Visual crowding in ANNs

Taekjun Kim

Abstract

I evaluated visual crowding, the inability to recognize an object in cluttered visual scene, in three architectures of artificial neural networks (ANNs) – AlexNet, VGG16, ResNet50. I manipulated stimulus parameters including the number of distractors, target-distractor distance, and target saliency to assess how feature selectivity of individual units in convolutional layers is systematically affected by those parameters. Results show that all tested ANNs suffer from visual crowding in a manner similar to what reported in humans. The feature selectivity was gradually degraded when i) more distractors were added, ii) target-distractor distance decreased, and iii) target was similar to distractors. And these visual crowding effects tended to be stronger in deeper layers as receptive field sizes get bigger. Several examples in which preferred feature of a unit is visualized using guided back propagation technique are also presented. This work will help to form a basis for the research for studying neural mechanisms of visual crowding.

Introduction

Visual crowding refers to the phenomenon in which an object (i.e., target) that is easily identified when viewed in isolation becomes unrecognizable when surrounded by other stimuli (i.e., distractors). Researchers have proposed that this phenomenon may reflect the processes of how our brain integrates the information provided by multiple sources over space to create a unified representation of the visual scene (Pelli et al., 2004; Van Den Berg et al., 2010). Through extensive psychophysical studies of visual crowding, we now understand how the strength of the crowding effect is influenced by the parameters of visual display (Whitney and Levi, 2011). For example, critical spacing, which is the minimum distance needed between a target and distractors to avoid crowding is proportional to the target eccentricity. At a given distance, crowding is reduced when target and flankers are dissimilar to each other. However, very few neurophysiological studies have been conducted to manipulate these factors, and our knowledge about the neuronal mechanisms of crowding still remains limited.

In this project, I examined how outputs from intermediate convolution layers of artificial neural networks (ANNs) are affected by various target-distractor relationships. In recent years, ANNs for image classification are considered as the best image computable models (i.e., generate responses for arbitrary input images) for neurons in the ventral visual pathway (Yamins et al., 2014; Pospisil et al., 2018). However, ANNs are also different from biological neural networks. Most ANNs are based on feedforward-only connections. Therefore, it needs to be tested if visual crowding can be achieved even in the absence of feedback and local recurrent connections. This project will be used to gain insight into brain function and to make a better model of biological object recognition.

Methods

Visual stimuli

One of the key characteristics of visual crowding is that the presence of distractors impairs target identification, but not target detection (Whitney and Levi, 2011). Therefore, it is more important to determine how feature selectivity of a unit is changed by distractors rather than to see whether the magnitude of unit response to a target is simply modulated by distractors. To measure feature selectivity, I used 100 animal icons (H: 32 pix, V: 32 pix) for target stimuli to stimulate the center region of the receptive field. Each icon is differently characterized by contour and color (Figure 1). The distractors were randomly selected from the same set of 100 animal icons. We expected to see feature selectivity could be systematically varied by controlling the distractor parameters. We considered three independent variables: i) the number of distractors (e.g., 2, 4, or 6), ii) target-distractor distance (e.g., near, middle, far), iii) target saliency (e.g., color target surrounded by non-color distractors).

