A coarse graining perspective of deep learning

Ellen de Mello Koch
Supervisor: Prof. Ling Cheng

Doctor of Philosophy

School of Electrical and Information Engineering
Faculty of Engineering and the Built Environment
University of the Witwatersrand

Abstract

Deep learning has shown remarkable success at a number of tasks in various applications. This success is well appreciated in practice but theoretical explanations for how deep learning works are in their infancy. Previous work suggests that a link may exist between deep learning and the renormalization group (RG). RG is a powerful tool with a strong theoretical foundation. It allows one to move from an enormous set of microscopic parameters which describe a system to a succinct macroscopic description. RG achieves this by repeatedly performing a local coarse graining. Deep learning can also be viewed as performing a form of averaging or coarse graining. Deep networks reduce a large number of inputs to a smaller number of outputs which summarize the input training data. This suggests that a link between RG and unsupervised deep learning may exist. Such a link would give a possible starting point from which to develop a theoretical framework for deep learning by exploiting the rich theory of RG. This thesis is focussed on establishing this link in a quantitative way.

The studies presented here are carried out on models of a magnet, namely the Ising model and long range spin chain, as well as the MNIST (Modified National Institute of Standards and Technology - a handwritten digit dataset), TensorFlow Flowers and SWIMSEG (Singapore Whole sky IMaging SEGmentation - a sky and cloud image dataset). We extend current work by probing the restricted Boltzmann machine (RBM) flow fixed point using two-point spin correlators. In the case of the Ising model, the RBM does not correctly reproduce properties related to longer ranged scaling dimensions. For the long range spin chain, the RBM does not correctly reproduce even the shortest distance scaling dimensions. This is an interesting result as this suggests RBMs are better at learning local features as opposed to long ranged interactions.

We introduce a method for comparing the coarse graining of RBMs and RG using correlation functions between inputs and outputs of both transformations. By studying spatial correlators of the input output correlators we learn about the local

patterns each output node encodes. Locality is a key characteristic of RG which motivates this method. Another key characteristic of the RG flow is the flow of temperature. We compare the flow of temperature in stacked RBM networks to several steps of RG. Here we see RG-like behaviour when studying the Ising model but when there are longer ranged interactions as in the case of the long range spin chain this no longer holds.

The final contribution relates to the generalization and noise rejection properties of RBMs and autoencoders. One of the main results we present shows that unsupervised deep learning is performing a coarse graining of the momentum space RG near the free field fixed point. By performing an eigen decomposition of the trained weight matrix we show that for large singular values the hidden singular vector is obtained by discarding the high Fourier modes of the visible singular vector. These large eigenvalues and associated visible and hidden singular vectors define the relevant modes of the training data. We also find strong agreement between the subspaces defined by the training data covariance matrix and the trained weights. Using an initial condition related to the training data covariance matrix we find training times improve by a factor between 4 and 100. This is shown not only on models of magnets but also using the MNIST, TensorFlow Flowers and SWIMSEG datasets.

Deep learning when viewed as curve fitting results in the generalization puzzle. There are millions of parameters which must be fitted using only tens or hundreds of thousands of training samples. The fact that deep networks do generalize goes against logic as we require more samples than parameters to fit the training data. The result we present provides a different framework in which to view unsupervised deep learning and shows that only a handful of parameters are needed to capture the relevant degrees of freedom. We show that deep networks are able to learn the underlying structure present in the training data and many of the parameters are irrelevant and can be set to 0. The reduction is so large that the number of training samples is greater than the number of parameters which need tuning. This result is a promising step forward for understanding unsupervised deep learning. In addition we can now phrase new questions in terms of a coarse graining framework. Future work includes extending these results to supervised networks.