Computational Brain & Behavior (2023) 6:503–512
https://doi.org/10.1007/s42113-023-00171-8

ORIGINAL PAPER

Feature Attention as a Control Mechanism for the Balance of Speed and Accuracy in Visual Search

Thom Griffith[1] · Florence J. Townend[1,2] · Sophie-Anne Baker[1] · Nathan F. Lepora[1]

Accepted: 6 March 2023 / Published online: 13 June 2023

Abstract

Finding an object amongst a cluttered visual scene is an everyday task for humans but presents a fundamental challenge to computational models performing this feat. Previous attempts to model efficient visual search have focused on locating targets as swiftly as possible, but so far have not considered balancing the costs of lengthy searches against the costs of making errors. Here, we propose a neuro-inspired model of visual search that offers an attention-based control mechanism for this speed-accuracy trade-off. The model combines a goal-based fixation policy, which captures human-like behaviour on a simple visual search task, with a deep neural network that carries out the target detection step. The neural network is patched with a target-based feature attention model previously applied to standalone classification tasks. In contrast to image classification, visual search introduces a time component, which places an additional demand on the model to minimise the time cost of the search whilst also maintaining acceptable accuracy. The proposed model balances these two costs by modulating the attentional strength given to characteristic features of the target class, thereby minimising an associated cost function. The model offers a method for optimising the costs of visual search and demonstrates the value of a decision theoretic approach to modelling more complex visual tasks involving attention.

Keywords Visual search · Feature attention · Speed-accuracy trade-off · CNN

Introduction

The goal in visual search is to locate a target in a distracting visual scene. Traditionally, visual search in humans has been studied using simple stimuli, such as a display of a target object amongst distractors. However, recent developments in image processing tools, specifically convolutional neural networks (CNNs), have enabled visual search models that can handle natural images. Traditional computer vision methods for visual search, such as instance segmentation and object detection, require meticulously labelled training data sets to localise objects in images (Hariharan et al., 2014; Dai et al., 2016; Zou et al., 2023). These systems use whole images or sliding windows to scan the image. However, humans use small, rapid eye movements called saccades to foveate on areas of interest because the fovea, a small area of the eye, has the highest visual resolution on the retina. The peripheral areas of the retina are much less effective as detectors, as demonstrated by retinotopic maps of target discriminability (Najemnik & Geisler, 2009). Therefore, effectively selecting regions of the image to fixate on is important for efficient visual search.

Efficient visual search is a hard problem because it requires economical use of computational and temporal resources; i.e., where to deploy the increased discriminability of the fovea (considered as the fixation policy) and how long to search for before being confident that the target has been found (considered as target detection). In realistic environments where mistakes could prove costly and time may be priced according to opportunity cost or metabolic expenditure, efficient visual search behaviour should minimise those

Thom Griffith and Florence J. Townend contributed equally to this work.

✉ Thom Griffith
thom.griffith@bristol.ac.uk

[1] Department of Engineering Maths, University of Bristol, University Walk, Bristol BS8 1TW, UK

[2] Centre for Medical Image Computing, Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK

costs (Verghese, 2001). How, then, should an agent optimally balance search time and accuracy?

In this paper, we propose a model of adaptive visual search that uses a model of feature attention, conceived in Lindsay and Miller (2018), that has previously only been applied to pre-trained target detectors on standalone classification tasks. Feature attention has similarly been applied, in a decision theoretic context, to simple image classification tasks (Lindsay, 2020a; Luo et al., 2021). Here, we build task complexity from these previous works by applying feature attention to the problem of visual search. We propose a neuro-inspired fixation policy that compares Convolutional Neural Network (CNN)-derived feature maps from the target and the search image in order to determine the visual search path. The model performance is shown to match human data from a similar task (Zhang et al., 2018). This fixation policy uses a novel application of the Structural Similarity Index Measure (SSIM), a perceptual metric, to select portions of the search image for classification by a CNN-based model for target detection. The target detector is based on a current model of human attention mechanisms during visual search by implementing a top-down feature attention mechanism in a CNN that controls the relative occurrence of true and false positives. We then demonstrate how the fixation policy and target detection step work together to provide a mechanism for controlling the speed-accuracy trade-off in a visual search task where the relevant control parameter is the attentional strength given to characteristic features of the target class.

Background and Related Work

Many models of visual search utilise the notion of a saliency map (Koch & Ullman, 1987; Itti & Koch, 2001), although other approaches have included stochastic accumulator models (Chen et al., 2011; Chen & Perona, 2017), optimal control of foveated vision (Zelinsky et al., 2005; Zelinsky, 2008; Akbas & Eckstein, 2017), and models of Bayesian ideal observers or other information maximising policies (Verghese, 2001; Renninger et al., 2004; Najemnik & Geisler, 2005, 2009; Rashidi et al., 2020). A saliency map is a spatial representation of how much regions of the visual scene "grab our attention". The values in the map determine the regions of interest for further image processing and in which order the model should "fixate" on the image (Itti & Koch, 2001; Li, 2002). As originally proposed, saliency maps are generated by decomposing the visual scene to a set of feature maps, and the spatial locations within each map compete to be the most salient. The saliency over different feature maps combine to give one "master" saliency map (Koch & Ullman, 1987). Later versions generated feature maps taking inspiration from the architecture of the early primate visual system (Itti et al., 1998; Miconi et al., 2016) but feature maps for visual search can also be handcrafted based on known features of the targets and distractors in a scene (Navalpakkam & Itti, 2005). These methods for extracting image features have since been superceded by convolutional neural networks (CNNs), which can be trained to recognise simple features, such as edges, as well as more abstracted features, in a hierarchical manner, at different scales and orientations (Lindsay, 2020b). This means they are able to capture not just local features but the relationships between information in different regions of the image. Whilst CNNs are inspired by the visual system in the human brain, research suggests that representations of saliency maps may also exist in the posterior parietal cortex in primates. For instance, it has been hypothesised that similar processes take place in stages in the human visual system from primary visual cortex VI (Li, 2002) to the lateral interparietal region (LIP) (Gottlieb et al., 1998; Goldberg et al., 2006).

