Visual Attention in Object Classification. The Relationship between Attention in Humans and Deep Networks

Michal Gacka
bfj541@alumni.ku.dk,
David Oppenberg
lrb737@alumni.ku.dk, and
Cristian Mitroi
dmn470@alumni.ku.dk

University of Copenhagen

Abstract. This work explores different attention mechanisms in an object recognition setting with the POET dataset. We introduce a novel approach to constructing heatmaps visualizing human attention, formalize object recognition as a sequential task, and employ an evaluation scheme that proves to distinguish between computational attention mechanisms in relation to human attention. Finally, we use sequential fixations to guide a machine learning model and draw conclusions about the foundational reasons for human effective attention range as a function of eccentricity from the fixation.

1 Introduction

Visual attention is one of the most fundamental cognitive processes. Recently, attention became a resource-efficient enhancement of machine learning algorithms. It is applied in machine translation, object recognition, image captioning and visual question answering.

There are many parallels between research theories of human visual attention and attentional computational models. By proposing a heatmap visualization technique based on gradient theory, we compare human eye-tracking data with attention of deep networks in the context of an object recognition task. We use the POET dataset, which provides data of sequential human gaze fixations.

We also propose a model that is guided and enhanced by data of human gaze. We show that it performs better than the baseline. From the experiments we make connections to the cognitive theories of human attention.

We note that due to the limited length of this report, we had to prioritize and limit the scope.

1.1 Attention

The brain's capacity for processing sensory information is limited. It therefore uses attention to focus on specific aspects of the realm of sensory input. Attention can therefore be defined as the ability to select a particular stimulus for increased scrutiny [1].

1.2 Visual attention in humans

As early as the 19th century, Von Helmholtz was the first to define eye movements as evidence of visual attention [2]. He was mostly concerned with externally observable manifestations of attention. William James pioneered the idea of attention as an internal process. James conceptualized a more active idea of attention, one that is guided by expectations and a desire to extract meaning from stimuli [3]. The two positions, by Von Helmholtz and James, fundamentally defined the contemporary notions of bottom-up, prioritizing intrinsic visual features, and top-down approaches, prioritizing more abstract features.

Anne Treisman united the top-down and bottom-up theories by proposing a two-fold system consisting of a bottom-up filter that first marks the sensory input and then passes it on to a threshold-based regulating system that processed the marked input according to high-level relevance and context [4].

1.3 Visual attention and Eye Tracking

Eye tracking experiments were conducted as a response to generate evidence for the various bottom-up and top-down theories. Another research question that inspired eye tracking experiments is the scene integration problem, which deals with the question of how the human mind takes the filtered patches of information and forms a coherent scene of the visual field [5].

Fig. 1: Early Eye-Tracking experiment by Yarbus visualizing sequential fixation points [5]

Posner's spotlight theory proposed the dissociation of the human attention mechanism and the ocular fixation of the eye. The idea that it is possible to visually fixate one location while simultaneously diverting attention to another poses a big problem for eye-tracking-based research [6]. Therefore, a well known limitation of contemporary eye-tracking devices is their inability to accurately track covert shifts of attention.

1.4 Visual attention in machine learning

Attention in the context of Machine Learning describes the ability of an algorithm to focus available processing power on a subset of the input data or to prioritize it in some way over the rest.

It can either be implicitly built into an algorithm (e.g. CNNs can make a decision based on high activations corresponding to a small region in the image) or it can be explicitly added to an algorithm (e.g. a weighted sum in the soft attention can prioritize a patch of the image over the rest of it).

1.5 Research questions and hypotheses

In this paper we first explore multiple computer vision attention mechanisms in order to: show how cognitive processes can inspire computational models and improve their performance and interpretability; and remark on similarities and differences between human and machine attention.

We hypothesize that visual search, underlying object recognition, requires the model to extract information from similar areas as humans do when performing visual search. Thus a significant correlation should exist between the areas that the computer and humans find important. We show that this is dependent on the model, prove the relevance of attention mechanisms in computer vision, and draw secondary conclusions about similarities and differences between human and machine attention.

Secondly, we hypothesize that, given the locality of important information in images in an object detection task, we can use human eye-tracking data to guide a computational model. We show that computational models can hint at why human attention is effective in a certain area around the fixation.

1.6 State of the Art

In the paper "Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions?", the Facebook AI Research team compares human and computational visual attention in the context of Visual Question Answering (VQA) tasks. The research compares human attention to the regions of interests generated by Stacked Attention Networks (SAN) and Hierarchical Co-Attention Networks (HieCoAtt). They conclude that VQA attention models such as SAN and HieCoAtt do not seem to be 'looking' at the same regions as humans to produce an answer [7]. Though attention-based VQA models become more accurate, they seem to be (slightly) better correlated with humans in terms of where they look[7].

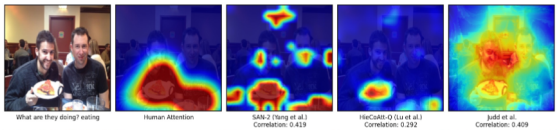


Fig. 2: Heatmap of attended regions for VQA models

The paper "Bottom-Up and Top-Down Attention for Image Captioning" by Anderson et al. proposes a model that combines the context-specific top-down soft attention mechanism in combination with a bottom-up R-CNN. Their model consistently outperforms the ResNet baseline on image captioning tasks [**?**].

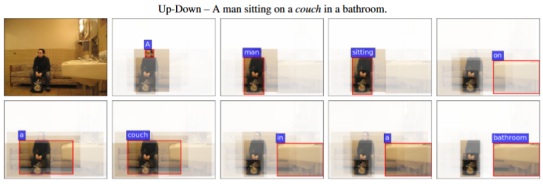


Fig. 3: Attended image regions of model by Anderson et al. for image captioning

2 Theoretical Background

2.1 Eye tracking metrics

Conventional eye-tracking metrics include the following relevant units [8]:

- Fixations: discrete samples of almost stable points where the eye is looking.
- Fixation duration: how long does the eye look at a stable point before it shifts attention.

– Time to first fixation: how long until the first fixation is recorded.
– Inter-observer consistency: measures similarity of observer fixation patterns on an image.
– Saccades: eye movement between fixations.
– Heatmaps are a popular technique for visualizing the distribution of gaze points.

2.2 Covert and overt attention

Shifts of attention can occur either covertly, through eye movement or overtly, during ocular fixations. Covert eye movement involves rapid, saccadic eye movements to bring the fovea onto an object in order to visually process it [9]. Overt attention can shift without any movement of the eye.

2.3 Gradient Theory

Humans are able to categorize complex natural images at a glance over a large extent of the visual field without eye movement. The field of attention (FA) surrounding the focus point of the eye allows humans to e.g. categorize briefly flashed photographs [10]. The same study provides evidence to prove that the quality of attention deteriorates as the function of eccentricity from the fixation point.

2.4 CNNs, RNNs, and transfer learning

We use Long-Short Term Memory units [11] and transfer learning techniques together with CNNs. We explore them in more detail in the appendix C due to the limited space of the report.

2.5 CNN interpretability

Despite the great performance of CNNs, their interpretation is an active research area. In this paper we focus on methods that allow for interpretation by highlighting the areas that the network decides are important in the task.

