岡山大学
OKAYAMA UNIVERSITY

Grenoble INP – ENSIMAG
École Nationale Supérieure d'Informatique et de Mathématiques Appliquées

Final year project report

Performed at Okayama University
Human Behavior Understanding Lab

Human Behavior Understanding

Analysis of gaze data in relation to internal state of humans and semantic properties of images

LANGUILLE Timothée
3rd year – MMIS Option

April 3, 2023 – September 22, 2023

Okayama University
Graduate School of Natural Science and Technology
Department of Computer Science
Division of Industrial Innovation Sciences
3-1-1 Tsushima-naka, Kita-ku, Okayama-shi, 700-8530 JAPAN

Internship Supervisor
Zeynep Yücel
School Supervisor
Olivier Muller

Table of Contents

1 Introduction

1.1 Context

This internship is carried out in an academic context. The host organization is the Graduate School of Natural Science and Technology of Okayama University. Particularly, the Human Behavior Understanding laboratory, in which the internship is taking place, is part of the Division of Industrial Innovation Sciences, in the Department of Computer Science.

Situated in the Okayama Prefecture within the Chugoku region of Honshu, Okayama University stands as a highly regarded in Japan. Established in 1870, the university is home to approximately 13,000 students, comprising 10,000 undergraduates and 3,000 postgraduates. The institution's motto is "Creating and fostering higher knowledge and wisdom".

Okayama University has a strong global engagement and welcomes between 700 and 800 international students every year. One of its important collaborative relationships involves Université Grenoble Alpes. Over the past decade, this partnership has been facilitating research-oriented internships for students from Grenoble.

Within the academic landscape, the Graduate School of Natural Science and Technology focuses on research in fundamental sciences such as global climate change, plant photosynthesis or supernova neutrinos. The Division of Industrial Innovation Sciences works especially on applied engineering in the field of computer science, robotics, material sciences among others.

In the Department of Computer Science, the topics of research are the basic theory and application of information technology, artificial intelligence and computer technology. Human Behavior Understanding Lab's focus is on studying and comprehending human behavior. Dr Zeynep Yücel is the professor of the laboratory and is particularly interested in human behavior in social contexts, and mechanisms of attention. For instance, recent papers from doctoral students in the laboratory discuss collision avoidance during pedestrian group movements [6] and the addition of audio elements to enhance the memorisation of visual stimuli in e-learning settings [18].

1.2 Motivation

As mentioned earlier, one of Dr Zeynep Yücel's main topics of interest involves mechanisms of attention. These mechanisms play indeed a crucial part in understanding how humans behave. Lawrence M. Ward [19] defines attention this way:

> "Attention refers to the process by which organisms select a subset of available information upon which to focus for enhanced processing and integration."

In particular, understanding and replicating human visual attention represents a whole area of research. Analysing and being able to predict where people look when confronted to a given visual stimulus also have possible industrial applications, for instance in design/advertising or in image compression.

Different types of stimuli will not have the same effect on people's gaze, because of particularities in their attentional capture. For instance, the literature [17] shows that people's gaze behaves differently when they look at a tool than when they look at a similarly shaped non-tool object. This study will revolve around visual attention when confronted to images of simple objects, including tools: our aim is to take into account the specific nature of tools in predictive technology for gaze behaviour. In this scope, rather than the industrial applications mentioned above, applications related to human-robot interaction are more relevant.

2 Background and related work

This section contains elements of previous works in domains related to this project.

2.1 Visual saliency

2.1.1 Where people look

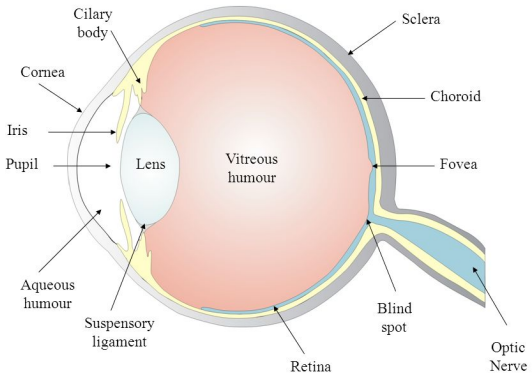

Figure 1: Human eye anatomy.

The human eye (see Figure 1) perceives images by letting light go through the *pupil* and be projected onto the *retina* (back of the eyeball), where light-sensitive cells, called *cones* and *rods* convert the incoming light into electrical signals sent to the visual cortex for further processing. Especially, *cones* are sensitive to visual detail, but are very sparsely distributed on most of the *retina*. There is only a small part at the center, called *fovea* and spanning less than 2° of the visual field, where they are actually over-represented: this results in human only being able to have full acuity in this small area [7].

Therefore, we need to shift our gaze in order to capture what is in front of our eyes. The process of deciding where we look depending on the stimuli in our visual field is what

Lawrence M. Ward refers to as *visual attention orienting* [19]. Humans successively stop their gaze for a moment at different locations in the visual scene, which is referred to as *fixations*. The position of these fixations is a topic that has been studied extensively by the scientific community. A key concept in studying the properties of a visual scene which influence human attention orienting is *salience*.

> "Visual salience (or visual saliency) is the distinct subjective perceptual quality which makes some items in the world stand out from their neighbors and immediately grab our attention." [8]

The key component of visual salience is bottom-up, or stimulus-driven, which means that the salience of an area basically corresponds to the extent to which this area differs (according to different visual features such as the color, orientation, shape, depth, etc.) from its surroundings in the visual scene.

Besides, this bottom-up contribution can be strongly modulated by a top-down, or user-driven component, which means that the internal state of the subjects influences where they look: for instance, if someone is searching for a specific object, their fixations are more likely to occur near shapes similar to this object.

In order to depict salience in various areas of a visual scene in an intuitive manner, *saliency maps* (see Figure 2) are used.

> "The Saliency Map is a topographically arranged map that represents visual saliency of a corresponding visual scene." [16]

Figure 2: Example of saliency map for a given image stimulus.

2.1.2 Saliency prediction

For almost twenty years, researchers have been designing algorithms to establish saliency maps. For a given input image, the goal is to create a heatmap having the same dimensions and depicting the extent to which each pixel is salient. Early on, models were basically sticking to the previously mentioned definition (2.1.1) and used hand engineered features

to compute the saliency value at each pixel. For instance, they could use color contrast or edge detection as components of their salience computation.

A fairly intuitive way of looking at the concept of saliency map is to think of it as a probability density $p(x, y|I)$ that is supposed to predict people's fixations on a given stimulus I, in a context of free-viewing (i.e. limited top-down component in visual salience). Thus, it makes sense to refer to the previously mentioned algorithms as *saliency prediction models*.

In order to evaluate these models, *ground truth* data are needed. For visual salience, these ground truth data consist in real fixations measured by gaze tracking on several subjects (see Figure 3).

In the remainder of this study, for a given stimulus, we will use the following terminology:

- *Ground truth fixation map* refers to the binary matrix where each coefficient indicates if a fixation was recorded at the corresponding pixel for at least one subject during the gaze tracking experiment.
- *Ground truth saliency map* refers to the probability density obtained by applying a Gaussian filter to the ground truth fixation map and normalizing the result.
- *Saliency map prediction* refers to the probability density outputted by a saliency prediction model.

