Is Deep Learning an RG Flow?

Ellen de Mello Koch, Robert de Mello Koch and Ling Cheng

***Abstract*—Although there has been a rapid development of practical applications, theoretical explanations of deep learning are in their infancy. A possible starting point suggests that deep learning performs a sophisticated coarse graining. Coarse graining is the foundation of the renormalization group (RG), which provides a systematic construction of the theory of large scales starting from an underlying microscopic theory. In this way RG can be interpreted as providing a mechanism to explain the emergence of large scale structure, which is directly relevant to deep learning. We pursue the possibility that RG may provide a useful framework within which to pursue a theoretical explanation of deep learning. A statistical mechanics model for a magnet, the Ising model, is used to train an unsupervised RBM. The patterns generated by the trained RBM are compared to the configurations generated through a RG treatment of the Ising model. We argue that correlation functions between hidden and visible neurons are capable of diagnosing RG-like coarse graining. Numerical experiments show the presence of RG-like patterns in correlators computed using the trained RBMs. The observables we consider are also able to exhibit important differences between RG and deep learning.**

***Index Terms*—restricted Boltzmann machines (RBMs), deep learning, deep neural networks, learning theory, renormalization group (RG).**

I. Introduction

THE power of machine learning and artificial intelligence is established: these are powerful methods that already outperform humans in specific tasks [1]–[3]. Much of the research carried out in machine learning is of an applied nature. It establishes the practical utility of the method but does not construct an understanding of how deep learning works or even if such an understanding is possible [4]–[9]. Consequently, deep learning remains an impressive but mysterious black box. A possible starting point for a theoretical treatment suggests that deep learning is a form of coarse graining [3], [10], [11]. Since there are more input than output neurons this is almost certainly true. The real question is then if this is a useful observation, one that might shed light on how deep learning works. This is the question we take up in this paper.

We argue that understanding deep learning as a form of coarse graining is a useful observation, and make the case by adopting and adapting several ideas from theoretical physics. Specifically, in theoretical physics there is a sound framework to carry out coarse graining, known as the renormalization group (RG) [12]. RG provides a systematic way to construct a theory describing large scale features from an underlying microscopic description, which can be understood as recognizing sophisticated emergent patterns, a routine achievement of deep learning. Further, RG is applicable to field theories, that is, to systems with a large number of degrees of freedom so that it seems that RG is well positioned to deal with massive data sets. Finally, the way in which RG works is, in contrast to deep learning, well understood and can be described in precise mathematical language. These features suggest that RG may provide a useful framework in which to describe deep learning and attempts to argue that this is the case have been made in [13], [15]–[17]. We focus on unsupervised learning by a restricted Boltzmann machine (RBM) in what follows.

Two distinct possible connections to RG have been attempted, both relevant to our study. The first [13] is an attempt to link deep learning to the RG flow. The RG flow is a smooth process during which degrees of freedom are continuously averaged out, so that we flow from the initial microscopic description to the final macroscopic description. In deep learning one stacks layers of networks to obtain a deep network. The proposed connection of [13] suggests that each layer in the stack performs a small step along the RG flow[1]. We contribute to this discussion by developing quantitative tools with which this proposal can be explored with precision. The basic objects that appear in our analysis are correlation functions between the visible and the hidden neurons. This allows us to decode the mechanics of the RBM's pattern generation and to compare it to what the RG is doing. Although there are important differences, our results indicate remarkable similarities between how the RBM and RG achieve their results. The second approach [15], [16] builds an RBM flow using the weight matrix of the neural network after training is complete. The results of [15], [16] suggest that the RBM flow is closely related to the RG flow. We carry out a critical examination of this conclusion. The central tools we employ are correlation functions defined using the patterns generated by the RBM. We give a detailed and precise argument showing that the largest scale features of RG and RBM patterns are in complete agreement. The correlation functions involved are non-trivial probes of the statistics of the generated pattern[2] so the conclusion we reach is compelling. We also find that if one probes smaller scale features there are important differences between the two patterns. We will comment on the interpretation of these results in the conclusions. Our basic

Robert de Mello Koch is with the School of Physics and Telecommunication Engineering, South China Normal University, Guangzhou 510006, China and National Institute for Theoretical Physics and the School of Physics and Mandelstam Institute for Theoretical Physics, University of the Witwatersrand, Wits, 2050, South Africa.

Ellen de Mello Koch and Ling Cheng are with the School of Electrical and Information Engineering, University of the Witwatersrand, Wits, 2050, South Africa

[1]For a critical discussion of this proposal, see [11], [14]

[2]The studies carried out in [15], [16] used averages of the RBM pattern. Our correlation functions provide more sensitive probes into the structure of the pattern.

message is that the RG and RBM are doing the same thing, but using different methods.

The setting for our study is the two dimensional Ising model [18]. This is a simple model of a magnet, built for many individual "spins" each of which should be thought of as a microscopic bar magnet. Each spin can be aligned "up" or "down". Spins align at low temperatures producing a magnet. At high temperatures, spins are aligned randomly and there is no net magnetic field. The spins themselves define a binary pattern (the two states are up or down) and it is these patterns that the RBM learns. An important motivation for this choice of model is that it is well understood. The theory exhibits a first order phase transition terminating at a critical point. The theory at the critical point enjoys a conformal symmetry so that it can be solved exactly. It exhibits many interesting observables which we use to explore how deep learning is working [19], [20]. For example, if a neural network generates a microstate of the model, we can ask what the corresponding temperature of the microstate is. At the critical point special observables known as primary operators appear. Their correlation functions are power laws with powers that are known. These are the natural variables which encode, completely, the long scale features of the patterns. In this way, the Ising model gives a framework to explore deep learning both through the results of numerical experiments and using the complete understanding of the large scale features of the coarse grained system. To probe whether deep learning is a type of coarse graining this knowledge of correlations on large length scales is a valuable tool.

Our study of an Ising magnet may seem rather far removed from more usual (and practical) applications, including for example image recognition and manipulation. However, one might be optimistic that lessons learned from the Ising model are applicable to these more familiar examples. Indeed, the energy function of the Ising model tries to align nearby spins with the result that nearby spins are correlated. This is not at all unlike an image for which the color of nearby pixels is likely to be correlated [21].

A description of deep learning in the RG framework would have important implications. RG explains how macroscopic physics emerges from microscopic physics. This understanding leads to an organization of the microscopic physics into features that are relevant or irrelevant, so that in the end the emergent patterns depend only on a small number of relevant parameters. Carried over to the deep learning context, a similar understanding will strive to explain what features of the data and which weights in the network are important for deep learning. Such an understanding would have implications for what architectures are optimal and how the learning process can be improved and made more efficient.

We now sketch the content of the paper and outline how it is organized. In Section II we give a quick review of RBMs, RG and the Ising model, providing the background needed to follow subsequent arguments. In Section III we consider the RBM flows defined using the matrix of weights learned by the network [15], [16]. By studying correlation functions of primary operators of the Ising conformal field theory, we argue that although the RBM and RG patterns agree remarkably well on the largest scales of the pattern they differ just as dramatically, on the short scale structure. In Section IV we examine the possibility that deep learning reconstructs an RG flow, with each layer of the deep network performing one step of the flow. Our discussion begins with a critical look at the argument given in [13] that claims that deep learning is mapped onto the RG flow. The argument shows a system of equations that is obeyed by both the RBM and a variational realization of the RG flow. Our basic conclusion is that the argument of [13] only shows that aspects of the RBM learning are consistent with the structure of the RG transformation. Indeed, we explicitly construct examples that satisfy the equations derived by [13] that certainly do not perform an RG flow or arise from an RBM. Nevertheless, the arguments of [13] are compelling and we find the possible connection between deep learning and RG fascinating and deserving of further study. Towards this end we couch some of the qualitative observations of [13] as statements about the behaviour of well chosen correlation functions. The form of these correlators, puts certain qualitative observations of Section IV.B. of [13] onto a firm quantitative footing. Finally we study the RG flow of the temperature. This turns out to be interesting as it reveals a further difference between the RBM patterns and RG. In the final Section of this paper, we discuss our results and suggest open directions that can be pursued.

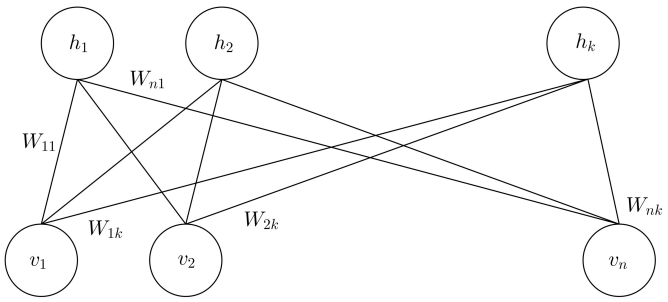


Fig. 1: An RBM network with visible nodes v_i and hidden nodes h_a where $N_v = n$ and $N_h = k$. Connections between input and output nodes are each associated with a weight $W_i a$.

II. RBM, RG and Ising

In this section we introduce the background material used in our study. The first subsection reviews RBMs emphasizing both the structure of the network and its implementation. Following that the RG is reviewed, emphasizing aspects relevant to the deep learning application. This section concludes with a review of the Ising model, motivating why the model is considered.

A. Restricted Boltzmann Machines

RBMs perform unsupervised learning to extract features from a given data set [22]–[24]. They have an input and an output layer. The input layer is made up of visible nodes, v_i with $i = 1, 2 \cdots, N_v$ and the output layer is made up of hidden nodes, h_a with $a = 1, 2, \cdots, N_h$ as illustrated in Figure 1.