They allow for interpretation as a byproduct of their main design goal. Class Activation Mapping [12] method is developed for the purpose of localizing objects within the scene and Soft Attention was made to allow the network to sequentially pay attention to different parts of the image and generate a locally-dependent output.

2.6 Attention in RNNs

For RNNs, the attention was first developed to prioritize different parts of the input sequence on different processing steps, giving the network the ability to

sequentially extract and relate information that was arranged spatially or temporally.

Later the concept was adapted to visual inputs, where the attention allows to prioritize information in an input that does not inherently have a temporal dimension and that would normally be processed all at once. In soft attention, as proposed in the paper "Show, Attend and Tell" [13], the network can reinforce the importance of a sub-region of the image on each step. The inputs are multiplied by an attention map to prioritize selected spatial locations depending on what has been seen and predicted in previous steps. We discuss the model in more detail in the subsection 4.4.

3 Dataset and Methodology

3.1 Data Collection

This work is based on eye tracking data from the Pascal Objects Eye Tracking dataset, where the participants were asked to perform a visual search task while their eye movements were tracked.

A visual search task was chosen as it encourages fixation on target objects. Participants were prompted with an image which contains one of two object classes (e.g. cow or horse). By the press of a button they could identify the correct class. By using an object discrimination setup, data for two classes could be collected in parallel.

We describe the details of the setup in appendix B.

3.2 Dataset

The dataset containts a total of 6270 unique images distributed over 10 corresponding object classes. Each image is annotated with the eye movement data of five participants. For each participant the following data is provided [14]:

- scrnCoord - screen coordinates of the eye tracking data
- imgCoord - image coordinates of the eye tracking data
- fixR - fixations of the right eye
- fixL - fixations of the left eye

  These then contain the following for each fixation:
  - time - initial and final time of the fixations in ms
  - pos - x and y coordinates of the fixation positions.

3.3 Statistics and data quality assessment

We observe a significant class imbalance (fig. 5). The cow class is the least represented. This is most likely one of the reasons for why the class is confused by the models (fig. 22 and 23).

Fig. 4: Example of image from POET dataset and annotated fixations

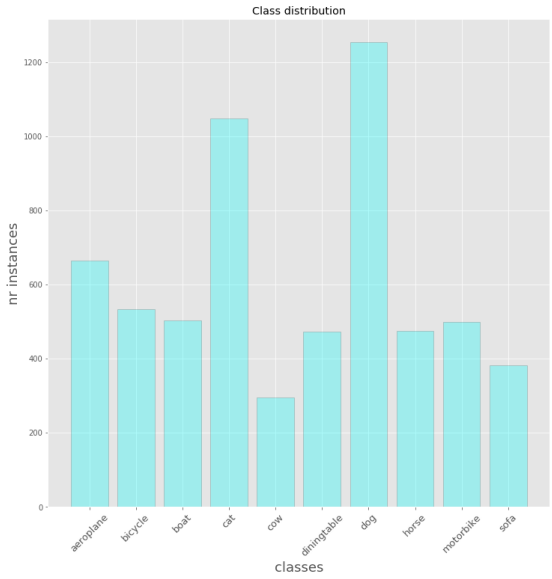


Fig. 5: Class distribution

We define usable fixations as within the bounds of the image and not *None*. We notice that cow and horse stand out in the data with a slightly higher number of fixations (fig. 6).

We see that images in the cow and horse classes have the longest response times 7. This correlates with the mean number of fixations for these classes. These two classes are perhaps easily mistaken by most users. We can see that this problem reflects in our models (fig. 22 and 23).

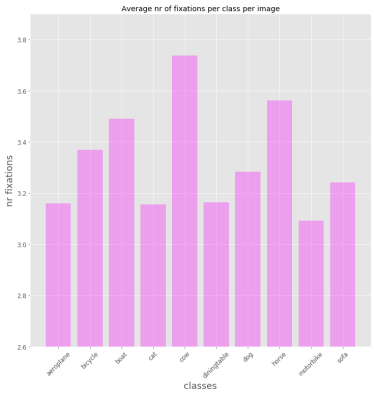


Fig. 6: Average number of fixations per image per class

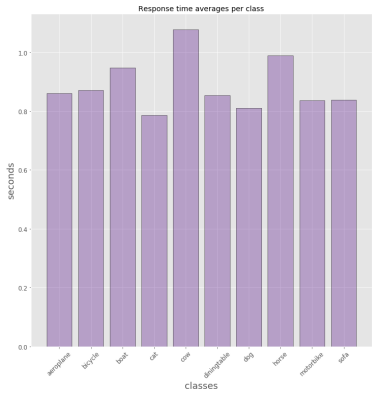


Fig. 7: Average response time per class per image

Quality of samples While browsing the collection of photos we notice problems with the data. Some of the examples provided were of poor quality or were not representative of the class. Consider the examples in appendix D. We believe this greatly hinders the learning capacity of our models. We also noticed that within the boat category there is a wide variety of boats: from rafts to cruise liners. This is surely confusing to the models: each type has different visual patterns.

Furthermore the authors of the dataset did not control for biases such as center bias, which is a common occurrence in human attention [15].

4 Field of attention, experimental setup, and models

The authors of the dataset use an arbitrary distribution relative to the image size around the fixation points when building the heatmaps, without accounting for the actual area from which the information was extracted. We try to improve on that. We explore a relevant study and propose a novel technique for constructing eye-tracking heatmaps and comparing them to machine attention.

4.1 Field of attention - ground truth heatmap

We use the concept of the field of attention as described by Yao et al. [10] to build a Gaussian distribution around fixations that approximates the area that was relevant to the performed cognitive task.

Participants of the experiments performed by Yao et al. were asked to recognize objects rapidly. Those were placed in different places on the screen, while participants were asked to keep their gaze in the center of the screen. That way, the authors were able to measure how the quality of the retrieved information deteriorates as a function of the eccentricity (angle between the fixation point, participant's eyes, and the image) from the fixation point.

In fig. 8 we see the gradient theory in practice. The accuracy of object recognition is the highest for images appearing up to around 2.5 degrees from the fixation point and then quickly deteriorates.

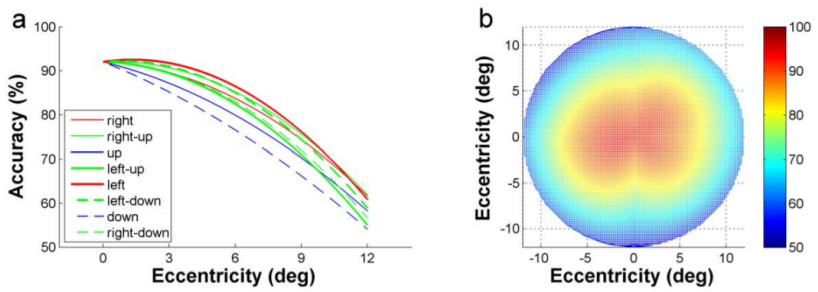


Fig. 8: Experiments by Yao et al. [10] show accuracy in object recognition task as a function of eccentricity - in practice that means how far relative to the fixation points images are presented to the participant.

In order to visualize a plausible area of effective attention around fixation points, we combine the above with information about the experimental setup in POET dataset, and some useful probability theory. It is important to note that the area from which the information was retrieved by a fixation should not be dependent on the image size.