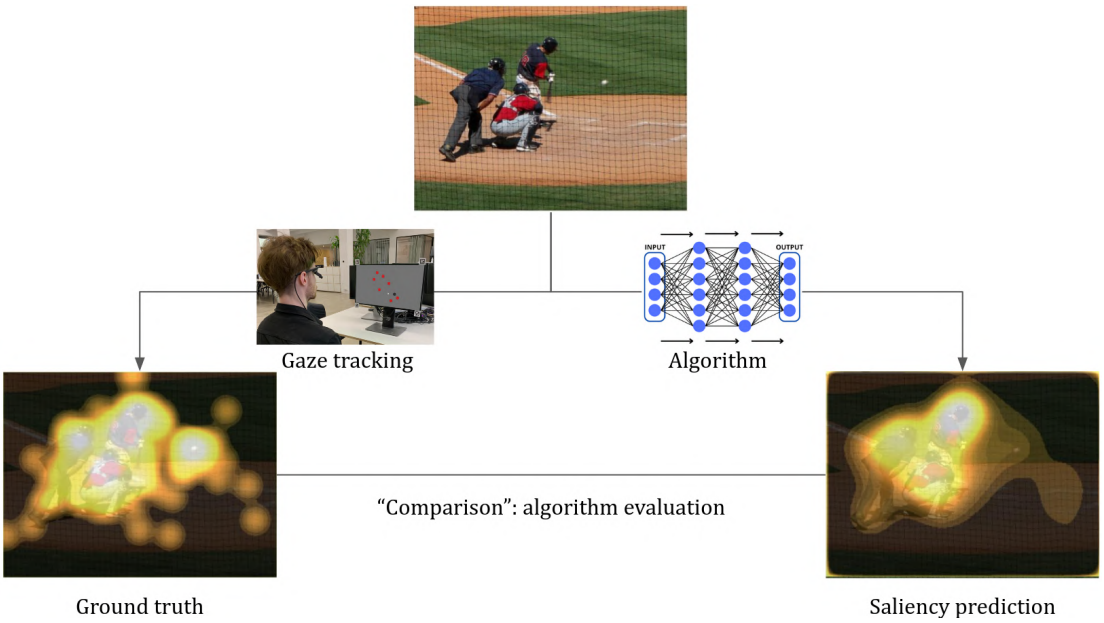

Figure 3: Evaluation of saliency prediction models.

Around 2014, saliency prediction has shifted from the classic models mentioned earlier towards deep learning models. However, since gaze tracking is time-consuming, requires

special equipment and must be carried out on a sufficient number of subjects, the cost and difficulty to obtain a large amount of ground truth data make it really hard to create high-performance deep learning models from scratch. Therefore, nowadays most models use transfer learning. In other words, they mainly use preexisting image classification/semantic segmentation models (usually based on ImageNet and trained over millions of data) as backbones, keep their convolutional layers, which extract features likely to play a role in visual salience, and add a few layers (fully-connected) on top of it to compute a salience value at each pixel. This enables to attain high-performance without as much training data since the main part of the model has already been trained in order to reach high-performance in its original task. For example, in 2017, DeepGaze II [12] used the VGG-19 deep neural network trained to identify objects in images as its backbone (see Figure 4) and achieved state-of-the-art performance.

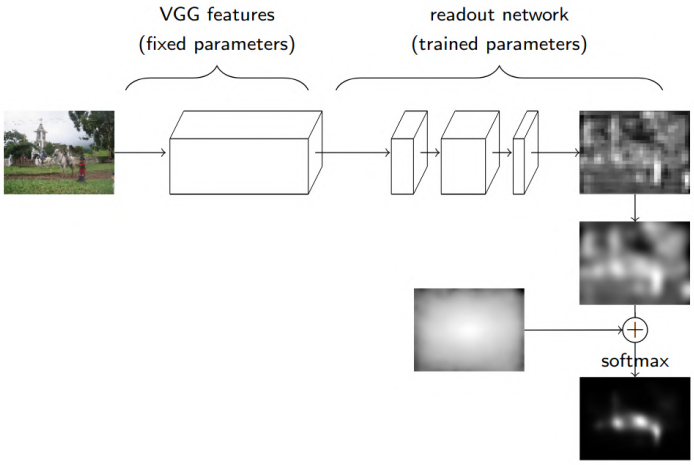

Figure 4: DeepGaze II architecture: transfer learning from VGG19.

One point to note is that human gaze tends to be biased towards the centre of the visual scene. It means that, for two identical stimuli appearing on an image, the one close to the center will be more salient. Therefore, DeepGaze II and other models add a prior stimulus-independent centerbias component on top of their raw prediction before outputting their true saliency map prediction. This stimulus-independent centerbias component is essentially a distribution in which the values are higher the closer a pixel is to the center.

2.1.3 Benchmarking

In order to evaluate saliency prediction models and to compare them with each other, the MIT Saliency Benchmark [9], which became the MIT/Tuebingen Saliency Benchmark [10] in 2019, was introduced in 2012. It is used to measure prediction models' performances over a dataset containing ground truth fixation data for 300 images, collected on 39 subjects.

Table 1: State of the art performances on the MIT/Tübingen Saliency Benchmark.

Name	IG	AUC	sAUC	NSS	CC	KL-Div	SIM
Gold Standard (3.3)	1.3239	0.8982		2.8481			
DeepGaze IIE [13]	1.0715	0.8829	0.7942	2.5265	0.8242	0.3474	0.6993
UNISAL [4]	0.9505	0.8772	0.7840	2.3689	0.7851	0.4149	0.6746

Since the benchmark is aimed at ranking the models, quantitative metrics are needed to evaluate performances: in visual salience, 7 metrics are commonly used (table 1), some of which have been created for this specific task such as *Normalized Scanpath Saliency* (NSS) while others come from information theory like *Kullback–Leibler divergence* (KL-Div).

A metric is basically a function M calculating a value $M[s(x, y|I); D]$ for an given saliency map $s(x, y|I)$ (obtained on a given stimulus I) and a given set of ground truth fixation positions $D = (x_1, y_1), ..., (x_n, y_n)$ (corresponding to this given stimulus I). The value displayed in the benchmark for a given metric averages its value on all stimuli of the evaluation dataset.

Because different metrics account for different properties and have different optimal saliency map predictions [3], a model's output can not be expected to yield best performances in all metrics as it stands, and for a long time it was not considered relevant to compare two models if they were not optimized for the same metrics.

However, in 2018, Kümmerer et al. [11] showed that it was possible to overcome this issue by considering the raw output of a model as the predictive probability density and applying transformations to obtain metric-specific versions of the saliency map prediction before computing the metrics. It enables to have good models perform well in all metrics, and makes it relevant to compare models according to any metric.

An interesting point to note is that *Information Gain* (IG) and *Normalized Scanpath Saliency* (NSS) already evaluate saliency map predictions as predictive probability densities, hence no transformation is needed before computing the metrics. Moreover, they operate directly on the ground truth fixation map (see 2.1.2). Therefore, they are convenient to use and have been popular recently. Thus, these are the metrics we will use the most in the remainder of this study.

2.2 Tools

2.2.1 Attentional capture

The original idea of this project's topic has been inspired by an article about human attentional capture for tool images by Skiba and Snow [**?**].

Affordances of an object refer to the possible and relevant actions the user can perform on this object. For example, a chair affords being sat on, a ball affords being held and

being thrown, etc. When it comes to tools, their affordances consist of highly specific motor routines and depend heavily on their function, which is tightly linked to the identity of the object. For instance, a saw affords being grasped in a specific way and moved in a back and forth manner, because it is the way it must be moved in order to fulfil its function of sawing material. In other words, what defines a tool, rather than its shape, is the recognition of its function through its shape, and this function implies the relevant actions that can be performed on it.

For a straight-shaped tool with a clear separation between the *head* (part of the tool affecting its environment) and the *handle* (part of the tool grasped by the user) such as a knife, a hammer or a screwdriver, attentional capture is expected to be biased towards either. According to the article [17], humans tend to orient their attention towards the head of a tool in priority. The experiment presented in Figure 5 consists in displaying a dot either to the right or left of a central stimulus: if the stimulus is a tool, the reaction time turns out to be significantly higher when the dot is displayed near the handle. The hypothesis formulated is that humans look at the head in priority, since it enables them to quickly grasp the function of the tool, and thus establish its affordances.

