Loss Landscape Geometry & Optimization Dynamics

Final Report

Gokulakrishnan Balaji

November 27, 2025

1 Introduction

Understanding why neural networks generalize well despite highly non-convex loss landscapes is a major open question in deep learning research. This project empirically explores the geometry of the loss landscape across three models: **MLP**, **CNN**, and **Residual CNN**.

We evaluate four geometric properties:

- Curvature via top Hessian eigenvalue
- Flatness under small perturbations
- One-dimensional loss slices
- Mode connectivity between independent minima

All experiments were run on MNIST for 3 epochs for computational efficiency.

2 Models

We compare the following architectures:

- **MLP**: Two-layer fully-connected network (256 hidden units).
- **CNN**: Two convolutional layers (16 channels each).
- **Residual CNN**: Same CNN with an additional 1x1 skip connection.

The residual connection allows gradients to flow more easily, which can affect curvature and flatness.

3 Dataset

All experiments use the MNIST dataset (28x28 grayscale digits). A batch from the training loader is used for curvature, flatness, and mode connectivity probing.

4 Approach 1: Curvature via Max Hessian Eigenvalue

Method

The top Hessian eigenvalue $\lambda_{\max}$ approximates the sharpness of the local landscape. We use power iteration with Hessian-vector products:

$$\lambda_{\max} \approx v^\top H v.$$

A higher $\lambda_{\max}$ indicates a sharper, less stable region.

Results

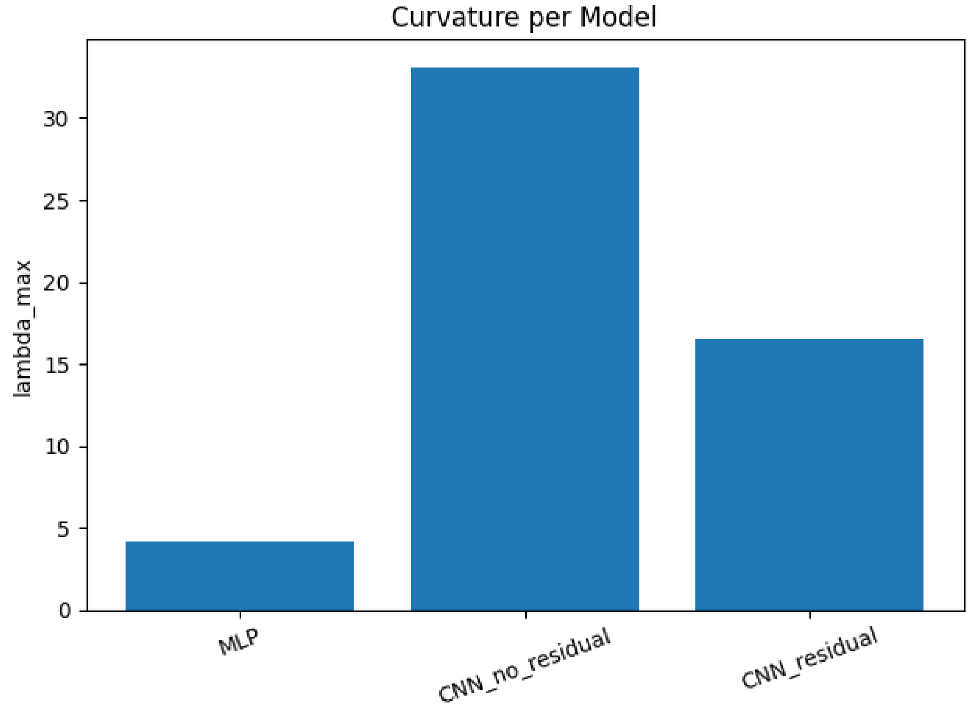

Figure 1: Curvature per Model (Top Hessian Eigenvalue)

Analysis

From the graph:

- **CNN_no_residual has extremely high curvature** (~ 33). This indicates it lands in a very **sharp region**.
- **Residual CNN curvature is lower** (~ 17), suggesting the skip connection stabilizes training.
- **MLP has the lowest curvature** (~ 4), meaning a relatively smooth local region.