In order to simplify our calculation we make the assumption that regardless of where in the screen the image is presented, the fixation point forms a 90 degree angle with the screen and human eye.

The participants in the POET dataset were seated 60cm from a 22inch screen. The size of a pixel in such screen is approximately 0.28mm [16]. We thus calculate the area in pixels around the fixation in 2.5cm eccentricity radius:

$$P = \frac{d \cdot \tan(\epsilon)}{s}, \tag{1}$$

where P is the effective field of attention in pixels, d is participant's distance from the screen, ϵ is the eccentricity (in radians), and s is the size of one pixel.

$$\frac{0.6 \tan(\frac{2.5\pi}{180})}{0.00028} = 93.56 \tag{2}$$

We conclude that the POET participants were accurately extracting information from an area of approximately 93 pixels. We use the full width half maximum parameter F of a Gaussian distribution, also known as the effective radius to form distributions around the fixation points [17]. Given that the relationship between the standard deviation of a distribution σ and F parameter is given by $F = 2\sqrt{2 \ln 2}\sigma$, for our distribution $\sigma = 39.72$ pixels.

It is important to note that the distributions are constructed, summed, and normalized for all participants. Thus the result is an average across participants.

4.2 Experimental setup

Firstly, we decide on a specific split between train and validation data to be able to make comparisons between the models. All models are trained using

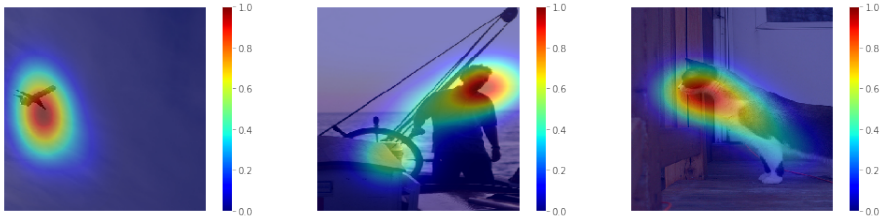


Fig. 9: Examples of the human attention visualized for an airplane, boat and cat images. Higher value corresponds to information being more critical for an object recognition task.

gradient descent, more specifically a well-established ADAM optimizer [18]. We also remove images that appear in more than one class.

In the subsequent experiments the computational model is solving a task of classifying images into 10 classes. We claim that both POET participants and our models have to fundamentally perform a similar visual search in order to solve their task accurately, which makes the attentions comparable.

We train four models divided into 2 experiments. Class Activation Map Model and Soft Attention Model for comparing human and machine attention; and the CNN Baseline Model and patch-based CNN+LSTM for using eye-tracking data in machine learning.

For CAM and patch-based LSTM we use *Keras*. For the Soft Attention model, because of its complexity, we use *Tensorflow*. We use 2 GPUs: GeForce 1070ti and GeForce 970.

4.3 CAM Model

First approach uses the fact that a CNN's activation is location-dependent throughout the layers: given a position in a feature map we know exactly which part of the image it had access to [19].

Zhou et al. [12] use the last layer containing N feature maps to visualize which areas of the image the network prioritizes on the last step before classification. A simple modification to the usual CNN architecture is being introduced at the end: Global Average Pooling (GAP), which summarizes each feature map with one number, resulting in a feature vector $V = \{v_0, v_1, ..., v_n\}$ of length N.

A dense layer can be thought of as computing a weighted sum of numbers in the feature vector for every class (fig. 10). Every class will therefore prioritize different locations in that feature vector, thus prioritizing different feature maps from which these vectors were computed.

For one class c the activation P_c (before the softmax function is applied) can be calculated as:

$$P_c = \sum_{i=0}^{N-1} w_{c,i} V_i \tag{3}$$

For every class a heatmap can be created by reversing the above logic and computing a sum over the feature maps before Global Average Pooling was

performed, weighted by the weights that the network assigned to the feature vector for that specific class (fig. 10). The heatmap is then resized back to the original image resolution.

We interpret both positive and negative neural activations as equally important and take an absolute value of the aforementioned weighted average.

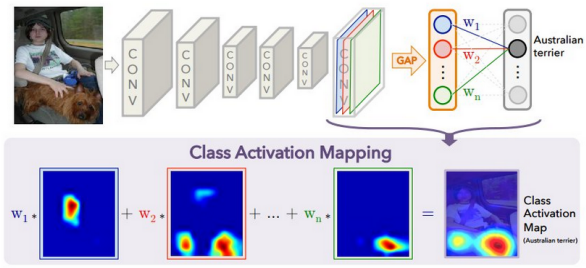


Fig. 10: Class Activation Map explained. Source: [12]

This way we arrive at a heatmap that defined which regions of the image the network prioritized to make a judgment about a specific class. In our later results we usually refer to the CAM heatmap as the heatmap of the class with highest probability.

A VGG16 network is used as the deep architecture before GAP and it produces 512 feature maps of size 28x28. We use weights from VGG16 pretrained on ImageNet without the last pooling layer. We add the GAP and a dense, softmax layer with 10 neurons that allow for performing classification and extracting CAM.

As a base for the model we used an open source repository and adaped it accordingly [20].

4.4 Soft Attention Model. Object recognition as a sequential task

The CAM Model is a slight modification of a usual CNN utilizing an implicit mechanism within it. Soft Attention, on the other hand, models attention explicitly, giving the network the ability to actually attend to different parts of the image depending on the stage it is at in the task.

Our approach is inspired by the "Show, Attend, and Tell" paper, which uses attention for image labelling [13]. Here the task enforces a temporal dependency and the network shifts attention depending on the input data and what words were generated before.

We focus on differentiable, Soft Attention where the entire image has to be processed and different regions are prioritized, because it can be optimized using gradient descent methods.

Attention is an inherently temporal process but object recognition as performed by CNNs is not. Thus the Soft Attention cannot be applied to object recognition task in the form that it is formalized in usually.

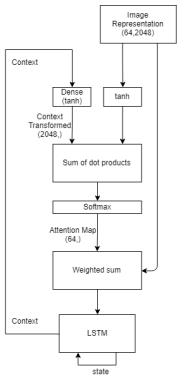


Fig. 11: Attention Module

Inspired by human attention, we model the object recognition task as a sequential process, where the network can "look" at the image multiple times. This then gives us the ability to see which regions of the image the network prioritizes at each step and how that influences the predicted probabilities over the classes.

We divide the image into 64 regions and use a ResNet50 model pretrained on ImageNet to extract a separate vector of 2048 numbers for each subregion. The image is thus represented as a 64x2048 matrix and the goal of the attention mechanism is to calculate a weighted average that will compress that to 2048 values at each step of the object recognition task.

In order to produce an importance distribution over the 64 areas in the image, the output of the LSTM (context) is fed through a dense layer and combined in a sum of dot products before applying softmax. The result is then used to compute a weighted average of the 64 feature vectors representing the image, before being delivered to the LSTM (fig. 11).

