Towards a Cortically Inspired Deep Learning Model: Semi-Supervised Learning, Divisive Normalization, and Synaptic Pruning

Tan Nguyen[1,2] Wanjia Liu[1] Fabian Sinz[2] Richard G. Baraniuk[1] Andreas S. Tolias[2] Xaq Pitkow[1,2] Ankit B. Patel[1,2]
[1]Rice University [2]Baylor College of Medicine
6100 Main Street, Houston, TX 77005 1 Baylor Plaza, Houston, TX 77030
{mn15, wl22, richb, xaq}@rice.edu {sinz, astolias, ankitp}@bcm.edu

Abstract

Deep learning has driven dramatic advances in performance on a wide range of difficult machine perception tasks, and its applications abound. Yet for many tasks it still lags far behind the mammalian brain in term of performance and efficiency in natural tasks. Building a brain-inspired learning system to narrow the gap between artificial and biological neural networks has been a long sought-after goal in both the neuroscience and machine learning communities. To take a step towards a neurally plausible learning system, we build a class of models that use functional elements and computational principles of the cortex for more robust and versatile machine learning. In particular, we incorporate the following three major neural features into the Deep Convolutional Networks (DCNs): semi-supervised learning, divisive normalization, and synaptic pruning. These neural features are derived from a recently developed generative model underlying DCNs - the Deep Rendering Mixture Model (DRMM). Our semi-supervised learning algorithm achieves state-of-the-art performance on the MNIST and SVHN datasets and competitive results on CIFAR10 amongst all methods that do not use data augmentation. Our divisive normalization enables faster and more stable training. Using our synaptic pruning method, we can compress the model significantly with little loss in accuracy.

Keywords: deep learning; semi-supervised learning; divisive normalization; synaptic pruning

Cortically Inspired Model

Deep Rendering Mixture Model

A fundamental hypothesis of our work is that deep neural networks in the brain are performing probabilistic inference with respect to a generative probabilistic model of the world (Lochmann & Deneve, 2011). But how can we link artificial networks to generative models? The Deep Rendering Mixture Model (DRMM) is a recent effort to *reverse-engineer* several classes of artificial networks, including the Deep Convolutional Networks (DCNs) (see Figure 1). The DRMM is a hierarchical generative model in which the image is generated iteratively in a coarse-to-fine manner. It has been shown that the bottom-up inference in the DRMM (after a discriminative relaxation), corresponds to the feedforward propagation in the DCNs (Patel, Nguyen, & Baraniuk, 2016). The DRMM allows us to subsume semi-supervised learning, divisive normalization, and synaptic pruning with DCNs in one theoretical framework.

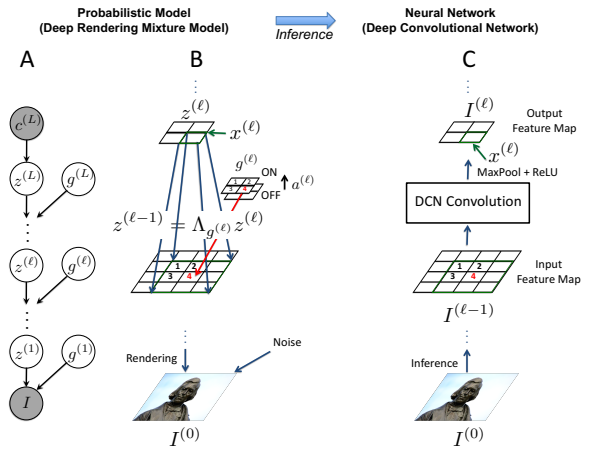

Figure 1: (A)The DRMM (B) Rendering from layer $\ell \to \ell - 1$ in the DRMM. (C) Bottom-up inference in the DRMM leads to processing identical to the DCNs.

Neural Features for Deep Convolutional Networks

Semi-supervised Learning Current state-of-the-art machine learning algorithms are trained on enormous sets of labeled examples presented in batches to the learning algorithm. In contrast to that, humans and animals learn from few instances of only weakly labeled examples. It is widely believed that the brain is building a rich model for the data in an un- or weakly supervised way such that novel categories can be constructed quickly. We develop a semi-supervised learning algorithm that allows DCNs to learn from both labeled and unlabeled data using the DRMM and variational inference (VI). In particular, our method performs bottom-up and top-down inference in the DRMM and then optimizes the loss function $\mathcal{L} \equiv \alpha_{CE}\mathcal{L}_{CE} + \alpha_{RC}\mathcal{L}_{RC} + \alpha_{KL}\mathcal{L}_{KL} + \alpha_{NN}\mathcal{L}_{NN}$ where α_{CE}, α_{RC}, α_{KL} and α_{NN} are the weights for the cross-entropy loss $\mathcal{L}_{CE}$, reconstruction loss $\mathcal{L}_{RC}$, variational inference loss $\mathcal{L}_{KL}$, and the non-negativity penalty loss $\mathcal{L}_{NN}$, respectively. The non-negativity penalty loss $\mathcal{L}_{NN}$ results from the assumption that the intermediate rendered templates in the DRMM are non-negative.

Divisive Normalization Divisive normalization is a phenomenological model describing the response behavior of

Table 1: Test error for semi-supervised learning on MNIST with $N_U = 60K$ unlabeled & $N_L \in \{50, 100\}$ labeled images.

