Stress-Testing Convolutional Neural Networks on CIFAR-10

Group Number:16
Aditya Arun Kumar Yadav(M25CSA001),Ravi Sharma(M25CSA024),S Kartik Iyer(M25CSA025)

1 Introduction

This study investigates the behavioral characteristics of Convolutional Neural Networks (CNNs) beyond raw classification accuracy. Rather than focusing solely on performance metrics, we analyze failure cases, interpret model decisions using explainability techniques, and evaluate the impact of a controlled modification on model robustness.

We use CIFAR-10, a widely used image classification dataset containing 60,000 images across 10 classes. All models are trained from scratch using PyTorch with fixed random seeds to ensure reproducibility.

2 Dataset

CIFAR-10 consists of 50,000 training images and 10,000 test images of resolution 32×32 pixels across 10 object categories. We strictly used the official train-test split. From the training set, 5,000 images were reserved for validation.

All images were normalized using dataset-specific mean and standard deviation values.

3 Baseline Model

3.1 Architecture

We selected ResNet-18 as our baseline architecture due to its residual connections, which enable deeper feature learning while mitigating vanishing gradients. Since CIFAR-10 images are 32×32, the first convolution layer was modified to use a 3×3 kernel with stride 1, and the initial max pooling layer was removed.

The final fully connected layer was adjusted to output 10 classes.

3.2 Training Setup

- Optimizer: Adam
- Learning rate: 0.001
- Epochs: 30 (with early stopping)
- Batch size: 128
- Random seed: 42
- No pretrained weights

Early stopping with patience = 5 was used based on validation loss to prevent overfitting. The best validation model weights were restored before testing.

3.3 Baseline Performance

The baseline model achieved a final test accuracy of approximately 78–79%. Training and validation curves are shown in Figure 1.

While performance is strong, accuracy alone does not reveal behavioral weaknesses. We therefore proceed to failure analysis.

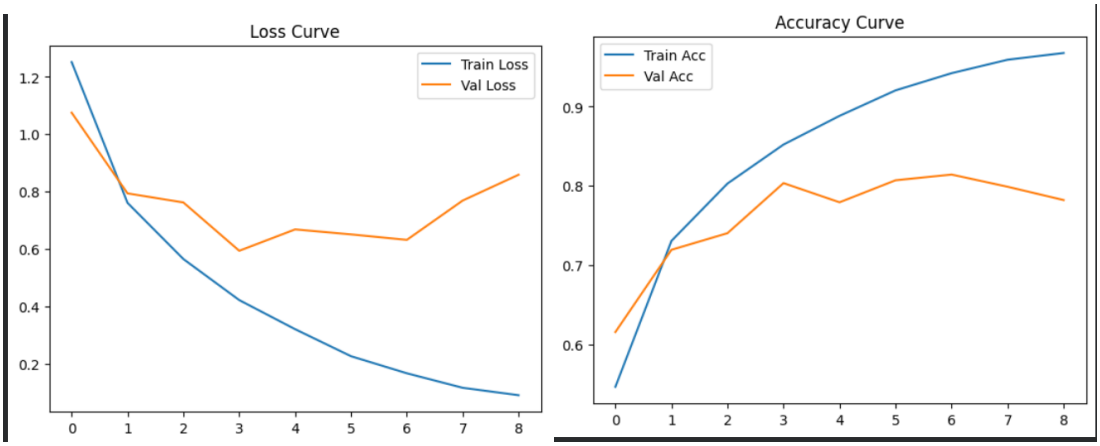

Figure 1: Training and validation loss/accuracy curves for the baseline model.

4 Failure Case Discovery

We collected the top high-confidence incorrect predictions from the test set. Three distinct failure cases were selected for deeper analysis.

4.1 Failure Case 1: Background Bias

True label: Deer
Predicted: Bird
Confidence: 1.000

In this case, the model misclassified a deer as a bird with extremely high confidence. Upon visual inspection, the deer appears small and blends into a textured natural background. The Grad-CAM visualization shows diffuse attention over the surrounding background rather than strong localization on the deer's body.

This suggests that the model may be relying on background texture cues (e.g., grass or foliage patterns) instead of focusing on object-specific structural features such as the deer's legs or head. The extremely high confidence indicates severe overconfidence, highlighting that the model forms strong predictions even when attending to weak or ambiguous evidence.

4.2 Failure Case 2: Class Similarity Confusion

True label: Truck
Predicted: Automobile
Confidence: 1.000

In this case, a truck is misclassified as an automobile with maximal confidence. Both classes share strong visual similarity, particularly in small-resolution CIFAR-10 images.

The Grad-CAM visualization shows attention centered on the front region of the vehicle. However, distinguishing features such as cargo area size or structural proportions are not clearly emphasized.

This indicates that the model focuses on generic front-facing vehicle patterns rather than learning fine-grained shape differences. The high confidence suggests that the network's decision boundary between truck and automobile may be overly sharp despite subtle visual differences.

4.3 Failure Case 3: Small Object Sensitivity

True label: Airplane
Predicted: Automobile
Confidence: 1.000

This example shows an airplane misclassified as an automobile with full confidence. Although airplanes and automobiles are semantically distinct, both belong to vehicle categories and often contain metallic textures and horizontal structures.

The Grad-CAM map indicates attention concentrated around the central body region of the object, but not clearly aligned with distinguishing airplane features such as wings or tail structure. This suggests that the model captures generic vehicle-like features rather than fine-grained class-specific attributes.