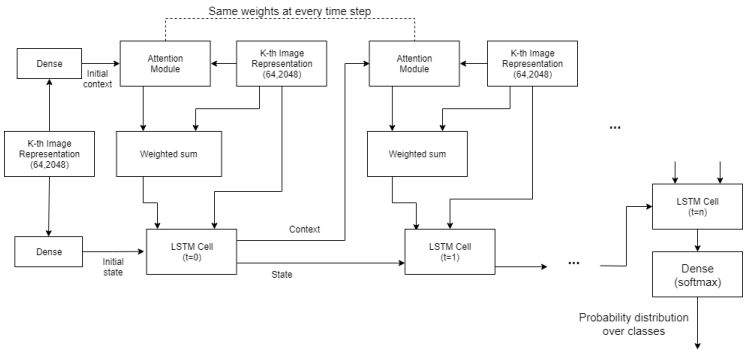


Fig. 12: Soft Attention model architecture unpacked in time. In this paper the results were obtained with n=3.

Dense layers are used to learn how to initialize the context and the state of the LSTM depending on the input. The attention module as described in fig. 11 produces an importance distribution over the input vectors that is used to

compute a weighted average later passed to the LSTM at each step. In our case, the model "looks" at the image 3 times (fig. 12).

In order to obtain a final heatmap, we average the attention maps over the 3 time steps that they were generated for.

4.5 Baseline CNN and Patch-based LSTM Model

Baseline We first construct a traditional, full-image, one-step, baseline model for comparison. We use a pretrained ResNet50 to extract features from full images and build a simple classification network out of 2 dense layers.

Patch-based LSTM We then propose a model that is guided by human attention. We use the fixations from the POET data set in order to extract square patches centered on the fixations. These are then passed to the pretrained ResNet50 in order to obtain a feature vector representing that patch. Finally, we build an LSTM that sequentially looks at each of the patch-representations and then outputs a probability distribution over the 10 classes.

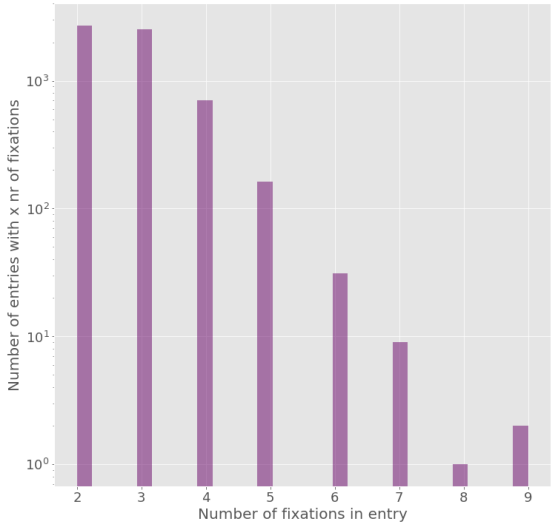


Fig. 13: Histogram: number of entries with x fixations

We believe this is a valid approach because information in an image is *localized*. This means that looking at the entire image is not *necessary* in order to make an accurate classification. Rather, looking at *small* areas of meaningful information is enough to reach the same performance as by looking at the *entire* image.

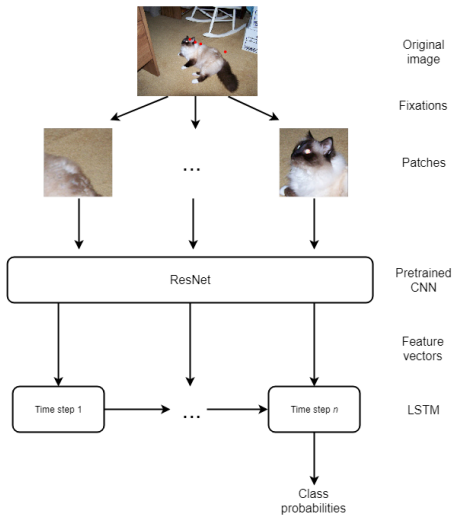


Fig. 14: LSTM model architecture

We did not observe any meaningful difference between the quality or the number of fastest and slowest users' fixations, thus, we choose fixations of the fastest user with at least two valid fixations, and then calculate an average between the left- and right-eye fixations. A valid fixation is not *None* and is within the bounds of the image. Given the histogram in fig. 13, we truncate or pad the sequences to the length of 5.

5 Results

In the first part of the experiments we train two types of networks exemplifying mechanisms that can be compared to human attention: CAM Model, where the CNN implicitly assigns different importance to different regions of the image depending on the class it is predicting and "looks" at it only once (explained in section 4.3); and Soft Attention Model, where the network has the ability to assign that importance explicitly at each step and has the chance to "look" at the image and shift its attention 3 times (explained in section 4.4).

In the second part of our experiments we use human eye-tracking data to guide model's focus (explained in section 4.5).

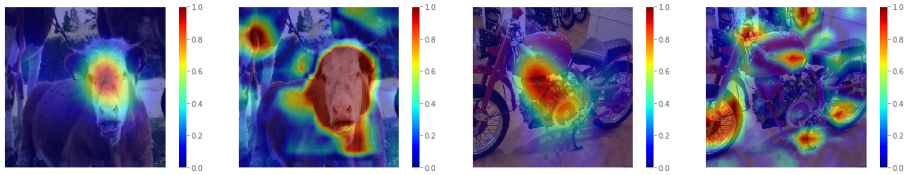
5.1 Human and machine attentions compared

Both the CAM and the Soft Attention Models achieve over 80% accuracy on the test set, thus have achieved a good generalization ability. Since we focus on the comparison of the attention mechanisms, we do not dive into model's training and performance.

Evaluation metric: Pearson Correlation Coefficient In order to perform a strong quantitative analysis of how similar human and machine attentions are, we define our evaluation metric, Pearson Correlation Coefficient PCC_{HG} between a model-generated heatmap H and human heatmap G (obtained as described in section 4.1) as follows:

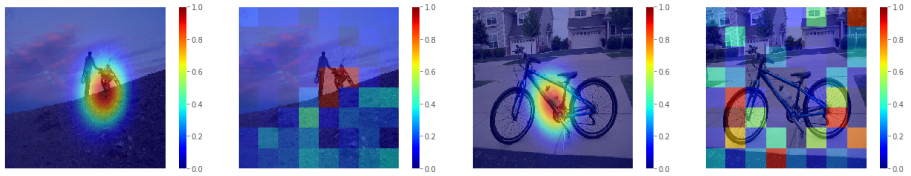
$$PCC_{HG} = \frac{\sum_{i=1}^{n} \left(h_i - \overline{h}\right)\left(g_i - \overline{g}\right)}{\sqrt{\sum_{i=1}^{n} \left(h_i - \overline{h}\right)^2} \sqrt{\sum_{i=1}^{n} \left(g_i - \overline{g}\right)^2}}, \tag{4}$$

where h_i and g_i are the elements of the heatmaps of size $W \times H$ flattened into a 1-dimensional vector, thus $n = W \cdot H$.



Baseline *PCC = 0.91* *Baseline* *PCC = -0.23*

Fig. 15: Examples of CAM heatmaps highly correlated (0.91) and negatively correlated (-0.23) with human attention



Baseline *PCC = 0.63* *Baseline* *PCC = -0.25*

Fig. 16: Examples of Soft Attention heatmaps highly correlated (0.63) and negatively correlated (-0.25) with human attention

In figures 15 and 16 examples of heatmaps with extreme values of PCC are presented next to their human baseline. The PCC by definition varies from 1 (perfectly positively correlated) to -1 (perfectly negatively correlated), thus value 0.91 is an almost perfect correlation. And since the heatmaps are composed only of positive values, negative correlation is equivalent to the attention being focused in completely different regions.

