Review-27

Using Fast Weights to Attend to the Recent Past

Improving neural activity representations and weights to capture regularities among inputs restricts development of artificial neural systems. A natural alternative to prevent this resitrction from arising is leveraging variable time-scale operations motivated by dynamics of synapses. Weights that change faster than standard weights and slower than neural activities present a neurally plausible scheme for attending to events in the short term. The work build on this insight and makes use of fast weights as an attention mechanism to learn neural activity patterns while preventing the need to store their large copies. Fast weight matrices constructed using hidden representations are leveraged efficiently without computing the complete weight matrix and allow learning models to efficiently attend towards complex inputs in different learning settings.

Utilization of associative memory is an essential component in scenarios wherein the model may need to attend to large memory inputs. These inputs can be represented as compact synaptic footprints which can be learnt efficiently using a faster temporal scheme. Evidence of fast neural computation may be observed in physiology wherein the human brain leverages a variety of short-term plasticity mechanisms that operate on different timescales. This gives rise to fast weight updates in artificial neural networks wherein fast weight matrices are updated at an increased timescale in comparison to standard weights. Fast weights update recent history using weighted outer products of hidden representations which serve as attention weights to the model. In order to eliminate extra computation of fast weights, outer products are directly computed using previous hidden representations. Stabilization of the product is carried out using Layer Normalization which prevents the hidden matrices from vanishing or exploding values.

Utilization of the fast weight scheme demonstrates suitability in different learning domains such as supervised and reinforcement learning. Efficient associative retrieval of targets from input strings and improved performance of A3C RL agent on Catch indicate that fast-moving temporal updates indeed facilitate efficient short-term storage. Furthermore, comparable performance of the scheme to ConvNet on attending to facial glimpses depicts its suitability as a neurally plausible mechanism of attention. While fast weights improve short-term attention of models towards previously seen examples, the work does not provide any insights towards their scalability to unseen examples which is indeed the more practical challenge. Furthermore, attending to glimpses demonstrates comparable performance to ConvNet which indicates that visual attention models consisting of larger neural representations would be able to perform better on the task. This leaves various questions related to the efficiency and applicability of fast associative memory unanswered.

Fast weight updates in conjunction with slow moving weights provides a unique temporal perspective towards artificial neural representations motivated by the human brain. The work presents adaptation of fast weights to complex visual tasks as a possible direction for future research. Furthermore, the scheme can be improved in light of its applicability to larger neural representations and other biologically inspired phenomenon such as the problem of forgetting.

Lifting off temporal restrictions on the construction of neural activities is essential for attending to short term events. To this end, the work presents the framework of fast weights which updates weight matrices constructed from outer product of hidden vectors at a faster temporal resolution. Motivated by neural activations observed in physiology, fast weights present a suitable mechanism for attention which is empirically validated on associative retrieval tasks and learning visual glimpses and sequential control.