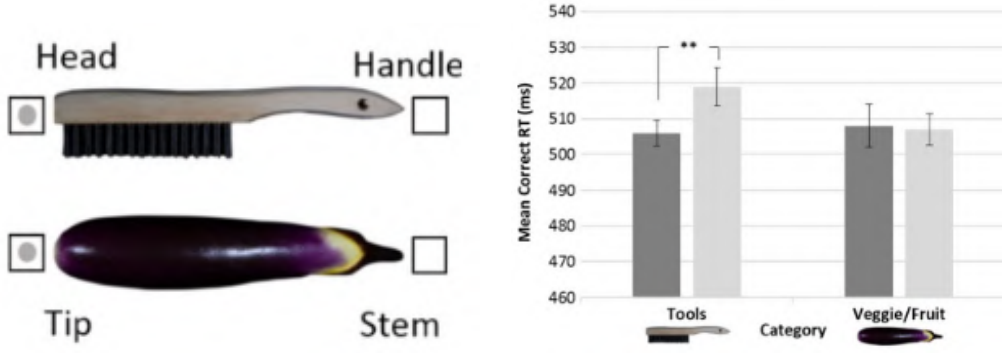

Figure 5: Dot detection time for tools and veggie/fruit depending on the side where the target appears.

With the terminology introduced in Section 2.1.1, when we look at a tool picture, a top-down component is induced in the corresponding saliency map: the recognition of its function contributes to orient our attention. Thus, we can expect existing saliency prediction models not to perform optimally on this particular image type.

2.2.2 Classification

For this project, even if we restrict the scope to hand tools, it still covers a wide spectrum of objects. The head/handle distinction established in the previous section does not always exist or can be rather ambiguous. Moreover, we can not expect human attention to be influenced in the same manner by every tool. Therefore, if we will eventually only consider simple straight-shaped tools for this project, establishing a comprehensive and relevant classification of tools would be an important step to deal with this study more in depth.

This kind of classification for tools in particular does not really exist in the literature. However, there are taxonomies classifying the different ways of grasping objects [5], [14].These classifications mostly divide grasp types between power and precision requirements, with some modulations depending on the position of the hand (see Figure 6). This could be a relevant idea for classifying tools because grasps are directly linked to the affordances and function of an object.

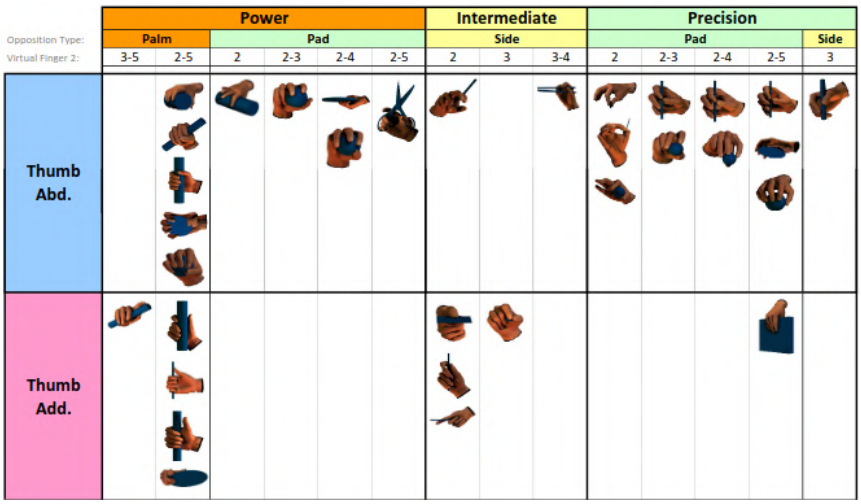

Figure 6: Comprehensive Grasp Taxonomy which includes 33 grasp types.

3 Collecting gaze data

Since the core of this project consists in retraining a preexisting deep learning saliency map prediction model in order to specialize it towards the type of image that we are interested in, collecting gaze data for this type of image represents a crucial part.

3.1 Custom dataset

Saliency prediction models are mostly designed to predict saliency for natural, complex stimuli, scenes directly corresponding to what humans can perceive at any time. For instance, these stimuli can be urban landscape, people interacting or the inside of a room (see Figure 7). Hence, the main datasets used in the field of saliency prediction are mostly made up of this type of stimulus: *MIT300* and *MIT1003* [9] are respectively the most used datasets to evaluate and to train prediction models. Both consist of fixation data for natural indoor and outdoor scenes, complex stimuli in which numerous elements can be noticed.

On the contrary, for the purpose of this project, we wanted to have access to fixation data for simple stimuli, consisting of individual objects on a plain background. For the

Figure 7: Images from the MIT1003 dataset.

most part, this object would be a hand tool.

Another commonly used dataset named *CAT2000* [1] has an "object" category containing fixation data for 100 stimuli similar to this characterisation, but few of these objects are actually tools. Therefore, we picked 115 images (100 for training, 15 for testing) from the *Bank Of Standardized Stimuli* (BOSS) [2], which is a dataset including high resolution images of all sorts of objects on a plain white background (see Figure 8), and carried out our own gaze tracking experiment to collect fixation data. Among the 115 selected stimuli, 60 are images of simple tools, 30 are images of non-tool but similarly shaped objects, and 25 are images of more ambiguous objects (either not clearly defined as tools, or not having a clear head/handle separation).

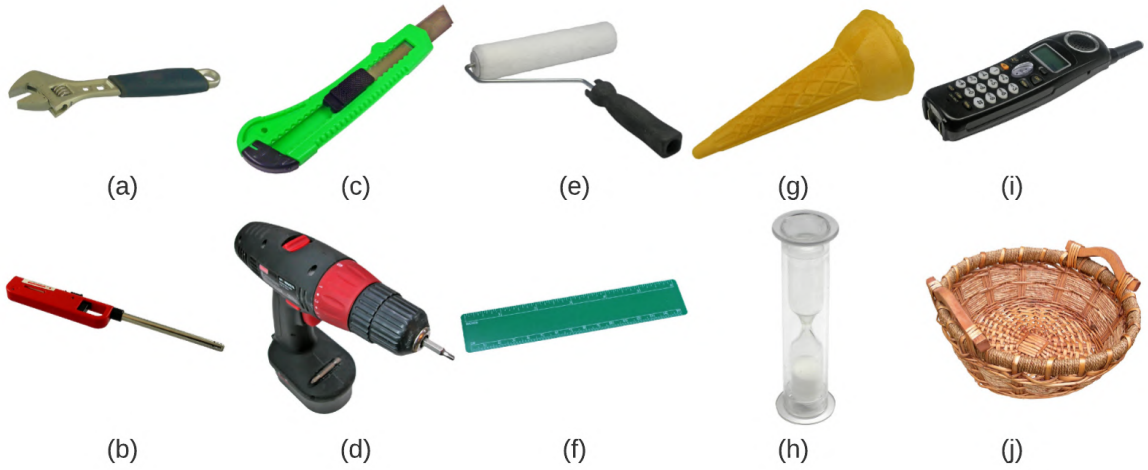

Figure 8: Images from the BOSS dataset. (a) to (e) are simple tools, (g)/(h) are non-tool objects, (f)/(i)/(j) are ambiguous objects.

A hundred images is very few for a deep learning training dataset, but, having carried out fine-tuning tests using the "object" category of *CAT2000*, it seems possible to

have a significant influence on the metrics, even with only hundred images in the training dataset. In addition, since collecting gaze data is quite time-consuming and requires a lot of attention from the subjects, it seems difficult to gather much more data than that.

3.2 Method

The methodology used to collect data is largely inspired by *MIT300* and *MIT1003* [9].

The subject sits 50cm away from a 51cm wide screen (1920x1080 pixels), wearing a *Pupil Core* eye tracking device (200Hz, 0.6° of gaze accuracy), with their head supported by a chin-rest for more stability.

The 115 stimuli selected from the BOSS dataset are randomly divided into 5 blocks of 25 images (15 for the last one) in order to enable the subject to have frequent breaks and reset their attention. Each block begins with a calibration (9 point calibration) before successively displaying 25 stimuli for 3 seconds each, with a 1.5 seconds interval between two stimuli, during which a red cross is displayed in the center of the screen. The subject are asked to freely look at the stimuli, and are told that there will be a memory test at the end, in order to enhance their engagement in the task.