Originally, saliency maps were generated in a bottom-up fashion only using information from the scene itself. In contrast, the ability to electively fixate on regions in a visual scene according to a specific goal or target is known as top-down attention. Top-down modulation can be integrated into bottom-up saliency maps to give a priority map for fixation. Top-down modulated saliency maps have an advantage over vanilla, bottom-up saliency maps in that they can adapt to the context of the search, enabling goal-based visual search. In contrast bottom-up saliency maps, due to their construction, are stable under different conditions and do not change based on the search context. Top-down guidance can be provided by features such as colour, orientation or shape (Wolfe & Horowitz, 2017). It has been found that object-based attention activates feature-based attention for the features of that object (Craven et al., 1999). Therefore, ideally a model of feature-based attention in visual search should be cued by the target class itself rather than the features of that class. An influential model of feature attention, the feature similarity gain model (FSGM) posits that a neuron's activity is multiplicatively scaled up (or down) according to how much the neuron preferentially activates (or does not activate) in response to the target stimulus (Treue & Martinez Trujillo, 1999). For example, if one was looking for a round, green bowl, the FSGM says that a top-down, goal-driven attention mechanism in the brain enhances activation of neurons in the visual cortex that respond preferentially to "circle" or "green" features. Conversely, neurons that do not respond preferentially to those features (such as those for "red" or "square" features) would have their activities suppressed. Evidence for such signals has been found in FEF during visual search (Zhou & Desimone, 2011).

More recently, many top-down attention models for visual tasks have used a similar idea, dynamically reweighting CNN feature maps in a downstream-dependent manner. In these top-down models, pre-trained CNNs were used in conjunction with feature biasing to successfully produce fixation behaviour that mirrors human performance on visual search on natural images (Miconi et al., 2016; Zhang et al., 2018; Lindsay, 2020a, b). The degree of feature bias, or feature attention strength, has been shown to control the classification performance of a pre-trained CNN in simple target detection paradigm (Lindsay & Miller, 2018) and over 1000 image classes (Luo et al., 2021).

Dynamic reweighting of CNN feature map schemes has also been successfully used on combined language processing and image recognition tasks to guide text generation (Chen & Perona, 2017; de Vries et al., 2017)

Methods

The goal of the proposed model is to efficiently locate the target class image within an array of sub-images that makes up the search image. The model takes the search image and target class as inputs and determines the location of the target class object in the search image. The general architecture and pipeline of the model are shown in Fig. 1.

In setting up the model, we aimed to stay consistent with the feature attention model in Lindsay and Miller (2018), which performed object classification on single images. The natural extension to a visual search task is to consider an array of images as a scene, with a search that iterates over single images in the array until the target image class is found. Whilst this extension is simpler than visual search within natural scenes, where much research has focused, it is a representative visual search task that enables investigation of decision theoretic aspects, such as decision speed versus accuracy.

The search images are generated from the CIFAR-10 imageset (MIT), which consists of 60,000 natural colour images that are split, equally, into 10 classes (Krizhevsky, 2009). In our study, the natural CIFAR-10 images are combined to create synthetic three-by-three search arrays. The sub-images within the array are chosen so that no more than one instance of the same class appears in each array, meaning nine of the 10 classes are present in each array but are otherwise randomly selected. An example synthetic search image is shown in Fig. 1.

The model implements two attention mechanisms: one for classification and one for the fixation policy. In the following two subsections, we detail the two attention mechanism implementations starting with details of the goal-based fixation policy of the model.

Goal-Based Fixation Priority Map of Search Image

To determine the fixation order for the target detector, a fixation priority map of the search image is generated, which depends on the target class. To calculate this map, the model generates feature maps derived from the relevant search and target class images using a CNN that has been pre-trained on the CIFAR-10 imageset. The network was trained using Adam optimizer (lr = 0.001) with a cross-entropy loss function and a batch size of 100. The most salient features of each image are then extracted for a meaningful spatial comparison to guide where best to direct attention for the target detection step. The CNN architecture comprises of five convolutional layers with max pooling between layers two and three and is shown in Fig. 1.

The Structural Similarity (SSIM) index is a commonly used perceptual metric designed to assess the similarity of two images in line with human judgement (Wang et al., 2003). It mimics the functionality of the human visual system in that it is sensitive to image luminosity in a manner consistent with Weber's law and is sensitive to structural variation but not non-geometric distortions, such as noise or blurring, in a way that traditional measures like l^2-norm are not. The model uses SSIM to quantify the saliency of a sub-image for a given target by calculating the SSIM between the sub-image and target image feature maps. Applying the SSIM to the CNN-extracted feature maps reduces the SSIM sensitivity to spatial variation due to the convolution and max pooling operations of the network. Although the SSIM is designed to

