UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science Curriculum

Tarun Khajuria

A unified account of visual search using a computational model

Master's Thesis (30 ECTS)

Supervisor: Raul Vicente, PhD
Supervisor: Jaan Aru, PhD

Tartu 2020

A unified account of visual search using a computational model.

Abstract:

Visual Search is a task ubiquitously performed by humans in everyday life. In the laboratory, to understand more about this process, experiments have characterised the time that humans need to locate a particular target object amongst others. Based on this search time's dependence on the number of objects in the image, it is believed that two kinds of search take place. Feature search, where the target pops-out of the search image and is instantly found using a parallel search mechanism, and conjunction search, with more complex objects where the search is serial and the search time increases with the number of objects. In this work, we use a computational model to propose a unified process that can result in feature or conjunction search characteristics depending on the precision of the attention guidance mechanism. We show that the search performance can be partly explained by the precision or capacity of the encoding of distinct features that is used to guide attention during the search process.

Keywords:
Visual Search, Attention, Computational Neuroscience, Deep Learning, Convolutional Neural Networks.

CERCS:
P176 - Artificial Intelligence , S260 - Psychology, P170 - Computer science, numerical analysis, systems, control, B640 - Neurology, neuropsychology, neurophysiology

Ühtne raamistik visuaalse otsingu mõistmiseks arvutusliku mudeli abil.

Lühikokkuvõte:

Visuaalne otsing on tegevus, mis omab keskset rolli ka igapäevaelus. Laboris selle protsessi kohta lisateabe saamiseks tehtud katsed on iseloomustanud aega, mis inimestel on vaja segavate objektide hulgas sihtobjekti leidmiseks. Selle otsinguaja seos pildil olevate objektide arvuga on aluseks arvamusele, et toimub kahte tüüpi otsing. Esiteks üksiktunnuse otsing, kus sihtmärk hüppab välja otsingupilt ja leitakse kohe, kasutades paralleelset otsingumehhanismi, ja teiseks konjunktsiooniotsing keerukamate objektidega, kus otsing on seriaalne ja otsinguaeg suureneb objektide arvuga. Selles töös kasutame arvutuslikku mudelit, et pakkuda välja ühtne protsess, mille tulemuseks võivad olla üksiktunnuse või konjunktsiooniotsingu omadused sõltuvalt tähelepanu juhtimismehhanismi täpsusest. Näitame, et otsingutulemusi saab osaliselt seletada eraldiseisvate tunnuste kodeerimise täpsuse või mahuga, mida kasutatakse otsingu ajal tähelepanu juhtimiseks.

Võtmesõnad:
visuaalne otsing, tähelepanu, arvutuslik neuroteadus, sügavõpe, konvolutsioonilised närvivõrgud

CERCS:
P176 - Tehisintellekt , S260 - Psühholoogia, P170 - Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine (automaatjuhtimisteooria), B640 - Neuroloogia, neuropsühholoogia, neurofüsioloogia

Contents

1 Introduction

Visual search for an object cluttered in between other objects is a task essential for everyday living [21, 22]. Looking for a book on a library shelf, remote of television in the house or the character Waldo in the finding waldo game image are all examples of this everyday process. All these tasks involve remembering the object to be searched and then matching it with objects as we scan through visual images. This process is called **visual search** [38, 42].

The guidance of attention is considered to be the central part of the visual search process [38, 42, 40]. Attention is considered to be guided by both salient features in the scene as well as by the current objective of the search [34, 10, 16].

The guidance mechanisms of visual attention in humans have been thoroughly studied (e.g. Yarbus, 1967 [44]; Itti and Koch, 2001 [15]; Tatler et al., 2005 [32]; Wolfe and Utochkin,2019 [43]). The main way of studying this process has been with the task for subjects to find a particular object called **target** amongst other objects called **distractors** and record the search time for various kinds of target and distractor shapes [7, 43]. In these experiments, it has been observed that simple target shapes that vary from the distractors in simple ways like orientation or color are easy to find [38]. They **pop-out** (like words in bold do) i.e. they are easily noticed amongst distractors and the search time does not change irrespective of the number of distractors on the search image. This condition is called **feature search**. When the objects are more complex and are varying in multiple dimensions in a search image, the search time is linearly dependent on the number of distractors [36]. This condition is called **conjunction search**.

Visual Search has been computationally modelled based on various methods (eg. Itti and Koch, 2001 [15]; Torralba et al., 2006 [35]). With advances in deep learning, artificial neural networks (ANNs) are used in computational modelling for many cognitive processes [1, 17]. High-level explanations for various processes involving vision are based on a class of deep learning models called Convolutional Neural Networks (CNNs). In CNNs, the feature extraction ranges from simpler features in earlier layers to complex features that are aggregated from the simpler features in the later layers. The resemblance of feature extraction to the human visual cortex enables us to use CNN as computational models of visual processing and other possible dynamics of the human visual system [24].

In this work, we build on the computational model by Zhang and colleagues [45] and use it to explain dynamics in visual search when comparing feature and conjunction search. Different variations of the model are considered that mimic resource constraints during

the search. Overall we propose that important characteristics of feature and conjunction search can be explained using a single process based on the effective correctness of attention guidance mechanism. In particular, we explore the importance of feature representation (encoding) and feature similarity between search objects and how they affect the guidance of attention.

2 Background

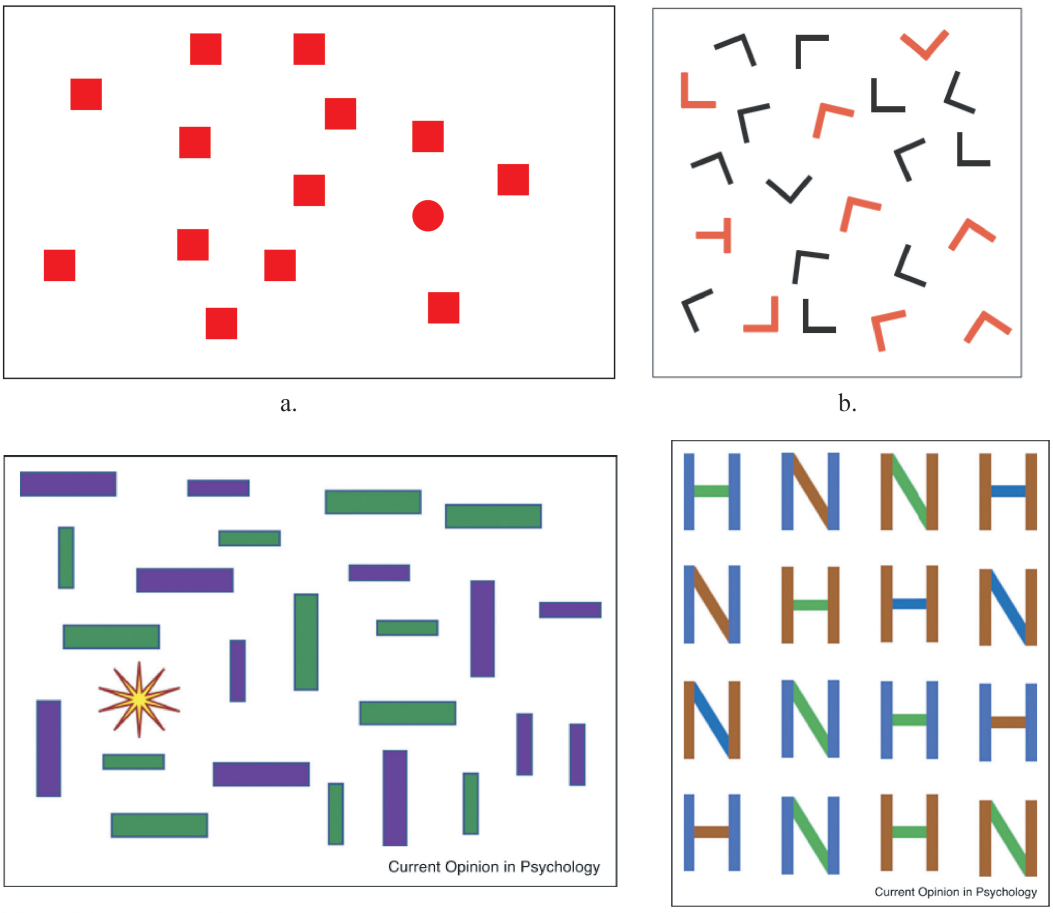
Visual search of objects has been studied in the laboratory setting for a long time [44, 41]. Most of the emphasis of studying visual search of objects has been on understanding the process and attributes that make something be found amongst other objects. Primarily, the experiments are conducted with human subjects in the laboratory. In these experiments, the subjects are shown some images on a screen. In these images, they are instructed to search for an object (the target) which is place amongst other objects (distractors). Figure 1 shows some examples of such search images. Some of the search images contain the target object and in others, the target is not present. The subject has to report when they find the target. The time taken by the subject to find the target is recorded as the **search time**. The absence of the target in some search images allows for verifying the overall correctness of the subject's response.

With new advances in technology more nuanced readings are taken of the subject's response during a search task. Head-mounted cameras and other new devices are used to record the gaze fixations of the subject with high precision during the experiment. Availability of gaze pattern data using these tools has allowed a better understanding of the gaze process and helped in better probing of theories of visual search [33, 39].

2.1 Visual Attention

Attention is a concept that appears in slightly different ways in many fields of study such as psychology and machine learning. But it roughly points to a process of control of limited resources [4, 3, 24]. Visual attention is a process that guides the allocation of resources used in tasks of vision. One particular example of this is the control of a limited area on the retina called the fovea. The fovea has a higher image resolution than the rest of the retina [24]. Control of eye gaze to select what part of the scene to attend to using fovea is an important part of the visual process. Primarily, humans control their gaze to point to important parts of the scene to be processed by the fovea. The implication of this for the process of perception is that more information can be extracted from this region.

The selection of points of gaze fixation in a scene is a much-studied topic. An aspect that has been widely debated is the relative contribution of feature salience in the scene and the search target, in the decision towards the guidance of foveal attention [12]. The salient points are defined in the scene based on extreme values of features such as intensity, color contrast, orientation and motion [15]. It is considered that in the absence of a target i.e. during 'free viewing', gaze is guided towards such salient points in the scene. Other research supports this notion by pointing towards a correlation between the salient



a. b.

Find a big green vertical target. c.

Find the two brown and green ‘N’s. d.

Figure 1. Search Images with a) circle as target b) red T as target [18] c) big green vertical target [43] d) Two brown and green Ns as target [43].

points in the scene and the actual points of fixation while viewing an image [28]. On the other side, a major counter-argument is that the predictive power of lower-level salience based fixation selection is low [32] and the found correlations cannot be interpreted as causation [14]. Further, a case for a greater role of target based selection of gaze position is that most objects of interest (targets) during the search, co-incidentally have salient features [11]. Another aspect to consider is that during viewing experiences the subject may randomly choose some target or bias and the explicit ‘free viewing’ experience results may be affected by this choice [22].

2.2 Memory and Attention

Memory plays an important role in attention, as it is one of the resources that is controlled by attention. In other cases, memory-based constraints play an important role in how attention is controlled [24].

Limitation of working memory has been modelled by Ma et al, 2014 [25]. **In this model, memory limitation causes a relative decrease in the precision of stored representations of the entities as the number of entities increase.** A choice of important features to attend to and to store is made during such conditions.

Another particular role that memory plays in visual attention is to remember the places that have been previously visited by attention or eye-movements [15]. These places that have previously been visited but have a high salience or target based activation score need to be remembered during a visual search process so that the attention does not keep getting allotted back to these points and the visual search process is efficient. This phenomenon is called **inhibition of return**.

2.3 Visual Search and Attention

Visual search is historically believed to take place either as a serial or parallel process. In serial search, attention is paid to one object at a time [9], where parallel search allows for parallel processing of multiple objects at a time. This processing at each time also determines the presence of the target in the searched objects [26]. If the target and the distractors differ in a single dimension, the target is observed to pop-out, due to a quick bottom-up feature search conducted in parallel in lower levels of the visual process. Feature integration theory (FIT) [38] points out that some features are processed in parallel across all objects. These features are called **preattentive features** and they shall be simple features that cannot be broken down into any component features. The search when driven by these features is quick and results in the pop-out effect. A limiting condition is that, within a single feature, the difference between the search objects shall be large enough to be easily distinguished for the pop-out effect to occur [8].

In case of a conjunction of features in the objects, the features of an object have to be integrated before processing and linear search is believed to take place, hence causing the search time to increase with the number of distractors.

Guided search model modifies FIT to propose that in case of both feature and conjunction search, pre-attentive guidance of attention takes place. The difference

in the search time comes from the correctness of attention guidance, based on the relative presence of guiding features [2]. After each gaze fixation, the search process is continued in case the target is not found or it is stopped if the target is found. In case of continuation of the search, the next position of gaze is selected based on the pre-attention selection process.

