Spatial Attention introduces Behavioral Trade-off in a Large-Scale Spiking Neural Network

Lynn K.A. Sörensen (l.k.a.sorensen@uva.nl)
Psychology Department, University of Amsterdam,
Nieuwe Achtergracht 129B, 1018 WT Amsterdam, Netherlands

Davide Zambrano (d.zambrano@cwi.nl)
Machine Learning Group,
Centrum Wiskunde & Informatica, Science Park 123,
1098 XG Amsterdam, Netherlands

Heleen A. Slagter (h.a.slagter@vu.nl)
Department of Experimental & Applied Psychology,
Vrije Universiteit Amsterdam, Van der Boechorstraat 7,
1081 BT Amsterdam, Netherlands

H. Steven Scholte* (h.s.scholte@uva.nl)
Psychology Department,
University of Amsterdam, Nieuwe Achtergracht 129B,
1018 WT Amsterdam, Netherlands

Sander M. Bohté* (s.m.bohte@cwi.nl)
Machine Learning Group,
Centrum Wiskunde & Informatica, Science Park 123,
1098 XG Amsterdam, Netherlands

* Shared last authorship

Abstract:

Visuo-spatial attention is a key mechanism for selecting goal-relevant information in natural scenes. We here implement a variant of the normalization model of attention into a spiking convolutional neural network, which approximates attentional gain with a change in firing rates. We apply this type of attention with different spatial extents to various levels in the processing hierarchy of a network performing object recognition in natural scenes. We find that close to the average object-size attentional kernels yield the best performance, equivalent to a rather focused attentional enhancement. Furthermore, manipulating spatial attention within a single level was ineffective as benefits of spatial attention only arose from the combination of early-to-mid level modulations in the network hierarchy. Our results demonstrate that one can efficiently boost performance on the challenging task of recognizing objects in cluttered environments in a large-scale vision model by understanding attentional gain as a more or less precise representation of sensory information.

Keywords: Spiking Neural Network; Spatial Attention; Object Recognition; Natural Scenes

Introduction

Visuo-spatial attention is a key mechanism for visual perception when dealing with the richness of natural scenes by allowing the system to prioritize processing at attended over unattended locations (Desimone & Duncan, 1995). Indeed, small, yet robust attention effects have been observed for a variety of performance measures such as reaction time, accuracy and perceptual sensitivity (Carrasco, 2011). While an extensive body of research found these attentional benefits for abstract, simple stimuli (e.g. Gabor patches in an orientation change task), the impact of spatial attention on natural scene processing is only started to be understood (Battistoni, Stein, & Peelen, 2017).

Behavioral changes for abstract stimuli have been associated with a range of attentional modulations of visual neural activity, such as altered firing rates, noise correlations and tuning properties (Maunsell, 2015). One proposal that integrates these diverse attention effects is the Normalization Model of Attention (NMA, Reynolds & Heeger, 2009). In this model, the changes in neural firing rates are described as a function of stimulus drive (sensory input within the receptive field modulated by its feature tuning properties), attention field (specified over the feature tuning and receptive field of the neuron) and pooling across space and features. Specifically, the stimulus drive is modulated by the attention field and then in turn normalized to neighboring locations and features. Using simple orientation discrimination tasks, this model has been related to human behavior and to changes in activation in fMRI by modeling the relationship between attention field and stimulus size (Herrmann et al., 2010). An outstanding question that remains is whether the modeled principles in attentional gain in the NMA critically account for behavior also in more ecological scenarios such as object recognition in natural images.

Lindsay & Miller (2018) have investigated the effect of attention in more complex visual tasks using convolutional neural networks (CNN) as a model of the visual system. Their work showed that feature similarity gain (Treue & Martínez Trujillo, 1999), especially when applied to the first network layers, has a smaller effect than formerly assumed in improving model performance. This work illustrates that principles derived from observations of single or populations of neurons not necessarily have to give rise to a behavioral effect in more complex visual scenarios.

Spiking convolutional neural networks (sCNNs) allow us to address the relationship between neural attention gain and behavior in a more biologically plausible way: By adjusting the firing threshold of individual neurons in the network, we can modulate firing rates as a function of attentional gain, paralleling observations from attention effects in visual neurons (Desimone & Duncan, 1995; Maunsell, 2015). Importantly, the encoded activation in the spike trains is equivalent to the activations in a CNN, while the precision of the encoded activation can be changed by using less or more spikes to represent this activation (Fig. 1A, Zambrano et al., 2018). Thereby, we can manipulate the information processing strategy in the network and examine how this affects classification behavior.

Here, we use sCNNs to investigate whether the principles from the NMA can affect performance during object recognition in natural scenes. Specifically, we simulate spatial attention to real-world images by selectively manipulating the firing threshold for a location at various levels of the network hierarchy, and probe behavior of the model as a function of valid or invalidly attended location.