The input nodes are set with values of ± 1 and the trained network generates a corresponding pattern by setting the output nodes to ± 1. The values of the output neurons are obtained by evaluating a non-linear function on a linear combination of the input neurons, perhaps offset by a constant bias. The specific linear combination of neurons is represented by connections between nodes, with a weight for each connection. For the RBM there are connections between every input node and every output node, while nodes belonging to the same layer are not connected. The "unrestricted" Boltzmann machines allow connections between any two nodes in the network [25], but this generality comes at a cost: training algorithms are much less efficient [22]–[24]. The connection between input node v_i and output node h_a is assigned a weight, W_{ia}, and visible (hidden) nodes are assigned a bias $b_i^{(v)}$ ($b_a^{(h)}$). Using these ingredients we define a Hamiltonian for the RBM

$$E = -\sum_a b_a^{(h)} h_a - \sum_{i,a} v_i W_{ia} h_a - \sum_i b_i^{(v)} v_i \tag{1}$$

where $h_a, v_i \in \{-1, 1\}$. The RBM defines the probability distribution for obtaining the pair of input and output vectors, $\mathbf{v}$ and $\mathbf{h}$, by [26]

$$p(\mathbf{v}, \mathbf{h}) = \frac{1}{\mathcal{Z}} e^{-E} \tag{2}$$

where $\mathcal{Z}$ is the partition function, obtained by summing over all possible hidden and visible vectors

$$\mathcal{Z} = \sum_{\{\mathbf{v},\mathbf{h}\}} e^{-E} \tag{3}$$

As usual, to determine the marginal distribution of a visible vector, sum over the state space of hidden vectors

$$p(\mathbf{v}) = \frac{1}{\mathcal{Z}} \sum_{\{\mathbf{h}\}} e^{-E} \tag{4}$$

Similarly, the marginal distribution of a hidden vector is

$$p(\mathbf{h}) = \frac{1}{\mathcal{Z}} \sum_{\{\mathbf{v}\}} e^{-E} \tag{5}$$

The weights, W_{ia} and biases $b_i^{(v)}$, $b_a^{(h)}$ are determined during training. Training strives to match the model distribution $p(\mathbf{v})$ to the distribution $q(\mathbf{v})$ defined by the data and it achieves this by minimising the Kullback-Liebler (KL) divergence. The KL divergence is a measure of how much information is lost when approximating the actual distribution with the model distribution [27]. Training adjusts $\{W_{ia}, b_i^{(v)}, b_a^{(h)}\}$ to minimize the KL divergence, which is given by

$$\begin{aligned} D_{KL}(q||p) &= \sum_{i=1}^{N_v} q(v_i) \left(\log\left(q(v_i)\right) - \log\left(p(v_i)\right)\right) \\ &= \sum_{i=1}^{N_v} q(v_i) \log\left(\frac{q(v_i)}{p(v_i)}\right) \end{aligned} \tag{6}$$

Gradients of the KL divergence used to update the parameters of the RBM are computed as follows

$$\frac{\partial D_{KL}(q||p)}{\partial W_{ia}} = \langle v_i h_a \rangle_{data} - \langle v_i h_a \rangle_{model} \tag{7}$$

$$\frac{\partial D_{KL}(q||p)}{\partial b_i^{(v)}} = \langle v_i \rangle_{data} - \langle v_i \rangle_{model} \tag{8}$$

$$\frac{\partial D_{KL}(q||p)}{\partial b_a^{(h)}} = \langle h_a \rangle_{data} - \langle h_a \rangle_{model} \tag{9}$$

where the expectation values appearing above are easily derived using (2). They are given explicitly in Appendix A. The data set contains an enormous number N_s samples implying that the method just outlined is numerically intractable: the sum over the whole state space of visible and hidden vectors is too expensive. To make progress, approximate the KL divergence by the contrastive divergence [24]. Rather than summing over the entire state space of visible and hidden vectors, one simply sets the states of the visible units to the training data [26]. This is an enormous simplification. Given a set of visible vectors, the hidden vectors are sampled by setting each h_a to 1 with probability

$$p(h_a|\mathbf{v}) = \tanh\left(\sum_i W_{ia} v_i + b_a^{(h)}\right) \tag{10}$$

Likewise, given a set of h_a, we are able to sample visible vectors by setting each v_i to 1 with probability

$$p(v_i|\mathbf{h}) = \tanh\left(\sum_a W_{ia} h_a + b_i^{v}\right) \tag{11}$$

Expectation values for the data are computed using $\hat{\mathbf{h}}$ and $\hat{\mathbf{v}}$. The set of hidden vectors $\hat{\mathbf{h}}$, is generated using (10). The training data set provides $\hat{\mathbf{v}}$.

To determine model expectation values, determine a sample of visible vectors $\tilde{\mathbf{v}}$ and a sample of the hidden vectors $\tilde{\mathbf{h}}$. Expectation values of the model are approximated using these sets. Again, summing these much smaller sets (and not the complete space of hidden and visible vectors) is an enormous simplification. The equations (11) and (10) determine the sets of visible and hidden vectors. The set $\tilde{\mathbf{v}}$, is calculated using $\hat{\mathbf{h}}$. We then determine $\tilde{\mathbf{h}}$, using $\tilde{\mathbf{v}}$. The equations for $\hat{\mathbf{h}}$, $\tilde{\mathbf{v}}$ and $\tilde{\mathbf{h}}$ are thus

$$\hat{h}_a^{(A)} = \tanh\left(\sum_i W_{ia} \hat{v}_i^{(A)} + b_a^{(h)}\right) \tag{12}$$

$$\begin{aligned} \tilde{v}_i^{(A)} &= \tanh\left(\sum_a W_{ia} \hat{h}_a^{(A)} + b_i^{(v)}\right) \\ &= \tanh\left(\sum_a W_{ia} \tanh\left(\sum_k W_{ka} \hat{v}_k^{(A)} + b_a^{(h)}\right) + b_i^{(v)}\right) \end{aligned} \tag{13}$$

$$\begin{aligned}\tilde{h}_a^{(A)} &= \tanh\left(\sum_i W_{ia}\tilde{v}_i^{(A)} + b_a^{(h)}\right)\\ &= \tanh\left(\sum_i W_{ia}\tanh\left(\sum_k W_{ik}\hat{h}_k^{(A)} + b_i^{(v)}\right) + b_a^{(h)}\right)\\ &= \tanh\left(\sum_i W_{ia}\tanh\left(\sum_k W_{ik}\tanh\left(\sum_m W_{mk}\hat{v}_m^{(A)} + \right.\right.\right.\\ &\quad \left.\left.\left. + b_k^{(h)}\right) + b_i^{(v)}\right) + b_a^{(h)}\right)\end{aligned} \tag{14}$$

Using this approximation the expressions used to train the RBM are

$$\langle v_i h_a\rangle_{data} = \frac{1}{N_s}\sum_A \hat{v_i}^{(A)}\hat{h_a}^{(A)} \tag{15}$$

$$\langle v_i h_a\rangle_{model} = \frac{1}{N_s}\sum_A \tilde{v_i}^{(A)}\tilde{h_a}^{(A)} \tag{16}$$

$$\langle v_i\rangle_{data} = \frac{1}{N_s}\sum_A \hat{v_i}^{(A)} \tag{17}$$

$$\langle v_i\rangle_{model} = \frac{1}{N_s}\sum_A \tilde{v_i}^{(A)} \tag{18}$$

$$\langle h_a\rangle_{data} = \frac{1}{N_s}\sum_A \hat{h_a}^{(A)} \tag{19}$$

$$\langle h_a\rangle_{model} = \frac{1}{N_s}\sum_A \tilde{h_a}^{(A)} \tag{20}$$

where $A = 1, 2, 3, \ldots, N_s$ denotes samples in the training data set made up of N_s samples. These approximations achieve a dramatic speed up in training. Although this method performs well in practice [1], [29], [30], it is difficult to understand when and why the approximations work [26], [28]. This approximation does not follow the gradient of any function [31].

B. RG

RG is a tool used routinely in quantum field theory and statistical mechanics [12]. RG coarse grains by first organizing the theory according to length scales and then averaging over the short distance degrees of freedom. The result is an effective theory for the long distance degrees of freedom. RG thus gives a systematic procedure to determine the dynamical laws governing macroscopic physics of a system with given microscopic laws, and it achieves this by employing coarse graining. The analogy to deep learning should be evident: deep learning also extracts regularities from massive data sets.

To illustrate RG consider the example provided by quantum field theory. Observables $\mathcal{O}$ are functions (usually polynomials) of the field and its derivatives. Examples of observables are the energy or momentum of the field. To calculate the expected value $\langle\mathcal{O}\rangle$ of observable $\mathcal{O}$, integrate (i.e. average) over all possible field configurations

$$\langle\mathcal{O}\rangle = \int [d\phi] e^{-S}\mathcal{O} \tag{21}$$

The factor e^{-S}, which defines a probability measure on the space of fields, depends on the theory considered. S is called the action of the theory and is also a polynomial in the field and its derivatives, with the coefficients of the polynomial providing the parameters of the theory, things like couplings and masses. A theory is defined by specifying S.

