Learning where to look for a target before seeing it

\begin{abstract}
A realistic system for object categorization should be able to find the target in possibly large images, independently of its position in visual space. Current solutions leverage this solution by processing the different hypothesis (classes) at all possible spatial configuration. However, this can be costly in terms of computing time especially without dedicated parallel hardware. %
We explore here a solution inspired by the anatomy of the human visual system, that is, the combination of a foveated sensor with the capacity of rapidly moving the center of fixation using saccades. Indeed, the position and category of objects in images are a priori independent and we hypothesize that the retinotopic map overlays a central area dedicated to object categorization and a peripheral area dedicated to
Using this hypothesis, we formalize this problem in a probabilistic setting which allows us to build two parallel but interactively connected system: a classical image classification algorithm assuming that gaze is centered on the object on one side and a system learning to infer the position of the target on the other. Until the classification is confident, the system performs a saccade to the most likely position in the image. Overall, the computational cost of this strategy is less than that in holistic methods. %
We tested this framework on a simple task of finding digits in a large, cluttered image. Results demonstrate that it is possible to correctly learn the position of a target from a given class, and this before actually seeing a foveated image of the target. We compare the results of this model with classical psychophysical results in visual search. This provides evidence of the importance of such strategies in computer vision and we highlight some predictions of our model.
\end{abstract}

Notes :

- differential processing
  - periphery : low resolution in space
    high in time
  - fovea : high spatial resolution
    low temporal

- saccadic suppression
  - Ziad Hafed : elevation of perceptual thresholds at the time of saccade
    → frequency tuning in SC

Notes existantes

2017-09-28

- mann → papier frontiers
- coarse periphery
- Hermann grid ?
- invariances

2018-03-15 sujet pierre

- def sujet / protocole
- def methode avec carte accuracy
- onal

2018-12-21

- outline paper
- Figures
- yaka

(1) INTRODUCTION

(1a) Active Inference in Machine Learning

The prominence of automatic methods to identify objects in natural images is ever increasing. The performance of such systems recently reached the same performance as human observers [ref needed]. Moreover, these systems which were trained on Energy greedy high performance computers are now designed to work on more common hardware such as desktop computers with a decent GPU. However, such methods are not yet available for mobile devices, as will be necessary for instance for the fast detection of visual objects in autonomous driving, as when has to identify a pedestrian from a sign pole. More importantly, the robustness of such methods is still lower than that of humans. Indeed, it is still difficult for such a system to learn to categorize a particular object class given all its posible spatial configurations and the respective geometrical visual transformations. This explosion

of combinations is currently handled by increasing accordingly the number of parameters, hence the energy consumption of such methods. As a consequence, state-of-the art classification architectures contain many millions parameters while still handling relatively small images.

On the contrary, the human visual system is able to perform such a feat very rapidly (less than 100 ms) [Kirchner, ...] and at a low energy cost (~5W). On top of that the system is mostly autonomous, robust to transforms or lighting conditions and can learn with a few examples. If many different features may explain such efficiency, a main difference of the human visual system with classical computer vision approaches is the fact that its sensor (the retina) combines

a non homogeneous sampling of the world with the capacity to rapidly change its center. Indeed, on the one hand the retina is composed of two separate system: a central, high definition fovea and a large peripheral area. On the other hand, the retina is attached on the back of the eye and the eye is capable of low latency, high speed (>500°/s) eye movements. In particular, saccades allow for efficient changes of the position of the center of gaze. The interplay of those 2 properties allow to engage observers in an action perception cycle which sequentially scans parts of the image. This behavior is prevalent during our lifetime (2/3 saccades per second = 2 billions per life). This behaviour is one type of active inference [Friston] and we will envision herein how to incorporate it to classical computer vision schemes.

To explain and take advantage of this visual behavior, it is of particular importance to understand its computational and biological (neurophysiological) principles. One main hypothesis regarding this active vision is that visual scenes most often consist of a single visual object of interest. Take for instance the case of a conversation with a friend in a noisy café. To ease the understanding of his voice and emotion you will track his face despite all the remaining visual clutter. Such a visual experience can be simplified in a manner reminiscent of psychophysical experiments. An observer is asked to classify digits (for instance as taken from the MNIST database) as they are shown on a computer display. However these digits can be placed at a random position on the display, and visual clutter is added as a background to the image (see Figure 1). This defines more precisely our problem: how do we

identify a small object in a large image while not knowing its position?

This joint problem of localization and identification found many solutions in computer vision. Notably, recent advances in deep-learning have provided with efficient solution such as Faster-RCNN [Ren 2015] or Yolo [Redmon 2015]. This last implementation is particularly interesting as it predicts in the image the probability of proposed bounding boxes around the visual object. While rapid, the amount of such boxes greatly increases with image size and necessicates a dedicated hardware.

When limiting our problem to few objects of interest in the image, this strategy amounts to a classical problem in neuroscience, that is, the transformation of a luminous image into a saliency map [Itti 2001]. Such a computation

is essential to understand and predict saccades but also as models of attention. Recently, deep learning methods have extended this model by learning the computation of saliency maps over large databases of natural images [Kummerer 2014]. While these methods are efficient at predicting the probability of fixation, they miss an essential point in the action perception cycle: They operate on the full image while the retina operates on the non-uniform, foveated sampling of visual space (Figure 1-B). As such we believe that this is an essential step to reproduce and understand this active vision process.

