A Renormalization Group Framework for Scale-Invariant Feature Learning in Deep Neural Networks (Student Abstract)

Sarah Liaw

Computing and Mathematical Sciences, California Institute of Technology
sliaw@caltech.edu

Abstract

We propose a framework that uses renormalization group (RG) theory from statistical physics to analyze and optimize the hierarchical feature learning process in deep neural networks. Here, the layer-wise transformations in deep networks can be viewed as analogous to RG transformations, with each layer implementing a coarse-graining operation that extracts increasingly abstract features. We propose an approach to enforce scale invariance in neural networks (NNs), introduce scale-aware activation functions, and derive RG flow equations for network parameters. We show that our approach leads to fixed points corresponding to scale-invariant feature representations. Finally, we propose an RG-guided training procedure that converges to these fixed points while minimizing the loss function.

Introduction

Deep neural networks have achieved success across various domains (image classification, sentiment analysis etc.), yet our understanding of their feature learning process remains limited. The process, fundamental to the networks' performance, involves the automatic discovery and extraction of relevant patterns from raw input data without explicit feature engineering. The hierarchical nature of this learning, where lower layers detect simple features and higher layers combine these into more complex representations, is believed to underpin the network's ability to handle complex tasks and generalize to unseen data (Cagnetta et al. 2024).

This hierarchical structure has similarities to renormalization group (RG) transformations in statistical physics, a framework for understanding system behavior across different scales. RG theory, which is used to critical phenomena in physical systems, provides a systematic approach to analyze scale-dependent behavior. Iterative coarse-graining of a system eliminates fine-grained details while preserving essential large-scale structures, thereby revealing scale-invariant properties and universal behaviors.

While previous works have explored the connection between deep learning and RG theory (Iso, Shiba, and Yokoo 2018; Erdmenger, Grosvenor, and Jefferson 2021), these approaches have largely been qualitative or limited to specific architectures. We establish a framework that unifies deep learning and RG theory to better understand the feature learning process and enable the design of robust and generalizable neural networks (NNs).

The primary motivations for this work are: (1) to develop NN architectures with inherent scale invariance, improving their ability to recognize patterns at multiple scales; (2) to improve the interpretability of deep networks, providing insights into their decision-making processes; (3) to improve the generalization capabilities of NNs, particularly with respect to out-of-distribution data. For (3), scale-invariant models are better at generalizing from training data to unseen data because they can recognize patterns across different scales. We aim to improve the understanding and capabilities of deep learning systems, especially with applications in lower-dimensional and scale-invariant data such as 1-D signals, time series analysis and wavelet analysis in signal processing.

Mathematical Framework

Let $x \in \mathbb{R}^d$ be an input vector and $f_\theta : \mathbb{R}^d \to \mathbb{R}^m$ be a NN with L layers parametrized by θ. We express f_θ as a composition of layer-wise transformations:

$$f_\theta(x) = (T_L \circ T_{L-1} \circ \cdots \circ T_1)(x)$$

where $T_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_{i+1}}$ is the transformation implemented by the i-th layer. To incorporate scale-awareness into our framework, we introduce a scale parameter s_i associated with each layer i and require the following scale covariance property for any $\lambda > 0$: $T_i(\lambda x; s_i) = \lambda^{\Delta_i} T_i(x; \lambda s_i)$, where Δ_i is the scaling dimension of the i-th layer transformation.

Definition 1 *A function $\sigma_s : \mathbb{R} \to \mathbb{R}$ is a scale-aware activation function if it satisfies $\sigma_s(\lambda x) = \lambda^\alpha \sigma_{\lambda s}(x)$ for some $\alpha > 0$ and all $\lambda > 0$.*

Building on this definition, we establish the scale covariance property for typical NN layers:

Theorem 1 *Let $T_i(x; s_i) = W_i \sigma_{s_i}(x) + b_i$, where W_i is a weight matrix, b_i is a bias vector, and σ_{s_i} is a scale-aware activation function. Then T_i satisfies the scale covariance property with $\Delta_i = \alpha$.*

RG Flow Equations

RG flows describe how a system's effective parameters change as we observe it at different scales. In NNs, RG flows

help to understand how the network's parameters evolve during training and how they contribute to the emergence of scale-invariant features. We derive RG flow equations for the network parameters θ as the scale s is varied: $\frac{d\theta}{ds} = \beta(\theta)$, where $\beta(\theta)$ is the beta function characterizing how the effective parameters change with scale. Here, this links the learning dynamics of NNs and the RG flows in physical systems.