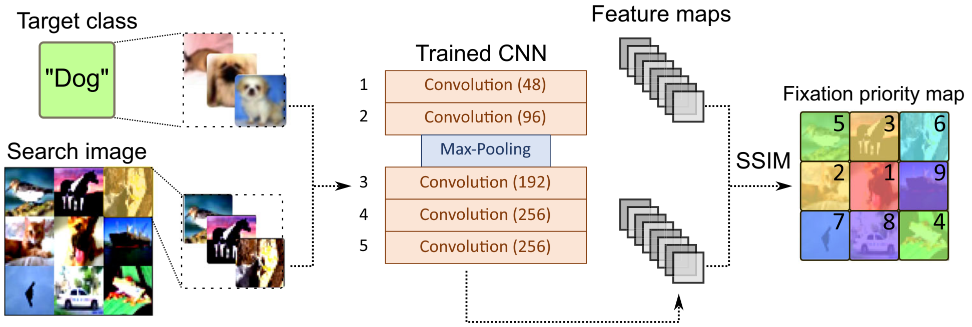

Fig. 1 Schematic of the model pipeline for generating fixation priority maps according to the fixation policy method. The numbers in brackets at each convolution layer represent the number of feature maps in that layer. The numbers on the fixation priority map represent the fixation order determined by the model

measure the structural similarity between images, in this case it can also be used to quantify the shared structure between two feature maps. Indeed feature maps are often represented visually as 2D grids of unit activations (where activations take the place of pixel values), and the SSIM can be applied to these representations to assess the similarity between two feature maps. SSIM values range between 0 and 1, where a SSIM of 1 indicates that the two compared images are identical. The SSIM of feature maps, x and y, corresponding to the search image and target image, is calculated using three functions, $l(x, y)$, $c(x, y)$ and $s(x, y)$, that compare the luminance, contrast and structure, respectively, of the two feature maps:

$$l(x, y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1}, \quad c(x, y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2}, \\ s(x, y) = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3}, \tag{1}$$

where $\mu_{x,y}$ and $\sigma_{x,y}$ are the mean and standard deviation of the pixel values in the two feature maps and the constants c_1, c_2, and c_3 (which depend on the range of the pixel values) ensure that the denominator is positive. The overall SSIM score between feature maps x and y is given by $\mathrm{SSIM}(x, y) = l(x, y)^\alpha \cdot c(x, y)^\beta \cdot s(x, y)^\gamma$, where α, β and γ weight the contribution of the comparison functions. We attribute no importance to one comparison function over another since doing so implies some assumptions on the characteristics of the images being compared, whereas we want the fixation policy method to be general. Equal weighting is therefore assumed so $\alpha = \beta = \gamma = 1$, giving the following expression for the SSIM calculation:

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}. \tag{2}$$

To capture the similarity of each sub-image to the target *class* rather than to a particular target *image*, fixation priority map scores are calculated by averaging the SSIM over 50 randomly selected images from the target class. For a search array of nine sub-images, the final fixation priority map is a three-by-three real-valued array with each element corresponding to the similarity between the target class and the sub-image in that position in the search array.

The fixation policy is "winner-takes-all" with infinite "inhibition-of-return". Fixations are determined by ordering SSIM values from the fixation priority map from highest to lowest and fixating to the corresponding sub-images in that order. A sub-image that has already been visited is never fixated upon again, implementing infinite inhibition-of-return.

The target detection step uses the same CNN used to generate the fixation priority map with three fully connected ReLU layers for classification. The convolutional layers of the CNN are patched with a feature-based attention mechanism, based on the FSGM (Treue & Martinez Trujillo, 1999). The model uses a mathematical implementation of the FSGM, first developed in Lindsay and Miller (2018), to dynamically reweight the network according to the visual search target class. Reweighting of the network is achieved by applying tuning values to each feature map in the convolutional layers. The same tuning value is used for all units in a feature map because feature attention has been shown to act uniformly over space (Zhang & Luck, 2009).

Calculation of Tuning Values The k^{th} feature map in the l^{th} layer is supposed to have an associated tuning value for the target class c and is denoted f_c^{lk}. Tuning values are calculated using:

$$f_c^{lk} = \frac{\frac{1}{N_c}\sum_{n\in c} r^{lk}(n) - \bar{r}^{lk}}{\sqrt{\frac{1}{N}\sum_{n=1}^{N}(r^{lk}(n) - \bar{r}^{lk})^2}}, \tag{3}$$

where $r^{lk}(n)$ is the average activity of the k^{th} feature map in the l^{th} layer in response to image n, N_c is the total number of images in class c shown during training and $\bar{r}^{lk} = \frac{1}{N}\sum_{n=1}^{N} r^{lk}(n)$ is the activity of the feature map averaged over all images where N is the total number of images (all classes) in the training set. Essentially, the tuning value, f_c^{lk}, is the expected deviation from the mean of the feature map's response to target class c, in units of standard deviation, and can be interpreted as a measure of how much the feature map preferentially activates in response to the target class, c.

Implementation of Dynamic Reweighting Through Tuning Values The network is cued by a target through dynamic reweighting of the convolutional layers using the tuning values for that target's class, c. All layers in the network are subject to feature-based modulation by the tuning values. Dynamic reweighting is implemented by scaling the unit activations x_{ij}^{lk}, by the tuning values, f_c^{lk}, for their feature map according to:

$$\hat{x}_{ij}^{lk} = \left(1 + \beta f_c^{lk}\right)\left[I_{ij}^{lk}\right]_+, \tag{4}$$

where β is a parameter controlling the strength of the attention and $\left[I_{ij}^{lk}\right]_+$ is the positive rectified input to that unit from layer $(l - 1)$. The net effect is to amplify responses of feature maps that preferentially respond to the target class (and hence the characteristic features of the target class) whilst inhibiting the response of feature maps that do not respond strongly to the target class (thereby suppressing feature signals in the model considered irrelevant to the target of the visual search). This framework implements the FSGM since it posits that cuing by a target stimulus prompts a similar

