Effect of Crowding in Deep Neural Networks

Aathira Manoj (am10245)

Andrei Kapustin (ak7671)

Jasper Duan (zd793)

Abstract

Crowding, the inability to recognize objects in clutter which could be easily recognized in isolation, is a widespread phenomena in humans, limiting object recognition and reading (Whitney & Levi, 2011). It sets a fundamental limit on conscious visual perception and object recognition. In this paper, we study the effect of crowding in different deep convolutional neural networks (DCNNs). In particular, we study 3 different neural network architectures: a simple deep convolutional neural network, eccentricity dependent network (Volokitin, Roig, & Poggio, 2017) and VOneNet (Dapello et al., 2020). Both the eccentricity model and VOneNet are inspired by human and primate brains. In addition to this, we collect data from humans and compare their performance with that of the model in the presence of crowding with different configurations of flankers.

Introduction

Crowding is a well known effect in human vision, in which targets that can be recognized in isolation can no longer be recognized in the presence of nearby objects (flankers), even though there is no occlusion. Crowding impairs the ability to recognize objects in clutter. It has been studied extensively many years and has important implications for patients with macular degeneration, amblyopia and dyslexia. It sets limits on object perception, eye and hand movements, visual search, reading and perhaps other functions in peripheral, amblyopic and developing vision. Crowding is neither masking nor surround suppression. It is possible to localize crowding to the V1 cortex; however, there is a growing consensus for a two-stage model of crowding in which the first stage involves the detection of simple features (perhaps in V1), and a second stage is required for the integration or interpretation of the features as an object beyond V1. (Levi, 2008).

In human studies conducted to study crowding, the subjects are asked to fixate at a cross at the center of a screen, and objects are presented at the periphery of their visual field in a flash such that the subject has no time to move their eyes. Based on such studies, factors such as the distance of the target and the flankers (BOUMA, 1970), eccentricity (the distance of the target to the fixation point), as well as the similarity between the target and the flankers (Tripathy & Cavanagh, 2002) or the configuration of the flankers around the target object are known to effect crowding in humans (BOUMA, 1970).

Many computational models of crowding have been proposed (Balas, Nakano, & Rosenholtz, 2009). In this paper,

- We study the effect of crowding on the task of recognizing even MNIST digits in the presence of flankers in different DNNs. We chose 5 different configurations of flankers, each of them varying in number and/or eccentricity.
- We evaluate 3 different models on their ability to perform well in the presence of flankers. The models which we evaluated are: basic 3 layer convolutional network, eccentricity dependent network and the basic convolutional network augmented with a VOneNet block.
- We compare this with data collected from human subjects on similar configuration of target and flankers.

Models

Deep Convolutional Neural Networks

We first evaluate the performance of a deep convolutional neural network where the image is processed by three rounds of convolution and max pooling across space, and then passed to one fully connected layer for the classification. Each round has 32 filters, kernel size 5×5 and stride 1. We use a 3×3 max pooling with a stride of 2. The data in each layer is a 4-dimensional tensor of minibatch size × x × y × number of channels, in which x defines the width and y the height of the input. The input image to the model is resized to 60×60 pixels. For training, we used mini batches of 128 images. This forms the base to both the VOneNet model and the Eccentricity model and are built on top of it.

VOne Block

Since one of the many theories for crowding localizes it to the V1 cortex, we wanted to see how a network which simulates the primate visual cortex

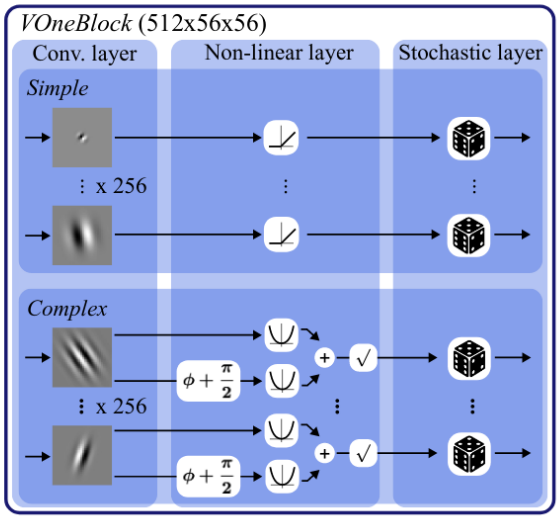


Figure 1: VOne Block.

(V1) in its initial layer would perform when presented with a crowded image. We use VOne Block as proposed in (Dapello et al., 2020) in the initial layer of our model to simulate V1 cortex and train it on different data sets and evaluate its performance.