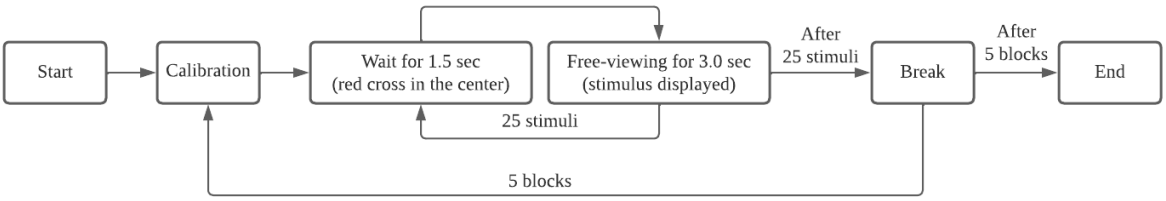

Figure 9: Gaze tracking experiment pipeline.

The recording software *Pupil Capture* stores data in a directory which is then read by the software *Pupil Player* in order to export it as a list of gaze positions associated with timestamps. Sometimes we need to manually adjust the calibration because the data are shifted: it is fairly easy to detect because we know that the gaze is supposed to be on the central red cross in between images.

We managed to get gaze data from 12 subjects, which is almost as many as the 15 subjects involved in *MIT1003*.

3.3 Data analysis

Once we get one subject's gaze data as a list of positions with timestamps, we have to process it to establish their fixations for each stimulus. Using the timestamps, we can extract the gaze data corresponding to one stimulus and apply an naive yet effective algorithm to compute fixations: we consider that there has been a fixation somewhere when

gaze position stays for a long enough period in a small enough area [7] . Common values for these temporal and spatial threshold are respectively 0.1 second and 1° of visual angle (in our configuration, this corresponds to around 20 pixels). We discard the first fixation since it usually stands right in the center because of the red cross cue displayed in between stimuli.

Then, by considering the fixations of all subjects for a given stimulus, we can establish its ground truth fixation map (see Section 2.1.2). To build the corresponding ground truth saliency map (see Section 2.1.2, see Figure 10), we must determine the size of the Gaussian filter we will apply. [9] mentions a cutoff frequency corresponding to 1° of visual angle, thus we chose a filter with a cutoff frequency of 23 pixels, which is a little bit bigger than 1° of visual angle for our images, but accounts for potential uncertainty in the measure.

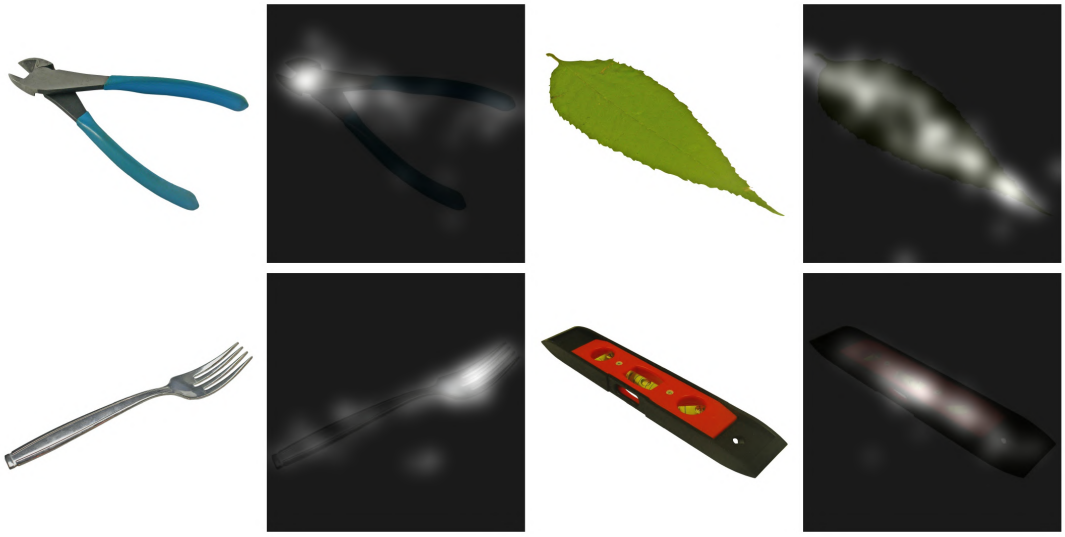

Figure 10: Ground truth saliency maps for stimuli in our custom dataset.

In addition to establishing ground truth fixation and saliency maps, an interesting concept is the *gold standard*. This refers to a line in the MIT/Tuebingen Saliency Benchmark (see Table 1) which does not correspond to an actual saliency prediction model, but to the ability of the ground truth data itself to predict human fixations. Basically, for a given metric and a given stimulus in the dataset, the gold standard value is calculated by leaving one subject out of the computation of the ground truth saliency map and evaluating this saliency map as a predictive probability density of the fixation of the subject left out. The metric is computed once with each subject being left out, then the mean value represents the gold standard of the metric for the given stimulus. Hence, in the benchmark, the gold standard value of a metric corresponds to the mean of this value for all stimuli in the MIT300 dataset.

For our custom dataset, it is possible to compute the gold standard value for both IG and NSS metrics.

Given a binary map of fixations Q^B, a saliency map P, and a centerbias baseline map B (see Section 2.1.2),

$$IG(P, Q^B) = \frac{1}{N} \sum_i Q_i^B [\log_2(\epsilon + P_i) - \log_2(\epsilon + B_i)], \tag{1}$$

where i indexes the i^{th} pixel, N is the total number of fixated pixels, ϵ is for regularization. This metric measures the average information gain of the saliency map over the prior centerbias baseline at fixated locations: a positive score in this metric means that the saliency map predicts fixated locations better than the image-independent prior centerbias distribution.

With the same notations, NSS is defined as

$$\begin{aligned} NSS(P, Q^B) &= \frac{1}{N} \sum_i \overline{P_i} \times Q_i^B \\ \text{where } N = \sum_i Q_i^B \text{ and } \overline{P} &= \frac{P - \mu(P)}{\sigma(P)}. \end{aligned} \tag{2}$$

NSS measures the average normalized (zero mean, unit standard deviation) saliency at fixated locations: it means that if actual fixations were distributed according to chance, NSS value would be 0, and the higher the value, the more fixations are gathered in salient area of the map P.

In the calculation of the gold standard for both IG and NSS, the binary map of fixations Q^B includes the fixations of the subject being left out, while the saliency map P is the ground truth saliency map made from the fixations of all other subjects.

The gold standard value can be interpreted as an indicator for the quality of collected data: if low values would not necessarily be a sign of error in data capture and could be due to specific features of the stimuli, high values show that the ground truth saliency maps are meaningful since their predictive power is significant. Another way of looking at these values is to consider them as potentially achievable values for predictive models. It is sometimes referred to as a lower bound for the *explainable information*.

For our custom dataset, gold standard values (see Table 2) are relatively close to the ones for MIT300: it makes us optimistic about the usability of our data to train a saliency prediction model. However, we have to be careful with theses values: because of the small sample size, we can not really interpret them quantitatively. For instance, even if we notice that there is no big difference between the three kinds of stimuli included in our dataset, no conclusion can be drawn.

Table 2: Gold standard IG and NSS values on custom dataset compared to MIT300.

Set	IG	NSS
All stimuli	1.191	2.501
Simple tools	1.200	2.521
Non-tool objects	1.244	2.529
Ambiguous objects	1.100	2.413
MIT300	1.3239	2.8481

4 Model fine-tuning

4.1 DeepGaze IIE

As the basis for this project, we select the saliency prediction model DeepGaze IIE [13] which achieves state-of-the-art performance in the MIT/Tuebingen Saliency Benchmark (see Section 2.1.3) for almost all metrics. Moreover, Deepgaze IIE is implemented in *Python* using the popular library *PyTorch*, which makes it relatively simple to get started with it and to integrate our custom dataset in its training process.

