TOWARD PREDICTIVE MACHINE LEARNING FOR ACTIVE VISION

Anonymous

ABSTRACT

We develop a comprehensive description of the active inference framework, as proposed by Friston (2010), under a machine-learning compliant perspective. Stemming from a biological inspiration and the auto-encoding principles, a sketch of a cognitive architecture is proposed that should provide ways to implement *estimation-oriented* control policies under a POMDP perspective. Computer simulations illustrate the effectiveness of the approach through a foveated inspection the input data. The pros and cons of the control policy are reviewed in details, showing interesting promises in term of processing compression, but also putative risks of a confirmation bias that may degrade the recognition performance if the model is too optimistic about its own predictions. The presented formalism is fully compliant with the auto-encoding framework and would deserve further developments under variational encoding architectures.

1 MOTIVATION

The oculo-motor activity is an essential component of man and animal behavior, subserving most of daily displacements and interactions with objects, devices or people. By moving gaze with the eyes, the center of sight is constantly and actively moving around during all waking time. Though ubiquitous in biology, object recognition through saccades is seldom considered in artificial vision. The reasons are many, of which the existence of high-performance sensors that provide millions of pixels at low cost [1]. Increasingly powerful computing devices are then assigned to compute in parallel those millions of pixels to perform recognition, consuming resources in a brute-force fashion.

The example of animal vision may allow to consider computer vision differently, and possibly develop parsimonious recognition algorithms that may encompass some of its aspects. A salient aspect of animal vision is thus the use of *active* sensing devices, capable of moving around under some degrees of freedom in order to choose a particular viewpoint. The existence of a set of possible sensor movements calls for the development of specific algorithms that should *solve the viewpoint selection problem*. A computer vision program should for instance look back from past experience to see which viewpoint to use to provide the most useful information about a scene. Optimizing the sensor displacements across time may then be a part of computer vision algorithms, in combination with traditional pixel-based operations.

More generally, the idea of viewpoints selection turns out to consider beforehand the computations that need to be done to achieve a certain task. A virtual sensing device should for instance act like a filter that would select which part of the signal should be worth considering, and which part should be bypassed. This may be the case for robots and drones that need to react fast with light and low-power sensing devices. Similarly, in computer vision, Mega-pixel high-resolution images appeals for selective convolution over the images, in order to avoid unnecessary matrix multiplications. Less intuitively, the ever-growing learning databases used in machine learning also suggest an intelligent scanning of the data, in a way that should retain only the critical examples or features, depending on context, before performing learning on it.

Behind the viewpoint selection problem thus lies a feature selection problem, which, in that case, should rely on a context. An evolving context over time would imply a changing feature selection. Tu put it clear, the visual features (or viewpoints) that should be used to recognize an armchair

1. on contrary to animals retina whose final design relies on a long optimization process under severe resource constraints.

should not be the same than the ones used to recognize a squirrel. If you are in a park, and there is a good chance to meet a squirrel, you should probably look around in the trees for something small and furry, whereas if you enter in a hotel, where there is a good chance to find an armchair, you may look sideways for something large and static, so you can ignore the "up in the tree" viewpoint and the small and furry features, in your computations, to confirm your hypothesis.

1.1 GAZE ORIENTATION IN BIOLOGY

The most documented case of active perception is gaze orientation, primarily studied in both man and animal Yarbus (1967); Robinson (1968). A nice review of principal promises of *animate* vision against passive vision is presented in Ballard (1991), in relation with eye-hand coordination in computer vision. A salient feature of superior vertebrates visual apparatus is the foveated retina that concentrates photoreceptors over a small central portion of the visual field. The scanning of the visual scene is principally done with high-speed targeted eye movements called saccades (Yarbus (1967)), that sequentially capture local chunks of the visual scene.

1.2 EXISTING WORK

The concept of active vision and/or active perception is present in robotic literature under different acceptances. In Aloimonos et al. (1988), the authors address the case of multi-view image processing of a scene, i.e. show that some ill-posed object recognition problems become well-posed problems as soon as several views on the same object are considered. The term was also proposed in Bajcsy (1988) as a roadmap for the development of artificial vision systems, that provides a first interpretation of active vision in the terms of sequential Bayesian estimation.