VOne Block is a constrained neural network that simulates the primate visual cortex (V1). Using this augmentation improves robustness to adversarial attacks and increases performance for corrupted images while maintaining high accuracy.

The VOne Block contains a convolutional layer, a nonlinear layer, and a stochastic layer.

Convolutional layer The convolutional layer of VOneBlock is a mathematically parameterized Gabor Filter Bank (GFB). The stride of the GFB is set to four, which creates a 56×56 spatial map of activations. Since the number of channels in most CNNs' first convolution is relatively small, a larger number is used in the VOne Block so that the Gabors would cover the large parameter space and better approximate primate V1. The models contain 512 channels equally split between simple and complex cells. Each channel in the GFB convolves a single color channel from the input image.

The Gabor function consists of a two-dimensional grating with a Gaussian envelope and is described by the following equation:

$$G_{\theta,f,\phi,n_x,n_y}(x,y) = \frac{\cos(2\pi f + \phi)}{2\pi\sigma_x\sigma_y} \exp\left[-0.5\left(\frac{x_{rot}^2}{\sigma_x^2} + \frac{y_{rot}^2}{\sigma_y^2}\right)\right]$$

Where

$$x_{rot} = x\cos(\theta) + y\sin(\theta)$$

$$y_{rot} = -x\sin(\theta) + y\cos(\theta)$$

$$\sigma_x = \frac{n_x}{f}$$

$$\sigma_y = \frac{n_y}{f}$$

x_{rot} and y_{rot} are the orthogonal and parallel orientations relative to the grating, θ is the angle of the grating orientation, f is the spatial frequency of the grating, ϕ is the phase of the grating relative to the Gaussian envelope, and σ_x and σ_y are the standard deviations of the Gaussian envelope orthogonal and parallel to the grating, which can be defined as multiples (n_x and n_y) of the grating cycle (inverse of the frequency).

VOne Block's nonlinear layer has two different non linearities that are applied to each channel depending on its cell type: a rectified linear transformation for simple cells, and the spectral power of a quadrature phase-pair for complex cells:

$$S^{nl}_{\theta,f,\phi,n_x,n_y} = \mathrm{ReLU}\left(S^{l}_{\theta,f,\phi,n_x,n_y}\right)$$

$$\mathrm{ReLU}(x) = \begin{cases} x, & \text{if } x > 0 \\ 0, & \text{otherwise} \end{cases}$$

$$C^{nl}_{\theta,f,\phi,n_x,n_y} = \frac{1}{\sqrt{2}}\sqrt{\left(C^{l}_{\theta,f,\phi,n_x,n_y}\right)^2 + \left(C^{l}_{\theta,f,\phi+\pi/2,n_x,n_y}\right)^2}$$

Where $S^{l}_{\dots}$ and $S^{nl}_{\dots}$ are are the linear and nonlinear responses of a simple neuron and $C^{l}_{\dots}$ and $C^{nl}_{\dots}$ are the same for a complex neuron.

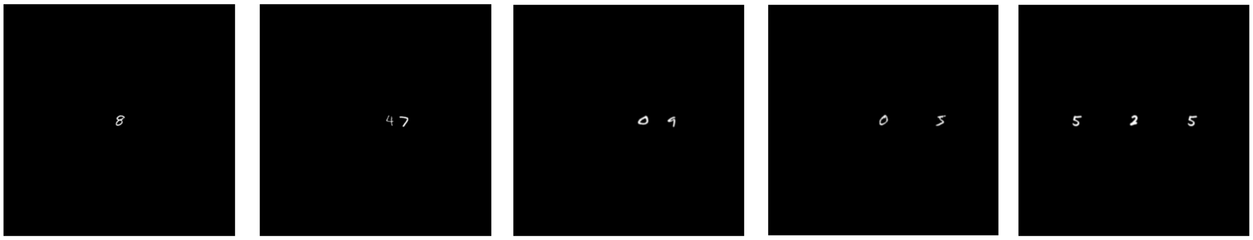
Stochastic layer In awake monkeys, spike trains of V1 neurons are approximately Poisson, i.e. the variance and mean of spike counts, in a given time-window, over a set of repetitions are roughly the same [82]. The stochasticity incorporated into the VOne Block emulates this property of neuronal responses. Since the Poisson distribution is not continuous, it breaks the gradients. In order to avoid this situation the neuronal stochasticity generator uses a continuous, second-order approximation of Poisson noise by adding Gaussian noise with variance equal to the activation:

$$R^{s} \sim \mathcal{N}(R^{ns}, R^{ns})$$

where R^{ns} and R^{s} are the non-stochastic and stochastic responses of a neuron.

