Q-AGREL: Biologically Plausible Attention Gated Deep Reinforcement Learning

Isabella Pozzi (i.pozzi@cwi.nl)
Machine Learning Group, Centrum Wiskunde & Informatica
Amsterdam, Netherlands

Sander M. Bohté (s.m.bohte@cwi.nl)
Machine Learning Group, Centrum Wiskunde & Informatica
Amsterdam, Netherlands

Pieter R. Roelfsema (p.roelfsema@nin.knaw.nl)
Vision & Cognition Group, Netherlands Institute for Neuroscience
Amsterdam, Netherlands

Abstract

The success of deep learning in end-to-end learning on a wide range of complex tasks is now fuelling the search for similar deep learning principles in the brain. While most work has focused on biologically plausible variants of error-backpropagation, learning in the brain seems to mostly adhere to a reinforcement learning paradigm, and while biologically plausible neural reinforcement learning has been proposed, these studies focused on shallow networks learning from compact and abstract sensory representations. Here, we demonstrate how these learning schemes generalize to deep networks with an arbitrary number of layers. The resulting reinforcement learning rule is equivalent to a particular form of error-backpropagation that trains one output unit at any time. We demonstrate the learning scheme on classical and hard image-classification benchmarks, namely MNIST, CIFAR10 and CIFAR100, cast as direct reward tasks, both for fully connected, convolutional and locally connected architectures. We show that our learning rule - Q-AGREL - performs comparably to supervised learning via error-backpropagation, requiring only 1.5-2.5 times more epochs, even when classifying 100 different classes as in CIFAR100. Our results provide new insights into how deep learning may be implemented in the brain.

Keywords: Reinforcement learning; biologically plausible deep learning

Introduction

Similarly to deep neural networks, the brain of humans and animals are composed of many layers between the sensory neurons that register the stimuli and the motor neurons that control the muscles. Hence it is tempting to speculate that the methods for deep learning that work so well for artificial neural networks also play a role in the brain (Marblestone, Wayne, & Kording, 2016; Scholte, Losch, Ramakrishnan, de Haan, & Bohte, 2017). A number of important challenges need to be solved, however. First of all, the error-backpropagation rule (i.e. the method typically used to train deep artificial neural networks) was argued to be neurobiologically unrealistic (Crick, 1989). Researchers have started to address this challenge by proposing ways in which learning rules that are equivalent to error-backpropagation might be implemented in the brain (Urbanczik & Senn, 2014; Schiess, Urbanczik, & Senn, 2016; Brosch, Neumann, & Roelfsema, 2015; Richards & Lillicrap, 2019; Scellier & Bengio, 2019; Amit, 2018; Sacramento, Costa, Bengio, & Senn, 2018), most of which were reviewed in (Marblestone et al., 2016). One of the main challenges remained to inform synapses at the lower network levels about the desired change in their strength, because the influence of changes in their strength on activity in the output layer is only indirect and depends on many intermediate synapses. In addition, most of the algorithms still focus on learning high-rank representations, while animals learning to select actions by trial-and-error is intrinsically low-rank.

Here we will focus on a particular type of learning rule known as AGREL (attention-gated reinforcement learning) and AuGMEnT (attention-gated memory tagging) (Roelfsema & Ooyen, 2005; Rombouts, Bohte, & Roelfsema, 2015), which provide us with a biologically plausible (in particular, low-rank learning) solution for the lower synapses-update challenge. These learning rules realized that in a reinforcement learning setting the synaptic error derivative can be split into two factors: a reward prediction error (RPE) which is positive if an action selected by the network is associated with more reward than expected or if the prospects of receiving reward increase while it is negative if the outcome of the selected action is disappointing. In the brain, the RPE is signaled by neuromodulatory systems that project diffusely to many synapses so that they can inform them about the RPE (Schultz, 2002); the second factor is an attentional feedback signal that is known to propagate from the motor cortex to earlier processing levels in the brain (Roelfsema & Holtmaat, 2018; Pooresmaeili, Poort, & Roelfsema, 2014). When a network chooses an action, this feedback signal is most pronounced for those neurons and synapses that can be held responsible for the selection of this action and hence for the resulting RPE. These two factors jointly determine synaptic plasticity. As both factors are available at the synapses undergoing plasticity, it has been argued that learning schemes such as AGREL and AuGMEnT are indeed implemented in the brain (Roelfsema & Holtmaat, 2018). However, the previous AGREL and AuGMEnT models used networks with a single hidden layer, and modeled learning in tasks with only a handful input neurons.

