Neural Collapse in Quantized Neural Networks

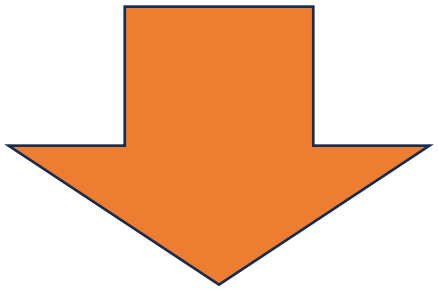
Gábor Balázs Régely, ABX5LX

Introduction

AI Boom

Black box behavior

Compute requirements



Interplay between neural collapse and quantization

Neural Collapse

Surprisingly simple geometric form of the features and of the classifier

During the **terminal phase of training** (TPT), when **zero error** is achieved

Four interconnected phenomena

NC1 - Within-class variability collapse

- Within-class variation of activations becomes negligible as they collapse to their class-means

- The within-class covariance matrix approaches zero:

$$\sum S_w \rightarrow 0$$

NC2 - Class mean convergence to simplex ETF

- Class-means (after centering) converge to an equiangular tight frame (ETF), maximizing pairwise angles and distances

$$\mu_c = \frac{< \mu_c^t, \mu_{c'}^t >}{|||\mu_c^t|| \cdot ||\mu_{c'}^t||} \rightarrow \begin{cases} 1, & \text{if } c = c' \\ -\frac{1}{C-1}, & \text{if } c \neq c' \end{cases}$$

NC3 - Self-dual alignment

- Columns of the last layer linear classifier matrix also form a simplex ETF in their dual vector space
- And converge to the simplex ETF (up to rescaling) of the penultimate layer features

$$\left|\left|\frac{\mu_c^t}{||\mu_c^t||} - \frac{w_c^t}{||w_c^t||}\right|\right| \to 0$$

NC4 - Nearest class center classification

- Last-layer classifier acts with the nearest class mean decision rule on the penultimate layer features

$$\hat{c}^t = \arg\min_{c'} ||h(x) - \mu^t_{c'}||$$

Quantization

- Comprehensive survey by Gholami et al., detailing the main quantization approaches

- EfQAT by Ashkboos et al. is a framework for QAT that reduced the computational overhead while maintaining accuracy

Machine Vision Experiments

- Metrics of Papyan et al. (2020)

- Convolutional neural networks
  - Custom CNNs
  - ResNet-18
  - MobileNetV3
  - Base ConvNeXt variants and their customized versions