This model is mostly based on the transfer learning architecture of its predecessor DeepGaze II (see Figure 4): it uses convolutional layers from preexisting state-of-the-art *ImageNet* models and applies a readout network on these layers. The key difference from DeepGaze II is that DeepGaze IIE combines backbones from different models, whereas DeepGaze II was selecting one classification model as its foundation. DeepGaze IIE leverages complementarity between its backbones to achieve more accurate predictions (figure 11.

One feature of DeepGaze II which made it stand out from other models was that it did not update the weights of its backbone model during training: only the added fully-connected layers were trained. Since the backbone's layers' weights were already optimal to extract image features, retraining them could hardly improve performances and might have caused overfitting (the model "memorising" the training dataset instead of learning from it). DeepGaze IIE is trained in the same way: only the readout network applied on top of the convolutional layers of the preexisting models is trained.

The core of the model's training involves MIT1003, which is the most commonly used training dataset for saliency prediction models. However, a thousand images is usually not enough for training layers from scratch and reaching high performance. Therefore, DeepGaze IIE's readout network is pre-trained using the SALICON dataset [15].

This dataset contains saliency data for 10,000 stimuli, yet these data are not obtained through the usual method which involves gaze tracking, rather, subjects are confronted with a blurred image displayed on their screen, and they can move their mouse over the image to see certain areas with high resolution. It basically emulates the process of atten-

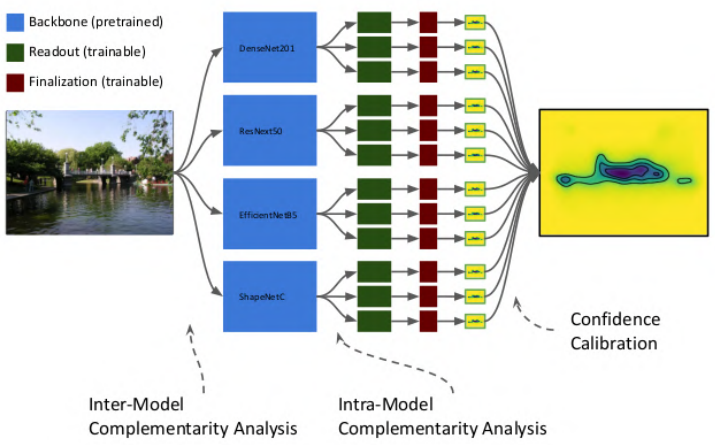

Figure 11: DeepGaze IIE architecture.

tion orienting using only a computer mouse, and the movements of the mouse have been shown to be correlated to actual fixation patterns. Even if these data remain inferior to traditionally obtained saliency dataset in terms of quality, this method greatly reduces the difficulty and cost to collect ground truth saliency data which can still be effectively used as pre-training data.

The pre-trained model then goes through a 10-fold cross-validation, using the MIT1003 dataset. It means that the 1000 images from MIT1003 are divided into 10 blocks of 100 images and the model is trained 10 times, every time using a different block of images as validation data and the 9 other blocks as training data. At the end, of the training, the 10 resulting models are combined to create the final model, which is then tested over the MIT300 dataset. This enables to use the whole dataset to test the model during training, without reducing the amount of training data.

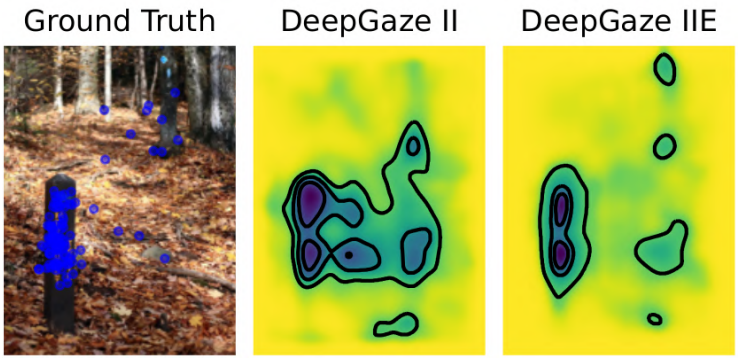

Figure 12: Example of saliency map prediction from DeepGaze IIE compared to its predecessor.

4.2 Hyperparameters

The crucial part in training a deep learning model, other than training data itself, is the choice of hyperparameters. Here, the structure of the model is already established so hyperparameters basically include the loss function, the learning rate, and the number of epochs.

The loss function evaluates the performances of the model on the training data as the training progresses. During training, the model's weights are continuously updated to try to minimize the loss function. For saliency prediction model, metrics used in the benchmark are likely to be relevant training loss functions (with a sign modulation if necessary). Before constituting our custom dataset, we carried out training experiments (see Figure 13) on DeepGaze IIE with training data from the CAT2000 dataset with different training losses, each one involving different common saliency map evaluation metrics. It turned out that the general behaviour of the model did not change drastically depending on the metrics used. That being said, it seems that using IG in the training loss improves the value of every metric on the validation dataset almost as much as any other training loss.

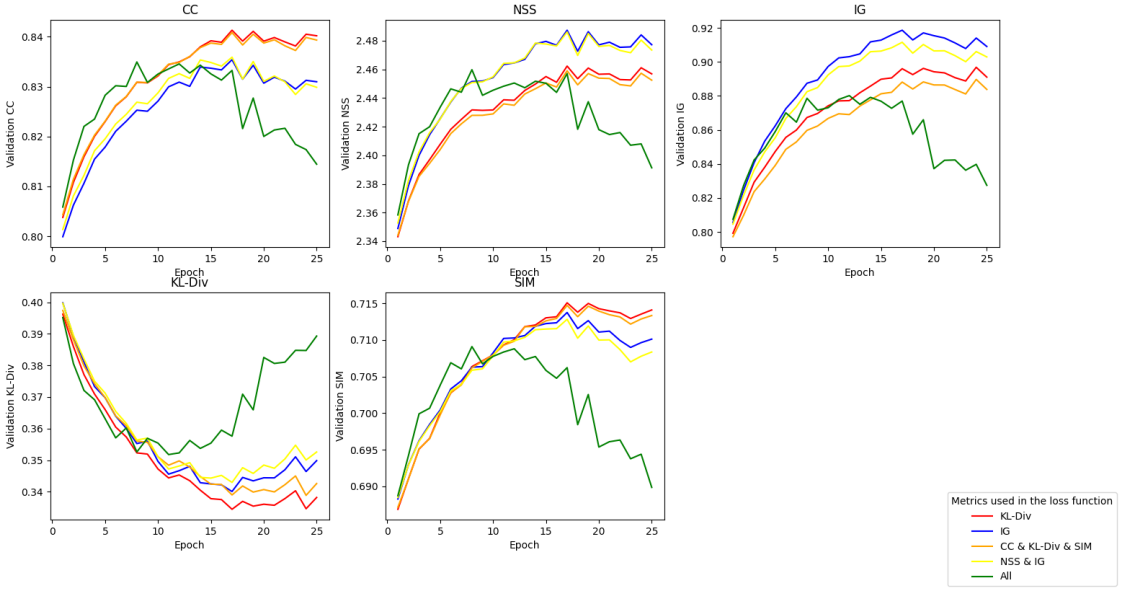

Figure 13: Evolution of validation metrics during training depending on the training loss.

Therefore, we decided to use IG as the training loss for the rest of this project. Actually, DeepGaze IIE was already trained using this metric as loss. Regarding the learning rate, which represents the rate at which the weights are updated during training, the model was originally trained with a learning rate starting at 0.001 and regularly divided by 10 after a fixed number of epoch. Since we are only fine-tuning an already trained model, we chose a learning rate of 0.0001 from the start. We tested a few learning rates around this value and

it seems that it only changes the epoch at which the optimum is reached, not the actual value of the optimum.

