Learning Stable Deep Predictive Coding Networks with Weight Norm Supervision

Ruohao Guo
University of Illinois Urbana-Champaign
ruohaog2@illinois.eduUniversity of Illinois Urbana-Champaign

ILLINOIS
UNIVERISTY OF ILLINOIS AT URBANA-CHAMPAIGN

Introduction

Datasets: MNIST, CIFAR-10, SVHN, 3D Chairs, CelebA .

Motivation: Predictive Coding Network (PCN) is a recurrent neural network inspired from neuroscience. However, training a PCN in finite-time with backpropagation through time (BPTT) can only approximate how a real PCN works. To further explore its potential, it is necessary to study its dynamics in infinite-time evolution. Moreover, the number of recurrent cells needed in PCN is big, and this long-term dependency could lead to vanishing and exploding gradient problems in BPTT training.

Theories

1. A sufficient condition for stable hierarchical RNNs.

Lemma 1. Let a hierarchical RNN consist of a stack of layers and be linearizd around some point, where at time step $t+1$, each layer receives input from the output of lower layers at step t, the feedback of the closest higer layer and the recurrent input of itself. For any $\varepsilon \in (0,1)$, there exists arbitrary $\delta > 0$ with finite value. The sufficient condition of ESP for the network is given by:

$$\delta < \frac{1 - \max_i \left(\rho\left(\frac{\partial \mathbf{S}^i(t+1)}{\partial \mathbf{S}^i(t)} \right) \right) - \varepsilon}{\sqrt{n} \max_i \left(\| \frac{\partial \mathbf{S}^i(t+1)}{\partial \mathbf{S}^{i+1}(t)} \|_\infty \right)} \tag{7}$$

where n is the dimension of global state space, $\mathbf{S}^i(t)$ is the state vector variable in layer i at step t, $\rho(\cdot)$ is the spectral radius, $||\cdot||_\infty$ is infinite norm. δ is the parameter of D_δ-Norm [4], which is finite value dependent of ε.

2. A sufficient condition for stable PCNs.

Theorem 2. Consider a PCN whose dynamics are defined in Section III-A. Its stability of dynamics is studied by investigating the approximation of the local linearized dynamic system in Eq. 6. Then, a sufficient condition for the stability of PCN dynamics around the linearizing point is given by Eq. 7, where

$$\frac{\partial \mathbf{S}^i(t+1)}{\partial \mathbf{S}^{i+1}(t)} = \theta' \cdot \mathbf{W}^i_{repfb} \mathbf{W}^{i+1}_{fb} \tag{19}$$

$$\frac{\partial \mathbf{S}^i(t+1)}{\partial \mathbf{S}^i(t)} = \theta' \cdot \mathbf{W}^i_{reps} - \theta' \cdot \mathbf{W}^i_{repi} \cdot \phi' \cdot \mathbf{W}^i_{pred} \tag{20}$$

θ is the activation of representation, while ϕ is the activation of prediction. θ' and ϕ' are diagonal matrices whose diagonal blocks are activation derivative at the linearizing point of each layer.

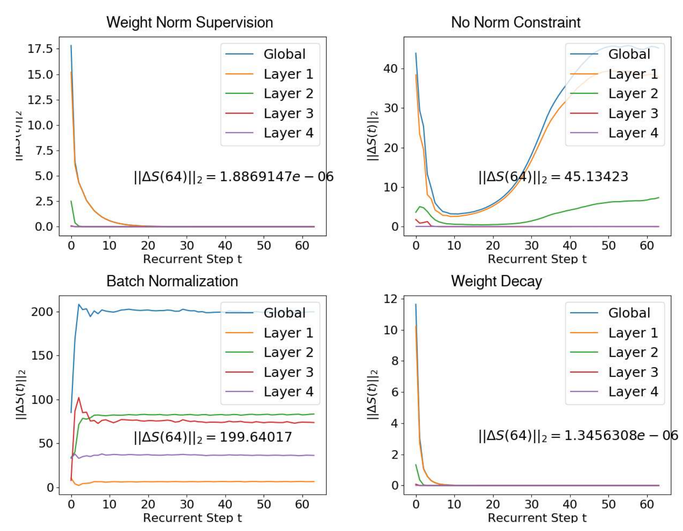
Method

Our study shows the global stability of a PCN is determined by stability of the local layers and the feedback between neighboring layers. Based on it, we further propose Weight Norm Supervision method to control the stability of PCN dynamics by imposing different weight norm constraints on different parts of the network.

Experiments & Results

Compared methods: weight norm supervision, no norm constraint, batch normalization, weight decay.

1. Metrics: State asymptotic.
Results: our weight norm supervision method achieved the best performance on stability. Please see the results on all the datasets in appendix.



2. Metrics: mean squared prediction errors in test set.
Results: weight norm supervision has the smallest prediction errors over 5 datasets, except that on CIFAR-10, weight decay performs a little better.

TABLE I: Mean Squared Prediction Errors in Test Set After 8 Cycles

	Weight Norm Supervision	No Norm Constraints	Batch Normalization	Weight Decay
MNIST	**0.04268**	0.14898	0.165877	0.068254
CIFAR-10	0.003129	0.023687	0.036623	**0.000522**
SVHN	0.000823	0.0027	0.087307	**0.000375**
3D Chairs	**0.000225**	0.000577	0.075244	0.000904
CelebA	**0.000631**	0.002483	0.068673	0.001025

TABLE II: Mean Squared Prediction Errors in Test Set After 64 Cycles

	Weight Norm Supervision	No Norm Constraints	Batch Normalization	Weight Decay
MNIST	**0.04293**	0.128875	0.120328	0.068254
CIFAR-10	0.002946	0.109168	0.03507	**0.000549**
SVHN	**0.001343**	0.002701	0.090878	0.214557
3D Chairs	**0.000214**	0.180429	11102707.0	0.02755
CelebA	**0.000402**	0.001173	0.106447	0.275487