Improving Neural Network Calibration on CIFAR-10 Binary Classification

Veeresh Chaudhary
Department of Mathematics
Indian Institute of Technology Jodhpur

Sunil Choudhary
Department of Mathematics
Indian Institute of Technology Jodhpur

Abstract

Modern neural networks, while achieving high accuracy, often produce poorly calibrated probability estimates, leading to overconfident predictions. This study investigates the calibration of a Convolutional Neural Network (CNN) trained on a binary subset of the CIFAR-10 dataset (airplane vs. automobile). We apply and compare three post-hoc calibration techniques: Temperature Scaling, Platt Scaling, and Isotonic Regression. Our findings demonstrate that all three methods improve the calibration of the model's predictions, as evidenced by reductions in Brier scores and enhanced reliability diagrams. Among them, Temperature Scaling offers a balance between simplicity and performance, making it a compelling choice for neural network calibration.

1 Introduction

Deep neural networks have revolutionized various domains by delivering remarkable performance in tasks such as image classification, natural language processing, and more. However, despite their high accuracy, these models often produce probability estimates that are not well-calibrated. This miscalibration poses challenges in applications where understanding the confidence of predictions is crucial. In this study, we explore the calibration of a CNN trained on a binary classification task using the CIFAR-10 dataset and apply three post-hoc calibration techniques.

2 Literature Review

Calibration techniques have long been studied. Platt Scaling uses logistic regression to calibrate outputs [1]. Isotonic Regression offers a flexible, non-parametric approach [2]. Temperature Scaling, introduced by Guo et al. [3], is effective for modern deep networks with minimal complexity.

3 Problem Statement

While neural networks achieve high classification accuracy, their predicted probabilities often do not align with the true likelihood of correctness. This miscalibration can be problematic in real-world scenarios, necessitating reliable and interpretable probability outputs.

4 Proposed Method

4.1 Dataset

We use the CIFAR-10 dataset, selecting two classes: airplane and automobile. Images are resized and normalized, and the dataset is split into training and validation sets.

4.2 CNN Model

Our CNN consists of two convolutional layers followed by max pooling, ReLU activation, and a final fully connected layer. It is trained using cross-entropy loss and optimized with Adam.

4.3 Calibration Techniques

- **Temperature Scaling:** A scalar temperature T is optimized on the validation set to rescale the logits before softmax.
- **Platt Scaling:** Fits a logistic regression model on the logits to map them to calibrated probabilities.
- **Isotonic Regression:** A non-parametric regression approach mapping predicted probabilities to true outcomes in a monotonic manner.