In such experiments, the search time is recorded for each trial as the measure of search performance. The slope of the line obtained by plotting search time plotted as function of number of distractors, is an important indicator of the search efficiency, with a lower slope indicating a more efficient search.

2.4 Computational Models

Attention in visual search has been modeled using various computational models (eg. Itti and Koch, 2001 [15]; Torralba et al., 2006 [35]; Ehinger et al., 2009 [10]). Primarily, bottom-up feature-based salience maps were modelled by various methods [15]. With the growing evidence that attention is guided by the interaction of bottom-up salience and top-down target, one such computational model proposed was to modify the salience maps using the output of an object detector [10, 16]. Olivia and Torralba, 2006 [27] proposed a model that uses a global feature detector to mark the likely location of fixation in the scene according to prior experience and the high-level target of the search. This information about prior experience and high level target, encodes the context for the search. In Figure 2, it can be seen how this context is combined with the low-level local feature detector's output salience map to generate a final attention map (white spots point to positions with higher attention score). A simple example of this can be that while looking for a pan in the kitchen, the salience map is modified such that the points in the middle area and top get more score than the ground as they have a higher probability to be the place where a pan might be located.

Another model of visual attention was proposed by Zang et al., 2018 [45], which uses a pre-trained Artificial Neural Network to extract the features from the search image. Figure 3 describes their model, where higher-level features are extracted from a CNN. As the features are extracted from higher layers, the spatial dimension of the feature map is greatly reduced from the original image. An attention map to determine the gaze order is obtained by convolution of the top-level target feature vector on the extracted feature map of the search image using a dot product.

The obtained dot products are used as a similarity score or activation for the attention map. Scores over the representative areas of the attention map corresponding to each object

Figure 2. Attention selection as an interaction of local salience and global context. Adapted from Olivia and Torralba, 2006 [27].

in the search image is averaged. The average score for a particular object in the search image is used to decide the next position of gaze fixation or attention using a winner takes all strategy. Inhibition of return allows the model to explore new points of fixation rather than returning to the previously visited ones. The results show that **guidance of attention by complex features in natural objects is possible**, as the search performance using this attention guidance model is better than a random serial search through the objects.

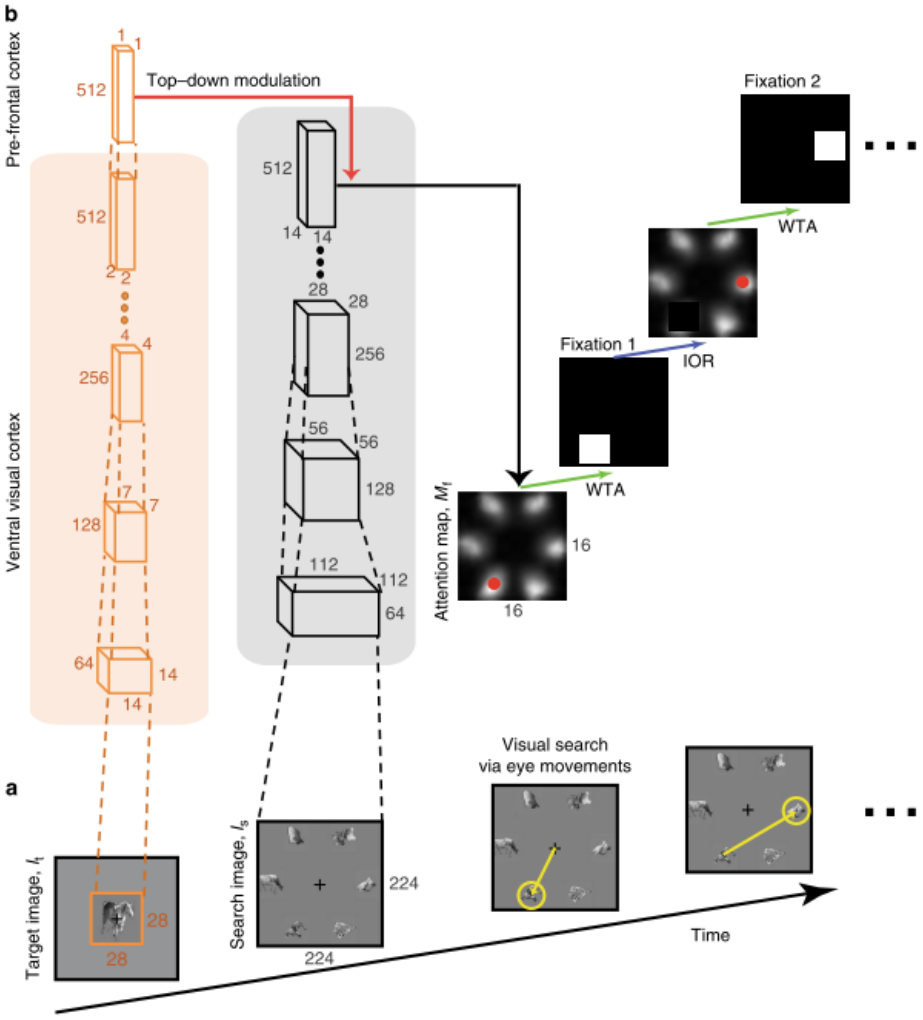


Figure 3. Computational Model of visual search using attention guided by top down target representation [45].

3 Methods and materials

The main objective of this work is to explain aspects of visual search by developing a computational model of the process based on the model by Zhang et al, 2018 [45]. **We aim to simulate visual search in experiments focused on comparing feature and conjunction search, to evaluate the performance of base computational model in these experiments and also evaluate our proposed modifications.** Particularly, the role of guidance of spatial visual attention in the visual search experiments will be explored. We seek to understand the variation of search time due to variations in the **search objects or stimuli** and their arrangement in the search image. The type of variations studied includes the change in the number of distractors, varying the complexity of objects, modulating difference between the target and distractor features.

We propose to use the main experimental paradigm (with our own variations) and model the search process computationally with the following steps.

- **Generation of search images** with a varying number of distractors for both single feature and conjunction of feature changes between target and distractors.
- **Extracting relevant visual features** that encode the target and search image using a pre-trained ANN. This process mimics the information extraction from images by the visual system in the brain.
- **Similarity score calculation** between target representation and each spatial position in search image representation that generates the resulting attention map.
- Mapping locations on the attention map to object locations in the original search image and **identifying the search order** for the target's position according to activation in the attention map.
- Repeat the process by **varying the number of distractors** and obtain the corresponding search time for each number of distractors.

This section expands on various aspects of the process covered above. It shall cover discussion on the neural network architectures used for extracting features from the images, the choice of stimuli, the choice of similarity measure and the implementation details for performing the visual search experiment with the computational model.

3.1 ANN architecture encoding image features

The computational visual search model proposed by Zhang et al, 2018 [45], generates an attention map for a search image with a particular target. The attention map is generated by convolving the higher-level neural network representations of the target on the search image's representation. Figure 4 shows the process adopted in our experiments.

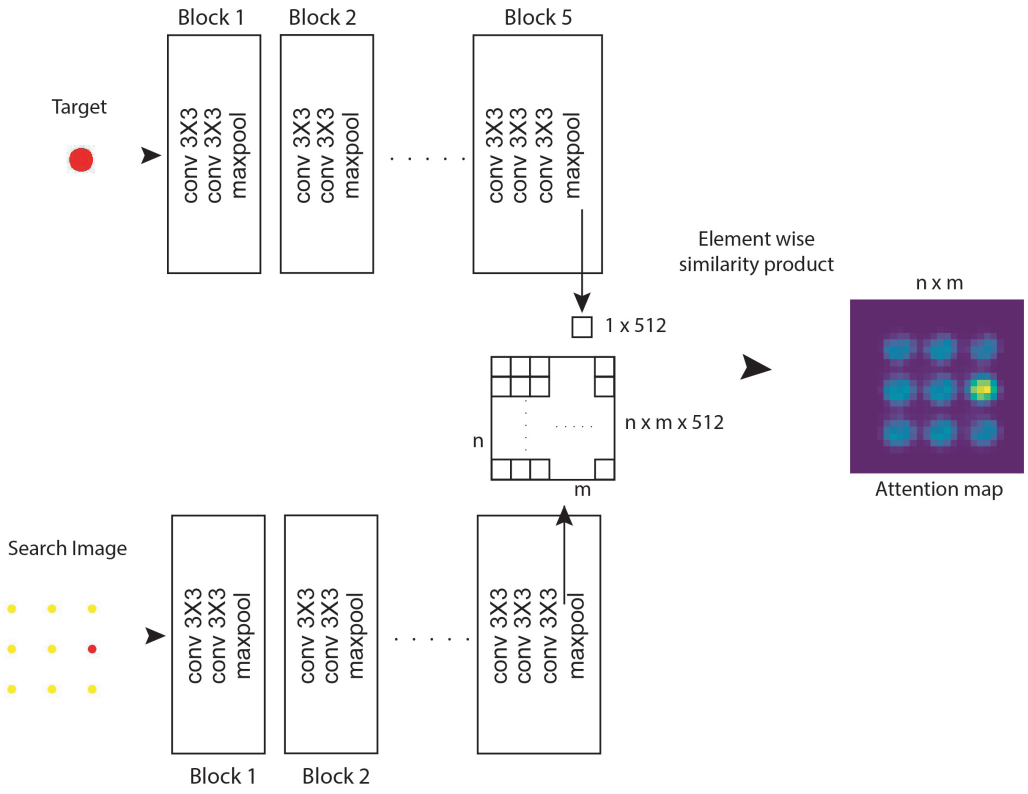


Figure 4. Feature extraction using CNN and attention map generation by convolution of target's features over search image's feature map.

.

As the raw inputs to the visual systems both in humans and machines in the form of retinal activations and pixels are very high dimensional, they need to be processed to extract useful information from them. Certain patterns of this information is called a **feature** that describes an aspect of the image.

In this text, a feature used by humans to describe an image will be called an **apparent feature** and the feature extracted by the ANN will be called an **encoded feature**.

The ANN architectures used to perform tasks on images like classification, segmentation or object detection are primarily based on Convolutional Neural Networks(CNNs) [31, 19, 29]. CNNs stack blocks of convolutional filters that are applied sequentially on the image to extract useful features. At the same time, they funnel the information to lower spatial dimensionality as the processing progresses. As a number of convolution operations combine the information deeper down the network, it learns higher-level features of the object or scene being represented. As the features learnt are used by the architecture for its final classification or segmentation task, the network tries to learn features that are helpful for its performance on that particular task.

Various factors affect the similarity of the computational model to the human visual search. This mostly depends on the base feature extracting CNN and the dataset used to pre-train it. Due to this, different CNN architectures were explored in this study, for their suitability as a base feature encoding architecture. The architectures used in this study are networks trained for the purpose of image classification. The networks are pre-trained with an ImageNet [6] dataset that consists of images of natural scenes and objects. The training of these architectures on natural scenes, in turn, is one possible reason that makes it possible to use these trained networks for making inferences about the human visual system. **It is expected from CNN to have learnt features similar to ones used by humans as the images used to train them are similar to the ones encountered by humans [13, 5]**.

In this study, two types of architectures have been tested. The VGG16 is a widely used image classification network, which has high test scores for the classification tasks on Imagenet Datasets. The Cornets are a family of architectures that modify the typical CNN architecture to use operations that are more plausible in the brain and better explain data recorded from the human visual cortex.

3.1.1 VGG16

VGG16 [31] is an architecture primarily used for image classification and detection tasks. Figure 5 shows the VGG16 architecture that consists of 16 weight layers (13 convolution + 3 dense fully connected layers). The spatial dimensionality reduction across an image is performed by 5 maxpool layers. The use of this architecture as a base for encoding target and search image representations has been proposed in Zang et al., 2018 [45]. The encoding of the target is obtained from the output of a selected convolution block. The encoding for the search image is obtained at a layer lower i.e. before the maxpool

operation is performed for the particular convolution block.

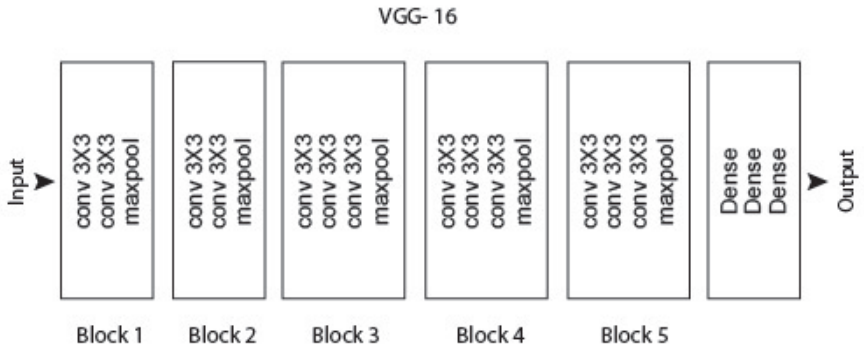


Figure 5. VGG16 Architecture. Adapted from Simonyan & Zisserman, 2014 [31].