The active vision paradigm was recently introduced in neuroscience through the work of Friston (2010); Friston et al. (2012). The general setup proposed by Friston and colleagues is that of a general tendency of the brain to counteract surprising and unpredictible sensory events through building generative models that improve their predictions over time and render the world more amenable. This improvement is mainly done through sampling the environment and extracting statistical invariants that are used in return to predict upcoming events. Building a model thus rests on extracting a repertoire of invariants and organizing them so as to process the incoming sensory data efficiently through predictive coding (see Rao & Ballard (1999)). This proposition, gathered under the "Variational Free Energy Minimization" umbrella, is reminiscent of the auto-encoding theory proposed by Hinton & Zemel (1994), but introduces a new perspective on coding for *it formally links dictionary construction from data and (optimal) motor control*. In particular, motor control is here considered as a particular implementation of a *sampling process*, that is at the core of the estimation of a complex posterior distribution.

The active inference approach relies on a longstanding history of probabilistic modelling in signal processing and control Kalman (1960); Baum & Petrie (1966); Friston et al. (1994). Put formally, the physical world takes the form of a generative process p that is the cause of the sensory stream. This process is not visible in itself but is only sensed through a (non reliable) measure process that provides an observation vector x. The inference problem consists in estimating the underlying causes of the observation, that rests on a latent state vector z and a control u. Then, the evolution of x relies on both u and z in the form of a stochastic dynamical system undergoing an external forcing command u, i.e. :

$$\dot{z} = A(z) + B(u) + \text{process noise} \tag{1}$$

$$x = C(z) + D(u) + \text{measurement noise} \tag{2}$$

where A, B, C and D constitute a generative model $p(x, u, z, \dot{z})$ that explicits the dependencies between u, x and z. The calculation of $\dot{z}$ from z and u and the calculation of z from u and x rely on a model $p = \{A, B, C, D, \text{noise models}\}$. The model can then be inverted in order to compensate the drift of the state estimate Kalman (1960); Baum & Petrie (1966). The more accurate the model, the better this estimation.

2 PERCEPTION-DRIVEN CONTROL

The question addressed by Friston et al. (2012) is the design a *controller C that outputs a control u from z* so as to maximize the accuracy of this state estimation process. This is the purpose of a *perception-driven* controller.

The logic behind the model is that of an external sensory scene X that is never undisclosed in full, but only sensed under a particular view x under sensor orientation u (like it is the case in foveated vision). We moreover consider an organization of the visual scene in objects (or causes), whose presence and position is continuously checked by visual inspection. The objects may be described by their identity o and position in space s, but for simplicity o and s are here reduced to a single variable $z = (o, s)$.

Knowing that z is invariant to changing the sensor position u, uncovering z should rest on collecting sensory patches x's through changing u (sensor orientation) across time in order to refine z's estimation. Considering now that a certain prior $\pi(z)$ has been formed about z, choosing u conducts the sight in a region of the visual field that provides x, which in turn allows to refine the estimation of z. Each saccade should consolidate a running assumption about z (scene constituents), that may be retained and propagated from step to step, until enough evidence is gathered.

Instead of choosing u at random, the general objective of an *active inference* framework is to choose u in a way that should minimize *at most* the current uncertainty about z. The knowledge about z can be reflected in an inference distribution $q(z)$. The better the knowledge (precision) about a sensory scene, the lower the *entropy* of q, with :

$$H(q) = E_z[-\log q(z)]$$

It is shown in Friston et al. (2012) that minimizing the entropy of the posterior through action can be linked to minimizing the variational free energy attached to the sensory scene.

The control u is thus expected to reduce at most the entropy of $q(z)$ at each step. This optimal u is not known in advance, because x is only read *after* u has been carried out. Then comes the predictive framework that identifies the effect of u with its most probable outcome, according to a *generative* model p.

