NVIDIA Academic Hardware Grant Request

Sethna Group - Cornell University Physics Department

Renormalization group, Restricted Boltzmann Machines and Spin Glasses:

The success of neural networks and Machine Learning still does not have a satisfactory theoretical explanation. The Renormalization Group of statistical physics is a framework for understanding physical systems by systematically ignoring (marginalizing) unimportant degrees of freedom and quantifying important degrees of freedom. Recent work by Mehta and Schwab [1] has introduced a connection between Restricted Boltzmann Machines (RBM)—the statistical models which spawned modern Artificial Neural Networks—and the Renormalization Group. We would like to explore this connection by training convolution RBMs on data generated from Ising models, a canonical model of magnetism in statistical physics. A nice property of the graph connections in an RBM is that it is easy to analytically marginalize over the bottom or top layer, producing an effective Hamiltonian for the data on which the RBM was trained. We will use this property to study the flows of model parameters as a function of layer depth, to glean the critical exponents which fully characterize the phase transition of a 2D Ising model. Training RBMs is a natural candidate for parallelization, so much so that they form a basis for training Deep Neural Networks on GPUs. Our current CPU solution takes hours to train RBMs on modestly sized Ising model data, and an NVIDIA Tesla K40 would allow us to train 10-100x faster. The double precision available on this device would also be very useful, as we will be maximizing likelihoods to fit models to data, and these numbers span many decades.

In addition to spawning new work, a new GPU would also accelerate current projects. We have been studying effective Hamiltonians of 2D spin glass clusters[2], where we use clustering methods and an inverse Ising algorithm from Machine Learning to find simpler models that preserve the correlations and thermodynamic behavior. We believe that the marginalized RBMs are a generalization of this approach, and we believe the development of the technology mentioned above will lead to a more beautiful and general approach to renormalizing systems that lack translation invariance.

---

[1] http://arxiv.org/abs/1410.3831

[2] Presented at APS March Meeting (http://meeting.aps.org/Meeting/MAR15/Session/B50.4)

Big data problems:
Sloppiness in the training and use of Neural Networks

Manifolds defined by nonlinear models with many parameters often admit lower-dimensional descriptions, wherein the relevant parameters are a set of linear combinations of naive model parameters. These models are called *sloppy*, since most parameter combinations have almost no influence on model predictions. Biological networks, particle accelerators, the Ising model, and neural networks have all been found to be sloppy[3]. Sloppiness has a couple complementary consequences: most parameter 'directions' are unimportant and so the model gives robust predictions, however optimizing such models (as in fitting to data) is difficult due to ill-conditioned Hessians. We want to study how these two properties might be used to enhance the trainability and robustness of neural networks. Recent work[4] has found deep neural networks to be incredibly susceptible to small noise perturbations at their input. We would explain this by saying that the input-output map of the network is *not sloppy*. As a result we want to understand how to make a neural network sloppy at its inputs so that we can get robust classification or feature recognition, while at the same time minimizing the sloppiness of the model weights and biases to enable fast training.

We have trained small feed-forward NNs on the MNIST handwritten digit classification problem, with $N_{parameters} \approx 5 \times 10^5$ parameters. In order to study the training sloppiness, we have to look at the singular value decomposition (SVD) of the Jacobian which has dimensions $(N_{parameters} \times N_{data})$. This is not a sparse matrix, and since $N_{data} = 5 \times 10^4$ it would require ≈ 200 GBs of memory, therefore we have implemented a streaming randomized SVD, which takes ≈ 10 hours and returns only a few hundred singular values. Since it was designed with big data applications in mind, the NVIDIA Tesla K40 would dramatically improve both the speed at which we could study these large dimensional models, and allow us to study even larger models.

---

[3] http://sethna.lassp.cornell.edu/research/why_is_science_possible

[4] http://arxiv.org/abs/1412.6572