In this work, the implementation of the feature extraction process using VGG16 architecture makes use of the default pre-trained VGG16 model in Keras library (used with Tensorflow backend).

3.1.2 Alternative Architectures considered: Cornets

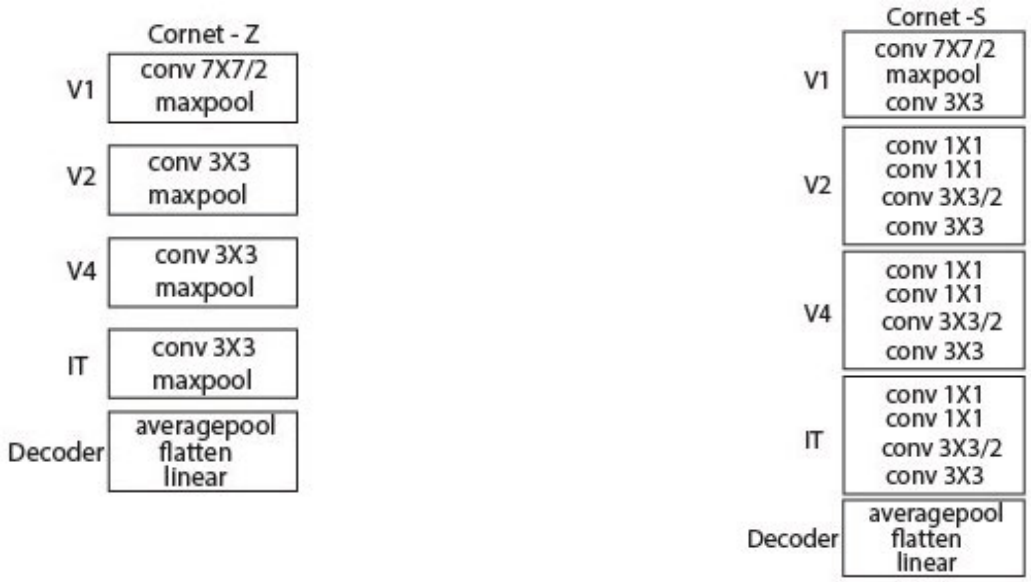
Cornets are a family of artificial neural networks based on CNNs using basic design principles of commonly used networks on image-based tasks [20]. Cornets optimise this base architecture design based on principles from the human visual cortex. The design is optimised based on the predictive ability of the network to encode for features encoded in V1, V2, V4 and IT areas of the human visual cortex. Networks with less complexity in terms of the number of convolution layers between the input and the end process are preferred. The authors of the Cornet study also explore the possibility of use of recurrences and lateral connections in the network. Other than the use of recurrence, these networks incorporate the use of strided convolution instead of max pool for spatial dimensionality reduction. Two of the network architectures that are evaluated in this study form the Cornet family are Cornet Z and Cornet S.

Cornet Z is one of the simplest cornet architecture with no use of recurrent network structures. In Figure 6a we see that the network tries to replicate the V1, V2, V3 and IT regions of the brain by simple feedforward convolutional layers. The transition into each

further logical block consists of a dimensionality reduction step using maxpool. The network has fairly little feedforward depth and complexity compared to the state of the art image classification networks.

In contrast to Cornet Z, Figure 6b shows the Cornet S which is a more complex architecture with weight sharing within each logical block, to reprocess the information over the same part of the architecture. It also consists of skip connections within the logical block. Though it has increased complexity form Cornet Z, Cornet S architecture gives the maximum classification accuracy amongst the Cornet family.

For the purpose of their use in the computational model in this study, the target representation is encoded as the output from the average pool layer of the decoder, while the search image is encoded as the output of the convolutional block corresponding to the V4 layer.



(a) Cornet Z

(b) Cornet S

Figure 6. a) Cornet Z and b) Cornet S Architectures . Adapted from Kubilius et al., 2018 [20].

The feature extraction process using Cornet Z and Cornet S makes use of the implementation provided by Kubilius et al., 2018 [20]. We use the implementation with pre-trained Imagenet weights using the PyTorch framework in python.

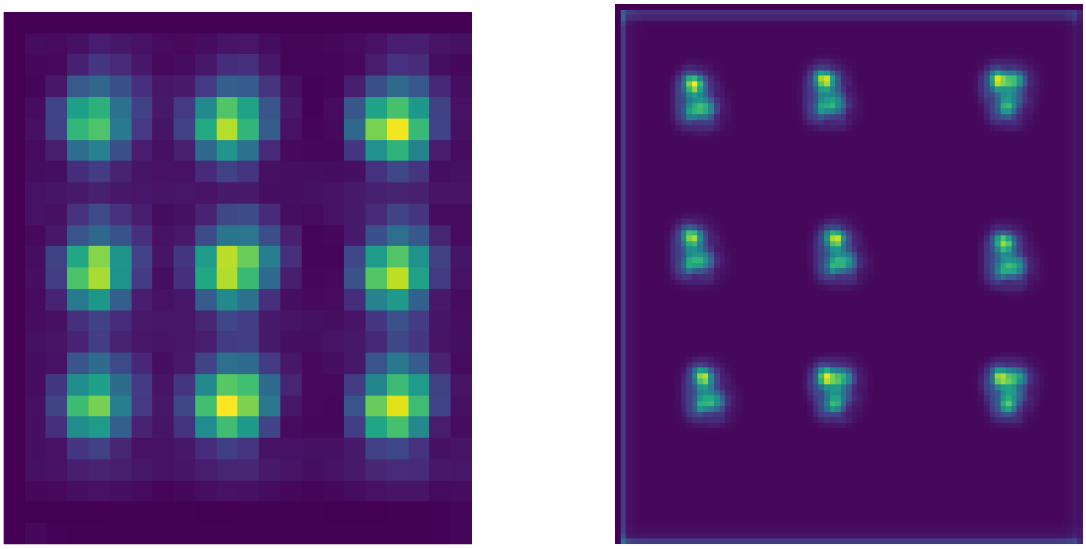
3.2 Stimuli Selection

The search stimuli or the objects in the search image used for the visual search experiments are expected to have certain properties so that their use in the experiment gives meaningful results. The **stimuli** or **search objects** shall be simple and have features that conform to the notion of a single apparent feature like color or orientation. Also, with a pre-trained CNN being used to extract embedded features, the features of the object have to be adequately represented by the CNN's filters. This is important so that the network mimics the human's performance and gives similar results as the stimuli used with humans. A baseline criterion used for selecting features of the stimuli is that the network's encoded features (or representations) vary enough with the variations along that apparent feature. **As for CNNs, encoding at a particular layer varies with the size of the objects in the image, we checked the variations on the CNN layer and then the size of objects before a final configuration and a range of stimuli are selected.**

3.2.1 Layer selection

The apparent features considered for the experiment are simple like shape, color and orientation. The encoding layer should be selected based on the proper encoding of these features by the layer. The agreement between CNN's encoding layer, stimuli and size of the stimuli is necessary for successful simulation of experiment. The network encodes more complex higher level features from the image with the increasing depth of the encoding layer [19]. Hence, we may want to explore lower layers for simple apparent features in our stimuli.

Lower layers of CNN encode for low-level features like the orientation of edges. On the other hand, lower level features capture only very local properties in space. Hence it is difficult to capture the overall feature representation of the object or object parts at a lower level. Figure 7 shows attention maps for search of target 'T' in a search image with objects 'T's and 'L's at two different depths. In Figure 7b, it is apparent that the attention map between the stimuli and target appears like the object itself. This indicates this phenomenon of non-aggregation of spatial features over the object. The higher layer output shown in Figure 7a show a more aggregated blur and are also more invariant to local perturbations around objects in the search image. Thus the highest level of CNN layer possible for a given architecture is used to extract the features from the images after testing that they could still encode most of the apparent features of our choice.



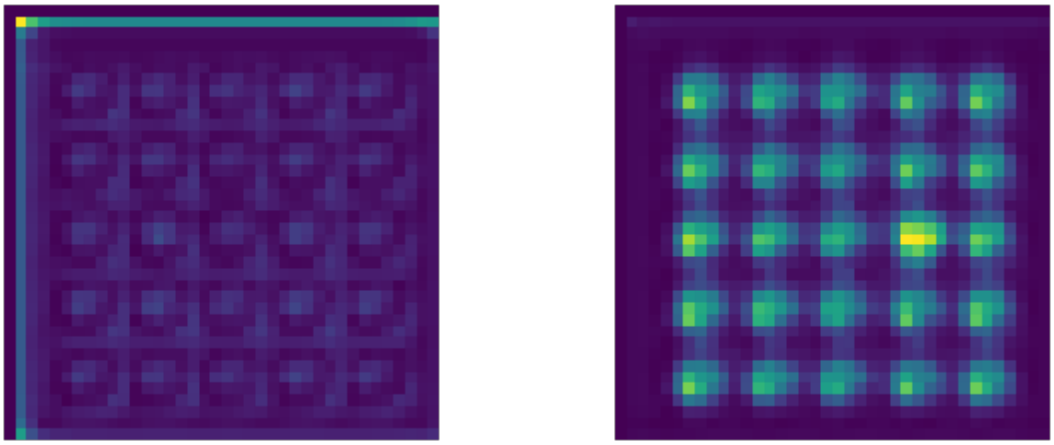
(a) Higher Layer (b) Lower Layer

Figure 7. Attention maps for target with stimuli 'T' and search objects 'T' and 'L', generated using features from different CNN layers, depicting better aggregation of local features in a) higher layer than b) lower layer.

3.2.2 Stimuli Size selection

CNN's extracted feature representations change with the size of the objects in the image [31]. Therefore, we tuned the size of the objects in the stimuli to meet some optimal range. A CNN can only encode apparent features using a filter at a certain layer depth for the object size. This size is dependent on the size of objects available during training. Selection of optimal size of objects is performed by checking the final attention maps to show the difference in object activations on change along an observed feature. Figure 8a and 8b shows the effect on attention maps for different sizes of search objects.

As the size of the encoded feature map is dependent on the input image. Also, the size of the target object is kept the same in the target image and the search image. Hence due to the constraints on the size of the target to produce a favourable 1x1 spatial feature map for easy interpretation and computation, the maximum size of the objects used is also constrained by the size of the target image.



(a) bar size 5 pixels (b) bar size 18 pixels

Figure 8. Attention maps generated with different size of stimuli. a) bar size 5 pixels show less difference between objects than b) bar size 18 pixels.

3.2.3 Choice of stimuli

The search objects or stimuli tested are the most commonly used ones in literature with human experiments. We chose the stimuli such that it should be easy to generate it computationally and vary the required apparent features. The choice of stimuli is also dependent on the apparent feature desired to be varied for the tests and the base network's ability to adequately represent the apparent feature in its encoded feature representation. Figure 9 shows examples of the stimuli used in the experiments.

3.3 Similarity score

A similarity score is calculated while convolving the target's encoded representation at each position on the search image's representation. Zhang et al., 2018 [45] describes this as a scalar product between the two vectors at each position (Equation 1).

$$a(x, y) = f_s(x, y).f_t \tag{1}$$

Here, $f_s(x, y)$is the encoded feature vector at position x, y of the feature map for the search image and $f_t$is the encoded feature vector for the target. $a(x, y)$ represents the obtained attention map value at position x, y. The operator "." represents a scalar or dot product between the two vectors of equal length.

Figure 9. Examples of objects used as stimuli in this work.

The **similarity score** obtained at each position is a measure of higher-level similarity between the target and the object at the position in the search image. This aggregate map of similarity scores is considered as the **attention map** to guide the attention in the target's search on the particular search image.

Each spatial position on the encoded feature map for target and the search image consists of a vector representing the CNN extracted encoded features. With the CNN architecture settings used in this work, the vector of features is of the same length for the target and search image encoding.

During the experiment, the input search image and its encoded spatial dimensions change with the changing number of distractors. As only the convolutional layers are used to obtain the output encoding, variable size stimuli images can be processed by CNN. This is possible because the convolution kernel performs the same operations over a predefined local area of its inputs and moves over the whole image repeating the operation.

The target image 's size is fixed at 32×32 pixels, that provides a 1×1 spatial dimension with our choice of encoding layer for VGG16 and Cornets. With a single vector representing the target, generation of attention map with its convolution over the search image' encoded feature representation is a scalar product between the target vector and the search feature vector at each location.

