Reproducing Calibration Results for Deep Neural Networks

Analysis of Guo et al. (2017) and Extensions

Madhur (25M0742) Monil (25M0744) Anurag (25M0770)
Aditya (25M0745)

CS725: Foundations of Machine Learning
IIT Bombay

November 25, 2025

Overview

Problem Statement

- Modern neural networks have achieved very high accuracy.
- However, they can become significantly *miscalibrated*.

What is Calibration?

- A model is calibrated if predicted confidence aligns with observed accuracy.
- Example: If a model predicts 100 images with 80% confidence, ≈ 80 of them should be correct.
- Perfect calibration is defined by:

$$P(\hat{Y} = Y \mid \hat{P} = p) = p, \quad \forall p \in [0, 1]$$

Goal: Reproduce Guo et al. (2017) to understand this phenomenon and evaluate post-processing calibration methods.

Why Miscalibration Occurs?

The paper identifies architectural trends responsible for overconfidence:

1. **Depth and Width:** Higher capacity models overfit to the NLL loss, pushing probabilities towards 1 (overconfidence) even after classification error saturates.
2. **Normalization:** Batch Normalization improves convergence but can disrupt logit scale.
3. **Weight Decay:** Modern training often uses less regularization, allowing models to fit training distributions too closely.

Why Miscalibration Occurs? (From the Guo et al. paper)

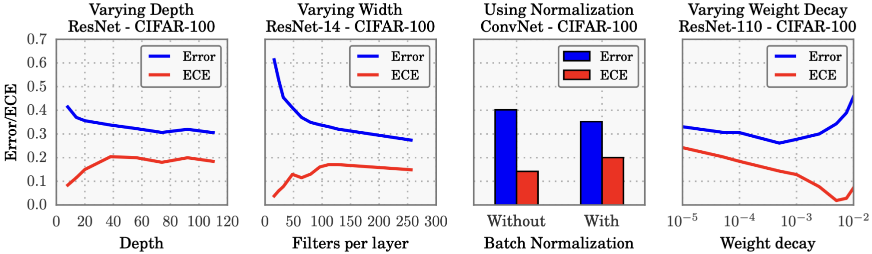


Figure: The effect of network depth (far left), width (middle left), Batch Normalization (middle right), and weight decay (far right) on miscalibration, as measured by ECE (lower is better).

Datasets and Models

We chose the following dataset-model pairs to capture the essence of the paper's findings including some latest models:

Dataset	Model	Type
CIFAR-100	ResNet-56	Standard
CIFAR-100	ResNet-164	Deep
CIFAR-100	DenseNet-190	Deep
CIFAR-100	WideResNet-28-10	Wide
CIFAR-10	ResNet-56	Standard
CIFAR-10	ResNet-164	Deep
Stanford Cars	**MobileNetV2**	**Fine-grained**
Birds-400	**InceptionV3**	**Fine-grained**

Table: Note the inclusion of MobileNetV2 and InceptionV3.

Calibration Metric: ECE

Expected Calibration Error (ECE)

- We partition predictions into M fixed bins.
- ECE is the weighted average of the difference between Accuracy and Confidence in each bin.

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |\mathrm{acc}(B_m) - \mathrm{conf}(B_m)| \tag{1}$$

- **acc(B_m)**: Average accuracy of samples in bin m.
- **conf(B_m)**: Average confidence of samples in bin m.
- **n:** $|Dataset|$

Calibration Methods (Post-Processing)

We implemented four methods requiring a held-out validation set:

1. **Histogram Binning (Non-Parametric):**
   - Assigns calibrated probability θ_i to each bin based on validation accuracy.
   - *Con:* Changes predicted probabilities for all classes; non-continuous.

$$\theta_m = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}(\hat{y}_i = y_i) \tag{2}$$

2. **Isotonic Regression (Non-Parametric):**
   - Learns a piecewise constant, monotonically increasing function.
   - *Con:* Prone to overfitting on small validation sets.

$$\min_f \sum_{i=1}^{N} (f(\hat{p}_i) - y_i)^2 \quad \text{subject to} \quad \hat{p}_i \leq \hat{p}_j \implies f(\hat{p}_i) \leq f(\hat{p}_j) \tag{3}$$

Calibration Methods (Post-Processing)

1. **Vector Scaling (Parametric):**
    - Applies linear transformation $\mathbf{Wz} + \mathbf{b}$ to logits ($\mathbf{W}$ is diagonal).
    - *Con:* $2K$ parameters can lead to overfitting.

$$\hat{q} = \max_k \sigma_{SM}(\mathbf{Wz} + \mathbf{b})^{(k)} \tag{4}$$

2. **Temperature Scaling (TS):**

$$\hat{q}_i = \max_k \sigma_{SM}(\mathbf{z}_i / T)^{(k)} \tag{5}$$

    - Single scalar parameter $T > 0$ for all classes.
    - Optimized via NLL on validation set.