If we take a little step back, the general formulation of the generative model is that of a feedback control framework, under a discrete Bayesian inference formalism. Given an initial state z_0, the prediction rests on two conditional distributions, namely $p(z|u, z_0)$ – the link dynamics that generates z – and $p(x|z, u)$ – the measure process that generates x – (the discrete analog of (1) and (2)). Then, the forthcoming posterior distribution is (Bayes rule) :

$$p(z|x, u, z_0) = \frac{p(x, z|u, z_0)}{p(x|u, z_0)} = \frac{p(x|z, u)p(z|u, z_0)}{\sum_{z'} p(x|z', u)p(z'|u, z_0)}$$

so that the forthcoming entropy is :

$$H(q)|_{u,z_0} = E_{x,z}\left[-\log p(z|x, u, z_0)\right]$$

and the optimal u is :

$$\hat{u} = \underset{u \in \mathcal{U}}{\text{argmin}}\, H(q)|_{u,z_0}$$

with :

$$q(z)|_{u,z_0} = p(z|x, u, z_0)$$

In practice, the analytic calculations are out of reach (in particular for predicting the next sensory field x). One thus need to consider an *estimate* $\tilde{u} \simeq \hat{u}$ that should rely on sampling from the generative process p to predict the effect of u, i.e.

$$\tilde{u} = \underset{u}{\text{argmin}} \frac{1}{N} \sum_{i=1..N, x_i \sim p, z_i \in \mathcal{Z}} -\log p(z_i|x_i, u, z_0)$$

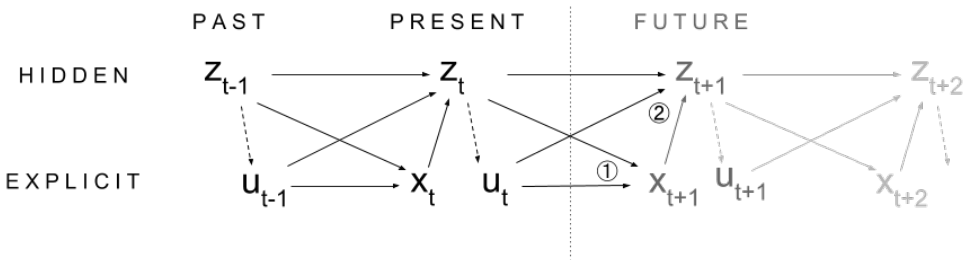

FIGURE 1 – Graphical model (see text)

or on an even sharper direct estimation through maximum likelihood estimates (point estimate) :

$$x_{\text{max}} = \underset{x}{\text{argmax}}\, p(x|u, z_0)$$
$$z_{\text{max}} = \underset{z}{\text{argmax}}\, p(z|x_{\text{max}}, u, z_0)$$
$$\tilde{u} = \underset{u}{\text{argmin}}\ -\log p(z_{\text{max}}|x_{\text{max}}, u, z_0)$$

This operation can be repeated in a sequence, where the actual control $u = \tilde{u}$ is followed by reading the actual visual field x, which in turn allows to update the actual posterior distribution over z. This updated posterior becomes the prior of the next decision step, i.e. $z_0' \sim p(z|x, u, z_0)$ so that a new control u' can be carried out etc.

If we denote T the final step of the process, with $u_{1:T}$ the actual sequence of controls and $x_{1:T}$ the actual sequence of observations, the final posterior estimate becomes $p(z_{1:T}|z_0, u_{1:T}, x_{1:T})$, which complies with a Hidden Markov Decision Process estimation (see fig. 1), whose policy would be defined by the entropy minimization principles defined above, precisely to facilitate the estimation process. The active inference framework thus appears as a *state-estimation oriented policy* (it has no other purpose than facilitate estimation).

If we now turn back to the active vision analogy, $q(z_t)$ is the current assumption about the visual scene under exploration, which is used as a context ("I'm in a park with many squirrels around" – z_t –) to form a prediction ("If I look up, ..." – u_t – "I may see a squirrel" – $\tilde{x}_{t+1}$ –) which would reduce the entropy of the posterior $q(z_{t+1})$ for it would reduce the uncertainty about the constituents of the scene. This incites me to look up (rather than looking down the grass) so that I can effectively process x_{t+1}, which may result in improving (or not) the entropy of the posterior depending whether my prediction was correct or not. I may then look to other places up around to precisely locate squirrels, but I probably don't need to inspect the grass or the lake for the same purpose, so I can currently avoid spending time looking there.