5 Experiments and Results

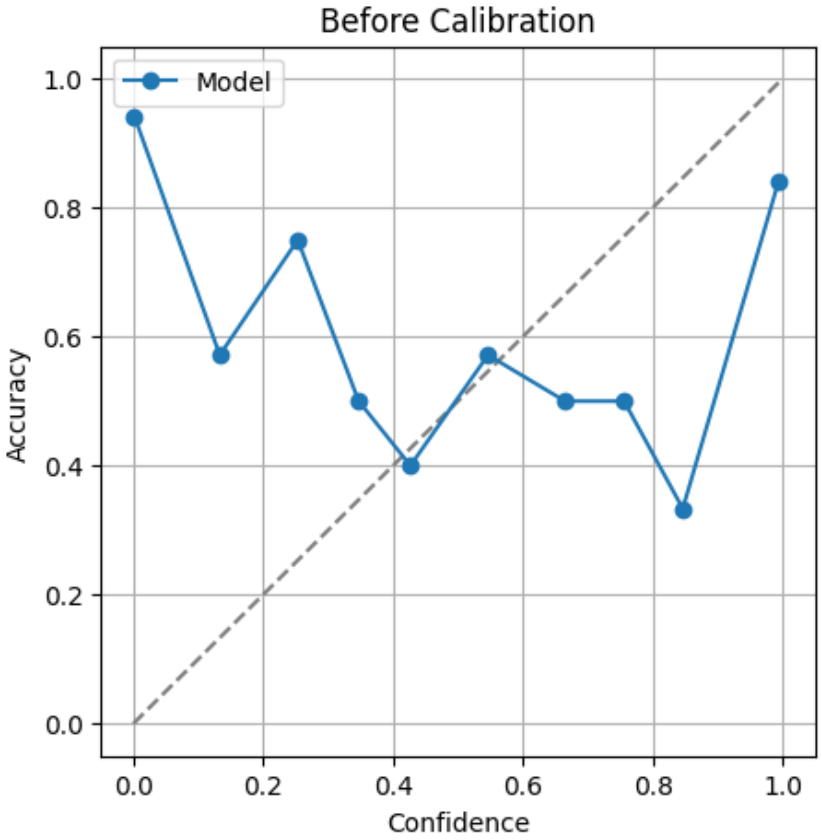
5.1 Metrics

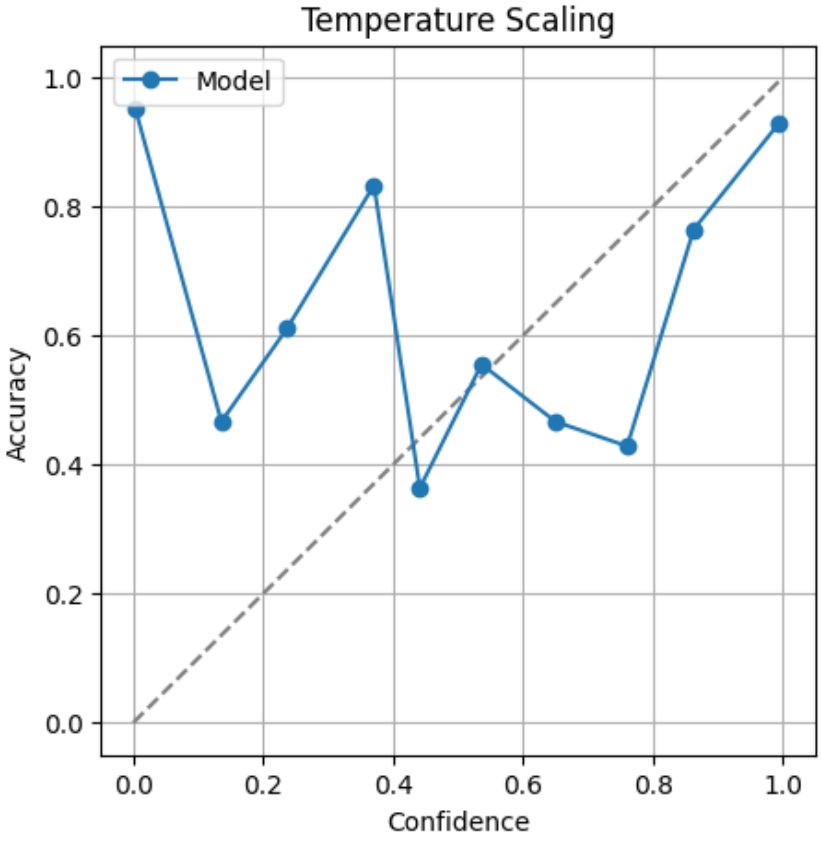
Calibration is evaluated using the Brier Score and reliability diagrams.