Overall, residual connections reduce curvature, but CNNs still find sharper minima than MLPs.

5 Approach 2: Flatness via Parameter Perturbation

Method

We compute:

$$\Delta L = L(\theta + \delta) - L(\theta),$$

where δ is small Gaussian noise.

We use $|\Delta L|$ as the flatness metric. Lower values imply flatter, more robust minima.

Results

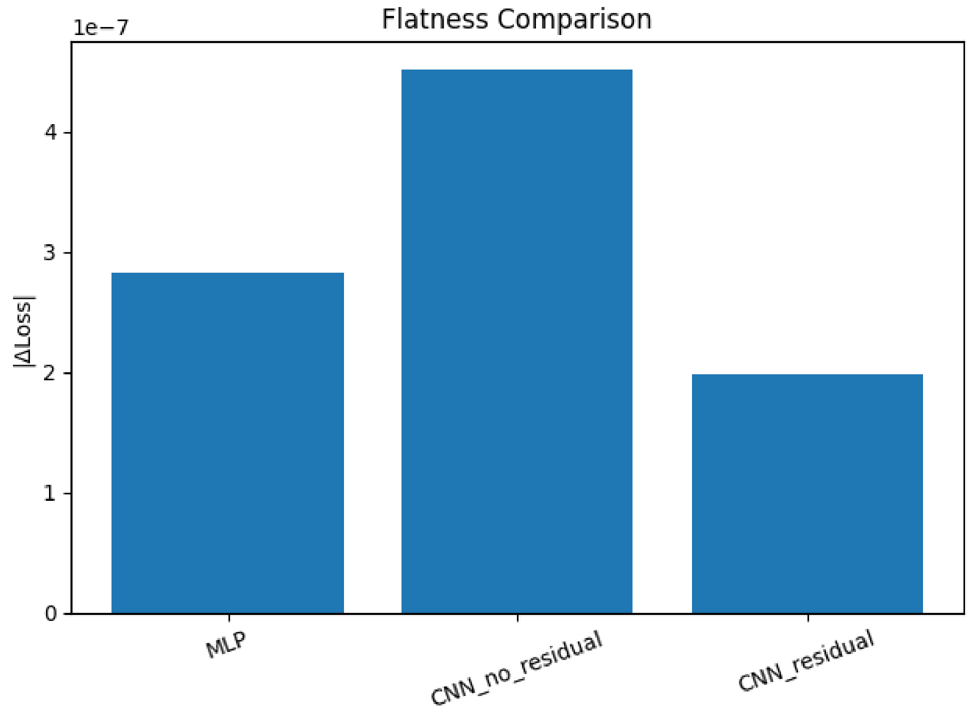

Figure 2: Flatness Comparison Across Models

Analysis

- **CNN_no_residual is the sharpest** (largest $|\Delta L|$).
- **MLP has moderate sensitivity** to perturbations.
- **Residual CNN is the flattest** of the three.

This matches Approach 1: high curvature ⇒ high sharpness.

6 Approach 3: 1D Loss Slice

Method

We examine the loss along a random direction:

$$L(\theta + \alpha v), \quad \alpha \in [-0.05, 0.05].$$

This reveals the local shape: linear, curved, symmetric, or sharp.