The active perception framework allows many relieving simplification from the general POMDP estimation framework, first in considering that changing u has no effect on the scene constituents, i.e. $p(z_{t+1}|u, z_t) = p(z_{t+1}|z_t)$. Then using the *steady state* assumption, that considers that no significant change should take place in the scene constituents during a saccadic exploration process, i.e. $\forall t, t', z_t = z_{t'} = z$. This finally entails a simplified chaining of the posterior estimation :

$$p(z|x_{t+1}, u_t) = \frac{p(x_{t+1}|z, u_t)p(z|x_t, u_t)}{\sum_{z'} p(x_{t+1}|z', u_t)p(z'|x_t, u_t)}$$

with $p(z|x_t, u_t)$ calculated at the previous step, issuing a final estimate $p(z|z_0, u_{1:T}, x_{1:T})$.

3 INTERPRETATION

The active inference framework, that is rooted on the auto-encoding theory (Free Energy minimization) and predictive coding, provides a clear roadmap toward an effective implementation in artificial devices. It should rely on three elements, namely a generative model p that should predict the next

Algorithm 1 Prediction-Based Policy

Require: p (generator), q (discriminator), π (prior), $\mathcal{U}$ (actions set)
predict $z \sim \pi$
$\forall u \in \mathcal{U}$, predict $\tilde{x}_u \sim p(x|z, u)$
return $\tilde{u} = \underset{u \in \mathcal{U}}{\text{argmax}}\ q(z|\tilde{x}_u, u)$

Algorithm 2 Scene Exploration

Require: p (generator), q (discriminator), π_0 (initial prior), $\mathcal{U}$ (actions set)
$\pi \leftarrow \pi_0$
while $H(\pi) > H_{\text{ref}}$ **do**
choose : $\tilde{u} \leftarrow$ Prediction-Based Policy$(p, q, \pi, \mathcal{U})$
read : $x_{\tilde{u}}$
update : $\forall z, \text{odd}[z] \leftarrow \log q(z|x_{\tilde{u}}, \tilde{u}) + \log \pi(z)$
$\pi \leftarrow \text{softmax}(\text{odd})$ {*posterior becomes prior*}
$\mathcal{U} \leftarrow \mathcal{U} \setminus \{\tilde{u}\}$
end while
return π

sensory scene under the current hypothesis z, an inference network q that should predict the next posterior under putative observation $\tilde{x}$, and a policy ρ that should use those predictions to issue a scene-estimation oriented control u.

Those "two-steps ahead" predictions entail a cognitive architecture based on a graphical model shown in figure 1. This formal similarity with auto-encoding architectures suggests that scene exploration may be learned in an unsupervised way, with both generative and inference models learned aside to give support to the scene exploration policy. It must be noticed that the generative model (1) stems from the present z_t and present u_t while the inference model (2) heads toward the future z_{t+1}. In the case of active vision however, we assume that $z_t = z_{t+1}$ (steady-state assumption – see previous section) so that the generative and inference models may be learned over the same encoding. The upper arrow represents the fact that the memory of the previous state participates in the present state estimate, and the dashed arrow represents the control policy, that is not inferred but optimized from two steps ahead predictions.

From the computer vision perspective, each different u corresponds to a different pose or a different viewpoint, so that a distinct set of inference/generative models should be learned for each different u. The idea of having many models to identify a scene complies with the weak classifiers evidence accumulation principle used in computer vision (see Viola et al. (2003)). In static image analysis, the weak classifiers generally reduce to a set of low-level filters, with u corresponding to the coordinates at which the filters are applied. They may also correspond to the first layer of convolution filters used in convolutional neural networks. Under that perspective, choosing u (or choosing a subset of u's) corresponds to choosing the set of coordinates at which the filter is applied, with each image patch obtained from a particular coordinate corresponding to a "viewpoint" over the whole image (like looking in a keyhole).

The active inference perspective also addresses important aspects of representation learning, for the scene constituents encoding z are learned as invariants over many viewpoints. The structure of the coding scheme imposes z not to (or minimally) vary across views (over the same object). Following the perspective proposed in Bengio et al. (2017), one may consider a dual encoding of a view $x : u$ and z, with z accounting for the view-independent scene identity information and u accounting for the gaze orientation (or subjective position) information. In our case the gaze orientation information is fully disclosed (as it belongs to the control parameters), while the scene identity is not. This separation of the encoding in two components is only viable if there is a cross-talk between components, i.e. if u informs z about the position at which x is seen as a side variable over the generator and the discriminator, i.e. $u, z \rightarrow x$ (generation) and $u, x \rightarrow z$ (inference).