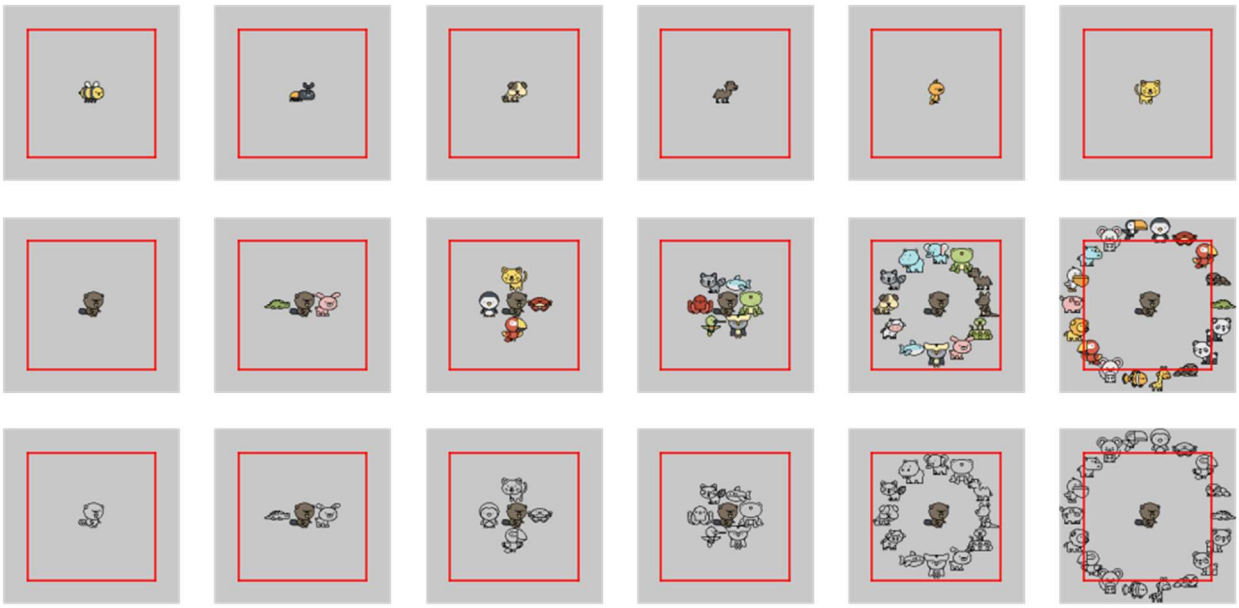


Figure 1. Example visual stimuli. The first row shows several examples of target alone condition. Targets (32 x 32 pix) were always placed at the center of the 224 x 224 pix visual field (i.e., gray background). The second row shows controls of 'distractor number (columns 2-4)' and 'target-distractor distance (columns 4-6)'. The third row has the same distractor configuration as the second row, but target is more salient because distractors do not have any colors. The size of red box is 163 x 163 pix, which is the receptive size of conv5 layer in AlexNet.

Tested ANNs

I looked into intermediate convolutional layers in pre-trained AlexNet, VGG16, and ResNet50 provided by PyTorch (https://pytorch.org/docs/stable/torchvision/models.html). The response a unit in convolutional layers only depends on the image within the receptive field. To ensure that the receptive fields of the tested units can cover at least a portion of the nearest distractors as well as the entire target stimulus, units with receptive fields smaller than 50 pix were excluded from the analysis. The table below shows the theoretical receptive field sizes for units in different layers in three ANNs used in this project. In case of AlexNet and VGG16, even the RF size of the top convolutional layer is smaller than the full visual field, but units in ResNet50 can cover much wider areas.

Table 1. Receptive field sizes of three ANNs

AlexNet		VGG16		ResNet50	
Layer	RF size	Layer	RF size	Layer	RF size
Conv2	51	Conv8	60	L2-2	59
Conv3	99	Conv9	76	L2-3	75
Conv4	131	Conv10	92	L2-4	91
Conv5	163	Conv11	132	L3-1	107
		Conv12	164	L3-2	139
		Conv13	196	L3-3	171
				L3-4	198
				L3-5	235
				L3-6	267
				L4-1	299
				L4-2	363
				L4-3	427

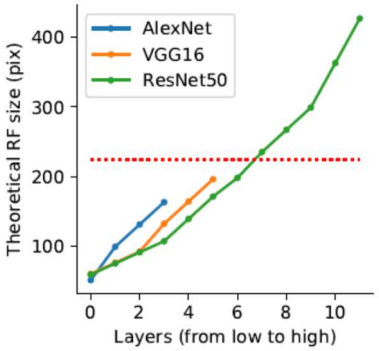


Figure 2. Theoretical RF sizes of three ANNs: AlexNet, VGG16, ResNet50. Red dotted line represents the size of full visual field (224 pixel)

Analysis

I first evaluated unit responses to 100 target stimuli under target alone condition, then excluded units that responded to less than 10 stimuli from the analysis. I used Pearson *r* correlation coefficient as the similarity measure of the feature selectivity between "target alone" condition and "target+distractor" "conditions. To measure the effects of distractors on response magnitude and dispersion, I also computed mean and standard deviation values for 100 target stimuli responses, then compared these values in target alone condition to those in target+distractor conditions,

Lastly, I used natural images from Caltech256 and COCO datasets and the guided backpropagation technique to visualize the most preferred and non-preferred features of representative units (Springenberg et al., 2015).

Results

Feature selective units across layers

I defined units that respond to 10 or more of the 100 target stimuli as feature selective units and used them for subsequent analyses. Figure 3 shows how the numbers of feature selective units differ depending on layers and networks. ResNet50 is composed of a much larger number of layers and units compared to the other two ANNs. One unexpected finding is that the proportion of feature selective units in ResNet50 is also much higher than AlexNet, VGG16. In AlexNet, the proportion of available units tends to decrease as RF size gets bigger. However, this tendency is less clear in VGG16 and ResNet50.

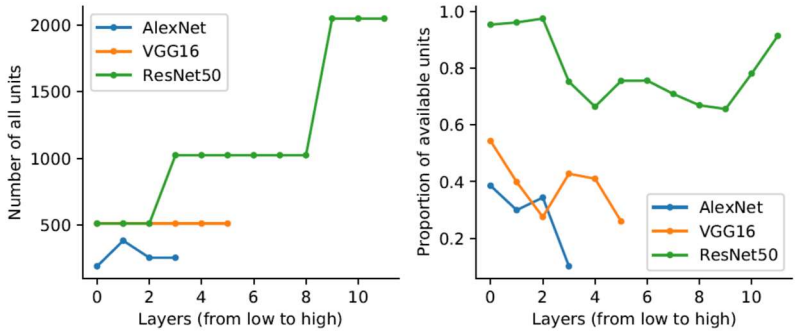


Figure 3. Feature selective units in three ANNs. Left: the numbers of all units in each layer of ANNs. Right: proportion of feature selective units in each layer of ANNs. In a given layer, 1.0 indicates that all units are feature selective (i.e., unit responds to >10 stimuli).