To coarse grain, express the position space field in terms of momentum space components

$$\phi(x) = \int dk e^{ik\cdot x}\phi(k) \tag{22}$$

$e^{ik\cdot x}$ oscillates in position space with wavelength $\frac{2\pi}{k}$. High momentum (big k) components have small wavelengths and encode small distance structure. Low momentum components have huge wavelengths and describe large distance structure. Declare there is a smallest possible structure, implemented by cutting off the momentum modes at a large momentum Λ as follows

$$[d\phi] = \prod_{k<\Lambda} d\phi(k) \tag{23}$$

RG breaks the integration measure into high and low momentum components $[d\phi] = [d\phi_<][d\phi_>]$ where

$$\begin{aligned}[d\phi_<] &= \prod_{k<(1-\epsilon)\Lambda} d\phi(k)\\ [d\phi_>] &= \prod_{(1-\epsilon)\Lambda<k<\Lambda} d\phi(k)\end{aligned} \tag{24}$$

RG considers observables that depend only on large scale structure of the theory, i.e. observables that depend only on $\phi_<$. In this case, when computing the expected value of $\mathcal{O}$ we can pull $\mathcal{O}$ out of the integral over $\phi_>$ and integrate over the high momentum components

$$\begin{aligned}\langle\mathcal{O}(\phi_<)\rangle &= \int [d\phi_<]\int [d\phi_>]\; e^{-S}\mathcal{O}(\phi_<)\\ &= \int [d\phi_<]\; e^{-S_{\text{eff}}}\mathcal{O}(\phi_<)\end{aligned} \tag{25}$$

This procedure of splitting momentum components into two sets and integrating over the large momenta defines a new action S_{eff}. Repeating the procedure many times defines the RG flow under which S_{eff} changes continuously. After the flow, one is left with an integral over the very long wavelength modes. This completes the coarse graining: we have a new theory defined by S_{eff}. The new theory uses only long wavelength components of the field and correctly reproduces the expected value of any observable depending only on long wavelength components. Values of the parameters of the theory, which appear in S_{eff} change under this transformation. In general, many possible terms are generated and appear in S_{eff}. Each possible term defines a coupling of the theory. Each coupling can be classified as marginal (the size of the coupling is unchanged by the RG flow), relevant (the coupling grows under the flow) or irrelevant (the coupling goes to

zero under the flow). It is a dramatic insight of Wilson that almost all couplings in any given quantum field theory are irrelevant and so the low energy theory is characterized by a handful of parameters. This is a dramatic (experimentally verified) simplicity hidden in the rather complicated quantum field theory. This simplicity explains why "simple large scale patterns" can emerge from "complicated short distance data". The possibility that the same simplicity is at work in deep learning is a key motivation of this paper.

Conceptually, the coarse graining performed by RG is well defined. Computationally, it is almost impossible to carry out. To develop a useful calculational scheme, partition the momentum components into tiny sets (i.e. follow (24) with ϵ infinitesimal) and ask what happens when we average over a single tiny set. Two things happen: couplings g_i change $\delta g_i \propto \beta_i$ and the strength of the field changes $\delta\phi \propto \gamma\phi$. One can prove that all observables built using n fields will obey the Callan-Symanzik equation

$$\left(\mu\frac{\partial}{\partial\mu} + \sum_i \beta_i \frac{\partial}{\partial g_i} + n\gamma\right)\langle\mathcal{O}\rangle = 0 \tag{26}$$

The parameter μ here defines the scale of the effective theory: the smallest wavelength in the effective theory is $\frac{2\pi}{\mu}$. This equation provides a remarkable and simple description of the RG coarse graining that captures the essential features of the long distance effective theory.

If RG (or a variant of it) is relevant to understanding deep learning, it makes concrete suggestions for the resulting theory. For example, is there an analogue of the Callan-Symanzik equation? One might assign beta functions $\beta_{ia}, \beta_i^{(v)}, \beta_a^{(h)}$ to the weights W_{ia} and biases $b_i^{(v)}, b_a^{(h)}$. These would determine which parameters of the RBM are relevant, irrelevant or marginal.

The RG flow halts at a fixed point, described by a conformal field theory. This field theory enjoys additional symmetries including scale invariance. It is interesting to note that the possibility that scale invariance plays a role in deep learning has been raised in [11], [13], [15]–[17].

C. Ising model

The Ising model is a model for a magnet. The two dimensional model has a discrete variable $\sigma_{\vec{k}} = \pm 1$ on each site of a rectangular lattice. The sites are labeled by a two dimensional vector $\vec{k}$, which has integer components. Spins on adjacent sites i and j interact with strength J_{ij}. Each spin will also interact with an external magnetic field h_j, with strength μ. The energy of a given configuration $\{\sigma_{\vec{k}}\}$ of spins is determined by the Hamiltonian

$$H = -\sum_{\langle i\,j\rangle} J_{ij}\sigma_i\sigma_j - \mu\sum_j h_j\sigma_j \tag{27}$$

where the first sum is over adjacent pairs, indicated by $\langle i\,j\rangle$. From now on we set $h_j = 0$ and what follows holds only for vanishing external magnetic field. The probability of configuration $\{\sigma_{\vec{k}}\}$ of spins is given by the Boltzmann distribution, with inverse temperature $\beta \geq 0$

$$P_\beta(\sigma) = \frac{e^{-\beta H(\{\sigma_{\vec{k}}\})}}{Z_\beta} \tag{28}$$

where the constant Z_β, the partition function, is given by

$$Z_\beta = \sum_{\{\sigma_{\vec{k}}\}} e^{-\beta H(\sigma)} \tag{29}$$

Averages of physical observables are defined by

$$\langle f\rangle_\beta = \sum_\sigma f(\sigma)P_\beta(\sigma) \tag{30}$$

We study unsupervised learning of the Ising Model by a RBM. The input data that is used to train the network is generated using the probability measure (28).

There are good reasons to focus on the Ising model. The model has a fixed point in its RG flow. The fixed point is described by a well known conformal field theory [32]. This fixed point is an unstable fixed point meaning that generic flows move away from the fixed point. We must tune things carefully if we are to terminate on the fixed point. This tuning is necessary because there is a relevant operator present and it tends to push us away from the fixed point. The papers [15], [16] argue that the RBM flow always flows to the fixed point. This challenges conventional wisdom and it suggests a different kind of coarse graining to that employed by RG, is at work. A distinct proposal [13] claims that the RG flow arises by stacking RBMs to produce a classic deep learning scenario. Each layer of the deep network performs a step in the flow.

At the Ising model fixed point, detailed checks of both proposals are possible. There are observables, known as primary operators, whose correlators are power laws of distance on the lattice. The power entering these power laws are known, so that we have a rich and detailed data set that the RBM must reproduce if it is indeed performing a RG coarse graining. This is a compelling motivation for the model. Another advantage of the model is simplicity: it is a model of spins which take the values ± 1 so it defines a simple model with discrete variables, well suited to numerical study and naturally accommodated in the RBM framework. Finally, the Ising model is not that far removed from real world applications: the Ising Hamiltonian favors configurations with aligned neighboring spins. Thus, at low enough temperatures "smooth" slowly varying configurations of spins are favored. This is similar to data defining images for example, where neighboring pixels are likely to have the same color. In slightly poetic language one could say that at low temperatures the Ising model favors pictures and not speckle.

We end this section with a summary of the most relevant features of the Ising model fixed point. At the critical temperature

$$T_c = \frac{2J}{k\ln(1+\sqrt{2})} \tag{31}$$

where J is the interaction strength and k is the Boltzmann constant, the Ising model undergoes a second order phase transition. There are two competing phases: an ordered (low temperature) phase in which spins align producing a macroscopic magnetization, and a disordered (high temperature) phase in which spins fluctuate randomly and the magnetization averages to zero. At the critical point the Ising model develops a full conformal invariance and one can use the full power of

conformal symmetry to tackle the problem. The field which takes values ± 1 in the Ising model is a primary field, of dimension $\Delta = \frac{1}{8}$. The two and three point correlation functions of primary fields are determined by conformal invariance to be

$$\langle \phi(\vec{x}_1)\phi(\vec{x}_2)\rangle = \frac{B_1}{|\vec{x}_1 - \vec{x}_2|^{2\Delta}} \tag{32}$$

$$\langle \phi(\vec{x}_1)\phi(\vec{x}_2)\phi(\vec{x}_3)\rangle = \frac{B_2}{|\vec{x}_1 - \vec{x}_2|^{\Delta}|\vec{x}_1 - \vec{x}_3|^{\Delta}|\vec{x}_2 - \vec{x}_3|^{\Delta}} \tag{33}$$

where B_1 and B_2 are constants. There is also a primary operator in the Ising model (which we describe below) with a dimension $\Delta = 1$. These correlation functions must be reproduced by the RBM if it is indeed flowing to the critical point of the Ising model.

III. Flows derived from learned weights

In this section we consider the RBM flows introduced in [15], [16]. These flows use the weight matrix W_{ia}, and bias vectors $b_i^{(v)}$ and $b_a^{(h)}$, obtained by training, to define a continuous flow from an initial spin configuration to a final spin configuration. The flow appears to exhibit a fascinating behavior: given any initial snapshot, the RBM flows towards the critical point of the Ising model. This is in contrast to the RG which flows away from the fixed point. In addition, the number of spins in the configuration is a constant along the RBM flow. In contrast to this, the number of spins in the configuration decreases along the RG flow, as high energy modes are averaged over to produce the coarse grained description. Despite these differences, the flow of [15], [16] appears to produce configurations ever closer to the critical temperature and these configurations yield impressively accurate predictions for the critical exponents of the Ising magnet. Our goal in this section is to further test if the RBM flow produces configurations at the critical point of the Ising model. We explore the spatial dependence of spin correlations in configurations produced by the flow. On large scales the Ising critical point configurations are correctly reproduced. However, as one starts to probe smaller scales there are definite quantifiable departures from the Ising predictions.