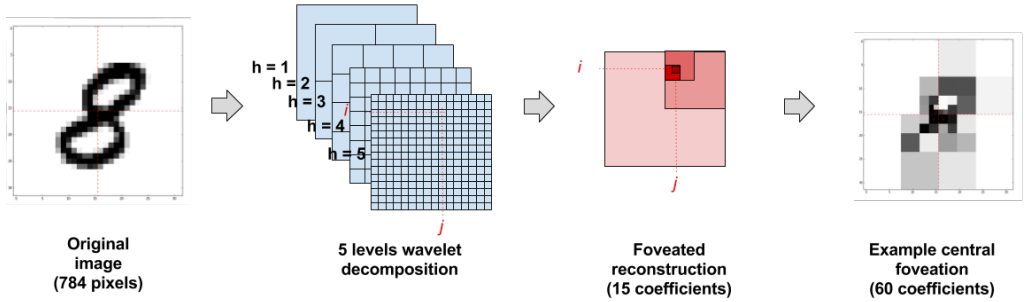

FIGURE 2 – Foveated image construction

4 IMPLEMENTATION

As a preliminary step here, we suppose the predictive and generative models are given to the controller. The discriminator and the generator are trained separately in a supervised fashion so that we can evaluate the properties of the control policy. This *model-based* approach to sequential view selection is provided in algorithms 1 and 2. It globally complies with the predictive coding framework (Rao & Ballard (1999)) with the predictions from the actual posterior estimate used to evaluate the prediction error and update the posterior.

A significant algorithmic add-on when compared with the original formulas is the use of a *dynamic actions set* : $\mathcal{U}$. At each turn, the new selected action $\tilde{u}$ is drawn off from $\mathcal{U}$, so that the next choice is made over fresh directions that have not yet been explored. This implements the inhibition of return principle stated in Itti & Koch (2001) and Friston et al. (2012).

A second algorithmic aspect is the use of a threshold H_{ref} to stop the evidence accumulation process when enough evidence has been gathered. This threshold is a free parameter of the algorithm that sets wether we priviledge a conservative (tight) or optimistic (loose) threshold. The stopping criterion is the pillar of estimation-based feature selection and needs to be optimized to arbitrate between resource saving and coding accuracy.

4.1 FOVEA-BASED IMPLEMENTATION

Natural scene exploration uses two principal tricks to minimize sensory resource consumption. The first trick is the foveated retina, that concentrates the photoreceptors at the center of the retina, with a more scarce distribution at the periphery. A foveated retina allows both treating central high spatial frequencies, and peripheral low spatial frequencies at a single glance (i.e process several scales in parallel). The second trick is the sequential saccadic scene exploration, already mentioned, that allows to grab high spatial frequency information where it is necessary (serial processing).

The baseline foveated vision model we propose relies first on learning local foveated viewpoints on images. Foveated image decomposition is rarely proposed in literature for machine learning purposes, at the exception of Simoncelli's group that has developed a framework to process centrally-magnified images in a bio-realistic fashion (see Freeman & Simoncelli (2011) and Deza et al. (2016)). For simplicity and consistency purposes, we restrain here the foveal transformation to its core algorithmic elements, i.e. the compression of an image according to a particular spatial focus. The essential properties of foveated images is that they retain high spatial frequency information at the center an keep only low-frequency information at the periphery. Our foveal image compression rests on a 2D wavelet decomposition of images according to the Haar wavelets dictionary. Taking the example of the MNIST database, we first transform the original images according to a 5-levels wavelet decomposition (see figure 2). We then define a viewpoint u as a set of 3 coordinates (i, j, h), with i the row index, j the column index and h the spatial scale. Each u may correspond with a view of three of wavelet coefficients $\boldsymbol{x}_{i,j,h} \in \mathbb{R}^3$, obtained from an horizontal, a vertical and an oblique filter at location (i, j) and scale h. The multiscale visual information $\boldsymbol{x}_{i,j} \in \mathbb{R}^{15}$ available at coordinates (i, j) corresponds to a set of 5 coefficient triplets, namely $\boldsymbol{x}_{i,j} = \{x_{i,j,5}, x_{i/2,j/2,4}, x_{i/4,j/4,3}, x_{i/8,j/8,2}, x_{i/8,j/8,1}\}$ (see figure 2), so that each multiscale view

