching behavior is observed more in FEF than V4 [96] and
350 that match information is more easily read out from perirhinal cortex than IT [63].

351 Finally, we investigated the extent to which measures of attention's neural effects
352 correlate with changes in performance (see Methods 4.8). For this, we used a measure
353 of FSGM-like activity that could be calculated on an image-by-image basis. We also
354 created a separate measure, inspired by our gradient approach, that considers activity
355 in light of its downstream effects. Specifically, we measure the extent to which activity
356 when attention is applied becomes more like activity when images (in the absence of
357 attention) are classified as containing the given orientation ("Vector Angle" method,
358 see Figure 7A and B). For the purposes of this analysis, we consider images that,
359 without attention, give false negative responses and measure performance as the rate
360 at which these are converted to true positives by attention. For both measures and
361 whether attention is applied according to tuning or gradients, activity changes are more
362 correlated with performance in later layers (Figure 7C). When attention is applied
363 with gradients, the gradient-inspired measure is better correlated with performance
364 changes than the feature similarity gain model. When recording activity from early
365 layers, this measure also performs better even when attention is applied according to
366 tuning curves. As this new measure is experimentally testable, it would be valuable
367 to see how well it predicts performance on real neural data.

368 3. Discussion

369 In this work, we utilized a deep convolutional neural network (CNN) as a model of
370 the visual system to probe the relationship between neural activity and performance.
371 Specifically, we provide a formal mathematical definition of the feature similarity gain
372 model (FSGM) of attention, the basic tenets of which have been described in several

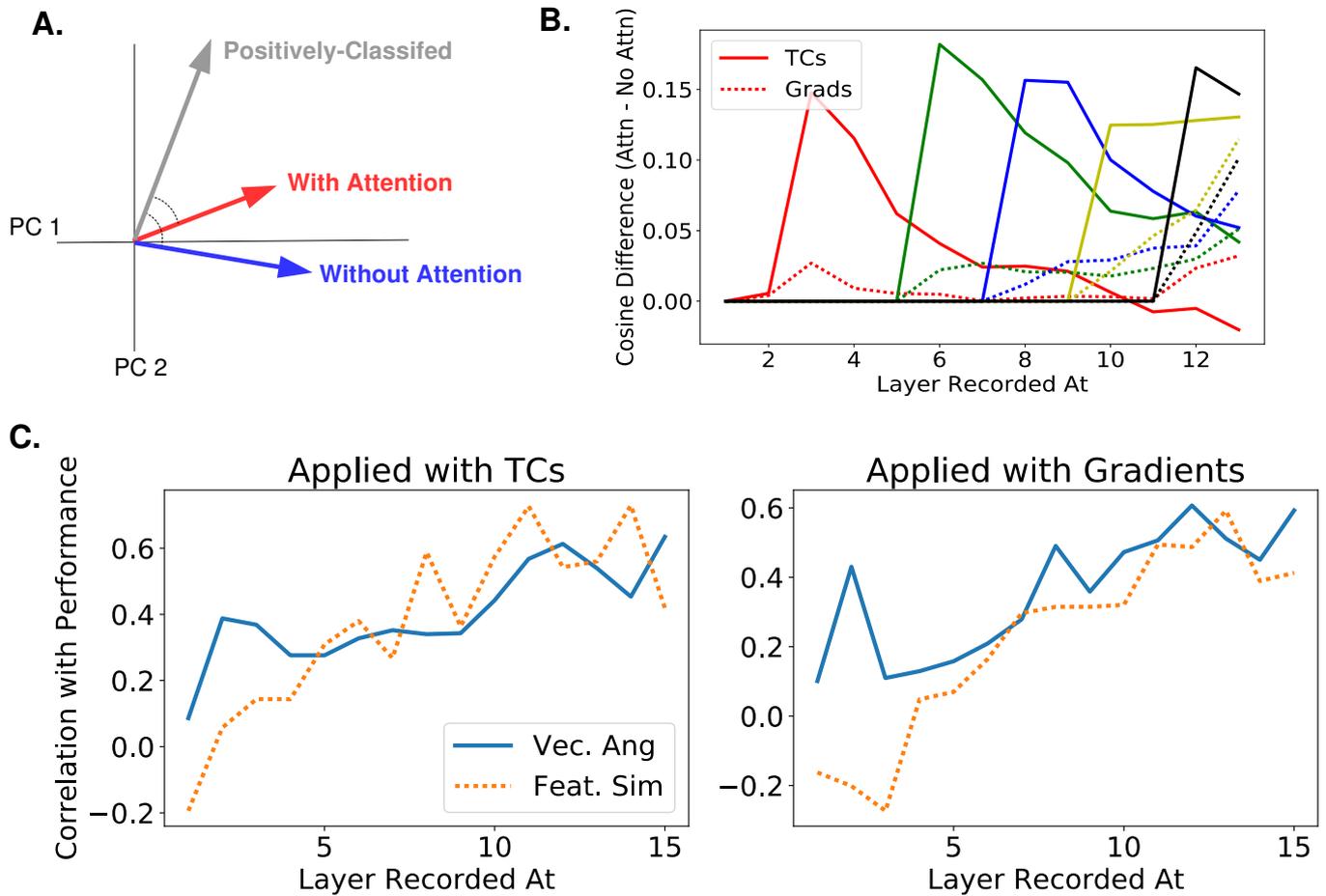


Figure 7: How Activity Changes Correlate with Performance Changes A.) A new measure of activity changes inspired by gradient values. The gray vector represents the average pattern of neural activity in response to images the classifier indicates as containing the given orientation (i.e., positively-classified in the absence of attention). The blue vector (activity without attention) and red vector (activity vector when attention is applied) are then made using images that contain the orientation but are not initially classified as containing it. Assuming that attention makes activity look more like activity during positive classification, this measure compares the angle between the positively-classified and with-attention vectors to the angle between the positively-classified and without-attention vectors. We use $\cos(\theta)$ as the measure, but results are similar using θ . B.) Using the same color scheme as Figure 6, this plot shows how attention applied at different layers causes activity changes throughout the network, as measured by the vector method introduced in (A). Specifically, the cosine of the angle between the positively-classified and without-attention vectors is subtracted from the cosine of the angle between the positively-classified and with-attention vectors. Solid lines indicate median value of this difference (across images) when attention is applied with tuning curves and dashed line when applied with gradients. C.) The correlation coefficient between the change in true positive rate with attention and activity changes as measured by: difference in cosines of angles (solid line) or feature similarity gain model-like behavior (dashed line, see Methods 4.8 for how this is calculated). Activity and performance changes are collected when attention is applied at different layers and various strengths according to tuning curves (left) or gradient values (right). Correlation coefficients calculated for activity changes from both application methods combined can be seen in Supplementary Figure 12

373 experimental studies. This formalization allows us to investigate the FSGM’s ability
374 to enhance a CNN’s performance on challenging visual tasks. We show that neural
375 activity changes matching the type and magnitude of those observed experimentally
376 can indeed lead to performance changes of the kind and magnitude observed experi-
377 mentally. Furthermore, these results hold for a variety of tasks. We also use the full
378 observability of the model to investigate the relationship between tuning and function.

379 A finding from our model is that the layer at which attention is applied can have
380 a large impact on performance: attention (particularly applied according to tuning)
381 at early layers does little to enhance performance while attention at later layers such
382 as 9-13 is most effective. According to [25], these layers correspond most to areas V4
383 and LO. Such areas are known and studied for reliably showing attentional effects,
384 whereas earlier areas such as V1 are generally not [47]. In a study involving detection
385 of objects in natural scenes, the strength of category-specific preparatory activity in
386 object selective cortex was correlated with performance, whereas such preparatory
387 activity in V1 was anti-correlated with performance [65]. This is in line with our
388 finding that feature-based attention effects at earlier areas can counter the beneficial
389 effects of that attention at later areas (Supplementary Figure 11).

390 While CNNs have representations that are similar to the ventral stream, they lack
391 many biological details including recurrent connections, dynamics, cell types, and noisy
392 responses. Preliminary work has shown that these elements can be incorporated into
393 a CNN structure, and attention can enhance performance in this more biologically-
394 realistic architecture [45]. Furthermore, while the current work does not include neural
395 noise independent of the stimulus, the fact that a given image is presented in many
396 contexts (different merged images or different array images) can be thought of as a
397 form of highly structured noise that does produce variable responses to the same image.

398 Another biological detail that this model lacks is "skip connections," where one
399 layer feeds into both the layer directly after it and deeper layers after that [30, 32]
400 as in connections from V2 to V4 or V4 to parietal areas [87]. Our results regarding
401 propagation of changes through the network suggest that synaptic distance from the
402 classifier is a relevant feature—one that is less straight forward to determine in a
403 network with skip connections. It may be that thinking about visual areas in terms of
404 their synaptic distance from decision-making areas such as prefrontal cortex [31] can
405 be more useful for the study of attention than thinking in terms of their distance from
406 the retina. Finally, a major challenge for understanding the biological implementation
407 of selective attention is determining how such a precise attentional signal is carried by
408 feedback connections. The machine learning literature on attention and learning may
409 inspire useful hypotheses on underlying brain mechanisms [91, 43].

