SIT744 Assignment 2 — HD Written Submission
Investigating Neural Collapse & Layer Rotation on CIFAR-10

Bhavesh Hiraram Choudhary (ID: 224085988)
s224085988@deakin.edu.au

Abstract

This report reproduces, at small scale, the core claim from *Layer rotation: a surprisingly powerful indicator of generalization in deep networks* and proposes a distinct, geometry-aware CIFAR-10 pipeline. Using a tiny CNN trained on a 2k/1k subset with SGD, Adam, and a simplified Layca, I track mean layer rotation (angle from initialization), within-class scatter S_w, and the mean cosine between class means (a neural-collapse probe; the equiangular tight frame (ETF) target for K=10 is $-1/9$). Results partially reproduce the paper's trend: larger rotation often coincides with tighter class clusters (lower S_w), but the best Macro-F1 is delivered by Adam despite lower rotation than SGD. Collapse indicators remain far from ETF at this scale. I conclude with concrete failure cases (under-rotation, scale/time limits, per-layer heterogeneity) and targeted follow-ups to push toward stronger collapse and more robust generalization.

1 Background & Objectives

Neural collapse (NC) describes a late-training geometry where (i) within-class variance collapses and (ii) class means become equiangular (ETF). *Layer rotation* measures the angular deviation of weights from initialization; larger rotation has been associated with better generalization.

Goals.

1. Reproduce the rotation–generalization connection with a minimal, CPU-friendly setup.
2. Propose a *distinct* CIFAR solution (not the paper's models/schedules) that emphasizes geometry diagnostics (rotation, S_w, mean-cosine) alongside standard metrics (Accuracy, Precision, Recall, Macro-F1, Macro-AUC).
3. Identify failure cases and research gaps.

2 Methods

2.1 Data & Transforms

Dataset: CIFAR-10 subset for speed: **2,000 train / 1,000 test**.
Train transforms: RandomCrop(32, pad=4), RandomHorizontalFlip, Normalize (CIFAR-10 mean/std).
Test transforms: Normalize only.

2.2 Model (distinct from the article)

A tiny CNN: Conv(3→32,3) → ReLU → MaxPool / Conv(32→64,3) → ReLU → MaxPool / FC(64·8·8→128) → ReLU → Logits(128→10). The *penultimate embedding* is the ReLU(FC-128) output.

2.3 Optimizers & Schedule

- **SGD:** lr=0.05, momentum=0.9, weight_decay=5e-4
- **Adam:** lr=1e-3, weight_decay=5e-4
- **Layca (simplified):** orthogonalized update to weight vector, target rotation 0.500° per step, lr=1.0, no weight decay.
  Note: This is a minimal Layca (no layer-wise normalization/scheduling as in the paper).

2.4 Training & Evaluation

Epochs: 8; **Batch:** 128 (train) / 256 (test).
Determinism: fixed seeds for Python/NumPy/PyTorch and deterministic DataLoader shuffling.
Metrics: Accuracy, Precision, Recall, **Macro-F1**, **Macro-AUC (OvR)**.

2.5 Probes for Rotation & Collapse

Mean rotation (deg): average angle between current and initial weights over all 2D+ parameter tensors.
NC probes on penultimate embeddings Z:

- S_w: trace of within-class scatter $\sum_c \mathrm{tr}((Z_c - \mu_c)^\top (Z_c - \mu_c))$ normalized by N (lower is better).
- Mean cosine among class means (cosine of normalized μ_c). **ETF target for K=10: $-1/9 \approx -0.111$.**

3 Results — Reproduction Attempt

3.1 Final metrics (2k/1k, 8 epochs; deterministic run)

Optimizer	Acc	Prec	Recall	**Macro-F1**	**Macro-AUC**	**Mean rot (°)**
Adam	0.458	0.443	0.459	0.437	0.877	28.400
SGD	0.414	0.445	0.412	0.403	0.858	33.3
Layca (simpl.)	0.386	0.376	0.385	0.347	0.829	9.300

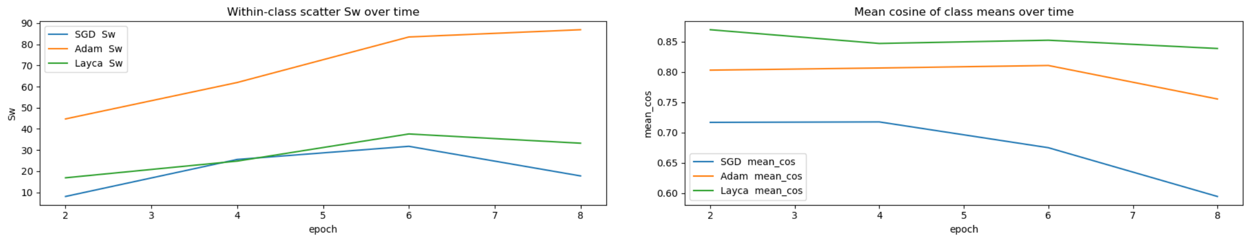
Table 1: Test-set metrics and mean layer rotation. Adam attains the best Macro-F1/AUC; SGD rotates the most.