Methods

Dataset

To obtain images with a set of potential target categories sharing a context, we first curated a dataset from the Common Objects in Context database (COCO, Lin et al., 2014). We selected images with target objects that were big enough (>0.05% of the image), placed in a not too complex scene (spatial coherence < 1.2; Scholte et al., 2009), were not too central (outside of a radius of 5% from the image center) and salient enough (summed object probability density from DeepGaze II > 0.04; Kümmerer, Wallis, & Bethge, 2016) resulting in 8 eligible target categories (Fig. 3A) with at least 50 images with a single target object. The spatial attention experiments were conducted on a category-balanced subset of the single-target images (N=526, 224x224 pixels). Due to imperfect COCO annotations, some images could also feature other target categories. For the network fine-tuning, we used images with more than one target object present (N_{train}=7103, $N_{validation}$=1776, N_{test}=2736).

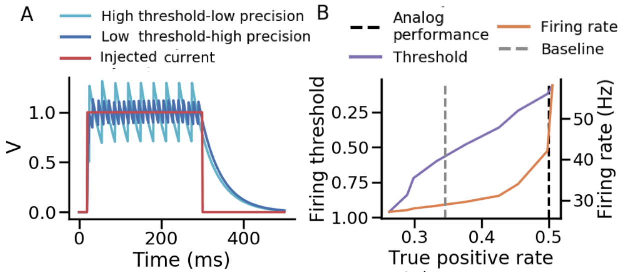

Figure 1: Properties & performance of a sCNN. (A) Depending on the firing threshold, the same current is encoded by a spiking unit with more or less precision, effectively resulting in more or less variation around the integrated current. (B) True positive rate at different firing thresholds (purple) for the fine-tuned model and corresponding firing rates (orange). All models had a precise encoding stage at 0.06 (cf. Baseline precision).

Deep Spiking Neural Networks

Fine-tuning A ResNet18 architecture (He et al., 2015) optimized for post-training conversion to an adaptive spiking network and trained on the ImageNet dataset served as a basis for all experiments (for details see Zambrano et al., 2018). The last fully-connected layer of this model was replaced with 8 fully-connected units computing a sigmoid activation function. This model was fine-tuned on the multi-object dataset with a binary cross-entropy loss function to optimize for independent class distributions while the rest of the network' weights stayed unchanged (lr: 0.0001, 200 epochs). The resulting model had a F1-score of 0.54 on the held-out multi-object test set.

Evaluation Due to incomplete image annotations (cf. Dataset), we evaluated model performance in true positive rate to capture the changes in likelihood of detecting the chosen target category for the attention experiments. A model detected a class as present when the mean prediction time course (200-700ms after stimulus onset) surpassed 0.5 (Fig. 3A). We computed firing rate (FR) as being the mean number of spikes emitted by the network over 4 randomly chosen images.

Spatial attention

Precision Modulation We focused on attention effects along two spatial dimensions and in turn did not include feature tuning in our interpretation of the NMA. We computed the modulation for a given layer based on the NMA by equating the stimulus drive with 1 (equal amount of precision at every location). Expanding the model to two spatial dimensions (x_1, x_2), we obtain:

$$R(x_1, x_2) = (A(x_1, x_2) / [S(x_1, x_2)]) - 1$$

$$S(x_1, x_2) = s(IxWidth) * A(x_1, x_2),$$

where $A(x_1, x_2)$ is the attention field described by a two-dimensional Gaussian centered at Ax with a width described by AxWidth, s(IxWidth) is the suppressive field capturing spatial pooling with IxWidth and * is a convolution. The resulting modulation matrix R is multiplied with the baseline-precision and normalized to the sum of an unmodulated layer (Fig. 2B-D).

Baseline Precision We kept precision high (firing threshold: 0.06) in the layers before the ResNet blocks and lowered overall precision in the remaining network[1]. The baseline for the remaining layers was determined by fitting a line to the decrease in top-1 accuracy as a function of precision. Taking this fit, we estimated the precision needed to perform at the mid-point between analog and chance performance (based on permutation-testing on the analog (non-spiking) model predictions), which served as a baseline (firing threshold: 0.55, FR: 29 Hz) for the precision modulations in the attention experiments (Fig. 1B).

Model Modulation Spatial attention was modulated for a given ResNet block [batch normalization (BN), activation, convolution (Conv), BN, activation, Conv]. To maintain the spatial resolution between these blocks, AxWidth$_{1,2}$ and IxWidth were divided by the total stride of a given layer in a given block [4, 4, 8, 8, 16, 16]. Attention was applied in isolation to a respective ResNet block, successively from lower to higher ResNet blocks (1-6/ 2-6), or successively from higher to lower blocks (6-1/ 5-1, Fig. 3C). We approximated stimulus extent with the radius describing the mean target object area and chose AxWidth and IxWidth to be related to stimulus size as modelled in Heeger & Reynolds, 2009 (cf. Exp. 2B, 5C, & 6C) resulting in either a focused, object-sized or distributed attentional field (AxWidth$_{1,2}$: 24, 40, 120, IxWidth: 80; Fig. 2B-D).