Robotics

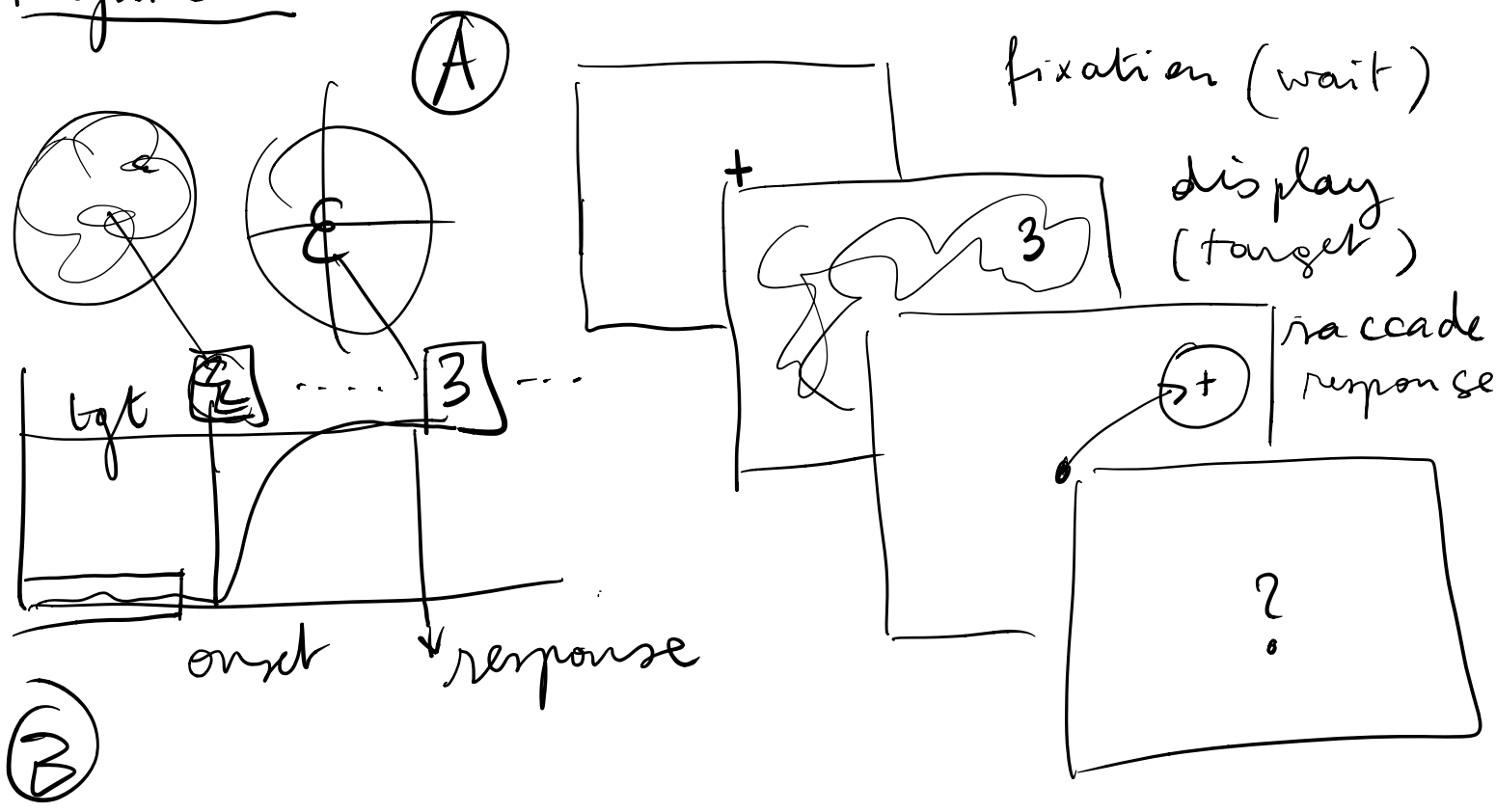
→ identify gap

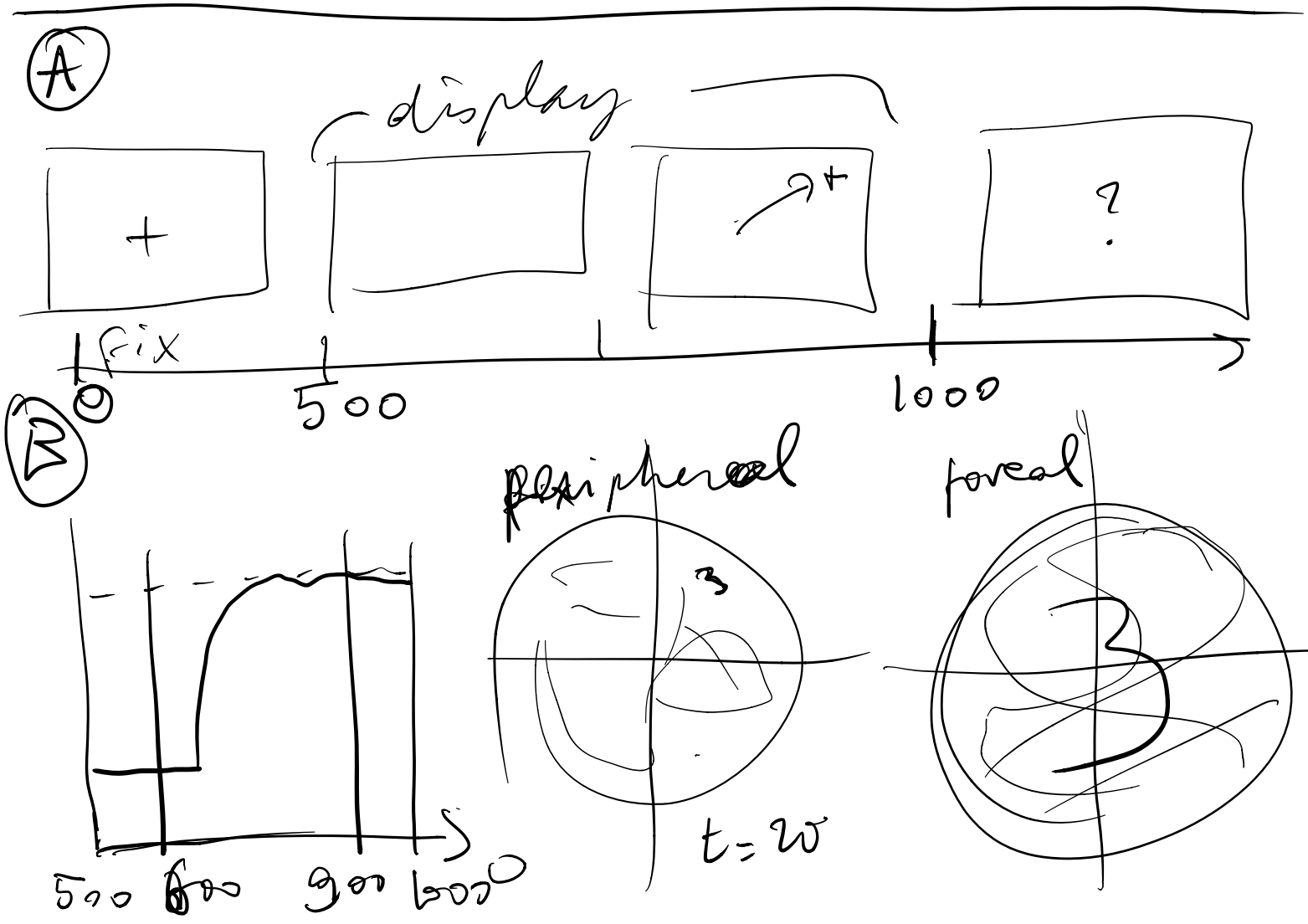
↓

= "categorize before seeing"

An interesting perspective is given with previous modeling of foveated sensors. Indeed the non-uniform sampling of vitual space is often modeled as a log-polar conformal mapping [Javier Traver 2010] which has a long history in computer vision and robotics. A first property of this mapping is the separation between the foveal and the peripheral areas as we defined above. This transformation has also other notable properties, such as the correspondance (by way of translations) in the radial and angular directions to rotations and scalings (respectively) in the vitual domain. However, this sensor is not coupled to an action (but see [ref needed]). This paper aims at addressing the fragmentation of these studies respective to their fields (Machine learning, neuroscience, robotics) to propose a novel computational model of this active vision behaviour.

Figure 1:


A
fixation (wait)
display (target)
saccade response
1
2
3
tgt
onset
response
?
B


A
display
fix
500
1000
?
B
peripheral
foveal
3
t=20
500
800
900
1000

(A) Problem setting: After a fixation period, an observer is presented with a luminous display which shows a target (here a digit) at a random position. The display is presented for a short period but enough to perform a saccade on the potential target. In particular, the configuration of the display is such that by adding clutter and reducing the size of the digit it may become necessary to perform a saccade to be able to see the digit.

Finally, the observer identifies the digit.

(B) We show a prototypical trace of a saccadic eye movement to the target position. In particular, we show the fixation window used to ensure fixation during that window (green shaded area). Overlaid is a simulation of the retinotopic map at the onset of the display and after a (successful) saccade. This demonstrates that the position of the target has to be inferred from a degraded (sampled) image and that a correct identification is mediated by the action to the location of the target * before seeing it *.

http://www.cnbc.cmu.edu/cns/papers/Najemnik-Geisler-05-N.pdf

https://cyber.sci-hub.se/MTAuMTAzOC9zNDE1OTMtMDE4LTAyNTUtNQ==/samonds2018.pdf

http://www.cs.colorado.edu/~mozer/Teaching/syllabi/ProbabilisticModels/readings/NajemnikGeisler2005Supplemental.pdf

"the eye movement is presented briefly such that no eye movement is possible."

(16) State-of-the-art

Several studies are relevant to our endeavour. First, one can consider optimal strategies to to resolve the problem of the visual search of a target [Najemnik 2005]. In a setting similar to that presented in Figure {fig: intro}, where the target is an oriented edge and the background pink noise, authors show first that a bayesian ideal observer provides with an optimal strategy and second that human observers are close to that optimal performance. Though they can predict a sequence of saccades in this perception/action loop, this model is limited by the simplicity of the display (elementary edges and a finite number of locations on a triangular grid) and by the abstract level of (Bayesian) modelling. Though its apparent simplicity, this important study could predicts some characteristics of visual scanning such as the trade-off between memory content and rapidity.

