Machine Cognitive Set-Shifting: Supplementary information

Yen Yu[1*], Acer Y.C. Chang[1], Ryota Kanai[1],

[1] Araya, Inc., Tokyo, Japan

* Corresponding author: Yen Yu (first.lastname@araya.org)

Abstract

Cognitive set-shifting as one of the constituent mechanisms of cognitive flexibility speaks to the shaping of perceptual processing. This shaping ensures different percepts are admissible such that alternative interpretations can be registered from the same sensory input. Set-shifting is therefore central to intelligent being's ability to escape monotonous, reflex-like stimulus-response mapping. This is because exhibiting diverse behavioural programmes is necessarily contingent upon the *constructional freedom* in perceptual contexts. In this study, we considered core functions of set-shifting identified in the past psychological and neurological findings and proposed a model for set-shifting in statistical machine learning terms. To this end, we first entertained the notion that perceptual inference, seen as the inversion of an agent's generative model of the world, is the pre-attentive aggregation of cross-dimensional latent causes that jointly explain away sensory surprisal. This inference process is automatic and activates all relevant perceptual dimensions. Cognitive sets can then be cast as a belief over *reconfiguration patterns* for weighing and routing a subset of perceptual dimensions. A reconfiguration pattern is an encoding schema for higher-level representations in working memory. Information available in working memory thus provides the substrate for task-oriented interpretation. This interpretation involves devising downstream behavioural decisions which we argued suffer from high *complexity* if the routing mechanism is otherwise disabled. Crucially, shifting is driven by admitting conflict-reporting cues and conflict resolution which amount to a conditional belief given past decisions and outcomes. We implemented this model using deep neural networks and trained the networks to solve a modified Wisconsin Card Sorting Task. Finally, we reported the comparisons between model-based and model-free variants of our networks. We called this novel network architecture the Conflict Monitoring Network (CMN).

Model

The objective of our model is to minimise the following free energy functional with respect to model parameters θ and φ:

$$\min_{\theta,\varphi} \mathcal{F}(y_t, x_t, q_\varphi, \theta) \tag{1}$$

where θ parametrises the *Predictor* and φ the *Shifter*. x_t and y_t are Predictor inputs and targets at time t. $q_\varphi(\cdot)$ is the variational density over *reconfiguration pattern* ξ_t.

The above free energy function can be unpacked to reveal the following form:

$$\begin{aligned}\mathcal{F}(y_t, x_t, q_\varphi, \theta) = \mathbb{E}_q\left[-\ln p\big(y_t|f_\theta^\xi(x_t)\big)\right] + \\ D_{KL}\big[q_\varphi(\xi_t)\|p(\xi_t)\big]\end{aligned} \tag{2}$$

where

$$q_\varphi := \text{Bern}\left(\xi_t; \pi_\varphi\big(y_{<t}, f_\theta^\xi(x_{<t})\big)\right) \tag{3}$$

Supervised learning

During training and testing phases, the Predictor received the same input of the following form:

$$\begin{aligned} x_t &\in \mathbb{R}^4 \\ x_t^{(i)} &= \cos(2^{i-1}t) \end{aligned} \tag{4}$$

where time t falls within the range $[-2, 2]$. When used for testing, the input was modified with an additive Gaussian i.i.d. noise $\eta \sim N(0, 0.1)$.

The training data points were generated according to the following transformation:

$$\begin{aligned} y_{i:j} = h(\lambda_1 x_{i:j}^{(1)} + \phi_1) - h(\lambda_2 x_{i:j}^{(2)} + \phi_2) \\ + h(\lambda_3 x_{i:j}^{(3)} + \phi_3) - h(\lambda_4 x_{i:j}^{(4)} + \phi_4) \end{aligned} \tag{5}$$

The tuple (i, j) belongs to a set $\mathcal{P}$ of non-overlapping segments whose overall length is constrained by x and satisfies $i < j$. For example, let an input x_t defined on discrete time have length T, then its bipartition would be identified by the segments $(0, \tau)$ and $(\tau + 1, T)$. Each segment would have different frequency modulation and phase shifting coefficients, λ and ϕ. The training sets were generated via sampling τ, λ, and ϕ from predefined probability distributions.

$h(\cdot)$ denotes an arbitrary nonlinear function. We used $\sin(\cdot)$ for training and $\tanh(\cdot)$ for testing.

In what way is the relevance between our task and WCST justified?

In a typical WCST setting, a cue card is dealt in each trial by the experimenter for the subjects, who based their course of action on a belief over most relevant perceptual dimension, to match the target decks.

An interesting situation happens when the same cue is dealt the second time but implicitly the target dimension has shifted to a different one. The subjects unconsciously perceive the exact same information nonetheless. However, they have to consciously divert only the relevant information—through selecting, filtering, and reconfiguring the composition of their percepts—to guide their actions. All these happens on the fly, triggered by some error feedback, without further learning, as if the brain already possesses bags of tools for such task. What needs to be done is only picking the right tools to improvise the desired trick.

Our Predictor input (i.e., the four dimensional cosine bases) is analogous to the repeatedly dealt cue as mentioned above. The same percept can be interpreted and reinterpreted in so many ways only to be limited by the capacity of the Predictor and the reconfiguration patterns triggered by the conflict monitoring unit, the Shifter.

We hypothesise that all the possible reconfiguration patterns form a space where similar functions cluster. Transitions within one cluster is associated with minute conflict signals, whereas a conflict orders of magnitude larger would result in a transition across clusters. The latter is perhaps a more detectable one in behavioural sense.

Funding

This study was funded by the Japan Science and Technology Agency (JST) under CREST grant number JPMJCR15E2.

References

Berlyne, Daniel E (1960). *Conflict, arousal, and curiosity.* McGraw-Hill Book Company.

Botvinick, Matthew M (2007). "Conflict monitoring and decision making: reconciling two perspectives on anterior cingulate function". In: *Cognitive, Affective, & Behavioral Neuroscience* 7.4, pp. 356–366.

Botvinick, Matthew M et al. (2001). "Conflict monitoring and cognitive control." In: *Psychological review* 108.3, p. 624.

Cho, Kyunghyun et al. (2014). "Learning phrase representations using RNN encoder-decoder for statistical machine translation". In: *arXiv preprint arXiv:1406.1078*.

Gal, Yarin (2016). "Uncertainty in deep learning". In: *University of Cambridge*.

Gal, Yarin and Zoubin Ghahramani (2016). "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning". In: *international conference on machine learning*, pp. 1050–1059.

Gal, Yarin, Jiri Hron, and Alex Kendall (2017). "Concrete dropout". In: *Advances in Neural Information Processing Systems*, pp. 3584–3593.

Miller, Earl K and Jonathan D Cohen (2001). "An integrative theory of prefrontal cortex function". In: *Annual review of neuroscience* 24.1, pp. 167–202.

Norman, Donald A and Tim Shallice (1986). "Attention to action". In: *Consciousness and self-regulation*. Springer, pp. 1–18.