An almost-exact map between the real-space renormalization group and restricted Boltzmann machines

Ziwei Li
Department of Earth, Atmospheric and Planetary Sciences
Massachusetts Institute of Technology
ziweili@mit.edu

Abstract

A series of recent papers attempted to understand the good performance of deep neural networks from the perspective of the Renormalization Group, a profound theoretical tool in physics. This report gives a critical review of the origination of this idea, current hypothesis, and numerical evidences of whether this connection is correct. Furthermore, we will use the inherent symmetries of the restricted Boltzmann machine and the Ising model, and show an almost-exact map between the two systems through analytical and numerical considerations.

1 Introduction

Neural networks have been widely used in image recognization, natural language processing, and early forms of artificial intelligent machines. It's been regarded and used as *the* machine learning algorithm because of its architectural flexibility and learning abilities. However, theoretical understanding of why this algorithm works in so many different types of problems is still limited. A prime example is the universal approximation theorem and its variations. Although it states that the functional class represented by neural networks are dense in a finite domain provided with a large number of neurons N, it does not give any reasonable constraint on N. In fact, the no-free-lunch theorem makes this theorem practically irrelevant in the real world where one has only limited storage and computational power. Furthermore, it still remains an open question why for some problems, deep networks work much better than shallow networks, i.e., require far less degrees of freedom given a fixed error tolerance.

The idea of the connection of RG and neural networks originated from the similarity between multiple stages of RG of decreasing degrees of freedom, and the layered structure of NN. Beny, 2013 proposed that multi-body problems in the quantum limit can be sufficiently learned by an algorithm which has a hierarchical "knowledge representation" that parameterizes the probability of observations. Provided with a suitable algorithm that minimizes certain cost function, the knowledge of the physical system can be learned from data. Although only a philosophical stand point was provided in the paper without numerical results or even referring to a particular algorithm, it was arguably the first paper to draw a conceptual analogy of RG and deep learning. Mehta and Schwab, 2014 proposed that, as their title suggests, an exact mapping between the Kanadoff's variational renormalization group and restricted Boltzmann machines (RBM). Their claim was a bit controversial: on one hand the "exactness" depend on the fact that condition (8) in the paper is satisfied, which awaits to be achieved via neural network; on the other hand, training RBMs using contrastive divergence were argued to be inappropriate (e.g., supplement in Koch-Janusz and Ringel, 2018).

MIT 9.520/6.860: Statistical Learning Theory and Applications (Fall 2016).

2 The setting of the Ising model and the restricted Boltzmann machine

The Ising model is a classical statistical model for the ferromagnetic phase transition. It is a lattice spin model with binary nearest-neighbor interactions between spins; its Hamiltonian is given by

$$H(\{v_i\}) = \sum_{n.n.} K v_i v_{i'}, \tag{1}$$

where $\{v_i\}$ is the *configuration* of spins, K represent the strength of interaction, and the summation is over nearest neightbors ($n.n.$). We denote the number of spins as N_v. We will constrain our discussion on 2-dimensional Ising model, in which there exists a critical temperature T_c such that the system is infinitely unstable: a small and local disturbance will result in large-scale responses (infinite susceptibility); and the correlation length is infinite (scale-free). The renormalization group (RG) approach is designed for such phenomena. Since one is interested in long-wavelength behaviors, the so-called RG flow can integrate out, or marginalize over, local degrees of freedom step by step, and leave essential quantities of physical relevance (order parameters).

Assuming thermo-equilibrium, The probability of $\{v_i\}$ is coded by the Boltzmann distribution

$$P(\{v_i\}) = \frac{1}{Z} e^{-H(\{v_i\})} = \frac{1}{Z} \exp\left(-\sum_{n.n.} K v_i v_{i'}\right), \tag{2}$$

where Z is the partition function, or the normalizing constant, which is the summation of $e^{-H(\{v_i\})}$ over all possible $\{v_i\}$. Equation 2 is hard to marginalize over, or extract useful statistical information because of the curse of dimensionality, in that one has to do 2^{N_v} summations to calculate Z and infer what the probability is for each configuration $\{v_i\}$. However, because of the compactness of the Hamiltonian $H(\{v_i\})$, we hope to find a set of other parameters $\{h_j\}$, such that it represents the physical system with reduced degrees of freedom. The physical approach is the real-space renormalization group proposed by Wilson, 1975, or the variational version by Kadanoff, 1978. A machine learning approach can be using the restricted Boltzmann machine (RBM) to find a reduced representation of the system; other learning schemes such as deep belief networks [Morningstar and Melko, 2018] and bijective map [Li and Huang, 2018] were used as well.

