THALAMOCORTICAL CIRCUIT INSIGHTS ENABLE ZERO-SHOT TRANSFER BY RECURRENT NETWORKS DURING HIERARCHICAL CONTROL OF CONTINUOUS MOTOR BEHAVIORS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study learning of recurrent neural networks that produce hierarchically organized continuous outputs consisting of the concatenation of re-usable 'motifs'. In the context of neuroscience or robotics, these motifs would be the motor primitives from which complex behavior is generated. Can a motif library be efficiently and extendably learned without interference between motifs, and can these motifs be chained in arbitrary orders without first learning the corresponding motif transitions during training? Two requirements enable this: *(i)* parameter updates while learning a new motif do not interfere with the parameters used for the previously acquired ones; and *(ii)* each motif can be generated when starting from the network states reached at the end of any of the other motifs, even if these states were not present during training (a case of out-of-distribution generalization). We meet the first requirement by designing artificial neural networks (ANNs) with specific architectures that segregate motif-dependent parameters, and try a standard trick to address the second by using random initial states during training. The performance is good when testing the ANNs for within-distribution generalization; however the performance drops during zero-shot transfer to motif chaining. This performance drop is a very robust feature which is observed for different motif types, randomness used during training, network designs; and which does not disappear when increasing the network size. We then use insights from an analytically tractable model whose architecture and dynamics are inspired by the motor thalamocortical circuit in mammals, featuring a specific module that shapes motif transitions. We develop a method to constrain the ANNs to function similarly to the thalamocortical model during motif transitions, while preserving the larger flexibility afforded by gradient-based training of non-analytically tractable ANNs. We then show that this inductive bias creates networks that can perform zero-shot transfer to motif chaining with no performance cost.