Model	Test error (%)	
	$N_L = 50$	$N_L = 100$
catGAN	-	1.39 ± 0.28
Skip DGM	-	1.32
LadderNetwork	-	1.06 ± 0.37
Auxiliary DGM	-	0.96
ImprovedGAN	2.21 ± 1.36	0.93 ± 0.065
DRMM 5-layer Supervised	-	22.98
DRMM 5-layer + VI	2.46	1.36
DRMM 5-layer + VI + $\mathcal{L}_{NN}$	**0.91**	**0.57**

populations of neurons to changes in the contrast of the signal (Heeger, 1992). Recently, a particular form of divisive normalization has been used to normalize the activations in the DCNs and shown promising results (Ren, Liao, Urtasun, Sinz, & Zemel, 2016). From the DRMM inference standpoint, divisive normalization can be interpreted as inferring a latent variable in a Gaussian scale mixture (GSM) (Schwartz & Simoncelli, 2001).

Intuitively, the Gaussian latent variable captures the pattern of an image (patch) while the scale variable describes the local contrast. For an inverse gamma scale distribution, the maximum *a posteriori* estimator (MAP) for the pattern given an image patch takes the form of divisive normalization (Lyu, 2011). In the context of the DCNs, the MAP is realized by normalizing the outputs of convolution operations at each layer with divisive normalization.

Pruning In its early development, the brain prunes synapses and neurons at a rapid rate. Some types of pruning are thought to follow a "use-it-or-lose-it" (UILI) rule: synapses that are not used regularly are pruned away (Allred, Kim, & Jones, 2014). Inspired by this pruning rule, we derive a novel UILI synaptic pruning algorithm in the DRMM framework. Particularly, we place a mixing parameter π^{ℓ}_{xy} on the presence or absence of a given weight $\lambda^{\ell}_{xy} \equiv (\Lambda_{g^{\ell}})_{xy}$. This hyper-prior on the weights controls *the probability* that a given weight is present or not. During learning, we will update our estimates of these weight presence parameters and then apply a statistical hypothesis test to decide whether the synapse should be there (using a threshold $\alpha_S > 0$). We use pruning rate schedules inspired by the developing rat cortex (Navlakha, Barth, & Bar-Joseph, 2015).

Experimental Results

Table 1 shows the test errors of our semi-supervised learning method on MNIST using all available unlabeled data and different amounts of labeled data. Our algorithm can learn more from fewer labeled examples, achieving state-of-the-art performance in semi-supervised learning in different setups on MNIST (see (Nguyen, Patel, & Baraniuk, 2017) for results on SVHN and CIFAR10). Note that all methods use the same baseline architecture.

Figure 2 compares the performance of divisive normalization (DN) and batch normalization (BN) on a shape classification task in the semi-supervised learning setup and on an object classification task in an online learning setup. The networks are trained using a synthetic dataset containing 110K rendered images of natural objects with different textures and nuisance configurations. The images are rendered from 1,085 shape models and 55 object models in the ShapeNet library (Chang et al., 2015). Overall, we find that DN converges as fast or faster, and yields more stable learning curves than BN. Importantly, we find that DN has an especially large advantage in the online learning setting. This is because, in contrast to BN, which requires a batch of images to compute a normalization, DN works with a single image and can thus be used in the online setting where one input is processed at a time.

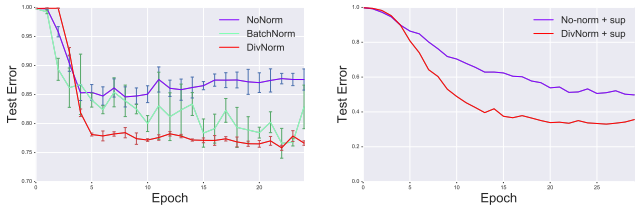

Figure 2: Comparison between the performance of divisive normalization and batch normalization for (left) the semi-supervised learning task on shape classification using 10% labeled data and (right) the online supervised learning task on object classification. Both are trained on our synthetic dataset.

To evaluate our synaptic pruning method, we apply synaptic pruning while training a 9-layer DRMM using our semi-supervised learning algorithm with divisive normalization for object classification task using the synthetic dataset described above. The model is prunned by more than 60% while still achieving good test accuracy (85%).

References

Allred, R. P., Kim, S. Y., & Jones, T. A. (2014). Use it and/or lose itexperience effects on brain remodeling across time after stroke. *Frontiers in Human Neuroscience*.

Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., ... others (2015). Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*.

Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual neuroscience*, *9*(02), 181–197.

Lochmann, T., & Deneve, S. (2011). Neural processing as causal inference. *Current opinion in neurobiology*, *21*(5), 774–781.

Lyu, S. (2011). Dependency reduction with divisive normalization: Justification and effectiveness. *Neural computation*, *23*(11), 2942–2973.

Navlakha, S., Barth, A. L., & Bar-Joseph, Z. (2015). Decreasing-rate pruning optimizes the construction of efficient and robust distributed networks. *PLoS Comput Biol*, *11*(7), e1004347.

Nguyen, T., Patel, A. B., & Baraniuk, R. G. (2017). Semi-supervised learning with deep rendering mixture model. *ICCV (Submitted)*.

Patel, A. B., Nguyen, T., & Baraniuk, R. G. (2016). A probabilistic framework for deep learning. *NIPS*.

Ren, M., Liao, R., Urtasun, R., Sinz, F. H., & Zemel, R. S. (2016). Normalizing the normalizers: Comparing and extending network normalization schemes. *ICLR*.

Schwartz, O., & Simoncelli, E. P. (2001). Natural signal statistics and sensory gain control. *Nat Neurosci*, *4*(8), 819–825. doi: 10.1038/90526