Then we carried out a 5-fold cross-validation on our custom dataset to check if the chosen hyperparameters made the training efficient and to estimate after how many epochs we should stop training to avoid overfitting (this phenomenon is visible when the evaluation metrics stop improving on the validation data but the training loss continues to decrease). For each fold, we record significant improvements for all metrics, especially between 10 and 15% for NSS, between 15 and 20% for IG, with best performances reached after around 40 epochs. Thus, it validates the hyperparameters.

4.3 Results

We fine-tuned DeepGaze IIE using the hyperparameters introduced in Section 4.2 and obtained the following results for IG and NSS metrics.

Table 3: Influence of fine-tuning on the model's performances on training set (100 stimuli).

DeepGazeIIE	IG	NSS
Original	1.432	2.307
fine-tuned	1.752	2.585

When computing the metrics on the dataset used for fine-tuning the model (see Table 3), it is not surprising to notice a drastic improvement (IG improved by 22%, NSS by 12%). However it is not necessarily a good news, since an overfitting model could result in such a result: if it memorizes the training dataset, it is likely to perform extremely well for stimuli inside the dataset and poorly for stimuli outside. Therefore, the most crucial part is to evaluate our model on data it has not seen during training.

Table 4: Influence of fine-tuning on the model's performances on testing set (15 stimuli).

DeepGazeIIE	IG	NSS
Original	1.356	2.271
fine-tuned	1.576	2.461

For stimuli from our custom dataset which were not involved in the fine-tuning (table 4), the model performs 16% better for IG and 8% better for NSS after fine-tuning. This is a reassuring result, showing there is a good chance that our training actually improved the model for the particular type of stimuli we are interested in. However, it is difficult to draw conclusions from such restricted sample.

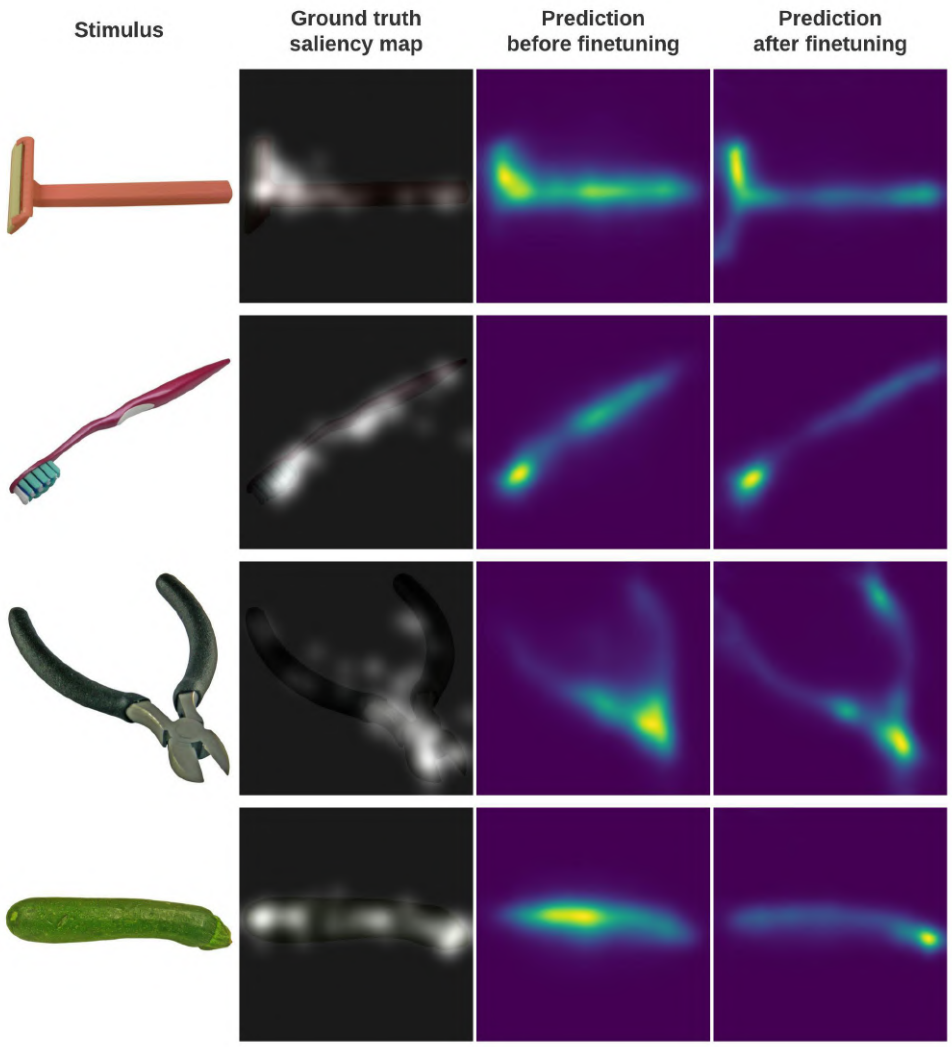

Figure 14: Comparison of model output for a few stimuli from the testing set before and after fine-tuning.

Moreover, the qualitative aspect (see Figure 14) of saliency predictions slightly improved, as it seems that the outputted saliency maps are more confident in fixation probability near the head on tool stimuli (the saliency value decreased in areas other than the head), but it can still hardly be seen as a significant change. Thus, we can question what improves the metrics. For instance, we could imagine that the fine-tuned model takes better account of the white background, rather than the specificity of tools.

Table 5: Influence of fine-tuning on the model's performances on a subset of CAT2000 (100 stimuli).

DeepGazeIIE	IG	NSS
Original	0.664	2.340
fine-tuned	0.462	2.079

Finally, we need to recompute the saliency metrics on data completely independent from our custom dataset, closer to MIT datasets, in order to test regression (see Table 5). It seems that the model's performances for stimuli outside of the particular type we are interested in have significantly dropped after fine-tuning. It is something that had to be expected, since we wanted to specialize the model to some extent, but it is important to note that.

5 Conclusion

To sum up the work exposed in this report, we reviewed the literature about key concepts of attention and visual salience, before establishing a particularity of tools' attentional capture that we would like to take into account in visual saliency prediction. Then we proposed a method to modulate the preexisting deep learning prediction model DeepGaze IIE in order to tackle this issue.

We managed to create a custom dataset of ground truth saliency data for tool and simple object stimuli. Metrics like gold standard IG and NSS tend to confirm the quality and relevance of these fixation data, although the sample size is quite small. Also, the contribution of tool head in attentional capture is noticeable on these saliency maps.

This dataset has been used to carry out fine-tuning on the state-of-the-art model in visual saliency prediction, and its performance quantitatively improved for the type of stimuli contained in our custom dataset. Yet, it remains unclear whether this improvement comes from tool specificity better taken into account or any other distinctive feature of this type of stimulus. It is possible that we have just slightly optimized the capture of some small components of visual saliency for a precise stimuli type. Then maybe we should have wandered a little bit further from the existing model than just fine-tuning it.

Another idea that was originally part of the topic was to add an intention component when capturing subjects' gaze data on tools (for instance, having them want to use the tool), but it would have complicated a lot of things. It remains something that could be studied someday. Finally, another possibility to deepen the study, and maybe get a better grasp on the concept of priority in attentional capture, would be to take into account the order of the fixations (notion of scanpath).