Looking more closely at neurophysiology, the study of [Samonds 2016] allows to go further in our understanding of the interplay between saccadic behaviour and the statistics of the input. In this study, authors were able to manipulate the size of saccades by manipulating key properties of the presented (natural) images. In particular, smaller images generate smaller saccades.

Interestingly, they also explained the size of saccades for different species, including mice which lack a foveal region, by the size of receptive fields. One key prediction of this study for our problem is the fact that saccades seem optimal to a priori decorelate the visual input, that is, to minimize redundancy, knowing statistics of natural inputs.

A further modelling perspective is given by [Friston12]. In this model, we first have a full description of the visual world as a generative process, here of the presentation of faces, knowing they are constituted of their components: mouth, eye, etc... An agent is completly described by giving the generative model governing the dynamics of its internal beliefs and is interacting with this image by scanning it through a foveated sensor, just as we described in Figure \ref{fig:intro}. Equipping the agent with the ability to actively sample the visual world enable us to explore the idea that actions (saccadic eye movements are optimal experiments, by which the agent seeks to confirm predictive models of the (hidden) world. One key ingredient to this process is the (internal) representation of counterfactual predictions, that is, of the probable consequences of possible hypothesis as they would be realized into actions.

Such a model constitutes an Active Inference scheme \cite{Mirza18} and simulations of the resulting optimization scheme reproduce sequential eye movements which fit well with empirical data. Compared to Najemnik (2005), saccades are not the output of a value-based cost function but a consequence of the urgency for the agent to minimize the uncertainty about his beliefs, knowing his priors on the generative model of the visual world. Such an approach applies well to our setting, as described in Fig \ref{fig:intro}. In particular, we will similarly include a generative process of the visual world as a digits at a random position and embedded in a cluttered noise. Then, we will equip the agent with a foveated sensor and with the ability to actively scan the visual image. Knowing such priors, we will optimize the behaviour of this agent and explore its properties.

