Models of Neural Systems (WiSe 2018/2019) Final Project - Computer Practical Biologically Plausible Deep Learning

Robert T. Lange *
Einstein Center for Neurosciences Berlin
robert.lange17@imperial.ac.uk
www.rob-lange.com

Abstract

Backpropagation provides a biologically implausible solution to the synaptic credit assignment problem in Deep Learning. Computational graphs and the chain rule successfully provide approximate gradients in deep layered structures. But the empirical success does not necessarily imply that the brain is capable of implementing such a procedure. In this report we review different proposed solutions to such problems that render backpropagation biologically implausible. More specifically, we focus on an approach which implements local plasticity rules in a neural network architecture with dendritic compartments [2]. Previously it has been argued that such an architecture overcomes multiple points of critique while accomplishing similar strong results. Our robustness checks reveal that such a claim is not justified.

1 Introduction

Deep Learning (DL) has been the poster child of Machine Learning success in the 21st century. It has dominated competitions and research across all domains (computer vision, natural language processing, robotics as well as computational neuroscience).

*This progress report was submitted as part of the final project of the "Models of Neural Systems" computer practical course.

2 Credit Assignment in Deep Layered Structures

Backpropagation: A Successful Deep Learning Perspective

Backpropagation: A Critical Neuroscience Perspective

ADD FIGURE WITH KEY POINTS OF CRITIQUE

Synaptic Integration via Compartmental Dendrites

3 Literature Review

ADD FIGURE WHICH COMPARES ALL APPROACHES - TREE OR TABLE

Feedback Alignment - Lillicrap et al. [4]

General Content: Introduce a first feedback alignment approach to solve the weight transport problem of backpropagation. Forward and backward weights are modeled separately - backward weights align with weight matrix transpose through learning process. Argument follows from positive definiteness of weight and random matrix product and a rotation line of thought.

Keypoints:

* Weight transport Problem: downstream errors are fed back to upstream neurons via exact symmetric copy of downstream synaptic weight matrix - neuron "deep" within network has to have precise knowledge of all downstream synapses!

* Possible solutions: 1. Retrograde transmission of info along axons - problem of slow timescale 2. Feedback of errors via second network - problem of symmetry assumption of feedforward and feedback connections 3. Here: Show that even fixed random connections can allow for learning - symmetry not required! Instead implicit dynamics lead to soft alignment between forward and backward weights

* Observations: * Feedback weights does not have to be exact: $B \approx W^T$ with $e^T W B e > 0$. rotation within 90 degrees of backprop signal. Learning speed depends on degree! * Alignment of B and W^T via adjustment of W (and B) possible

* Feedback alignment: * Modulator signal (error-FA) does not impact forward pass post-synaptic activity bu acts to alter plasticity at the forward synapses. * FA may encourage W to align with Moore-Penrose pseudoinverse of B - approximate functional symmetry * Inference vs learning - towards bayesian approaches

* Experiments: * Learns linear function with single hidden layer - learning not slower than backprop * Sigmoid nonlinearity and classification task - altered function of post-synaptic activity - learned also to communicate info when 50* More layers 3 hidden layers - as well as backprop and making use of depth - froze layers and trained alternatingly - positive/negative phase? * Neurons that integrate activity over time and spike stochastically - synchronous pathways

* Possible Extensions: * Fixed spike thresholds/refractory period * Dropout/stochasticity

Questions:

* Still signed error signal has to be transferred which remains illusive - see target propagation. * Is result related to Johnson-Lindenstrauss concentration ineq ideas? * Usage of intricate/more complex architectures of communication of backward error - relation to multi-agent RL * Relationship to predictive coding

Target Propagation - Lee et al. [3], Bartunov et al. [1]

Bartunov et al. [1] - Simplified Difference

General Content: Extent the target propagation algorithm to not require exact gradient at penultimate layer. Test alternative learning rules in more complicated settings (CIFAR/ImageNet) and differentiate between locally and fully connected architectures. Very good review but not much additional innovation. Behavioral + Physiological Realism