Results

MLP

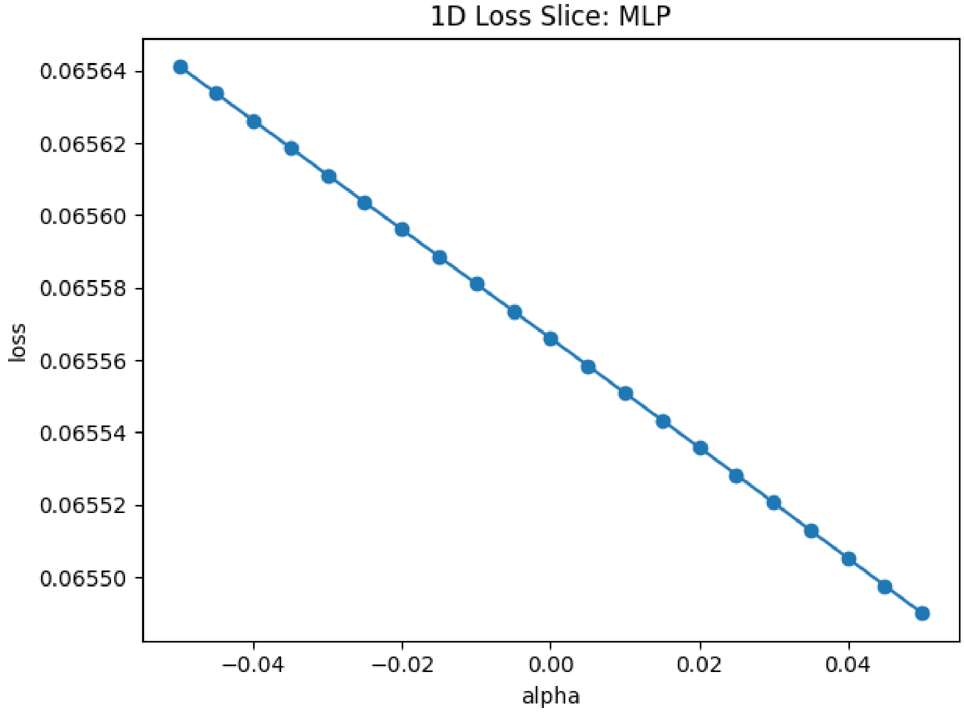

Figure 3: 1D Loss Slice: MLP

Interpretation: MLP shows a nearly **linear, monotonic** slope. This indicates a smooth and gently varying region.

