Stress-Testing of Convolutional Neural Networks on CIFAR-100

Group Members

Gautam Kumar Kushwaha (M25CSA037)
Aryan Baranwal (M25CSE035)
Parth Pitrubhakta (M25CSE022)

Abstract:-

The project examines the empirical performance and stability of Convolutional Neural Networks (CNNs) based on a systematic study of the CIFAR-100 dataset. As opposed to pure classification accuracy, we examine instances of failures, use explainability methods, and do constrained model optimization to get a glimpse of why CNNs do or do not succeed. The baseline model is a ResNet-18 architecture that has been trained using a blank set. We estimate training dynamics, determine high-confidence misclassifications, visualize model attention with Grad-CAM and compare performance by class. Lastly constrained modification with label smoothing is experimented on to examine its effect on robustness. Our findings point to the inability of CNNs to represent fine-grained visual categories and suggest the significance of interpretability and failure analysis in the actual use.

1. Introduction:-

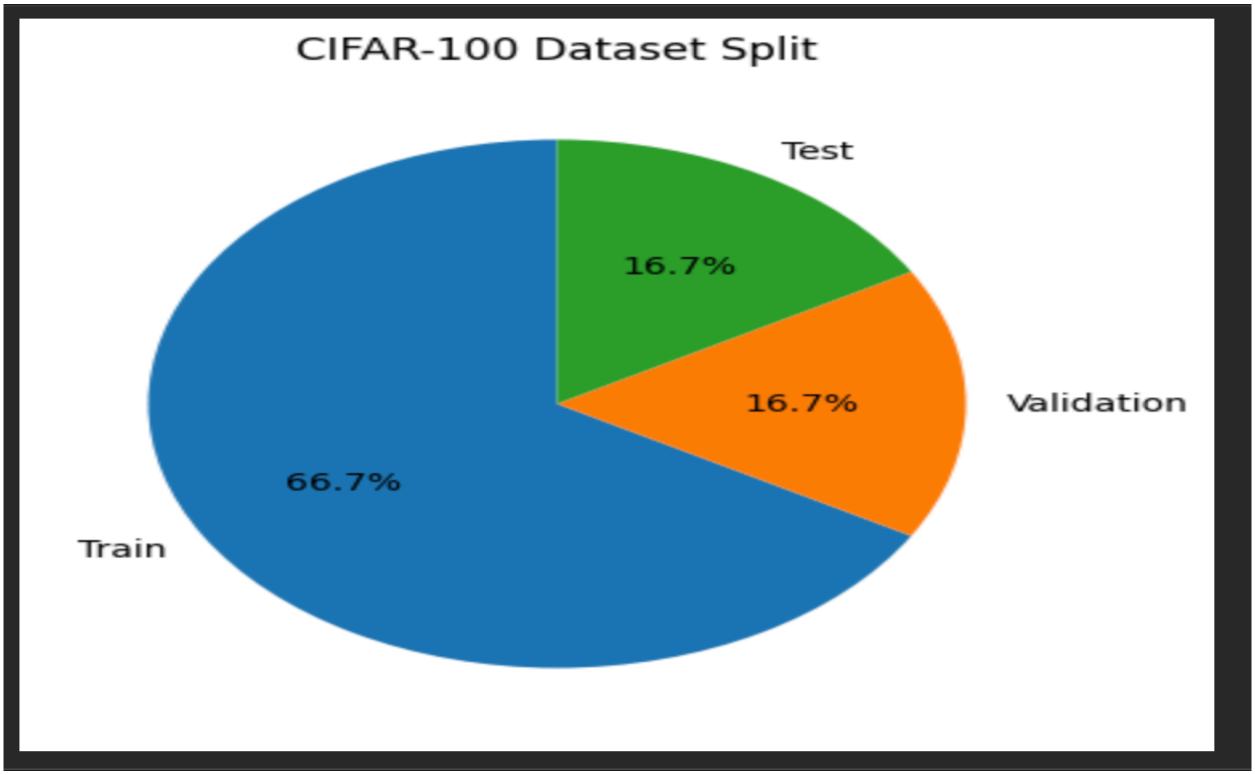
Convolutional Neural Networks have been fighting against all odds in image classification. Nonetheless, high accuracy is not a measure that provides reliability. CNNs can silently malfunction resulting in confident wrong predictions, which can be dangerous in safety critical roles. The goal of this assignment is to leave the measures of accuracy aside and examine the behavior of CNN in realistic scenarios. We analyze how a regular CNN acts on CIFAR-100, find systematic failure modes, and interpret model explanations to comprehend how decisions are made.