RBM flows [16] are generated using equation (13) together with the trained weight matrix, W_{ia}, and bias vectors, $b_i^{(v)}$ and $b_a^{(h)}$. Apply equation (13) to the Ath input data sample (taken from the training set) $\hat{v}^{(A)}$, to produce a single step in the RBM flow, $\tilde{v}_1^{(A)}$. The flow proceeds by repeatedly applying equation (13). Concretely, the flow of length n is

$$\begin{aligned}
\tilde{v}_1^{(A)} &= \tanh\left(W \cdot \left[\tanh\left(\hat{v}^{(A)} \cdot W + b^{(h)}\right)\right]^T + b^{(v)}\right) \\
\tilde{v}_2^{(A)} &= \tanh\left(W \cdot \left[\tanh\left(\tilde{v}_1^{(A)} \cdot W + b^{(h)}\right)\right]^T + b^{(v)}\right) \\
&\vdots \\
\tilde{v}_n^{(A)} &= \tanh\left(W \cdot \left[\tanh\left(\tilde{v}_{n-1}^{(A)} \cdot W + b^{(h)}\right)\right]^T + b^{(v)}\right)
\end{aligned} \tag{34}$$

Note that the length of the vector $\tilde{v}_k^{(A)}$ is a constant of the flow and consequently there is not obviously any coarse graining implemented.

A. Numerical results

This section considers statistical properties of configurations produced by the RBM flow. At the Ising critical point, the theory enjoys a conformal invariance. Using this symmetry a special class of operators with a definite scaling dimension Δ can be identified. The utility of these operators is that their spatial two point correlation functions drop off as a known power of the distance between the two operators, as reviewed above in equation (32). These two point functions can be evaluated using the RBM flow configurations and, if these configurations are critical Ising states, they must reproduce the known correlation functions. This is one of the checks performed and it detects discrepancies with the Ising model predictions. There are two primary operators we consider. This first is the basic spin variable minus its average value $s_{ij} = \sigma_{ij} - \bar{\sigma}$. The prediction for the two point function is (32) with $\Delta_s = \frac{1}{8}$. This correlator falls off rather slowly, so that this two point function probes the large scale features of the RBM flow configurations. The RBM flow nicely reproduces this correlator and in fact, this is enough to reproduce the critical exponent for the Ising model consistent with the results of [16]. One should note however that our computation and those of [16] could very well have disagreed, since they probe different things. The critical exponent evaluated in [16] uses the magnetisation computed from different flows generated by the RBM, at temperatures around the critical temperature. Magnetisation measures the average of the spin in the lattice. It is blind to the spatial location of each spin. On the other hand, the two-point correlation function is entirely determined by the spatial location of spins in a single flow configuration. We also consider a second primary operator

$$\epsilon_{ij} = s_{ij} \cdot (s_{i+1,j} + s_{i-1,j} + s_{i,j+1} + s_{i,j-1}) - \bar{\epsilon}_{ij}, \tag{35}$$

which has $\Delta_\epsilon = 1$. This correlation function falls off much faster and is consequently a probe of shorter scale features of the RBM configurations. The RBM flow fails to reproduce this correlation function, indicating that the RBM configurations differ from those of the critical Ising model.

Consider an RBM network with 100 visible nodes and 81 hidden nodes. The number of visible and hidden nodes is chosen to match [16], so that we can compare our results to existing literature. The network trains on data generated by Monte Carlo simulations with the Boltzmann distribution in equation (28). The training data set includes 20000 samples at each temperature, ranging from 0 to 5.9 in increments of 0.1. This gives a total of 1200000 configurations. Training uses 10000 iterations of contrastive divergence, performed with the update equations (7), (8) and (9). Once the flow configurations are generated, following [15], a supervised network is used to measure the temperature of each flow. The supervised network is trained to measure discrete temperatures of $T = 0, 0.1, \ldots, 5.9$.

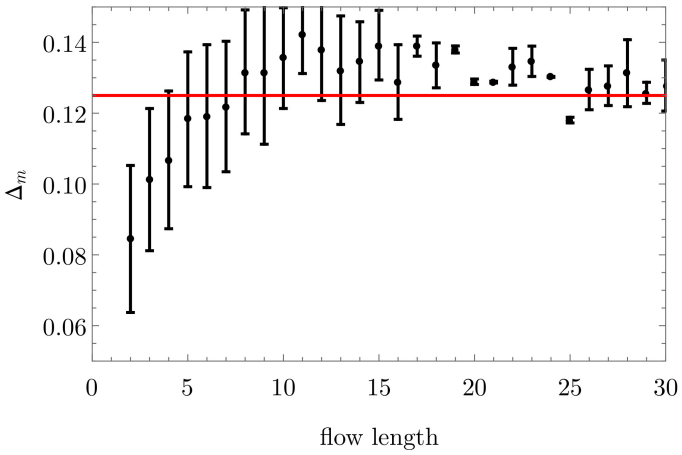


Fig. 2: Estimates of the scaling dimension versus flow length, obtained using the average magnetisation of flows at temperatures $T = 2.1$, 2.2 and 2.3. The red line indicates the value of $\Delta = 0.125$ at the critical point. After approximately 8 flows, Δ_m converges to this critical value.

To estimate Δ using magnetisation [16], select flows at temperatures close to T_c, where the average magnetisation m depends on temperature as

$$m \approx \frac{A|T - T_c|^{\Delta_m}}{T_c} \tag{36}$$

We denote Δ obtained by this fitting as Δ_m. The fit also determines A and the critical temperature is $T_c = 2.269$. The fit uses the magnetization computed at temperatures 2.1, 2.2 and 2.3. A plot of Δ_m versus flow length is given in figure 2. Our results indicate that we converge to the correct critical value $\Delta_m = 0.125$ for flows of length 8. This is completely consistent with the results of [16], who find convergence for flow lengths of 9.

We now shift our focus to consider spatial two point correlation functions computed using the configurations generated by the RBM flows. The correlators are calculated using the flow configuration and the result is then fitted to the function in equation (32) to estimate Δ. We denote this estimate by Δ_s as it is determined using spatial information. For RBM flows at the critical temperature, the prediction is $\Delta_s = 0.125$ at $T = 2.269$, as explained above. Even with a limited lattice of 10 by 10 spins, we find $\Delta_s = 0.1263$ using Monte Carlo Ising model configurations. A plot showing this estimate can be seen in figure 3b. The point of this exercise is to demonstrate that a lattice of size 10 by 10 is large enough to estimate the scaling dimension of interest.

Figure 3a shows the scaling dimension Δ_s versus the flow length, for RBM flows at temperatures of 2.2 (in gray) and 2.3 (in black). The red horizontal line indicates the scaling dimension at the critical point. The results are intuitively appealing. The gray points in figure 3a show estimates of Δ_s from flows slightly below the critical temperature, where the scaling dimension is slightly underestimated. Below the critical temperature spins are more likely to align and so the correlator should fall off more slowly than at the critical temperature. This is what our results shows. The black points in figure 3a show Δ_s estimated using flows slightly above the critical temperature. The scaling dimension is over estimated, again as expected. Selecting flows at T_c would determine the scaling dimension in between the values shown in figure 3a. This gives a value very close to the prediction of $\Delta_s = 0.125$.

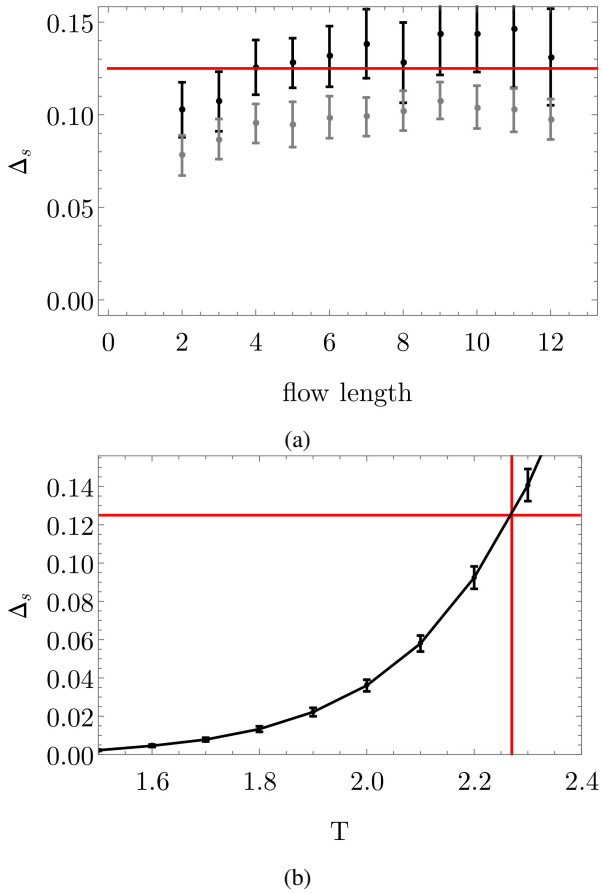
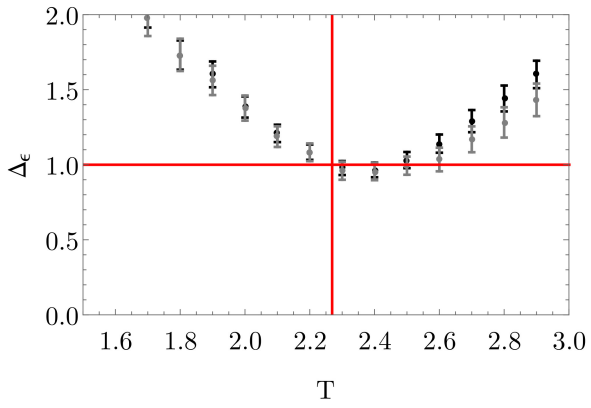