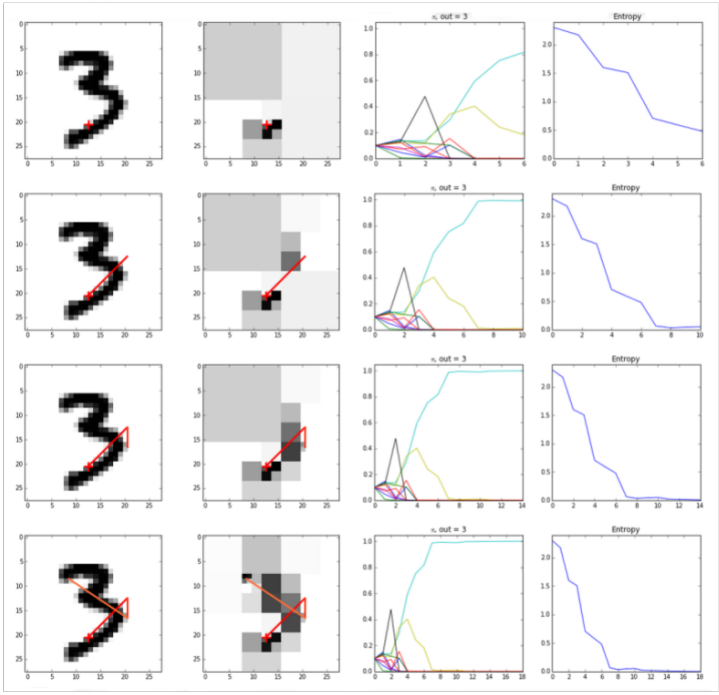

FIGURE 3 – Image exploration through saccades in the foveated vision model. The left column gives the succesive saccades position over the original image. The second column shows the progressive image reconstruction from the coefficients gathered from the saccadic exploration. The third column shows the evolution of the posterior distribution in function of the number of wavelet coefficients triplets read out. The last columns provides the update of the entropy of the posterior distribution

owns 15 coefficients (as opposed to 784 pixels in the original image). The rightmost image in figure 2 displays a reconstructed image from the 4 central viewpoints at coordinates (7, 7), (7, 8) (8, 7) and (8, 8).

A weak generative model is learned for each $u = (i, j, h)$ (making a total of 266 weak models) over 55,000 examples of the MNIST database. For each category z and each position u, a generative model is built over parameter set $\Theta_{z,u} = (\rho_{z,u}, \boldsymbol{\mu}_{z,u}, \boldsymbol{\Sigma}_{z,u})$, so that $\forall z, u, \tilde{x}_{z,u} \sim \mathcal{B}(\rho_{z,u}) \times \mathcal{N}(\boldsymbol{\mu}_{z,u}, \boldsymbol{\Sigma}_{z,u})$ with $\mathcal{B}$ a Bernouilli distribution and $\mathcal{N}$ a multivariate Gaussian. The Bernouilli reports the case where the coefficient triplet is null in the considered portion of the image (which is quite common in the peripheral portions of the image), which results in discarding the corresponding triplet from the Gaussian moments calculation. Each resulting weak generative model $p(.|z, u)$ is a mixture of Bernouilli-gated Gaussians over the 10 MNIST labels. For the generative model, a posterior can here be calculated explicitly using Bayes rule, i.e. $q(.|x, u) = \text{softmax} \log p(x|., u)$.

The saccade exploration algorithm is an adaptation of algorithm 2. The process starts from a loose assumption based on reading the baseline lower-level coefficient of the image, from which an initial guess π_0 is formed. Then, each follow-up saccade is calculated on the basis of final coordinates $(i, j) \in [0, .., 15]^2$, so that the posterior calculation is based on several coefficient triplets. After selecting (i, j), all the corresponding coordinate triplets (h, i, j) are discarded from $\mathcal{U}$ and can not be reused for upcoming posterior estimation (for the final posterior estimate may be consistent with a uniform scan over wavelet coefficients).