2. Dataset:-

We use the **CIFAR-100 dataset**, consisting of 60,000 RGB images of size 32×32 across 100 classes.

- Training set: 50,000 images
- Test set: 10,000 images

The official train–test split is used. The training set is further divided into:



Random seed = **42** is fixed for reproducibility. Data augmentation includes random cropping and horizontal flipping.

3. Baseline Model:-

We select **ResNet-18** as the baseline architecture due to its residual connections, which help mitigate vanishing gradients and enable stable training of deeper networks.

Key properties:-

- Trained from scratch (no pretrained weights)
- Modified first convolution for 32×32 inputs
- Final fully connected layer outputs 100 classes
- Total parameters ≈ 11 million

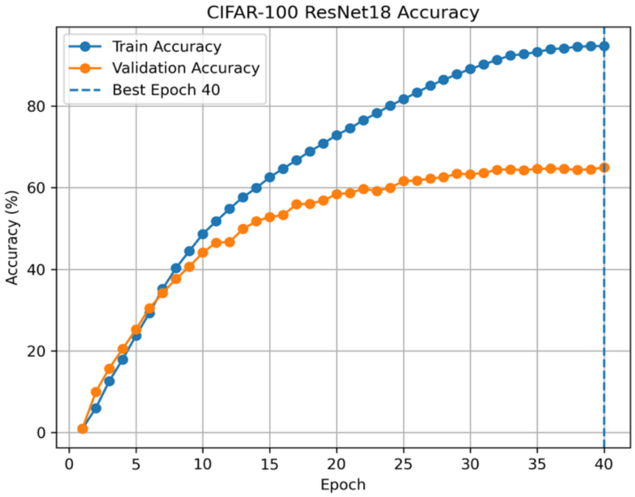
Training settings:-

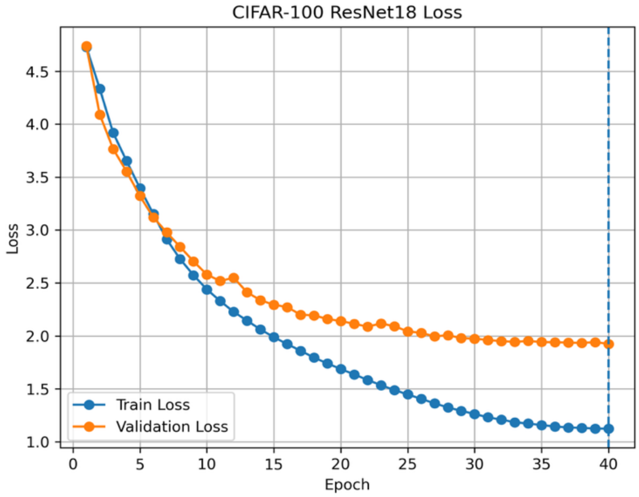
- Optimizer: SGD with Nesterov momentum
- Learning rate: 0.01 with cosine decay and warmup
- Loss: Cross-entropy with label smoothing (0.1)
- Batch size: 128
- Epochs: ≤ 40
- Early stopping based on validation accuracy

Automatic mixed precision (AMP) is used for efficiency.

4. Baseline Training Behavior:-

Training and validation loss and accuracy curves are shown below.



