[Paper Review]
Learning to Compress: Local Rank and Information Compression in Deep Neural Networks

Paul Jason Mello
Department of Computer Science and Engineering
University of Nevada, Reno
pmello@unr.edu

Abstract

”Deep neural networks tend to exhibit a bias toward low-rank solutions during training, implicitly learning low-dimensional feature representations. This paper investigates how deep multilayer perceptrons (MLPs) encode these feature manifolds and connects this behavior to the Information Bottleneck (IB) theory. We introduce the concept of local rank as a measure of feature manifold dimensionality and demonstrate, both theoretically and empirically, that this rank decreases during the final phase of training. We argue that networks that reduce the rank of their learned representations also compress mutual information between inputs and intermediate layers. This work bridges the gap between feature manifold rank and information compression, offering new insights into the interplay between information bottlenecks and representation learning.” [2]

1 Summary

In this work, the authors investigate the relationship between MLPs and the manifold hypothesis by exploring a novel concept called ”local rank”, a measure that defines the feature manifold dimensionality. In this process, they develop a theoretic and empirical justification for this new metric and demonstrate that increases in the manifold dimensionality result in worse model performance on accuracy, further supporting the manifold hypothesis and illustrating a connection between it and the IB principle for internal information compression.

2 Introduction

The manifold hypothesis describes that real-world datasets lie on a manifold which captures the intrinsic dimensions of a system in a far lower dimensional space. Recent work has shown MLPs learn to compress input data into lower-dimensional feature manifolds. In this work, the authors aim to understand how MLPs compress feature representations. To understand this concept, the authors propose the concept of ”Local Rank” as a way to measure feature manifold dimensionality and identify its relationship to information compression, a phase which occurs during training.

3 Background and Motivation

At the heart of modern AI research is discovering the ”substance” through which models learn and attempting to distill this process into a highly optimized approach to train models better and faster. This paper explores this direction through understanding the fundamental processes that guide a models information compression from the data manifold it learns. They generate a metric to capture the data dimensionality from the feature manifold and illustrate empirical results directly relating local rank to prior works in implicit regularization, the IB principle, and Jacobian matrix ranks.

3.1 Key Concepts

- **Concept 1:** They propose the novel concept of ”Local Rank” as a metric to quantify the dimensionality of feature manifolds.
- **Concept 2:** Neural networks implicitly seek to learn low-rank solutions to solve a problem. This is akin to learning the properties of a variable X and its target Y through its mutual information, where MI is highly concentrated in low dimensional manifold spaces.
- **Concept 3:** They demonstrate, through the IB principle, that as one increases local rank, equivalent to increasing the dimensionality of the manifold, the accuracy of a model on classification tasks significantly reduces. Conversely, when the manifold dimension (local rank) is kept low the model is able to attain a high accuracy.

4 Theory

4.1 Local Rank

”Local rank is a measure of the dimensionality of feature manifolds learned by a neural network” [2]. The local rank LR is defined on a per-layer-basis l where the first layer is written as the expected rank of the Jacobian p_l. We approximate this local rank, since all matrices must respect this rank, to be $\epsilon > 0$

$$LR_l^\epsilon = \mathbb{E}_{x\sim\text{Data}}\left[\text{rank}_\epsilon\left(J_x p_l\right)\right]. \tag{1}$$

4.2 Theoretical Analysis

Under the assumption of certain ODE based gradient flows, solutions have been shown to converge to a Karush-Kuhn-Tucker (KKT) point. Succinctly, KKT is a set of conditions, consisting of stationarity, primal feasibility, complementary slackness, and dual feasibility, that are required to find an optimal nonlinear solution. Particularly, KKT minimizes a norm which, under the manifold hypothesis, should show dimensionality reduction in the rank weight of matrices through the training process. This implies the existence of a process which bottlenecks feature translation from a higher dimensional data manifold to a lower dimension.

They leverage results from a prior work titled ”Implicit regularization towards rank minimization in relu networks” [4] which illustrate two findings. Firstly, that gradient flow on small ReLU networks fail to minimize rank matrices, and secondly, that adding sufficient depth to these ReLU networks result in a bias to low-rank solutions. In this work, Patel et al. [2], they expand on these notions and derive two (informal) proofs. The first of which describes that deep neural network layers are bounded by an upper bound on local rank 2:

$$LR_l^\epsilon \leq \frac{2}{\epsilon^2}\left(\frac{B}{\sqrt{2}}\right)^{\frac{2K}{L}} \cdot \frac{L+1}{L}\|W_l\|_\sigma^2, \tag{2}$$

where,

- L: number of layers in the neural network.
- ϵ: small positive threshold for rank approximation.
- LR_l^ϵ: local rank at layer l with threshold ϵ.
- $\|W_l\|_\sigma$: operator norm of the weight matrix W_l (largest singular value of W_l).
- B: uniform bound on the weight matrices $W_1, \ldots, W_L$.
- $\frac{2K}{L}$: a constant depending on the network depth L and the problem’s specific properties.