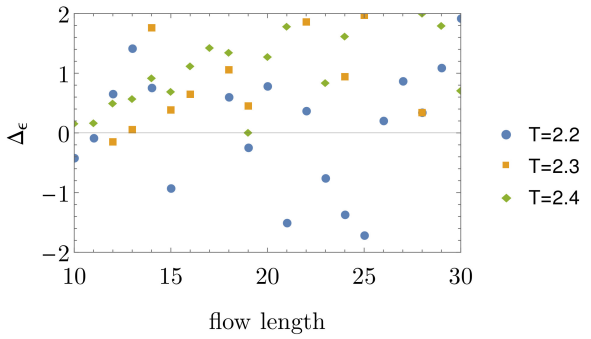
Fig. 3: Plot showing the estimated scaling dimension versus flow length using the two-point correlation function for (a) flows at $T = 2.2$ (in gray) and flows at $T = 2.3$ (in black). Plot (b) shows the estimated scaling dimension versus temperature for the Ising model data used for training.

The two point correlation functions for the spin variable establish that the critical Ising states and the states produced by the RBM flow share the same large scale spatial features. We will now consider the two point correlation function of the ϵ_{ij} field, which probes spatial features on a smaller scale. Using critical Ising data generated using Monte Carlo, on a lattice of size 10 by 10 and 9 by 9, we estimate Δ_ϵ at various temperatures as shown in figure 4a. The intersection of the red horizontal and vertical lines cross the critical temperature and prediction $\Delta_\epsilon = 1$. Interpolating the Ising data with a continuous curve, we would pass through the intersection point, as predicted. The RBM flows are unable to confirm this prediction. Indeed, the RBM flows near T_c are summarized in figure 4b. None of the three temperatures shown have a value of Δ_ϵ that converges with flow length.

The fact that the RBM produces configurations that correctly reproduce the correlation function of the spin field s_{ij} but not of the ϵ_{ij} implies that although the spatial correlations



(a)



(b)

Fig. 4: Plots showing Δ_ϵ calculated using (a) Monte Carlo Ising model data on a 10 by 10 lattice (in black) and a 9 by 9 lattice (grey), (b) RBM flows at a temperature of 2.2, 2.3 and 2.4. Error bars in plot (a) indicate a 90% confidence interval. No error bars are shown in (b) as the error bars are larger than the y range.

encoded into the RBM flow configurations agree with those of the critical Ising configurations at long length scales, the two start to differ on smaller length scales. This conclusion agrees with [16] which also finds differences between the RBM flow and RG. [16] considers $h \neq 0$ and uses different arguments to reach the conclusion.

IV. Flows derived from deep learning

The RBM flows of the previous section provide one possible link to RG. An independent line reasoning, developed in [13], claims a mapping between deep learning and RG. The idea is not that there is an analogy between deep learning and RG, but rather, that the two are to be identified. The argument for this identity starts from the energy function of the RBM, which is

$$E(\{v_i, h_a\}) = b_a h_a + v_i W_{ia} h_a + c_i v_i \tag{37}$$

This energy determines the probability of obtaining configuration $\{v_i, h_a\}$ as

$$p_\lambda(\{v_i, h_a\}) = \frac{e^{-E(\{v_i,h_a\})}}{\mathcal{Z}} \tag{38}$$

Marginal distributions for hidden and visible spins are defined as follows

$$\begin{aligned} p_\lambda(\{h_a\}) &= \sum_{\{v_i\}} p_\lambda(\{v_i, h_a\}) = Tr_{v_i} p_\lambda(\{v_i, h_a\}) \\ p_\lambda(\{v_i\}) &= \sum_{\{h_a\}} p_\lambda(\{v_i, h_a\}) = Tr_{h_a} p_\lambda(\{v_i, h_a\}) \end{aligned} \tag{39}$$

The equations (39) are key equations of the RBM and [13] essentially use these to characterize the RBM. The comparison to RG is made by Next, as explained in Appendix B, the variational RG uses an operator $T(\{v_i, h_a\})$ defined as follows

$$\frac{e^{H_\lambda^{RG}(\{h_a\})}}{\mathcal{Z}} = Tr_{v_i} \frac{e^{T(\{v_i,h_a\})-H(\{v_i\})}}{\mathcal{Z}} \tag{40}$$

In this formula, $H(\{v_i\})$ is the microscopic Hamiltonian describing the dynamics of the visible spins and $H_\lambda^{RG}(\{h_a\})$ is the coarse grained Hamiltonian describing the hidden spins. The operator $T(\{v_i, h_a\})$ is required to obey

$$Tr_{h_a} e^{T(\{v_i,h_a\})} = 1 \tag{41}$$

which obviously implies that

$$Tr_{h_a} e^{T(\{v_i,h_a\})-H(\{v_i\})} = e^{-H(\{v_i\})} \tag{42}$$

Notice that (40) and (42) exactly match (39) as long as we identify

$$T(\{v_i, h_a\}) = -E(\{v_i, h_a\}) + H(\{v_i\}) \tag{43}$$

This then implies that

$$\begin{aligned} \frac{e^{H_\lambda^{RG}(\{h_a\})}}{\mathcal{Z}} &= Tr_{v_i} \frac{e^{T(\{v_i,h_a\})-H(\{v_i\})}}{\mathcal{Z}} \\ &= Tr_{v_i} \frac{e^{-E(\{v_i,h_a\})}}{\mathcal{Z}} \\ &= \frac{e^{-H_\lambda^{RBM}(\{h_a\})}}{\mathcal{Z}} \end{aligned} \tag{44}$$

which is the central claim of [13].

The above argument proves an equivalence between deep learning and RG if and only if the equations (39) provide a unique characterization of the joint probability function $p_\lambda(\{v_i, h_a\})$. This is not the case: it is easy to construct functions $p_\lambda(\{v_i, h_a\})$ that obey (39), but are nothing like either the RBM or RG functions. As an example, define

$$\begin{aligned} \rho(\{v_i\}) &= \frac{Tr_{h_a}\left(e^{T(\{v_i,h_a\})-H(\{v_i\})}\right)}{\mathcal{Z}} \\ \tilde{\rho}(\{h_a\}) &= \frac{Tr_{v_i}\left(e^{T(\{v_i,h_a\})-H(\{v_i\})}\right)}{\mathcal{Z}} \end{aligned} \tag{45}$$

where

$$\mathcal{Z} = \sum_{v_i, h_a} e^{T(\{v_i,h_a\})-H(\{v_i\})} \tag{46}$$

We clearly have $Tr_{v_i}(\rho(\{v_i\})) = 1 = Tr_{h_a}(\tilde{\rho}(\{h_a\}))$ which implies that

$$A_\lambda(\{v_i, h_a\}) = \tilde{\rho}(\{h_a\})\rho(\{v_i\}) \tag{47}$$

obeys (39). It is quite clear that in $A_\lambda(\{v_i, h_a\})$ there are no correlations between the hidden and visible spins

$$\begin{aligned}\langle v_j h_b\rangle =& Tr_{v_i,h_a}(v_j h_b\, A_\lambda(\{v_i,h_a\})\,)\\ =& Tr_{v_i}(\rho(\{v_i\})v_j) Tr_{h_a}(\tilde{\rho}(\{h_a\})h_b)\\ =& \langle v_j\rangle\langle h_b\rangle\end{aligned} \tag{48}$$

so that we would reject it as a possible model of either the RG quantity

$$\mathcal{Z}^{-1} e^{T(\{v_i,h_a\})-H(\{v_i\})} \tag{49}$$

or of the RBM quantity

$$\mathcal{Z}^{-1} e^{-E(\{v_i,h_a\})} \tag{50}$$

In addition to clarifying aspects of the argument of [13], the joint correlation functions between visible and hidden spins can be used to characterize the RG flow, as we now explain. The RG flow "coarse grains" in position space: a "block of spins" is replaced by an effective spin, whose magnitude is the average of the spins it replaces. Since correlations between microscopic spins fall off with distance, an RG coarse graining implies that because the hidden spin is a linear combination of nearby visible spins, the correlation function between hidden and visible spins reflects a correlation between a hidden spin and a cluster of visible spins. We will search for this distinctive signal in the $\langle v_i h_a\rangle$ correlator, to find quantitative evidence that deep learning is indeed performing an RG coarse graining.

A. *Numerical results*

Our numerical study aims to do two things: First, we establish whether there are RG-like patterns present within the correlator $\langle v_i h_a\rangle$, for correlators computed using the patterns generated by an RBM flow. If these patterns are indeed present, this constitutes strong evidence in favor of the connection between RG and deep learning. We find that RG-like patterns do indeed emerge. Second, according to the proposal of [13], in a deep network each layer that is stacked to produce the depth of the network performs one step in the RG flow. With this interpretation in mind, it may be useful to compare how a network with multiple stacked RBMs learns as compared to a network with a single layer. We do not have anything concrete to report on this question.