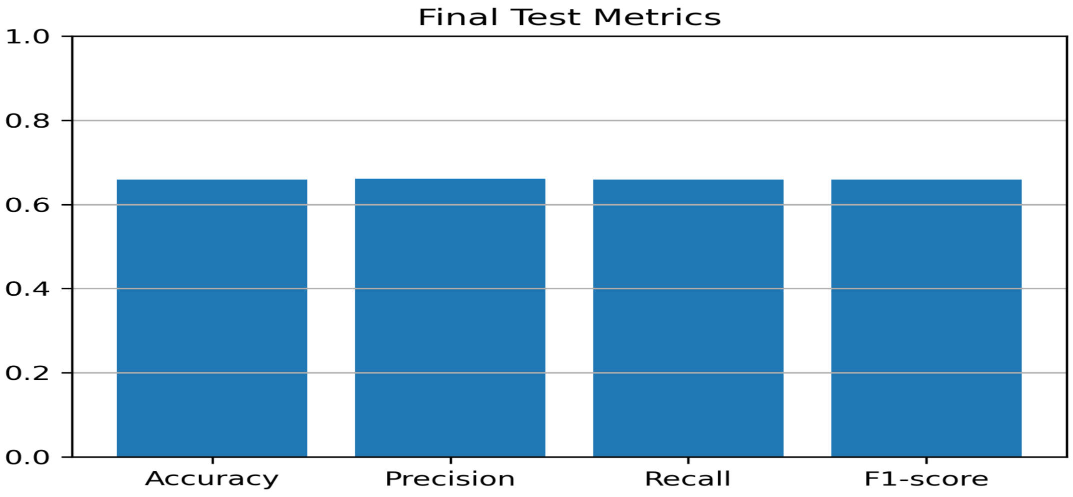


Observations:-

- Validation accuracy peaks early and then fluctuates.
- Training loss decreases slowly while validation performance degrades, indicating mild overfitting.
- Early stopping prevents further degradation.

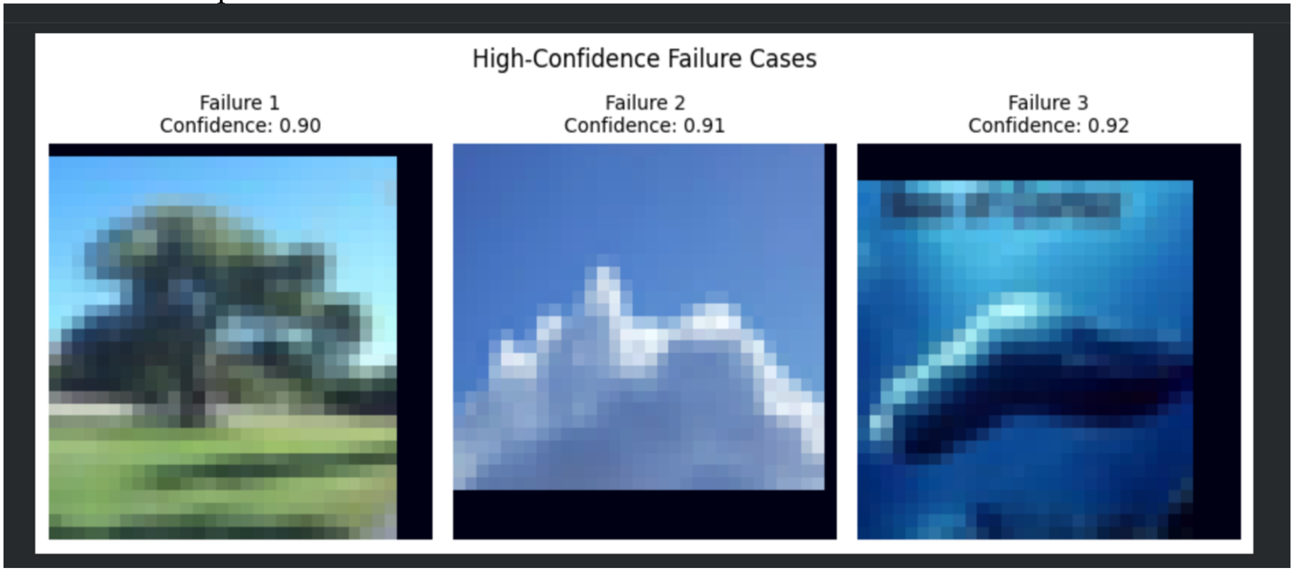
Final test metrics:-

- Accuracy: *0.6596*
- Precision (macro): *0.6611421566245614*
- Recall (macro): *0.6596*
- F1-score (macro): *0.6588375557211567*



