Machine Cognitive Set-Shifting

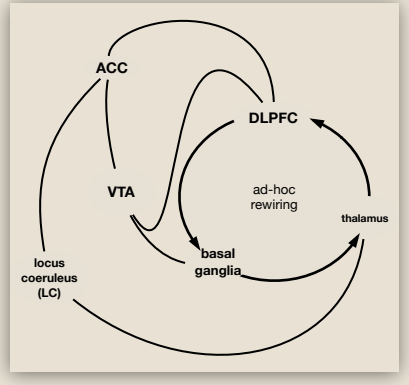
Yen Yu, Acer Y.C. Chang, Ryota Kanai

Araya, Inc., Tokyo

OVERVIEW & TAKEAWAYS

- **Key message:** We presented a novel network architecture capable of self-monitoring and self-regulation by learning to adaptively reconfigure network connections on the fly. We called this architecture the Conflict Monitoring Network (CMN).
- **Relation to set-shifting:** A set entails a class of representation-transforming functions. Set-shifting then rests on categorical invocation of functional regimes, giving rise to constructional and interpretational freedom on internal representations.

NEUROSCIENTIFIC MOTIVATION



- **Prefrontal cortex (PFC):** The *Guided Activation* theory proposed that the PFC is not responsible for input-output mappings for behavioural purposes; instead, it modulates other units responsible for said mappings. This had led to the interpretation that the PFC is an attentional device for ad hoc rewiring or re-routing of information processing in its subordinate units.
- **Conflict monitoring:** The *conflict monitoring theory* proposed the anterior cingulate cortex monitors conflicts in information processing, leading to downstream alterations in arousal, attentional, and working memory states.

MACHINE LEARNING MOTIVATION

- **Universal approximator:** Any multilayer feedforward artificial neural network can potentially represent a wide range of continuous functions, given carefully selected network weights.
- **Limiting factor:** In spite of its immense potential, many use cases, came with specific sets of training data and objective functions, require neural networks to learn *one*, albeit highly complex, function.
- Approaches: The subject of *meta-learning* studies how machines can learn to learn, thereby enabling of acquisition of multiple functions within one architecture. Famous cases include MAML. However, MAML requires task labels during training for post-training adaptation.
- Our idea: We let the neural network learn a distribution of functions and leave until later the decision of which function to reconstitute. A trained network adapts by muffling connections instead of tuning its weights.

FORMALISM

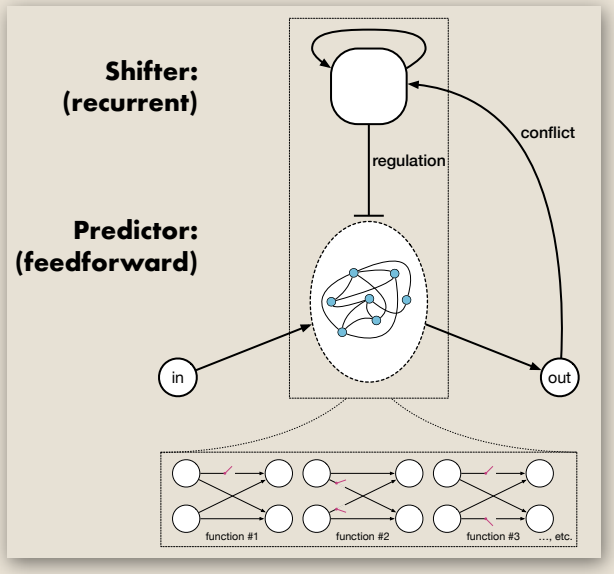
$$\min_{\theta,\varphi} \mathcal{F}(y_t, x_t, q_\varphi, \theta)$$

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

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

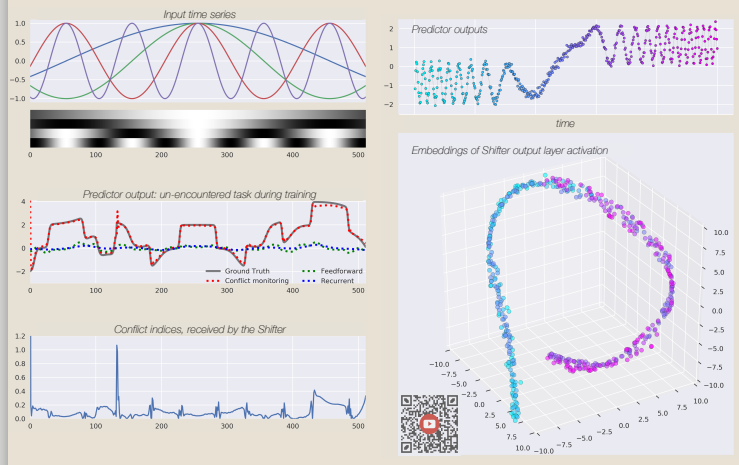
FOLLOW-UP

NETWORK ARCHITECTURE



- **Predictor:** May assume any architecture; here we used a multilayer perceptron. The input and output are two time series, though a feedforward Predictor does not model any temporal structure latent in the data. Instead, at each time point the Predictor is contextualised by the modulation from the Shifter.
- **Shifter:** Learns the causal structure of conflicts, as measured by the past performance of the Predictor (e.g., prediction-target mismatch). It then outputs a reconfiguration pattern which works like on/off switches on the Predictor connectivity.

EXPERIMENTS & VIDEO



- **Left:** This demonstrates the ad hoc rewiring of the Predictor network, as regulated by the Shifter which receives conflict indices from the Predictor. The CMN was being tested on functions it had never been trained on, thus evidently reflected on the conflict indices. The sharp peak indicates two distinct segments of target.
- Right: Using the same trained CMN and the set of inputs as shown in the left panel, the network was given a novel time series as the target for prediction. The top insert shows the predictor outputs, superimposed on the ground truth (grey dashed line). The bottom plot shows the shift between two dynamic regimes of the Shifter outputs. Each dot represents at one time point the reconfiguration pattern, embedded in 3D space. Video included.