The present work has two goals. The first is to establish the relation between the biologically realistic learning rules and

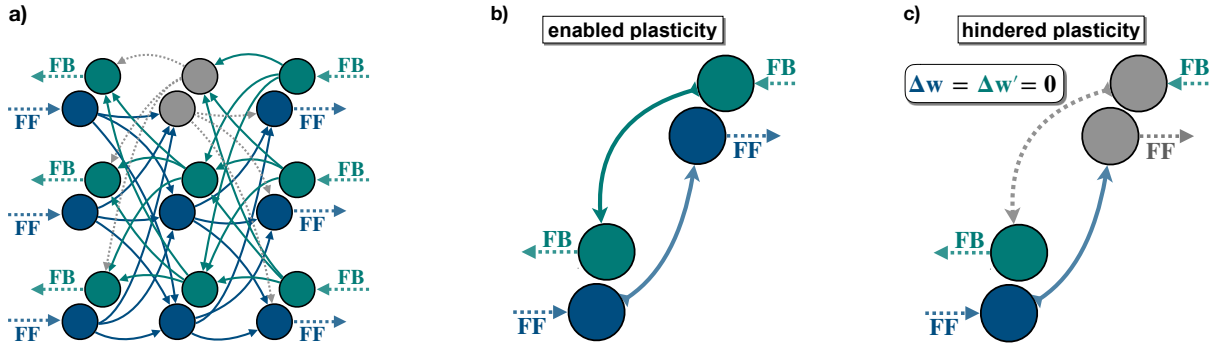

Figure 1: Q-AGREL algorithm plasticity gating. **a)** Example hidden layers of a network; **b)** when the activity of the feedforward neuron is above the threshold, the feedback signal is propagated to lower neurons and plasticity is enabled; **c)** otherwise, the feedback signal is not propagated to the lower layer and plasticity is hindered.

error-backpropagation for deep networks composed of multiple layers between the input and output layer in a reinforcement learning setting. Can the brain, with its many layers between input and output indeed solve the credit-assignment problem in a manner that is equivalent to deep learning? The second goal is to compare trial-and-error learning with biologically plausible learning rules to learning with error-backpropagation in more challenging problems. To this aim we investigated if and how the biologically learning rules cope with different datasets, namely MNIST, CIFAR10 and CIFAR100, trained as direct reward reinforcement learning tasks.

Biologically plausible deep reinforcement learning

We here generalize and extend AGREL to networks with multiple layers with two modifications of the previous learning schemes. Firstly, we use rectified linear (ReLU) functions as activation function of the neurons in the network. This simplifies the learning rule, because the derivative of the ReLU is equal to zero for negative activation values, and has a constant positive value for positive activation values. Note however that this can easily be generalized to other activation functions. Secondly, we assume that network nodes correspond to cortical columns with feedforward and feedback subnetworks: in the present implementation we use a feedforward neuron and a feedback neuron per node.

Overall, the network learning goes through five phases upon presentation of an input image: 1) the signal is propagated through the network by feedforward connections to obtain activations for the output units where the Q-values are computed; 2) in the output layer one output unit wins in a stochastic, competitive action selection process; 3) the selected output unit causes (attention-like) feedback to the feedback unit of each node (note that this feedback network propagates information about the selected action, just as in the brain, see e.g. (Roelfsema & Holtmaat, 2018), and that it does not need to propagate error signals, which would be biologically implausible); 4) a RPE δ is globally computed after the outcome of the action is evident; 5) the strengths of the synapses (both feedforward and feedback) are updated. The role of the feedback units is to gate the plasticity of feedforward connections (as well as their own plasticity, see Fig. 1). There is neuroscientific evidence for the gating of plasticity of feedforward connections by the activity of feedback connections, as was reviewed by (Roelfsema & Holtmaat, 2018).