3.4 Experiment Details

3.4.1 Search Image Generation

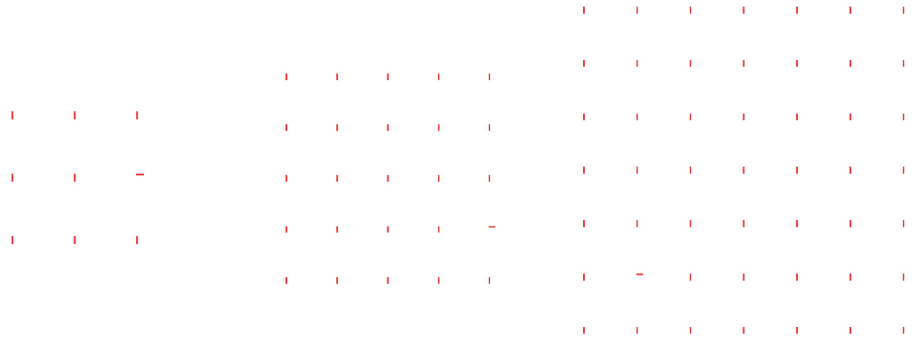
The base experiment conducted in this work is to compare the search time for our computational model with variations in the search image. We develop an image generator to obtain the required search images. The requirement from the search image generator is that it should create images with different basic stimuli types, allowing for changing the orientation, color, shape, size and distance between the objects. As the visual search experiment checks for the increase in search time with the increasing number of distractors, the generator shall allow for increasing the number of distractors, hence generating images of varied size.

Figure 10a and 10b show examples of generated search images for feature and conjunction search. **The generator places the target object at a random position in the search image while also randomly allocating positions to various distractors.** The output of the generator is a search image along with two numbers denoting the position of the target in the search image and the size of the generated image. These outputs are further used by other modules, the image and its width for processing the attention map and the target position is used for confirming the result of the experiment. The arrangement of the shapes or objects on the search image is done in the form of a grid where the input to the generator specifies the number of objects in each side of the grid. The color of the background is set to white. The final size in terms of pixels of the search image increases with the increase in the number of distractors.

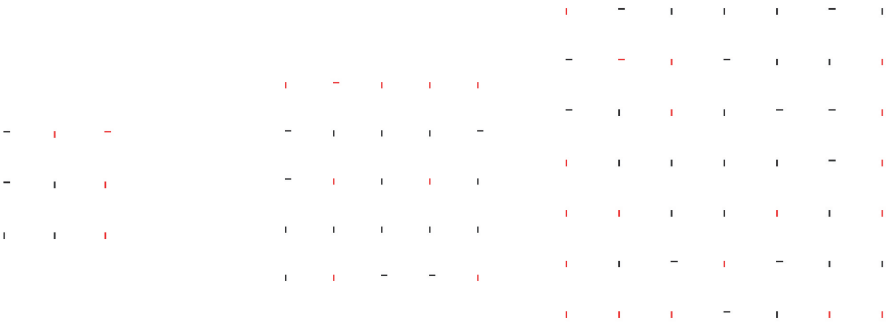
The generation of the target image is done using a simple code derived from the generator. The object in target image has the same size as the size of the objects in the search image. The dimension of the target image is fixed to 32X32 pixels. White background fills the rest of the image. The image generator uses python's PIL module for creating the various stimuli and generating a final image as an output.

3.4.2 Attention-map processing

The attention map is generated after the convolution of the target‘s encoded feature representation on the search image's encoded feature map using a particular choice of similarity score. The attention map is a scaled-down version of the search image and corresponds more closely to the scale of search image's encoded feature map. After a manual identification of the scaling transform between the search image and its attention map, objects can be mapped to their respective positions on the attention map. The



(a) Feature Search



(b) Conjunction Search

Figure 10. Example of search images with various number of objects for a) Feature Search b) Conjunction Search, generated using the image generator.

attention map is easy to parse as the objects in the search image are arranged uniformly, hence the positions of their activations on the attention map are uniformly distributed.

The clusters of activation on the attention map depict that only positions corresponding to the various object's positions on the search image show activations. As shown in Figure 11, the object with the highest activation is identified as the first gaze position. The maximum activation amongst the cluster representing an object is chosen as the activation for the object. Using this rule an activation score for each object on the search image is obtained.

The sequence of gaze made on search objects is determined by the descending order

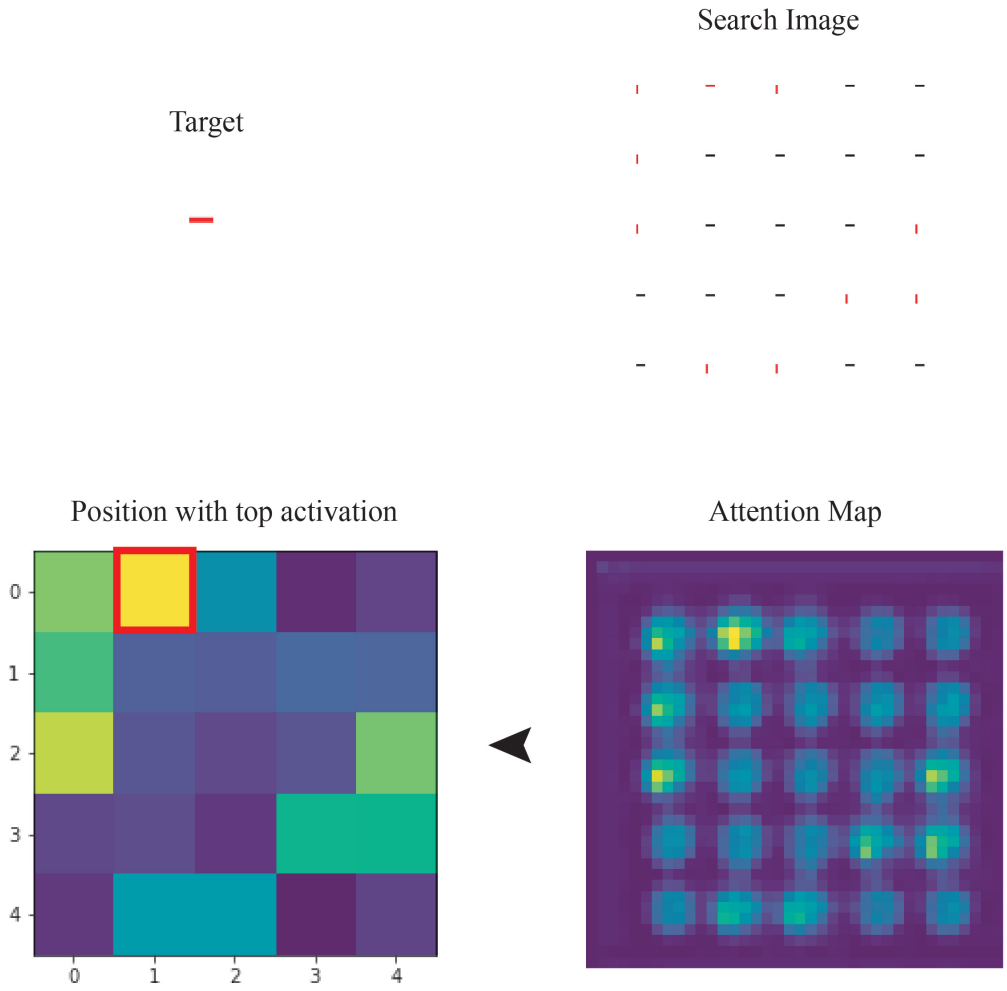


Figure 11. Attention map processing to generate per object activation for gaze point selection. Top activation is marked with red enclosure, the corresponding object position will be the point of first gaze fixation.

of activation score representing their positions. This rule is based on the inhibition of return in human gaze patterns, which suggests subjects remember and avoid previously attended objects. Accordingly, the final search time would be noted as the target's position in gaze sequence times a constant time (T_{const}) to process objects at each position. In results and analysis, the constant time is not represented and the final search time is given by the number of gaze shifts before the target is identified.

3.4.3 Evaluating search time for an increasing number of distractors

The result from the processed attention map i.e the position of target in the gaze sequence is used as the corresponding search time. The search time is the basic measure used in this work to check the performance in the search task. The search images with varying numbers of distractors are generated using the image generator. As the image generator arranges images in a square grid pattern with n as the number of images per side of the square. Variation is made on n from 2 to 10 to generate search images for the experiments. For each number of distractors, the experiment is repeated 30 times so that search time is averaged over varying random positions of target and distractors in the search image. Finally, $n \times n$ i.e number of objects in the search image vs search time graph is generated.

4 Results

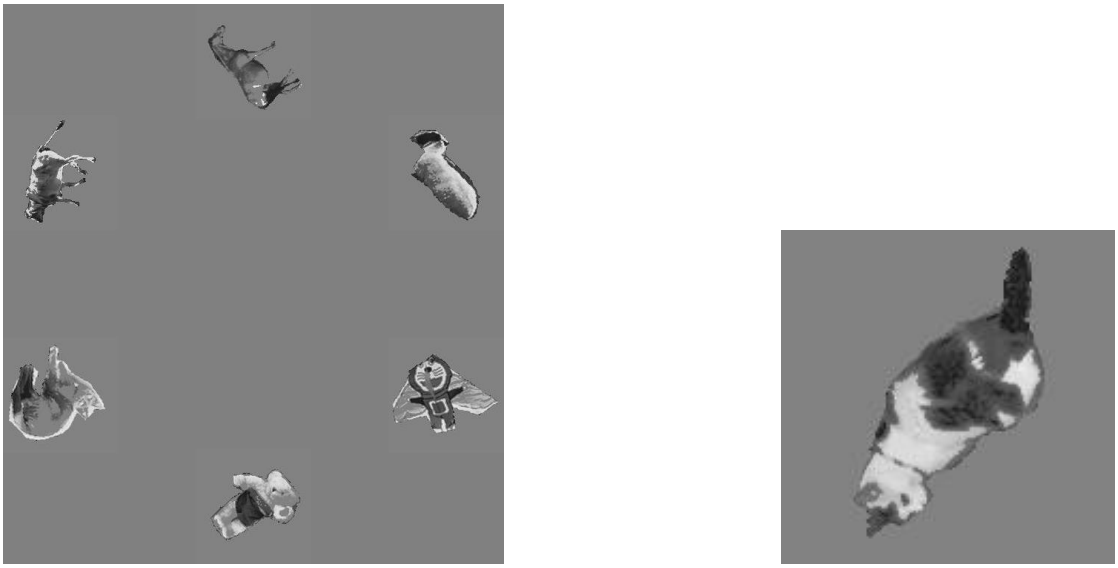
We stimulate the use of the computational model proposed by Zhand et al., 2018 [45] in visual search experiments. Particularly, in experiments comparing between feature and conjunction search. The complete pipeline to simulate various experiments including generators for search images, processing of attention maps and finally obtaining the search time graphs has been adapted and implemented. There are some new components such as image generators that have been designed from scratch. Further, we aim to propose modifications to the base model by trying different base CNN architectures and new methods to introduce resource constraints in the model, particularly related to limitations in working memory.

This section describes the experiments and its results, conducted for comparing feature and conjunction search. It also proposes modifications to Zhang and collegues's visual search model and discusses its effect on the search performance. Finally, we try to understand the role of feature similarity in the search objects and its impact on attention guidance and search efficiency.

4.1 Selection of base CNN architecture

First, we evaluate the use of the architectures from the Cornet family as the base CNN architecture for encoding features. The suitability of architectures is evaluated based on the search performance of the models using these architectures on visual search tasks. The target and search images for this experiment are used from the MS COCO Dataset [23]. Examples of these images are given in Figure 12a and 12b. The base model, search and target images used for this evaluation are the same as used in the original paper [45], hence the search performance can be compared to their results. The replication of results from the original paper is performed by using VGG16 as the base feature encoding network.

Figure 13 shows the cumulative search performance quantified as a proportion of times target is found when a certain number of gaze fixations have been made for VGG16, Cornet Z and Cornet S networks. It is calculated over the 300 search image dataset and the graph is plotted against an increasing number of fixations. The random search line (dotted orange) indicates the search performance in case of a random selection of search order for the objects on the search image. The curve for VGG16 shows above random performance using this architecture as in about 30 % of the times the target image can be identified in the first gaze.



(a) Search Image (b) Target

Figure 12. Examples of a) search and b) target images from MS COCO dataset.