1c outline of the paper : the uncoupling hypothesis

The goal of this work is to emulate a model of active vision which is able to find a visual target from a known target, while the position is unknown. This comes with the constraint that the visual sensor is foveated. We will use this constraint as an asset to minimize the overall computational cost of finding the target. This generic visual search problem is of broad general interest in machine learning, computer vision and robotics, but also to neuroscience, as it speaks to the mechanisms underlying foveation and more generally to low-level attention mechanisms.

Our aim is to produce a generic principled model and to implement a fast computer implementation. As such, we will start as in \citep{Friston12} by a probabilistic formulation and use the fundamental hypothesis outlined in Figure \ref{fig:intro}: The position of an object is independent from its identity. This property is strictly true in our setting and is very generic in vision for simple classes such as digits and simple displays (but see [Melissa Le-Hoa Võ and Wolfe, 2012] for more complex visual scene grammars). From this independence hypothesis, we may separate the two tasks (identification vs localization) in two separate pathways with different morphologies (respectively foveal and peripheral). Note that from the retinotopic projection of the visual information, this independence is conditionnal on action: These pathways should update their beliefs upon decisions made in each respective pathway.

This is paper is organized as follows. After this introduction, we will define the notations, variables and equations for the generative process governing the experiment and the generative model for the active vision agent. In particular, we will derive our method to simplify the learning of an optimal agent given these definitions. In section ~\ref{sec:results}, we will then show results of numerical simulations of this agent. We will first demonstrate some applications of this framework to different levels of complexity of the problem. This will allow us to derive some limits of this agent and, as in ~\citet{Najemnik05}, we will draw some analogies with biologically observed eye movements. Finally, in section ~\Ref{sec:discussion}, we will summarize these results in comparison with other similar schemes. We will conclude by showing the relative advantages of using this active inference approach.

Theory (methods)

1 Introduction

1.1 Issue

1.2 State of the art

1.3 Outline

1.3.1 Notations

- $\boldsymbol{x}$: visual field (image)
- $\boldsymbol{y}$: target category (categorical)
- $\boldsymbol{u}$: target position (real coordinates or categorical, retinocentric referential)

Generative model :

$$\boldsymbol{x} \sim P(X|\boldsymbol{y}, \boldsymbol{u})$$

Full inference (posterior):

$$P(Y, U|\boldsymbol{x}) \propto P(\boldsymbol{x}|Y, U)$$

Independence assumptions :

$$P(Y, U) = P(Y)P(U) \text{ } \textit{(toujours vrai)} \tag{1}$$

$$P(Y, U|X) = P(Y|X)P(U|X) \text{ } \textit{(faux s'il y a plusieurs cibles)} \tag{2}$$

Partial inference on object category:

$$P(Y|\boldsymbol{x}, \boldsymbol{u}) \propto P(\boldsymbol{x}|Y, \boldsymbol{u})$$

Partial inference on object position:

$$P(U|\boldsymbol{x}, \boldsymbol{y}) \propto P(\boldsymbol{x}|U, \boldsymbol{y})$$

Marginals:

- $P(Y|\boldsymbol{x}) = \int P(Y|\boldsymbol{x}, \boldsymbol{u})d\boldsymbol{u}$
- $P(U|\boldsymbol{x}) = \int P(U|\boldsymbol{x}, \boldsymbol{y})d\boldsymbol{y}$

1.3.2 What we did so far...

Consider a view $\boldsymbol{x}$ that contains a single target $\boldsymbol{y}$ at unknown retinocentric position $\boldsymbol{u}$. The brain needs to guess both $\boldsymbol{y}$ and $\boldsymbol{u}$ with limited computational resources.

We assume here that the brain adopts independence assumption (2), making a separation between the "Where" and the "What" pathways, forming separate (and cheaper) inferences :

- $p(Y|\boldsymbol{x})$
- $p(U|\boldsymbol{x})$

Another assumption is that the category $\boldsymbol{y}$ is *translationally invariant*: given a transformation $\mathcal{T}$,

$$\mathcal{T}(\boldsymbol{u}, \boldsymbol{y}) = (\mathcal{T}(\boldsymbol{u}), \boldsymbol{y})$$

Now, given $\boldsymbol{x}$ and the separation assumption, it is sensible to change the viewpoint to better estimate $\boldsymbol{y}$, because $\boldsymbol{y}$ is invariant to the viewpoint transformation.

This is where *active inference* comes into the play:

- Consider that the true target is $\hat{\boldsymbol{y}}$
- Consider that the target current retinocentric position is $\boldsymbol{u}$
- Then, for any translation $\delta\boldsymbol{u}$, the future posterior on the true target is estimated by: $\mathbb{E}_{\boldsymbol{x}'\sim p(X|\hat{\boldsymbol{y}},\boldsymbol{u}+\delta\boldsymbol{u})}p(\hat{\boldsymbol{y}}|\boldsymbol{x}')$
- And the optimal translation is: $\underset{\delta\boldsymbol{u}}{\operatorname{argmax}}\ \mathbb{E}_{\boldsymbol{x}'\sim p(X|\hat{\boldsymbol{y}},\boldsymbol{u}+\delta\boldsymbol{u})}p(\hat{\boldsymbol{y}}|\boldsymbol{x}')$

If now $\boldsymbol{u}$ is unknown and needs to be guessed from $\boldsymbol{x}$, the optimal translation is:

$$\underset{\delta\boldsymbol{u}}{\operatorname{argmax}}\ \mathbb{E}_{\boldsymbol{u}\sim p(U|\boldsymbol{x})}\mathbb{E}_{\boldsymbol{x}'\sim p(X|\hat{\boldsymbol{y}},\boldsymbol{u}+\delta\boldsymbol{u})}p(\hat{\boldsymbol{y}}|\boldsymbol{x}')$$

with :

- $p(U|\boldsymbol{x})$ the inferred target position
- and $\mathbb{E}_{\boldsymbol{x}'\sim p(X|\hat{\boldsymbol{y}},\boldsymbol{u}+\delta\boldsymbol{u})}p(\hat{\boldsymbol{y}}|\boldsymbol{x}')$ the expected inference on the actual target.