We will use RBM to approximate the probability distribution of the Ising model in our analysis. A RBM comprises a visible layer $\{v_i\}$ of number N_v, and a hidden layer $\{h_j\}$ of number N_h. The joint probability distribution of $\{v_i\}$ and $\{h_j\}$ is encoded by

$$P(\{v_i\},\{h_j\}) = \frac{1}{Z_{RBM}} e^{-E_\theta(\{v_i\},\{h_j\})}, \tag{3}$$

where $E_\theta(\{v_i\},\{h_j\})$ is called *energy function* of the RBM, and Z_{RBM} is the normalizing constant. It takes the form

$$E_\theta(\{v_i\},\{h_j\}) = \sum_j (b_j h_j + C_j) + \sum_{i,j} w_{ij} h_j v_i + \sum_i c_i v_i. \tag{4}$$

The subscript θ refers to the parameter set $\{b_j, C_j, w_{ij}, c_i\}$ of the RBM. Note that we introduced an extra term C_j compared to the traditional definition, the reason of which will become obvious in our later discussion. The RBM is usually trained using the contrastive divergence algorithm, in which the KL divergence of the probability distribution of the visible layer, $P_\theta(\{v_i\})$ and the probability of training data is minimized. In practice, the training data are probed by sampling from 2. The distribution over $\{v_i\}$ is sampled by running a Markov Chain over 3.

$P_\theta(\{v_i\})$ can also be calculated from marginalizing over all hidden spins. If a RBM is performing RG transformation, the hidden layer should ideally have local interactions, and take binary values ± 1 just like $\{v_i\}$. Taking this into account, we can marginalize over the hidden spins

$$P_\theta(\{v_i\}) = \frac{1}{Z} \exp\left[-\sum_i c_i v_i + \sum_j \ln 2\cosh(\sum_i w_{ij} v_i + b_j) - \sum_j C_j\right] \tag{5}$$

It's interesting in equation 5 that by introducing conditional independence of the hidden units from visible units, the marginalization becomes surprisingly easy. The above equation is similar to equation (5) in [Huang and Wang, 2016]. Note that in general the binary value assumption of $\{h_j\}$ is not true, because RBM is only trained to learn the distribution of $\{v_i\}$, and is not given any requirement on the behaviors of the hidden units. However, this is a very important criterion regarding whether neural network is performing RG, and one can easily force $\{h_j\}$ to take binary values during sampling.

3 Some analytical considerations

The perfect RG achieves: 1) the Helmholtz free energy is conserved [Mehta and Schwab, 2014]; 2) the transformed Hamiltonian is compact and local, in a sense that only local interactions of the hidden variables are present. We show in this section that there exists an analytical solution of RBM, whose visible layer reconstructs the probability distribution of the Ising model *almost exactly*, and has a compact Hamiltonian for the hidden variables as well. The free energy is also approximately conserved in our solution.

In order to reconstruct the probability of Ising model faithfully, we need equation 5 to involve only nearest-neighbor interactions of visible spins $\{v_i\}$. At the same time, we can also marginalize equation 3 over the visible spins to obtain the distribution over hidden spins. We also require that hidden spins have local and only nearest-neighbor interactions. The marginalized distribution of the hidden spin is given by

$$P_\theta(\{h_j\}) = \frac{1}{Z}\exp\left[-\sum_j b_j h_j + \sum_i \ln 2\cosh(\sum_j w_{ij}h_j + c_i) - \sum_j C_j\right] \tag{6}$$

It seems that there's no way the additive terms in equations 5 and 6 to feature multiplicative interactions. However, this is not true because, as is shown in [Lin et al., 2017], additive terms can capture multiplicative functions when variables are discrete.