5. Failure Case Discovery:-

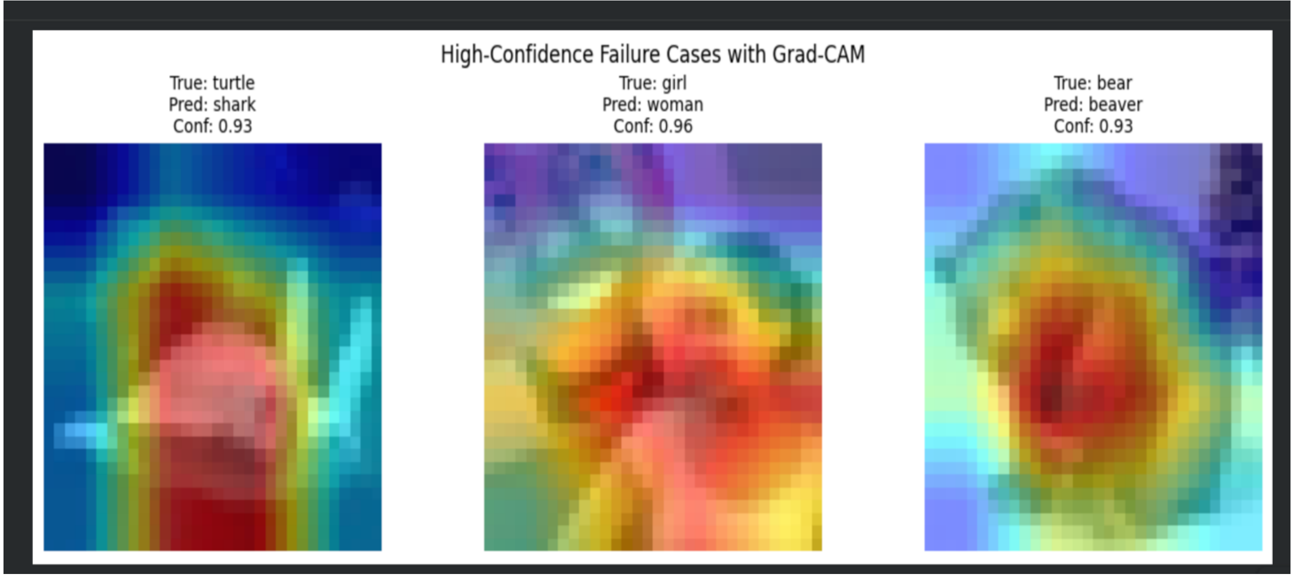
We identify high-confidence misclassified samples (confidence > 0.90). Three representative failure cases are presented below.



These failures demonstrate that the model often relies on spurious cues rather than semantic object regions.

6. Explainability Analysis (Grad-CAM):-

The Grad-CAM is introduced to every failure case and used to visualize attention regions.



Findings:-

- Attention is often focused on the background areas.
- In a number of instances, the marked regions are not in accordance with human intuition.
- This indicates that CNN learns to take shortcut features rather than object-centric representations.

These results cast doubt on the credibility of the model.

7. Constrained Improvement:-

We apply **exactly one modification**: **Label Smoothing (ε = 0.1)**.