1.3.3 Accuracy maps

In practice, it is computationally impossible to make exact guesses about the future observation $\boldsymbol{x}'$. Our second assumption is that instead of predicting future inferences on true target, the brain trains a *parametric accuracy map* by experience (trial and error).

In a model-based approach, the *accuracy maps* can be calculated using a parametric classifier :

- Given a training set $\{(x_1, u_1, y_1), ..., (x_n, u_n, y_n)\}$:
  - Train a classifier p_θ that estimates $p(Y|\boldsymbol{x})$.
- Then, for each class $\hat{\boldsymbol{y}}$, taking $\tilde{\boldsymbol{y}} \sim p_\theta(Y|\boldsymbol{x})$, *the classification rate* $r_\theta(\boldsymbol{u})$ *is an estimator of the posterior expectation :*

$$\begin{aligned} r_\theta(\boldsymbol{u}) &= \mathbb{E}_{\boldsymbol{x}\sim p(X|\hat{\boldsymbol{y}},\boldsymbol{u})}\mathbb{E}_{\tilde{\boldsymbol{y}}\sim p_\theta(Y|\boldsymbol{x})}\delta_{\hat{\boldsymbol{y}}=\tilde{\boldsymbol{y}}} \\ &= \mathbb{E}_{\boldsymbol{x}\sim p(X|\hat{\boldsymbol{y}},\boldsymbol{u})}p_\theta(\hat{\boldsymbol{y}}|\boldsymbol{x}) \\ &\simeq \mathbb{E}_{\boldsymbol{x}\sim p(X|\hat{\boldsymbol{y}},\boldsymbol{u})}p(\hat{\boldsymbol{y}}|\boldsymbol{x}) \end{aligned}$$

  that forms an *accuracy map* for each target position $\boldsymbol{u}$.

1.3.4 Parametric transformation (Colliculus?) map

One can now select $\delta\boldsymbol{u}$ with the parametric estimator:

$$\begin{aligned} \widehat{\delta\boldsymbol{u}} &\simeq \underset{\delta\boldsymbol{u}}{\operatorname{argmax}}\ \mathbb{E}_{\boldsymbol{u}\sim p(U|\boldsymbol{x})}r_\theta(\boldsymbol{u}+\delta\boldsymbol{u}) \\ &= \underset{\delta\boldsymbol{u}}{\operatorname{argmax}}\ Q(\delta\boldsymbol{u}|\boldsymbol{x}) \end{aligned}$$

with $Q(\delta\boldsymbol{u}|\boldsymbol{x})$ the *transformation* map, given the view $\boldsymbol{x}$ and the marginal posterior estimate $p(U|\boldsymbol{x})$.

It must be noticed that, given $\hat{\boldsymbol{u}} = \underset{\boldsymbol{u}}{\operatorname{argmax}}\ r_\theta(\boldsymbol{u})$), the transformation map is maximal at $\delta\boldsymbol{u} = \hat{\boldsymbol{u}} - \boldsymbol{u}$. Each initial $\boldsymbol{u}$ provides a different transformation map, that is a shift of the original accuracy map (*Ergodic assumption??*).

We assume in the following that a parametric action value map Q_ψ can be trained on top of the parametric classifier p_θ and its accuracy map r_θ. The training set is $\{(\boldsymbol{x}_1, \boldsymbol{u}_1), ..., (\boldsymbol{x}_n, \boldsymbol{u}_n)\}$ and the accuracy map classifier learns to associate each $\boldsymbol{x}$ with its full transformation map $Q(.|\boldsymbol{x})$.

1.3.5 Algorithms

Once p_θ and Q_ψ are trained, the recognition algorithm is straightforward:

Single saccade algorithm:

1. Read the view $\boldsymbol{x}$
2. Choose $\delta\boldsymbol{u}$ according to $Q_\psi(.|\boldsymbol{x})$
3. Move the eye
4. Update the view $\boldsymbol{x}'$
5. Identify the target with $\tilde{\boldsymbol{y}} \sim p_\theta(Y|\boldsymbol{x}')$

Multi saccades algorithm:

1. $q(Y) \leftarrow$ uniform distribution
2. Read the view $\boldsymbol{x}$
3. Choose $\delta\boldsymbol{u}$ according to $Q_\psi(.|\boldsymbol{x})$
4. Repeat several times up to some posterior confidence threshold:
   (a) Move the eye
   (b) Read $\boldsymbol{x}$
   (c) $q(Y) \leftarrow q(Y) \times p_\theta(Y|\boldsymbol{x})$
   (d) normalize q
   (e) Choose $\delta\boldsymbol{u}$ according to $Q_\psi(.|\boldsymbol{x})$ (with some inhibition of return mechanism)
5. Identify the target with $\tilde{\boldsymbol{y}} \sim q(Y)$