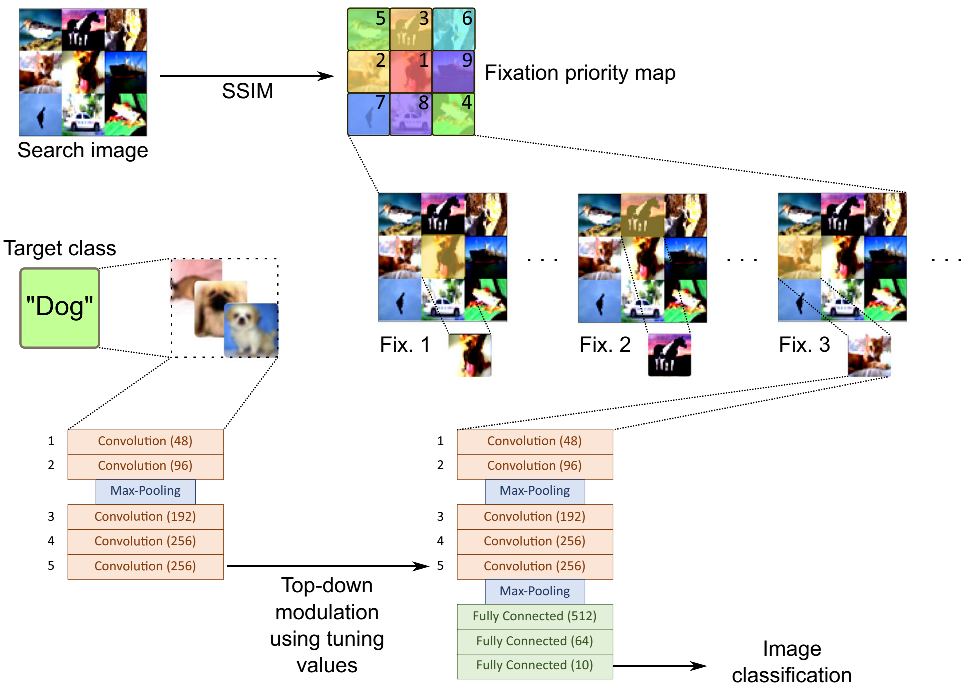

Fig. 2 Schematic of model pipeline for target detection step with feature-based attention mechanism applied. Sub-images from the search image are presented in turn to the CNN acting as target detector according to the fixation policy. The network is cued by the target using a feature attention mechanism, which is implemented by applying previously learned tuning values for that target class to dynamically reweight feature maps during detection step

process in the visual system of the brain via top-down modulation of signals in the visual system (Fig. 2).

Results

We evaluated the performance of the fixation policy by estimating the likelihood of finding the target by each fixation number during the search. Estimates were calculated from 500 visual search trials on randomly generated search images that all contained one target image. To assess performance of the fixation policy in isolation from the target detection step, we consulted an “oracle” at each fixation, that is, we simply checked the sub-image and target class labels for a match (using label information not available to the model). We used the cumulative performance as a function of the fixation number to gauge effectiveness of the policy. Cumulative performance was defined as the proportion of searches that had found the target by that fixation number.

We assessed performance of the fixation policy against a baseline random strategy for eye movement to demonstrate the importance of top-down guidance for target search (Fig. 3). When applied to Layer 5 feature maps, the fixation

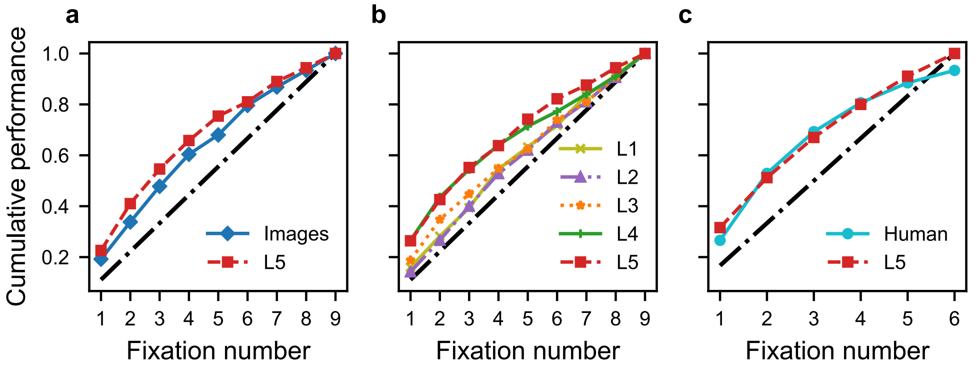

Fig. 3 Evaluation of fixation policy by plotting the cumulative performance as a function of fixation number. Cumulative performance is the estimated probability of having fixated to the target class by that fixation number (500 trials used for estimate). (**a**) Fixation policy based on Layer 5 feature maps (red curve) outperforms policy that uses raw images (blue curve). Both methods perform better than the baseline random strategy (dashed black line). (**b**) Fixation policies based on feature maps from later layers in the network perform better than those from earlier layers. (**c**) Fixation policy performance (using Layer 5 feature maps) (red curve) resembles human data (light blue curve) from target-cued visual search experiments performed in Zhang et al. (2018) (Mean squared error = 0.0014)

policy resulted in more than half of all searches concluding within three fixations; however, a random fixation policy would need five fixations to reach the same level of performance (Fig. 3a). To measure the effect of feature extraction by the CNN on the performance of the fixation policy we removed image preprocessing by the CNN from the method and used the SSIM comparison on the raw target and search images. We found that applying the SSIM comparison to the raw images rather than the feature maps degraded the performance of the fixation policy markedly at earlier fixations (Fig. 3a, blue curve), suggesting that feature maps result in more effective fixation priority maps. Both the SSIM and raw image fixation policies have perfect performance by the ninth fixation because we simply required a target image be reached and the target class was always present in the search array.

To examine the effect of each successive feature map layer on the performance of the fixation policy, we used feature maps from convolutional layers 1 to 5 of the network to generate five corresponding fixation priority maps. We found that visual search was more efficient when later layers were used to generate the fixation policy, indicating that each successive feature extraction step in the CNN results in more useful fixation priority maps (Fig. 3b), as expected from a similar experiment carried out for the FSGM on classification performance (Lindsay & Miller, 2018). Each successive convolutional layer extracts more salient features (the underlying structural information from the images that is most useful to the classification layers) thus improving the performance of the SSIM when scoring feature maps from the same image class. At earlier layers, performance trends closer to random fixations because less salient features are being compared, such as the background colour.