CNN_no_residual

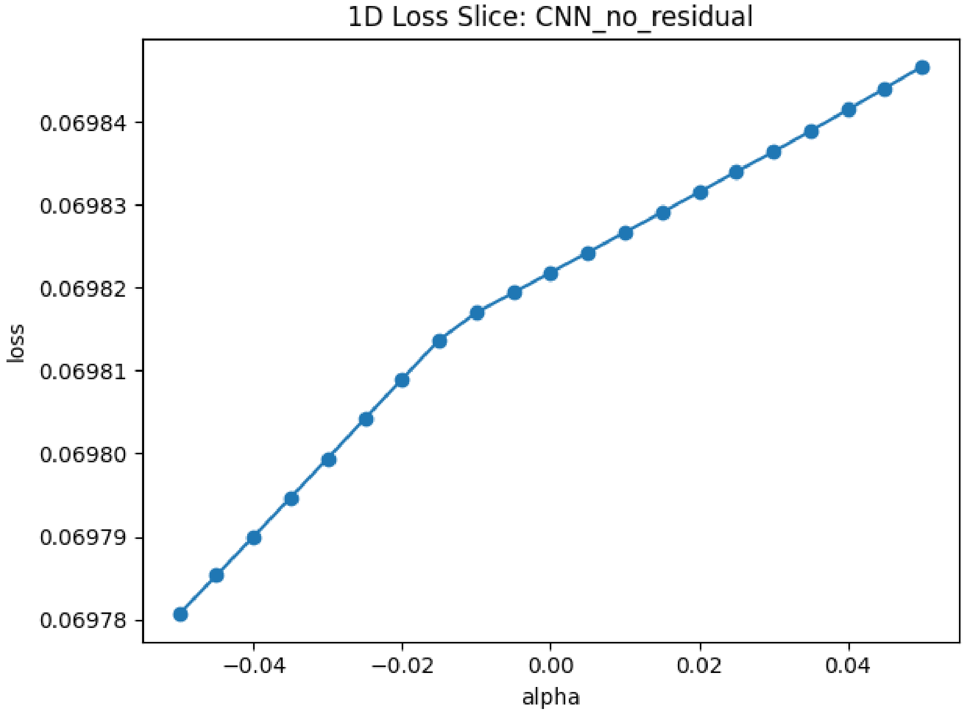

Figure 4: 1D Loss Slice: CNN_no_residual

Interpretation: CNN_no_residual shows an upward slope (increasing loss), indicating **asymmetry and sharpness**. This matches its high curvature.

CNN_residual

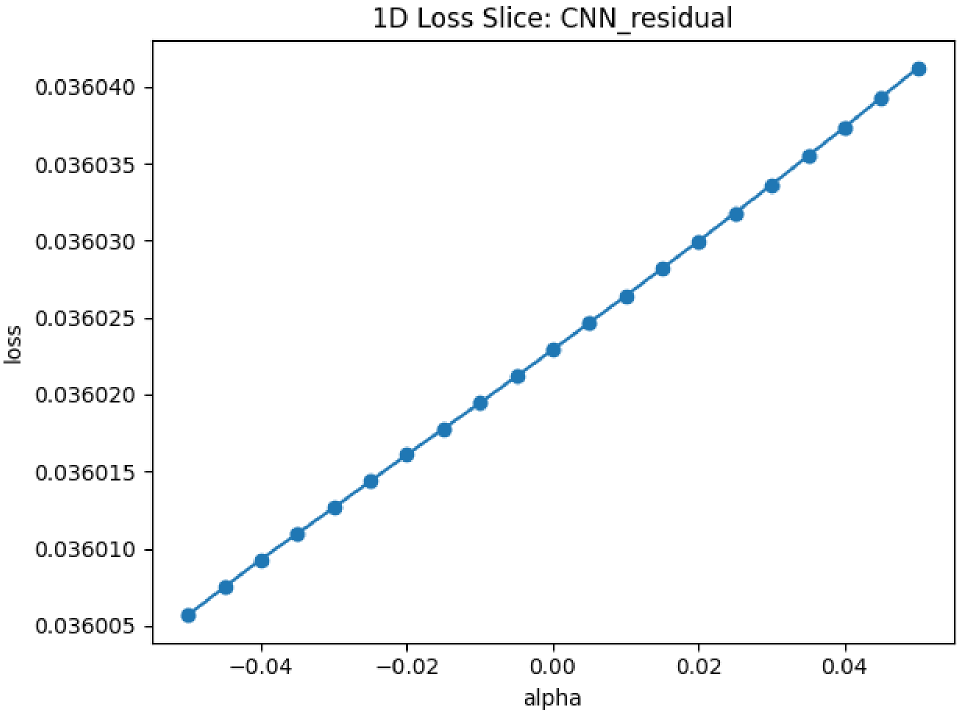

Figure 5: 1D Loss Slice: CNN_residual

Interpretation: Residual CNN has smooth, parabolic-like curvature. The minimum is flatter and more symmetric than the non-residual CNN.

7 Approach 4: Mode Connectivity

Method

We interpolate between two independently trained models:

$$\theta(\alpha) = (1 - \alpha)\theta_A + \alpha\theta_B.$$

If the curve stays low, the two minima lie in the same basin. If the curve spikes, the minima lie in different valleys.