410 While CNNs lack certain biological details, a benefit of using them as a model is
411 the ability to backpropagate error signals and understand causal relationships. Here
412 we use this to calculate gradient values that estimate how attention should modulate
413 activity, and compare these to the tuning values that the FSGM uses. The facts that
414 attention performs better in middle layers when guided by gradients than by tuning
415 values, while the two have correlated values and behave similarly at late layers, raise
416 an obvious question: are neurons really targeted according to their tuning, or does the
417 brain use something like gradient values? In [12] the correlation coefficient between an
418 index of tuning and an index of attentional modulation was .52 for a population of V4
419 neurons, suggesting factors other than selectivity influence attention. Furthermore,
420 many attention studies, including that one, use only preferred and anti-preferred stim-

421 uli and therefore don't include a thorough investigation of the relationship between
422 tuning and attentional modulation. [51] uses multiple stimuli to provide support for
423 the FSGM, however the interpretation is limited by the fact that they only report
424 population averages. [72] investigated the relationship between tuning strength and
425 the strength of attentional modulation on a cell-by-cell basis. While they did find a
426 correlation (particularly for binocular disparity tuning), it was relatively weak, which
427 leaves room for the possibility that tuning is not the primary factor that determines
428 attentional modulation.

429 There is a simple experiment that would distinguish whether factors beyond tuning,
430 such as gradients, play a role in guiding attention. It requires using two tasks with
431 very different objectives, which should produce different gradients, but with the same
432 attentional cue. An example is given by comparing Figure 5C to Supplementary Figure
433 10A: various gratings of various colors are simultaneously shown, and the task is either
434 to report whether a vertical (or other orientation) grating is present, or to report
435 the color of the vertical grating, with attention being cued in both cases for vertical
436 orientation. Gradient-based attention will produce different neural modulations for
437 the two tasks, while the FSGM predicts identical modulations.

438 A related finding from comparing gradient values with tuning values is that tuning
439 does not always predict how effectively one unit in the network will impact downstream
440 units or the classifier. In particular, applying attention according to gradient values
441 leads to changes that are hard to interpret when looking through the lens of tuning,
442 especially at earlier layers (Figure 6). However these changes eventually lead to large
443 and impactful changes at later layers. Because experimenters can easily control the
444 image, defining a cell's function in terms of how it responds to stimuli makes practical
445 sense. However, studies looking at the relationship between tuning and choice proba-
446 bilities also suggest that a neuron's preferred stimulus is not always an indication of
447 its causal role in classification [93, 67]. Studies that activate specific neurons in one
448 area and measure changes in another area or in behavioral output will likely be of
449 significant value for determining function. Thus far, coarse stimulation protocols have
450 found a relationship between the tuning of neural populations and their impact on
451 perception [56, 18, 75]. Ultimately though, targeted stimulation protocols and a more
452 fine-grained understanding of inter-area connections will be needed.

453 4. Methods

454 4.1. Network Model

455 This work uses a deep convolutional neural network (CNN) as a model of the
456 ventral visual stream. Convolutional neural networks are feedforward artificial neural
457 networks that consist of a few basic operations repeated in sequence, key among them
458 being the convolution. The specific CNN architecture used in the study comes from
459 [79] (VGG-16D) and is shown in Figure 1A (a previous variant of this work used
460 a smaller network [44]). For this study, all the layers of the CNN except the final
461 classifier layer were pre-trained using backpropagation on the ImageNet classification
462 task, which involves doing 1000-way object categorization (weights provided by [23]).
463 The training of the top layer is described in subsequent sections. Here we describe the
464 basic workings of the CNN model we use, with details available in [79].

465 The activity values of the units in each convolutional layer are the result of applying
466 a 2-D spatial convolution to the layer below, followed by positive rectification (rectified

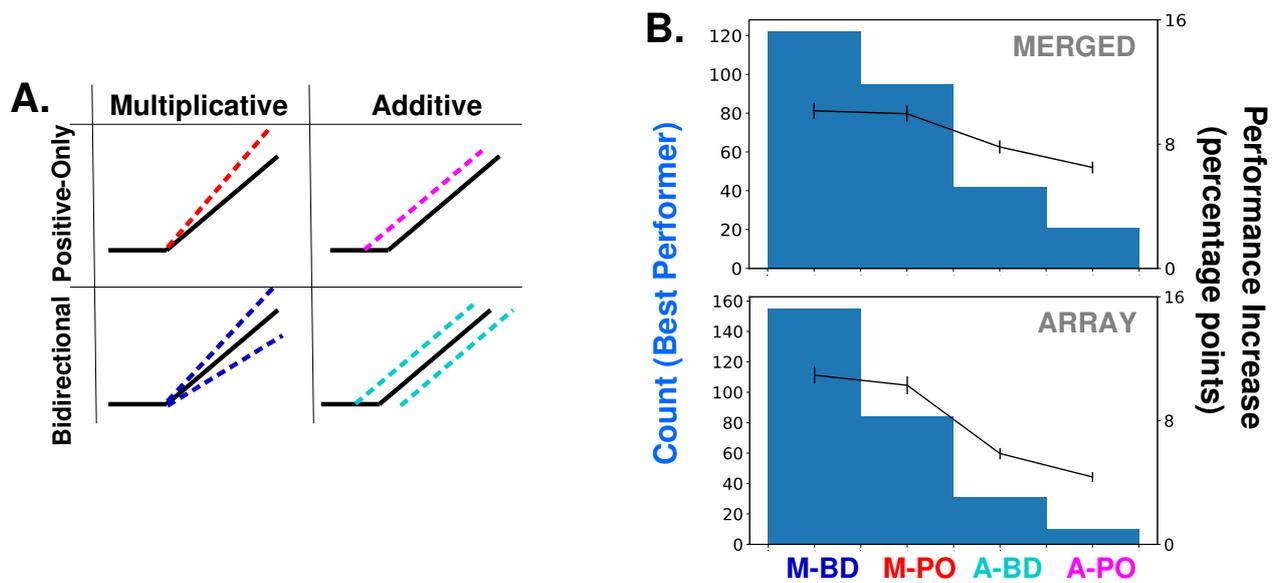


Figure 8: Supplementary Figure Associated with Figure 3. A.) Schematics of how attention can modulate the activity function. Feature-based attention modulates feature maps according to their tuning values but this modulation can scale the activity multiplicatively or additively, and can either only enhance feature maps that prefer the attended category (positive-only) or also decrease the activity of feature maps that do not prefer it (bidirectional). See Methods 4.5.4 for details of these implementations. The main body of this paper only uses multiplicative bi-directional. B.) Comparison of binary classification performance when attention is applied in each of the four ways described in (A). Considering the combination of attention applied to a given category at a given layer/layers 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 \pm S.E.M.).