They then apply these notions to a theoretical regression task for neural network. The theory of this approach is that neural networks are capable of being trained to exactly fit a regression datasets. However, implicit regularization will drive the network to solutions where the local rank is reduced causing a collapse from an exact solution to approximate over time. With this in mind, they propose 3, serving as a theoretical justification. They posit that a deep neural network should implicitly find a reduced local rank that should be visible during neural network training to fit a regression task, which it should do perfectly.

$$LR_l^\epsilon \leq \frac{\|W_l\|_\sigma^2 \cdot B^{\frac{2K}{L}}}{\epsilon^2}, \tag{3}$$

where,

- $\mathcal{N}$: Fully connected neural network with weight matrices $\theta = [W_1, \ldots, W_L]$.
- $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^{n_0} \times \mathbb{R}_+$: Regression dataset where x_i are inputs and y_i are the corresponding outputs.

Here, $B^{\frac{2K}{L}}$, describes how a networks layer depth influences compression of the feature space. Through enough depth, one can amplify the effect of implicit regularization. Meanwhile, $\|W_l\|_\sigma^2$, is the operator norm variable which represents the largest singular value of a matrix, and thus its upper bound. By bounding the norm of the Jacobian, the authors of this work demonstrate guarantees that the local rank decreases during training and illustrate a relationship between the Jacobian matrix of the layer's output and its weight matrices.

5 Experiments and Results

To validate these insights they measure the local rank during training of MLPs on synthetic and real datasets. For synthetic data, they generate Gaussian data with random covariance matrices. A 3-layer MLP must then learn to map between correlated Gaussian distributions. For real data, they select MNIST and train a 4-layer MLP on cross entropy loss. Utilizing these models and dataset, they track the local rank through training of each layer.

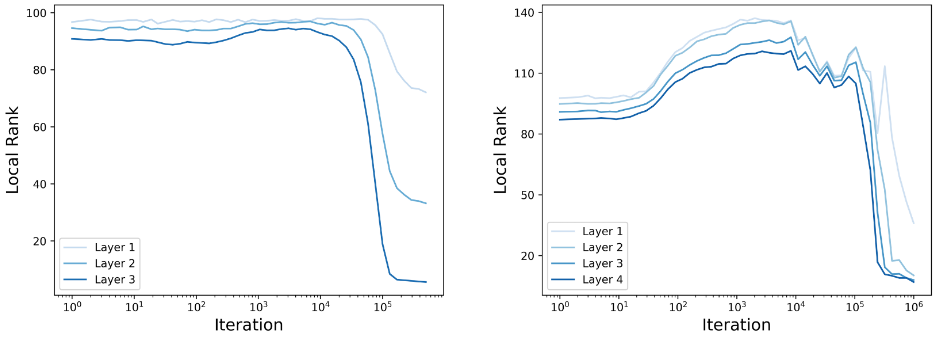

Figure 1: Local Rank Training Reduction by Layer. The **Left** contains the Gaussian data, 3-layer MLP. The **Right** contains the MNIST data, 4-layer MLP.

Following this demonstration of local lank decay during training, they explore local rank's relationship to information compression through the IB principle. They show in theorem 4 that there exists a trade-off parameter β which can apply optimal linear transformations to a matrix rank and thus add some controllability to the manifolds dimensionality.

Assume a noisy linear transformation $T = A_\beta X + \eta$. Critical values exist for β_n^c such that $0 \leq \beta_c^i \leq \beta_c^j$ where $i < j$, and:

$$\text{rank}(A_\beta) = n \quad \text{for} \quad \beta \in (\beta_n^c, \beta_{n+1}^c). \tag{4}$$

These critical points correspond to dimensionality changes of T, which constitutes a feature representation, and thus develop the local rank.

To explore this, they establish experimentation with β on both synthetic and complex datasets utilizing a Deep Variational Information Bottleneck (DVIB) model [1]. For datasets, they continue with synthetic Gaussian and real MNIST. They select β values to be critical points for the changing of T and measure the gradual changes of β and its effects on the local rank.

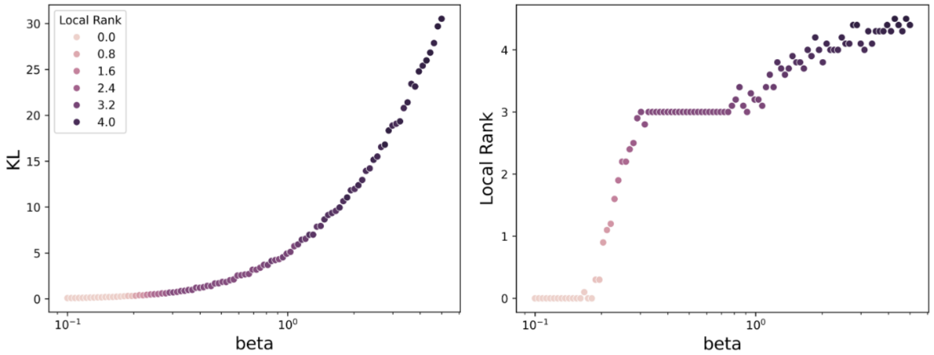

Figure 2: Gaussian dataset using a DVIB. The **Left** measures the KL divergence with each point being β which are critical values. The **Right** identifies the "local rank as a function of β, showing an increase with β and distinct phase transitions" [2].