An example of such saccadic image exploration is presented on figure 3. The first row shows the state of the process after one saccade. If we look at the central right figure, we see the accumulation of evidence over the scale sub-steps. There is a total of six sub-steps (not including the uniform basis step), over which a posterior is formed on the basis of 15 wavelet coefficients (plus a baseline wavelet coefficient summing all the pixels), that provide a recognition probability around 80% for the category 3 (that is correct here). The next saccade (second row) heads toward a region of the

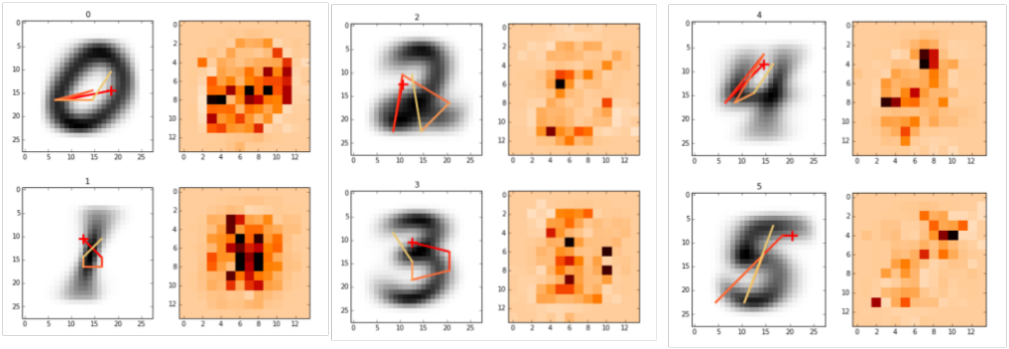

FIGURE 4 – Saliency maps inferred from the model and corresponding saccades prototypic trajectories for the six first categories.

image that is expected to help confirm the guess, which is the case for the recognition climbs up to about 99%. The next saccade (third row) allows to reach 99.9% in favor of class 3, and the final saccade (last row) allows to reach the threshold, set at $H_{\text{ref}} = 1e^{-5}$ for the entropy.

The model provides apparently realistic saccades, for they cover the full range of the image and tend to point over regions that contain class-characteristic pixels. The image reconstruction after 4 saccades allow to visually recognize a "fuzzy" three, while it would not have been the case if the saccades had been chosen at random. The observed trajectory illustrates the *guess confirmation* logic that is behind our active vision framework. Every saccade heads toward a region that is supposed to confirm the current hypothesis. This confirmation bias appears counter-intuitive at first sight, for some would expect the eye to head toward places that may *infirm* the assumption (to challenge the current hypothesis). This is actually not the case for the class-confirming regions are more scarce than the class-infirming regions, so that heading toward a class-confirming region may bring more information in the case it would, by surprise, invalidate the initial assumption.

A strong aspect of the model is the search for an optimal image processing compression, that is reflected in the number of wavelet coefficients used in the reconstruction. The average number of saccades is between 9 and 12 in our setting, corresponding to a compression rate of about 85 %. It can be more if the threshold is more optimistic, and less if it is more conservative. The number of saccades is representative of the recognition difficulty. Fewer saccades reflect a “dead-easy” recognition, while many saccades reflect a cumbersome recognition.

With a posterior approaching $1 - 1e^{-5}$, the model is overconfident in its own predictions which introduces a confirmation bias that tends to harm the final recognition rate of the model (which is here around 88% classification accuracy). It must be noticed however that the confirmation bias is probably a more general feature of the active inference framework that should be specifically addressed. For there is no free lunch, it corresponds to the "price to pay" for reducing the observation range of the image at the risk of potentially neglecting critical information.