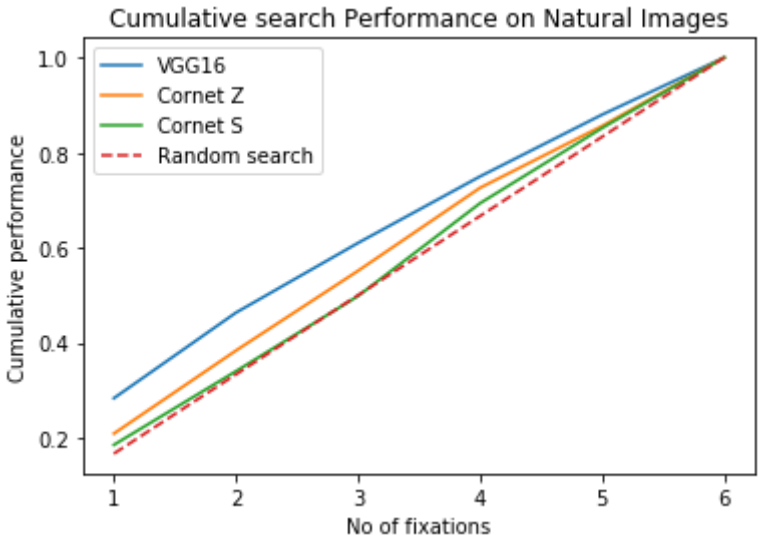


Figure 13. Cumulative performance as a function of completed gaze, calculated over 300 search images.

The same curves compared for Cornet Z and Cornet S do not show comparable performance to VGG16. The graph for Cornet Z shows above random search performance, but that for Cornet S the performance is similar to a random selection of gaze fixations.

The reason for lower search performance in the case of Cornet Z and Cornet S was investigated by finding these networks' response to the encoding of an object at different positions and rotations in the search image. In a search task the encoding of the features of an object, at object's corresponding position in the feature map, should be invariant to the part of the image in which the of the objects is located in the search image. Otherwise, the position in the image shall affect the results more than the guidance by object's similarity of features with the target. For this test, the feature vector is obtained from the final average pool layer of both Cornet Z and Cornet S, which gives a single vector representing the image. The cosine similarity of the vector with other configurations in position and rotation is plotted to understand the changes in encoding.

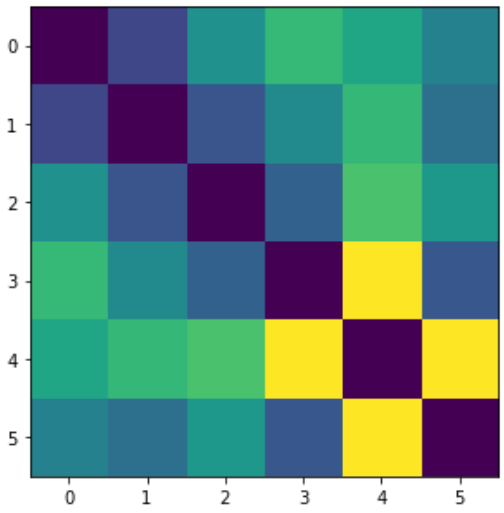
Figures 14a & 14c show that Cornet Z's similarity matrices have some similar cells, depicting some translation and rotation invariance. In comparison, Cornet S in Figure 14b and 14d has very less translation and rotation invariance. The results correspond to the search results shown by the two networks and may be the possible reasons for their non-suitability for use in modelling visual search.

Due to these reasons, the usability of networks from the Cornet family is not explored further in this work. In the case of VGG16, some variance can be seen in the encoded feature with the varying position in the search image but the variance is fairly low. This allows for the generation of attention maps that can be used for efficient search.

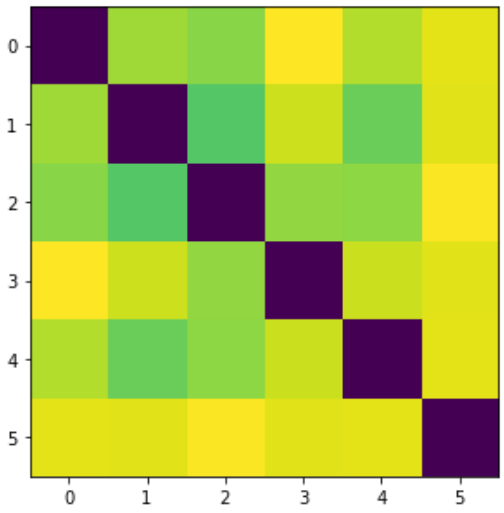
4.2 Similarity score using the full feature vector dot product

The first object used in the simulation of visual search experiment to compare feature search with conjunction search is target shape 'T' in a field of distractors 'L'. These shapes are themselves considered to be conjunctions or complex features as they are composed of two bars placed at different positions [8]. In the single feature variation, color is varied, where the target red 'T' is searched from black 'T' distractors as can be seen in Figure 15a. The search uses the base attention guidance model with full scalar product across encoded features as the similarity score. The search time estimation, using the target position's activation rank is plotted as a function of the number of objects in the search image.

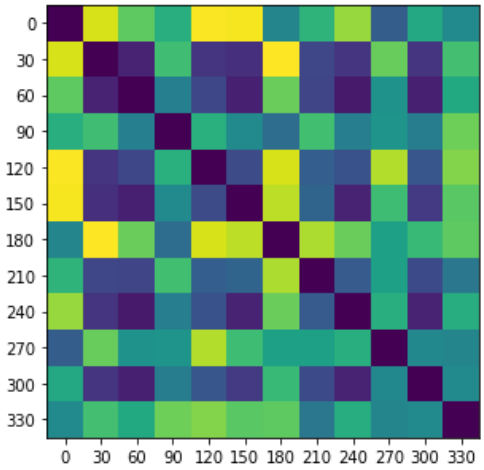
Figure 15a and 15b show the search and target images used for this particular experiment. **Figure 15c gives the search time graph averaged over 30 random trials for each number of distractors in the feature search and conjunction search conditions.** The extent of the error bars denote the standard deviation of the readings during these 30 trails. In the case of the conjunction search, the search time increases with the number of



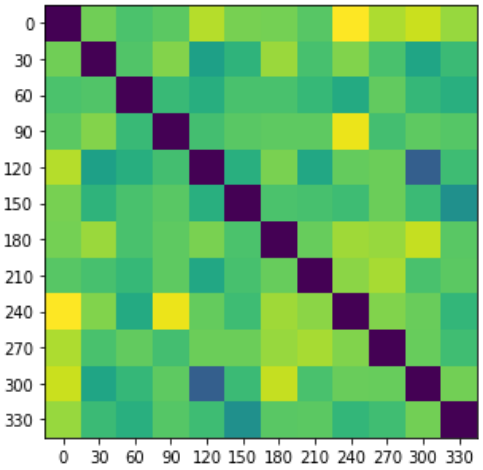
(a) Cornet Z: 6 different positions of an object



(b) Cornet S: 6 different positions of an object



(c) Cornet Z: rotation at given angles (in degrees)

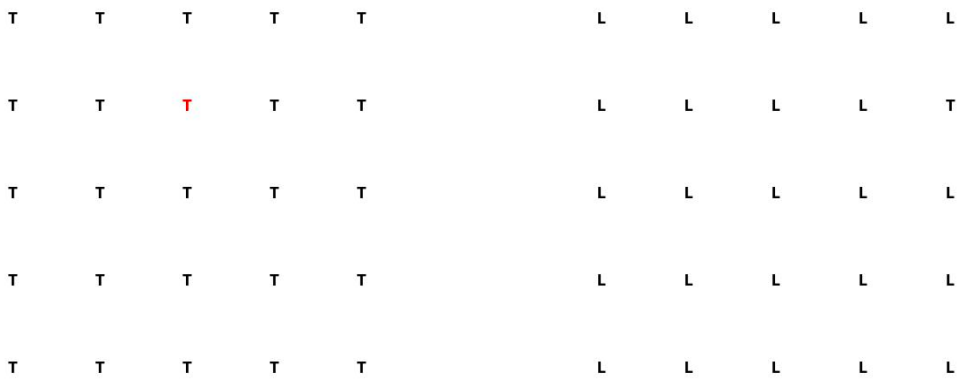


(d) Cornet S: rotation at given angles (in degrees)

Figure 14. Similarity matrix of feature encoding in Cornet Z with a) different object positions, c) different orientations. Similar matrix for Cornet S are given in b) with different positions and c) with different orientations. Higher similarity means higher feature invariance and better suitability for the search model.

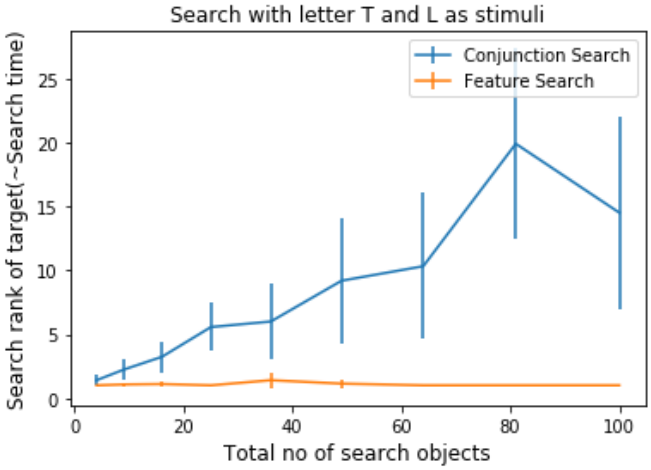
distractors, while a near-constant search time can be noticed in case of feature search.

To check whether the computational search model shows the same result with other



(a) Feature Search Image

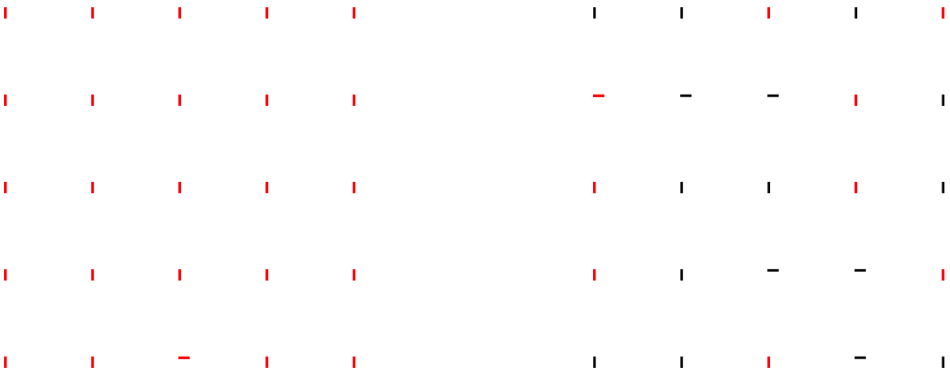
(b) Conjunction Search Image



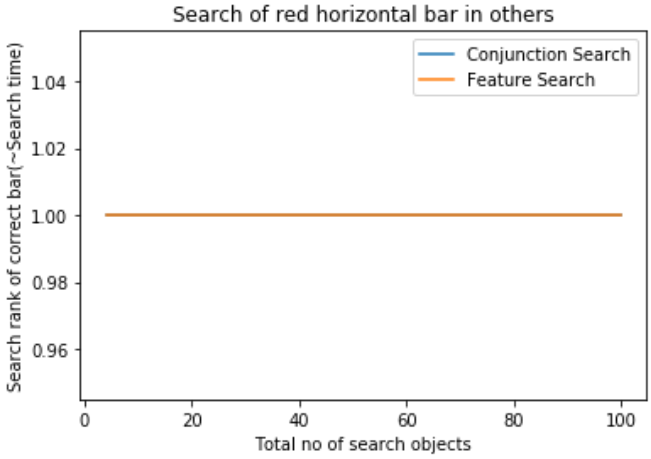
(c) Search time vs number of distractors for using given stimuli

Figure 15. Example of Search Images with objects 'T' and 'L' as stimuli for a) Feature Search and b) Conjunction Search. c) Search time experiments with given stimuli using full feature vector scalar product as similarity score.

stimuli, the experiment is then conducted using bars as shown in Figures 16a and 16b. In this case, the orientation and color of the bars are apparent features. In Figure 16c, it can be seen that the search time using the full vector product always remains constant i.e. the target is found in the first gaze. **This implies that the features are encoded**



(a) Feature Search Image (b) Conjunction Search Image



(c) Search time vs number of distractors while using model with full feature vector scalar product as similarity score.

Figure 16. Example of Search images with bars as stimuli for a) Feature Search and b) Conjunction Search. c) Search time experiments with given stimuli using full feature vector scalar product as similarity score.

optimally to cause different activations for different objects. In case of both feature and conjunction search, a pop-out effect is occurring as the target position has the highest activation in the attention map.

4.3 Exploring effects of limited representation precision in a CNN model

As the operations in the human brain have various resource constraints such as limited memory, we propose modifications to the computational model to better demonstrate the effect of these constraints. While formulating these changes in the model, we try to use operations which we can possibly take place in the human brain. To introduce these resource constraints, three different modifications in ways of generating attention maps are proposed. These proposed changes may be better suited to explain the different nature of search time results in the feature search and conjunction search over a larger range of stimuli.

4.3.1 Using higher order norms in similarity score

The original feature similarity is determined using a scalar product that takes equal contributions from each feature component in the activation score. The simple scalar product can be defined as given in equation (2).

$$d = \sum_{i=1}^{n} f_s(i) \times f_t(i) \tag{2}$$

Where d is the scalar product for feature vectors f_s(search) and f_t(target) with n elements.