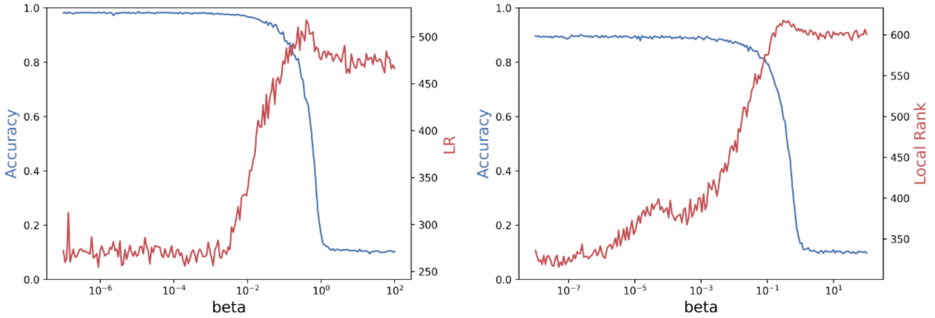

Figure 3: The **Left** MNIST data and **Right** Fashion-MNIST. In both instances, as β increases accuracy decreases.

These findings in fig. 3 illustrate that as β increases, the local rank increases, and accuracy inevitably decreases. In other words, as we increase the local rank, we effectively move the data to a higher-dimensional manifold which is effected by the curse of dimensionality. Conversely, reducing the local rank maps our model to a lower dimensional manifold which can help learn the compressed representations easier.

They note these findings coincide with the findings of the author in another one of his works titled "Representation compression and generalization in deep neural networks" [3]. In this prior work, the author describes a two phase training system consisting of diffusion and compression related to the mutual information between variables over layers. Effectively, the first phase of training consists in aggregating all the relevant mutual information between variables and between layers, while the much longer second phase of compression learns to identify and compress better patterns to generalize properties of the collected feature representations.

5.1 Evaluation Metrics

- **Metric 1:** To evaluate their theory on local rank, they train two models. One to a regression task and one to MNIST data, then measure the local rank through time. This metric becomes an information geometric measure to understand the unfolding dimensionality of manifolds through training.

5.2 Key Results

- **Result 1:** The local rank decreases during the later stages of training in both synthetic and real-world data, supporting the hypothesis that neural networks compress feature representations and show that compressions on manifold dimensions are measurable.
- **Result 2:** This metric can be used in future work to approximate the bounds of the manifold dimensionality. Fundamentally, helping to train and run better models by making progress on understanding the exact properties which allow neural networks to exhibit these phase transitions.

6 Discussion and Critique

This work introduces a novel metric which can be used to understand the internal representations of datasets on a feature manifold. This opens a new direction to exploring the data manifold and reinforces prior theory like that of the IB principle and manifold hypothesis. A few strengths and weaknesses are listed below

6.1 Strengths

- **Strength 1:** Introduction of local rank as a novel measure for understanding the internal mechanics of neural networks from an information-theoretic and geometric manifold perspective.
- **Strength 2:** They show strong knowledge of the problem space and empirical solutions. They start from first principles expand into strong evidence which demonstrates clear phases transitions and dimensionality reductions implicit in training.
- **Strength 3:** They open a new path to study the feature manifold from control over local rank.

6.2 Weaknesses

- **Weakness 1:** It would be interesting to see if these results hold under models and architectures beyond MLPs. Particularly, CNNs, GNNs, Transformers, etc... as the inductive biases tend to be stronger, it would be interesting to see how local rank behaves under these different information compression landscapes.

7 Future Directions

- It would be very interesting to see ablation studies on the optimal control of local rank.
- Utilizing local rank to precisely explore the data manifold of a model and how it dynamically shapes its dimensionality during training. Exploring this direction could be of particular use for optimization problems.
- It would be very interesting to see what happens to local rank when a model begins to undergo grokking.

8 Conclusion

Utilizing prior work on ReLU inductive biases, the researchers introduce a novel metric called ”local rank” and demonstrate its usefulness in identifying the dimensionality of a learned feature manifold. The manifold hypothesis posits that all data lies on a low rank manifold and this work expands the theory further into a meaningful metric. Through a simple MLP and complex DVIB model, they make evident that not only are models learning in lower dimensional manifolds, but that it is necessary for models to learn at all, as higher dimensional manifolds fail to classify MNIST data properly. In all, the exploration of local rank unlocks a new direction to study neural networks and their feature manifolds in a more precise and controllable manner.

References

[1] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck, 2019.

[2] Niket Patel and Ravid Shwartz-Ziv. Learning to compress: Local rank and information compression in deep neural networks, 2024.

[3] Ravid Shwartz-Ziv, Amichai Painsky, and Naftali Tishby. Representation compression and generalization in deep neural networks, 2019.

[4] Nadav Timor, Gal Vardi, and Ohad Shamir. Implicit regularization towards rank minimization in relu networks, 2022.