Paper Review: A Geometric Analysis of Neural Collapse with Unconstrained Features[4]

Qiulin Fan

1 Motivation

Deep neural networks have achieved remarkable success across various tasks, yet many aspects of their training dynamics remain poorly understood. Early stopping is often used in machine learning to prevent overfitting, but modern approaches of training classification deepnets involves a *Terminal Phase of Training* (TPT), i.e., even if the training error is already zero, continued training to improve generalization and counter robustness. In this process, a striking phenomenon observed, termed *Neural Collapse*, is the collective name for: that, though *TPT* does not explicitly control the shape of the classifier, according to the original research of Papyan et al.[2], the model is regularly directed towards a specific structure of convergence, termed *Neural Collapse* (NC). This is the collective name of four deeply interconnected phenomena, namely:

- **(NC1) Variability collapse:** The within-class variation of the activations becomes negligible as these activations collapse to their class-means.
- **(NC2) Convergence to Simplex ETF:** The vectors of the class-means converge to having equal length, forming equal-sized angles between any given pair, identical to mathematical structure known as **Simplex Equiangular Tight Frame (ETF)**.
- **(NC3) Convergence to self-duality:** The class-means and linear classifiers, although mathematically quite different objects, living in dual vector spaces. Combined with NC2, this implies that each iso-classifier-decision region is isometric to any other such region by rigid Euclidean motion.
- **(NC4) Simplification to Nearest Class-Center (NCC):** For a given deepnet activation, the network classifier converges to choosing whichever class has the nearest train class-mean (in standard Euclidean distance).

Studying the behavior of NC produce important benefits, including better generalization performance, better robustness, and better interpretability. And it has broad applications in many fields of deep learning. For instance, Md Yousuf Harun et al.[1] recently showed that in *Out-of-distribution* (OOD) detection, stronger NC improves OOD detection but degrades generalization, while weaker NC enhances generalization at the cost of detection, thus by linking them in some framwork, we can mitigates NC to improve generalization in OOD.

2 Supporting Materials

The article “Prevalence of Neural Collapse during the terminal phase of deep learning training"[2] is the earliest to introduce the term *Neural Collapse*, and to categorize NCs into the above four categories. In this article, the authors analyze the reasoning behind NC, show that NC2 through NC4 are roughly NC1-induced, and demonstrate that the mathematical structure of the Simplex ETF is the optimal geometry for this process, and that NC naturally converges to this structure.

This paper was far-reaching and widely cited, and it was followed by many excellent studies discussing *Neural Collapse*, such as the paper we will be reviewing: “A Geometric Analysis of Neural Collapse with Unconstrained Features”. The paper studies the problem based on a simplified unconstrained feature model, which isolates the topmost layers from the classifier of the neural network. It proves that the global optima correspond exactly to the NC phenomenon while establishing that all non-global critical points are strict

saddles. Empirical validations on datasets like MNIST and CIFAR-10 using architectures such as ResNet18 further corroborate the theoretical insights, showing that NC emerges regardless of the optimization algorithm used (e.g., SGD, Adam, LBFGS), and the cross-entropy loss will **efficiently escape from the saddle points** and **guarantee convergence to the global optimum**. This demonstrates the potential of utilizing the neural collapse phenomenon to improve the network design (e.g., fixing the last layer of classifiers to be the Simplex ETF, and lowering the feature dimensionality to reduce the computational and memory overheads).

3 Plans

Our review is organized into four periods:

In the first period during the week of spring break (from 3/1 to 3.7), we will scrutinize "Prevalence of Neural Collapse during the terminal phase of deep learning training", using ResNet, to recover NC generation on FashionMNIST in order to understand the behavior of NC2 (converges to Simplex Equiangular Tight Frame) in the TPT phase.

In the second period during the two weeks from 3/8 to 3/22, we will read through the key theoretical and experimental results from "A Geometric Analysis of Neural Collapse with Unconstrained Features", focusing on understanding and recapitulating the proofs of important theorems, such as Global Optimality Conditions and No Spurious Local Minima and Strict Saddle Property, as well as using SGD algorithms to rehabilitate the Experiments.

In the third period during the week from 3/23 to 4/2, we will be looking for more papers on Neural Collapse, in particular the followup paper citing "A Geometric Analysis of Neural Collapse with Unconstrained Features", and explore potential applications of NC insights. For instance, we have "Imbalance Trouble: Revisiting Neural-Collapse Geometry" by Christos Thrampoulidis et al [3].

During the rest time, we will finish compiling our review.

References

[1] Md Yousuf Harun, Jhair Gallardo, and Christopher Kanan. *Controlling Neural Collapse Enhances Out-of-Distribution Detection and Transfer Learning*. 2025. arXiv: 2502.10691 [cs.LG]. URL: https://arxiv.org/abs/2502.10691.

[2] Vardan Papyan, X. Y. Han, and David L. Donoho. "Prevalence of neural collapse during the terminal phase of deep learning training". In: *Proceedings of the National Academy of Sciences* 117.40 (Sept. 2020), pp. 24652–24663. ISSN: 1091-6490. DOI: 10.1073/pnas.2015509117. URL: http://dx.doi.org/10.1073/pnas.2015509117.

[3] Christos Thrampoulidis et al. "Imbalance Trouble: Revisiting Neural-Collapse Geometry". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 27225–27238. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/ae54ce310476218f26dd48c1626d5187-Paper-Conference.pdf.

[4] Zhihui Zhu et al. *A Geometric Analysis of Neural Collapse with Unconstrained Features*. 2021. arXiv: 2105.02375 [cs.LG]. URL: https://arxiv.org/abs/2105.02375.