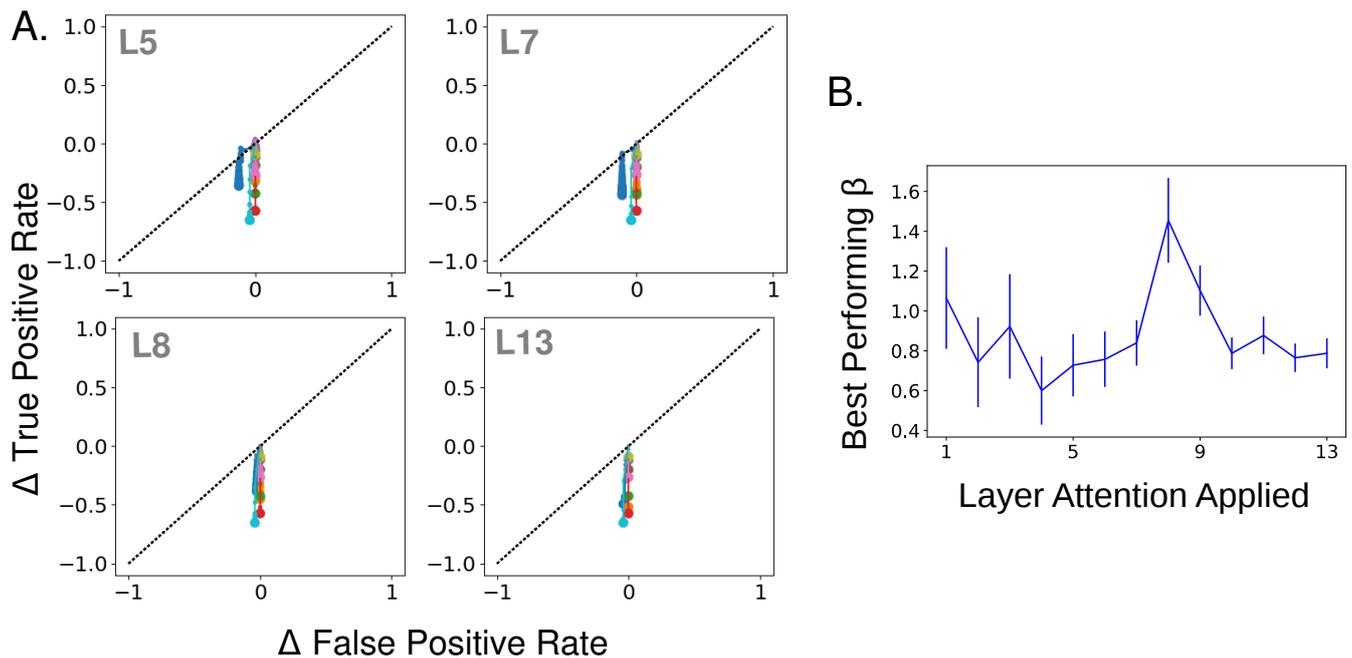


Figure 9: Supplementary Figure Associated with Figure 4. A.) Effect of strength increase in true and false positive rate space when tuning values are negated. Negated tuning values have the same overall level of positive and negative modulation but in the opposite direction of tuning for a given category. Plot same as in Figure 4A. Layer attention applied at indicated in gray. Attention applied in this way decreases true positives, and to a lesser extent false positives (the initial false positive rate when no attention is applied is very low). B. Mean best performing strength (β value, using regular non-negated attention) across categories as a function of the layer attention is applied at, according to merged images task. Errorbars \pm S.E.M.

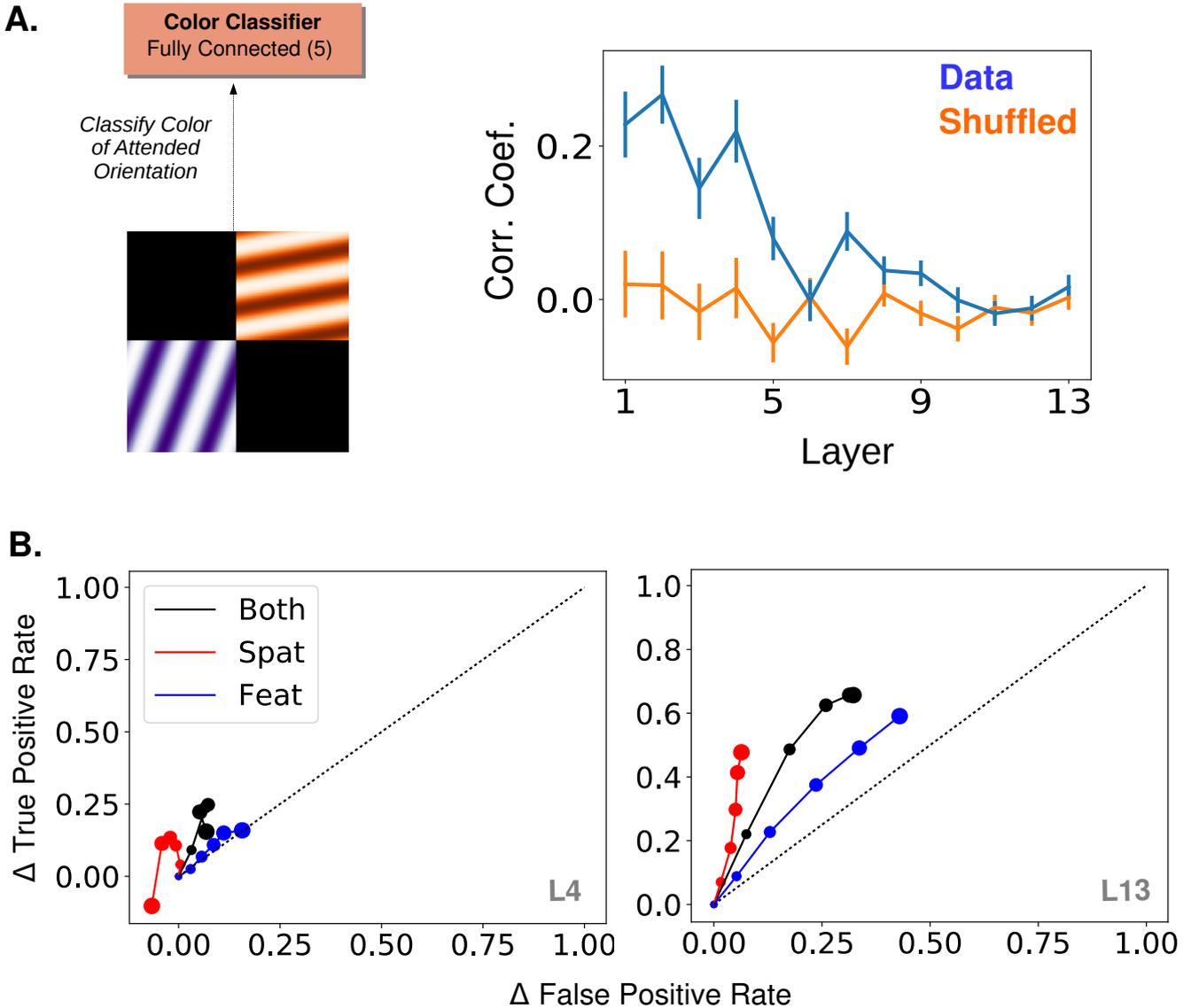


Figure 10: Supplementary Figure Associated with Figure 5. A.) "Cross-featural" attention task (left). Here, the final layer of the network is replaced with a color classifier and the task is to classify the color of the attended orientation in a two-orientation stimulus. Gradient values calculated for this task are correlated with orientation tuning values, and the mean correlation is plotted per layer (right, as in Figure 5C) B.) Effect of strength increase in true and false positive rate space when attention is applied according to quadrant, orientation, or both in the orientation detection task. Rates averaged over orientations/locations. Increasing dot size corresponds to .2 increase in β each. No-attention rates are subtracted and the black dotted line indicates equal increase in true and false positives. Layer attention applied at indicated in gray.

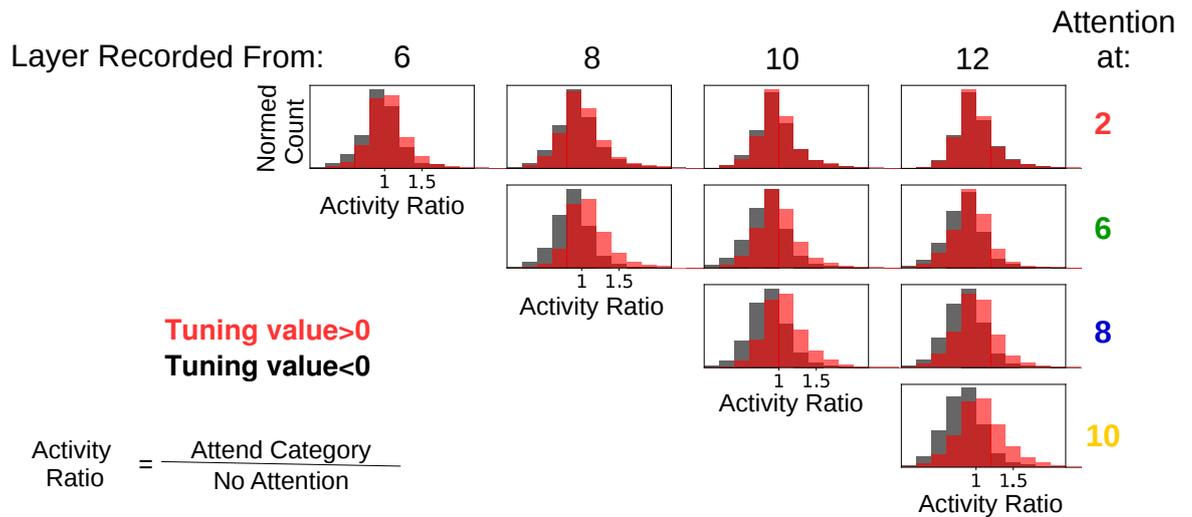


Figure 11: Supplementary Figure Associated with Figure 6. Feature attention at one layer often suppresses activity of the attended features at later layers. Activity ratios are shown for when attention is applied at various layers individually and activity is recorded from later layers. In all cases, the category attended was the same as the one present in the input image (standard ImageNet images used to ensure that these results are not influenced by the presence of other category features in the input). Histograms are of ratios of feature map activity when attention is applied to the category divided by activity when no attention is applied, split according to whether the feature map prefers (red) or does not prefer (black) the attended category. In many cases, feature maps that prefer the attended category have activity 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 demonstrates the ability of attention at a lower layer to change activity in ways opposite of the effects of attention at the recorded layer.