Deterioration of feature selectivity: number of distractors

If the perception of ANNs, like human perception, suffers from visual crowding in a cluttered scene situation, feature selectivity in "target+distractor" condition will be impaired compared to "target alone" condition. And the deterioration of feature selectivity will lead to a low correlation coefficient between "target alone" responses and "target+distractor" responses. I first evaluated how the correlation coefficient (i.e., similarity of feature selectivity) changed as the number of distractors increased.

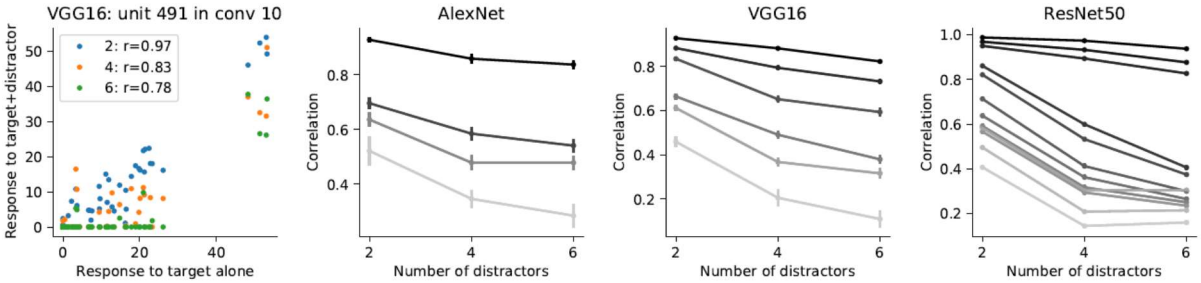


Figure 4. Distractor number effects on feature selectivity. Left: Responses to 'target alone' and 'target + distractor' conditions were compared in an example unit in VGG16. As more distractors were additionally presented, correlation between 'target alone' and 'target + distractor' conditions decreased. Right: Distractor number effects on the correlation between 'target alone' and 'target+distractor' conditions were compared across layers and networks. The brighter color represents the higher layer. In all networks, the correlation coefficient monotonically decreased as the number of distractors increased, and as layer went higher.

Deterioration of feature selectivity: target-distractor distance

Psychophysical experiments have demonstrated that crowding is strong when the distractors are nearest to the target and the effect is getting weaker as target-distractor distance increases (Toet and Levi, 1992). Here, I assessed whether the feature selectivity of ANNs is also systematically affected by target-distractor distance.

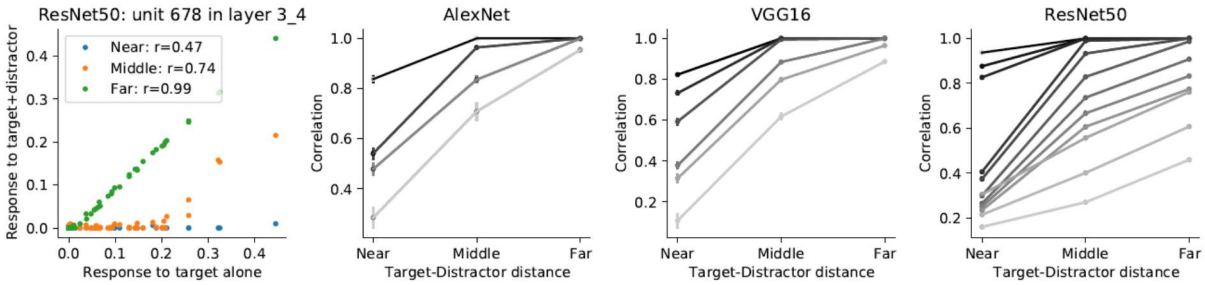


Figure 5. Target-Distractor distance effects on feature selectivity. Left: Responses to 'target alone' and 'target + distractor' conditions were compared in an example unit in ResNet50. Correlation between 'target alone' and 'target + distractor' conditions increased with the target-distractor distance. Right: Distractor distance effects on the correlation between 'target alone' and 'target+distractor' conditions were compared across layers and networks. The brighter color represents the higher layer. In all networks, the correlation coefficient gradually decreased as the target-distractor distance got closer, and as layer went higher.

Deterioration of feature selectivity: target saliency

Crowding is also known to depend on the similarity between the target and distractors (Bernard and Chung, 2011). When target and distractors are similar, they are more likely to be grouped together and produce stronger crowding. To test whether the processing of visually salient target is less affected by distractors in ANNs, I compared "the condition where the color target is surrounded by color distractors" with "the condition where color target is surrounded by line distractors".