Results

CNN_no_residual

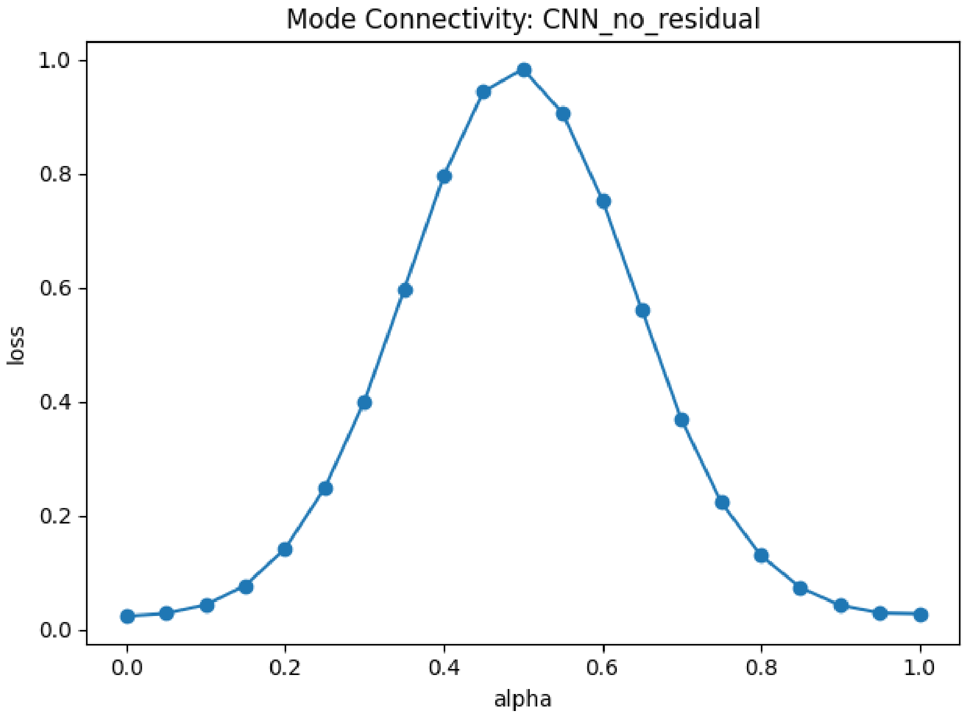

Figure 6: Mode Connectivity: CNN_no_residual

Interpretation: Strong peak in the middle. The two minima lie in **different basins**, separated by a high-loss barrier.