+ Parler de Mirza 2018 → MDP

2 Methods

2.1 Visual transformation

2.1.1 Wavelets

2.1.2 Log Gabor

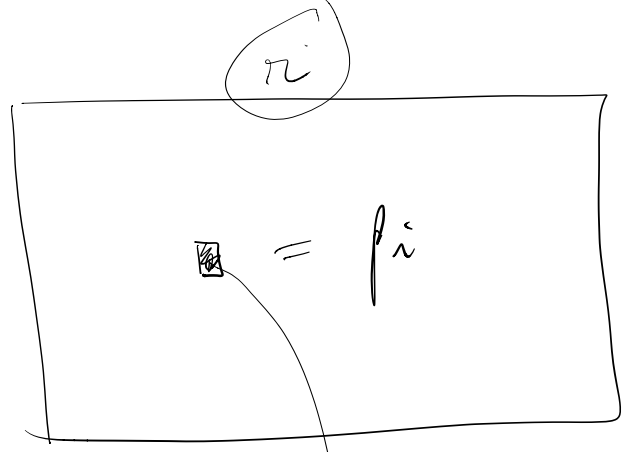
2.2 Accuracy map

2.3 Network architecture

3 Results

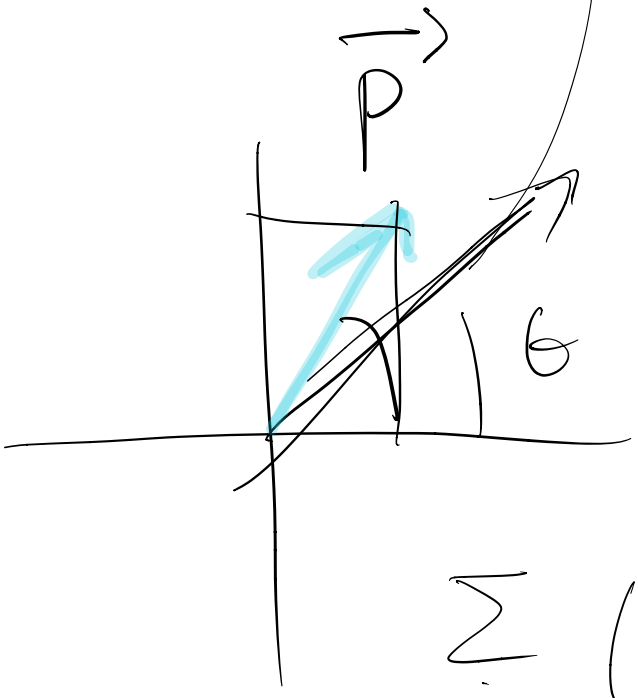
https://github.com/laurentperrinet/WhereIsMyMNIST/blob/master/2018-11-13-Where%20recap%20(clut

STATS IN CIRCULAR MAPS



J. Kaplan 13 (Eq. 10)

https://laurentperrinet.github.io/publication/kaplan-13/



1. make a 2D vector from the polar coord
2. make the mean

$$\sum_i \left(p_i \cos\theta_i \, , \, p_i \sin\theta_i \right)$$

$$\overline{\theta} = \mathrm{ArcTan2}\left(p_y \, , \, p_x \right)$$

Similar for radius

$$\overline{r} = \sum_i p_i r_i \Big/ \sum r_i$$

$$\log \overline{r} = \sum_i p_i \log(r_i)$$

4 Discussion

4.1 Summary

4.2 Limits

4.3 Perspectives

Predictions of the model → discussion

- Reaction times versus proba of detection

- can we get fooled by "fake" digits?

- bigger digits ⇒ bigger saccades.

↓

the type of database ⇒ context influences size of saccades of Yarbus

Interaction What and Where

$P(u, y \mid x)$?

trop coûteux

"what" avant la saccade $P(y \mid x, u=0)$

$\max_y P(y \mid x, u)$ on l'approche

en sortie on a un vecteur

$P_{Y=y} = [\ldots\ldots]$ de probas

en fonction d'un critère (entropie, seuil) on peut choisir

$a_0 = \max_y \left(P_{Y=y}\right)$

- on accepte STAY → évaluation
- sinon, signal Go

prédiction locale du what sur le where :

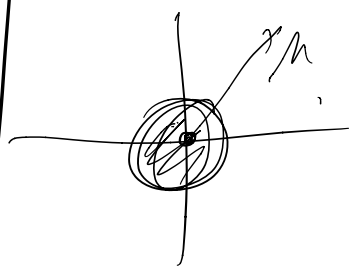
pachant G_0,

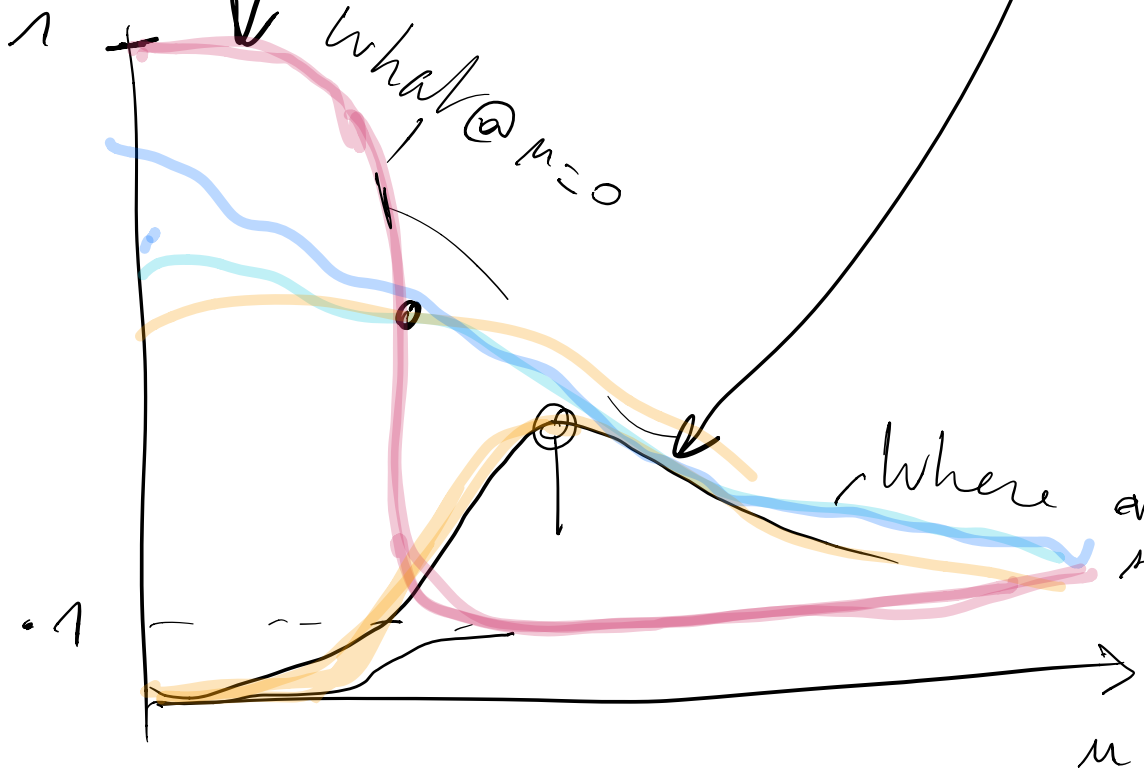
$$P(a \mid x, u) = \max_y \left[P(y \mid x, u) \right] \left[1 - P(a_u \mid a_{u=0}) \right]$$