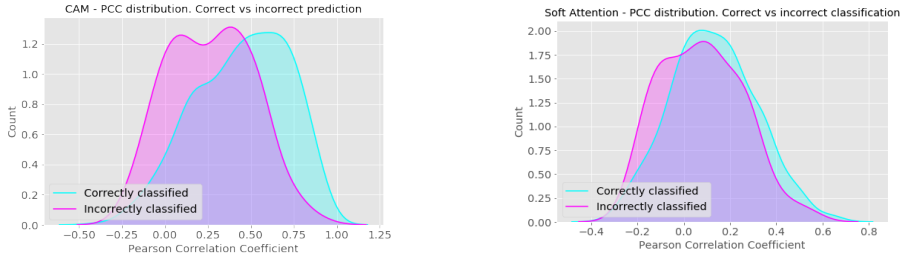
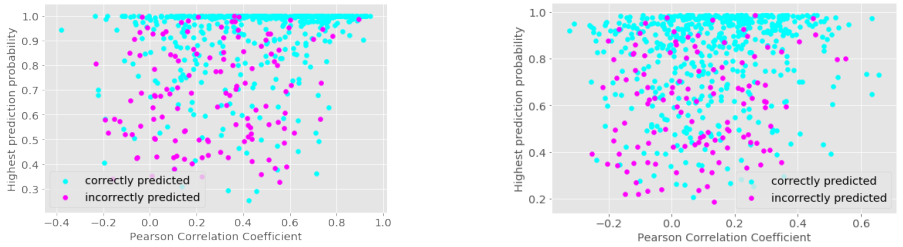


Fig. 17: Distributions of PCC labeled by the correctness of classification. CAM compared to human attention heatmap on the left; and Soft Attention compared to human attention on the right



CAM *Soft Attention*

Fig. 18: Join distributions of PCC and highest predicted probability in CAM and Soft Attention

Quantitative analysis Fig. 17 clearly indicates that CAM attention is much closer to the human attention than the Soft Attention with a difference on average of approx. 0.3.

Furthermore, CAM has the mean of the distributions for incorrectly and correctly classified samples shifted significantly relative to each other. The heatmaps of the correctly classified samples are much more correlated with human heatmaps with approx. 0.35 higher PCC on average. Thus the error of the network is highly correlated to the network focusing on the wrong part of the image.

This is not the case for the Soft Attention, where the means are almost equal, thus indicating that when Soft Attention makes a correct prediction (which it does in over 80% of the cases) it must be focusing on different information than humans are.

From the joint distributions (fig. 18) we see that there is no significant correlation between the confidence of the model (highest predicted probability) and the PCC in the Soft Attention Model. It is the case though, as expected, for the CAM Model, where the PCC between the highest probabilities for each samples and PCC of their heatmaps is equal to approx. 0.3, indicating a relevant correlation. This can be seen on the leftmost plot, where the population is denser as we move from the bottom-left to the top-right corner of the plot.

The joint distributions also reveal an interesting property of the Soft Attention Model: the probability of the predicted class is more uniformly spread over the entire range and there are fewer errors with high confidence. The CAM Model tends to be overly certain about its predictions (high accumulation of correctly predicted samples on the top of the plot and errors spread uniformly along the y axis).

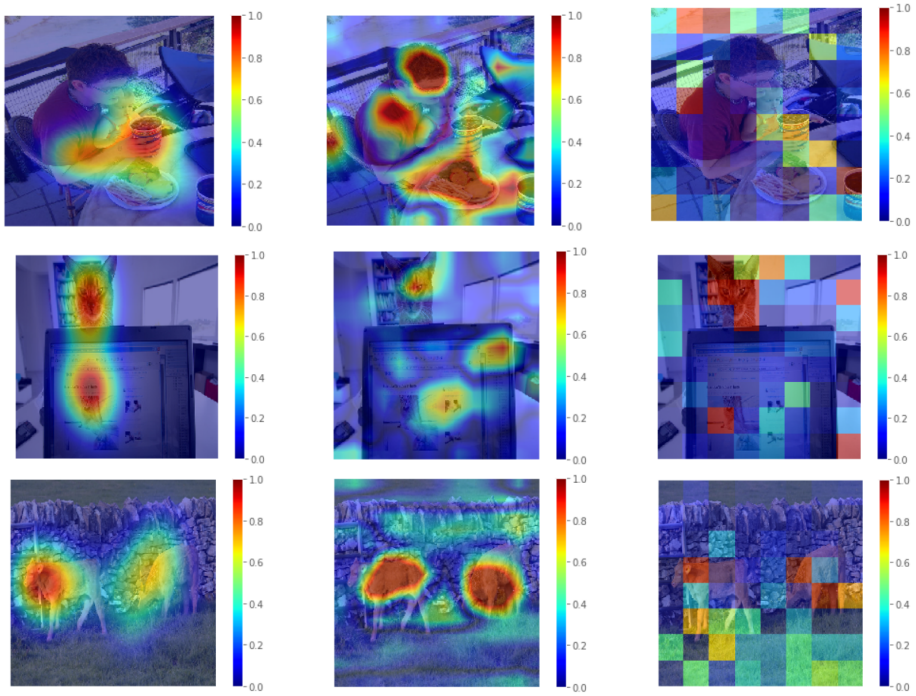


Fig. 19: Correctly classified examples of (from the top) dinning table, cat and cow classes. From the left to right, presented are: human attention, CAM attention, Soft Attention. PCC for the CAM heatmaps (top to bottom): 0.34, 0.19, 0.48; PCC for the Soft Attention heatmaps: 0.02, 0.23, 0.22

	CAM	Soft Attention
PCC	0.30, 0.32, 0.02	-0.07, -0.07, 0.05
Highest probability	0.90, 0.91, 0.79	0.68, 0.35, 0.67
True class probability	0.04, 0.02, 0.20	0.26, 0.18, 0.27

Table 1: PCC and confidence for the incorrectly predicted samples (top to bottom) from fig. 20

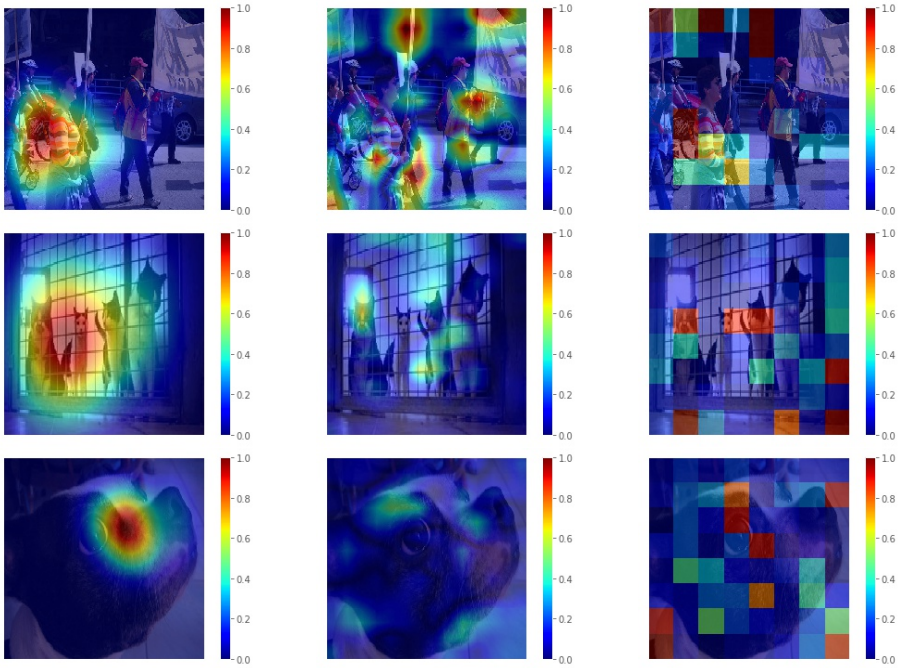


Fig. 20: Incorrectly classified examples of (from the top) bicycle, cat and dog classes. From the left to right, presented are: human attention, CAM attention, Soft Attention. PCC and certainty are summarized in tab. 1

Qualitative analysis We first analyze the correctly predicted samples in fig. 19 and note that the CAM attention indeed is more similar to human attention than Soft Attention.