Experiments We probed model performance for valid and invalid modulations of spatial attention location on the same single-target dataset. While a valid modulation was when Ax$_{1,2}$ described the center of mass of the target object, an invalid modulation referred to a target location based on another image, which was at least 100 pixels away from the valid location. Results were compared to the neutral baseline, where precision was the same at all locations. All presented experiments were evaluated with equal spike numbers.

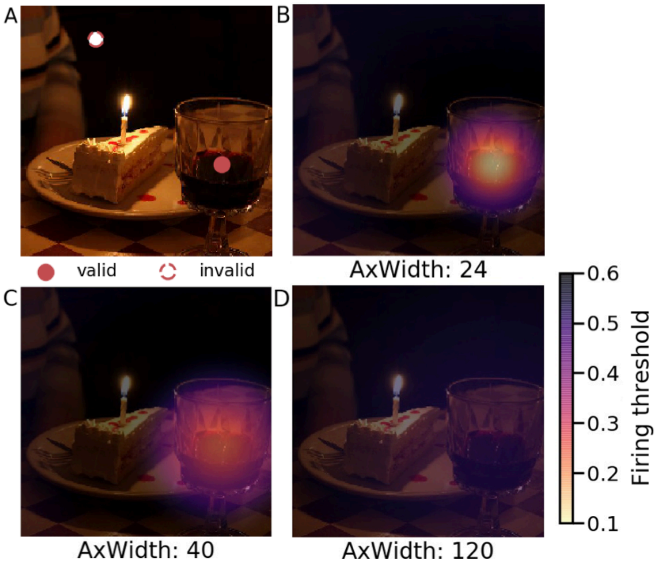

Figure 2: Spatial attention experiments. (A) Example image of the dataset. Valid and invalid locations are marked for illustration. (B-D) Spiking threshold maps projected on the image for the three modelled conditions.

Results & Discussion

Attention Improves Object Recognition

The experiments showed that spatial attention can introduce a behavioral trade-off, boosting object recognition for valid locations and moderately hampering performance for invalidly attended locations, with an equal number of spikes (Figure 3B). Notably, spatial attention had to be orchestrated across multiple subsequent blocks and have a width smaller than or equal to the average object to evoke this behavioral modulation (4.2% performance gain, 7.9% attentional modulation, Fig. 3). The observed moderate performance gains here are in line with earlier work manipulating attention in the early-to-mid layers of a CNN (Lindsay & Miller, 2018) and findings from electrophysiological and psychophysics studies of spatial attention (e.g. Chica et al. 2015). Our results extend this work by showing that one can boost performance by understanding attentional gain as a more or less precise representation of sensory information while keeping energy expenses constant. Future work can expand on this by also evaluating the prediction time courses as a proxy for decision making.

Spatial attention as a hierarchical mechanism

In line with earlier research (e.g. Herrington & Assad, 2010), we found that effects of spatial attention are likely to act throughout the hierarchy and not in an isolated manner (Fig. 3D): effects of attention added up mainly across the early-to-mid layers (up to block 5). At the last stage, representations might be too coarse to be meaningfully targeted by spatial attention.

Object recognition in natural scenes

Our primary goal was to understand the impact of different attentional settings of the NMA on object recognition in natural scenes. We observed that an attentional width close to the average target object size was the most effective, followed by an even narrower width. A wider attentional enhancement carried no

[1] Pilot experiments on manipulating precision in isolated layers showed that lowering precision before the ResNet blocks resulted in more severe performance losses as compared to other consecutive layers, indicating that the underlying activation distributions are qualitatively different from those in the ResNet blocks.