In the opposite direction, the feedforward units gate the activity of the feedback units. Feedback gating is shaped by the local derivative of the activation function, which, for a unit with a ReLU activation function, corresponds to an all-or-nothing gating signal: for ReLU feedforward units, the associated feedback units of a node are only active if the feedforward units are activated above their threshold, otherwise the feedback units remain silent and they do not propagate the feedback signal to lower processing levels. Gating of the activity of feedback units by the activity of feedforward units is also in accordance with neurobiological findings: attentional feedback effects on the firing rate of sensory neurons are pronounced if the neurons are well driven by a stimulus and much weaker if they are not (Van Kerkoerle, Self, & Roelfsema, 2017; Roelfsema, 2006; Treue & Trujillo, 1999).

In general, for deep networks, updates of feedforward synapses $\Delta w_{p,m}$ from p-th neuron in the n-th hidden layer onto m-th feedforward neuron in the $(n+1)$-th hidden layer are computed as:

$$\Delta w_{p,m} = \alpha \delta y_p^{(n)} g_{(n+1)_m} fb_{y_m^{(n+1)}}, \quad (1)$$

and it is equal to the update of the corresponding feedback synapse $\Delta w'_{m,p}$, where the activity of the feedback unit is determined by the feedback signals coming from the $(n+2)$-th hidden layer as follows:

$$fb_{y_m^{(n+1)}} = \sum_q g_{(n+2)_q} v'_{q,m} fb_{y_q^{(n+2)}}, \quad (2)$$

with q indexing the units of the $(n+2)$-th hidden layer. The update of a synapse is thus expressed as the product of four factors: the RPE δ, the activity of the presynaptic unit, the activity of postsynaptic feedforward unit and the activity of feedback unit of the same postsynaptic node. Notably, all

Table 1: Results (averaged over 10 different seeds, the mean and standard deviation are indicated; in some cases - indicated with "*" - only 9 out of 10 seeds converged).

	Rule	1^{st} layer	Hidden units	α	Epochs [#]	Accuracy [%]
MNIST	Q-AGREL	Full	1500-1000-500	5e-01	130 (54)	98.33 (0.09)
	Error-BP	Full	1500-1000-500	1e-01	92 (11)	98.32 (0.04)
	Q-AGREL	Conv	21632-5408-500	1e+00	44 (10)	99.17 (0.05)
	Error-BP	Conv	21632-5408-500	1e-02	26 (12)	99.19 (0.10)
	Q-AGREL	LocCon	21632-5408-500	1e+00	83 (13)	99.04 (0.14)
	Error-BP	LocCon	21632-5408-500	1e-02	31 (10)	98.82 (0.20)
CIFAR10	Q-AGREL	Conv	28800-7200-1000-500	1e+00	115 (23)	73.54 (1.35)
	Error-BP	Conv	28800-7200-1000-500	1e-03	83 (21)	71.25 (1.08)
	Q-AGREL	LocCon	28800-7200-1000-500	1e+00	173 (36)	64.37 (2.41)
	Error-BP	LocCon	28800-7200-1000-500	1e-03	145 (16)	64.65 (1.16)
CIFAR100	Q-AGREL	Conv	28800-7200-1000-500	1e+00	230 (30)	34.90 (1.49)*
	Error-BP	Conv	28800-7200-1000-500	1e-03	104 (24)	36.79 (1.78)
	Q-AGREL	LocCon	28800-7200-1000-500	1e+00	343 (68)	29.39 (2.38)
	Error-BP	LocCon	28800-7200-1000-500	1e-03	156 (13)	32.73 (0.78)

the information necessary for the synaptic update is available locally, at the synapse. Moreover, simple inspection shows that the identical update for both feedforward and corresponding feedback synapses can be computed locally.