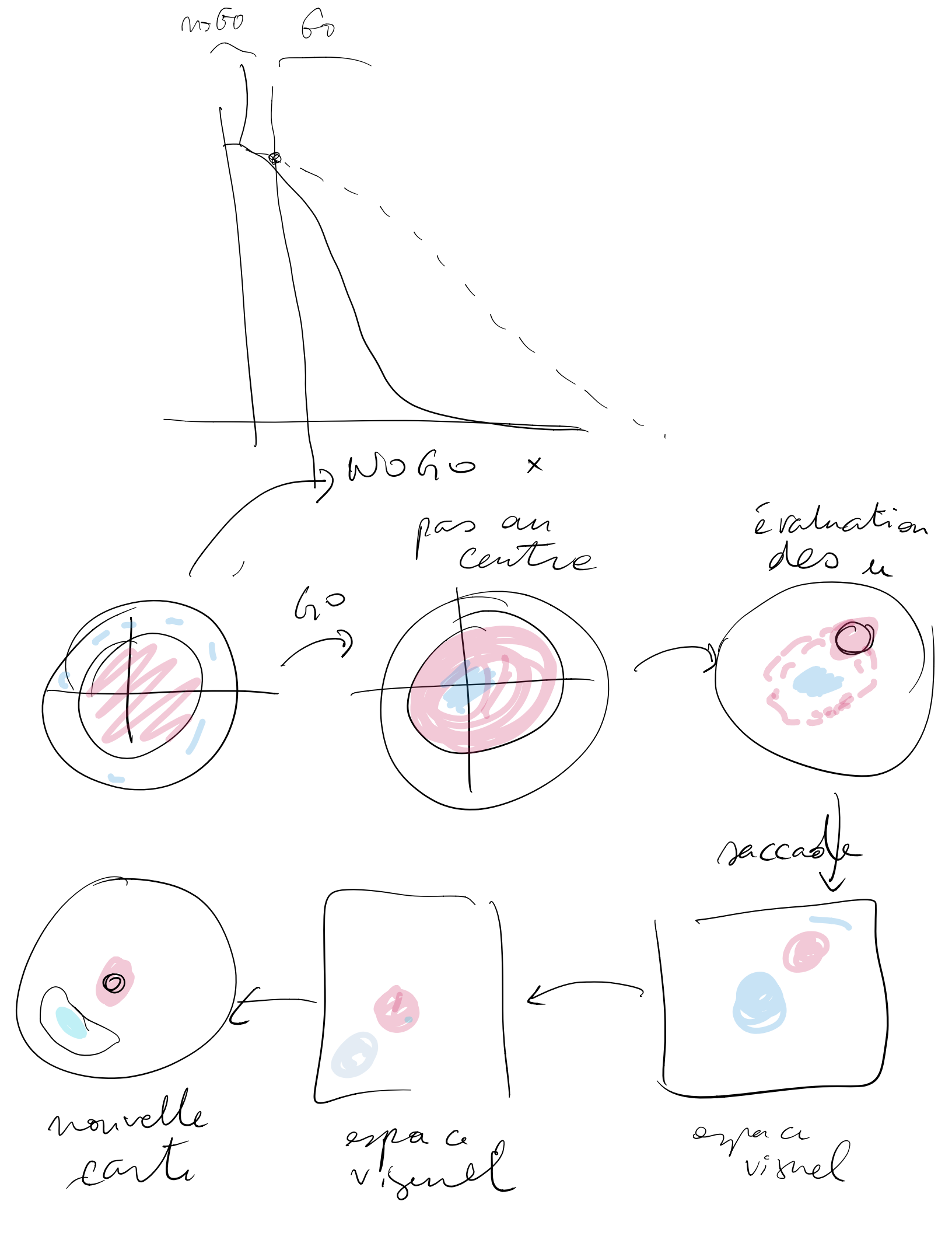
G_0

$$P(a \mid x, u=0)$$

$$P(a_u \mid a_{u=0})$$






Go
pas au centre
évaluation des u
saccade
nouvelle carte
espace visuel
espace visuel

General case: Visual information gain maximization

Consider a view $\boldsymbol{x}$ generated from a target $\boldsymbol{y}$ viewed at retinocentric position $\boldsymbol{u}$.

Consider first that :

- The generative model $p(X|\boldsymbol{y}, \boldsymbol{u})$ is known
- The retinocentric position $\boldsymbol{u}$ is known.
- The view $\boldsymbol{x}$ is known.
- The target category $\boldsymbol{y}$ is unknown.

The question comes how to choose the new retinocentric position $\boldsymbol{u}'$ in order to maximize the *mutual information* between $\boldsymbol{x}|\boldsymbol{u}$ (current view) and $\boldsymbol{x}'|\boldsymbol{u}'$ (future view).