Comparison of search performance with human performance on the same visual search task shows that fixation policy behaviour resembles human behaviour. We compared performance of the fixation policy on two-by-three search arrays to data from a human eye-tracking experiment where subjects were presented with an array of six images and instructed to find an image of the target class (Zhang et al., 2018) (Fig. 3c). Interestingly, more than half of searches using the SSIM fixation policy on this task, even when applied to Layer 5 feature maps, took longer than 3 fixations to find the target sub-image, in line with the behaviour demonstrated by human subjects. Overall, the model and human cumulative performance curves show a similar trend except that humans performed sub-optimally after six fixations, whereas the fixation policy had 100% cumulative performance by that same stage of the search. This is because the human subjects will occasionally return to a previously visited location in the search image whereas our model has an infinite inhibition-of-return mechanism. This means the model performance is perfect after as many fixations as there are sub-images in the search array, because the goal sub-image is always present.

Feature-Based Attention Controls the Expected Cost on Standalone Classification Tasks

To examine the effect of the feature attention mechanism on the target detector's performance, we carried out a single image classification task using just the attention-patched CNN, analogously to a prior investigation of the FSGM (Lindsay & Miller, 2018). The CNN was cued with a target class by dynamically reweighting the convolutional layers (see methods) and the model then had to determine if random images from the CIFAR-10 dataset represented that target class or not. We estimated the true positive rate (TPR) and false positive rate (FPR) at 20 different attention strengths by averaging the outcomes for each target class over 500 trials and then plotted those 20 datapoints to obtain a receiver operating characteristic (ROC) curve (Fig. 4a). We found that the attentional strength parameter, β, controlled the TPR/FPR ratio during target detection (Fig. 4a). Moderate increases in the attentional strength resulted in initially moderate increase in FPR and larger increase in TPR, but at larger attentional strengths, TPR actually trended downward whilst FPR increased dramatically (Fig. 4b). Equal changes in TPR and FPR due to feature attention are indicated by the diagonal dashed blue line in Fig. 4b. Points plotted above this line give a net benefit if costs of false positives and false negatives are equal, and points plotted below this line mean a net loss from applying the feature attention. Most points were found to lie below this line (Fig. 4b), which makes sense because the objective the network was trained on is equivalent to using an objective that assumes equal error costs for true and false positives. In what situations, then, is boosting feature attention advantageous?

To address this question, we then analysed the effect of the attentional strength on the expected cost of a single classification trial for asymmetric error costs. We defined the expected cost of the decision, $\bar{C} = C_0 + \sum C_\mathrm{m} P_\mathrm{m}$, where $m \in \{\mathrm{TP, FN, FP, TN}\}$ are the four possible outcomes of the decision, C_m is the cost of that outcome, P_m is the probability of that outcome and C_0 is the cost of a trial regardless of the outcome. Assuming costs only arise from errors ($C_0 = C_\mathrm{TP} = C_\mathrm{TN} = 0$), factorising the probabilities and rearranging gives the following cost function:

$$\bar{C}(\beta) = C_\mathrm{FP} P^- \,\mathrm{FPR}(\beta) - C_\mathrm{FN} P^+ \,\mathrm{TPR}(\beta) + C_\mathrm{FN} P^+, \quad (5)$$

where P^+ and P^- are the prior probabilities of "target" and "not target" stimulus conditions, respectively. This function is plotted in Fig. 4d for different cost ratios, $(C_\mathrm{FP}/C_\mathrm{FN}) = (0.01, 1, 10)$. We found that the attentional strength parameter, β, controlled the minimum of the cost function at different

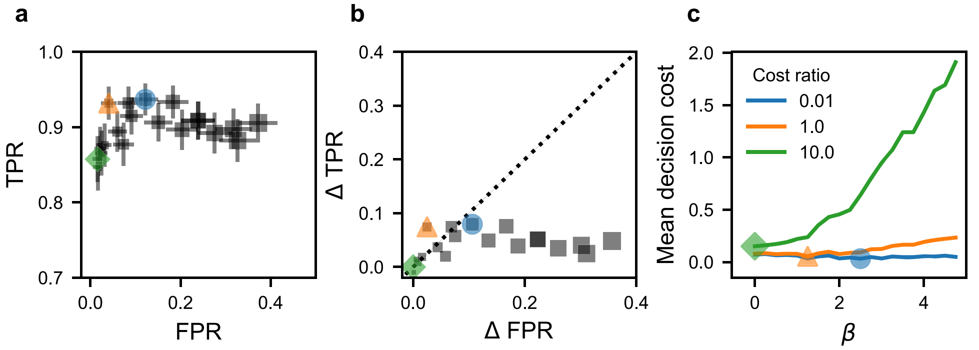

Fig. 4 Effect of feature attention on standalone classification task. (**a**) ROC curve for target detection model illustrating classification performance as attentional strength parameter β is varied. Error bars are 95% confidence intervals. Marker size corresponds to β value. Data points further along the FPR axis correspond to larger β values. (**b**) The change in TPR plotted against the change in FPR after feature-based attention application. Changes are relative to TPR and FPR when no feature attention modulation is applied ($\beta = 0$). Marker size corresponds to β value. (**c**) The cost function in Eq. 5 plotted as a function of the attentional strength parameter for different cost ratios. Triangle, circle and diamond markers indicate function minima. The same markers are used to highlight the corresponding coordinates in plots **a** and **b**

cost ratios, which suggested a control mechanism for minimising the expected cost of visual classification tasks. For example, if, in a given context, false positives were relatively harmless compared to false negatives, then the expected cost would be minimised by increasing the attentional strength given to features of the target.