Experiments

We tested the performance of Q-AGREL on the MNIST, CIFAR10 and CIFAR100 datasets. The MNIST dataset consists of 60,000 training samples (images of 28 by 28 pixels), while the CIFAR datasets comprise 50,000 training samples (images of 32 by 32 by 3 pixels), of which 1,000 were randomly chosen for validation at the beginning of each experiment. We use a *batch gradient* to speed up the learning process (but the learning scheme also works with learning after each trial): 100 samples were given as an input, the gradients were calculated, divided by the batch size, and then the weights were updated, for each batch until the whole training dataset was processed (i.e. for 590 or 490 batches in total), indicating the end of an *epoch*. At the end of each epoch, a validation accuracy was calculated on the validation dataset. An early stopping criterion was implemented: if for 20 consecutive times the validation accuracy had not increased, learning was stopped.

We ran the same experiments with Q-AGREL and with error-backpropagation for neural networks with three and four hidden layers. The first layer could be either convolutional or locally connected, the second layer was convolutional but with a stride of 2 in both dimensions, to which a dropout of 0.8 (i.e. 80% of the neurons in the layer were silent) was applied, then either only one fully connected layer or two followed (with the last layer having a dropout rate of 0.3). At the level of the output layer (which had 10 neurons for MNIST and CIFAR10, while it was 10 times bigger for CIFAR100) for error-backpropagation a softmax was applied and a cross-entropy error function was calculated. We decided to test networks with locally connected layers because such an architecture could represent the biologically plausible implementation of convolutional layers in the brain (since shared weights are not plausible). Moreover, the convolutional layers with stride 2 were used instead of max pooling layers (which are not biologically plausible) to reduce the dimensionality of the layer following the convolutions, as described in (Springenberg, Dosovitskiy, Brox, & Riedmiller, 2014). As argued by Hinton (Hinton et al., 2016), dropout is biologically plausible as well: by removing random hidden units in each training run, it simulates the regularisation process carried out in the brain by noisy neurons.

In summary, we ran experiments with the architectures: `conv32 3x3` or `loccon32 3x3; conv32 3x3 stride2; drop.8; (full1,000;) full500; drop.3`, with 10 different seeds for synaptic weight initialization. All weights were randomly initialized within the range $[-0.02, 0.02]$ and the feedback synapses were identical to the feedforward synapses (strict reciprocity). For MNIST only we also performed a few experiments with fully connected networks, of which the weights were initialized in $[-0.05, 0.05]$.

Results

Table 1 presents the results of simulations with the different learning rules. We trained networks with only three hidden layers and networks with an extra hidden layer with 1000 units. We used 10 seeds for each network architecture and report the results as *mean (standard deviation)*. Our first result is that Q-AGREL reaches a relatively high classification accuracy of 99.17% on the MNIST task, obtaining essentially the same performance as standard error-backpropagation both with the architectures with convolutions and straightforward fully connected networks. The convergence rate of Q-AGREL was a factor of 1.5 to 2 slower than that of error-backpropagation for networks with convolutional layers, while it was a factor of 2.5 slower in networks for locally connected layers, but performing slightly better than error-backpropagation.

The results obtained from networks trained on the CIFAR10 dataset show that networks trained with Q-AGREL reached the same accuracy (if not higher) than with error-backpropagation. Additionally, the number of epochs required for the networks to

meet the convergence criterion was also comparable.

Table 1 also shows the results obtained from networks trained on CIFAR100. The final accuracy obtained with Q-AGREL was somewhat lower than with error-backpropagation. However, we still see that Q-AGREL is able to learn the CIFAR100 classification task with a convergence rate only 2 to 2.5 times slower than error-backpropagation and the rate for CIFAR10. These results shows that such trial-and-error learning rule can scale up to a 10 times higher number of classes with a penalty relatively small.

Discussion

We implemented a deep, biologically plausible reinforcement learning scheme called Q-AGREL and found that it was able to train networks to perform the MNIST, CIFAR10 and CIFAR100 tasks as direct reward problems with performance that was nearly identical to error-backpropagation. We also found that the trial-and-error nature of learning to classify with reinforcement learning incurred a very limited cost of 1-2.5x more training epochs to achieve the stopping criterion, even for classifying objects in 100 classes.