6 Work organization

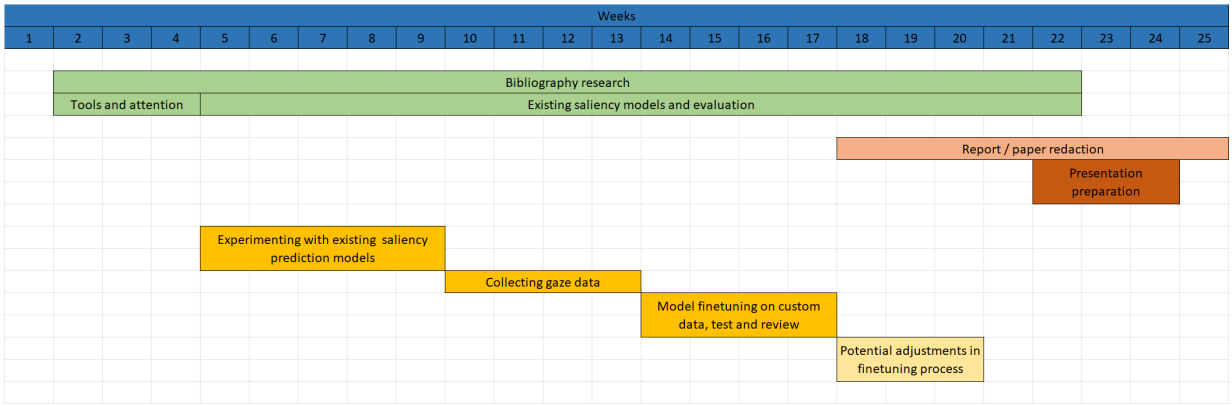

Figure 15: Early estimation of project schedule.

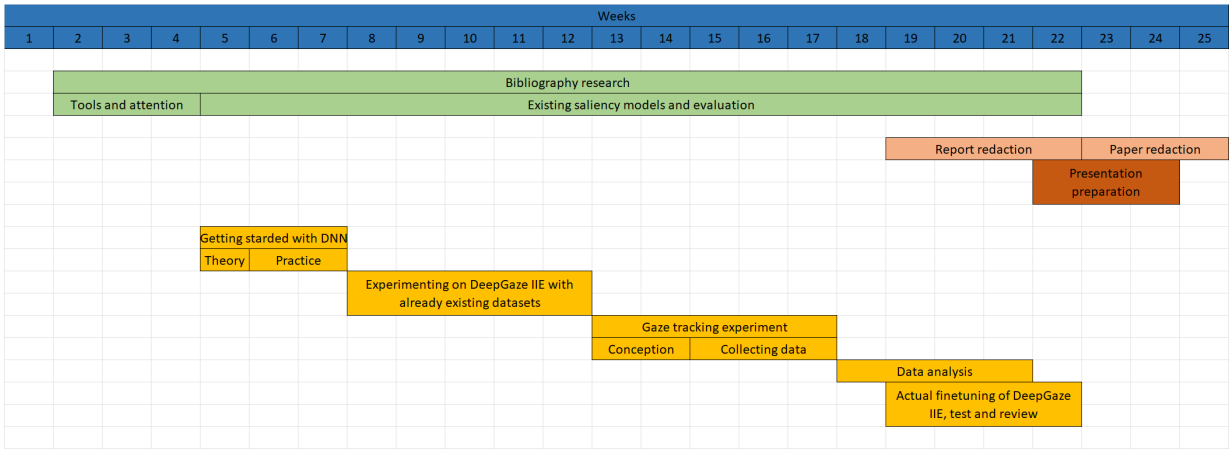

Figure 16: Effective final schedule of the project.

As introduced in section 1.1, this internship was carried out in an academic context. Thus, if the field of interest and the goal of the project were well defined, the outlines contained some uncertainty in the precise way to tackle the problem.

Figure 15 shows an estimated planning of the tasks involved in the project, established early on in the internship. Then, Figure 16 corresponds to the actual scheduled followed.

For the most part, tasks were completed within a timeframe close to what was initially estimated. A few parts were slowed down a little because of equipment availability constraints (or participant availability constraints for the data gathering part), while others were simply more complex than expected. The main addition in the effective planning is a specific period to get familiar with deep learning and PyTorch (which I used for the rest of the project) before experimenting on actual saliency prediction related models. This part was needed since my prior knowledge in this field was very limited. I had some issues manipulating deep learning models at first because I was using my personal laptop with no GPU but I had access to a more powerful computer after a few weeks.

If the necessity to carry out bibliography research throughout the course of the internship was planned, something I did not expect to be as prominent was the need to constantly return to previously read papers after learning something new about the field.

7 Environmental and societal impact

7.1 Personal environmental impact

At the laboratory, my work merely involved the usage of my personal laptop for programming and basic research as well as a more powerful computer with a GPU for the all the deep learning model training and gaze tracking experiments. Therefore, these are the main sources of electric consumption in my project.

According to the specifications of my laptop, its power should be around 65W. If used 40 hours a week, it represents a 2.6kWh consumption, which means 65kWh over the 25 weeks of the internship. On the other hand, I used the GPU-equipped PC (power around 700W) for around 50 hours, which represents 35kWh. These 100kWh are equivalent, for Japanese production, to 48kg of emitted CO_2.

Besides, the internship was entirely carried out at the laboratory and I used the bicycle to go around Okayama, therefore the environmental cost of commuting was limited.

However, even though I was already in Japan for an exchange semester before the internship began, it is impossible to overlook the fact that a flight from France to Japan emits 1.5 tons of CO_2 equivalent per passenger, which means 3 tons for the round trip. It is huge, over the annual carbon limit of 2 tons of equivalent CO_2 we are supposed not to exceed in order to achieve carbon neutrality by 2050. Moreover, even if it is not included in my personal carbon footprint, my family visiting me in Japan a few months ago add to this already poor environmental impact. It made me realize that settling in this far away is not ecologically sustainable unless we do not come back in Europe for a few years.

7.2 Global impact of the project

As an academic research project, this study does not really have immediate industrial applications, especially as there is still a significant amount of uncertainty about the result obtained. Therefore, it is quite tricky to assess both the environmental and societal impact it could have.

Generally speaking, deep learning models are known to consume more resources during the training phase than during the use phase (until they are deployed on a very large scale). In the case of DeepGaze IIE, being made up of several state-of-the-art ImageNet backbones makes it sizeable: it takes almost half a gigabyte to store all its weights. Depending on the extent to which it is intended to be used, its usage consumption can become significant.

For potential applications of visual saliency prediction mentioned earlier such as marketing and design, the model is likely to be used on an ad hoc basis. On the other hand, we also mentioned possible applications in image compression: this would represent a more frequent use case, but also contribute in decreasing the size of images on servers, representing a trade-off. Finally, for our modified version of the model, we previously stated human-robot interaction as an application field, but it is still very vague and makes it difficult to assess environmental impact.

For the time being, thus, it seems that the societal reflexion on potential applications of the project is more relevant in this section. The aim of this project being to predict human gaze movement when confronted to a precise type of visual stimulus (a tool picture), it basically falls under behavior understanding, prediction and replication. Its purpose, apart from providing insight about human visual attention and computer's ability to quantify and mimic it, is rather vague. We might consider that it could simply be applied to improve saliency prediction performances within the applications mentioned earlier (design, marketing, compression) on a type of stimulus which was not processed optimally until then. Another possibility, which seems more appealing to me, could be to use it on a medical level to detect potential attention anomalies among subjects. Also, if we were able to add an intention parameter in saliency prediction as mentioned at the end of Section 5, we could think of intention estimation technologies.

7.3 Okayama University policy

Okayama University implements basic yet essential measures, such as a strict waste sorting, and states "Building up a new paradigm for a sustainable world" as one of its purposes. It stresses the need to raise environmental and societal awareness among students, on global issues related to environment, energy issues, food supplies, economics, health, security and education.

8 Personal feedback

8.1 Working experience

At the end of my second year at Ensimag, I had the opportunity to spend two months in an IT company as part of my Assistant Engineer internship. During this experience, what was expected of me was clearly defined: I had to implement a solution to automate the deployment of features on one of the company's products. In other words, I had to implement a solution to a specific need.

I really enjoyed that first experience and learned a lot, but when I started my third year, I still did not really have any idea of what I wanted to do next. Academic research was a subject that came up regularly in discussions with my friends; it was still something very vague to me, but I was quite curious about it.

Ultimately, I am glad that I applied for this internship: it was a totally different experience and working on a project whose outlines were not as precisely defined as for my second year internship proved to be very instructive. On top of the technical knowledge I was able to develop on concepts involved in the project, such as deep learning, computer vision and gaze tracking, I had to learn to be proactive, test things out and draw conclusions to guide the next steps.