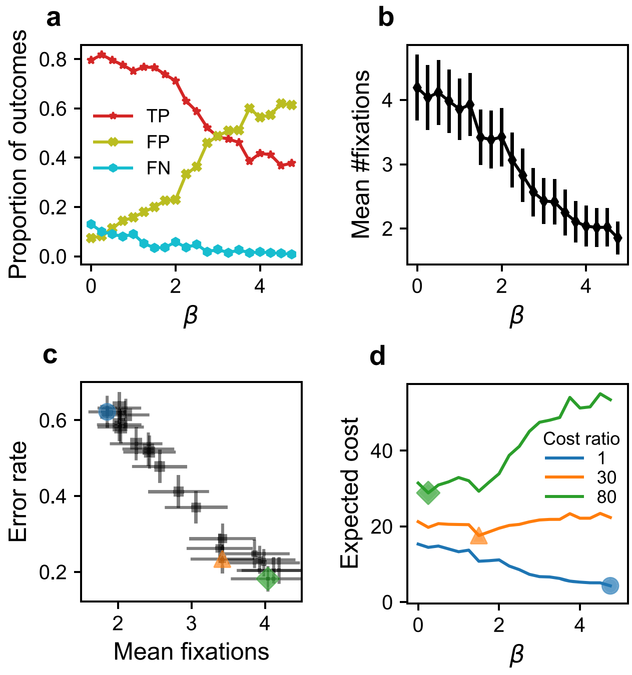

Fig. 5 Feature-based attention controls speed-accuracy trade-off in visual search. (**a**) Estimates of the probabilities of visual search outcomes (as a proportion of outcomes from 500 trials for each data point) for different values of the attentional strength parameter β. (**b**) Mean number of fixations before a search concludes plotted as a function of β. Higher feature attention strength results in faster searches. (**c**) The visual search error rate (combined false positives and false negatives as a proportion) plotted against the mean fixations for 20 values of β. Size of marker corresponds to β value. The relationship between these two metrics defines the speed-accuracy trade-off curve for the model. (**d**) Dependence of the cost function in Eq. 6 on attention strength β. Triangle, circle and diamond markers indicate function minima. The same markers are used to highlight the corresponding coordinates in plot **c**. All error bars are 95% confidence intervals

Feature-Based Attention Controls a Speed-Accuracy Trade-off in Visual Search

To examine how the two goal-based attention mechanisms interacted during a visual search task, we used the fixation policy and CNN target detector in combination on a visual search task (Fig. 2). For the fixation policy, we generated priority maps using feature maps from the final convolutional layer of the CNN to give the best performance (see Fig. 3b). For the target detection step, we implemented feature attention by applying tuning values to all feature maps in all convolutional layers. We classified the trial outcomes so that if a sub-image was correctly identified as the target, then that was regarded as a successful trial (true positive), but if a sub-image was incorrectly classified as the target, then that was regarded as an error (false positive). Another possible outcome was when the model did not detect the target at all, even once all sub-images had been fixated upon — in this case we stopped the search and took it as a false negative since the target was present in all search images.

The estimated probabilities of the three visual search outcomes (estimated from 500 visual search experiments) are plotted as functions of the attentional strength parameter β in Fig. 5a. As expected, the proportion of searches where the search concluded in a false positive increased as the attentional strength increased. Although increased attentional strength increased the likelihood of a positive classification on any given fixation, this actually had the effect of reducing the proportion of searches that concluded in a true positive. This reduction was also the case for false negatives, although to a lesser extent. Figure 5b clearly demonstrates that the average time to conclude a search decreases as the attentional strength increases due to the increased likelihood of a positive classification (thereby stopping the search) at each fixation.

Varying the attentional strength changes the error rate and the mean time taken to conclude the search (Fig. 5a, b). The relationship between these two metrics, the error rate and the mean fixation number, defines the speed-accuracy trade-off in the model (Fig. 5c). Data plotted in Fig. 5c shows that as the mean fixation number increases, the error rate decreases — longer searches tend to be more accurate, whereas faster searches tend to be more prone to error. Since the error rate and mean fixation number depend on the attentional strength, dynamically reweighting the CNN according to β is a control mechanism for the speed-accuracy trade-off in the visual search model.

We then analysed the effect of the attentional strength on the expected cost of a visual search trial. The cost function included a time cost $C(T) = \sum_{t=1}^{T} c(t)$, where $c(t)$ is the cost per unit time as a function of fixation number, t, and T is the total number of fixations in that search. We assumed that a true positive incurred no cost. The expected cost was calculated using:

$$\bar{C}(\beta) = C_{\mathrm{FP}}\,\mathrm{FPR}(\beta) + C_{\mathrm{FN}}\,\mathrm{FNR}(\beta) + \langle C(T)\rangle(\beta), \qquad (6)$$

The expected time cost is given by $\langle C(T)\rangle = \sum_k \sum_{t=1}^{T_k} c(t)p(T_k)$, where $p(T)$ is the probability of the search concluding after T fixations, which was estimated from the visual search experiments. We assumed a linear dependence on time for the cost per unit time, $c(t) = \alpha t$, meaning $C(T)$ is quadratic in T, which is consistent with cost per unit time functions that have been inferred from evidence accumulation experiments in monkeys (Drugowitsch et al., 2012).

The expected cost function (Eq. 6) is plotted in Fig. 5d for cost ratios, $(C_{\mathrm{FP}}/C_{\mathrm{FN}}) = (1, 30, 80)$, where we have assumed cost-per-unit-time parameter, $\alpha = 1$. Again, the attentional strength parameter, β, controlled the minimum of the cost function. Crucially if the cost of a false positive was low, attentional strength on the target features needed to be high to minimise the search time. However, if false positives came at higher cost, then attentional strength should be lower so there was a higher likelihood of a true positive but also a higher likelihood of a longer search. Therefore, target feature attention can optimise the balance of speed and accuracy to minimise the expected cost of a visual search.