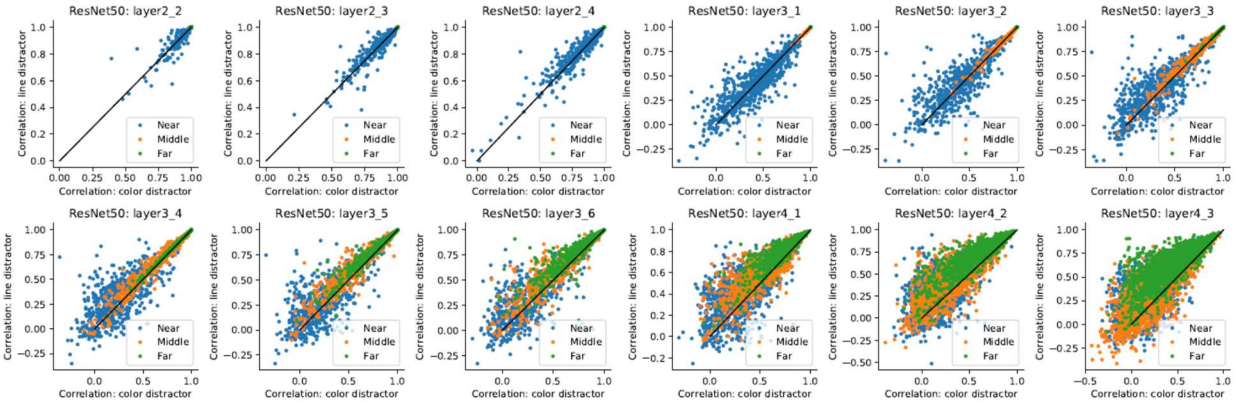


Figure 6. Target saliency effects on feature selectivity in ResNet50. X axis values indicate correlation between 'color target alone' and 'color target+color distractor' conditions, and y axis values indicate correlation between 'color target alone' and 'color target+line distractor' conditions. Blue, orange, green dots represent near, middle, far distractor conditions, respectively. It is observed that the higher the layer, the more data points lies above the diagonal line. This means that identification of crowded target is better when the target and its distractors are dissimilar.

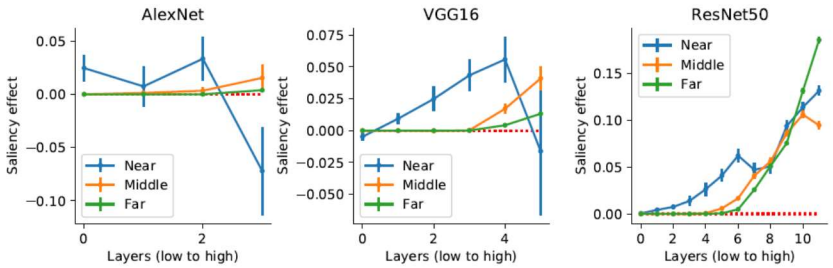


Figure 7. Target saliency effects on feature selectivity in 3 ANNs. Saliency effect was quantified by subtracting correlation in color distractor condition (x axis value in Figure 5) from that in line distractor condition (y axis value in Figure 5). Blue, orange, green dots represent near, middle, far distractor conditions, respectively. Saliency effects seemed to occur in deeper layers in deeper neural networks. In ResNet50 and VGG16, the saliency effect tended to get stronger in higher layers. But AlexNet didn't show a clear effect.

Role of color in feature selectivity

Each icon in target stimulus set is differently characterized by contour and color (Figure 1). To test the role of color in feature selectivity in individual units, I computed the correlation coefficient between 'color target' responses and 'line target' responses. If a unit is purely processing shapes regardless of color information, high correlation coefficient will be obtained. On the other hand, if color is an important feature for a unit, correlation coefficient will be low.

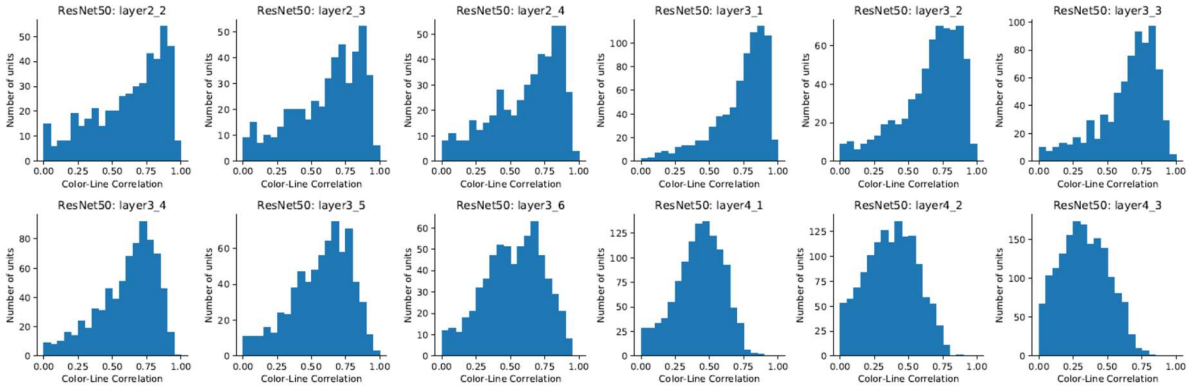


Figure 8. Histograms of color target vs. line target correlation for layers in ResNet50. In low layers, many units showed strong correlation between responses to color targets and responses to line targets, suggesting that processing of contour is independent of color. However, in higher layers, the correlation monotonically decreased. It may be due to the increased role of color information in feature selectivity.