Which notion of similarity (or distance) between representations is more adequate to describe how neural systems guide some decisions is still open. Hence, here we explore some variations of the standard correlation or scalar product. We introduce higher order norms as a measure to calculate similarity with the expression in equation (3)

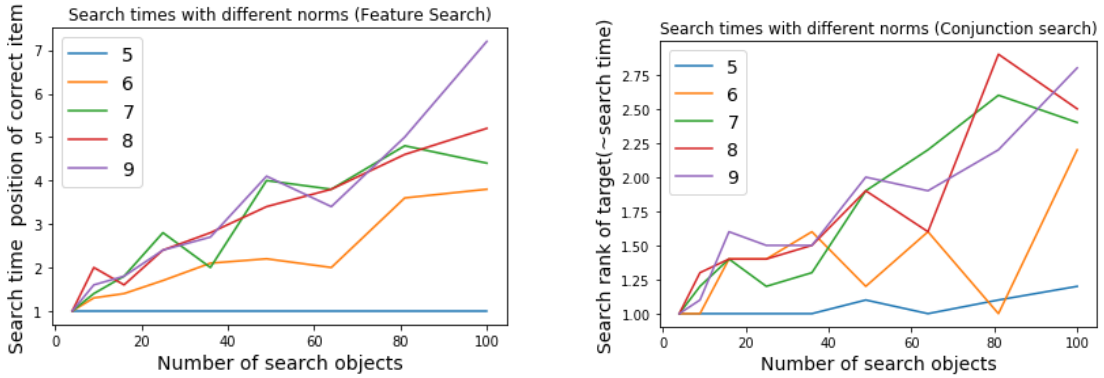
$$d = \sqrt[m]{\sum_{i=0}^{n} (f_s(i) \times f_t(i))^m} \tag{3}$$

In this equation m denotes the order of norm. The value for m if selected as 1 we get the original model that uses equation (2). With higher norms the notion of similarity is such that it can give more consideration to indices where individual similarity of features have higher magnitude.

To test this notion of similarity, the use of bars as a search stimuli is continued. With the

introduction of higher norms, if the object is primarily defined by features e.g. color and orientation, then its scalar product with the target will have a higher score contributions from the features that are common in between the target and the search object. It is expected from this method to prioritise the use of such defining features of the objects.

We try different order of norms in both feature search and conjunction search conditions and obtain the search time graphs in figure 17a and 17b.



(a) Feature Search

(b) Conjunction Search

Figure 17. Search experiments with use of increasing norms in similarity score in a) Feature Search and b) Conjunction Search. Experiment performed using bars as stimuli.

It can be seen that norms higher than order 5 introduce some noise to the search process as the search process can no longer identify the target in the first place. So higher norms can simulate resource constraints in representation, in a certain manner. But the increase in search time is much greater in feature search than in conjunction search. Hence, this similarity score, does not satisfy the observed characteristic patterns of feature and conjunction search in human vision.

4.3.2 Noise addition to target's representation

In the next step, we propose another method to introduce resource constraints to the visual search model. The target representation must be encoded and stored somewhere in the brain as it is compared to the representation of the stimuli objects. **As the number of objects to be stored increases, there is some loss of precision in the representation of those objects [25]. Hence in case of conjunction search when more type of objects need to be considered, the target's representation is added some Gaussian noise to represent this loss of precision.** During the computational implementation, Gaussian noise of a specific variance is added to the target's representation as shown in Figure 18.

The scalar product to obtain the attention maps is calculated with this noisy representation in case of conjunction search. The degree of noise can be controlled by changing the variance for the additive noise. No extra noise is added to the target representation in case of feature search.

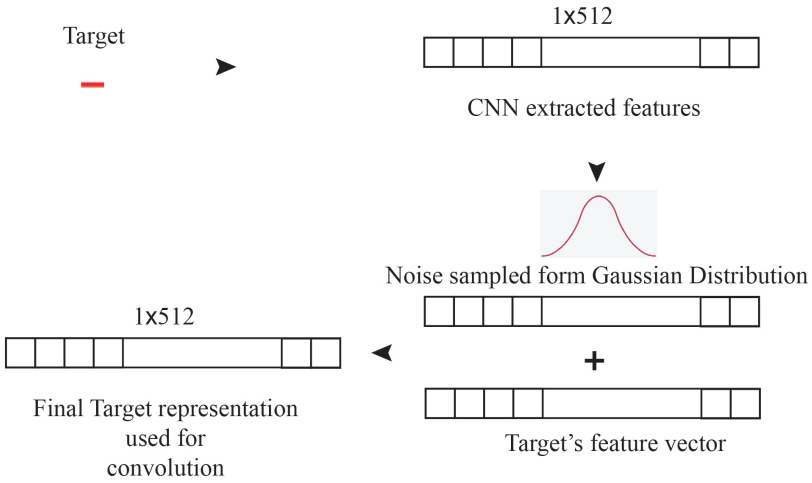
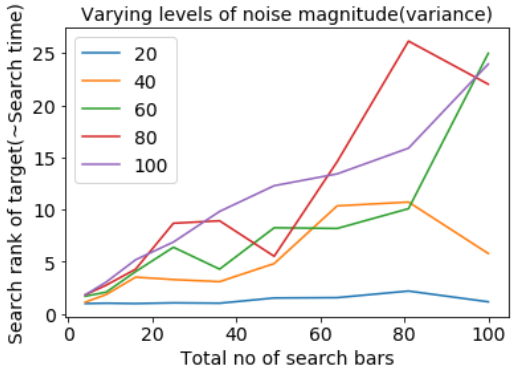


Figure 18. Noise addition to target's representation in the case of conjunction search. This represents decreasing precision of feature representations due to memory limitations.

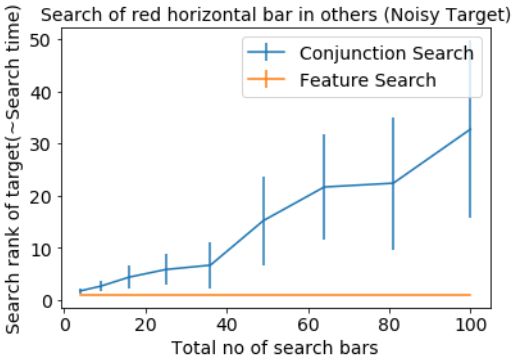
The effect of noise levels was tested on the search time with bars as stimuli. We used a conjunction of two features (orientation and color) for objects in this search image, as shown in Figure 16b. For each feature value in the target's encoding vector, a noise value sampled from Gaussian distribution with variance (v) and mean 0 is added to the feature. The search time graph is then obtained from the attention map using the scalar product for each location between the feature vector of the search image and the noisy target feature vector.

Increase in the variance (v) of noise is used to see the effect of increased noise on the search process. Five different noise levels are used from 20 to 100. The obtained search time graphs in Figure 19.a shows that the slope of the search time graph increases with the increase in noise levels.

Using this understanding of the effect of noise on search time slopes, a noise of variance 80 is added to the target's representation in conjunction search and search time plots for conjunction and feature search are obtained in Figure 19b. This noise level can be compared to the average inherent noise in the feature encoding in our system. This noise



(a) Variation of search time with different levels(variance) of noise.



(b) Search results using 80 as variance of added noise in case of conjunction search.

Figure 19. Addition of noise in conjunction search using bars as stimuli. a) Experiment with different noise levels b) Search time graph with a particular variance of added noise in conjunction search.

is due to the changes of the same object's position on the search image and has a variance of around 20 units.

4.3.3 Similarity score with feature selection based on high variance

Another modification to the model, to introduce and modulate memory-based resource limitations can be to decrease the precision of encoded feature representation by limiting the length of encoded features vector that is used for similarity score calculation. **With the introduction of this memory-based limitation to feature vector, there must be some mechanisms that rate the encoded features in terms of their importance and use the most important features. The chosen features are then used for the final similarity product.**

The importance of features in this case, can be decided based on which features help the most to differentiate between the objects in the attention map. Computationally, variance in contribution to the scalar product across different types of search objects marks the importance of a feature index. The selection procedure can be seen in Figure 11, where the ability of an encoded feature to cause the maximum variation in the top activation in different types of search objects in the image is used to prioritise its selection. Top N variant features are then used to calculate the similarity product used in the attention map. Feature importance in this case is dependent on the types of objects used in the search image, hence the importance is recalculated when the types of objects on the search

image changes. Therefore the feature importance and thus the selection of encoded features that contribute to the final similarity score shall be different for the feature and conjunction search images.

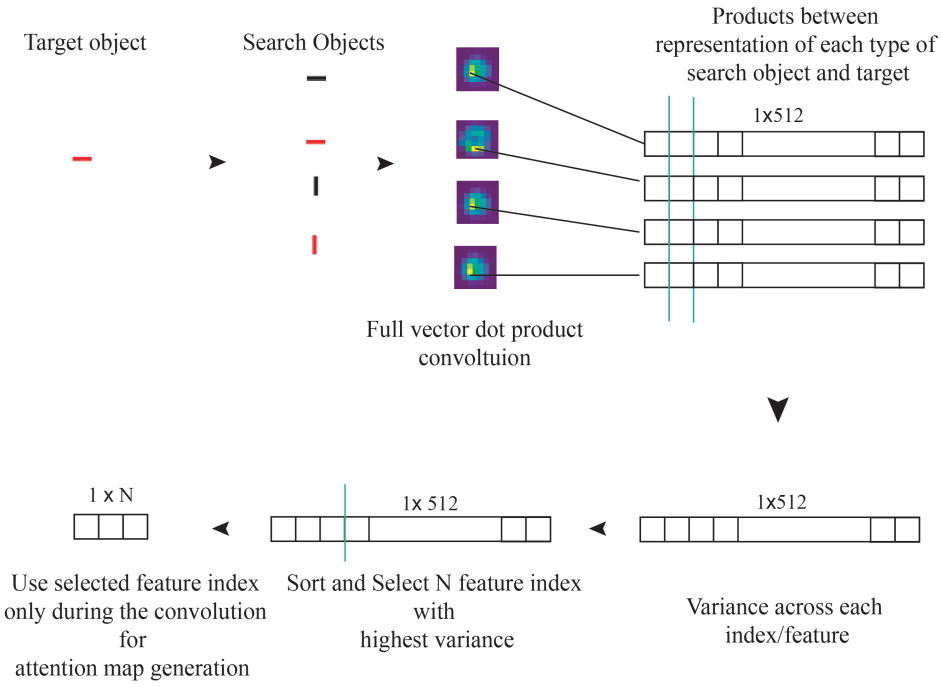


Figure 20. Feature selection procedure based on feature's contribution to variance in object activations across different objects. Only top N features are used for generation of attention map to guide visual attention to simulate limited memory. Using this process, same vector length is used in case of both feature and conjunction search but the selected features are different.

We continue to use bars (given in Figures 16a and 16b) as search objects for further experiments with this proposed method. As we do not know a valid length of feature vector limitations for the system, it is determined based on the point where the modified system can emulate the characteristic search performance for feature and conjunction search. To determine the vector length cut off, once the order of feature importance is determined for the feature and conjunction images, a cut off (N) is selected where the search performance in the conjunction case starts deteriorating with the decrease in encoded features. The cutoff in this case with the use of bars as stimuli is 3 top features (i.e. 3 features with most variance in activations over the objects in the search image).

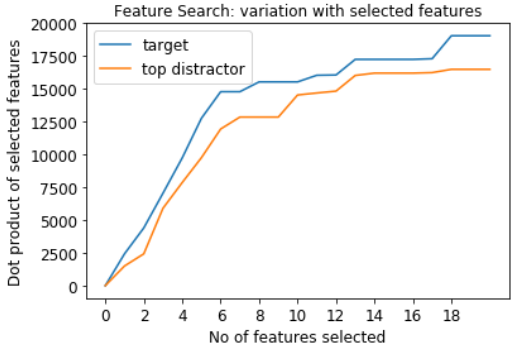
In Figure 21a and 21b, activations for the correct target position and the top distractor

activation from the attention map is plotted against the number of top important features used in the final dot product. The difference between target and distractor activations decreases as less number of features are selected. This difference in case of feature search does not become zero until the last point as it has only one type of distractor. The feature used to differentiate between that distractor is always represented in the selected features. In the conjunction search, two kinds of distractors differ in two different apparent features from the target, so as the number of features decreases, there will be a point where the feature distinguishing between the target and one kind of distractor will be underrepresented in the selected feature vector. In the curve, we see this as the position where the line representing the target and top distractor activations cross each other. Any number of features just below this point is chosen as the feature vector cutoff length.

