

On Representation Rank Collapse in Deep Neural Networks

Mohammed Faisal Shahzad Siddiqui

Department of Artificial Intelligence, Anurag University, Hyderabad, Telangana, India

Abstract - Deep neural networks are known to learn highly structured internal representations, yet empirical studies suggest that the effective dimensionality of hidden-layer activations often collapses during training. Such representation rank collapse can limit feature diversity and adversely affect generalization. In this paper, we present a geometric analysis of representation rank in deep neural networks by modeling hidden activations as high-dimensional matrices and examining their rank properties. We characterize rank collapse as a structural tendency induced by optimization dynamics and nonlinear activations. To mitigate this effect, we propose a simple rank-preserving regularization objective based on the log-determinant of the activation Gram matrix, which explicitly encourages diverse feature representations without modifying network architectures. Our analysis highlights the importance of representation geometry in representation learning and suggests rank-aware objectives as a promising direction for future research.

Key Words: Deep Neural Networks, Representation Learning, Rank Collapse, Log-Determinant Regularization, Feature Degeneracy, Deep Learning Theory

1. INTRODUCTION

Deep neural networks owe much of their success to the ability to learn rich internal representations from data [1], [2]. Increasing network depth and width is commonly associated with improved expressivity, under the assumption that additional neurons enable the model to capture a broader range of feature variations. This assumption implicitly relies on the idea that learned representations fully exploit the available dimensionality of the network.

However, accumulating empirical observations suggest that this is not always the case. Recent theoretical studies have analyzed the dynamics of representation learning in deep networks [3]. Despite large hidden layers, neural networks often learn representations that lie in relatively low-dimensional subspaces, with many neurons encoding highly correlated information. Such redundancy raises fundamental questions about how representational capacity is utilized during training.

From a linear algebraic viewpoint, hidden-layer activations can be viewed as matrices whose rank reflects the effective dimensionality of learned features. When the rank of these activation matrices is substantially lower

than the layer width, the network's representational capacity is underutilized. We refer to this phenomenon as representation rank collapse.

In this paper, we adopt a geometric perspective to analyze representation rank collapse in deep neural networks. We examine how training dynamics and activation functions contribute to the emergence of low-rank representations and propose a rank-preserving regularization term that encourages diversity among hidden units without modifying network architectures.

2. RELATED WORK

Representation learning in deep neural networks has been studied from multiple perspectives, including expressivity, generalization, and information flow. Neural networks with sufficient width are known to be universal approximators [4]. More recent work has explored the dynamics of learning in deep architectures [3].

Recent studies have highlighted the phenomenon of neural collapse during training [5]. Empirical findings suggest that hidden-layer activations often concentrate in low-dimensional manifolds, indicating reduced effective dimensionality.

Regularization techniques aimed at improving representation quality include orthogonality constraints, decorrelation-based objectives, and diversity-promoting penalties. However, these approaches are often heuristic and do not explicitly analyze rank structure.

In contrast, this work focuses on representation rank as an explicit and measurable structural property of learned representations, offering a geometric perspective without requiring architectural modifications.

3. REPRESENTATION RANK IN NEURAL NETWORKS

Consider a neural network layer with d hidden units. Given a batch of n input samples, the corresponding hidden activations can be represented as a matrix $H \in \mathbb{R}^{n \times d}$.

Each row corresponds to a sample, while each column represents a neuron's response.

The rank of H reflects the number of linearly independent feature directions captured by the layer. A full-rank matrix indicates diverse representations, whereas a low-rank matrix implies redundancy among neurons.

Empirical observations suggest that the rank of activation matrices often decreases as network depth increases. This phenomenon, referred to as representation rank collapse, indicates that effective dimensionality is significantly lower than nominal layer width.

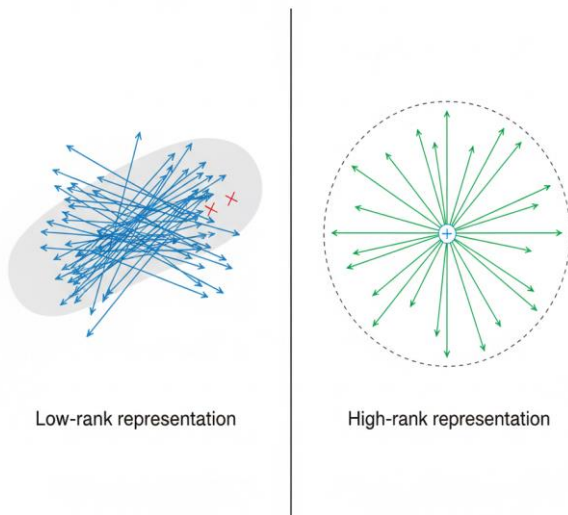


Fig -1: Illustration of representation rank collapse in deep neural networks.