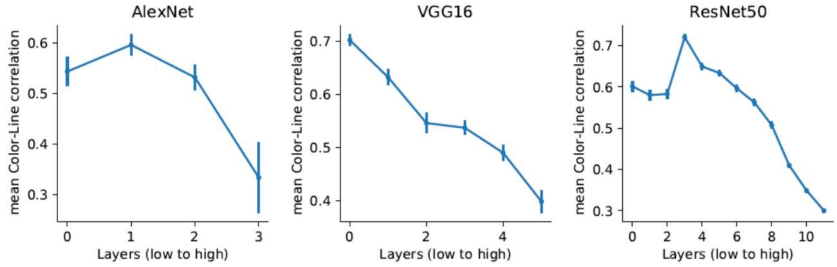


Figure 9. Color target vs. line target correlation across layers in three ANNs. Each data point represents mean and standard error of mean computed from the correlation histogram of each layer. Consistent with result from Figure 7, color target vs. line target correlation gradually decreased as layer got higher in all of three tested ANNs.

Distractor effect on response magnitude

Neurons in visual cortex respond best to local image features that fall within their classical receptive fields (CRFs). Stimulation outside the CRF cannot independently activate the neuron, but it can modulate (suppress in many cases) the neuronal output to stimuli inside the CRF (Cavanaugh et al., 2002). Surround suppression and visual crowding have similarities in terms of the change in response to center stimulus caused by surround stimuli. This suggests that a common mechanism may contribute to both phenomena. Here, I examined whether surround suppression is also observable in ANNs, and if so, whether its strength is associated with a decrease in feature selectivity.

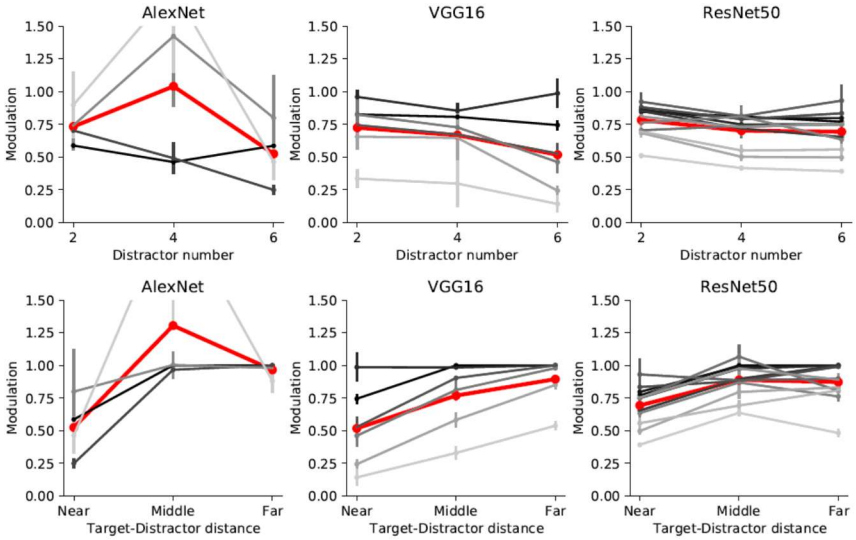
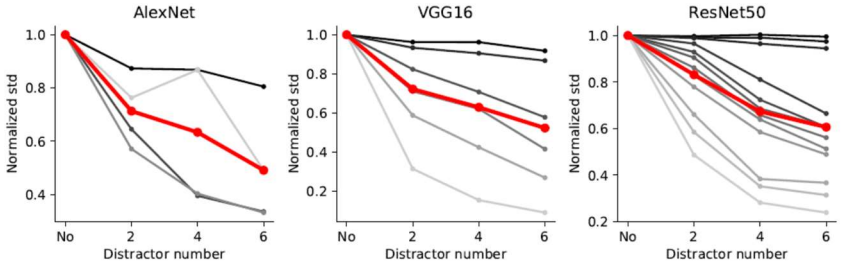


Figure 10. Surround modulation in three ANNs. Top: Effect of distractor number on response magnitude. Bottom: Effect of target-distractor distance on response magnitude. The brighter color represents the higher layer. In almost all conditions, responses to target stimuli were suppressed by surrounding distractors (i.e., modulation < 1.0). However, unlike the deterioration of feature selectivity (Figure 4 & 5), modulation strength was not systematically varied depending on layers, distractor numbers, and target-distractor distance.