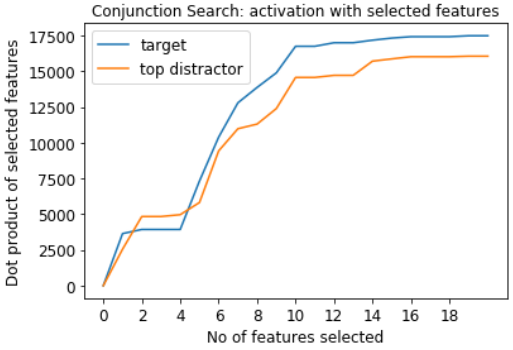
To check this modified model with bars as stimuli, the search time results were obtained with the number of feature vector length cutoff at 3 top features. This gives the search time graph in Figure 21c. The search time for the conjunction search rises with the number of distractors but for single-feature search, the search time remains constant with the number of distractors.

For other stimuli with different shapes given in Figure 22a and 22b, we also tested the same similarity score. The search time graph for this experiment is given in Figure 22c. It shows a similar characteristic result for search time variation with the number of distractors in feature and conjunction search.

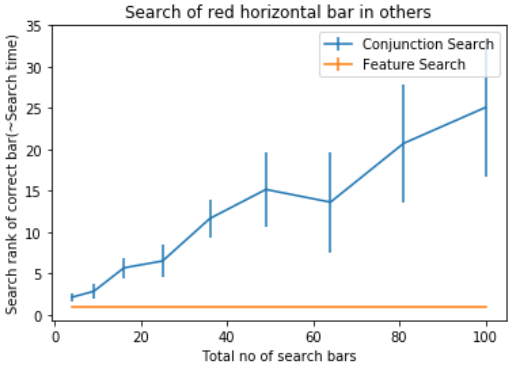
In this case, the experiment was started with the use of color yellow (RGB: 255,255,0) for the distractors and red (RGB: 255,0,0) for the target but we could not find a cut-off point for limiting the number of features. The reason for this is entangled encoded features and a high magnitude of difference in the apparent feature and its encoded value. Entanglement of features is a case in which apparent features, say color and shape are represented in same encoded features (i.e. the value of an encoded feature changes with a change in value of either of these two apparent features). In such cases if the objects are very different in their color and the shape, their values in the encoded feature will also be very different. These encoded features with very different values will also be the ones to be selected for final scalar product due to their high variance in the feature selection process. Due to this even when limited to a single encoded feature, the search process could distinguish between the target and the distractors in both conjunction and feature search. We finally bring the value of feature color between the target (RGB: 255,0,0) and distractor (RGB: 255,110,0) closer for obtaining the desired search time characteristics in conjunction and feature search. The effect of feature magnitude on the visual search, has been studied in detail in the next section.



(a) Feature Search



(b) Conjunction Search

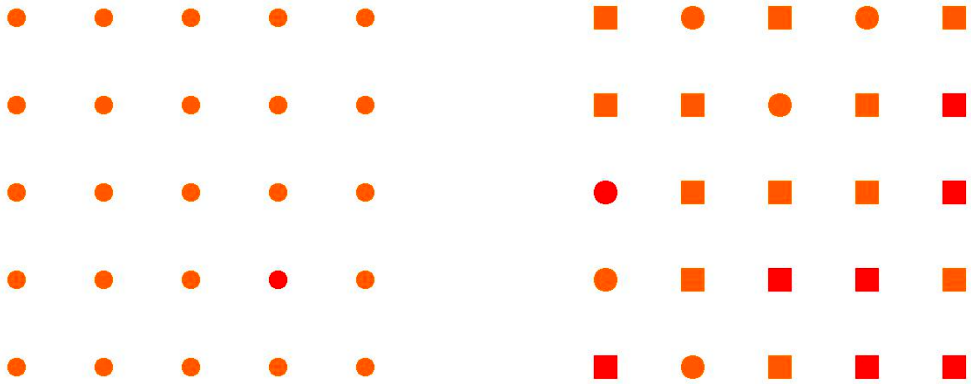


(c) Search time results in Conjunction and Feature Search

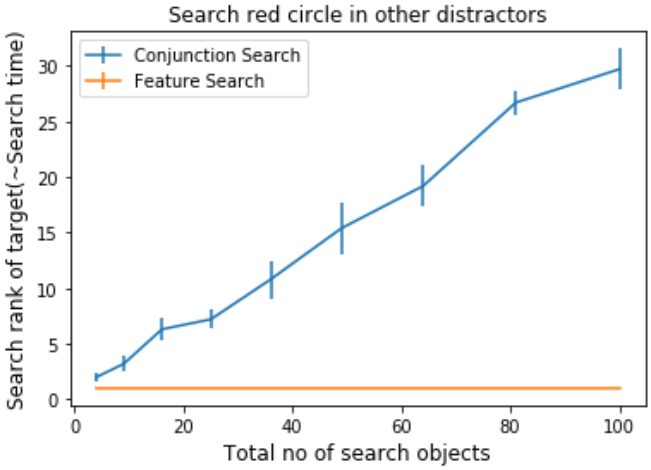
Figure 21. Object activation variation with the increasing number of most variant features used in search a) in Feature Search and b) in Conjunction Search. c) Search time results in Feature and Conjunction Search using bars as stimuli.

4.4 Search time dependence on stimuli's similarity.

In the previous experiments, we noticed that the attention activations of an object are dependent on the neural network's feature encoding of the object in the images. The results have shown that change in the number of apparent features characterising an object, affects the object's ability to be distinguished by the attention guiding mechanism. But, as this distinguishing ability is based on the encoded feature values we explore if the search shall also depend on the degree of dissimilarity of the target from the distractors within a single apparent feature.



(a) Feature Search Image (b) Conjunction Search Image



(c) Search time results in Conjunction and Feature Search

Figure 22. Search for red circle in distractors with different a) color (orange) in Feature Search b) color and shape (square) in Conjunction Search. c) Search time graph for the stimuli when search is modeled with limited capacity feature encoding.

In Figure 23, the plot shows the effect of change of similarity along a dimension in the search objects i.e color. The color of the target circle is selected to be red encoded in RGB scale as (255,0,0) and all the distractors are given the same red color at the start i.e. (255,0,0). The value of the green component for the distractors is then increased from 0 to 250 by steps of 10. The activation map is created for each step of change in color.

A plot for the difference between the target activation and the top distractor activation while increasing the intensity of the green color is obtained.

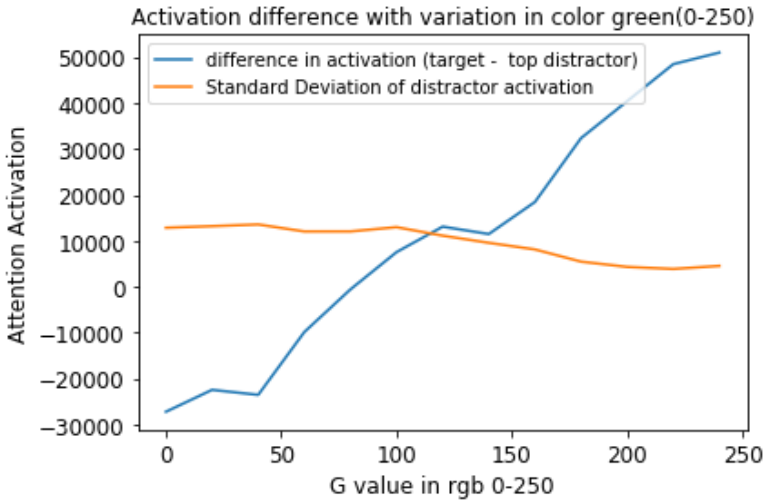


Figure 23. Change in attention activation similarity between target and distractor with the change in similarity of their color. Experiment uses circles as search objects.

The plot shows that with the increased difference between the color of the distractor and the target, the difference between the target's attention activation and the distractor's attention activation increases. The orange line marks the standard deviation due to positional changes in the distractor's representation. This can be considered as one source of noise for the system. Hence, if the difference between the target and the top distractor activation is less than the noise levels the attention selection will be confused for distractors and the slope of the search time curve will be higher than 0.

The significance of this result is that even within a single feature of distinction, it is possible to not have a pop-out condition when the difference between the distractors and the target is less than the noise. On the other hand if the activations of distractors and the target are very different in case of conjunction search, the target may still be identified in first gaze, irrespective of the number of distractors.

5 Discussion

The historical understanding of the visual search through Feature Integration Theory (FIT) [38], explains feature search as a parallel processing mechanism across all objects whereas a serial search mechanism works for conjunction search as the features need to be integrated for the object to be identified. The guided search model [2], built on top of FIT and proposed that the guidance of attention takes place in both feature and conjunction search. According to the guided search model only the quality of guidance changes due to the change in presence of guiding features. In Zhang et al, 2018 [45] they build a computational model in which attention is guided by the target even for the search of complex real-world objects. The results in our work show that both feature and conjunction search can be performed by a single mechanism. It is the difference in effective guidance of attention towards the target that causes the difference in feature and conjunction search. If the attention is effectively guided by the target features towards the target's position (i.e. activation for target position is more than others in the attention map) in the search image in the first instance, a pop-out condition occurs as the object is found at the first gaze. In other conditions, the rank of attention activation that the target object gets, determines the search time. Further, effective guidance is dependent on the effective encoding of features and the difference between the representation of different search objects emerges as a strong influencer of the search efficiency.

5.1 Understanding optimal feature representation

In search of the letter 'T' amongst 'L', the full feature dot product used as a similarity measure could replicate the effect of pop out in feature search while a linear time increase was observed for conjunction search as the number of distractors were increased. The search for T in L with the same color is itself considered conjunction search because it is a complex shape consisting of two bars and their relative placement. As this search has a non-zero slope for the search time curve, it means that their respective activations in the attention map and hence their encoded features were not different enough to discriminate between them. This may indicate that more complex features, i.e. conjunction of features, may require better representation than available to the search process.

In tasks such as search within bars as stimuli, using a full feature dot product as a similarity measure allowed the target to be identified easily in the first place during both feature and conjunction search conditions. It can be observed that in this case, the target's activation in the attention map is much higher than of distractors. The confident identification or pop-out phenomenon in conjunction search, in this case, can be attributed

to the optimal representation of the object's apparent features like color and orientation in the encoded features that makes the discrimination effective between the target and the distractors.

For the purpose of visual search, we consider **a feature's representation is optimal if the difference in the encoded representation shall be enough that the two objects with a difference in apparent feature shall be able to be distinguished by the search process.** So the difference has to be enough that the resulting attention map activations have a difference more than whatever background noise exists in the process.

5.2 Effects of limited capacity of feature representations on visual search

As the model shows perfect search results even for conjunction search with bars as search objects, we introduced additional resource-based limitations. This limitation of memory is manifested in the model via modifying similarity score in two forms, both of which limited the precision of the feature representation.

Addition of noise to the target's representation is based on the limitation of precision in representing objects when more objects need to be remembered. The target's embedded feature representation is considered noisier during conjunction search, as more types of objects need to be remembered and compared. During the search experiment, it is noticed that this noise can cause the target to be confused with distractors in the attention map, resulting in loss of search efficiency and the increasing slope of the search time curve.

Another way we modelled the memory-based resource limitation is by limiting the length of the vector with encoded features. The dot product for generation of attention map then uses this truncated vector with lower number of encoded features. As part of the vector encoding features is truncated, the precision of the representation decreases. Further if the truncation is random, important encoded features that represent the variation amongst search objects can be lost.

We proposed a mechanism based that may possibly be used by the brain to retain important features that are relevant to the search. We use a process that pre-selects the features before the search in an image takes place. Those features are selected that produce the highest variance in the dot product, across various types of objects in the search image. In everyday life, this process is equivalent to remembering the most distinguishing aspect of the target from the objects anticipated in the scene, in case the

complete target shape is too complex to be remembered. These selected features are then used to calculate the similarity score using the dot product of the chosen features from the target and search image's encoding.

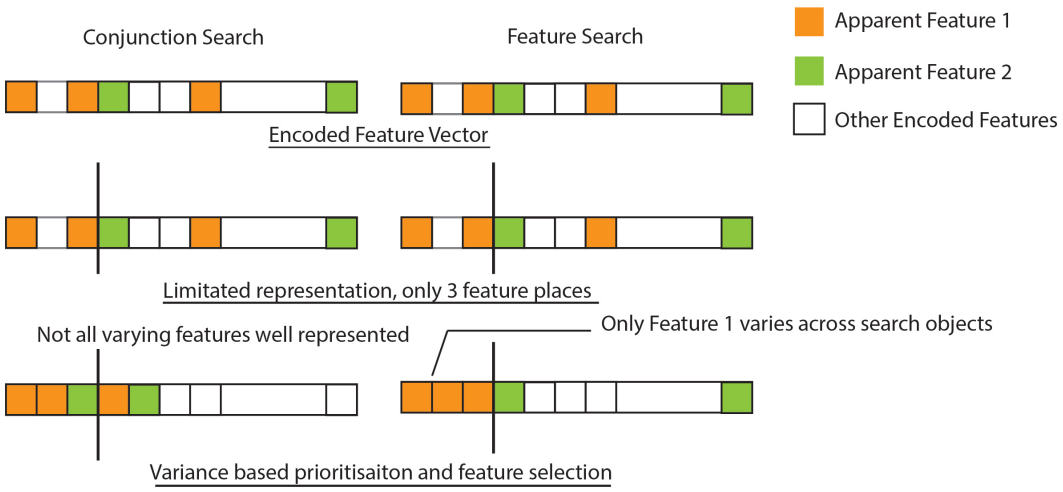


Figure 24. Explanation of search time characteristics in feature and conjunction search based on their different size requirements to encode features. Variance based feature selection allows for pop out effect in feature search even during resource constrains.