Here we pursue an approximate analytical solution on the visible layer. Our first consideration is symmetry. Given a configuration $\{v_i\}$, its probability should be the same as when all the spins are flipped, i.e., $\{-v_i\}$. Therefore we have from equation 5:

$$-\sum_i c_i v_i + \sum_j \ln 2\cosh(\sum_i w_{ij}v_i + b_j) = \sum_i c_i v_i + \sum_j \ln 2\cosh(-\sum_i w_{ij}v_i + b_j). \tag{7}$$

C_j does not come into this expression because they are canceled from both sides. This gives

$$\sum_j \ln\frac{\cosh(\sum_i w_{ij}v_i + b_j)}{\cosh(-\sum_i w_{ij}v_i + b_j)} = 2\sum_i c_i v_i \tag{8}$$

for all configurations. This is 2^{N_v} equations for $N_v + N_h + N_v N_h$ unknowns, which is overdetermined. However, looking at it again, one notice that if $c_i = 0, b_j = 0$, we can then leverage the symmetry of cosh function: $\sum_j \ln\frac{\cosh(\sum_i w_{ij}v_i)}{\cosh(-\sum_i w_{ij}v_i)} \equiv 0$. The first consideration is then satisfied.

Our second consideration is that the RBM has to perform a kind of "decimation", in that it reduces the degrees of freedom in a structured fashion. Suppose we have a 2D Ising grid of 8-by-8 spins of doubly-periodic condition. From equation 5, we have N_h components inside the exponent:

$$\ln 2\cosh(\sum_i w_{ij}v_i) - C_j. \tag{9}$$

Note that the biases b_j are already set to zero. We view this as N_h filters that take information from the visible spins. We assign $N_h = N_v/2 = 32$ filters to this system. Each filter takes a particular visible spin in the visible layer, and its four nearest neighbors (see left panel, figure 1). We think of vector $w_{:,j}$ as a 2D matrix multiplied on the matrix of the visible spins element-wise. Then all weights w_{ij} can be viewed as 32 matrices of size 8-by-8 (see right panel, figure 1). This way, we arrive at a much simpler expression of 9:

$$\begin{aligned}&\ln 2\cosh(w^{(j)}_{k,l}v_{k,l} + w^{(j)}_{k-1,l}v_{k-1,l} + w^{(j)}_{k+1,l}v_{k+1,l} + w^{(j)}_{k,l-1}v_{k,l-1} + w^{(j)}_{k,l+1}v_{k,l+1}) - C_j\\ &= \ln 2\cosh(w_1(v_{k-1,l} + v_{k+1,l} + v_{k,l-1} + v_{k,l+1}) + w_0 v_{k,l}) - C_j,\end{aligned} \tag{10}$$

where indices k, l indicate the position of the visible spin in 2D, and j represents the index of the filter. The second step of the above expression is due to the symmetry that all four nearest neighbors of the center spin $v_{k,l}$ should be treated equally (rotational symmetry). We also require that w_0 and w_1 should not be a function of j (translational symmetry). A schematic of this consideration is illustrated in the right panel in figure 1. All situations considered, there are 5 different values that

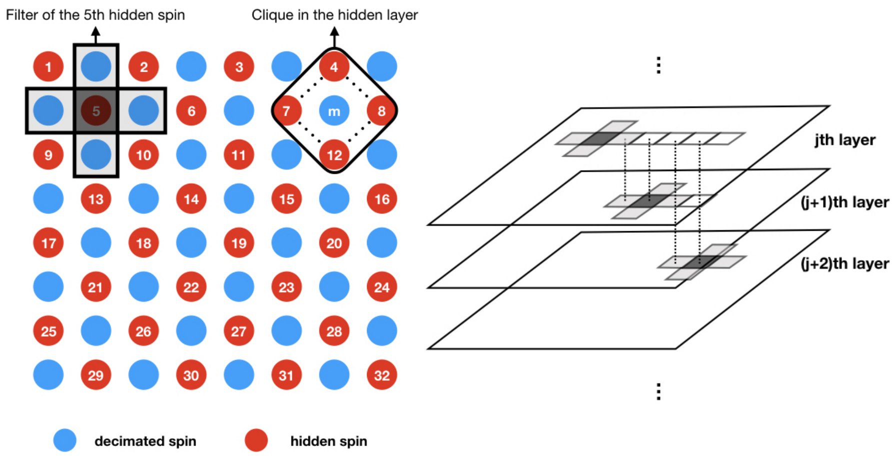

Figure 1: **Left panel**, configuration of the Ising model. Blue dots are the decimated spins after the transformation process, and red dots are hidden spins that remained after the transformation. The combination of blue and red dots are visible spins. The cross-like window represent a filter for, e.g., the 5^{th} hidden spin. The structure circled by the rounded rectangle on the right represent the four-spin clique in the visible layer centered around the visible spin labeled m. **Right panel**, layered structure of the weight matrices $w_{k,l}^{(j)}$.