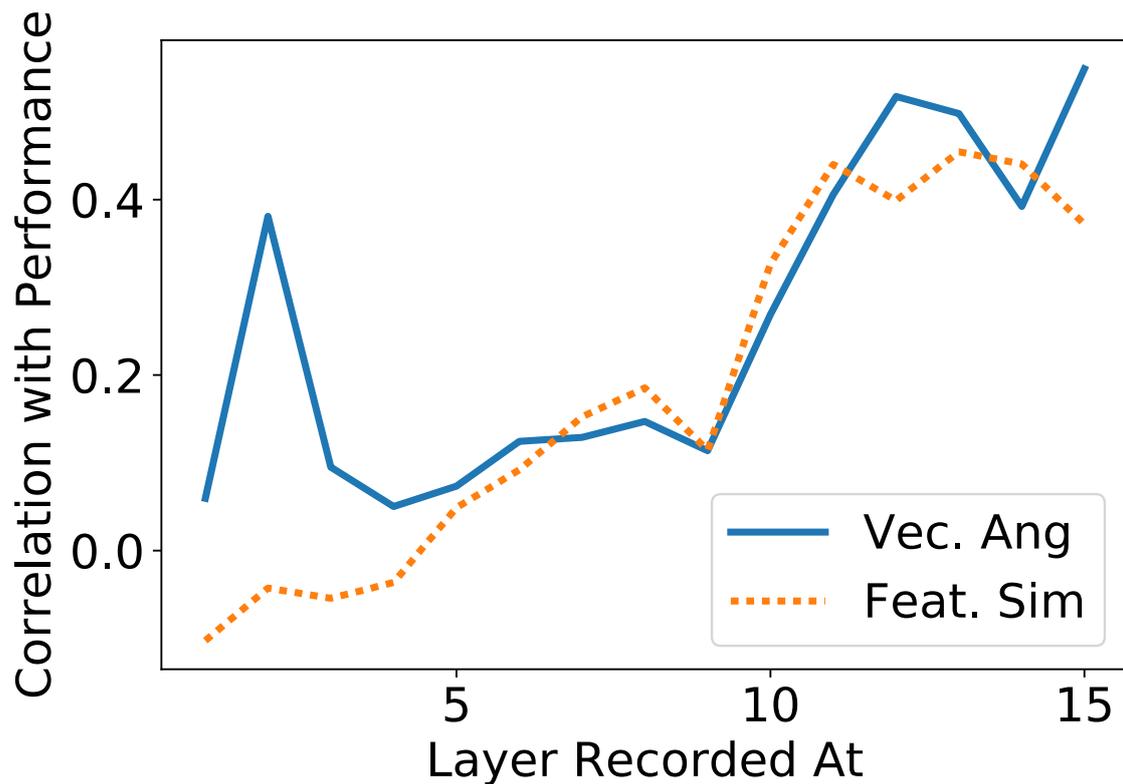


Figure 12: Supplementary Figure Associated with Figure 7. The increase in true positive rate with attention is correlated with activity changes as measured by: difference in cosines of angles (solid line) or feature similarity gain model-like behavior. Activity and performance changes are collected when attention is applied (at different layers and various strengths and according to tuning curves or gradient values (that is, all the data generated by these means are combined, and correlation coefficients are calculated; whereas in Figure 7C correlation coefficients were calculated separately for instances when attention was applied according to tuning or according to gradients).

467 linear 'ReLU' nonlinearity):

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

468 where \star indicates convolution, and $[x]_+ = x$ if $x > 0$, otherwise $x = 0$. W^{lk} is the
469 k^{th} convolutional filter at the l^{th} layer. The application of each filter results in a 2-D
470 feature map (the number of filters used varies across layers and is given in parenthesis
471 in Figure 1A). x_{ij}^{lk} is the activity of the unit at the i, j^{th} spatial location in the k^{th}
472 feature map at the l^{th} layer. X^{l-1} is thus the activity of all units at the layer below
473 the l^{th} layer. The input to the network is a 224 by 224 pixel RGB image, and thus the
474 first convolution is applied to these pixel values. Convolutional filters are 3x3. For the
475 purposes of this study the convolutional layers are most relevant, and will be referred
476 to according to their numbering in Figure 1A (numbers in parentheses indicate number
477 of feature maps per layer).

478 Max pooling layers reduce the size of the feature maps by taking the maximum
479 activity value of units in a given feature map in non-overlapping 2x2 windows. Through
480 this, the size of the feature maps decreases after each max pooling (layers 1 and 2: 224
481 x 224; 3 and 4: 112 x 112; 5, 6, and 7: 56 x 56. 8, 9, and 10: 28 x 28; 11, 12, and 13:
482 14 x 14).

483 The final two layers before the classifier are each fully-connected to the layer below
484 them, with the number of units per layer given in parenthesis in Figure 1A. Therefore,
485 connections exist from all units from all feature maps in the last convolutional layer
486 (layer 13) to all 4096 units of the next layer, and so on. The top readout layer of
487 the network in [79] contained 1000 units upon which a softmax classifier was used to
488 output a ranked list of category labels for a given image. Looking at the top-5 error
489 rate (wherein an image is correctly labeled if the true category appears in the top five
490 categories given by the network), this network achieved 92.7% accuracy. With the
491 exception of the gradient calculations described below, we did not use this 1000-way
492 classifier, but rather replaced it with a series of binary classifiers.

493 4.2. Object Category Attention Tasks

494 The tasks we use to probe the effects of feature-based attention in this network
495 involve determining if a given object category is present in an image or not, similar to
496 tasks used in [81, 66, 39]. To have the network perform this specific task, we replaced
497 the final layer in the network with a series of binary classifiers, one for each category
498 tested (Figure 1B). We tested a total of 20 categories: paintbrush, wall clock, seashore,
499 paddlewheel, padlock, garden spider, long-horned beetle, cabbage butterfly, toaster,
500 greenhouse, bakery, stone wall, artichoke, modem, football helmet, stage, mortar,
501 consomme, dough, bathtub. Binary classifiers were trained using ImageNet images
502 taken from the 2014 validation set (and were therefore not used in the training of
503 the original model). A total of 35 unique true positive images were used for training
504 for each category, and each training batch was balanced with 35 true negative images
505 taken from the remaining 19 categories. The results shown here come from using
506 logistic regression as the binary classifier, though trends in performance are similar if
507 support vector machines are used.

508 Once these binary classifiers are trained, they are then used to classify more chal-
509 lenging test images. Experimental results suggest that classifiers trained on unat-
510 tended and isolated object images are appropriate for reading out attended objects in
511 cluttered images [95]. These test images are composed of multiple individual images

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

522 4.3. Object Category Gradient Calculations

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

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

532 where H and W are the spatial dimensions of layer l and N_c is the total number of
533 images from the category (here $N_c = 35$, and the merged images used were generated
534 from the same images used to generate tuning curves, described below). $E(n)$ is
535 the error of the 1000-way classifier in response to image n , which is defined as the
536 difference between the activity vector of the final layer (after the soft-max operation)
537 and a one-hot vector, wherein the correct label is the only non-zero entry. Because
538 we are interested in activity changes that would decrease the error value, we negate
539 this term. The gradient value we end up with thus indicates how the feature map's
540 activity would need to change to make the network more likely to classify an image as
541 the desired category. Repeating this procedure for each category, we obtain a set of
542 gradient values (one for each category, akin to a tuning curve), for each feature map:
543 \mathbf{g}^{lk} . Note that, as these values result from applying the chain rule through layers of
544 the network, they can be very small, especially for the earliest layers. For this study,
545 the sign and relative magnitudes are of more interest than the absolute values.

546 4.4. Oriented Grating Attention Tasks

547 In addition to attending to object categories, we also test attention on simpler
548 stimuli. In the orientation detection task, the network detects the presence of a given
549 orientation in an image. Again, the final layer of the network is replaced by a series
550 of binary classifiers, one for each of 9 orientations (0, 20, 40, 60, 80, 100, 120, 140,
551 and 160 degrees. Gratings had a frequency of .025 cycles/pixel). The training sets
552 for each were balanced (50% had only the given orientation and 50% had one of 8
553 other orientations) and composed of full field (224 by 224 pixel) oriented gratings in

554 red, blue, green, orange, or purple (to increase the diversity of the training images,
555 they were randomly degraded by setting blocks of pixels ranging uniformly from 0%
556 to 70% of the image to 0 at random). Test images were each composed of two oriented
557 gratings of different orientation and color (same options as training images). Each
558 of these gratings were of size 112 by 112 pixels and placed randomly in a quadrant
559 while the remaining two quadrants were black (Figure 5A). Again, the test sets were
560 balanced and performance was measured as the average of the true positive and true
561 negative rates (100 test images per orientation).

562 These same test images were used for a task wherein the network had to classify the
563 color of the grating that had the attended orientation (cross-featural task paradigms
564 like this are commonly used in attention studies, such as [74]). For this, the final layer
565 of the network was replaced with a 5-way softmax color classifier. This color classifier
566 was trained using the same full field oriented gratings used to train the binary classifiers
567 (therefore, the network saw each color at all orientation values).