Taking an example of shape and color being the apparent features varying across the search objects. In Figure 24, we can see that the pop-out (feature search) and the conjunction search effects can be explained by the noticing that **a single apparent feature shall ideally need less representation space in the encoding vector than a conjunction of features.** So if only the most varying features are selected, in case of feature search the chosen features shall contain most of the encoded features representing the single apparent feature say color, hence be able to discriminate with the distractors. But in conjunction search, the high variance features can be from both shape and color, hence the complete encoding of those features that makes them different from distractors cannot be represented, hence the difference between the target object's activation and the distractor's activation will be lower, allowing for a possibility for distractor to have higher activation, when the inherent noise in the search process is taken into account.

Finally, the explanations for search time curves also need to take into account the actual difference in between the target and the distractors along a feature. If the target and the object are very close to each other along the apparent feature, the difference between their encoded features will also be very less and hence they will have similar attention activations. It is very likely that the attention selection process will not be able to clearly

discriminate amongst them, even if they differ by a single apparent feature. On the other hand, a very strong difference in features may cause the search to be fairly efficient even in the conjunction case and show a pop-out effect as happened in the case of bars using full dot product as similarity score.

Overall, **for each stimulus used, we were able to find possible constraints or conditions in which the same mechanism can produce the characteristic search time curves for feature and conjunction search. All the constraints and experiments point towards the role of encoded features quality and the ability to use the features to obtain higher target activation as causes for the difference in the search efficiency in different cases.**

5.3 Limitations of our Work

On the other hand, some of the limitations of this model and our proposed modifications are that, though we were able to create possible conditions that were able to achieve the desired search time results, they occured in a narrow range of conditions. The number of limited features in the encoded vector, the magnitude of added noise or the difference in magnitude of a feature had to be tuned. The tuned value of the limitations is somewhat different from the expected magnitudes in humans. As an example, for limitations introduced in length of vector with encoded features, the cut-off is set at 3 features in case of 'bars', which is very low as the original length of the vector is 512, and the non-zero valued encoded features in the encoding are also around 20 features. Another case of difference is that the magnitude of apparent feature, color in case of circles and rectangles for the distractors had to be tuned to be fairly close to the target to get the desired search time trends for feature and conjunction search. It is difficult that with the tuned colour for distractors in this case, that the same pop-out effect can be achieved with human subjects as the color can hardly be distinguished between the target and the distractors (Figure 21.a). Another possible limitation of this work is that we have not explored bottom-up salience, i.e. pop-out effect of certain objects in an image in the absence of a target.

On a high level, the narrow range constraints can be explained and justified by differences in encoding of features between humans and CNNs. CNNs may not exactly encode for the apparent features of interest to humans in the objects. They are trained for classification tasks and may **encode entirely different features in which we can only observe some projection of the simple apparent features.** A possible reason for a very short vector cutoff is the entanglement of apparent features in encoded feature vector, where a single encoded feature may encode values from two apparent features. Due

to this, **effectively the conjunction of apparent features in the search objects may be represented as a single encoded feature or an overlap in a vector of encoded features, making two features effectively closer to a single feature as represented for the use of search process.**

Irrespective of the limitation in correspondence to human feature representation and replicating the exact human performance, the computational model with the modelled resource limitations show that search performance is dependent on the correct guidance of attention, by the top-down target. Various effects such as feature and conjunction search are a manifestation of change in the ability to guide attention using the encoded object features. **A loss in precision or limitation in the feature encoding capacity when more features need to be encoded is a possible underlying principle that gives rise to the increased search time in conjunction search.**

6 Conclusion

This work builds on the computational model of visual search proposed by Zhang et al., 2018 [45]. By simulating visual search experiments, we explored the model's performance to search for a target in feature and conjunction search. We modified the similarity product used in the computational model, to explore manifestations of resource constraints. The modified similarity product also tried to model how the system would cope up with such resource constraints. Overall we explored a possible way in which feature search and conjunction search can occur using the same search process and try to probe factors that affect attention guidance and how it affects the search performance.

We started with the FIT [38] that proposes parallel processing of all objects in feature search and a serial search process in conjunction search. The guided search model by Cave& Wolfe, 1990 [2] and the computational model by Zang et al, 2018 [45] pointed out that search performance depends on correct guidance of visual attention. Our simulation results show that **the ability to discriminate between the objects using the object's feature representation has a major influence on attention guidance and how efficiently the attention is guided towards the target affects the search performance. Further, if the resources to store feature representations are limited, this may lower the discriminative power for the attention guidance mechanism. This occurs because the limited representation capacity decreses the precision of the feature encoding. In both our proposed modifications to the attention map generation process that produce the correct characteristic graphs for feature and conjunction search, it is the lower precision of feature encoding in conjunction search due to more types of objects in the image or more number of discriminating features that cannot be properly represented that ultimately cause its lower search efficiency than in feature search.** Overall, the quality of the feature encoding was identified as playing the most influential role in correctly being able to guide the visual attention towards the target. The search performance in turn is directly dependent on this correct guidance of attention towards the target in the search image.

To build on this work, more notions of object similarity can be explored which may not directly depend on correlation of features but may explore other notions of distance between representations. A good guide in this direction can be to understand other known and plausible notions of distance or similarity calculations in the brain. To overcome our model's limitations, further efforts can be made on making the feature encoding process more similar to humans with disentangled feature representation, which shall enhance the coverage of the results. More experiments can be done to try and replicate the results for a range of experiments in literature, such as the easy finding of target if distractors are more similar to each other [30], easier detection of the presence of feature than absence

[37] and other such experiments that depend on variation in feature and its effect on search performance. Work in this direction can facilitate a more robust understanding of the used computational model and its applicability, and further help to understand the process of attention selection in how it guides the visual search.

References

[1] Charles F Cadieu, Ha Hong, Daniel LK Yamins, Nicolas Pinto, Diego Ardila, Ethan A Solomon, Najib J Majaj, and James J DiCarlo. Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS Comput Biol*, 10(12):e1003963, 2014.

[2] Kyle R Cave and Jeremy M Wolfe. Modeling the role of parallel processing in visual search. *Cognitive psychology*, 22(2):225–271, 1990.

[3] Kyunghyun Cho, Aaron Courville, and Yoshua Bengio. Describing multimedia content using attention-based encoder-decoder networks. *IEEE Transactions on Multimedia*, 17(11):1875–1886, 2015.

[4] Marvin M Chun. Visual working memory as visual attention sustained internally over time. *Neuropsychologia*, 49(6):1407–1409, 2011.

[5] Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6:27755, 2016.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[7] Don C Donderi and Dorothy Zelnicker. Parallel processing in visual same-different decisions. *Perception & Psychophysics*, 5(4):197–200, 1969.

[8] John Duncan and Glyn W Humphreys. Visual search and stimulus similarity. *Psychological review*, 96(3):433, 1989.

[9] Howard E Egeth. Parallel versus serial processes in multidimensional stimulus discrimination. *Perception & Psychophysics*, 1(4):245–252, 1966.

[10] Krista A Ehinger, Barbara Hidalgo-Sotelo, Antonio Torralba, and Aude Oliva. Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual cognition*, 17(6-7):945–978, 2009.

[11] Wolfgang Einhäuser, Merrielle Spain, and Pietro Perona. Objects predict fixations better than early saliency. *Journal of Vision*, 8(14):18–18, 2008.

[12] Lior Elazary and Laurent Itti. A bayesian model of visual search and recognition. *Journal of Vision*, 8(6):841–841, 2008.

[13] Umut Güçlü and Marcel AJ van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

[14] John M Henderson, James R Brockmole, Monica S Castelhano, and Michael Mack. Visual saliency does not account for eye movements during visual search in real-world scenes. In *Eye movements*, pages 537–III. Elsevier, 2007.

[15] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203, 2001.

[16] Christopher Kanan, Mathew H Tong, Lingyun Zhang, and Garrison W Cottrell. Sun: Top-down saliency using natural statistics. *Visual cognition*, 17(6-7):979–1003, 2009.

[17] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11), 2014.

[18] Arni Kristjánsson. Reconsidering visual search. *i-Perception*, 6(6):2041669515614670, 2015.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[20] Jonas Kubilius, Martin Schrimpf, Aran Nayebi, Daniel Bear, Daniel L. K. Yamins, and James J. DiCarlo. CORnet: Modeling the Neural Mechanisms of Core Object Recognition. *bioRxiv*, page 408385, 2018.

[21] Michael Land, Neil Mennie, and Jennifer Rusted. The roles of vision and eye movements in the control of activities of daily living. *Perception*, 28(11):1311–1328, 1999. PMID: 10755142.

[22] Michael Land and Benjamin Tatler. *Looking and Acting*, volume 1. Oxford University Press, 2009. An optional note.

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[24] Grace W Lindsay. Attention in Psychology , Neuroscience , and Machine Learning. 14(April):1–21, 2020.

[25] Wei Ji Ma, Masud Husain, and Paul M. Bays. Changing concepts of working memory. *Nature Neuroscience*, 17(3):347–356, 2014.

[26] Ulric Neisser. Decision-time without reaction-time: Experiments in visual scanning. *The American Journal of Psychology*, 76(3):376–385, 1963.

[27] Aude Oliva and Antonio Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–36, 2006.

[28] Laura W Renninger, James M Coughlan, Preeti Verghese, and Jitendra Malik. An information maximization model of eye movements. In *Advances in neural information processing systems*, pages 1121–1128, 2005.

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[30] Ruth Rosenholtz. Search asymmetries? what search asymmetries? *Perception & Psychophysics*, 63(3):476–489, 2001.

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[32] Benjamin W Tatler, Roland J Baddeley, and Iain D Gilchrist. Visual correlates of fixation selection: Effects of scale and time. *Vision research*, 45(5):643–659, 2005.

[33] Benjamin W Tatler and Nicholas J Wade. On nystagmus, saccades, and fixations. *Perception*, 32(2):167–184, 2003.

[34] Antonio Torralba and Aude Oliva. Statistics of natural image categories. *Network: Computation in Neural Systems*, 14(3):391–412, 2003.

[35] Antonio Torralba, Aude Oliva, Monica S Castelhano, and John M Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, 113(4):766, 2006.

[36] Anne Treisman. The binding problem. *Current opinion in neurobiology*, 6(2):171–178, 1996.

[37] Anne Treisman and Janet Souther. Search asymmetry: A diagnostic for preattentive processing of separable features. *Journal of Experimental Psychology: General*, 114(3):285, 1985.

[38] Anne M. Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97 – 136, 1980.

[39] Nicholas Wade, Benjamin W Tatler, et al. *The moving tablet of the eye: The origins of modern eye movement research*. Oxford University Press, USA, 2005.

[40] David Whitney and Dennis M. Levi. Visual crowding: A fundamental limit on conscious perception and object recognition. *Trends in Cognitive Sciences*, 15(4):160–168, 2011.

[41] LG Williams. The effect of target specification on objects fixated during visual search. *Perception & Psychophysics*, 1(5):315–318, 1966.

[42] Jeremy M. Wolfe. Chapter 8 - visual attention. In Karen K. [De Valois], editor, *Seeing*, Handbook of Perception and Cognition, pages 335 – 386. Academic Press, San Diego, 2000.

[43] Jeremy M. Wolfe and Igor S. Utochkin. What is a preattentive feature? *Current Opinion in Psychology*, 29:19–26, 2019.

[44] A. L. Yarbus. *Eye Movements and Vision*. Plenum. New York., 1967.

[45] Mengmi Zhang, Jiashi Feng, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Gabriel Kreiman. Finding any Waldo with zero-shot invariant and efficient visual search. *Nature communications*, 9(1):3730, 2018.

Appendix

I. Access to code

The code used to obtain the results can be found in this GitHub repository given below:

`https://github.com/tarunkhajuria42/Attention_V_Search.git`

II. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Tarun Khajuria**,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   A unified account of visual search using a computational model,

   supervised by Raul Vicente and Jaan Aru.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Tarun Khajuria
15/05/2020