Plastic Bistable Recurrent Cells

Amiani Johns

August 12, 2021

Abstract

This is the abstract.

Introduction

This is the introduction.

Background

Recurrent Neural Networks

A recurrent neural network (RNN) is a type of neural network that excels at temporal prediction tasks. Inputs are fed into the network one at a time, and the model maintains a memory in the form of a hidden state vector $\mathbf{h}_t$ that is updated at each timestep. The model can then make predictions based on the contents of the input as well as the memory. At each timestep, the model updates the hidden state using the rule

$$\mathbf{h}_t = \sigma(W_h \mathbf{h}_{t-1} + W_x \mathbf{x}_t + b_h)$$

The output of the model can then be computed as

$$\mathbf{y}_t = \sigma(W_y \mathbf{h}_t + b_y)$$

where σ is a nonlinear activation function, W_h, W_x and W_y are weight matrices learned by the model, b_h and b_y are learned bias vectors and $\mathbf{x}_t$ is the input at time t. These weight matrices and bias vectors are typically learned using gradient descent.

While RNNs have proven effective for many tasks, in this standard form their ability to capture long-term dependencies is limited. This is because of the so-called vanishing/exploding gradient problem, which is a result of the repeated application of the weight matrix W_h at each timestep. Various approaches have been proposed to address this problem. One approach is to use a modified update rule that avoids changing the hidden state unless necessary. This is the approach used by the Gated Recurrent Unit.

Gated Recurrent Units

Gated Recurrent Units (GRUs) [1] are a type of RNN that preserve the gradient using a modified update rule. At each timestep the model calculates

$$\mathbf{z}_t = \sigma(W_{zh}\mathbf{h}_{t-1} + W_{zx}\mathbf{x}_t + \mathbf{b}_z)$$
$$\mathbf{r}_t = \sigma(W_{rh}\mathbf{h}_{t-1} + W_{rx}\mathbf{x}_t + \mathbf{b}_r)$$

known as the update and reset gate, respectively. These are then used to calculate

$$\tilde{\mathbf{h}}_t = \tanh(\mathbf{r}_t \odot W_{hh}\mathbf{h}_{t-1} + W_{hx}\mathbf{x}_t + \mathbf{b}_h)$$
$$\mathbf{h}_t = \mathbf{z}_t \odot \tilde{\mathbf{h}}_t + (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1}$$

where $W_{zh}, W_{zx}, W_{rh}, W_{rx}, W_{hh}, W_{hx}$ are weight matrices, b_z, b_r, b_h are bias vectors, and σ is a nonlinear activation function. Since at each timestep $\mathbf{h}_t$ is updated using linear interactions only, the vanishing gradient problem is much less pernicious, and the model is able to learn much longer term dependencies than the standard RNN model.

Bistable Recurrent Cells

Bistable Recurrent Cells (BRCs) [4] are another recurrent model similar in form to the GRU that allow for each individual unit of the memory vector to hold onto a value for an arbitrarily long time. The BRC features only local interaction of the hidden state, that is, that each hidden state neuron computes its activation value based only on the activation values of the neurons connected to it in the previous layer, the synaptic strengths of those connections and the activation value of the neuron itself at the previous timestep. Models that feature such local computations are interesting subjects of study because biological neural networks cannot do the global computations typically required for the operation of modern artificial neural networks. The BRC therefore modifies the update and reset gate equations to

$$\mathbf{z}_t = \sigma(\mathbf{w}_z \odot \mathbf{h}_{t-1} + W_z\mathbf{x}_t + \mathbf{b}_z)$$
$$\mathbf{r}_t = 1 + \tanh(\mathbf{w}_r \odot \mathbf{h}_{t-1} + W_r\mathbf{x}_t + \mathbf{b_r}$$

where $\mathbf{w}_z$ and $\mathbf{w}_r$ are now weight vectors multiplied elementwise with the previous hidden state. The equation for $\tilde{\mathbf{h}}_t$ is also modified, to