Distractor effect on response dispersion

A unit that shows strong feature selectivity for the target stimulus set will exhibit a large dynamic range in its responses. On the other hand, if a neuron has a weak feature selectivity, it will exhibit similar responses across the stimulus set, thus the dispersion (e.g., standard deviation) of responses will be small. I analyzed whether standard deviation of responses to the target stimulus set is gradually decreased as more distractors are added or target-distractor distance gets closer.



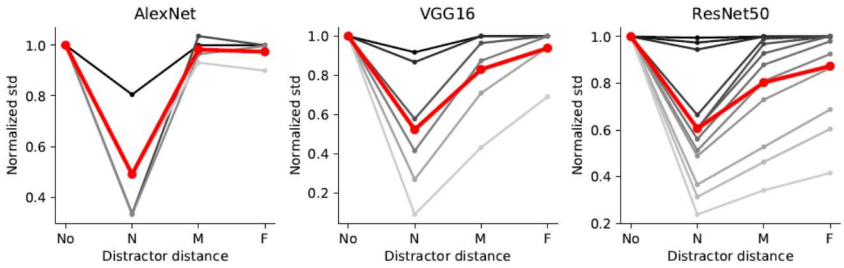


Figure 11. Distractor effects on response dispersion in three ANNs. Top: Effect of distractor number on response dispersion. Bottom: Effect of target-distractor distance on response dispersion. The brighter color represents the higher layer. Degree of response dispersion was quantified with standard deviation, then normalized to the value in 'target alone' condition (1.0). Surrounding distractors systematically reduced the dispersion of the responses to the target stimulus set. The magnitude of the reduction increased as the distractor number increased and the target-distractor distance got closer. This tendency was identical to that observed in the effect of distractors on correlation (Figure 4-5).

Visualization: unit with a weak visual crowding

This example unit showed a strong correlation between "target alone" responses and "target+distractor" responses, that is, a weak visual crowding effect. Feature visualization revealed the preference for tiny circular shape near the center of the receptive field.

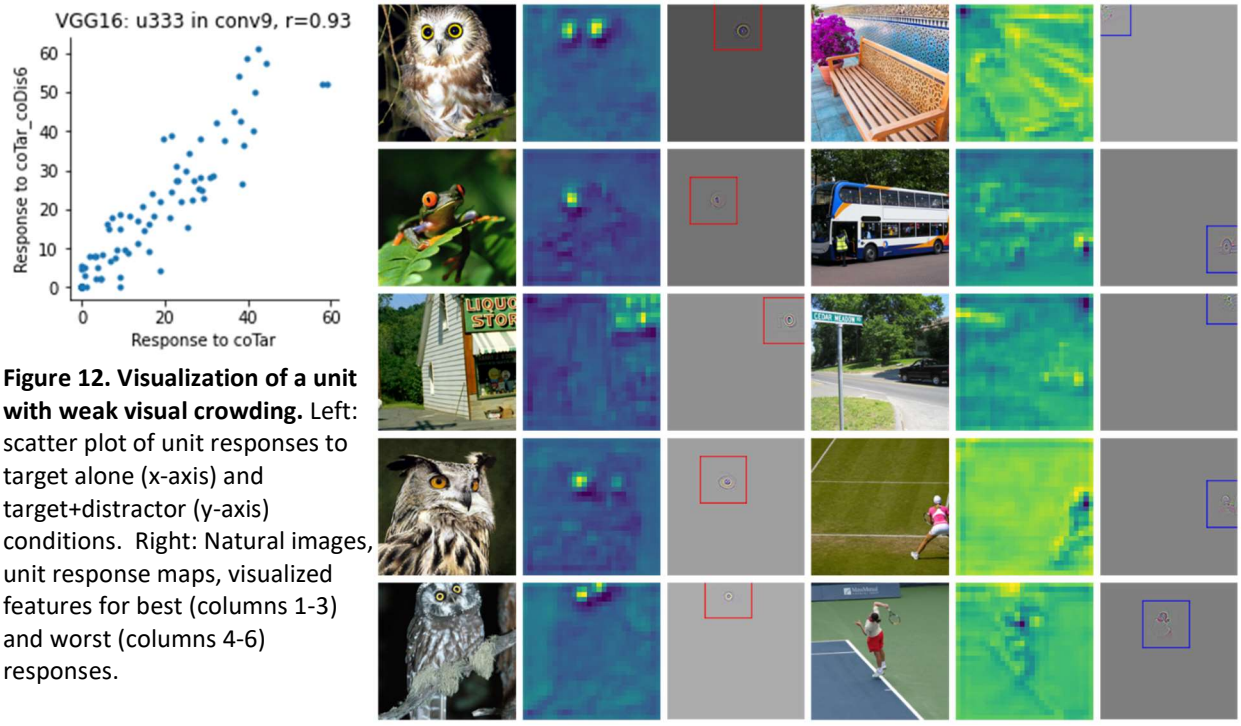


Figure 12. Visualization of a unit with weak visual crowding. Left: scatter plot of unit responses to target alone (x-axis) and target+distractor (y-axis) conditions. Right: Natural images, unit response maps, visualized features for best (columns 1-3) and worst (columns 4-6) responses.