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

573 4.5. How Attention is Applied

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

580 4.5.1. Tuning Values

581 To determine tuning to the 20 object categories used, we presented the network
582 with images of each object category (the same images on which the binary classifiers
583 were trained) and measured the relative activity levels. Because feature attention is
584 a spatially global phenomena [94, 73], we treat all units in a feature map identically,
585 and calculate tuning by averaging over them.

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

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

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

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

592 That is, a feature map's tuning value for a given category is merely the average
593 activity of that feature map in response to images of that category, with the mean

594 activity under all image categories subtracted, divided by the standard deviation of
595 the activity across all images. These tuning values determine how the feature map is
596 modulated when attention is applied to the category. Taking these values as a vector
597 over all categories, \mathbf{f}_{lk} , gives a tuning curve for the feature map. We define the overall
598 tuning quality of a feature map as its maximum absolute tuning value: $\max(|\mathbf{f}_{lk}|)$. To
599 determine expected tuning quality by chance, we shuffled the responses to individual
600 images across category and feature map at a given layer and calculated tuning quality
601 for this shuffled data.

602 We also define the category with the highest tuning value as that feature map's
603 most preferred, and the category with the lowest (most negative) value as the least or
604 anti-preferred.

605 We apply the same procedure to generate tuning curves for orientation by using
606 the full field gratings used to train the orientation detection classifiers. The orientation
607 tuning values were used when applying attention in these tasks.

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

612 4.5.2. Gradient Values

613 In addition to applying attention according to tuning, we also attempt to generate
614 the "best possible" attentional modulation by utilizing gradient values. These gradient
615 values are calculated slightly differently from those described above (4.3), because they
616 are meant to represent how feature map activity should change in order to increase
617 binary classification performance, rather than just increase the chance of classifying
618 an image as a certain object.

619 The error functions used to calculate gradient values for the category and orienta-
620 tion detection tasks were for the binary classifiers associated with each object/orientation.
621 A balanced set of test images was used. Therefore a feature map's gradient value for
622 a given object/orientation is the averaged activity change that would increase binary
623 classification performance for that object/orientation. Note that on images that the
624 network already classifies correctly, gradients are zero. Therefore, the gradient values
625 are driven by the errors: false negatives (classifying an image as not containing the
626 category when it does) and false positives (classifying an image as containing the cate-
627 gory when it does not). In our detection tasks, the former error is more prevalent than
628 the latter, and thus is the dominant impact on the gradient values. Because of this,
629 gradient values calculated this way end up very similar to those described in Methods
630 4.3, as they are driven by a push to positively classify the input as the given category.

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

638 In the orientation detection task, the test images used for gradient calculations (50
639 images per orientation) differed from those used to assess performance. For the object
640 detection task, images used for gradient calculations (45 per category; preliminary

641 tests for some categories using 90 images gave similar results) were drawn from the
642 same pool as, but different from, those used to test detection performance. Gradient
643 values were calculated separately for merged and array images.

644 4.5.3. Spatial Attention

645 In the feature similarity gain model of attention, attention is applied according to
646 how much a cell prefers the attended feature, and location is considered a feature like
647 any other. In CNNs, each feature map results from applying the same filter at different
648 spatial locations. Therefore, the 2-D position of a unit in a feature map represents
649 more or less the spatial location to which that unit responds. Via the max-pooling
650 layers, the size of each feature map shrinks deeper in the network, and each unit
651 responds to a larger area of image space, but the "retinotopy" is still preserved. Thus,
652 when we apply spatial attention to a given area of the image, we enhance the activity
653 of units in that area of the feature maps and decrease the activity of units in other
654 areas. In this study, spatial attention is applied to a given quadrant of the image.

655 4.5.4. Implementation Options

656 The values discussed above determine how strongly different feature maps or units
657 should be modulated under different attentional conditions. We will now lay out the
658 different implementation options for that modulation. In the main body of this work,
659 the multiplicative bidirectional form of attention is used. Other implementations are
660 only used for the Supplementary Results.

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

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

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

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

666 Strength of attention is varied via the weighting parameter, β . For the additive effect,
667 manipulations are multiplied by μ_l , the average activity level across all units of layer
668 l in response to all images (for each of the 13 layers respectively: 20, 100, 150, 150,
669 240, 240, 150, 150, 80, 20, 20, 10, 1). When gradient values are used in place of tuning
670 values, we normalize them by the maximum value at a layer, to be the same order of
671 magnitude as the tuning values: $\mathbf{g}^l / \max(|\mathbf{g}^l|)$.

672 Recall that for feature-based attention all units in a feature map are modulated
673 the same way, as feature attention has been found to be spatially global. In the case
674 of spatial attention, however, tuning values are not used and a unit's modulation is
675 dependent on its location in the feature map. Specifically, the tuning value term is set
676 to +1 if the i, j position of the unit is in the attended quadrant and to -1 otherwise.
677 For feature attention tasks, β ranged from 0 to a maximum of 11.85 (object attention)
678 and 0 to 4.8 (orientation attention). For spatial attention tasks, it ranged from 0 to 1.

679 Next, we chose whether attention only enhances units that prefer the attended
680 feature, or also decreases activity of those that don't prefer it. For the latter, the

681 tuning values are used as-is. For the former, the tuning values are positively-rectified:
682 $[\mathbf{f}^{lk}]_+$.

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

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

690 4.6. *Signal Detection Calculations*

691 For the joint spatial-feature attention task, we calculated criteria (c , "threshold")
692 and sensitivity (d') using true (TP) and false (FP) positive rates as follows [48] :

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

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

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

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

700 To prevent the individual terms in these expressions from going to $\pm\infty$, false
701 positive rates of $< .01$ were set to $.01$ and true positive rates of $> .99$ were set to $.99$.

702 4.7. *"Recording" Procedures*

703 We examined the effects that applying attention at certain layers in the network
704 (specifically 2, 6, 8, 10, and 12) has on activity of units at other layers. Attention
705 was applied with $\beta = .5$ unless otherwise stated. The recording setup is designed
706 to mimic the analysis of [51]. Here, the images presented to the network are full-
707 field oriented gratings of all orientation-color combinations. Feature map activity is
708 measured as the spatially averaged activity of all units in a feature map in response to
709 an image. Activity in response to a given orientation is further averaged over all colors.
710 We calculate the ratio of activity when attention is applied to a given orientation
711 (and the orientation is present in the image) over activity in response to the same
712 image when no attention is applied. These ratios are then organized according to
713 orientation preference: the most preferred is at location 0, then the average of next
714 two most preferred at location 1, and so on with the average of the two least preferred
715 orientations at location 4 (the reason for averaging of pairs is to match [51] as closely
716 as possible). Fitting a line to these points gives a slope and intercept for each feature
717 map (lines are fit using the least squares method). FSGM predicts a negative slope
718 and an intercept greater than one.

719 To test for signs of feature matching behavior, each feature map's preferred (most
720 positive tuning value) and anti-preferred (most negative tuning value) orientations
721 are determined. Activity is recorded when attention is applied to the preferred or

722 anti-preferred orientation and activity ratios are calculated. According to the FSGM,
723 activity when the preferred orientation is attended should be greater than when the
724 anti-preferred is attended, regardless of whether the image is of the preferred or anti-
725 preferred orientation. According to the feature matching (FM) model, however, ac-
726 tivity when attending the presented orientation should be greater than activity when
727 attending an absent orientation, regardless of whether the orientation is preferred or
728 not. Therefore, we say that a feature map is displaying feature matching behavior
729 if (1) activity is greater when attending the preferred orientation when the preferred
730 is present versus when the anti-preferred is present, and (2) activity is greater when
731 attending the anti-preferred orientation when the anti-preferred is present versus when
732 the preferred is present. The second criteria distinguishes feature matching behavior
733 from FSGM.

734 *4.8. Correlating Activity Changes with Performance*

735 We use two different measures of attention-induced activity changes in order to
736 probe the relationship between activity and classification performance. In both cases,
737 the network is performing the orientation detection task described in Figure 5A.

738 The first measure is meant to capture feature similarity gain model-like behavior on
739 an image-by-image basis (the measure illustrated in 6B is calculated over a population
740 of images of different stimuli). Images that contain a given orientation are shown
741 to the network and the spatially-averaged activity of feature maps is recorded when
742 attention is applied to that orientation and when it is not. The ratio of these activities
743 is then plotted against each feature map's tuning value for the orientation. According
744 to the FSGM, this ratio should be above 1 for feature maps with positive tuning values
745 and less than one for those with negative tuning values. Therefore, we use the slope of
746 the line fitted to these ratios plotted as a function of tuning values as an indication of
747 the extent to which activity is FSGM-like (with positive slopes more FSGM-like). The
748 median slope over a set of images of a given orientation is paired with the change in
749 performance on those images with attention. This gives one pair for each combination
750 of orientation, strength, and layer at which attention was applied (activity changes are
751 only recorded if attention was applied at or before the recorded layer). The correlation
752 coefficient between these value pairs is plotted as the dashed line in Figure 7C.