In general, the visual Information Gain between two visual fields $\boldsymbol{x}|\boldsymbol{u}$ and $\boldsymbol{x}'|\boldsymbol{u}'$ is:

$$\mathrm{IG}(\boldsymbol{x}|\boldsymbol{u}; \boldsymbol{x}'|\boldsymbol{u}') = -\log p(\boldsymbol{x}|\boldsymbol{u}) + \log p(\boldsymbol{x}|\boldsymbol{u}, \boldsymbol{x}', \boldsymbol{u}')$$

Information Gain Lower Bound Consider now that given $\boldsymbol{x}$ and $\boldsymbol{u}$, the target category $\boldsymbol{y}$ can be *inferred* using Bayes rule, i.e.:

$$P(Y|\boldsymbol{x}, \boldsymbol{u}) \propto P(\boldsymbol{x}|Y, \boldsymbol{u})$$

Then, it can be shown (see [**?**]) that :

$$\mathrm{IG}(\boldsymbol{x}|\boldsymbol{u}; \boldsymbol{x}'|\boldsymbol{u}') \geq \mathbb{E}_{\boldsymbol{y}\sim p(Y|\boldsymbol{x},\boldsymbol{u})} \left[\log p(\boldsymbol{y}|\boldsymbol{x}', \boldsymbol{u}') - \log(\pi(\boldsymbol{y}))\right]$$

with $\pi(\boldsymbol{y})$ the prior over the $\boldsymbol{y}$'s . When the prior is uniform, the information gain lower bound (IGLB) simplifies to $\mathbb{E}_{\boldsymbol{y}\sim p(Y|\boldsymbol{x},\boldsymbol{u})} \left[\log p(\boldsymbol{y}|\boldsymbol{x}', \boldsymbol{u}')\right] + c$, with c a constant.

Predictive approach One can adopt a *predictive* approach to choose the new eye orientation $\boldsymbol{e}'$:

- First choose a new retinocentric position $\boldsymbol{u}'$ that will maximize the information gain.
- Then choose $\boldsymbol{e}'$ such that
$$\boldsymbol{z} - \boldsymbol{e}' = \boldsymbol{u}'$$
i.e.
$$\boldsymbol{e}' = \boldsymbol{e} + \boldsymbol{u} - \boldsymbol{u}'$$

The predictive approach needs three predictive steps:

- $p(Y|\boldsymbol{x}, \boldsymbol{u})$ is the current posterior over the target category inferred from the current observation,
- $\boldsymbol{x}' \sim p(X|\boldsymbol{y}, \boldsymbol{u}')$ is the predicted view generated by the model assuming that the target $\boldsymbol{y}$ is seen from from $\boldsymbol{u}'$,
- and $p(\boldsymbol{y}|\boldsymbol{x}', \boldsymbol{u}')$ is the predicted posterior for assumption $\boldsymbol{y}$, given $\boldsymbol{x}'$ and $\boldsymbol{u}'$.

Then the optimal new retinocentric position is:

$$\hat{\boldsymbol{u}}' = \underset{\boldsymbol{u}'}{\operatorname{argmax}}\, \mathbb{E}_{\boldsymbol{y}\sim p(Y|\boldsymbol{x},\boldsymbol{u})} \left[\mathbb{E}_{\boldsymbol{x}'\sim p(X|\boldsymbol{y},\boldsymbol{u}')} \left[\log p(\boldsymbol{y}|\boldsymbol{x}', \boldsymbol{u}')\right]\right]$$

Taking $\delta\boldsymbol{e} = \boldsymbol{u} - \boldsymbol{u}'$, the optimal eye displacement is:

$$\widehat{\delta\boldsymbol{e}} = \underset{\delta\boldsymbol{e}}{\operatorname{argmax}}\, \mathbb{E}_{\boldsymbol{y}\sim p(Y|\boldsymbol{x},\boldsymbol{u})} \left[\mathbb{E}_{\boldsymbol{x}'\sim p(X|\boldsymbol{y},\boldsymbol{u}-\delta\boldsymbol{e})} \left[\log p(\boldsymbol{y}|\boldsymbol{x}', \boldsymbol{u} - \delta\boldsymbol{e})\right]\right]$$

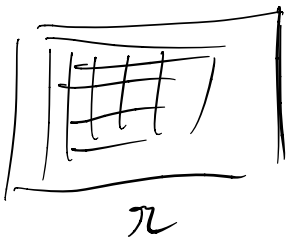
(TODO : Attention il faudrait à partir de maintenant une carte qui moyenne les log posteriors car l'espérance du log n'est pas égale au log de l'espérance, i.e. $r_\theta^{log}(\boldsymbol{u}|q) = \mathbb{E}_{\boldsymbol{y}\sim q(Y)} \left[\mathbb{E}_{\boldsymbol{x}\sim p(X|\boldsymbol{y},\boldsymbol{u})} \log p_\theta(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{u})\right]$*).*

Plan Theory { notations : proba, convolution, ...

(a) generative process -

- $x = \underbrace{\text{digit}_y * \delta_u}_{\text{in } [0,1]} + \underbrace{\text{noise}}_{\text{motion cloud.}}$

- retinal transform

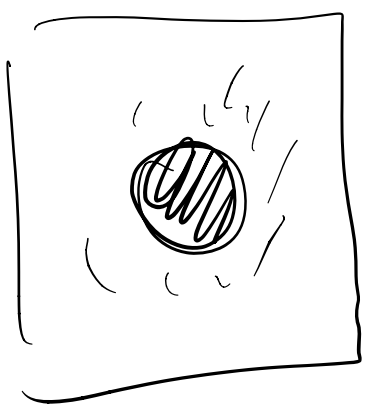
$\mathcal{R}_u(x) \longrightarrow$  θ

$\mathcal{R}$

(b) classifier = LeNet

- 98.1% accuracy a = % correct

- accuracy map



(c) where detection

given $\mathcal{Z}_0(x)$, can we estimate the accuracy in the visual space

$W(x) = P(a \mid \mathcal{Z}_0(x)) =$ function of coordinates

- this is supervised by trial and error (corrective saccades)

- we can have an ergodic hypothesis and use the accuracy map

- it's a fitting problem = deep learning
  - generative model = N-layered model
  - "cost function" = $KL(p \| q)$ = free energy (of THiOG)
- Action $u^* = \text{ArgMin } F(u)$

d) inhibition of return