expression 10 can take, and it should be at least approximately equal to the energy value given by the true Hamiltonian:

$$\begin{aligned}
\ln 2\cosh(4w_1 + w_0) - C_j &\approx 4K; & +,+,+,+,+ \\
\ln 2\cosh(2w_1 + w_0) - C_j &\approx 2K; & +,-,+,+,+ \\
\ln 2\cosh(w_0) - C_j &\approx 0; & +,-,-,+,+ \\
\ln 2\cosh(-2w_1 + w_0) - C_j &\approx -2K; & +,-,-,-,+ \\
\ln 2\cosh(-4w_1 + w_0) - C_j &\approx -4K; & +,-,-,-,-
\end{aligned} \tag{11}$$

The plus sign and minus signs correspond to the signs of spins $v_{k,l}, v_{k-1,l}, v_{k+1,l}, v_{k,l-1}, v_{k,l+1}$,

$\boxed{e^{-w}} + e^{w} = 2\cosh w$

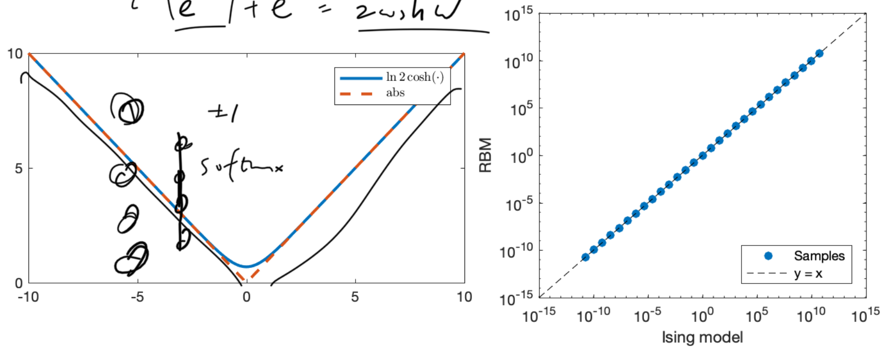

Figure 2: **Left panel**, comparison between function $\ln 2\cosh$ (blue) and the absolute function (red): $\ln 2\cosh(\cdot)$ can be seen as an approximation to the absolute function, if the absolute value of its argument is large enough. **Right panel**, a comparison of probability values of different configurations calculated using Ising model and the marginalized distribution of RBM on the visible layer (equation 12). 10^6 configurations are plotted. Each configuration is generated by individually sample each spin from a uniform distribution. The fractional error of RBM predictions compared to the Ising model is less than 1×10^{-12} for all shown configurations.

respectively. Note that if all spins flip sign, the above equation still holds, so that equations 11 accounts for all configurations of the five spins at question. Looking at the behavior of function $\ln 2\cosh(\cdot)$ (left panel, figure 2), we know that it's approximating the absolute value function when the absolute value of its input is large enough. As a first trial, we let $C_j = \ln 2\cosh(w_0)$ for all j,

$w_0 = 40$, and $w_1 = K$. It should be noted that the exact value of w_0 is not important, we are just leveraging the fact that function $\ln 2\cosh(\cdot)$ approximates the absolute function; in fact, the larger w_0 is, the closer the approximation will be. The expression of RBM over the visible layer is then

$$P_\theta(\{v_i\}) = \frac{1}{Z}\exp\left[\sum_j \ln 2\cosh(\sum_i w_{ij}v_i) - N_h \ln 2\cosh(w_0)\right], \tag{12}$$

where $\sum_i w_{ij}v_i$ is given by expression 10 with $w_0 = 40$, $w_1 = K$. This marginalized distribution can *almost exactly* reproduce the probability of each configuration $\{v_i\}$ (see right panel of figure 2). The fractional error is less than 10^{-12} for all configurations tested.