Keypoints:

* Problems with backpropagation * Feedback connections require exact copy of feedforward connections = Weight transport * Info propagation does not influence "neural activity" - does not conform to any known biological mechanism

* Feedback alginment: Use random weights in backward pass to deliver info to earlier layers * Still requires delivery if signed error via distinct pathway * Direct/Broadcast FA - connect feedback from output layer directly to all previous ones

* Contrastive Hebbian Learning/Generalized Recirculation: Use top-down feedback connections to influence neural activity and differences to locally approx gradients * Positive/negative phase - need settling process - Likely to slow for brain to compute in real time

* Target Propagation: Trains distinct set of feedback connections defining backward activity propagation * Connections trained to approximately invert feedforward connections to compute target activites for each each layer by successive inversion - decoders * Reconstruction + Forward loss * Different target constructions * Vanilla TP: Target computation via propagation from higher layers' targets backwards through layer-wise inverses * Difference TP: Standard delta rule with additional stabilization from prev reconstruction error. Still needs explicit grad comp at final layer * Not tested on data more complex than MNIST

* Simplified Difference Target Propagation: Computation also for penultimate layer with help of correct label distribution - removes implausible gradient communication * Need diversity in targets - problem of low entropy of classification targets * Need precision in targets - poor inverse learned * Combat both problems/weakness of targets with help of auxiliary output resembling random features from penultilmate hidden layer * Parallel vs alternating inverse training - simultaneous more plausible

* Weight-Sharing is not plausible - regularizes by reducing number of free parameters

* Experiments - Mostly negative results: 1. None of existing algos is able to scale up - Good performance MNIST/Somewhat reasonable on CIFAR/Horrible on ImageNet - Seems like weight-sharing is not key to success 2. Need for behavioral realism - judged by performance on difficult tasks 3. Hyperparameter Sensitivity * First fix "good" architeture and then optimize * Use hyperbolic tanh instead of ReLu - work better

Questions:

* How could the brain do weight sharing - is approximate again satisfactory/functional approx? * Think more about communication: MARL agents learning communication channels

Local Synaptic Rules with Dendritic Integration

Guerguiev et al. [2] - A Plausible Alternative?

Sacramento et al. [5] - Dendritic Microcircuits

General Content: MLP with simplified dendritic compartments learned in local PE plasticity fashion. No separate phases needed. Errors represent mismatch between pre input from lateral interneurons and top-down feedback. First cortical microcircuit approach. Analytically derive that such a setup/learning rule approximates backprop weight updates and proof basic performance on MNIST.

Keypoints:

* Hypothesis: Pred errors are encoded at distal dendrites of pyramidal neurons - receive input from downstream neurons - in model: error arise from mismatch of lateral local interneuron inputs (SST - somatostatin) - Learning via local plasticity

* 3 Compartment Neuron: * Soma + Integration zones: Basal/Apical - convergence of top-down/bottum-up synapses on different compartments - Larkum (2013): Preferred connectivity patterns of cortico-cortical projections

* 2nd Population within hidden layer - Interneurons = lateral + cross-layer connectivity: cancel t-d input - only backprop errors remain as apical dendrite activity * Predominantly driven by same layer but cross-layer feedback provides weak nudge for interneurons = modeled as conduc-based somatic input current * Modeled as one-to-one between layer interneuron and corresponding upper-layer neuron * Empirically justified by monosynaptic input mapping experiments: weak interneuron teaching signal

* Neuron/network Model: - Simplifications: 1. Membrane capacity to 1 and resting potential 0; Background activity is white noise 2. Modeling of layer dynamics - where vectors represent units 3. No apical compartment in pyramidal output neurons - 3 compartments seem to suffice as comparison mechanism - Qualitative dynamics: error = apical voltage deflection -> propagates down soma -> modulates somatic firing rate -> plasticity at bottom-up synapses - Somatic conductance acts as nudging conductance - Lateral dendritic projections: interneuron is nudged to follow corresponding next layer pyramidal neuron