The results were obtained with relatively simple network architectures (i.e. not very deep) and learning rules (no optimizers or data augmentation methods were used). These additions would almost certainly further increase the final accuracy of the Q-AGREL learning scheme.

The present results demonstrate how deep learning can be implemented in a biologically plausible fashion in deeper networks and for tasks of higher complexity by using the combination of a global RPE and "attentional" feedback from the response selection stage to influence synaptic plasticity. Importantly, both factors are available locally, at many, if not all, relevant synapses in the brain (Roelfsema & Holtmaat, 2018). We demonstrated that Q-AGREL is equivalent to a version of error-backpropagation that only updates the value of the selected action. Q-AGREL was developed for feedforward networks and for classification tasks where feedback about the response is given immediately after the action is selected.

We find it encouraging that insights into the rules that govern plasticity in the brain are compatible with some of the more powerful methods for deep learning in artificial neural networks. These results hold promise for a genuine understanding of learning in the brain, with its many processing stages between sensory neurons and the motor neurons that ultimately control behavior.

Acknowledgments

I.P. was supported by NWO NAI grant 656.000.002, P.R.R. by HBP FP7 grant 7202070 and ERC grant 339490.

References

Amit, Y. (2018). Biologically plausible deep learning. *arXiv preprint arXiv:1812.07965*.

Brosch, T., Neumann, H., & Roelfsema, P. R. (2015). Reinforcement learning of linking and tracing contours in recurrent neural networks. *PLoS computational biology*, *11*(10).

Crick, F. (1989). The recent excitement about neural networks. *Nature*, *337*(6203).

Hinton, G., et al. (2016). Can the brain do back-propagation. In *Invited talk at stanford university colloquium on computer systems.* (`https://www.youtube.com/watch?v=VIRCybGgHts`)

Marblestone, A. H., Wayne, G., & Kording, K. P. (2016). Toward an integration of deep learning and neuroscience. *Frontiers in computational neuroscience*, *10*.

Pooresmaeili, A., Poort, J., & Roelfsema, P. R. (2014). Simultaneous selection by object-based attention in visual and frontal cortex. *Proceedings of the National Academy of Sciences*.

Richards, B. A., & Lillicrap, T. P. (2019). Dendritic solutions to the credit assignment problem. *Current opinion in neurobiology*, *54*.

Roelfsema, P. R. (2006). Cortical algorithms for perceptual grouping. *Annu. Rev. Neurosci.*, *29*.

Roelfsema, P. R., & Holtmaat, A. (2018). Control of synaptic plasticity in deep cortical networks. *Nature Reviews Neuroscience*, *19*(3).

Roelfsema, P. R., & Ooyen, A. v. (2005). Attention-gated reinforcement learning of internal representations for classification. *Neural computation*, *17*(10).

Rombouts, J. O., Bohte, S. M., & Roelfsema, P. R. (2015). How attention can create synaptic tags for the learning of working memories in sequential tasks. *PLoS computational biology*, *11*(3).

Sacramento, J., Costa, R. P., Bengio, Y., & Senn, W. (2018). Dendritic cortical microcircuits approximate the backpropagation algorithm. In *Advances in neural information processing systems.*

Scellier, B., & Bengio, Y. (2019). Equivalence of equilibrium propagation and recurrent backpropagation. *Neural computation*, *31*(2).

Schiess, M., Urbanczik, R., & Senn, W. (2016). Somatodendritic synaptic plasticity and error-backpropagation in active dendrites. *PLoS computational biology*, *12*(2).

Scholte, H. S., Losch, M. M., Ramakrishnan, K., de Haan, E. H., & Bohte, S. M. (2017). Visual pathways from the perspective of cost functions and multi-task deep neural networks. *Cortex*.

Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron*, *36*(2).

Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*.

Treue, S., & Trujillo, J. C. M. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, *399*(6736).

Urbanczik, R., & Senn, W. (2014). Learning by the dendritic prediction of somatic spiking. *Neuron*, *81*(3).

Van Kerkoerle, T., Self, M., & Roelfsema, P. (2017). Effects of attention and working memory in the different layers of monkey primary visual cortex. *Nat. Commun*, *8*.