753 The second measure aims to characterize activity in terms of the outcome of the
754 classification, rather than the contents of the input (see Figure 7A for a visualization).
755 First, for a particular orientation, images that both do and do not contain that orien-
756 tation are shown to the network. Activity (spatially-averaged over each feature map)
757 in response to images classified as containing the orientation (i.e., both true and false
758 positives) is averaged in order to construct a vector in activity space that represents
759 positive classification for a given layer. To reduce complications of working with vec-
760 tors in high dimensions, principal components are found that capture at least 90% of
761 the variance of the activity in response to all images, and all computations are done in
762 this lower dimensional space. The next step is to determine if attention moves activity
763 in a given layer closer to this direction of positive classification. For this, images that
764 contain the given orientation (but were not positively-classified without attention) are
765 used. For each image, the cosine of the angle between the positive-classification vector
766 and the activity in response to the image is calculated. The median of these angles
767 over a set of images is calculated separately for when attention is applied and when it
768 is not. The difference between these medians (with-attention minus without-attention)

769 is paired with the change in performance that comes with attention on those images.
770 Then the same correlation calculation is done with these pairs as described above.

771 For activity recorded from the fully-connected layers (14 and 15), each of the indi-
772 vidual units is used in place of spatially-averaged feature map activity.

773 *4.9. Experimental Data*

774 Model results were compared to previously published data coming from several
775 studies. In [50], a category detection task was performed using stereogram stimuli
776 (on object present trials, the object image was presented to one eye and a noise mask
777 to another). The presentation of the visual stimuli was preceded by a verbal cue
778 that indicated the object category that would later be queried (cued trials) or by
779 meaningless noise (uncued trials). After visual stimulus presentation, subjects were
780 asked if an object was present and, if so, if the object was from the cued category
781 (categories were randomized for uncued trials). In Experiment 1 ('Cat-Drawings' in
782 Figure 4B), the object images were line drawings (one per category) and the stimuli
783 were presented for 1.5 sec. In Experiment 2 ('Cat-Images'), the object images were
784 grayscale photographs (multiple per category) and presented for 6 sec (of note: this
785 presumably allows for several rounds of feedback processing, in contrast to our purely
786 feedforward model). True positives were counted as trials wherein a given object
787 category was present and the subject correctly indicated its presence when queried.
788 False positives were trials wherein no category was present and subjects indicated that
789 the queried category was present.

790 In [49], a similar detection task was used. Here, subjects detected the presence of
791 an uppercase letter that (on target present trials) was presented rapidly and followed
792 by a mask. Prior to the visual stimulus, a visual ('Letter-Vis') or audio ('Letter-Aud')
793 cue indicated a target letter. After the visual stimulus, the subjects were required to
794 indicate whether any letter was present. True positives were trials in which a letter was
795 present and the subject indicated it (only uncued trials or validly cued trials—where
796 the cued letter was the letter shown—were considered here). False positives were trials
797 where no letter was present and the subject indicated that one was.

798 The task in [39] was also an object category detection task ('Objects'). Here, an
799 array of several images was flashed on the screen with one image marked as the target.
800 All images were color photographs of objects in natural scenes. In certain blocks,
801 the subjects knew in advance which category they would later be queried about (cued
802 trials). On other trials, the queried category was only revealed after the visual stimulus
803 (uncued). True positives were trials in which the subject indicated the presence of the
804 queried category when it did exist in the target image. False positives were trials in
805 which the subject indicated the presence of the cued category when it was not in the
806 target image. Data from trials using basic category levels with masks were used for
807 this study.

808 Finally, we include one study using macaques ('Ori-Change') wherein both neural
809 and performance changes were measured [53]. In this task, subjects had to report a
810 change in orientation that could occur in one of two stimuli. On cued trials, the change
811 occurred in the cued stimulus in 80% of trials and the uncued stimulus in 20% of tri-
812 als. On neutrally-cued trials, subjects were not given prior information about where
813 the change was likely to occur (50% at each stimulus). Therefore performance could
814 be compared under conditions of low (uncued stimuli), medium (neutrally cued stim-
815 uli), and high (cued stimuli) attention strength. Correct detection of an orientation

816 change in a given stimulus (indicated by a saccade) is considered a true positive and a
817 saccade to the stimulus prior to any orientation change is considered a false positive.
818 True negatives are defined as correct detection of a change in the uncued stimulus
819 (as this means the subject correctly did not perceive a change in the stimulus under
820 consideration) and false negatives correspond to a lack of response to an orientation
821 change. While this task includes a spatial attention component, it is still useful as a
822 test of feature-based attention effects. Previous work has demonstrated that, during a
823 change detection task, feature-based attention is deployed to the pre-change features
824 of a stimulus [15, 54]. Therefore, because the pre-change stimuli are of differing orien-
825 tations, the cueing paradigm used here controls the strength of attention to orientation
826 as well.

827 In cases where the true and false positive rates were not published, they were ob-
828 tained via personal communications with the authors. Not all changes in performance
829 were statistically significant, but we plot them to show general trends.

830 We calculate the activity changes required in the model to achieve the behavioral
831 changes observed experimentally by using the data plotted in Figure 4B. We determine
832 the average β value for the neutral and cued conditions by finding the β value of the
833 point on the model line nearest to the given data point. Specifically, we average the β
834 values found for the four datasets whose experiments are most similar to our merged
835 image task (Cat-Drawings, Cat-Images, Letter-Aud, and Letter-Vis).

836 5. Acknowledgements

837 We are very grateful to the authors who so readily shared details of their behavioral
838 data upon request: J. Patrick Mayo, Gary Lupyan, and Mika Koivisto. We further
839 thank J. Patrick Mayo for helpful comments on the manuscript.

840 6. References

- 841 [1] Ji Won Bang and Dobromir Rahnev. Stimulus expectation alters decision criterion
842 but not sensory signal in perceptual decision making. *Scientific reports*, 7(1):
843 17072, 2017.
- 844 [2] Jalal K Baruni, Brian Lau, and C Daniel Salzman. Reward expectation differ-
845 entially modulates attentional behavior and activity in visual area v4. *Nature*
846 *neuroscience*, 18(11):1656, 2015.
- 847 [3] Narcisse P Bichot, Matthew T Heard, Ellen M DeGennaro, and Robert Desimone.
848 A source for feature-based attention in the prefrontal cortex. *Neuron*, 88(4):832–
849 844, 2015.
- 850 [4] Ali Borji and Laurent Itti. Optimal attentional modulation of a neural population.
851 *Frontiers in computational neuroscience*, 8, 2014.
- 852 [5] Geoffrey M Boynton. A framework for describing the effects of attention on visual
853 responses. *Vision research*, 49(10):1129–1143, 2009.
- 854 [6] David A Bridwell and Ramesh Srinivasan. Distinct attention networks for feature
855 enhancement and suppression in vision. *Psychological science*, 23(10):1151–1158,
856 2012.

- 857 [7] Elizabeth A Buffalo, Pascal Fries, Rogier Landman, Hualou Liang, and Robert
858 Desimone. A backward progression of attentional effects in the ventral stream.
859 *Proceedings of the National Academy of Sciences*, 107(1):361–365, 2010.
- 860 [8] Claus Bundesen. A theory of visual attention. *Psychological review*, 97(4):523,
861 1990.
- 862 [9] Santiago A Cadena, George H Denfield, Edgar Y Walker, Leon A Gatys, An-
863 dreas S Tolia, Matthias Bethge, and Alexander S Ecker. Deep convolutional
864 models improve predictions of macaque v1 responses to natural images. *bioRxiv*,
865 page 201764, 2017.
- 866 [10] Marisa Carrasco. Visual attention: The past 25 years. *Vision research*, 51(13):
867 1484–1525, 2011.
- 868 [11] Kyle R Cave. The featuregate model of visual selection. *Psychological research*,
869 62(2):182–194, 1999.
- 870 [12] Leonardo Chelazzi, John Duncan, Earl K Miller, and Robert Desimone. Responses
871 of neurons in inferior temporal cortex during memory-guided visual search. *Jour-
872 nal of neurophysiology*, 80(6):2918–2940, 1998.
- 873 [13] Sharat Chikkerur, Thomas Serre, Cheston Tan, and Tomaso Poggio. What and
874 where: A bayesian inference theory of attention. *Vision research*, 50(22):2233–
875 2247, 2010.
- 876 [14] Marlene R Cohen and John HR Maunsell. Attention improves performance pri-
877 marily by reducing interneuronal correlations. *Nature neuroscience*, 12(12):1594–
878 1600, 2009.
- 879 [15] Marlene R Cohen and John HR Maunsell. Using neuronal populations to study
880 the mechanisms underlying spatial and feature attention. *Neuron*, 70(6):1192–
881 1204, 2011.
- 882 [16] Trinity B Crapse, Hakwan Lau, and Michele A Basso. A role for the superior
883 colliculus in decision criteria. *Neuron*, 97(1):181–194, 2018.
- 884 [17] Tolga Çukur, Shinji Nishimoto, Alexander G Huth, and Jack L Gallant. Atten-
885 tion during natural vision warps semantic representation across the human brain.
886 *Nature neuroscience*, 16(6):763–770, 2013.
- 887 [18] Gregory C DeAngelis, Bruce G Cumming, and William T Newsome. Cortical
888 area mt and the perception of stereoscopic depth. *Nature*, 394(6694):677, 1998.
- 889 [19] Cathryn J Downing. Expectancy and visual-spatial attention: effects on percep-
890 tual quality. *Journal of Experimental Psychology: Human perception and perfor-
891 mance*, 14(2):188, 1988.
- 892 [20] Miguel P Eckstein, Matthew F Peterson, Binh T Pham, and Jason A Droll.
893 Statistical decision theory to relate neurons to behavior in the study of covert
894 visual attention. *Vision research*, 49(10):1097–1128, 2009.