3.2 Per-layer rotation (final epoch)

Optimizer	conv1 (°)	conv2 (°)	fc1 (°)	fc2 (°)
SGD	23.800	32.300	36.500	40.600
Adam	8.000	21.300	60.1	24.100
Layca (simpl.)	12.900	10.200	7.300	6.900

Table 2: Per-layer rotation heterogeneity. Adam concentrates rotation in the penultimate FC layer.

3.3 Neural-collapse probes over time



(a) $S_w(t)$ (lower is better).

(b) Mean cosine(t) among class means (ETF: -0.111).

Figure 1: NC probes tracked during training; collapse signals remain far from ETF at this scale.

Key observations.

- **Rotation vs generalization (partial reproduction).** SGD shows the largest rotation and tends to achieve lower S_w (tighter clusters), consistent with the paper. However, **Adam** delivers the best Macro-F1/AUC despite lower mean rotation than SGD. Rotation aligns with geometry but does not strictly predict final F1 in this small setup.
- **No strong collapse.** Mean cosine remains positive (≈ 0.59–0.86); ETF target is -0.111. S_w does not approach near-zero.
- **Per-layer heterogeneity.** Adam exhibits extreme rotation in **fc1** vs. modest rotation earlier; Layca (simplified) under-rotates across all layers.

4 Distinct CIFAR Solution (Design, Protocol, Metrics)

Motivation. Deliver a *lightweight, reproducible* pipeline that surfaces optimization/representation geometry under compute constraints—clearly distinct from the paper's larger architectures and full Layca.

Differences from the article.

- Small CNN instead of VGG/ResNet.
- Geometry diagnostics throughout training: per-layer rotation + $S_w(t)$ + mean-cosine(t).
- Expanded evaluation: Macro-AUC in addition to Accuracy/Precision/Recall/Macro-F1.

Protocol. Data/transforms as above; 8 epochs; optimizers as listed; metrics per §2.4; probes per §2.5; deterministic seeds.

Findings.

- **SGD**: largest rotation, lowest S_w; Macro-F1 below Adam.
- **Adam**: best Macro-F1/AUC, but higher mean-cosine (class means more aligned, less ETF-like).
- **Layca (simplified)**: under-rotates and under-performs, highlighting the importance of the *full* Layca formulation.

5 Critical Observations & Connections to Unit Content

- **Optimization geometry matters.** Mean rotation and orthogonalized updates influence generalization; rotation is a useful *diagnostic*, not a guarantee.

- **Representation geometry matters.** S_w and class-mean angles expose feature quality beyond accuracy. Evaluating **Macro-F1** and **Macro-AUC** alongside geometry aligns with unit themes on balanced metrics and explainable diagnostics.
- **Scale & schedule** strongly affect collapse. ETF-like behaviour generally emerges with larger models, full data, and longer training.

6 Research Gaps & Failure Cases

1. **Under-rotation with naïve Layca.** Implement *full Layca* (layer-wise normalization + target-degree schedule); sweep target degrees (e.g., 1.000° to 3.000°).
2. **Scale/time limitations.** Move to full CIFAR-10 and 100.000 to 120.000 epochs with cosine LR + weight decay to elicit late-phase geometry.
3. **Per-layer heterogeneity.** Try per-layer target-degree schedules, layer-wise LR, or weight-norm constraints.
4. **Geometry vs metrics divergence.** Cases where S_w improves but Macro-F1 does not motivate multi-objective tuning and early-stopping on Macro-F1 (used here) rather than accuracy only.
5. **Augmentation/BN effects.** Ablate stronger augmentation and BN placement; both influence rotation and collapse.
6. **Evaluation breadth.** Add linear-probe and k-NN on frozen embeddings; sweep class imbalance to connect geometry with robustness.

Reproducible Code & Artifacts

All code, figures, and TensorBoard event files are publicly available at ¡https://github.com/Bhavesh2030/SIT744-Assignment2-224085988¿.[1] To reproduce: (i) create a Python ≥3.10 environment and run `pip install -r requirements.txt`; (ii) open the notebook `224085988_SIT744_assignment2_solution.ipynb` and execute cells in order. For the HD tasks, set `ULTRA_FAST=True` for the 8-epoch demo or `False` for longer runs; figures `set4_Sw_over_time.png` and `set4_meanCos_over_time.png` are generated automatically.

References (short)

- Li et al. *Layer rotation: a surprisingly powerful indicator of generalization in deep networks?*
- Papyan, Han, Donoho. *Prevalence of Neural Collapse during the terminal phase of deep learning training.*
- Krizhevsky. *CIFAR-10 dataset.*

Conclusion. With a compact CNN and a simplified Layca, I *partially reproduce* the rotation–generalization link: more rotation (SGD) tracks tighter within-class clusters (S_w), but Macro-F1/AUC peak with Adam at this small scale. ETF-like geometry does not emerge under 2k/1k data and 8 epochs. Implementing full Layca, scaling data/epochs, and using per-layer rotation schedules are the most actionable steps to sharpen collapse signals and clarify when rotation predicts generalization versus when it merely correlates.

[1] Author: Bhavesh Hiraram Choudhary (Student ID: 224085988, s224085988@deakin.edu.au). .