$$\tilde{\mathbf{h}}_t = \tanh(\mathbf{r}_t \odot \mathbf{h}_{t-1} + W_h\mathbf{x}_t + \mathbf{b}_h)$$

with the update to $\mathbf{h}_t$ being the same as the GRU. Since these equations change all matrix multiplications with the hidden state vector to elementwise multiplications, all interactions between elements of the hidden state are removed. Instead, each unit of the hidden state interacts only with itself and the input. Since it features only local computations, the BRC in this form is a much more plausible model of how biological neural networks might function than either RNNs or GRUs.

Neuromodulated Bistable Recurrent Cells

Vecoven et al. [4] also introduced another form of the BRC that reintroduces the interaction between elements of the hidden state in the equations for the update and reset gates. By relaxing the local computation requirement, this modified BRC showed improved performance on a number of tasks. The update and reset gate equations are now

$$\begin{aligned}\mathbf{z}_t &= \sigma(W_{zh}\mathbf{h}_{t-1} + W_{zx}\mathbf{x}_t + \mathbf{b}_z)\\ \mathbf{r}_t &= \sigma(W_{rh}\mathbf{h}_{t-1} + W_{rx}\mathbf{x}_t + \mathbf{b}_r)\end{aligned}$$

whereas the equations for $\tilde{\mathbf{h}}_t$ and the update to $\mathbf{h}_t$ are the same as the standard BRC. The hidden state units are modulated by the activations of the input units and hidden state units, thus the name Neuromodulated Bistable Recurrent Cell (nBRC).

Differentiable Plasticity

One of the ways that it is believed that biological neural networks update the connection strengths between neurons is through a process called synaptic plasticity. Plasticity is the ability of a synapse (connection between two neurons) to change its strength based on the activations of the neurons it connects. The most widespread theory of plasticity is called Hebb's rule [2], which was proposed as an explanation for learning and memory in the brain. Hebb's rule states that neurons that fire together, wire together i.e. that the strength of a synapse is increased when the activation of its presynaptic neuron is (perhaps along with other neurons) the cause of the activation of its postsynaptic neuron. Although many learning rules based on this fundamnetal idea are possible, Miconi et al. [3] proposed a plastic learning rule that is differentiable, a desirable property as it allows for the efficient training of neural networks with many parameters. In their formulation, synapses have two weights, one that is static during each episode (i.e. lifetime of the model), and one that is plastic, being updated based on its pre- and post-synaptic neuron activations. The plastic weight at each timestep according to the recursive formula

$$H_{i,j}(t) =$$

Plastic Recurrent Cells

Method

Plastic Bistable Recurrent Cells

In this paper we propose the Plastic Bistable Recurrent Cell (PBRC) model, which is a combination of the nBRC and diffentiable plasticity methods.

The Copy First Task

The copy first task is a temporal prediction task that is an effective test of a model's ability to maintain long term memories. In this task, the model is presented with a sequence of T inputs drawn from a standard multivariate normal distribution. At time T the model must output the first input of the series. All other outputs are discarded. Since the model is presented with a new random input at each timestep, it must learn not only to remember the first input, but also to ignore all subsequent inputs. When T is large (i.e. ¿ 5), this task is very difficult for most recurrent models, including the GRU.

Results

These are the results.

Conclusion

This is the conclusion.

References

[1] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[2] Donald Olding Hebb. *The organisation of behaviour: a neuropsychological theory*. Science Editions New York, 1949.

[3] Thomas Miconi, Kenneth Stanley, and Jeff Clune. Differentiable plasticity: training plastic neural networks with backpropagation. In *International Conference on Machine Learning*, pages 3559–3568. PMLR, 2018.

[4] Nicolas Vecoven, Damien Ernst, and Guillaume Drion. A bio-inspired bistable recurrent cell allows for long-lasting memory. *Plos one*, 16(6):e0252676, 2021.