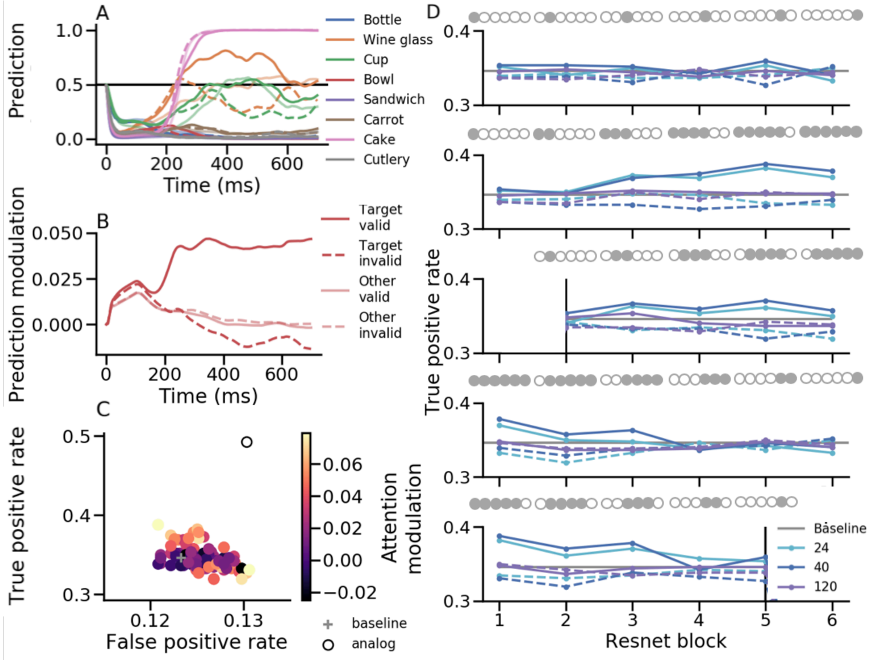

Figure 3: Focussed spatial attention introduces behavioural trade-off.
(A) Example prediction time course for the example image in 2A. Solid, dashed and transparent lines represent the valid, invalid and neutral condition, respectively. (B) Average prediction modulation compared to baseline for AxWidth 40 and bottom-up modulation up to Resnet block 5. (C) True and false positive rates for all experiments (valid, invalid) in comparison to baseline and analogue performance. Color represents the strength of attention modulation in true positive rates, (valid – invalid) / (valid + invalid) (D) True positive rates shown as a function of the spatial attention manipulation. Dashed lines are the invalid conditions. Filled grey circles represent the targeted Resnet blocks. For repeated experiments, estimates were averaged across repetions.

effects. This not only highlights the importance of taking into account the relationship between attentional field and stimulus size for interpreting neural attention gains (Herrmann et al., 2010; Reynolds & Heeger, 2009), but also makes the prediction that if an object in a natural scene is not easily recognized (here equivalent to a low precision mode in the current experiments), it will be the most efficient for an observer to leverage expected object size within the scene context to scale spatial attention.

In sum, we show that manipulating spatial attention based on the NMA in a deep spiking neural network affects object recognition in cluttered natural scenes.

Acknowledgments

This work was supported by an NWO Talent Grant.

References

Battistoni, E., Stein, T., & Peelen, M. V. (2017). Preparatory attention in visual cortex. Annals of the New York Academy of Sciences, 1396(1), 92–107.

Carrasco, M. (2011). Visual attention: the past 25 years. Vision Research, 51(13), 1484–1525.

Chica, A. B., Martín-Arévalo, E., Botta, F., & Lupiánez, J. (2014). The Spatial Orienting paradigm: How to design and interpret spatial attention experiments. Neuroscience & Biobehavioral Reviews, 40, 35-51.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. Annual Review of Neuroscience, 18, 193–222.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. Retrieved from http://arxiv.org/abs/1512.03385

Herrington, T. M., & Assad, J. A. (2010). Temporal sequence of attentional modulation in the lateral intraparietal area and middle temporal area during rapid covert shifts of attention. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 30(9), 3287–3296.

Herrmann, K., Montaser-Kouhsari, L., Carrasco, M., & Heeger, D. J. (2010). When size matters: attention affects performance by contrast or response gain. Nature Neuroscience, 13(12), 1554–1559.

Kümmerer, M., Wallis, T. S. A., & Bethge, M. (2016). DeepGaze II: Reading fixations from deep features trained on object recognition. Retrieved from http://arxiv.org/abs/1610.01563

Lindsay, G. W., & Miller, K. D. (2018). How biological attention mechanisms improve task performance in a large-scale visual system model. eLife, 7. https://doi.org/10.7554/eLife.38105

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., … Dollár, P. (2014). Microsoft COCO: Common Objects in Context. Retrieved from http://arxiv.org/abs/1405.0312

Maunsell, J. H. R. (2015). Neuronal Mechanisms of Visual Attention. Annual Review of Vision Science, 1, 373–391.

Reynolds, J. H., & Heeger, D. J. (2009). The normalization model of attention. Neuron, 61(2), 168–185.

Scholte, H. S., Ghebreab, S., Waldorp, L., Smeulders, A. W. M., & Lamme, V. A. F. (2009). Brain responses strongly correlate with Weibull image statistics when processing natural images. Journal of Vision, 9(4), 29.1–15.

Treue, S., & Martínez Trujillo, J. C. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. Nature, 399(6736), 575–579.

Zambrano, D., Nusselder, R., Scholte, H. S., & Bohté, S. M. (2018). Sparse Computation in Adaptive Spiking Neural Networks. Frontiers in Neuroscience, 12, 987.