From the top example we see that where the human fixates on the salient parts in the middle of the image, the network "notices" all important objects separately. We can see a similar pattern in fig. 16 and 15, where the model notices both wheels of bikes separately. Humans look between the wheels and extract information about the entire object.

This indicates that the decay of the quality of attention with larger eccentricities might be dependent on the object or task: since looking in the middle of a bicycle, we know what to expect on its edges, it is easier to fill in the blanks even far from the fixation point.

We also note, that both CAM and Soft Attention extract much more background information explicitly, while humans focus on the most important objects and most likely extract the background information implicitly, without looking at it directly.

Both CAM and Soft Attention can successfully identify most important objects in the image but Soft Attention seems to give more importance to the background information.

We look into the incorrectly classified examples in fig. 20 and their statistics in table 1 and note that Soft Attention, even though qualitatively more successful at finding the important objects, has a lower PCC, due to prioritizing the background too.

From the table, we confirm that the Soft Attention, when wrong, assigns lower certainty to the predicted class and a relatively high probability to the true class, while CAM tends to be overconfident in its prediction.

Fig. 21: Per-step analysis of an example going through the Soft Attention Model. Top row: subsequent human fixations. Middle row: Soft Attention. Bottom row: Prediction. The rightmost is the final prediction.

Detailed look at the Soft Attention Model In figure 21 we present an example of a detailed analysis of the Soft Attention Model. We show that the network is not limited to paying attention to a small neighborhood, like humans are.

The model first takes a glimpse at the image and makes an initial prediction based on a small area and then given that guess, it proceeds to analyze the input in a methodical way. It first analyzes the background and decides the image is

a bit more likely to be an airplane than a boat (the asphalt that the network focused on in the bottom of the image is specific to airplanes).

In the last step, the network merges the information acquired previously and focuses directly on the most salient part of the image. Assigns almost 94% probability to the airplane class.

The example demonstrates how powerful visual attention can be in image analysis. Instead of having to integrate different kinds of abstract concepts, e.g. background and foreground information in one step, the network can sequentially extract relevant information and abstract representations and then merge them into a coherent prediction.

5.2 Machine Learning guided by human attention

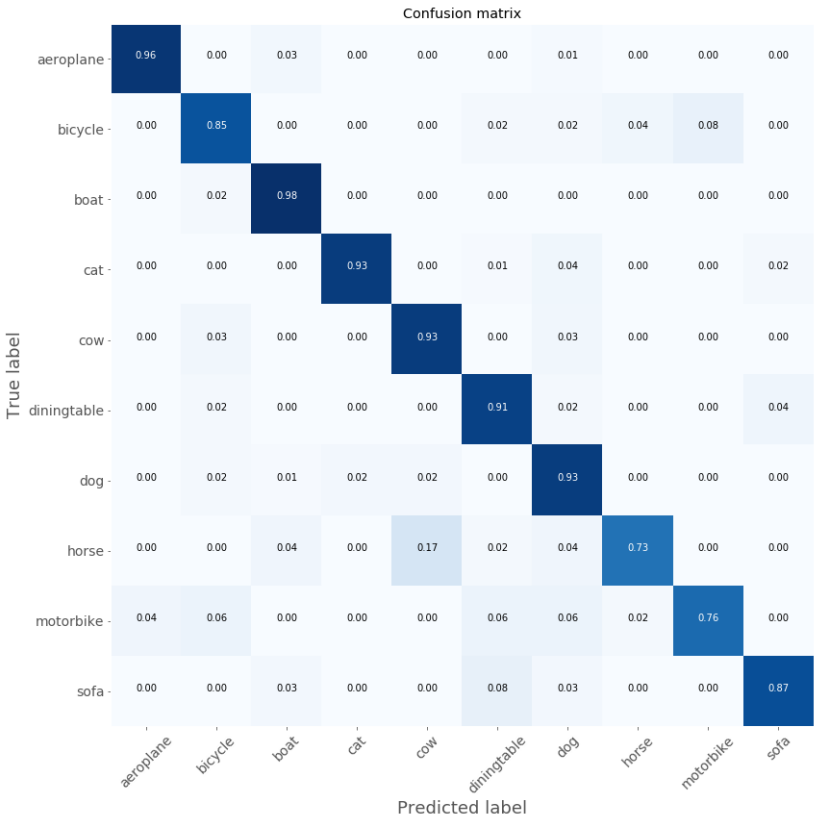


Fig. 22: Confusion matrix for baseline model

Baseline In the provided confusion matrix (figure 22) we can see that the model achieves an overall good performance, with the notable exception of horse being

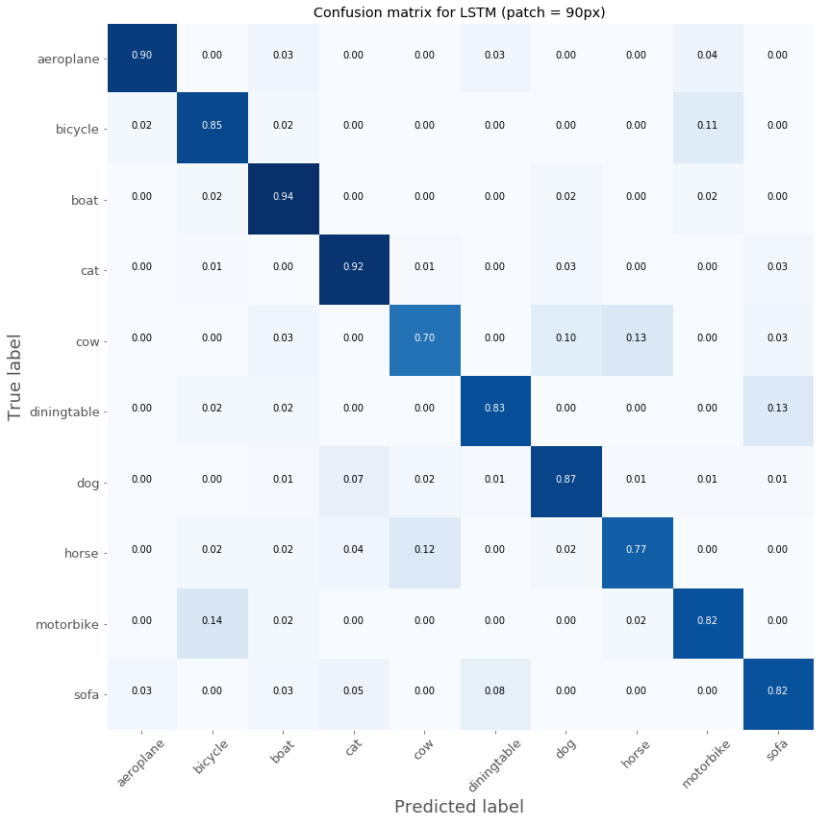


Fig. 23: LSTM confusion matrix. Patch size = 90px

confused with cow; bicycle with motorbike (both ways); and diningtable with boat. The cow-horse problem is observable in the eye-tracking dataset itself: the slow response times and large number of fixations indicate that humans also have a problem with identifying these objects (fig. 7).