The training data is a set of 30000 configurations of Ising model 32 by 32 lattices, near the critical temperature $T = 2.269$. The dataset is generated using Monte Carlo simulations. An input lattice length of 32 allows a large enough final configuration even after two steps of RG, corresponding to stacking two RBMs. In each step of the RG, the number of lattice sites is reduced by a factor of 4. Thus, we flow from lattices with 1024 sites to lattices with 64 sites. We enforce periodic boundary conditions so that the a maximum possible distance between two spins is half the lattice size. To find signals of RG in the correlation functions this maximum distance must be large enough so that the spin-spin correlation has dropped to zero. Our lattice is large enough.

Having described the conditions of our numerical experiment, we consider the correlators $\langle v_i h_a\rangle$ generated when the hidden neurons h_a are generated from the visible neurons v_i using RG. Our goal is to understand the patterns appearing in correlation functions, that are a signature of the RG. In Figure 5d the process of decimation used in our RG is explained. The red dots, representing the input lattice, are averaged (coarse grained) to produce the blue dots which is the lattice after a single step of the RG. The four spin values located at the red dots surrounding each blue dot are averaged and then normalised to ± 1, to obtain the value of the spin at the new (blue) lattice point. This process clearly reduces the number of lattice sites by a factor of four.

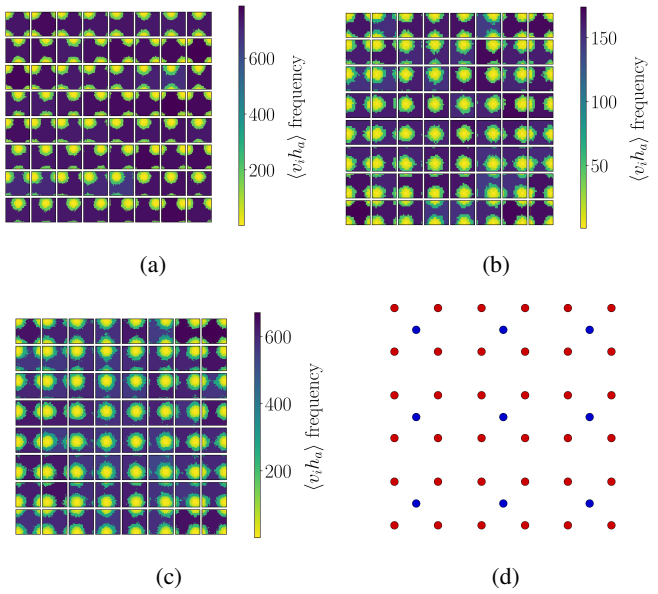


Fig. 5: Frequency correlation plots for Ising input data with lattice size 32 by 32 at T_c and RG decimated Ising data of sizes 16 by 16 (one step of RG) and 8 by 8 (two steps of RG). The colours represent the frequency of correlators occurring. (a) shows input Ising data correlated with configurations resulting after one step of RG. (b) shows correlations between configurations resulting after one step of RG and configurations resulting after two steps of RG. (c) shows correlations between Ising input data and configurations resulting after two steps of RG. (d) shows one step of RG. The red dots show the original input lattice and the blue dots show the lattice obtained after one step of RG. Each blue dot is surrounded by four red dots. The value of the blue dot is determined by averaging the surrounding four red dots.

Using the input data which populates a 32 by 32 lattice, we populate lattices of size 16 by 16 and 8 by 8 spins by applying the RG and then calculate the various possible $\langle v_i h_a\rangle$ correlations. Figure 5a shows the $\langle v_i h_a\rangle$ correlation function that results from a single RG step. To display the matrix of N_h by N_v correlation values, an informative representation is obtained by plotting the frequency with which values in the correlation matrix occur. Each panel of the three figures 5a, 5b and 5c, shows how a given hidden spin is correlated with the visible spins. We can clearly see a peak in correlation values around the spatial location of the hidden spin. This is the signal of RG coarse graining: small spatially localized collections of spins are replaced by their average value. We can go into a

little more detail: the patches of large correlation in Figures 5b and 5c are larger in size than those of figure 5a. This makes sense since each step of the RG implies ever larger spatial regions of the spins are being averaged to produce the coarse grained variables. The fact that the spins that are averaged are spatially localized is a direct consequence of the fact that the Ising model Hamiltonian is local in space so that spatially adjacent spins have similar behaviors. In more general settings it may be harder to decide if the coarse graining is RG-like or not.

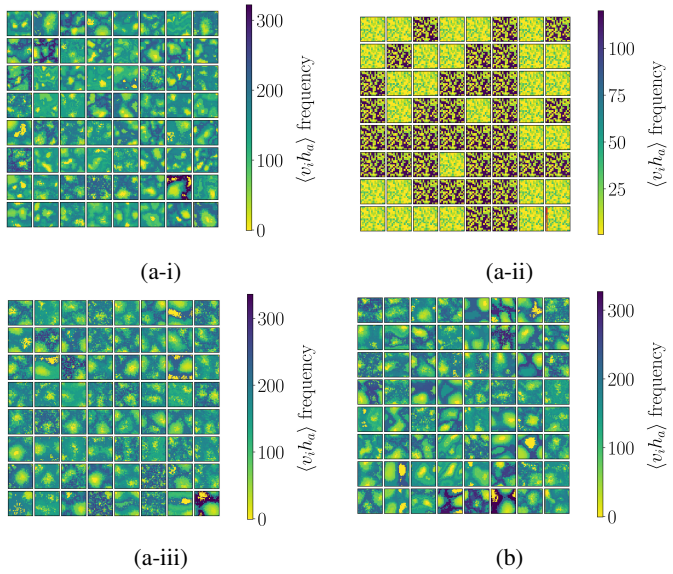


Fig. 6: Plots showing the frequency of correlation values for (a) the stacked RBMs various layers and (b) the single RBM. (a-i) shows correlations between input Ising data (1024 nodes) at T_c and outputs from the first stacked RBM (256 nodes). (a-ii) shows correlations between outputs from the first stacked RBM (256 nodes) and outputs from the second stacked RBM (64 nodes). (a-iii) shows correlations between input Ising data (1024 nodes) at T_c and outputs from the second stacked RBM (64 nodes). (b) shows correlations between input Ising data (1024 nodes) at T_c and outputs from the single RBM (64 nodes).

Having established the signal characteristic of the RG flow, we will now search for this signal in the $\langle v_i h_a \rangle$ correlators computed using the configurations generated from the RBM flow. We consider configurations generated by a stacked network with an RBM with 1024 visible nodes and 256 hidden nodes cascading into a second RBM with 256 visible nodes and 64 hidden nodes. We also consider configurations generated by a single RBM network with 1024 visible nodes and 64 hidden nodes. The factor of 4 relating the number of visible to hidden nodes is chosen to mimic the decimation of lattice sites in each step of the RG. The networks are trained on the same data used as input for the RG considered above. Training is through 10000 steps of contrastive divergence [27].

Figure 6 shows plots for the stacked RBM and figure 6b for the single RBM network. Figure 6a-iii shows the correlation functions and is to be compared to the corresponding RG result in Figure 6b. The two patterns are very similar suggesting that the trained RBM is indeed performing something like the RG coarse graining.

There is one more interesting comparison that can be carried out and it quantitatively tests the flow. The temperature is a relevant coupling so it grows as the flow proceeds. In the block spin RG that we are considering, the length of the lattice keeps halving. In contrast to this, the temperature of the system, which in suitable units has a dimension of inverse length, will roughly double. There will be small departures from precise doubling due to interactions, but the temperature must increase by roughly a factor of 2 as each new layer is stacked. If the RBM is performing an RG-like coarse graining, the temperature should grow in a similar way as we pass through the layers of the deep network. Figure 7a plots the temperature of coarse grained lattices, generated by applying three steps of RG to an input lattice of size 64 by 64, at a temperature of $T = 2.7$. There is a clear increase in the measured temperature as the number of RG steps increase. The temperature of each layer is roughly 2.3, 4.8 and 11 for layers 1, 2 and 3 respectively, which is indeed consistent with the rough rule that the temperature doubles with each step.

Now consider a deep network made by stacking three RBMs. The first network has 4096 visible nodes and 1024 hidden nodes, the second 1024 visible nodes and 256 hidden nodes and the third 256 visible nodes and 64 hidden nodes. The network is trained on Ising data at the critical temperature, as described above. Figures 7b-i, 7b-ii and 7b-iii give the temperatures of the outputs of the layers of the RBM, given input lattices at temperatures of 2.269, 2 and 2.7 respectively. Temperatures of $T = 2$ and $T = 2.269$ lead to the same behavior for the temperature flow, as exhibited in Figures 7b-i and 7b-ii. The temperature jumps rapidly to a high temperature in the first step of the flow, and remains fixed when the second step is taken. This is an important difference that deserves to be understood better. It questions the identification of layers of a deep network with steps in an RG flow.

Figure 7b-iii shows different characteristics to those of 7b-i and 7b-ii. Here the temperature of the input is above T_c at 2.7. Layer 1 is not as sharply peaked near T_c as observed in figures 7b-i and 7b-ii. In addition to this, layers 2 and 3 are not at the same temperature but rather layer 2 is at a higher temperature than layer 3. This is also different to what is observed in the RG measurements where layer 2 is at a lower temperature than layer 3. Figure 7b-iii shows a decrease in temperature from layer 2 to layer 3 rather than an increase. These plots show that a flow through multiple layers in a "deep" network are not consistent with RG.