Visualization: unit with a strong visual crowding

This example unit showed a very weak correlation between “target alone” responses and “target+distractor” responses, that is, a strong visual crowding effect. Feature visualization revealed that best natural images share some yellow stuff, and that worst natural images all include regular circular dot texture pattern.

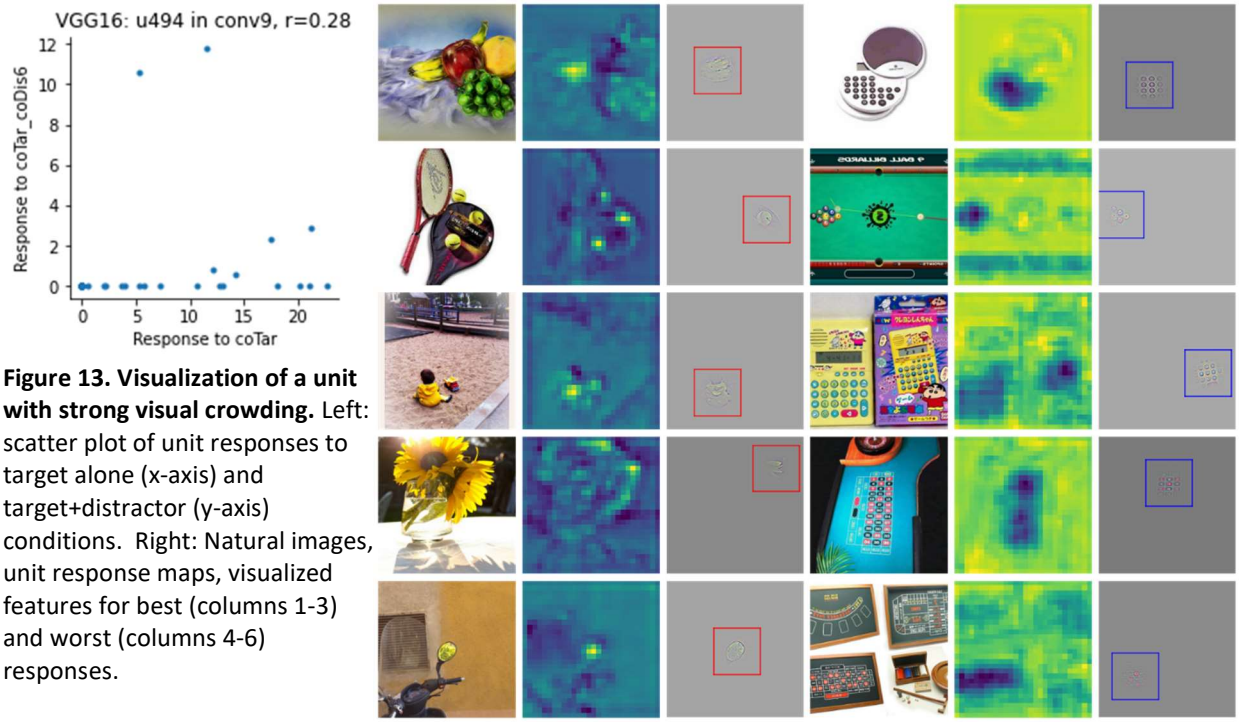


Figure 13. Visualization of a unit with strong visual crowding. Left: scatter plot of unit responses to target alone (x-axis) and target+distractor (y-axis) conditions. Right: Natural images, unit response maps, visualized features for best (columns 1-3) and worst (columns 4-6) responses.

Visualization: unit with a contour (rather than color) processing

This example unit showed strong correlation between "color target" responses and "line target" responses suggesting that it may be processing contour rather than color information. Feature visualization revealed similar contour feature (i.e., convex upward curve) from best natural scene images.