- MNIST, CIFAR-10, CIFAR-100

ResNet-18, MNIST Example

200 epochs ≈ 4 hours

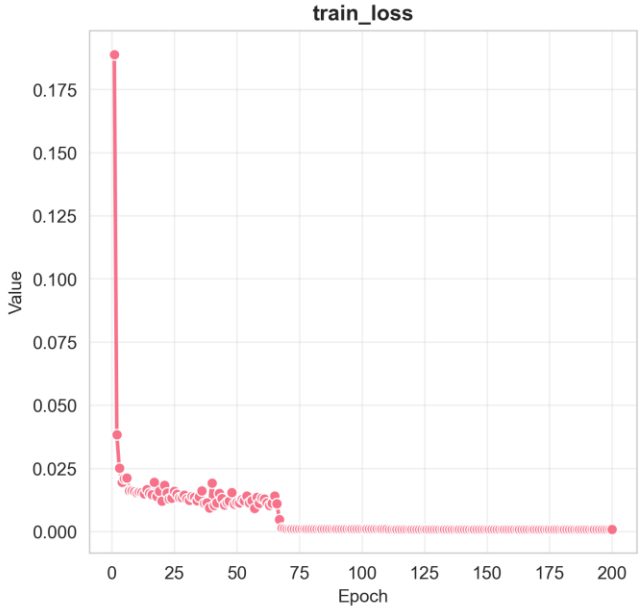
SGD with LR scheduling

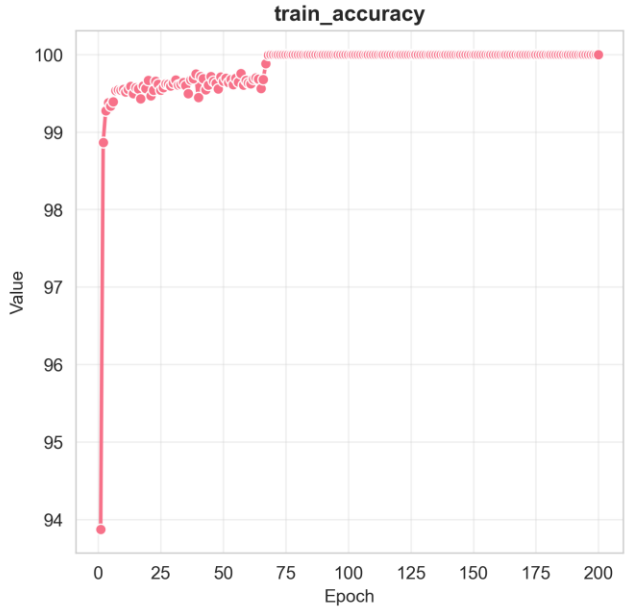
0.9 momentum

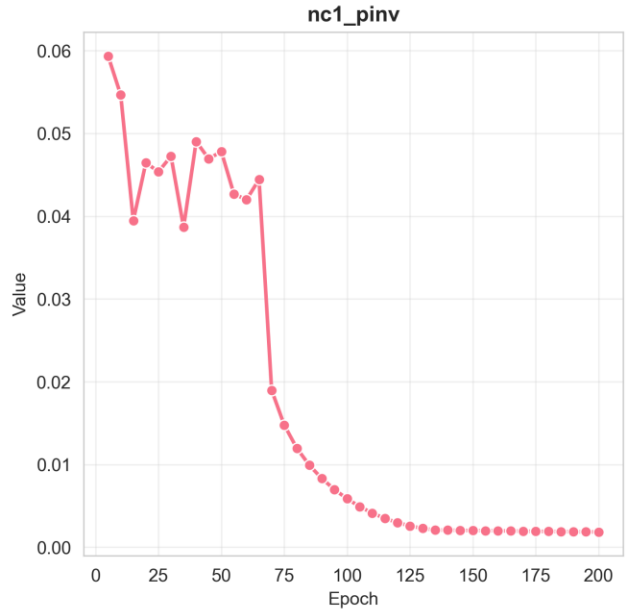
5e-4 weight decay

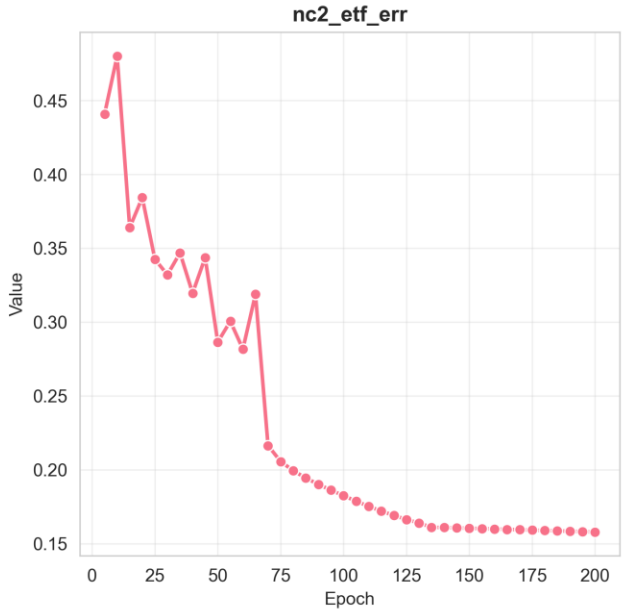
128 batch size

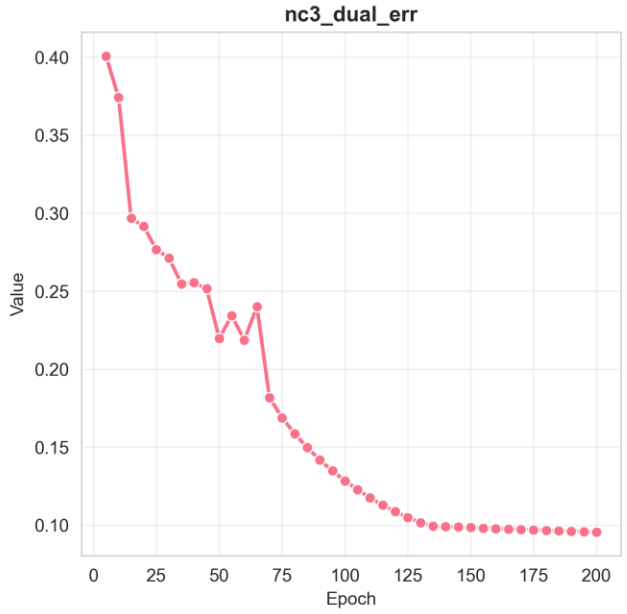
float32 precision

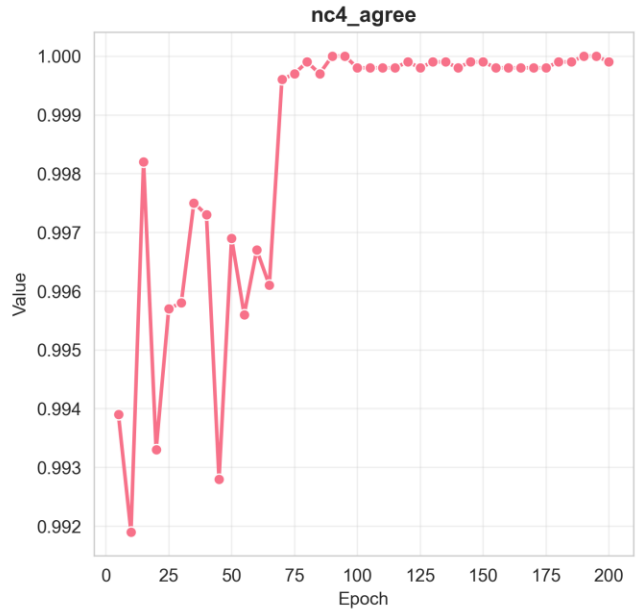












Linguistic Collapse

NLP requires a new approach as the conditions that give rise to NC in computer vision differ here

- LMs are typically undertrained
- Classes are imbalanced
- Class numbers exceed the embedding dimension
- Contexts can be ambiguous to the next token prediction

Language Modeling

- Metrics of Wu and Papyan (2024)

- Preliminary trials with small transformer models
- Character modeling on Shakespeare with nanoGPT

Future Work

- Improving theoretical and practical foundation

- Trials with quantization configurations
- Different Language Models

- Interaction of NC with post-training techniques
- Benchmarking downstream performance on different LLM metrics

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

Thank you for your attention!

AI Usage

- During the project, I trained CV and NLP AI models on publicly available datasets.
- Gemini 3.0 Pro: Correcting LaTeX page structure and citation formatting.
- Claude Sonnet 4.5: Generating Python functions for visualization, performance improvements on various functions.