Neural Collapse in Quantized Neural Networks

Gabor Balazs Regely, ABX5LX

1 Introduction

During the terminal phase of training (TPT), Deep Neural Networks reach zero error and various architectures exhibit the Neural Collapse (NC) properties. Neural collapse represents a state at which the within-class variability of the final hidden layer outputs is infinitesimally small, their class means form a simplex equiangular tight frame and the last layer acts as a nearest-class center classifier. [1, 2]

Quantization is a widely used technique to reduce the memory footprint and computational requirements of deep neural networks. However, quantization introduces noise and can potentially affect the training dynamics and final performance of neural networks.

This project work aims to investigate the interplay between neural collapse and quantization in deep neural networks. The project's source code is available at: `github.com/eRGiBi/QuantizedNeuralCollapse`.

2 Literature Review

2.1 Neural Collapse

Papyan et al. [1] introduced NC and described the four main properties across different architectures and datasets, focusing on the computer vision domain. NC is characterized as the co-occurrence of:

1. NC1 - Within-class variability collapse: within-class variation of activations becomes negligible as they collapse to their class-means.
2. NC2 - Class mean convergence to simplex ETF: vectors of class-means (after centering) converge to an equiangular tight frame (ETF), maximizing pairwise angles and distances.
3. NC3 - Self-dual alignment (convergence to self-duality): columns of the last layer linear classifier matrix also form a simplex ETF in their dual vector space and converge to the simplex ETF.
4. NC4 - Nearest class center classification: the last-layer classifier acts with the nearest class mean decision rule on the penultimate layer features.

Kothapalli [2] gives a broad review on the principles behind NC and its implications for generalization.

NC properties can also be identified in Language Models in Natural Language Processing tasks, as shown by Wu et al. [3]. However, NLP requires a new approach as the conditions that give rise to NC in computer vision differ in this domain. Language Models are typically undertrained, classes are imbalanced and their numbers exceed the embedding dimension, and contexts can be ambiguous to the next token prediction. The authors introduce modified metrics to measure NC in language models, show that scaling models makes the NC properties emerge, and there is a correlation between NC and generalization.

2.2 Quantization

Gholami et al. [4] provide a comprehensive survey of quantization methods for efficient neural network inference, covering the advantages and disadvantages of different methods.

Ashkboos et al. present EfQAT [5], an efficient framework for quantization-aware training that reduces the computational overhead of QAT while maintaining model accuracy.

3 Practical Results

3.1 Computer Vision

I ran experiments with multiple convolutional neural network architectures, including simple custom CNNs, ResNet-18, MobileNetV3, different ConvNeXt variants and their customized versions. I reproduced Neural Collapse on the CIFAR-10 and CIFAR-100 datasets with different models using the NC metric calculations provided in [1], but success was heavily dependent on the given architecture and hyperparameters.

3.2 Natural Language Processing

I also conducted a few preliminary experiments with simple transformer models from the nanoGPT [6] repository and metrics from the Linguistic Collapse paper, modeling character level Shakespeare. For quantization, I used the default 32 and 16 bit floating point precisions.

4 Future Work

So far, I have limited my experimentation to simple models and haven't tested the large amount of possible quantization configurations. In NLP, the interaction of NC with post-training techniques is still an active research area.

References

[1] 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 (2020), 24652–24663. DOI: 10.1073/pnas.2015509117. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2015509117.

[2] Vignesh Kothapalli. *Neural Collapse: A Review on Modelling Principles and Generalization*. 2023. arXiv: 2206.04041 [cs.LG]. URL: https://arxiv.org/abs/2206.04041.

[3] Robert Wu and Vardan Papyan. *Linguistic Collapse: Neural Collapse in (Large) Language Models*. 2024. arXiv: 2405.17767 [cs.LG]. URL: https://arxiv.org/abs/2405.17767.

[4] Amir Gholami et al. *A Survey of Quantization Methods for Efficient Neural Network Inference*. 2021. arXiv: 2103.13630 [cs.CV]. URL: https://arxiv.org/abs/2103.13630.

[5] Saleh Ashkboos et al. *EfQAT: An Efficient Framework for Quantization-Aware Training*. 2024. arXiv: 2411.11038 [cs.LG]. URL: https://arxiv.org/abs/2411.11038.

[6] Andrej Karpathy et al. *nanoGPT: The simplest, fastest repository for training/finetuning medium-sized GPTs*. https://github.com/karpathy/nanoGPT. 2025.