Another follow-up aspect is the bad scaling of the full predictive model when large images are considered. A critical component of the predictive model is the two-steps ahead prediction that is necessary for the future posterior estimate. It seems in practice unrealistic to make prediction over the whole action set to foresee every possible consequences of every action. Rather a parametrized policy would be much preferable, allowing for a single draw over the action set. Luckily, this parametrized policy is relatively straightforward to compute in advance, given a generative model p (and a discriminator model q). If we note $\mu_{z,u}$ the generative prediction of x when the object z is seen under visual orientation u, then :

$$\forall z, q(z|.) = \text{softmax} \log p(\mu_{z,.}|z, .)$$

with the dot “.” a placeholder for the visual orientation. Then a rough estimate of the expected posterior in every viewpoint is provided, allowing to bypass all expensive online predictive compu-

tations. Each saliency map then provides an optimal pathway through the image that, given a certain assumption z, provides the most promising next viewpoint regarding its confirmation.

Examples of such saliency maps are provided in figure 4, for categories 0 to 5. The saliency maps allow to analyze in more detail the class-critical positions (that appear brownish) as opposed to the class-unspecific locations (pale orange to white). First to be noticed s the relative scarceness of the salient locations for most of the class. Those "highly informative" locations appear, as expected, mutually exclusive from category to category. A small set of saccades is expected to provide most of the classification information while the rest of the image is putatively uninformative (or even counter informative if whitish). A second aspect is that the class-relevant locations are all located in the central part of the images, so there is very few chance for the saccades to explore the periphery of the image where little information is expected to be found. This indicates that the model has captured the essential concentration of class-relevant information in the central part of the images for that particular training set.

5 OUTLOOK AND PERSPECTIVES

We developed a comprehensive description of the active inference framework, as proposed by Friston (2010) and Friston et al. (2012), under a machine-learning compliant perspective. Stemming from a biological inspiration and the auto-encoding principles, we develop a sketch of a cognitive architecture that should provide ways to implement *estimation-oriented* control policies. This may be particularly useful for developing active information search in the case high dimensionality input data. The pros and cons of the approach are reviewed in details, showing interesting promises in term of computation compression, but also putative risks of a confirmation bias that may degrade te recognition performance if the model is too optimistic about its own predictions. The presented formalism is fully compliant with the auto-encoding framework and would deserve further developments with deep variational auto-encoding architectures Makhzani et al. (2015).

RÉFÉRENCES

John Aloimonos, Isaac Weiss, and Amit Bandyopadhyay. Active vision. *International journal of computer vision*, 1(4) :333–356, 1988.

Ruzena Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8) :966–1005, 1988.

Dana H Ballard. Animate vision. *Artificial intelligence*, 48(1) :57–86, 1991.

Leonard E Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, pp. 1554–1563, 1966.

Emmanuel Bengio, Valentin Thomas, Joelle Pineau, Doina Precup, and Yoshua Bengio. Independently controllable features. *arXiv preprint arXiv :1703.07718*, 2017.

Arturo Deza, Emre Akbas, and Miguel P. Eckstein. Piranhas toolkit : Peripheral architectures for natural, hybrid and artificial systems. `https://github.com/ArturoDeza/Piranhas`, 2016.

Jeremy Freeman and Eero P Simoncelli. Metamers of the ventral stream. *Nature neuroscience*, 14 (9) :1195–1201, 2011.

Karl Friston. The free-energy principle : a unified brain theory ? *Nature Reviews Neuroscience*, 11 (2) :127–138, 2010.

Karl Friston, Rick Adams, Laurent Perrinet, and Michael Breakspear. Perceptions as hypotheses : saccades as experiments. *Frontiers in psychology*, 3 :151, 2012.

Karl J Friston, Andrew P Holmes, Keith J Worsley, J-P Poline, Chris D Frith, and Richard SJ Frackowiak. Statistical parametric maps in functional imaging : a general linear approach. *Human brain mapping*, 2(4) :189–210, 1994.

Geoffrey E Hinton and Richard S Zemel. Autoencoders, minimum description length and helmholtz free energy. In *Advances in neural information processing systems*, pp. 3–10, 1994.

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

Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering*, 82(1) :35–45, 1960.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv :1511.05644*, 2015.

Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex : a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1), 1999.

David A Robinson. Eye movement control in primates. *Science*, 161(3847) :1219–1224, 1968.

M Viola, Michael J Jones, and Paul Viola. Fast multi-view face detection. In *Proc. of Computer Vision and Pattern Recognition*. Citeseer, 2003.

Alfred L Yarbus. Eye movements during perception of complex objects. In *Eye movements and vision*, pp. 171–211. Springer, 1967.