Eccentricity Dependent Network

Eccentricity dependent model was proposed by (Poggio, Mutch, & Isik, 2014). It models the human visual cortex. Its eccentricity dependence is based on the human retina, which has receptive



(a) No flankers

(b) 1 flanker with ecc = 240 px

(c) 1 flanker with ecc = 360 px

(d) 1 flankers with ecc = 480 px

(e) 2 flankers with ecc = 480 px

Figure 2: Different flanker configurations used

fields which increase in size with eccentricity. According to (Poggio et al., 2014), computational reason for this property is the need to compute a scale and translation-invariant representation of objects and conjectures that this model is robust to clutter when the target is near the fixation point. The set of all scales and translations for which invariant representations can be computed lie within an inverted truncated pyramid shape, as shown in Fig 4. The width of the pyramid at a particular scale is roughly related to the amount of translation invariance for objects of that size. Scale invariance is prioritized over translation invariance in this model, in contrast to classical DNNs. From a biological point of view, the limitation of translation invariance can be compensated for by eye movements, whereas to compensate for a lack of scale invariance the human would have to move their entire body to change their distance to the object.

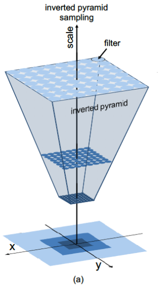


Figure 3: Eccentricity-dependent model: Inverted pyramid

Experiment Set-up

Models are trained with back-propagation to recognize a set of objects, called targets. During testing, we present the models with images which contain a target object as well as other objects which the model has not been trained to recognize, which we call flankers. The flanker acts as clutter with respect to the target object. We train our models to recognize even MNIST digits shifted at different locations of the image along the horizontal axis. The flankers are odd MNIST digits. We compare performance when we use images with the target object in isolation, or when flankers are also embedded in the training images. We experiment with different number of flankers (0 , 1 or 2) and different eccentricities of flankers (240px, 360px, 480px) as shown in Figure 2.

We used the same setup described in (Whitney & Levi, 2011). The images are of size 1920 squared pixels, in which we embedded target objects of 120 squared px, and flankers of the same size. Images are resized to 60 × 60 as input to the networks. We keep the training and testing splits provided by the MNIST dataset, and use it respectively for training and testing.

We also collect data from 5 human participants to get an estimate of human target recognition performance for each of the configurations of flankers. We show them 12 images from each of the category and record their response as well as the time taken by them to respond.

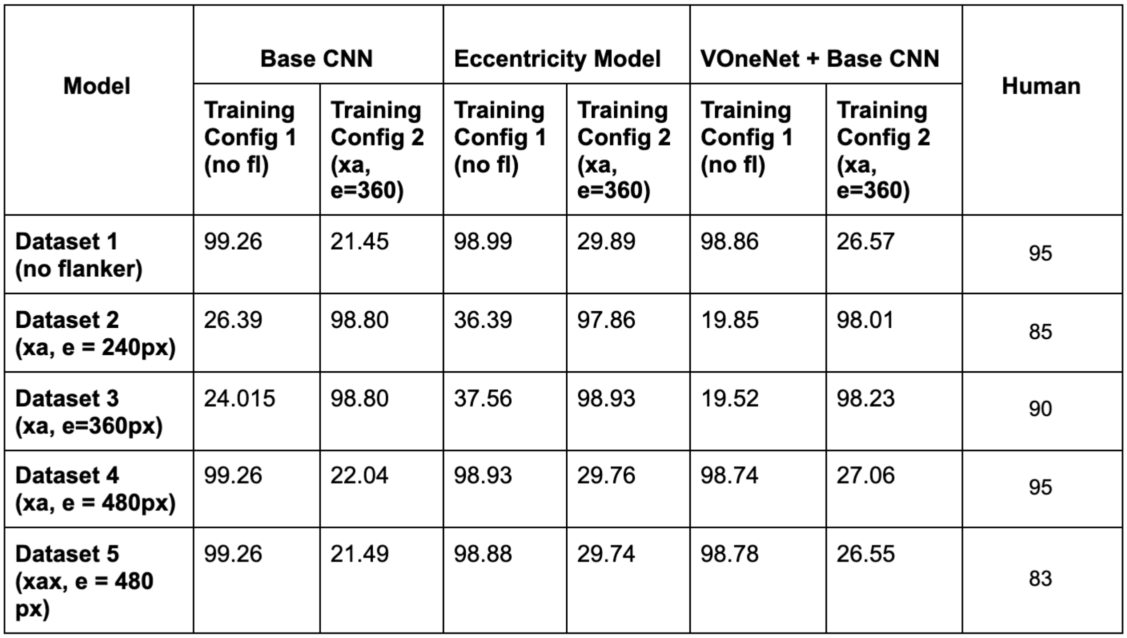
Experiments

Experiments are conducted on the 3 models that have been trained with images containing both targets and flankers. We then repeat our analysis with the models trained with images of the targets in isolation, shifted at all positions in the horizontal axis.