* Synaptic learning rules = Dendritic Predictive Plasticity Rules - Originally: reduction of somatic spiking error - conductance based normalization of lateral projections based on dendritic attenuation factors of different compartments - Implementation requires subdivision of apical compartment into two distal parts (t-d input and lateral input from interneurons)

* Prev work: * Guergiev: View apical dendrites as integration zones - temp difference between activity of apical dendrite in presence/absence of teaching input = error inducing plasticity at forward synapses. Used directly for learning b-u synapses without influencing somatic activity. HERE: apical dendrite has explicit error represnentation by sim integration of t-d excitation and lateral inhibition - No need for separate temporal phases - continuous operation with plasticity always turned on * PC based work - Whittington and Bogacz: Only plastic synapses are those connecting prediction and error neurons. HERE: all connections plastic - errors are directly encoded in dendritic compartments

* Main Results/Experiments: * Analytic derivation: Somatic MP at layer k integrate feedforward predictions (basal dendritic potentials) and backprop errors (apical dendritic potentials) * Analytic derivation: Plasticity rule converges to backprop weight change with weak feedback limit * Random/Fixed t-d weights = FA * Learned t-d weights minimizing inverse reconstruction loss = TP * Experiments: * Non-Linear regression task: Use soft rectifying nonlinearity as transfer fct - Tons of hyperparameters - injected noise current (dropout/regularization effect?) * MNIST - Deeper architectures: Use convex combination of learning/nudging

* General notes: * Kriegeskorte/DiCarlo/RSA - DNNs outperform alternative frameworks in accurately reproducing activity patterns in cortex - What does this mean? Is DL just extremely flexible/expressive? * bottom-up = feedforward, top-down = feedback

Questions:

* Neural transfer fct = Activation fct! * Again tons of hyperparameters to be chosen - How? * Think of learning (accurate gradient approx) vs architecture (depth, number of hyperparameters) complexity * Different interneuron types (PV = parvalbumin-positive) - different types of errors (generative)

4 Empirical Investigations

Scalability Across Datasets

Learning Dynamics

Hyperparameter Robustness

5 Outlook and Conclusion

In this report we have empirically investigated the robustness and learning dynamics of an alternative learning rule in deep layered structures. We first reviewed and formalized the classical backpropagation algorithm. Afterwards, we put on computational neuroscience googles and highlighted several short-comings such as the weight transport problem as well the necessity to propagate signed errors. In Section 3 of this report we then introduced the methodology outlined by Guerguiev et al. [2] which intends to overcome such limitations. Inspired by dendritic compartments and information integration at different sites, the algorithm solves the weight transport problem. In Section 4 we reviewed more current approaches and compared their benefits and limitations. Thereby, we highlight the difference between behavioral and neurophysiological realism. Furthermore, we discuss the differences between learning and architecture complexity across the different approaches. Afterwards, we implement the approach by Guerguiev et al. [2] and compare model selection as well as hyperparameter robustness across different popular datasets. Our experiments reveal major performance decreases. This brings up the following question: Why should the brain implement a suboptimal **and** non-robust learning rule on a neurophysiological level? A simple answer to this is the flexibility that such an alternative architecture comes with.

References

[1] BARTUNOV, S., A. SANTORO, B. RICHARDS, L. MARRIS, G. E. HINTON, AND T. LILLICRAP (2018): "Assessing the scalability of biologically-motivated deep learning algorithms and architectures," in *Advances in Neural Information Processing Systems*, 9389–9399.

[2] GUERGUIEV, J., T. P. LILLICRAP, AND B. A. RICHARDS (2017): "Towards deep learning with segregated dendrites," *ELife*, 6, e22901.