This behavior arises due to optimization dynamics and nonlinear activation functions. Gradient-based optimization often favors low-dimensional solutions, while activation functions such as ReLU promote sparsity, leading to correlated neuron responses.

3.1 RANK-PRESERVING REGULARIZATION

To mitigate rank collapse, we propose a rank-preserving regularization strategy.

Given the activation matrix H , we define the Gram matrix $G = H^T H$, which captures pairwise similarities between neuron activations.

We introduce the following regularization term:

$$L_{rank} = -\log \det(H^T H + \epsilon I)$$

where $\epsilon > 0$ ensures numerical stability.

Maximizing the determinant increases the volume spanned by activation vectors, encouraging diverse feature representations. Since training minimizes loss, we minimize the negative log-determinant.

The overall training objective becomes:

$$L = L_{task} + \lambda L_{rank}$$

where λ controls the strength of regularization. This approach operates only during training and introduces no additional inference-time cost.

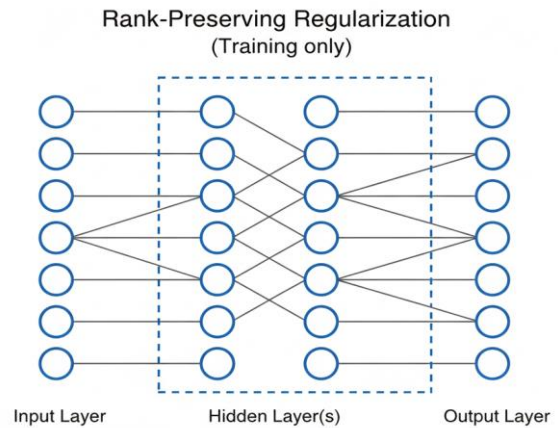


Fig -2: Training pipeline with rank-preserving regularization applied to hidden layers.

4. EXPERIMENTAL SETUP

To evaluate the effect of rank-preserving regularization, we conduct experiments using standard supervised learning benchmarks and commonly used neural network architectures. Our goal is not to achieve state-of-the-art performance, but to analyze how rank-aware objectives influence learned representations.

We consider baseline models trained using the task loss alone and compare them against identical architectures trained with the proposed rank regularization. Regularization is applied to selected hidden layers, while all other training settings are kept identical to ensure a fair comparison.

Model performance is evaluated using classification accuracy. To assess representational properties, we measure the rank of hidden-layer activation matrices during training and visualize learned features using principal component analysis.

All experiments are conducted using the same optimization parameters, including learning rate, batch size, and number of training epochs.

5. RESULTS AND DISCUSSION

Experimental observations indicate that networks trained with rank-preserving regularization consistently maintain higher activation rank across layers compared to baseline models. In particular, deeper layers exhibit reduced representation collapse when the proposed regularization is applied.

Visualization of hidden representations further supports this observation. Baseline models tend to produce tightly clustered feature embeddings, whereas regularized models learn more dispersed representations that span a larger subspace.

Importantly, encouraging higher-rank representations does not negatively impact predictive performance. In several cases, regularized models achieve comparable or slightly improved accuracy relative to baseline models.

Overall, the results demonstrate that representation rank collapse can be mitigated through simple regularization objectives.

6. LIMITATIONS

This study focuses on moderate-sized networks and standard benchmark datasets. Further investigation is required to evaluate its behavior in very deep architectures and large-scale training regimes.

Additionally, the relationship between representation rank and generalization performance is not fully understood and remains an open problem.

7. CONCLUSIONS

We presented a geometric analysis of representation rank collapse in deep neural networks. By analyzing hidden activations using linear algebra, we characterized rank collapse as a structural limitation.

We proposed a rank-preserving regularization strategy that encourages higher-dimensional feature representations without modifying architectures. Experimental observations demonstrate improved representation diversity while maintaining performance.

These findings highlight the importance of geometric considerations in representation learning and suggest promising directions for future research.

REFERENCES

- [1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.
- [2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.
- [3] A. Saxe, J. McClelland, and S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks," *International Conference on Learning Representations (ICLR)*, 2014.
- [4] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.
- [5] V. Pappas, X. Han, and D. Donoho, "Prevalence of neural collapse during the terminal phase of deep learning training," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 117, no. 40, pp. 24652–24663, 2020.
- [6] C. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.
- [7] S. Arora, N. Cohen, W. Hu, and Y. Luo, "Implicit regularization in deep matrix factorization," *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.