We analyze the effect of flanker configuration, model architecture and model type by evaluating accuracy recognition of the target objects.

DNNs trained with Images without flankers

Models are trained with target object in isolation and in different positions of the image horizontal



Model	Base CNN		Eccentricity Model		VOneNet + Base CNN		Human
	Training Config 1 (no fl)	Training Config 2 (xa, e=360)	Training Config 1 (no fl)	Training Config 2 (xa, e=360)	Training Config 1 (no fl)	Training Config 2 (xa, e=360)	
Dataset 1 (no flanker)	99.26	21.45	98.99	29.89	98.86	26.57	95
Dataset 2 (xa, e = 240px)	26.39	98.80	36.39	97.86	19.85	98.01	85
Dataset 3 (xa, e=360px)	24.015	98.80	37.56	98.93	19.52	98.23	90
Dataset 4 (xa, e = 480px)	99.26	22.04	98.93	29.76	98.74	27.06	95
Dataset 5 (xax, e = 480 px)	99.26	21.49	98.88	29.74	98.78	26.55	83

Figure 4: Model accuracy with different configurations of flankers

axis. We test the models on images with: (i) No flankers, (ii) 1 flanker with eccentricity of 240px, (iii) 1 flanker with eccentricity of 360px, (iv) 1 flanker with eccentricity of 480px, (v) 2 flankers with eccentricity of 480 px.

DNNs trained with targets and flankers

Models are trained with images in which there were two identical flankers randomly chosen from the training set of MNIST odd digits, placed at a distance of 120 pixels on either side of the target. We again test the models on images with: (i) No flankers, (ii) 1 flanker with eccentricity of 240px, (iii) 1 flanker with eccentricity of 360px, (iv) 1 flanker with eccentricity of 480px, (v) 2 flankers with eccentricity of 480 px.

Results

DNNs trained with Images without flankers

When models are trained with images of objects in isolation, adding flankers harms recognition. For all the 3 models, DNNs trained on the dataset without flankers perform well when evaluated on the test dataset without any flankers or when the flankers are far away from the target (eccentricity = 480px).

Human data collected also show similar patterns. The recognition accuracy is among the lowest when flankers are close to the target.

Within the three models, Eccentricity model produces the best accuracy with flankers (37%). VOneNet performs the worst. The reason could be the pooling layers, which merge response, used in VOne Block.

DNNs trained with targets and flankers

When models are trained with targets and flankers, the performance of the models on the test datasets flips. We see that the models are better at recognizing objects in clutter than isolated objects for all image locations tested, especially when the configuration of target and flanker is the same in the training images as in the testing images. For example, model trained on a dataset with 1 flanker of eccentricity 360px performs better on a test dataset with similar configuration of flankers like 1 flanker, e=240px rather than a test dataset with no flankers. Thus, in order for a model to be robust to all kinds of clutter, it needs to be trained with all possible target-flanker configurations, which is infeasible in practice.

This reflects the stage in the human trial when the participants have become accustomed to the flankers. It is observed that their reaction time is much quicker then.

Again, the Eccentricity model produces the best accuracy across all the test datasets. This could be because of the multi-scale crops that divide the image into discrete regions, letting the model learn from image parts as well as the whole image.

Unlike the previous case, VOneNet performs better than the Base DCNN, which implies that pooling in the initial layers helps when trained on flankers to recognize flankers.

Human Response time outliers

In our human trials, we realized that once the subject had been primed to look for a number (either 0,2,4,6, or 8), recognition errors were minimal. Furthermore, subjects had varying base reaction times (1-2s) due to factors outside of the image noise, such as age. In order to account for the response variation in participants, we calculate the number of outliers using the $1.5\times$IQR rule. That is, any reaction time that is slower than the third quartile + 1.5 × the interquartile range (outside 2.7σ) is identified as a delayed response in recognition.

