THE PERCEPTUAL BOOST OF VISUAL ATTENTION IS TASK-DEPENDENT IN NATURALISTIC SETTINGS

Freddie Bickford Smith, Xiaoliang Luo, Brett D Roads, Bradley C Love
University College London
{f.bickfordsmith.18,xiao.luo.17,b.roads,b.love}@ucl.ac.uk

ABSTRACT

Top-down attention allows people to focus on task-relevant visual information. Is the resulting perceptual boost task-dependent in naturalistic settings? We aim to answer this with a large-scale computational experiment. First, we design a collection of visual tasks, each consisting of classifying images from a chosen task set (subset of ImageNet categories). The nature of a task is determined by which categories are included in the task set. Second, on each task we train an attention-augmented neural network and then compare its accuracy to that of a baseline network. We show that the perceptual boost of attention is stronger with increasing task-set difficulty, weaker with increasing task-set size and weaker with increasing perceptual similarity within a task set.

1 INTRODUCTION

By modulating neural representations of visual stimuli, top-down attention enables prioritised processing of task-relevant information. Past work has shown that this mechanism enhances people's perceptual abilities and that its influence varies between visual tasks (Carrasco, 2011). Does this hold for a broad range of naturalistic tasks, like recognising an object in an everyday scene? To our knowledge, the literature lacks a systematic characterisation of this issue. At the same time, thanks to faster computers and larger sets of annotated data, it has become straightforward to train deep convolutional neural networks to perform such visual tasks (Krizhevsky et al, 2012). Analyses have shown that these networks are state-of-the-art computational models of the human visual system (Yamins & DiCarlo, 2016). Hence there is a gap in the visual-attention literature, and there is a computational-modelling technique that has only recently become practically usable and validated by neuroscience. Our work addresses the gap with the modelling technique. By studying the performance of neural networks on a number of visual tasks, we seek to establish how the perceptual boost of attention varies with the nature of a task.

Each task we consider consists of classifying images from a chosen *task set*, a subset of categories from the ImageNet dataset (Russakovsky et al, 2015). We define three quantitative dimensions along which a task set can vary: difficulty, size and perceptual similarity (see Section 2.1 for definitions). Our hypothesis is that each of these *task-set properties* is important in determining the perceptual boost of attention on a task. We base this on observations and intuitions. First, ImageNet categories are known to range in difficulty, likely due to variation in factors such as object scale, image clutter and shape distinctiveness (Russakovsky et al, 2015). If attention interacts with such factors, its impact on perception could be expected to vary with task-set difficulty. Second, increasing the breadth of a task (eg, performing classification on examples from a larger number of categories) generally introduces more severe tradeoffs when learning to weight the features of neural representations. If this is true, and if attention is framed as modulating representations, the size of a task set should influence how effectively attention can act. Third, defining attention as before, the similarity of representations within a task set might affect how usefully attention can shape them.

Our experiment[1] is based on a collection of task sets that vary with respect to the task-set properties we define. For each task set, we construct an attention-augmented neural network (more concisely, an *attention network*) by taking an ImageNet-pretrained VGG16 (Simonyan & Zisserman, 2015),

---

[1]Code and data are available at github.com/fbickfordsmith/attention-iclr.

fixing its weights and inserting a trainable attention mechanism. Then we train the attention network solely on images from the task set. The perceptual boost of attention is measured by comparing the accuracy of this optimised attention network to the accuracy of a baseline network. In this way, we gather evidence to assess how attention's impact on perception varies with task-set properties.

2 METHOD

2.1 TASK-SET PROPERTIES

Let task set $\mathcal{C}$ be a subset of the 1000 image categories in the ImageNet dataset. We define three task-set properties: difficulty, size and perceptual similarity. By choosing which categories to include in a task set, we control its properties. The *difficulty* of $\mathcal{C}$ is the mean error rate of an ImageNet-pretrained VGG16 on categories in $\mathcal{C}$:

$$\text{difficulty}(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum_{c_i \in \mathcal{C}} (1 - \text{accuracy}_{\text{VGG}}(c_i)) \tag{1}$$

where accuracy takes values between 0 and 1. The *size* of $\mathcal{C}$ is the number of categories it contains:

$$\text{size}(\mathcal{C}) = |\mathcal{C}| \tag{2}$$

The *perceptual similarity* of $\mathcal{C}$ is the mean pairwise similarity of the categories in it:

$$\text{similarity}(\mathcal{C}) = \frac{1}{|\mathcal{C}|^2 - |\mathcal{C}|} \sum_{c_i \in \mathcal{C}} \sum_{c_j \in \mathcal{C}} \mathbb{I}(i \neq j) s(c_i, c_j) \tag{3}$$

where $\mathbb{I}$ is an indicator function and $s(c_i, c_j)$ is the cosine similarity between the average representations of categories c_i and c_j. That is,

$$s(c_i, c_j) = \frac{r(c_i) \cdot r(c_j)}{\|r(c_i)\| \|r(c_j)\|} \qquad r(c) = \frac{1}{|\mathcal{X}_c|} \sum_{x_i \in \mathcal{X}_c} \text{VGG}'(x_i) \qquad r \in \mathbb{R}^{4096} \tag{4}$$

where $\mathcal{X}_c$ is the subset of the training images for which the label is c, and VGG' denotes VGG16 with its final layer removed.

2.2 NEURAL NETWORKS

We use an ImageNet-pretrained VGG16 as a foundational neural network, and incorporate visual attention as a linear modulation of neural representations (Lindsay & Miller, 2018). VGG16 computes a probability distribution over the 1000 possible states of the category label, c, for a given image, x:

$$p(c|x) = \text{VGG}(x) \tag{5}$$

The network can be decomposed into two parts. The convolutional layers, VGG_1, transform x to a latent representation, z. The fully-connected layers, VGG_2, transform z to $p(c|x)$. That is,

$$\text{VGG}_1 : x \to z \qquad \text{VGG}_2 : z \to p(c|x) \qquad x \in \mathbb{R}^{224 \times 224 \times 3} \qquad z \in \mathbb{R}^{7 \times 7 \times 512} \tag{6}$$

We define attention to be the multiplication of z by nonnegative attention weights, a. Our attention network computes

$$p(c|x, a) = \text{VGG}_2(a \odot \text{VGG}_1(x)) \qquad a \in \mathbb{R}_{\geq 0}^{7 \times 7 \times 512} \tag{7}$$

where $\odot$ denotes an elementwise multiplication. We treat the pretrained VGG_1 and VGG_2 as fixed functions. The attention weights are the only trainable parameters and are initialised to 1.

2.3 EXPERIMENT DESIGN

Using the task-set properties defined in Section 2.1, we assemble three groups of task sets. *Difficulty-based* task sets vary substantially in difficulty but not in size or perceptual similarity. *Size-based* task sets are diverse in size but similar in difficulty and perceptual similarity. *Similarity-based* task sets have a wide range of perceptual-similarity values but constant size and near-constant difficulty.

For each task set, we construct a new attention network and train it solely on images from the task set. Then we compare the accuracy of this attention network to the accuracy of a baseline network (a task-agnostic attention network trained on all ImageNet categories). We make this comparison on examples both from within the task set (giving the change in *in-set* accuracy) and from outside the set (giving the change in *out-of-set* accuracy). Our primary interest is the perceptual boost that attention produces on the task for which it is optimised (ie, the change in in-set accuracy), but for completeness we also include results for out-of-set accuracy. Figure 1 summarises our approach.
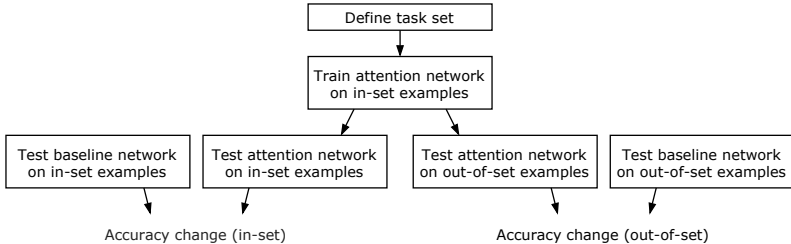

Figure 1: Procedure for measuring how the nature of an image-classification task, controlled by the choice of task set, affects the accuracy change produced by attention. This is performed for each task set.