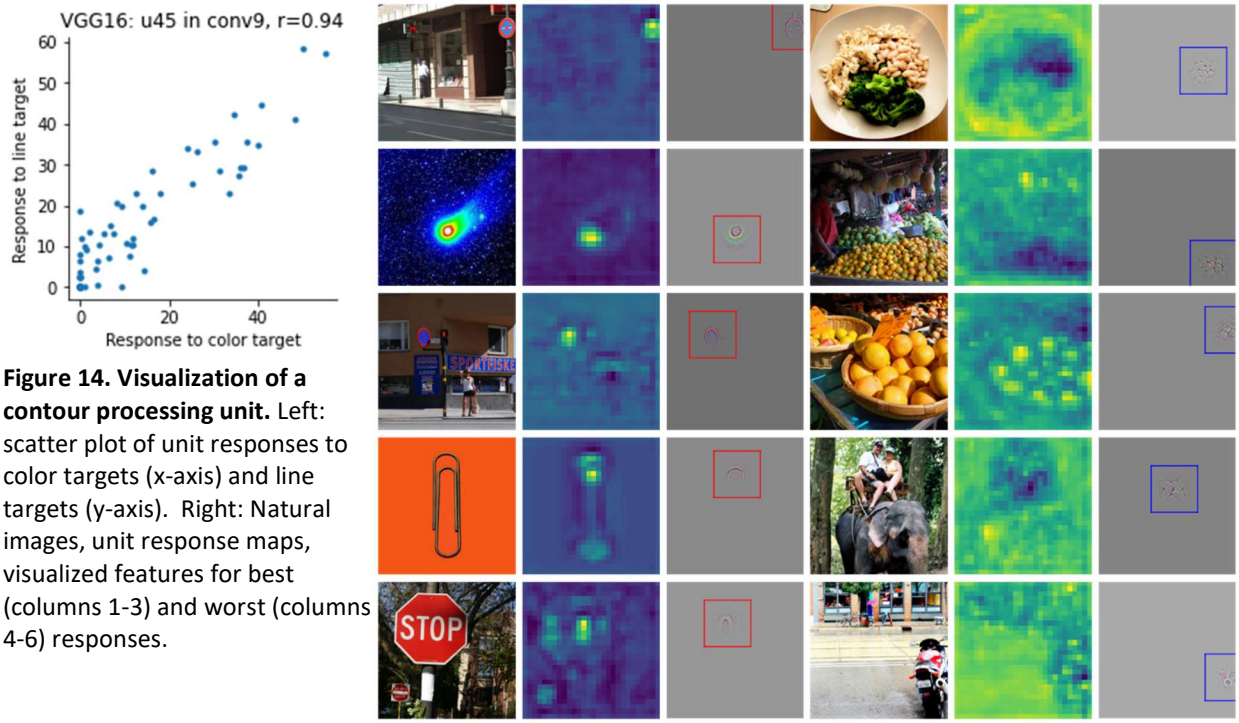


Figure 14. Visualization of a contour processing unit. Left: scatter plot of unit responses to color targets (x-axis) and line targets (y-axis). Right: Natural images, unit response maps, visualized features for best (columns 1-3) and worst (columns 4-6) responses.

Discussion

Early theories about visual crowding explain it using local mechanisms, where degraded target processing is due to the inappropriate pooling of target and surround elements. However, more recent psychophysical studies suggest that visual crowding may involve complex global and feedback processing, because crowding effects can be significantly reduced when multiple distractors form a perceptual pattern that is excludes the target.

Here, I revealed that ANNs, which lack feedback connections, suffer from visual crowding. I found the three following factors influenced the amount of crowding: the number of distractors, target-distractor distance, and target saliency. However, these results do not necessarily mean that visual crowding can occur without any feedback connections in biological neural networks. Modern ANNs often employ several tens to hundreds hierarchical layers between input and output to improve the classification performance, whereas the primate brain can achieve high level performance with a much shallower hierarchy. This may be largely due to massive recurrent and feedback connections, which are dominant in the real brain, but very limited in ANNs. Indeed, any recurrent model can be unfolded into a mathematically equivalent very deep network (Hornik et al., 1989). In Figure 7, the target saliency effect tended to get stronger in higher layers in ResNet50 and VGG16. But this effect was not observable in

AlexNet. This may mean that AlexNet does not have a sufficient number of layers to implement the role of feedback connections with only feedforward connections.

Future work may need to consider more advanced ANNs with recurrent connections and examine whether the global effect that the perceptual grouping of distractors reduces visual crowding can be observed even at low level convolutional layers. Furthermore, through the process of directly comparing the single unit responses of ANNs and the primate brain, we will be able to better understand the limitations of the existing neural network model and to design an improved model.

Discussion

Bernard JB, Chung STL (2011) The dependence of crowding on flanker complexity and target-flanker similarity. J Vis 11:1–16.

Cavanaugh JR, Bair W, Anthony Movshon J (2002) Nature and interaction of signals from the receptive field center and surround in macaque V1 neurons. J Neurophysiol 88:2530–2546.

Hornik K, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal approximators. Neural Networks 2:359–366.

Pelli DG, Palomares M, Majaj NJ (2004) Crowding is unlike ordinary masking: Distinguishing feature integration from detection. J Vis 4:1136–1169.

Pospisil DA, Pasupathy A, Bair W (2018) 'Artiphysiology' reveals V4-like shape tuning in a deep network trained for image classification. Elife 7:e38242.

Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M (2015) Striving for simplicity: The all convolutional net. In: 3rd International Conference on Learning Representations, ICLR 2015 - Workshop Track Proceedings.

Toet A, Levi DM (1992) The two-dimensional shape of spatial interaction zones in the parafovea. Vision Res 32:1349–1357.

Van Den Berg R, Roerdink JBTM, Cornelissen FW (2010) A neurophysiologically plausible population code model for feature integration explains visual crowding. PLoS Comput Biol.

Whitney D, Levi DM (2011) Visual crowding: A fundamental limit on conscious perception and object recognition. Trends Cogn Sci:160–168.

Yamins DLK, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc Natl Acad Sci U S A 111:8619–8624.