Discussion

CNN performance

We investigated the performance of different DNNs with different configurations of flankers to identify the effect of crowding. We make the following observations:

- Testing models trained on images without flankers on images with flankers harms recognition. Adding two flankers is the same or worse than adding just one and the smaller the spacing between flanker and target, the more crowding occurs. This could be because of the Max Pool operation that merges nearby responses, such as the target and flankers if they are close.
- The eccentricity of the flankers effect the overall performance. The larger the eccentricity, the less is its effect on the target and accuracy.
- Flankers more similar to targets cause more crowding, because of the selectivity property of the learned DNN filters.
- Overall, the eccentricity model performs the best. VOneNet performs better than Base CNN when trained on dataset with flankers.

Human performance

Human trials show that there is a consistent delay in user response in three main cases. When the image is below a certain resolution all participants were unable to recognize the number; after some delay participants said they cannot see the image and gave a guess. When the image is ambiguous to the subject the guesses were influenced by personal experience (ie: a handwritten ”6 or 8” when the top is unclosed). Finally when flankers are first encountered, there was a small delay in recognition. The effect flankers had on recognition was small as the participants became accustomed to looking for an even number.

Overall, the set of generated test images was small enough that participants were able to quickly learn to recognize the numbers (0,2,4,6,8) after the first trial. For all users, there was a drop in recognition time within the first trial. Afterwards, there was a consistent base recognition time (1-2s depending on the participant) largely irregardless of flankers.

Underlying Differences

As suggested in a study on crowding in CNNs, ”DCNNs, while proficient in object recognition, likely achieve this competence through a set of mechanisms that are distinct from those in humans. They are not necessarily equivalent models of human or primate object recognition and caution must be exercised when inferring mechanisms derived from their operation.” (Lonnqvist, Clarke, & Chakravarthi, 2019)

Specifically, we observed the following phenomenon.

1. Consistent with the article (Lonnqvist et al., 2019), the change in accuracy when flanker distances were under 360px suggests the mechanism of pooling layer effect had a large influence on the CNN that is not present in human trials.
2. Furthermore, while flankers can be trained on, human participants were much better at combining their experiences to handle both images with flankers and without flankers. The CNNs, on the other hand, required separate models for the two different configurations when flankers were less than 360px away.
3. Finally, even though the VOneNet’s architecture is supposed to simulate a primary visual cortex, this alone does not make a model achieve ”closer” to human performance and actually performs worse when combined with the base CNN. This suggests that the nature of the solution to flankers within DCNNs is different from those within humans.

The human trials emphasizes a fundamental difference from the DCNNs provided in the original paper. The observation that human recognition in images with flankers is time-dependent while DCNNs have an accuracy metric (DNNs would identify incorrectly regardless of the amount of time given), reveals different mechanisms in play.

References

Balas, B., Nakano, L., & Rosenholtz, R. (2009, 11). A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision*, *9*(12), 13-13. Retrieved from https://doi.org/10.1167/9.12.13 doi: 10.1167/9.12.13

BOUMA, H. (1970). Interaction effects in parafoveal letter recognition. *Nature*, *226*(5241), 177–178. Retrieved from https://doi.org/10.1038/226177a0 doi: 10.1038/226177a0

Dapello, J., Marques, T., Schrimpf, M., Geiger, F., Cox, D. D., & DiCarlo, J. J. (2020). Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations. *bioRxiv*. doi: 10.1101/2020.06.16.154542

Levi, D. M. (2008). Crowding—an essential bottleneck for object recognition: A mini-review. *Vision Research*, *48*(5), 635-654. doi: https://doi.org/10.1016/j.visres.2007.12.009

Lonnqvist, B., Clarke, A. D. F., & Chakravarthi, R. (2019). Object recognition in deep convolutional neural networks is fundamentally different to that in humans. *CoRR*, *abs/1903.00258*. Retrieved from http://arxiv.org/abs/1903.00258

Poggio, T., Mutch, J., & Isik, L. (2014, 06). Computational role of eccentricity dependent cortical magnification.

Tripathy, S. P., & Cavanagh, P. (2002). The extent of crowding in peripheral vision does not scale with target size. *Vision Research*, *42*(20), 2357-2369. doi: https://doi.org/10.1016/S0042-6989(02)00197-9

Volokitin, A., Roig, G., & Poggio, T. A. (2017). Do deep neural networks suffer from crowding? In I. Guyon et al. (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc.

Whitney, D., & Levi, D. M. (2011). Visual crowding: a fundamental limit on conscious perception and object recognition. *Trends in Cognitive Sciences*, *15*(4), 160-168. doi: https://doi.org/10.1016/j.tics.2011.02.005