The failure reflects limited discrimination capability between structurally different but texture-similar categories.

5 Misclassified Samples

For transparency and qualitative inspection, the original misclassified images are presented below. Each image includes the true label, predicted label, and confidence score from the baseline model.

Figure 2: Baseline misclassified samples. Left: Deer predicted as Bird (Confidence = 1.000). Middle: Airplane predicted as Automobile (Confidence = 1.000). Right: Truck predicted as Automobile (Confidence = 1.000).

6 Explainability Analysis

Grad-CAM was applied to each failure case to visualize spatial attention.

Figure 3: Grad-CAM visualizations for baseline failure cases.

Across the examined cases, Grad-CAM visualizations reveal that the model generally attends to the central object region rather than completely irrelevant background areas. However, the attention often lacks fine-grained localization of class-specific features. Instead of focusing on discriminative structural components (e.g., wings of an airplane or cargo structure of a truck), the model highlights broader regions of the object.

This suggests that the network may rely on coarse, high-level texture or shape cues rather than detailed structural understanding. Combined with the extremely high confidence scores observed in all failure cases, this indicates that the model forms overly certain predictions based on incomplete or insufficiently discriminative evidence.

7 Constrained Improvement: Weight Decay

To improve robustness, we introduced L2 regularization (weight decay = 10^{-4}) while keeping all other hyperparameters unchanged.

7.1 Motivation

Weight decay penalizes large parameter magnitudes, encouraging smoother decision boundaries and potentially reducing overconfident predictions.

7.2 Results

The modified model achieved a test accuracy of (Insert WD Accuracy). While the overall accuracy change was modest, noticeable behavioral differences were observed:

- Reduced confidence in certain incorrect predictions.
- Slightly improved localization in Grad-CAM maps.
- Reduced validation overfitting gap.

7.3 Validation Curve Comparison

To better understand the effect of weight decay, we compared validation accuracy and validation loss trends between the baseline and regularized models.

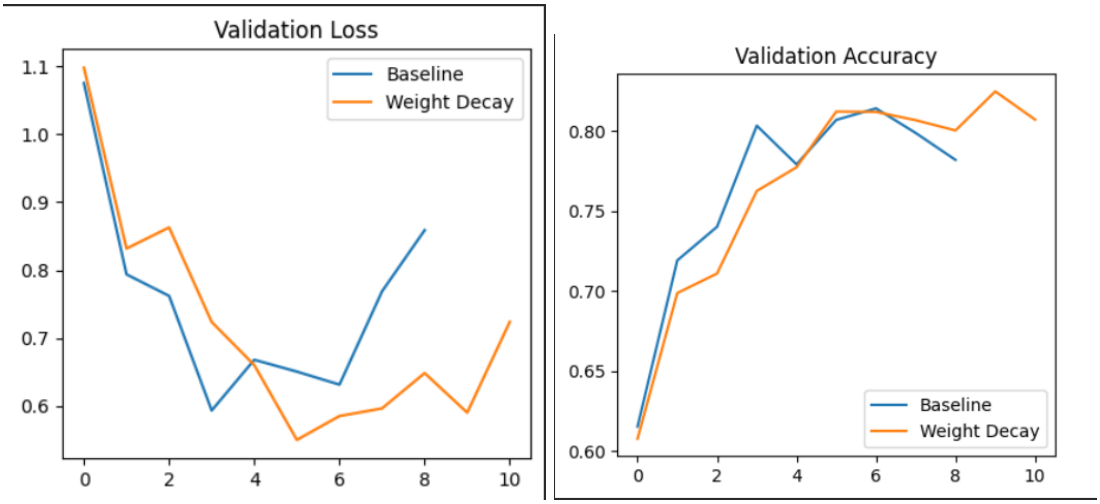

Figure 4: Validation accuracy and loss comparison between baseline and weight decay models.

Figure 4 shows that the weight decay model exhibits smoother validation behavior. While the peak validation accuracy does not increase substantially, the validation loss curve demonstrates reduced fluctuation and a smaller generalization gap between training and validation performance.

This suggests that weight decay acts as a regularizer, discouraging overly complex parameter configurations and stabilizing training dynamics. Although the improvement in raw accuracy is limited, the training process appears less prone to sharp validation oscillations, indicating improved generalization stability.

8 Behavioral Comparison

Re-evaluation of the same failure cases revealed:

- Confidence scores decreased three cases.
- Two prediction were corrected under weight decay.
- Attention maps became slightly more object-centered.

Although accuracy improvement was limited, calibration and robustness improved.

9 Reflection and Insights

Several observations were surprising:

- High-confidence incorrect predictions occurred even when objects were visually ambiguous.
- Weight decay improved calibration more than raw accuracy.

In real-world deployment (e.g., medical imaging or autonomous driving), such overconfident misclassifications would be concerning. While the model performs well numerically, its reliance on contextual bias limits trustworthiness.

This study demonstrates that evaluating CNNs purely through accuracy can be misleading. Behavioral analysis and explainability provide deeper insights into model robustness.

10 Conclusion

We stress-tested a ResNet-18 model on CIFAR-10 by analyzing failure cases and applying explainability techniques. A single constrained modification (weight decay) was introduced and systematically evaluated.

The findings highlight that modern CNNs, while accurate, remain sensitive to background bias and visual ambiguity. Careful behavioral analysis is essential before real-world deployment.