The baseline model reaches **0.8957**% accuracy on the validation set.

Eye-tracking-guided LSTM We experiment with multiple sizes for the patches around fixations, and provide a scatter plot of the validation accuracy (fig. 24). The initial increase in patch size yields great improvement in the performance. As we reach 90 and 100 a stagnation occurs.

This supports the calculations in section 4.1, where we observe that the human attention is most effective for approx. 2.5 degrees eccentricity from fixation, corresponding to 93 pixels for the POET setup and diminishes quickly after 5 degrees eccentricity, corresponding to approx. 187 pixels.

The last model before the plateau starts, achieves **0.8583**% accuracy on the validation set, and is the closest one to the human effective attention size. It achieves a similar performance to the baseline model, using less information.

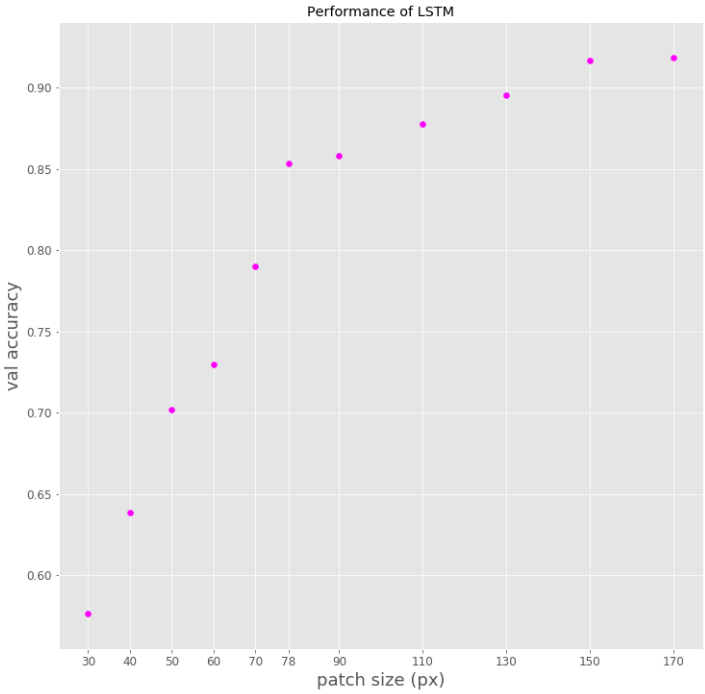


Fig. 24: LSTM performance

The model performs better than baseline for large patches but they require more processing given that the model has to process 5 large patches before making a prediction. Thus the improvement comes at a computational cost.

6 Discussion

6.1 Limitations

We acknowledge that although an improvement on generating the human attention heatmaps, our approach does not provide a 100% correct visualization of where the information has been extracted from. We can see that e.g. in the case of the bicycle in fig. 16, where human participants most likely extracted information about the entire bicycle and not just the middle part as the heatmap would indicate. Information is most likely also extracted about the background, implicitly informing the participants decision, which is also not indicated by the heatmap.

We note that the dataset is small for a deep model to be trained and the results, although relatively satisfying, might be improved with more data, and

without the class imbalances and unrepresentative images. Also, due to the computational complexity, we used transfer learning in all our experiments. Better results could be achieved with CNNs trained for the task. ResNet50 is trained on images of size 224x224, thus is not performing well for smaller patches.

It is difficult to say if the difference between the Soft Attention and human attention is due to the inherent differences between the mechanisms or is it due to the task being so simple that the human can perform it in "one go" without having to reason separately about different abstract features of the image.

6.2 Future work

It would be interesting to further investigate the possible relation between the soft attention mechanism and the theory of human scene integration being a parallel one-step process, as suspected by Gestalt theorists. To specifically research this question, a more suited experiment could be based on images that highlight the scene integration process such as the Kaniza illusion [21].

To compare the models and human data more directly on the object recognition task, the humans should perform a 10-class classification instead of the 2-class discrimination task.

It could also be interesting to employ some additional mechanism of inquiring about the overt attention in order to improve on the human attention heatmap generation algorithm.

It should also be relevant to use the eye-tracking data in pretraining the attention mechanisms in CAM and Soft Attention models.

Finally, it would be interesting to apply our model to other tasks that are widely adopted for modelling human visual attention, such as image captioning [22] and Visual Question Answering(VQA) [23].

6.3 Conclusions

We have proven that our novel way of generating human attention heatmaps together with the PCC as the evaluation metric can successfully distinguish between attention mechanisms in relation to human attention.

We have shown how various attention mechanisms, inspired by the cognitive mechanisms known from humans, can be employed to improve the interpretability of deep models and draw attention to issues of the dataset.

We have proven that there exist attention mechanisms in computational models that work similarly to the human attention and ones that employ more elaborate schemes, because they are not limited by extracting information around a specific fixation point.

Given the similarity between CAM and human visual attention it is plausible to assume that the latter has a strong bottom-up component. This would be in line with Von Helmholtz's theory of human visual attention prioritizing intrinsically interesting visual features.

Finally, we have given circumstantial evidence to further support the gradient theory and the thesis that human attention is most efficient up to 2.5 - 5 degrees eccentricity from fixation.