Theorem 2 *For a NN with scale-aware activation functions, the beta function for the weights W_i and biases b_i of layer i is given by:*

$$\beta(W_i) = (\alpha - 1)W_i, \quad \beta(b_i) = -\alpha b_i.$$

Proof Sketch: Consider the transformation $T_i(x; s) = W_i\sigma_s(x) + b_i$. Under a scale transformation $s \to \lambda s$, $T_i(x; \lambda s) = W_i\sigma_{\lambda s}(x) + b_i$, which simplifies to $W_i\lambda^{-\alpha}\sigma_s(\lambda x)+b_i$, and so $\lambda^{-\alpha}(W_i\sigma_s(\lambda x)+\lambda^{\alpha}b_i)$. For this to be consistent with the scale covariance property, we have $W_i(\lambda s) = \lambda^{\alpha-1}W_i(s)$, $b_i(\lambda s) = \lambda^{-\alpha}b_i(s)$. We yield $\beta(\theta)$ when we take the logarithmic derivative with respect to λ and evaluate at $\lambda = 1$.

Fixed Points and Scale-Invariant Representations

We consider fixed points and their relationship to scale-invariant representations when studying RG flows in NNs to improve the stability and generalization capabilities of NNs. Fixed points represent stable configurations of network parameters, which indicate optimal solutions, while scale-invariant representations are likely to generalize better across different scales of input data.

Definition 2 *A fixed point θ^* of the RG flow is a set of parameters that satisfies $\beta(\theta^*) = 0$.*

Theorem 3 *Fixed points of the RG flow correspond to scale-invariant feature representations.*

Proof Sketch: At a fixed point θ^*, we have $\beta(\theta^*) = 0$. This implies that the network parameters do not change under scale transformations. Thus, the features extracted by the network at this fixed point are invariant to changes in scale.

RG-Guided Training

We propose a training procedure incorporating RG flow to guide the optimization towards scale-invariant representations. Define η as the learning rate and ϵ controls the influence of the flow.

1. Initialize network parameters θ_0
2. For each batch of data:
   (a) Forward pass: Compute $f_\theta(x)$ for inputs x
   (b) Backward pass: Compute gradients $\nabla_\theta L$ for loss L
   (c) RG step: Update $\theta \leftarrow \theta - \eta\nabla_\theta L - \epsilon\beta(\theta)$

Theorem 4 *Under suitable conditions on the loss landscape and learning rate schedule, the RG-guided training procedure converges to a fixed point of the RG flow while minimizing the loss function.*

Proof Sketch: We analyze the dynamics of the combined gradient descent and RG flow in parameter space. We show that the procedure converges to a point where both the loss gradient and $\beta(\theta)$ disappears, corresponding to a minimum.

Information-Theoretic Analysis

To analyze the emergent hierarchical structure in the trained network, we introduce an information-theoretic measure.

Definition 3 *The layer information I_i is defined as the mutual information between the representation at layer i and the input: $I_i = I(T_i \circ \cdots \circ T_1(X); X)$ where X is the random variable representing the input distribution.*

Theorem 5 *For a network trained with the RG-guided procedure, I_i decreases monotonically with i.*

Proof Sketch: We use the data processing inequality and the properties of the RG flow in the proof. Each layer transformation may be a noisy channel, and the RG flow ensures that information is systematically coarse-grained as it propagates through the network. Since information can only decrease through successive transformations, $I_{i+1} \leq I_i$.

Theorem 5 indicates that each layer performs a coarse-graining operation. Thus, fine-grained details (noise or less relevant features) are progressively filtered out, leaving behind more abstract and 'essential' features.

Connections to Statistical Mechanics

We establish a connection between our framework and statistical mechanics. RG flow of NN parameters can be mapped to the flow of coupling constants in statistical physics models near critical points.

Theorem 6 *There exists a mapping between the RG flow of NN parameters and the flow of coupling constants in the Ising model near its critical point.*

Proof Sketch: We construct a mapping between the weight matrices of the NN and the coupling constants of the Ising model and show that $\beta(\theta)$ derived for the NN parameters have the same functional form as the $\beta(\theta)$ of the Ising model couplings near criticality. Thus, the feature representations learned by NNs share connections with the behavior of physical systems near phase transitions.

Conclusion and Future Work

In this work, we present a framework that unifies deep learning and renormalization group theory, in order to better understand the feature learning process of NNs. Our approach focuses on scale invariance, leading to more interpretable hierarchical representations and improving generalization to out-of-distribution data. Future work includes empirically validating the proposed framework through simulations.

References

Cagnetta, F.; Petrini, L.; Tomasini, U. M.; Favero, A.; and Wyart, M. 2024. How Deep Neural Networks Learn Compositional Data: The Random Hierarchy Model. *Physical Review X*, 14(3).

Erdmenger, J.; Grosvenor, K. T.; and Jefferson, R. 2021. Towards quantifying information flows: relative entropy in deep neural networks and the renormalization group. arXiv:2107.06898.

Iso, S.; Shiba, S.; and Yokoo, S. 2018. Scale-invariant feature extraction of neural network and renormalization group flow. *Phys. Rev. E*, 97: 053304.