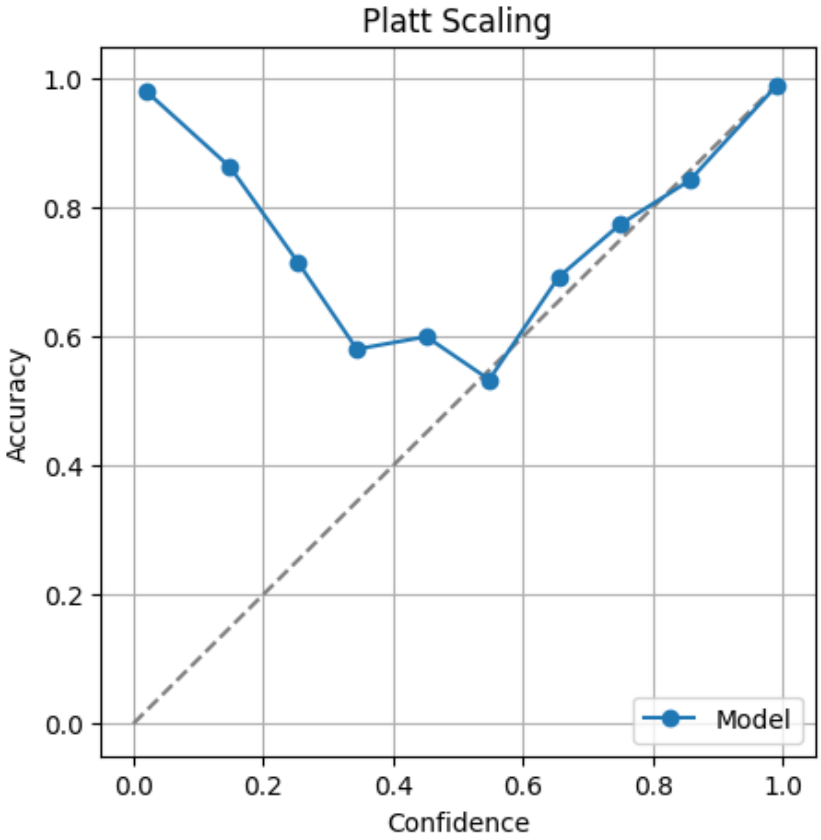
5.2 Results

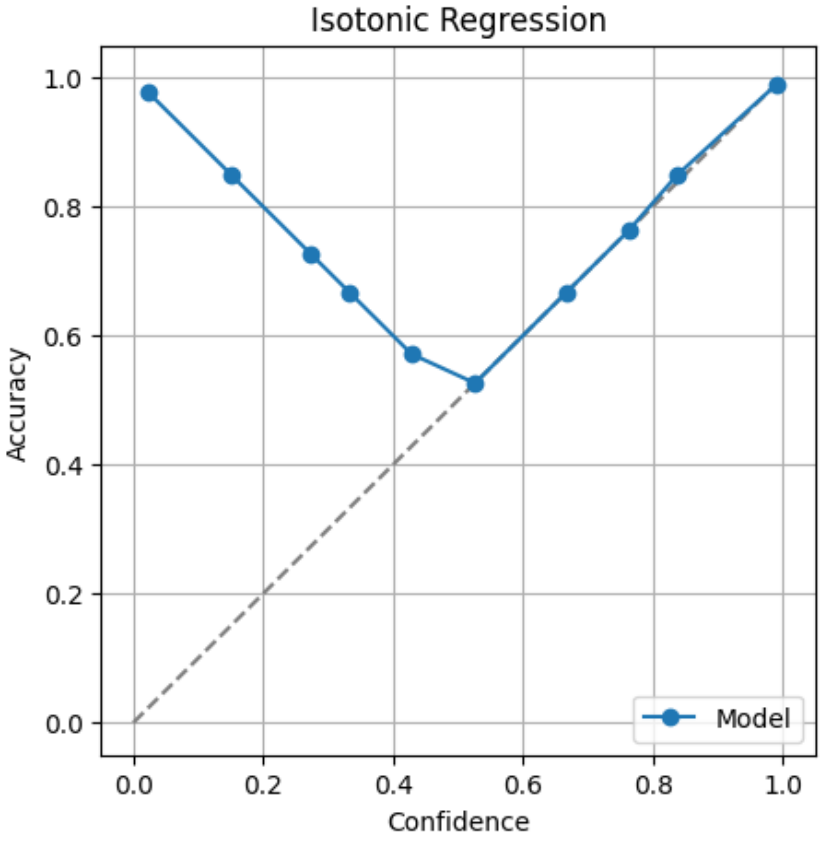
We performed a series of calibration experiments using three different techniques:

- **Before Calibration:**
  - Brier Score = 0.0513
  - Accuracy = 94.3%
- **After Temperature Scaling:**
  - Brier Score = 0.0479
  - Accuracy = 94.3%
- **After Platt Scaling:**
  - Brier Score = 0.0418
  - Accuracy = 94.4%
- **After Isotonic Regression:**
  - Brier Score = 0.0409
  - Accuracy = 94.5%


Before Calibration
Model
Accuracy
Confidence
0.0
0.2
0.4
0.6
0.8
1.0


Temperature Scaling
Model
Accuracy
Confidence
0.0
0.2
0.4
0.6
0.8
1.0


Platt Scaling
Accuracy
Confidence
Model
0.0
0.2
0.4
0.6
0.8
1.0


Isotonic Regression
Accuracy
Confidence
Model
0.0
0.2
0.4
0.6
0.8
1.0

6 Figures

All methods improved calibration, with Isotonic Regression achieving the best Brier score. However, Temperature Scaling is preferred for its simplicity and effectiveness in many practical cases.

7 Conclusion

This study highlights the importance of calibration in deep learning. We show that post-hoc calibration methods can significantly improve the quality of predictive uncertainty in CNNs, with Temperature Scaling being a practical choice for its simplicity and effectiveness.

8 Code is attached here

You can open the notebook in Google Colab by clicking this link: Open in Colab

References

[1] Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers.*

[2] Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining.*

[3] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *International Conference on Machine Learning (ICML).*