The results above have shown that the correlator $\langle vh \rangle$ exhibits RG-like characteristics. This is evident from the comparison between the $\langle vh \rangle$ frequency plots from RG, a stacked RBM network and a network with a single RBM. We can see RG-like patterns in the correlators produced by the two RBM networks. This is a promising result that demonstrates that a form of coarse graining is taking place when networks are stacked.

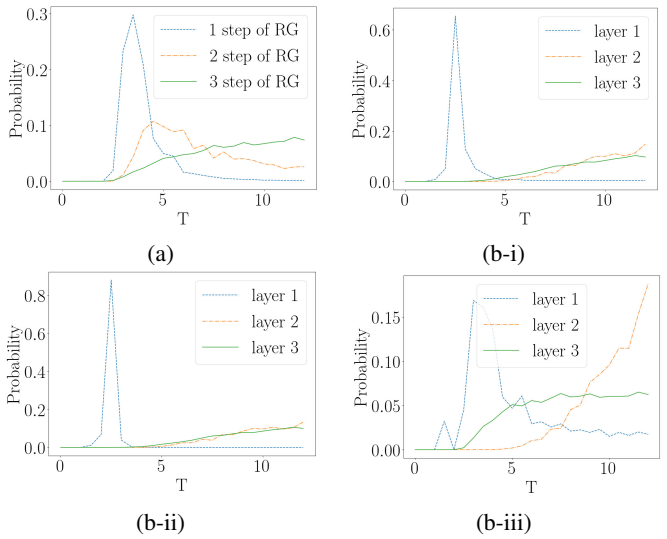


Fig. 7: (a) shows the average probability of the measured temperature of lattices resulting after 3 steps of RG, applied to an input lattice at T_c with 4096 sites. (b) shows the average probability plot of the measured temperature of outputs produced by a stacked RBM with 4096 input nodes, 1024 nodes in the first layer, 256 nodes in the second layer and 64 nodes in the output layer. (b-i) is given input Ising samples at $T = 2.269$, (b-ii) is given input Ising samples at $T = 2$ and (b-iii) is given input Ising samples at $T = 2.7$.

V. Conclusions and Discussion

Our main goal has been to explore the possibility that RG provides a framework within which a theoretical understanding of deep learning can be pursued. Our data set contains the possible states of an Ising magnet, generated using Monte Carlo simulation. This is an interesting data set, since we know that there is a well defined theory for the magnet defined on large length scales. The existence of this long distance theory guarantees that there is some emergent order for the unsupervised learning to identify. Another point worth stressing is that the RG treatment of this system is well understood and is easily implemented numerically. It is therefore an ideal setting in which both deep learning and RG can be implemented and their results can be compared. At the critical temperature, where the system is on the verge of spontaneous magnetization, there is an interesting scale invariant theory which is well understood. By working at this critical point, we have managed to probe the pattern generated by the RBM at different length scales and to compare it to the expected results from an RG treatment.

Our first set of numerical results compare the RBM flow introduced in [15] and further pursued in [16]. From a theoretical point of view the RBM flow looks rather different to RG since the RBM flow appears to drive configurations towards the critical temperature. The RG would drive configurations to ever higher temperatures due to the fact that the temperature corresponds to a relevant perturbation. Another important difference between the RBM flow and RG is that the number of spins is a constant of the RBM flow, but decreases with RG flow. Our numerical results confirm that the RBM flow does indeed generate RG-like Ising configurations and we have reproduced the scaling dimension of the spin variable from the spatial statistics of the patterns generated by the RBM. This is a remarkable result as reported and discussed in [15], [16]. The spin variable has the smallest possible scaling dimensions and consequently probes the largest possible scales in the pattern. When considering correlation functions of the next primary operators we find that the RBM data does not reproduce the correct scaling dimension, proving that the spatial statistics of the patterns generated by the RBM flow and those generated by RG start to differ. We therefore conclude that the RBM flow and RG are distinct, but they do agree on the largest scale structure of the generated patterns. This is a hint into the mechanism behind the RBM flow and it deserves an explanation.

Our second numerical study has explored the idea that deep learning is an RG flow with each stacked layer performing a step of RG. We have explained why correlation functions between the visible and hidden neurons, $\langle v_i h_a \rangle$ are capable of diagnosing RG-like coarse graining and we have computed these correlation functions using the patterns generated by the RBM. The basic signal of RG coarse graining is a "bright spot" in the $\langle v_i h_a \rangle$ correlation function, since this indicates that spins in a localized region were averaged to produce the coarse graining spin. The numerical results do indeed show a dark background with emerging bright spots. It would be interesting if the emergent patterns again guarantee agreement on the largest length scales, similar to what was found for the RBM flows, but we can not confidently make this assertion yet.

Our final numerical study considered the flow of the temperature, a relevant operator according to the RG. We find three distinct behaviors. Section III-A reviewed that RBM flows converge to the critical temperature. This is borne out in our results. The RG flows to ever higher temperatures, with (roughly) a doubling in temperature for each step. Again, this is precisely what we observe. Finally, for a deep network made by stacking three RBMs, the temperature appears to flow when moving between the first and second layers of a deep network, but is fixed when moving between the second and third layers. This is an important difference that deserves to be understood better. It questions the identification of layers of a deep network with steps in an RG flow.

By using Ising model data, generated by Monte Carlo simulation, starting from a local Hamiltonian we know how a coarse graining capable of identifying emergent patterns should proceed: spatially neighboring spins should be averaged. For more general data sets, this may not be the case. It is fascinating to ask what the rules determining the correct coarse graining are and in fact, with respect to this question, deep learning has the potential to shed light on RG.

Apart from the exciting possibility that the link to RG might contribute towards a theoretical understanding of deep learning, one might also ask if the connection would have any practical applications. One possibility that we are currently pursuing, is a Callan-Symanzik like equation governing the learning process. Roughly speaking, one might mimic RG by

dividing the weights to be learned into relevant, marginal and irrelevant parameters, depending on gross statistical properties of the training data. If this classification is itself not too expensive, one could pursue a more efficient approach towards training, since the classification of weights would provide an understanding of which weights are important, and which can simply be set to zero. We hope to report on this possibility in the future.

ACKNOWLEDGEMENT

This work is supported by the South African Research Chairs Initiative of the Department of Science and Technology and National Research Foundation as well as funds received from the National Institute for Theoretical Physics (NITheP). We are grateful for useful discussions to Mitchell Cox and Dimitrios Giataganas.

APPENDIX A
RBM EXPECTATION VALUES

The expectation values quoted in equations (7), (8) and (9) are derived using (2). Data expectation values are evaluated by summing over all samples, $\hat{v_i}^{(A)}$ in the training set. On the other hand model expectation values employ sums over the entire space of visible and hidden vectors. This is such an enormous sum that its numerically intractable. The complete set of expectation values needed to describe the RBM are given by

$$\langle v_i h_a \rangle_{data} = \frac{1}{N_s} \sum_{A=1}^{N_s} \hat{v_i}^{(A)} \tanh\left(\sum_i W_{ia} \hat{v_i}^{(A)} + b_a^{(h)} \right) \quad (51)$$

$$\langle v_i h_a \rangle_{model} = \sum_{\{\mathbf{v},\mathbf{h}\}} \tanh\left(\sum_i W_{ia} v_i + b_a^{(h)} \right) \cdot \tanh\left(\sum_a W_{ia} h_a + + b_i^{(v)} \right) \quad (52)$$

$$\langle v_i \rangle_{data} = \frac{1}{N_s} \sum_{A=1}^{N_s} \hat{v_i}^{(A)} \quad (53)$$

$$\langle v_i \rangle_{model} = \sum_{\{\mathbf{h}\}} \tanh\left(\sum_a W_{ia} h_a + b_i^{(v)} \right) \quad (54)$$

$$\langle h_a \rangle_{data} = \frac{1}{N_s} \sum_{A=1}^{N_s} \tanh\left(\sum_i W_{ia} \hat{v_i}^{(A)} + b_a^{(h)} \right) \quad (55)$$

$$\langle h_a \rangle_{model} = \sum_{\{\mathbf{v}\}} \tanh\left(\sum_i W_{ia} v_i + b_a^{(h)} \right) \quad (56)$$

with $\hat{v}_i^{(A)}$ the Ath sample of the data set, $\hat{\mathbf{v}}$.

APPENDIX B
TWO VERSIONS OF RG

In this section we review two versions of the RG that are needed in this article. The first of these, the variational renormalization group, was introduced by Kadanoff [33]–[35] as a method to approximately perform the renormalization group in practise.