Discussion

We have shown that a feature attention mechanism previously only applied to standalone classification tasks (Lindsay & Miller, 2018) can, in combination with a fixation policy, be extended to visual search tasks. Furthermore, the feature attention confers the model with a mechanism for controlling the speed-accuracy trade-off for the search. The relevant parameter is the attentional strength on target features: higher attentional strength leads to faster searches but lower accuracy, and, conversely, lower attentional strength leads to longer searches but higher accuracy.

Similar speed-accuracy trade-offs are seen in decision-making models, such as the drift diffusion model (DDM), where thresholds on integrated evidence control the balance of speed and accuracy (Bogacz et al., 2006; Gold & Shadlen, 2007; Griffith et al., 2021). These models usually appear within a Bayesian framework. In our model, the target cuing, which leads to dynamic reweighting of CNN in the model, could be viewed as modulating the prior for the target features. Attentional mechanisms have previously been considered as a potential candidate for modifying priors and resolving uncertainty in a perception-as-inference context (Rao, 2005) and in a Bayesian neural architecture (Yu & Dayan, 2005). Optimal speed-accuracy trade-offs have been found using reinforcement learning on decision model parameters (Lepora, 2016; Pedersen & Frank, 2020); a similar approach for visual search tasks could be applied to the attentional strength parameter in our model. However, our visual search model differs from models like the DDM in that there is no integration of information favouring the target object. This would be interesting to explore in future experimental and computational neuroscience work, for example by examining how previous fixations can affect the saliency map.

Our work builds on several previous studies: implementation of the FSGM by dynamic reweighting of a CNN (Lindsay & Miller, 2018; Luo et al., 2021), and a top-down modulated fixation policy that uses CNN-derived feature maps as a basis (Miconi et al., 2016; Zhang et al., 2018). The main contribution of this paper is proposing the combination of these components and successfully demonstrating the modulatory effects possible on performance in visual search. According to this view, increases in attention on a given set of features (directed by the context-dependent goal of the search) would result in higher error rates when distracting objects have similar features to the target. This phenomenon is evident in the popular book "Where's Waldo", where the visual scene is cluttered with Waldo-like features, leading to difficulty in finding the target of the search. The degree of attention given to features balances the speed and accuracy of the search, and so can be seen as representing the importance or urgency with which the target object is being sought.

In addition, we take the novel perspective of using decision theoretic approach to assess the costs and benefits of the top-down feature attention mechanism that dynamically reweights the target detector CNN. This decision theoretic perspective considers the possibility of asymmetric error costs and also prices in the cost of the search time. As such, it reveals the benefit of using feature attention on search images from an image set that the neural network has already been trained on, rather than using a pre-trained, "off-the-shelf" network, that is often the starting point for other studies.

Our visual search task of finding an image within an array was appropriate for extending the feature attention model in Lindsay and Miller (2018) to visual search. An important future direction would be to extend the model to naturalistic scenes. This could be achieved by using a foveation mechanism whereby a window is moved within the scene to search for an object on a natural background. This raises challenges, such as the appropriate choice of neural network classifier and how to introduce invariance to how an object is viewed. Another improvement could be to adopt an inhibition-of-return model to replicate human fixation behaviour where repeat visits are made to the same image locations. It would also be interesting to apply the feature attention model to those experimental tasks used to study crowding, a phenomenon observed in humans where target detection is compromised by nearby distractors that share features with the target (Manassi et al., 2013; Herzog, 2015). This could be a further avenue for exploring the model's ability to replicate human behaviour on visual tasks. In our view, the way forward is to combine the advances made in other models (as discussed in the background) with the feature attention and decision theoretic model introduced here.

In conclusion, we have proposed a model that offers a method for optimising the costs of visual search and demonstrated the value of a decision theoretic approach that gives a new way of modelling complex visual tasks involving attention.

Author Contributions All authors contributed to the study conception and design. Computational simulation and data collection and analysis were performed by Florence J. Townend and Thom Griffith. The first draft of the manuscript was written by Thom Griffith and Florence J. Townend and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This work was supported by a Leverhulme Trust Research Leadership Award (RL-2016-039) to Prof. Nathan Lepora.

Data Availability The datasets and code generated during and/or analysed during the current study are available from https://bitbucket.org/leporalab/2022-cbb-feature-attention.

Declarations

Ethics Approval Not applicable

Consent to Participate Not applicable

Consent for Publication Not applicable

Conflict of Interest The authors declare no competing interests.

References

Akbas, E., & Eckstein, M. P. (2017). Object detection through search with a foveated visual system. *PLoS Computational Biology, 13*(10)

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review, 113*(4), 700–765.

Chen, B., Navalpakkam, V. & Perona, P. (2011). Predicting response time and error rates in visual search. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011, NIPS 2011*.

Chen, B., & Perona, P. (2017). Speed versus accuracy in visual search: Optimal performance and neural implementations. In Zhao, Q. (ed.), *Computational and Cognitive Neuroscience of Vision. Cognitive Science and Technology,* pp. 105–140. Springer, Singapore.

Craven, K. M. O., Downing, P. E., & Kanwisher, N. (1999). fMRI evidence for objects as the units of attentional selection. *Nature, 401*, 584–587.

Dai, J., He, K., & Sun, J. (2016). Instance-aware semantic segmentation via multi-task network cascades. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 3150–3158.

de Vries, H., Strub, F., Mary, J., Larochelle, H., Pietquin, O., & Courville, A. (2017). Modulating early visual processing by language. arXiv:1707.00683

Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N., & Pouget, A. (2012). The cost of accumulating evidence in perceptual decision making. *Journal of Neuroscience, 32*(11), 3612–3628.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience, 30*(1), 535–574.