3   RESULTS

For all tasks, applying attention results in higher in-set accuracy and lower out-of-set accuracy (see Figure 2 and Table 1). This effect is stronger with increasing task-set difficulty, weaker with increasing task-set size and weaker with increasing perceptual similarity within a task set.
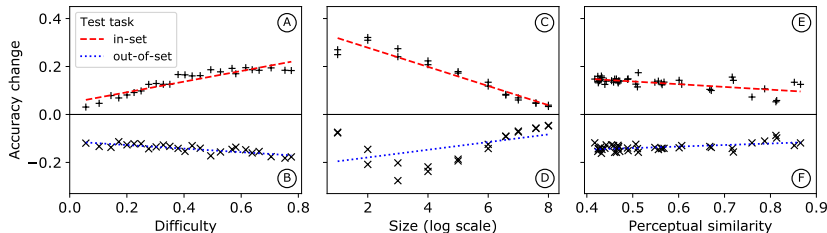

Figure 2:   Accuracy change produced by attention on 25 difficulty-based task sets (left), 20 size-based task sets (middle) and 40 similarity-based task sets (right). Task-set size is transformed logarithmically with base 2. Least-squares linear regression is applied to each subset of results, from Ⓐ to Ⓕ; predictions of the linear models are shown as broken lines.

	Property	Test task	ρ	τ_b	β_0	β_1	R^2
Ⓐ	difficulty	in-set	0.93	0.79	0.05	0.22	0.85
Ⓑ	difficulty	out-of-set	-0.83	-0.65	-0.11	-0.07	0.68
Ⓒ	size	in-set	-0.97	-0.91	0.36	-0.04	0.93
Ⓓ	size	out-of-set	0.69	0.59	-0.21	(0.02)	0.27
Ⓔ	similarity	in-set	-0.53	-0.38	0.19	-0.11	0.39
Ⓕ	similarity	out-of-set	(0.37)	(0.27)	-0.17	0.06	0.31

Table 1:   Relationships between the three task-set properties defined in Section 2.1 (the controlled experimental variables) and the accuracy change produced by attention (the measured experimental variable). Five statistics are shown for each subset of results, from Ⓐ to Ⓕ: Spearman's ρ; Kendall's τ_b; coefficient of determination, R^2; regression intercept, β_0; regression slope, β_1. For each value of ρ, τ_b, β_0 and β_1, except those in parentheses, $p < 0.001$.

4 DISCUSSION

The task-set properties defined in Section 2.1 have significant influence on the perceptual boost of visual attention (see Section 3). First, the more difficult a task set is, the greater the impact of attention on in-set accuracy (the performance measure we focus on in this discussion). This is perhaps because attention mitigates the effects of factors that contribute towards difficulty (eg, object scale, image clutter and shape distinctiveness). Second, increasing the number of categories in a task set results in a weaker attention-induced boost to in-set accuracy. The mechanism underlying this might be to do with the weight tradeoffs involved when training a neural network (Sutton et al, 2006). A visual feature might be discriminative for images from some categories but confounding for images from others; the weight placed on that feature is determined by its net contribution to accuracy across the whole training set. Optimising weights on a restricted set of categories (all tasks sets in our experiment contain at most 256 categories) corresponds to a relaxation of this weight tradeoff, allowing the network to make use of features that previously might have been dominated by others. (Similar logic might explain the difficulty result: high-difficulty categories are those whose most indicative features are most severely drowned out in the training process; specialising on fewer categories has an outsized effect for more difficult task sets.) Third, the greater the perceptual similarity of image categories within a task set, the less pronounced the impact of attention on performance. This could be because images from similar categories are represented alike in neural activations (this is how we define perceptual similarity) and a linear reweighting by attention does not much enhance the network's ability to discriminate between them.

5 CONCLUSION

The perceptual boost of visual attention varies substantially between naturalistic tasks. As well as informing further basic research on attention, this finding has practical implications for neural-network designers deciding whether to use a visual-attention mechanism. Empirically testing our suggested explanations of the results would be an interesting direction for future work. So would a study of the perceptual boost of attention when there is covariance between task-set properties (our experiment was designed to minimise this).

ACKNOWLEDGEMENTS

REFERENCES

Carrasco (2011). Visual attention: the past 25 years. Vision Research.

Krizhevsky, Sutskever, Hinton (2012). ImageNet classification with deep convolutional neural networks. Neural Information Processing Systems.

Lindsay, Miller (2018). How biological attention mechanisms improve task performance in a large-scale visual system model. eLife.

Russakovsky, Deng, Su, Krause, Satheesh, Ma, Huang, Karpathy, Khosla, Bernstein, Berg, Fei-Fei (2015). ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision.

Simonyan, Zisserman (2015). Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations.

Sutton, Sindelar, McCallum (2006). Reducing weight undertraining in structured discriminative learning. Human Language Technology Conference of the NAACL.

Yamins, DiCarlo (2016). Using goal-driven deep learning models to understand sensory cortex. Nature Neuroscience.