References

1. T. Buschman and E. Miller, "Shifting the spotlight of attention: Evidence for discrete computations in cognition," *Frontiers in Human Neuroscience*, vol. 4, p. 194, 2010. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnhum.2010.00194
2. L. T. Troland, "Helmholtz's treatise on physiological optics," *Science*, vol. 63, no. 1641, pp. 597–599, 1926. [Online]. Available: http://science.sciencemag.org/content/63/1641/597.2
3. W. James, *The principles of psychology*. Read Books Ltd, 2013.
4. A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive psychology*, vol. 12, no. 1, pp. 97–136, 1980.
5. A. Yarbus, "Eye movements and vision. 1967," *New York*, 1967.
6. M. I. Posner, Y. Cohen, and R. D. Rafal, "Neural systems control of spatial orienting," *Phil. Trans. R. Soc. Lond. B*, vol. 298, no. 1089, pp. 187–198, 1982.
7. J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," *CoRR*, vol. abs/1412.7755, 2014. [Online]. Available: http://arxiv.org/abs/1412.7755
8. M. Borys and M. Plechawska-Wójcik, "Eye-tracking metrics in perception and visual attention research," *EJMT*, vol. 3, p. 16, 2017.
9. L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision research*, vol. 40, no. 10-12, pp. 1489–1506, 2000.
10. J.-G. Yao, X. Gao, H.-M. Yan, and C.-Y. Li, "Field of attention for instantaneous object recognition," *PLOS ONE*, vol. 6, no. 1, pp. 1–8, 01 2011. [Online]. Available: https://doi.org/10.1371/journal.pone.0016343
11. F. GERS, "Long short-term memory in recurrent neural networks," 2001. [Online]. Available: http://www.felixgers.de/papers/phd.pdf
12. B. Zhou, A. Khosla, L. A., A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization." *CVPR*, 2016.
13. J. K. R. C. K. C. A. S. R. Z. R. B. Y. Xu, Kelvin; Ba, "Show, attend and tell: Neural image caption generation with visual attention," 2015. [Online]. Available: https://arxiv.org/abs/1502.03044
14. "Pascal objects eye tracking (poet) v 1.1," http://calvin.inf.ed.ac.uk/wp-content/uploads/eyetrackdataset/data/README.txt, accessed: 2018-12-20.
15. T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *2009 IEEE 12th International Conference on Computer Vision*, Sep. 2009, pp. 2106–2113.
16. "Pixel size as a function of screen size," http://www.prismo.ch/comparisons/desktop.php.
17. "Full width half maximum," https://en.wikipedia.org/wiki/Full_width_at_half_maximum.
18. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980
19. H. Le and A. Borji, "What are the receptive, effective receptive, and projective fields of neurons in convolutional neural networks?" *CoRR*, vol. abs/1705.07049, 2017. [Online]. Available: http://arxiv.org/abs/1705.07049
20. "Keras cam model," https://github.com/jacobgil/keras-cam.
21. A. T. Duchowski, *Eye Tracking Methodology: Theory and Practice*. Berlin, Heidelberg: Springer-Verlag, 2007.

22. J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via A visual sentinel for image captioning," *CoRR*, vol. abs/1612.01887, 2016. [Online]. Available: http://arxiv.org/abs/1612.01887
23. J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," *CoRR*, vol. abs/1606.00061, 2016. [Online]. Available: http://arxiv.org/abs/1606.00061
24. D. P. Papadopoulos, A. D. F. Clarke, F. Keller, and V. Ferrari, "Training object class detectors from eye tracking data," in *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, 2014, pp. 361–376. [Online]. Available: https://doi.org/10.1007/978-3-319-10602-1_24
25. D. Scherer, A. Müller, and S. Behnke, "Evaluation of pooling operations in convolutional architectures for object recognition," in *Artificial Neural Networks – ICANN 2010*, K. Diamantaras, W. Duch, and L. S. Iliadis, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 92–101.
26. D. C. Ciresan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," *CoRR*, vol. abs/1202.2745, 2012. [Online]. Available: http://arxiv.org/abs/1202.2745
27. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

Appendix

A Individual contributions

Report writing:

- David Oppenberg: wrote or contributed to subsections: 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 2.1, 2.2, 2.3, 3.1, 3.2, 6.2, 6.3, Appendix B.
- Michal Gacka: structured and edited the report; wrote or contributed to subsections: 1.5, 2.3, 2.4, 2.5, 2.6, 4.1, 4.2, 4.3, 4.4, 5.1, 6.1, 6.2, 6.3, Appendix C.
- Cristian Mitroi: wrote or contributed to subsections: 3.3, 4.5, 5.2, 6.1, Appendix D; edited sections: 1, 2, 3.

Other:

- David Oppenberg: introduced the group to cognitive theories of attention; guided the programming efforts towards cognition-related research questions; helped in the early experiments with the CAM Model.
- Michal Gacka: designed the approaches and experiments; authored sec.4 approach; programmed/trained: CAM Model (Keras), Soft Attention Model (Tensorflow); distribution plots and attention results visualizations.
- Cristian Mitroi: extracted, pre-processed and visualized POET dataset; programmed/trained patch-based LSTM and baseline CNN (Keras); visualized patch-based LSTM results.

B POET experiment details

Before the start of the experiment, a standard nine-point calibration and validation was performed. The sequence of images was randomized for each participant. Before each trial a centrally located cross was displayed. Drift correction was performed after every 20 images, re-calibration after every 200 images and a break was offered every 30 minutes.

The images were presented with a random offset from the center of the screen, in order to avoid the participants developing viewing strategies. A total of 28 students at the University of Edinburgh, 11 male and 17 female, took part in the data collection. While the confounding variable of gender was accounted for, it can be criticized that the age range of subjects is too narrow and potentially resulting in artificial inter-ob- server consistency. Every participant saw one block of 1000 images corresponding to one pair of images, with the exception of five participants who saw more than one block [24].

Participants were seated 60 cm from a 22" LCD screen, while their eye movements were recorded using an Eyelink 2000 eye tracker, which sampled both eyes at a rate of 1,000 Hz, with Training object class detectors from eye tracking data

5 a typical spatial resolution of 0.25 to 0.5. A head rest was used. Button presses were recorded using a Logitech gamepad that offers millisecond accuracy. The experiment was controlled using Psychophysics Toolbox Version 3 [24].

C Machine Learning background

C.1 Convolutional Neural Networks for object recognition

Convolutional Neural Networks (CNN) are commonly employed for computer vision tasks, due to their ability to detect spatial relations within data. The concept of CNNs, as well as their architecture, are based on the model of the mammal visual cortex proposed by Hubel and Wiesel. Their model distinguishes between simple cells, responsible for feature extraction, and complex cells which combine local features of spatial proximity [25].

Similarly, the layers of CNNs consist of convolutional filters that extract low level features such as edges and curves, which are then abstracted to a higher level representation throughout the network. Instead of hand-crafting the filters, they are learned from data. Intermediary feature matrices are called feature maps.

CNNs match human performance on benchmark tasks such as the recognition of handwritten digits or traffic signs [26].

One of the most prominent deep CNN architectures is the Residual Network (ResNet). ResNet tackles the design problem of extremely deep CNNs which often suffers from an increasingly down-scaled gradient that gets backpropagated [27]. It solves the "vanishing gradient" problem by introducing shortcut connections between layers [27]. ResNet50 is a Residual Network with 50 layers. VGG16 is one of the earliest architectures, with 16 convolutional, max pooling and dense layers. It won the ILSVR ImageNet competition in 2014 [**?**].

C.2 Transfer Learning

Transfer Learning is the concept of applying a pretrained model to a new domain or problem. A model is first fitted to one problem and then its weights are reapplied to another task. Usually it involves training a new classification layer that uses features maps extracted by that model, or fine-tuning the weights in the model to make it fit the new task better.

It is especially successful in object recognition, because frameworks like *Keras* and *Tensorflow* have models available that are pretrained on the ImageNet dataset to solve 1000-class classification task.

C.3 Recurrent Neural Networks

Recurrent Neural Networks excel at processing interdependent sequential data. RNNs are able to memorize information and interpret new input based on that memory. We use a Long Short-Term Memory unit [11].

D Poor quality data examples

Fig. 25: Example of "cow". This only shows the back area of a cow

Fig. 26: Example of "cow". This only shows the head

Fig. 27: Example of "horse". Only the back of the horse is visible

Fig. 28: Example of "aeroplane". The plane is too small

Fig. 29: Example of "boat". This is a very different kind of boat, as compared to the photo on the right

Fig. 30: Example of "boat". This is a very different kind of boat, as compared to the photo on the left