- 895 [21] Michael Eickenberg, Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion.
896 Seeing it all: Convolutional network layers map the function of the human visual
897 system. *NeuroImage*, 152:184–194, 2017.
- 898 [22] Pascal Fries, John H Reynolds, Alan E Rorie, and Robert Desimone. Modulation
899 of oscillatory neuronal synchronization by selective visual attention. *Science*, 291
900 (5508):1560–1563, 2001.
- 901 [23] Davi Frossard. *VGG in TensorFlow*. <https://www.cs.toronto.edu/frossard/post/vgg16/>
902 Accessed: 2017-03-01.
- 903 [24] Kunihiro Fukushima. Neocognitron: A hierarchical neural network capable of
904 visual pattern recognition. *Neural networks*, 1(2):119–130, 1988.
- 905 [25] Umut Güçlü and Marcel AJ van Gerven. Deep neural networks reveal a gradient
906 in the complexity of neural representations across the ventral stream. *Journal of*
907 *Neuroscience*, 35(27):10005–10014, 2015.
- 908 [26] FH Hamker. The role of feedback connections in task-driven visual search. In
909 *Connectionist models in cognitive neuroscience*, pages 252–261. Springer, 1999.
- 910 [27] Fred H Hamker and James Worcester. Object detection in natural scenes by
911 feedback. In *International Workshop on Biologically Motivated Computer Vision*,
912 pages 398–407. Springer, 2002.
- 913 [28] Harold L Hawkins, Steven A Hillyard, Steven J Luck, Mustapha Mouloua,
914 Cathryn J Downing, and Donald P Woodward. Visual attention modulates sig-
915 nal detectability. *Journal of Experimental Psychology: Human Perception and*
916 *Performance*, 16(4):802, 1990.
- 917 [29] Benjamin Y Hayden and Jack L Gallant. Combined effects of spatial and feature-
918 based attention on responses of v4 neurons. *Vision research*, 49(10):1182–1187,
919 2009.
- 920 [30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning
921 for image recognition. In *Proceedings of the IEEE conference on computer vision*
922 *and pattern recognition*, pages 770–778, 2016.
- 923 [31] Hauke R Heekeren, Sean Marrett, Peter A Bandettini, and Leslie G Ungerleider.
924 A general mechanism for perceptual decision-making in the human brain. *Nature*,
925 431(7010):859–862, 2004.
- 926 [32] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten.
927 Densely connected convolutional networks. In *Proceedings of the IEEE conference*
928 *on computer vision and pattern recognition*, volume 1, page 3, 2017.
- 929 [33] Daniel Kaiser, Nikolaas N Oosterhof, and Marius V Peelen. The neural dynamics
930 of attentional selection in natural scenes. *Journal of neuroscience*, 36(41):10522–
931 10528, 2016.
- 932 [34] Kohitij Kar, Jonas Kubilius, Elias Issa, Kailyn Schmidt, and James DiCarlo.
933 Evidence that feedback is required for object identity inferences computed by the
934 ventral stream. COSYNE, 2017.

- 935 [35] Sabine Kastner and Mark A Pinsk. Visual attention as a multilevel selection
936 process. *Cognitive, Affective, & Behavioral Neuroscience*, 4(4):483–500, 2004.
- 937 [36] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but
938 not unsupervised, models may explain it cortical representation. *PLoS computa-*
939 *tional biology*, 10(11):e1003915, 2014.
- 940 [37] Seyed-Mahdi Khaligh-Razavi, Linda Henriksson, Kendrick Kay, and Nikolaus
941 Kriegeskorte. Fixed versus mixed rsa: Explaining visual representations by fixed
942 and mixed feature sets from shallow and deep computational models. *Journal of*
943 *Mathematical Psychology*, 76:184–197, 2017.
- 944 [38] Saeed Reza Kheradpisheh, Masoud Ghodrati, Mohammad Ganjtabesh, and Tim-
945 othée Masquelier. Deep networks can resemble human feed-forward vision in
946 invariant object recognition. *Scientific reports*, 6:32672, 2016.
- 947 [39] Mika Koivisto and Ella Kahila. Top-down preparation modulates visual catego-
948 rization but not subjective awareness of objects presented in natural backgrounds.
949 *Vision Research*, 133:73–80, 2017.
- 950 [40] Simon Kornblith and Doris Y Tsao. How thoughts arise from sights: inferotem-
951 poral and prefrontal contributions to vision. *Current Opinion in Neurobiology*,
952 46:208–218, 2017.
- 953 [41] Richard J Krauzlis, Lee P Lovejoy, and Alexandre Zénon. Superior colliculus and
954 visual spatial attention. *Annual review of neuroscience*, 36:165–182, 2013.
- 955 [42] Jonas Kubilius, Stefania Bracci, and Hans P Op de Beeck. Deep neural networks
956 as a computational model for human shape sensitivity. *PLoS computational biol-*
957 *ogy*, 12(4):e1004896, 2016.
- 958 [43] Timothy P Lillicrap, Daniel Cownden, Douglas B Tweed, and Colin J Akerman.
959 Random synaptic feedback weights support error backpropagation for deep learn-
960 ing. *Nature communications*, 7, 2016.
- 961 [44] Grace W Lindsay. Feature-based attention in convolutional neural networks. *arXiv*
962 *preprint arXiv:1511.06408*, 2015.
- 963 [45] Grace W Lindsay, Dan B Rubin, and Kenneth D Miller. The stabilized supralinear
964 network replicates neural and performance correlates of attention. *COSYNE*,
965 2017.
- 966 [46] Bradley C Love, Olivia Guest, Piotr Slomka, Victor M Navarro, and Edward
967 Wasserman. Deep networks as models of human and animal categorization. In
968 *CogSci*, 2017.
- 969 [47] Steven J Luck, Leonardo Chelazzi, Steven A Hillyard, and Robert Desimone. Neu-
970 ral mechanisms of spatial selective attention in areas v1, v2, and v4 of macaque
971 visual cortex. *Journal of neurophysiology*, 77(1):24–42, 1997.
- 972 [48] Thomas Zhihao Luo and John HR Maunsell. Neuronal modulations in visual
973 cortex are associated with only one of multiple components of attention. *Neuron*,
974 86(5):1182–1188, 2015.