We have shown that RBM can *almost exactly* approximate the probability distribution of the Ising model; a more pressing question is, what is the behavior of the hidden layer? Answering this question will help us assess whether RBMs can perform the RG transformation. Rewriting equation 6:

$$P_\theta(\{h_j\}) = \frac{1}{Z}\exp\left[\sum_i \ln 2\cosh(\sum_j w_{ij}h_j) - N_h \ln 2\cosh(w_0)\right] \tag{13}$$

The expression $\ln 2\cosh(\sum_j w_{ij}h_j)$ means that for each visible spin indexed i, we collect hidden spins that are associated with it by the filters. There are two kinds of visible spins: the ones that were "decimated" out, and the ones that remained and became a hidden spin. The ones that remain only give $\ln 2\cosh(w_0 h_j) \equiv \ln 2\cosh(w_0)$; when added up, they cancel $-N_h \ln 2\cosh(w_0)$ in the exponent. The ones that are "decimated" out would form a clique, for example, see left panel of figure 1,

$$\ln 2\cosh(w_1(h_4 + h_7 + h_8 + h_{12})) \tag{14}$$

Surprisingly, this is just the result of integrating out the spin in the middle of the four hidden spins (noted by m in figure 1) from the original probability distribution! To be more explicit, we write out expression 14 as:

$$\begin{aligned}\ln 2\cosh(w_1(h_4 + h_7 + h_8 + h_{12})) &= e^{w_1(h_4+h_7+h_8+h_{12})} + e^{-w_1(h_4+h_7+h_8+h_{12})} \\ &= \sum_{v_m=\pm1} \exp\left(-w_1(h_4 + h_7 + h_8 + h_{12})v_m\right) \\ &= \sum_{v_m=\pm1} \exp\left(-K\sum_{n.n.} v_i v_m\right).\end{aligned} \tag{15}$$

This mechanism is similar to the example in chapter 14.2.C, Pathria, 2011. The above analysis means that, not only does the RBM almost exactly reproduces the probability distribution of the Ising model in the visible layer, it's performing an *accurate* real-space Renormalization Group transformation in the hidden layer. In the transformed hidden spin system, the interactions of the spins are quadruple instead of nearest-neighbor, which means that it will faithfully produce the probability distribution of the physical system made up of the remaining spins.

4 Conclusion and discussions

We showed that the restricted Boltzmann machine has a solution that almost exactly performs the real-space renormalization group transformation of the Ising model. Amazingly, the solution not only faithfully captures the probability distribution of the Ising model when marginalized on the visible layer, but also performs one decimation step over to the hidden layer. This proves a connection between the renormalization group transformation and the restricted Boltzmann machine. It will be interesting to see the generated RG flow if this process is performed multiple times.

With the aid of the symmetries in RBM and the Ising model, the solution in this paper is reduced to only two parameters, w_0 and w_1, in which w_0 is only used to ensure that $\ln 2\cosh$ approximates the absolute value function. This shows an interesting paradigm of machine-learning from physical systems: leveraging the inherent symmetries in learning algorithms and physical systems will greatly reduce the number of parameters. Conversely, one can expand the number of parameters in a structured way, just like multiple filters in convolutional neural networks, so that RBM can incorporate more than nearest-neighbor interactions. One can imagine how powerful RBMs can be, if more parameters were used in this "simple" system.

References

[1] Beny, C. (2013) Deep learning and the renormalization group. *arXiv*:1301.3124v4.

[2] Huang, L. & Wang, L. (2016) Accelerate Monte Carlo Simulations with Restricted Boltzmann Machines. *arXiv*:1610.02746v2.

[3] Iso, S., Shiba, S. & Yokoo, S. (2018) Scale-invariant feature extraction of neural network and renormalization group flow. *Physical Review E*, **97**:053304.

[4] Kadanoff, L.P. (1976) Variational Approximations for Renormalization Group Transformations. *Journal of Statistical Physics*, **14**(2):171-203.

[5] Koch-Janusz, M. & Ringel, Z. (2018) Mutual information, neural networks and the renormalization group. *Nature Physics*, **15**:578-582.

[6] Li, S.H., & Wang, L. (2018) Neural Network Renormalization Group. *arXiv*: 1802.02840v3.

[7] Lin, H.W., Tegmark, M. & Rolnick, D. (2017) Why does deep and cheap learning work so well? *Journal of Statistical Physics* **168**(6):1223-1247.