Besides, being surrounded by people who work on more or less similar topics, explaining your work to others and exchanging feedback, taking part in each other's data-gathering experiments, are all enriching aspects of academic research.

8.2 Personal experience

Before starting my internship at Okayama University, I had already been in Japan for six months, spending an exchange semester at Kyoto University. There, I really enjoyed Japanese culture, and the prospect to extend the experience while discovering a new facet of Japan in a less touristic city was part of the reason that made me want to apply.

Being in Japan for a whole year, I have had the opportunity to spend a lot of time with both Japanese and other international students: in Japan, university students typically join a laboratory in their fourth year of their bachelor's program (bachelor's degrees last for four years, and master's degrees for two) to conduct research alongside their coursework, and my laboratory had the particularity to frequently welcome international research students, creating a good balance in my opinion. I also got the opportunity to frequently meet Japanese students in the context of extracurricular activities. Culturally, this stay has been an immensely valuable experience.

Also, living that far away from France and from my relatives for a whole year, in a largely different environment with a completely different language, was something I was experiencing for the first time, and although it was a little intimidating at the beginning, I think that it helped me mature as a person to some extent.

8.3 Acknowledgment

I would like to express my heartfelt gratitude to Dr. Zeynep Yücel for making this internship possible by proposing original research topics and welcoming me in her laboratory, as well as for her availability during the whole project.

I would also like to thank Dr. Bernard Chenevier for passing on the research topics to Ensimag and giving me the opportunity to apply, as well as Dr. Olivier Muller, who supervised my internship from Ensimag, for his responsiveness and insight despite my late requests. Current PhD student Adrien Gregorj also spent time answering my questions during the application period and helped me a lot during the actual internship.

Finally, I would like to express my appreciation to all the students from the lab for making this experience both extremely enriching and enjoyable.

9 French abstract

Ce rapport rend compte du travail effectué dans le cadre d'un PFE en contexte académique, à l'Université d'Okayama au Japon. Le laboratoire du Dr. Zeynep Yücel, au sein duquel ce stage s'est déroulé, a pour objet d'étude la compréhension du comportement humain. En particulier, mon sujet traite des mécanismes liés à l'attention visuelle chez l'être humain, ainsi que des concepts et technologies utilisés pour la représenter et la prédire.

La notion de *saillance* (*salience* ou *saliency* en anglais) est un concept clé du sujet. Elle désigne la mesure dans laquelle une chose est susceptible de retenir l'attention par rapport aux autres choses présentes dans son environnement. Lorsque l'on parle de saillance visuelle d'une zone au sein d'une image, on cherche à quantifier sa faculté à attirer notre regard. Une *saliency map* est une représentation topographique de la saillance des différentes zones d'une image, souvent interprétée comme une densité de probabilité supposée prédire les zones fixées par l'être humain. Aujourd'hui, certains modèles d'apprentissage profond sont capables de prédire une saliency map pour une image en entrée de façon relativement précise.

Skiba et Snow [17] introduisent l'idée que la capture attentionnelle des objets fonctionnels, tels que les outils, serait davantage influencée par la reconnaissance de leur fonction que par leurs simples caractéristiques visuelles. Cela orienterait notre regard en priorité vers la tête d'un outil plutôt que vers sa poignée. L'objectif du projet est la prise en compte de cette spécificité des outils dans la prédiction de saliency maps.

Afin de réentraîner un modèle existant sur le type de stimuli qui nous intéresse, nous avons collecté les données de regard de 12 personnes pour 115 images, mettant chacune en scène un objet central (la plupart du temps un outil) sur fond uni. Nous avons utilisé des indicateurs usuels d'évaluation de saliency map afin de confirmer la qualité et la pertinence de notre dataset. Sur les saliency maps empiriques, établies à l'aide des données récoltées, on constate effectivement une prédominance de la tête sur la poignée dans le cas des images d'outils.

Ce dataset a été utilisé pour réentraîner le modèle DeepGaze IIE, qui réalise à ce jour l'état de l'art pour la prédiction de saliency maps. Les performances du modèle ont augmenté pour le type de stimuli d'intérêt (simples images d'objets) selon les indicateurs généralement utilisés en saillance visuelle pour les analyses comparatives. Cependant, nous n'avons pas la certitude que cette amélioration soit due à la nature particulière des objets fonctionnels, elle pourrait par exemple être liée à une meilleure prise en compte des arrière-plans unis, qui sont omniprésents dans le dataset d'entraînement.

Une pistes intéressante mentionnée au cours de l'étude, mais n'ayant finalement pas été traitée, serait de prendre en compte l'intention du sujet qui regarde l'image d'outil dans la prédiction de saliency map.

References

[1] Ali Borji and Laurent Itti. CAT2000: A large scale fixation dataset for boosting saliency research. *CoRR*, abs/1505.03581, 2015.

[2] M. Brodeur, G. Dion-Lessard, M. Chauret, E. Dionne-Dostie, T. Montreuil, and M. Lepage. The bank of standardized stimuli (BOSS): a new normative dataset of 480 visual stimuli to be used in visual cognition research. *Journal of Vision*, 11(11):825–825, September 2011.

[3] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(3):740–757, 2019.

[4] Richard Droste, Jianbo Jiao, and J. Alison Noble. Unified image and video saliency modeling. *CoRR*, abs/2003.05477, 2020.

[5] Thomas Feix, Javier Romero, Heinz-Bodo Schmiedmayer, Aaron M. Dollar, and Danica Kragic. The grasp taxonomy of human grasp types. *IEEE Transactions on Human-Machine Systems*, 46(1):66–77, 2016.

[6] Adrien Gregorj, Zeynep Yücel, Francesco Zanlungo, Claudio Feliciani, and Takayuki Kanda. Social aspects of collision avoidance: a detailed analysis of two-person groups and individual pedestrians. *Scientific Reports*, 13(1), April 2023.

[7] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. van de Weijer. *Eye Tracking: A comprehensive guide to methods and measures*. OUP Oxford, 2011.

[8] Laurent Itti. Visual salience. *Scholarpedia*, 2(9):3327, 2007. revision #72776.

[9] Tilke Judd, Frédo Durand, and Antonio Torralba. A benchmark of computational models of saliency to predict human fixations. In *MIT Technical Report*, 2012.

[10] Matthias Kümmerer, Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. Mit/tübingen saliency benchmark. https://saliency.tuebingen.ai/.

[11] Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge. Saliency benchmarking made easy: Separating models, maps and metrics. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, Lecture Notes in Computer Science, pages 798–814. Springer International Publishing.

[12] Matthias Kümmerer, Tom Wallis, and Matthias Bethge. DeepGaze II: Predicting fixations from deep features over time and tasks. *Journal of Vision*, 17(10):1147, August 2017.

[13] Akis Linardos, Matthias Kümmerer, Ori Press, and Matthias Bethge. Deepgaze iie: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. *International Conference on Computer Vision (ICCV)*, pages 12899–12908, 2021.

[14] Christine L. MacKenzie and Thea Iberall. *The Grasping Hand*. ISSN. Elsevier Science, 1994.

[15] Jiang Ming, Huang Shengsheng, Duan Juanyong, and Zhao Qi. Salicon: Saliency in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[16] Ernst Niebur. Saliency map. *Scholarpedia*, 2(8):2675, 2007. revision #147400.

[17] Rafal M. Skiba and Jacqueline C. Snow. Attentional capture for tool images is driven by the head end of the tool, not the handle. *Attention, Perception, &amp Psychophysics*, 78(8):2500–2514, July 2016.

[18] Parisa Supitayakul, Zeynep Yücel, and Akito Monden. Artificial neural network based audio reinforcement for computer assisted rote learning. *IEEE Access*, 11:39466–39483, 2023.

[19] Lawrence. M. Ward. Attention. *Scholarpedia*, 3(10):1538, 2008. revision #185343.