- 975 [49] Gary Lupyan and Michael J Spivey. Making the invisible visible: Verbal but not
976 visual cues enhance visual detection. *PLoS One*, 5(7):e11452, 2010.
- 977 [50] Gary Lupyan and Emily J Ward. Language can boost otherwise unseen objects
978 into visual awareness. *Proceedings of the National Academy of Sciences*, 110(35):
979 14196–14201, 2013.
- 980 [51] Julio C Martinez-Trujillo and Stefan Treue. Feature-based attention increases the
981 selectivity of population responses in primate visual cortex. *Current Biology*, 14
982 (9):744–751, 2004.
- 983 [52] John HR Maunsell and Erik P Cook. The role of attention in visual processing.
984 *Philosophical Transactions of the Royal Society of London B: Biological Sciences*,
985 357(1424):1063–1072, 2002.
- 986 [53] J Patrick Mayo and John HR Maunsell. Graded neuronal modulations related to
987 visual spatial attention. *Journal of Neuroscience*, 36(19):5353–5361, 2016.
- 988 [54] J Patrick Mayo, Marlene R Cohen, and John HR Maunsell. A refined neuronal
989 population measure of visual attention. *PloS one*, 10(8):e0136570, 2015.
- 990 [55] Carrie J McAdams and John HR Maunsell. Effects of attention on orientation-
991 tuning functions of single neurons in macaque cortical area v4. *Journal of Neu-*
992 *roscience*, 19(1):431–441, 1999.
- 993 [56] Sebastian Moeller, Trinity Crapse, Le Chang, and Doris Y Tsao. The effect of face
994 patch microstimulation on perception of faces and objects. *Nature Neuroscience*,
995 20(5):743–752, 2017.
- 996 [57] Ilya E Monosov, David L Sheinberg, and Kirk G Thompson. The effects of pre-
997 frontal cortex inactivation on object responses of single neurons in the inferotem-
998 poral cortex during visual search. *Journal of Neuroscience*, 31(44):15956–15961,
999 2011.
- 1000 [58] Tirin Moore and Katherine M Armstrong. Selective gating of visual signals by
1001 microstimulation of frontal cortex. *Nature*, 421(6921):370, 2003.
- 1002 [59] Ari S Morcos, David GT Barrett, Neil C Rabinowitz, and Matthew Botvinick.
1003 On the importance of single directions for generalization. *arXiv preprint*
1004 *arXiv:1803.06959*, 2018.
- 1005 [60] Sancho I Moro, Michiel Tolboom, Paul S Khayat, and Pieter R Roelfsema. Neu-
1006 ron activity in the visual cortex reveals the temporal order of cognitive opera-
1007 tions. *Journal of Neuroscience*, 30(48):16293–16303, 2010.
- 1008 [61] Brad C Motter. Neural correlates of feature selective memory and pop-out in
1009 extrastriate area v4. *Journal of Neuroscience*, 14(4):2190–2199, 1994.
- 1010 [62] Vidhya Navalpakkam and Laurent Itti. Search goal tunes visual features optimally.
1011 *Neuron*, 53(4):605–617, 2007.

- 1012 [63] Marino Pagan, Luke S Urban, Margot P Wohl, and Nicole C Rust. Signals
1013 in inferotemporal and perirhinal cortex suggest an untangling of visual target
1014 information. *Nature neuroscience*, 16(8):1132–1139, 2013.
- 1015 [64] William K Page and Charles J Duffy. Cortical neuronal responses to optic flow
1016 are shaped by visual strategies for steering. *Cerebral cortex*, 18(4):727–739, 2007.
- 1017 [65] Marius V Peelen and Sabine Kastner. A neural basis for real-world visual search in
1018 human occipitotemporal cortex. *Proceedings of the National Academy of Sciences*,
1019 108(29):12125–12130, 2011.
- 1020 [66] Marius V Peelen, Li Fei-Fei, and Sabine Kastner. Neural mechanisms of rapid
1021 natural scene categorization in human visual cortex. *Nature*, 460(7251):94, 2009.
- 1022 [67] Gopathy Purushothaman and David C Bradley. Neural population code for fine
1023 perceptual decisions in area mt. *Nature neuroscience*, 8(1):99, 2005.
- 1024 [68] Dobromir Rahnev, Hakwan Lau, and Floris P de Lange. Prior expectation mod-
1025 ulates the interaction between sensory and prefrontal regions in the human brain.
1026 *Journal of Neuroscience*, 31(29):10741–10748, 2011.
- 1027 [69] Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image
1028 classification: A comprehensive review. *Neural Computation*, 2017.
- 1029 [70] Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recog-
1030 nition in cortex. *Nature neuroscience*, 2(11), 1999.
- 1031 [71] Edmund T Rolls and Gustavo Deco. Attention in natural scenes: neurophysio-
1032 logical and computational bases. *Neural networks*, 19(9):1383–1394, 2006.
- 1033 [72] Douglas A Ruff and Richard T Born. Feature attention for binocular disparity in
1034 primate area mt depends on tuning strength. *Journal of neurophysiology*, 113(5):
1035 1545–1555, 2015.
- 1036 [73] Melissa Saenz, Giedrius T Buracas, and Geoffrey M Boynton. Global effects of
1037 feature-based attention in human visual cortex. *Nature neuroscience*, 5(7):631,
1038 2002.
- 1039 [74] Melissa Saenz, Giedrius T Buraças, and Geoffrey M Boynton. Global feature-
1040 based attention for motion and color. *Vision research*, 43(6):629–637, 2003.
- 1041 [75] C Daniel Salzman, Kenneth H Britten, and William T Newsome. Cortical mi-
1042 crostimulation influences perceptual judgements of motion direction. *Nature*, 346
1043 (6280):174–177, 1990.
- 1044 [76] K Seeliger, M Fritsche, U Güçlü, S Schoenmakers, J-M Schoffelen, SE Bosch, and
1045 MAJ van Gerven. Cnn-based encoding and decoding of visual object recognition
1046 in space and time. *bioRxiv*, page 118091, 2017.
- 1047 [77] John T Serences, Jens Schwarzbach, Susan M Courtney, Xavier Golay, and Steven
1048 Yantis. Control of object-based attention in human cortex. *Cerebral Cortex*, 14
1049 (12):1346–1357, 2004.

- 1050 [78] Thomas Serre, Lior Wolf, Stanley Bileschi, Maximilian Riesenhuber, and Tomaso
1051 Poggio. Robust object recognition with cortex-like mechanisms. *IEEE transac-*
1052 *tions on pattern analysis and machine intelligence*, 29(3):411–426, 2007.
- 1053 [79] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for
1054 large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- 1055 [80] Devarajan Sridharan, Nicholas A Steinmetz, Tirin Moore, and Eric I Knudsen.
1056 Does the superior colliculus control perceptual sensitivity or choice bias during
1057 attention? evidence from a multialternative decision framework. *Journal of Neu-*
1058 *roscience*, 37(3):480–511, 2017.
- 1059 [81] Timo Stein and Marius V Peelen. Object detection in natural scenes: Indepen-
1060 dent effects of spatial and category-based attention. *Attention, Perception, &*
1061 *Psychophysics*, 79(3):738–752, 2017.
- 1062 [82] Anne M Treisman and Garry Gelade. A feature-integration theory of attention.
1063 *Cognitive psychology*, 12(1):97–136, 1980.
- 1064 [83] Stefan Treue. Neural correlates of attention in primate visual cortex. *Trends in*
1065 *neurosciences*, 24(5):295–300, 2001.
- 1066 [84] Stefan Treue and Julio C Martinez Trujillo. Feature-based attention influences
1067 motion processing gain in macaque visual cortex. *Nature*, 399(6736):575, 1999.
- 1068 [85] Bryan P Tripp. Similarities and differences between stimulus tuning in the
1069 inferotemporal visual cortex and convolutional networks. In *Neural Networks*
1070 *(IJCNN), 2017 International Joint Conference on*, pages 3551–3560. IEEE, 2017.
- 1071 [86] John K Tsotsos, Scan M Culhane, Winky Yan Kei Wai, Yuzhong Lai, Neal Davis,
1072 and Fernando Nuflo. Modeling visual attention via selective tuning. *Artificial*
1073 *intelligence*, 78(1-2):507–545, 1995.
- 1074 [87] Leslie G Ungerleider, Thelma W Galkin, Robert Desimone, and Ricardo Gattass.
1075 Cortical connections of area v4 in the macaque. *Cerebral Cortex*, 18(3):477–499,
1076 2007.
- 1077 [88] Preeti Verghese. Visual search and attention: A signal detection theory approach.
1078 *Neuron*, 31(4):523–535, 2001.
- 1079 [89] Louise Whiteley and Maneesh Sahani. Attention in a bayesian framework. *Fron-*
1080 *tiers in human neuroscience*, 6, 2012.
- 1081 [90] Jeremy M Wolfe. Guided search 2.0 a revised model of visual search. *Psychonomic*
1082 *bulletin & review*, 1(2):202–238, 1994.
- 1083 [91] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan
1084 Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural
1085 image caption generation with visual attention. In *International Conference on*
1086 *Machine Learning*, pages 2048–2057, 2015.

- 1087 [92] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seib-
1088 ert, and James J DiCarlo. Performance-optimized hierarchical models predict
1089 neural responses in higher visual cortex. *Proceedings of the National Academy of*
1090 *Sciences*, 111(23):8619–8624, 2014.
- 1091 [93] Adam Zaidel, Gregory C DeAngelis, and Dora E Angelaki. Decoupled choice-
1092 driven and stimulus-related activity in parietal neurons may be misrepresented
1093 by choice probabilities. *Nature Communications*, 8, 2017.
- 1094 [94] Weiwei Zhang and Steven J Luck. Feature-based attention modulates feedforward
1095 visual processing. *Nature neuroscience*, 12(1):24–25, 2009.
- 1096 [95] Ying Zhang, Ethan M Meyers, Narcisse P Bichot, Thomas Serre, Tomaso A Pog-
1097 gio, and Robert Desimone. Object decoding with attention in inferior temporal
1098 cortex. *Proceedings of the National Academy of Sciences*, 108(21):8850–8855,
1099 2011.
- 1100 [96] Huihui Zhou and Robert Desimone. Feature-based attention in the frontal eye
1101 field and area v4 during visual search. *Neuron*, 70(6):1205–1217, 2011.