[8] Mehta, P. & Schwab D.J. (2014) An exact mapping between the Variational Renormalization Group and Deep Learning. *arXiv*:1410.3831v1.

[9] Morningstar, A., & Melko, R.G. (2018) Deep Learning the Ising Model Near Criticality. *Journal of Machine Learning Reseach*, **18**(1):1-17

[10] Pathria, R.K. & Beale, P.D. (2011) *Statistical Mechanics (Third Edition)* Academic Press, ISBN: 9780123821881.

[11] Poggio, T., Mhaskar, H., Rosasco, L. et al. (2017) Why and When Can Deep - but Not Shallow - Networks Avoid the Curse of Dimensionality: a Review. *International Journal of Automation and Computing*, **14**(5):503-519.

[12] Saremi, S. & Sejnowski, T.J. (2013) Hierarchical model of natural images and the origin of scale invariance. *Proceedings of the National Academy of Sciences*, **110**(8):3071-3076

[13] Wilson, K. (1975) The renormalization group: Critical phenomena and the Kondo problem. *Reviews of Modern Physics*, **47**(4):773-840.

[14] Zeiler M.D., & Fergus R. (2014) Visualizing and Understanding Convolutional Networks. In *Fleet, D. et al., (Eds): Computer Vision - ECCV 2014. LNCS*. **8689**:818-833.

A Previous works

In order to understand the performance of deep neural networks, considering the structural properties seems to be a good way to go. A strong difference between deep and shallow networks is that deep networks usually incorporates multiple layers of neurons. Zeiler and Fergus, 2013 showed evidence that deep convolutional neural networks learns classification of images via extracting feature information hierarchically. The first few layers of the convolutional neural network learns to extract reasonable filters to detect edges and brightness of images, and latter fully-connected layers that operate on top of the previous layers learns more abstract connections of the low-level features. Saremi and Sejnowski, 2012 proposed a novel view that natural images incorporates scale-invariance and can be decomposed into a hierarchy of layers, which supported the hypothesized connection of the renormalization group (RG) and neural networks.

Numerical results showed empirical connection between the RG and neural networks. Most of the works used Ising model to illustrate this connection. These results differ somewhat largely, mainly because of their different approaches to the problem. It's therefore crucial to take a close look at their choices of architecture. Koch-Janusz and Ringel, 2018 trained RBM through maximizing the mutual information between visible spins and showed that it can perform RG on the Ising model and the dimmer model. They used a localized RBM as a filter to "distill" information from the visible layer to the hidden layer, which is a stage of RG. This approach is very similar to the weight sharing concept used in convolutional neural networks, in that it uses the translational invariance of the physical system. They found that the RBM filter is very similar to Kanadoff's block spin model when trained on the Ising model.

Iso et al., 2018 found that RBMs trained using contrastive divergence is performing a transformation that is opposite to the RG flow. This is a very interesting yet confusing result. The authors used a global RBM, which acts to transform any observed state $\{v_i\}$ into the average value of its hidden counterparts $\{\langle h_i \rangle\}$, and then calculate the visible spin from these hidden units. The transformed state $\{v_i'\}$ is of the same dimensionality of the previous $\{v_i\}$, hence this type of transformation can be iterated to reach a steady state. They used data from a range of temperatures around the critical temperature T_c. Hence the RBM incorporates the full probability distribution of the training data. It's argued in the paper that the internal dynamics of the one layer network form an attractor that drives the inputs of a higher/lower temperature to the critical temperature. But this is expected because if the RBM simulates the probability distribution well, the attracting point will be somewhere near the highest probability in the temperature space. Therefore, given an input from a low/high temperature, the iterations performed with the RBM will follow an attracting trajectory to the attractor. This also explains other results in the paper: low temperature configurations will be transformed to high temperature if the RBM is trained on data points sampled from high temperature only, and vice versa.

Morningstar and Melko, 2018 looked specifically at whether generative neural networks can reproduce the critical behavior of the Ising model (e.g., the magnetic susceptibility at the critical temperature T_c). They found, interestingly, that the most efficient representation of 2D Ising model is a single RBM with only one hidden layer. This partly motivates my analytical treatment in the next section to RBM. Another line of thinking that I drew inspiration from focuses on compositional functions [Poggio et al., 2017 and Lin et al., 2017]. Their argument is that the hierarchical architecture of neural networks are helps to approximate functions that are compositional.