Goldberg, M. E., Bisley, J. W., Powell, K. D., & Gottlieb, J. (2006). Saccades, salience and attention: The role of the lateral intraparietal area in visual behavior. *Progress in Brain Research, 155*, 157–175.

Gottlieb, J. P., Kusunoki, M., & Goldberg, M. E. (1998). The representation of visual salience in Monkey Parietal Cortex. *Nature, 391*, 481–484.

Griffith, T., Baker, S.-A., & Lepora, N. F. (2021). The statistics of optimal decision making: Exploring the relationship between signal detection theory and sequential analysis. *Journal of Mathematical Psychology, 103*, 102544.

Hariharan, B., Arbeláez, P., Girshick, R., & Malik, J. (2014). Simultaneous detection and segmentation. In Fleet, D., Arbeláez, P., Girschick, R., & Tuytelaars, T. (eds.) *Computer Vision - ECCN 2014. ECCV 2014. Lecture Notes in Computer Science*, vol 8695, pp. 297–312. Cham, Springer International Publishing.

Herzog, M. H., Sayim, B., Chicherov, V., & Manassi, M. (2015). Crowding, grouping, and object recognition: A matter of appearance. *J Vis, 15*, 1–18.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience, 2*(February), 1–11.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(11), 1254–1259.

Koch, C. & Ullman, S. (1987). Shifts in selective visual attention: towards the underlying neural circuitry. In Vaina, L.M. (ed.) *Matters of Intelligence. Synthese Library,* vol 188, pp. 115–141. Springer, Dordrecht.

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images.

Lepora, N. F. (2016). Threshold learning for optimal decision making. *Nips*, 3756–3764.

Li, Z. (2002). A saliency map in primary visual cortex. *Trends in Cognitive Sciences, 6*(1), 9–16.

Lindsay, G. W. (2020a). Attention in psychology, neuroscience, and machine learning. *Frontiers in Computational Neuroscience, 14*, 1–21.

Lindsay, G. W. (2020b). Convolutional neural networks as a model of the visual system: past, present, and future. *Journal of Cognitive Neuroscience*, (Feb), 1–15.

Lindsay, G. W. & Miller, K. D. (2018). How biological attention mechanisms improve task performance in a large-scale visual system model. *eLife*, *7*, 1–29.

Luo, X., Roads, B. D., & Love, B. C. (2021). The costs and benefits of goal-directed attention in deep convolutional neural networks. *Computational Brain & Behavior*.

Manassi, M., Sayim, B., & Herzog, M. H. (2013). When crowding of crowding leads to uncrowding. *Journal of Vision, 13*(10), 1–10.

Miconi, T., Groomes, L., & Kreiman, G. (2016). There's waldo a normalization model of visual search predicts single-trial human fixations in an object search task. *Cerebral Cortex, 26*(7), 3064–3082.

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature, 434*(7031), 387–391.

Najemnik, J., & Geisler, W. S. (2009). Simple summation rule for optimal fixation selection in visual search. *Vision Research, 49*(10), 1286–1294.

Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research, 45*, 205–231.

Pedersen, M. L. & Frank, M. J. (2020). Simultaneous hierarchical Bayesian parameter estimation for reinforcement learning and drift diffusion models: a tutorial and links to neural data. *Computational Brain & Behavior*, *3*, 458–471.

Rao, R. P. (2005). Bayesian inference and attentional modulation in the visual cortex. *NeuroReport, 16*(16), 1843–1848.

Rashidi, S., Ehinger, K. A., Turpin, A., & Kulik, L. (2020). Optimal visual search based on a model of target detectability in natural images. In *34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.*

Renninger, L. W., Coughlan, J., Verghese, P., & Malik, J. (2004). An information maximization model of eye movements. *Advances in Neural Information Processing Systems, 17*, 1121–1128.

Treue, S., & Martinez Trujillo, J. C. (1999). Feature-based attention influences motion processing gain in Macaque Visual Cortex. *Nature, 399*, 575–579.

Verghese, P. (2001). Visual search and attention: A signal detection theory approach. *Neuron, 31*(4), 523–535.

Wang, Z., Simoncelli, E. P. & Bovik, A. C. (2003). Multi-scale structural similarity for image quality assessment. In *The Thirty-seventh Asilomar Conference on Signals, Systems & Computers*, pp. 1398–1402

Wolfe, J. M., & Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nature Human Behaviour, 1*(3), 1–8.

Yu, A. J. & Dayan, P. (2005). Inference, attention, and decision in a Bayesian neural architecture. In *Advances in Neural Information Processing Systems 17 (NIPS 2004)*.

Zelinsky, G. J. (2008). A theory of eye movements during target acquisition. *Psychological Review, 115*(4), 787–835.

Zelinsky, G. J., Zhang, W., Yu, B., Chen, X., & Samaras, D. (2005). The role of top-down and bottom-up processes in guiding eye movements during visual search. *Advances in Neural Information Processing Systems 18*.

Zhang, M., Feng, J., Ma, K. T., Lim, J. H., Zhao, Q., & Kreiman, G. (2018). Finding any Waldo with zero-shot invariant and efficient visual search. *Nature Communications*, *9*(1).

Zhang, W., & Luck, S. J. (2009). Feature-based attention modulates feedforward visual processing. *Nature Neuroscience, 12*(1), 24–25.

Zhou, H., & Desimone, R. (2011). Feature-based attention in the frontal eye field and area V4 during visual search. *Neuron, 70*(6), 1205–1217.

Zou, Z., Chen, K., Shi, Z., Guo, Y., & Ye, J. (2023). Object detection in 20 years: A survey. arXiv:1905.05055

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.