A. Variational RG

Consider a system of N spins $\{v_i\}$ which each take the values ± 1. The partition function describing the system is given by

$$Z = \sum_{v_i} e^{-H(\{v_i\})} \quad (57)$$

Here the sum is over all possible configurations of the system of spins and the function $H(\{v_i\})$, called the Hamiltonian, gives the energy of the system. This would include the energy of each individual spin as well as the energy associated to the fact that the collection of spins is interacting. The Hamiltonian $H(\{v_i\})$ can be an arbitrarily complicated function of the spins

$$H(\{v_i\}) = -\sum_i K_i v_i - \sum_{i,j} K_{ij} v_i v_j - \sum_{i,j,k} K_{ijk} v_i v_j v_k + \cdots \quad (58)$$

The RG flows maps the original Hamiltonian to a new Hamiltonian with a different set of coupling constants. The new Hamiltonian

$$H(\{h_a\}) = -\sum_a K'_a h_a - \sum_{a,b} K'_{ab} h_a h_b - \sum_{a,b,c} K'_{abc} h_a h_b h_c + \cdots \quad (59)$$

gives the energy for the coarse grained spins h_a. After many RG iterations many coupling constants (the so called irrelevant terms) flow to zero. A much smaller number may remain constant (marginal terms) or even grow (relevant terms). To implement this conceptual framework a concrete RG mapping is needed. Variational RG provides a mapping which is not exact but can be implemented numerically. It does this by introducing an operator $T_\lambda(\{v_i, h_a\})$ which is a function of a set of parameters $\{\lambda\}$. The Hamiltonian after a step of RG flow is

$$e^{-H_{RG}(\{h_a\})} = \sum_{v_i} e^{T_\lambda(\{v_i,h_a\}) - H(\{v_i\})} \quad (60)$$

The form of $T_\lambda(\{v_i, h_a\})$ must be chosen cleverly, for each problem we consider. This is the tough step in variational RG and it is carried out using physical intuition, but essentially on a trial and error basis. Once a given $T_\lambda(\{v_i, h_a\})$ has been chosen, we minimize the following quantity by choosing the parameters $\{\lambda\}$

$$\log(\sum_{v_i} e^{-H(\{v_i\})}) - \log(\sum_{h_a} e^{iH_{RG}(\{h_a\})}) \quad (61)$$

The minimum possible value for this quantity is zero. Notice that when

$$\sum_{h_a} e^{T_\lambda(\{v_i,h_a\})} = 1 \tag{62}$$

(61) attains its minimum value of 0 and the RG transformation is called exact.

B. Block Spin Averaging

Block spin averaging is a pedagogical version of RG. To illustrate the method, consider a rectangular lattice of interacting spins. Divide the lattice into blocks of 2×2 squares. Block spin averaging describes the system in terms of *block variables*, which are variables describing the average behavior of each block. The "block spin" is literally the average of the four spins in the block. The plots shown in Figures 5a use block spin averaging. The block spins h_a and averages of the spins v_i.

REFERENCES

[1] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[2] L. Deng, D. Yu *et al.*, "Deep learning: methods and applications," *Foundations and Trends® in Signal Processing*, vol. 7, no. 3–4, pp. 197–387, 2014.

[3] Y. Bengio *et al.*, "Learning deep architectures for ai," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[4] N. Le Roux and Y. Bengio, "Deep belief networks are compact universal approximators," *Neural computation*, vol. 22, no. 8, pp. 2192–2207, 2010.

[5] N. Le Roux and Y. Bengio, "Representational power of restricted boltzmann machines and deep belief networks," *Neural computation*, vol. 20, no. 6, pp. 1631–1649, 2008.

[6] Y. Bengio, Y. LeCun *et al.*, "Scaling learning algorithms towards ai," *Large-scale kernel machines*, vol. 34, no. 5, pp. 1–41, 2007.

[7] A. Paul and S. Venkatasubramanian, "Why does deep learning work?-a perspective from group theory," *arXiv preprint arXiv:1412.6621*, 2014.

[8] Y. Bengio, A. C. Courville, and P. Vincent, "Unsupervised feature learning and deep learning: A review and new perspectives," *CoRR, abs/1206.5538*, vol. 1, p. 2012, 2012.

[9] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," *arXiv preprint arXiv:1611.03530*, 2016.

[10] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Advances in neural information processing systems*, 2007, pp. 153–160.

[11] H. W. Lin, M. Tegmark, and D. Rolnick, "Why does deep and cheap learning work so well?" *Journal of Statistical Physics*, vol. 168, no. 6, pp. 1223–1247, 2017.

[12] K. G. Wilson and J. Kogut, "The renormalization group and the ϵ expansion," *Physics reports*, vol. 12, no. 2, pp. 75–199, 1974.

[13] P. Mehta and D. J. Schwab, "An exact mapping between the variational renormalization group and deep learning," *arXiv preprint arXiv:1410.3831*, 2014.

[14] M. Koch-Janusz and Z. Ringel, "Disordered Systems and Neural Networks Mutual Information, Neural Networks and the Renormalization Group," *arXiv preprint arXiv:1704.06279*, 2017.

[15] S. Iso, S. Shiba, and S. Yokoo, "Scale-invariant feature extraction of neural network and renormalization group flow," *Physical Review E*, vol. 97, no. 5, p. 053304, 2018.

[16] S. S. Funai and D. Giataganas, "Thermodynamics and feature extraction by machine learning," *arXiv preprint arXiv:1810.08179*, 2018.

[17] C. Bny, "Deep learning and the renormalization group," *arXiv preprint arXiv:1301.3124 [quant-ph]*, Jan. 2013.

[18] B. M. McCoy and T. T. Wu, *The two-dimensional Ising model*. Courier Corporation, 2014.

[19] J. Carrasquilla and R. G. Melko, "Machine learning phases of matter," *Nature Physics*, vol. 13, no. 5, p. 431, 2017.

[20] A. Morningstar and R. G. Melko, "Deep learning the ising model near criticality," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 5975–5991, 2017.

[21] S. Saremi and T. J. Sejnowski, "Hierarchical model of natural images and the origin of scale invariance," *Proceedings of the National Academy of Sciences*, vol. 110, no. 8, pp. 3071–3076, 2013.

[22] P. Smolensky, "Information processing in dynamical systems: Foundations of harmony theory," Colorado Univ at Boulder Dept of Computer Science, Tech. Rep., 1986.

[23] Y. Freund and D. Haussler, "Unsupervised learning of distributions on binary vectors using two layer networks," in *Advances in neural information processing systems*, 1992, pp. 912–919.

[24] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[25] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the national academy of sciences*, vol. 79, no. 8, pp. 2554–2558, 1982.

[26] G. E. Hinton, "A practical guide to training restricted boltzmann machines," in *Neural networks: Tricks of the trade*. Springer, 2012, pp. 599–619.

[27] M. A. Carreira-Perpinan and G. E. Hinton, "On contrastive divergence learning." in *Aistats*, vol. 10. Citeseer, 2005, pp. 33–40.

[28] T. Tieleman, "Training restricted boltzmann machines using approximations to the likelihood gradient," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 1064–1071.

[29] R. Salakhutdinov, A. Mnih, and G. Hinton, "Restricted boltzmann machines for collaborative filtering," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 791–798.

[30] A. Krizhevsky, G. Hinton *et al.*, "Factored 3-way restricted boltzmann machines for modeling natural images," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 621–628.

[31] I. Sutskever and T. Tieleman, "On the convergence properties of contrastive divergence," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 789–795.

[32] D. Poland, S. Rychkov, and A. Vichi, "The conformal bootstrap: Theory, numerical techniques, and applications," *Reviews of Modern Physics*, vol. 91, no. 1, p. 015002, 2019.

[33] E. Efrati, Z. Wang, A. Kolan, and L. P. Kadanoff, "Real-space renormalization in statistical mechanics," *Reviews of Modern Physics*, vol. 86, no. 2, p. 647, 2014.

[34] L. P. Kadanoff, *Statistical physics: statics, dynamics and renormalization*. World Scientific Publishing Company, 2000.

[35] L. P. Kadanoff, A. Houghton, and M. C. Yalabik, "Variational approximations for renormalization group transformations," *Journal of Statistical Physics*, vol. 14, no. 2, pp. 171–203, 1976.

Ling Cheng (M10-SM15) received the degree B. Eng. Electronics and Information (cum laude) from Huazhong University of Science and Technology (HUST) in 1995, M. Ing. Electrical and Electronics (cum laude) in 2005, and D. Ing. Electrical and Electronics in 2011 from University of Johannesburg (UJ). His research interests are in Telecommunications and Artificial Intelligence. In 2010, he joined University of the Witwatersrand where he was promoted to Associate Professor in 2015. He has served as the Vice-chair of IEEE South African Information Theory Chapter. He has been a visiting professor at five universities and the principal advisor for over forty full research post-graduate students. He has published more than 80 research papers in journals and conference proceedings. He was awarded the Chancellors medals in 2005, 2019 and the National Research Foundation rating in 2014. The IEEE ISPLC 2015 best student paper award was made to his Ph.D. student in Austin.

Ellen de Mello Koch obtained her BSc(Eng) degree in 2014 and MSc(Eng) degree in 2016 in Electrical Engineering at the University of the Witwatersrand. She is currently a Ph.D. student and lecturer in the department of Electrical and Information Engineering. Her doctoral research investigates the link between deep learning and the renormalization group, as an attempt to develop a theoretical framework for deep learning. Her research interests lie in unsupervised learning algorithms and their application to the real world.

Robert de Mello Koch obtained his BSc(Eng) degree in 1992, his BscHon(Phys) in 1993, his MSc(Phys) in 1994 and his PhD (Phys) in 1998, all from the University of the Witwatersrand. He is a Professor at the University of the Witwatersrand where he holds the DST/NRF Research Chair in Fundamental Physics and String Theory and is a Distinguished Visiting Professor at South China Normal University. He is a fellow of the Durham Institute for Advanced Studies, the Stellenbosch Institute for Advanced Studies and the Academy of Science of South Africa. His most recent research interests include the gauge theory/string theory duality, the application of representation theory of discrete groups and Lie groups to quantum field theory and deep learning.