Contribution ID: **131** Type: **not specified**

Attack based identification of the most informative patterns encoding the past in low rate hippocampal activity

Monday, November 28, 2022 2:55 PM (15 minutes)

Over the past decade, advances in computational power have vastly increased the interest in deep neural networks (DNNs) as an efficient machine learning method. Despite their impressive performance on state-of-the-art classification tasks, our understanding of DNNs remains incomplete, mainly owing to their black-box nature, which prohibits understanding of the geometry of the decision boundary. This short-coming particularly hampers the application of DNNs in neuroscience, where the decision boundary needs to be interpreted in terms of anatomical units such as neurons or brain regions, and further linked to cognitive functions and behavior. Here we explore working memory related activity in the hippocampus . We used a linear neural network to examine whether the previous behavior of an animal could be decoded from 100ms long time bins. To that end, permutation tests were performed, while shuffling randomly the labels of left and rightward trials in combination with a 2-fold cross-validation protocol . To directly identify the neuronal basis of the prediction scores of the classifier we visualized its decision boundary (DB) by applying adversarial attack techniques from machine learning. From this set of boundary positions we then constructed most informative directions (MIDs) as clusters of orthogonal vectors to the boundary. For low signal strength and artificial surrogate data, we show that the method outperforms estimating the weight vector by bootstrapping.

Presenter: ATHANASIADIS, M.

Session Classification: Session IV