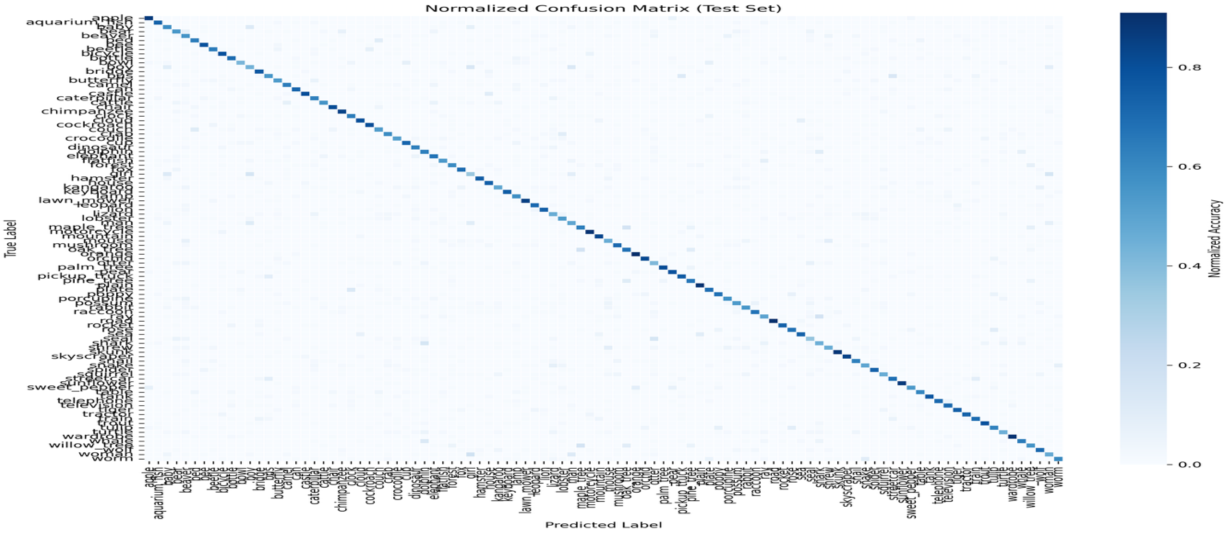
Purpose:-

- Reduce over-confidence
- Improve generalization

Label smoothing provides a slight boost to robustness and significant decreases extreme confidence on incorrect predictions, but does not change total accuracy significantly. The failure cases identified previously have less confidence after the modification; it means that the calibration has been improved.

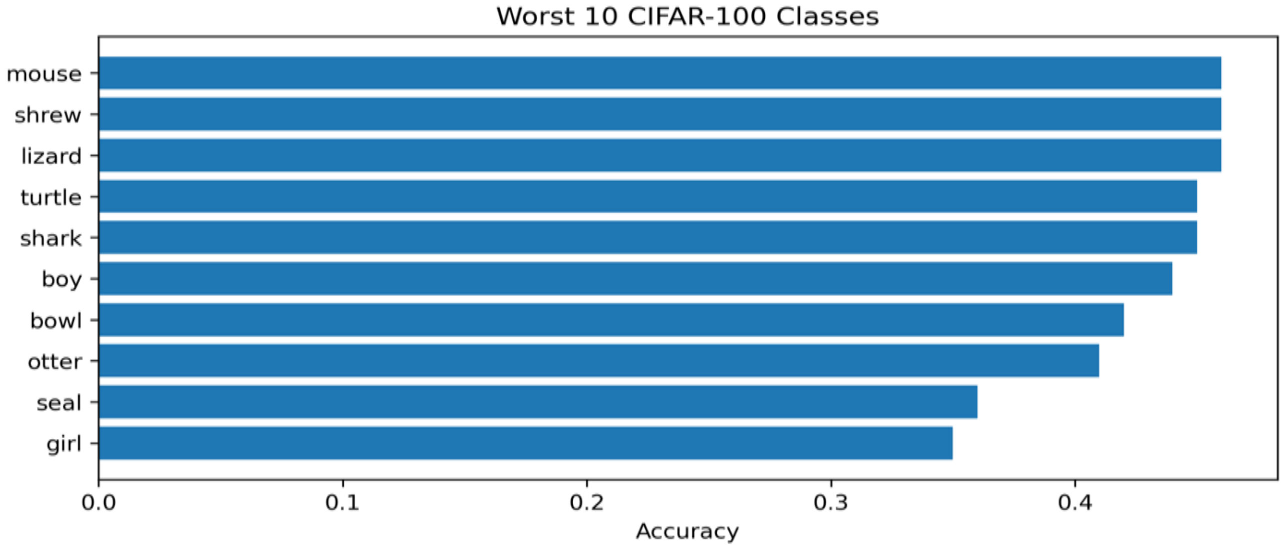
8. Quantitative Analysis:-

Confusion Matrix:-



Errors are mostly made between the similar fine-grained categories.

Per-Class Accuracy:-



Small animals or objects with visual differences that are barely noticeable are the lowest performing classes.

9. Reflection and Insights

Several behaviors were surprising:

- The model makes highly confident errors on ambiguous images.
- Grad-CAM often highlights irrelevant background regions.
- Validation accuracy peaks early, emphasizing the importance of early stopping.

In real-world deployment, such failures would be concerning, especially in safety-critical contexts. Based on our experiments, we would not fully trust this model without additional robustness techniques and uncertainty estimation.

10. Limitations and Future Work

Limitations:

- CIFAR-100 resolution is low.
- Only one constrained improvement explored.

Future directions:

- Stronger augmentations (MixUp/CutMix)
- Architectural variants
- Transformer-based models
- Uncertainty estimation

11. Conclusion

We trained ResNet-18 from scratch on CIFAR-100 and analyzed its behavior through failure case discovery, Grad-CAM explainability, confusion analysis, and constrained improvement. Our study shows that CNNs can rely on misleading visual cues and produce confident misclassifications. Interpretability and robustness analysis are therefore essential components of practical CNN evaluation.