[3] LEE, D., S. ZHANG, A. BIARD, AND Y. BENGIO (2014): "Target Propagation," *CoRR*, abs/1412.7525.

[4] LILLICRAP, T. P., D. COWNDEN, D. B. TWEED, AND C. J. AKERMAN (2016): "Random synaptic feedback weights support error backpropagation for deep learning," *Nature communications*, 7, 13276.

[5] SACRAMENTO, J., R. P. COSTA, Y. BENGIO, AND W. SENN (2018): "Dendritic cortical microcircuits approximate the backpropagation algorithm," in *Advances in Neural Information Processing Systems*, 8735–8746.

Supplementary Material

6 Biological Plausible Deep Learning

6.1 Author: Robert Tjarko Lange | December 2018

This project analyzes different learning rules in deep layered structures. More specifically, we explore alternatives to backpropagation (aka the chain rule). Weight transport (access to all weights at every layer of the backward pass) renders backpropagation biologically implausible. Recent alternatives explore local learning rules and draw inspiration from the compartmental design of pyramidal neurons.

6.2 DONE:

- [x] PyTorch MLP/CNN baseline for MNIST
- [x] Create remote repo
- [x] Generalize network architecture to variable inputs
- [x] Write update_logger, process_logger function
- [x] Plot learning curves - output from logger
- [x] Add Xavier init for networks
- [x] Rewrite architecture and simplify code
- [x] Tried running in colab
- [x] Set up bayesian optimization pipeline - BayesianOptimization
  - [x] implement cross-validation with torch data/skorch
  - [x] one fct taking in hyperparams, return objective
  - [x] write fct that transforms cont variables to discrete
  - [x] check how to add folds/add input to eval_nn, BO pipeline
  - [x] Generalize BO pipeline to CNN
  - [x] Write fct that checks if BO CNN proposal is valid (kernel/in/out)
  - [x] Add logging to BO pipeline

6.3 TODO - CODING:

- [ ] get_data - Different datasets - FashionMNIST, CIFAR 10
- [ ] Evaluate the model more frequently - not only once per epoch
- [ ] Record weight changes
- [ ] Get Guergiev Code running/understand
- [ ] Restructure Guergiev code and integrate into current pipeline
- [ ] Add comments! - Look up pep8 standard for fcts/classes
- [ ] Work on weight visualization/changes in weights!
- [ ] Work on error propagation comparison/delta W (||W_t - W_t-1||/||W_t||)
- [ ] Runs Bayesian Opt for 10 Epochs and 50 evaluations/BO iterations for 3 datasets
- [ ] Get best/worst performance, standard dev - plot as bar chart across approaches DNN/CNN/Guergiev

6.4 TODO - REPORT:

- [ ] Read papers/Add first notes of papers
  - [x] Lillicrap et al (2016)
  - [ ] Guergiev et al (2017)
  - [x] Bartunov et al (2018)
  - [x] Sacramento et al (2018)
  - [ ] Larkum (2013)
- [ ] Add first skeleton of report/sections - max 10 pages
  - [ ] Backprop/Notation
  - [ ] Literature Notes
- [ ] Overview figures (Problems with backprop, Solution approaches)

6.5 Repository Structure

```
Bio-Plausible-DeepLearning
+- workspace.ipynb: Main workspace notebook - Execute for replication
```

6.6 How to use this code

1. Clone the repo.

```
git clone https://github.com/RobertTLange/Bio-Plausible-DeepLearning
cd Bio-Plausible-DeepLearning
```

2. Create a virtual environment (optional but recommended).

```
virtualenv -p python BPDL
```

Activate the env (the following command works on Linux, other operating systems might differ):

```
source BPDL/bin/activate
```

3. Install all dependencies:

```
pip install -r requirements.txt
```

4. Run the main notebook:

```
jupyter notebook workspace.ipynb
```