Contribution ID: **12**

Type: **Talk**

Explainable deep learning inference to decode decision-making processes from multidimensional patterns of neural activities

Wednesday 26 October 2022 15:15 (15 minutes)

Connecting effective behavioural models to high-dimensional experimental data is one of the most important challenges in neuroscience. In this work we show how a link can be inferred between intracortical neural recordings performed during a simple decision task and the ramping variable postulated by a threshold decision model exploiting an artificial neural network (ANN).

We recorded the multi unit activity (MUA) from a 96-channel array in the dorsal premotor cortex of two monkeys performing a countermanding reaching task that requires, in a subset of trials, to cancel the planned movement before its onset. We trained a WaveNet-inspired and a multilayer perceptron causal ANN architecture to map this complex data to an accumulation process as derived from theoretical models.

Our results show that neural recordings can be, to a large extent, mapped at the single trial level to a ramping process, whose angular coefficient is tied to the inverse of the movement reaction time (RT). From the predicted ramp it is possible to perform an early estimation of the RT and the outcome of the stop trials with a good level of accuracy. Moreover, the network generalises nicely when tested on sessions having large temporal gaps with the ones used for training (generalisation through time).

We then applied explainability techniques (xAI) to our network to extract insights of the information hidden in the input data. By combining already established xAI algorithms based on Gradient methods and a newly proposed xAI method, that we call “functional explainability”, we show how, by perturbing the output function in principled ways, we obtain different spatio-temporal patterns of “saliency”in the input. Notably, our results suggest that the information used to build the ramps (so the decision rate and consequently the RT) can be found very early after, if not before, the onset of the visual clue indicating the target.

We also employed methods of training influence (TracIn) to find the training examples most relevant for a given prediction. The analysis confirms that the performance of the network on a given test trial is positively influenced by training examples recorded weeks and even months before or after it. Furthermore, when the network is trained on data from two monkeys, TracIn highlights a substantial inter-subject influence, thus hinting at a partial shared representation of the hypothesised ramping process at the neuronal level.

Primary authors: Dr CIARDIELLO, Andrea (INFN sezione di Roma); Dr BUONFIGLIO, Antonio (Department of Biomedical and Neuromotor Sciences, University of Bologna and Institute of Cognitive Sciences and Technologies (ISTC), National Council of Research of Italy (CNR)); Dr BARDELLA, Giampiero (Department of Physiology and Pharmacology, Sapienza University); Prof. PANI, Pierpaolo (Department of Physiology and Pharmacology, Sapienza University); Prof. FERRAINA, Stefano (Department of Physiology and Pharmacology, Sapienza University); Dr MATTIA, Maurizio (Natl. Centre for Radiation Protection and Computational Physics, Istituto Superiore di Sanita, Italian Institute of Health); Dr GIGANTE, Guido (Natl. Centre for Radiation Protection and Computational Physics, Istituto Superiore di Sanita, Italian Institute of Health)

Presenter: Dr CIARDIELLO, Andrea (INFN sezione di Roma)

Session Classification: Session 2: Image Processing and 3D Reconstruction