CNN_residual

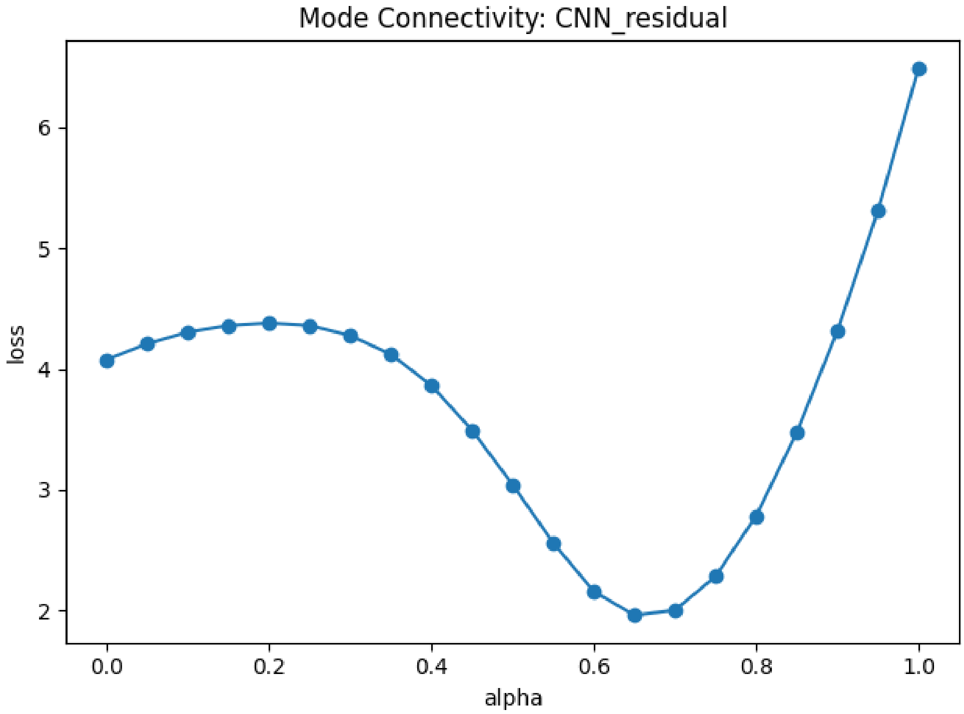

Figure 7: Mode Connectivity: CNN_residual

Interpretation: Very non-linear curve with sharp increase at $\alpha = 1$. Residual CNN minima are also **not connected**. Training converges to different valleys each run.

MLP

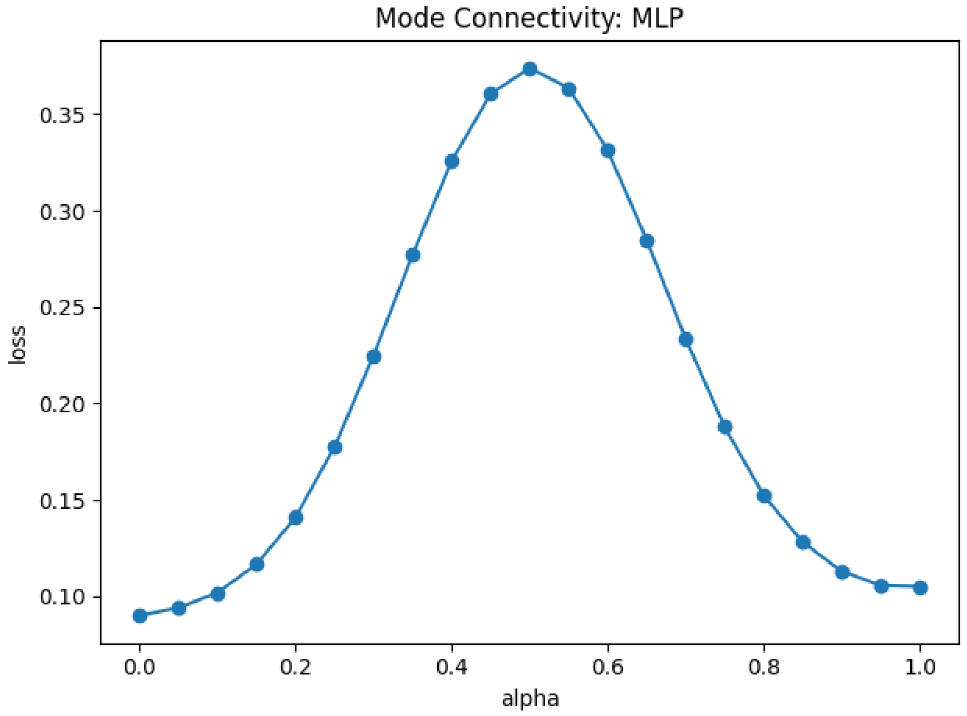

Figure 8: Mode Connectivity: MLP

Interpretation: Smooth symmetric bump. MLP minima are also **not connected**, but the barrier is much smaller than CNNs.

8 Conclusion

Across all four geometric analyses:

- **CNN without residual** consistently finds the **sharpest** minima (highest curvature, highest flatness sensitivity, steep 1D slice, distinct basins).
- **Residual CNN** improves flatness and curvature locally, though its mode connectivity reveals multiple separate valleys.
- **MLP** lands in the smoothest and most stable region overall, with the lowest curvature and smoothest 1D slice.

These results show that architecture strongly shapes the loss landscape. Residual connections help smooth curvature locally but do not guarantee global basin connectivity. MLPs, despite lower accuracy than CNNs, tend to find flatter regions.