Key Advantages:

- **Preserves Accuracy:** Does not change the maximum of the softmax.
- **Efficiency:** Only 1 parameter to learn.
- **Effectiveness:** Smoothens the distribution, avoiding overconfidence.

Calibration Performance (ECE %)

Dataset	Model	Uncalibrated ECE	Temp. Scaling (TS)	Hist. Binning	Isotonic Reg.	Vector Scaling
CIFAR-100	ResNet-164	15.07%	1.36%	**1.33%**	1.90%	1.99%
CIFAR-100	ResNet-56	15.87%	2.68%	2.03%	**1.94%**	3.45%
CIFAR-100	DenseNet-190	7.34%	2.88%	**1.01%**	1.61%	3.45%
CIFAR-100	WideResNet-28-10	6.45%	3.53%	1.53%	**1.32%**	4.27%
CIFAR-10	ResNet-164	3.75%	0.95%	**0.79%**	0.95%	1.18%
CIFAR-10	ResNet-56	3.96%	1.21%	**0.358%**	0.84%	1.63%
Stanford Cars	MobileNetV2	9.42%	2.20%	**1.26%**	1.52%	2.94%
Birds 400	InceptionV3	0.50%	0.85%	**0.354%**	0.82%	0.50%

Table: Comparison of ECE (%) across all models and calibration methods.

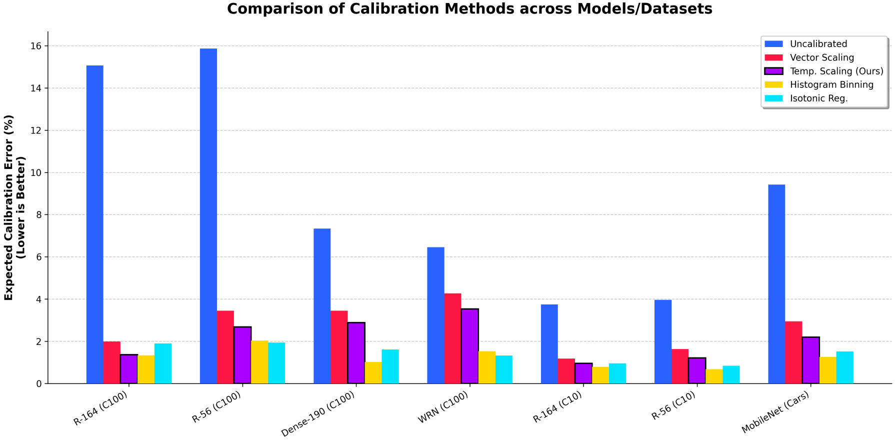
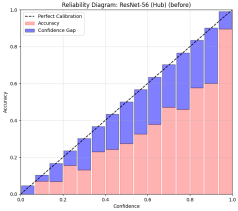


Figure: Bar chart comparing the ECE reduction across different methods. Temperature Scaling (Blue) consistently provides excellent reduction compared to the uncalibrated baseline (Gray) and outperforms Vector Scaling (Orange).

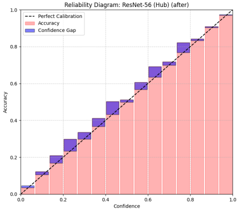
Visual Analysis: Reliability Diagrams

Before Calibration
(ResNet-56, CIFAR-100)



Gap indicates Overconfidence

After Temperature Scaling
(ResNet-56, CIFAR-100)

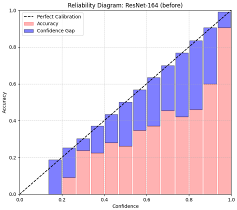


Better Diagonal Alignment

Note: In uncalibrated models (left), accuracy (blue) is consistently lower than confidence (red gap).

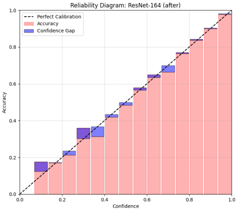
Visual Analysis: Reliability Diagrams

Before Calibration
(ResNet-164, CIFAR-100)



Gap indicates Overconfidence

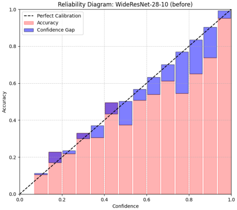
After Temperature Scaling
(ResNet-164, CIFAR-100)



Better Diagonal Alignment

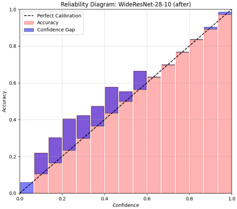
Visual Analysis: Reliability Diagrams

Before Calibration
(WideResNet-28-10, CIFAR-100)



Gap indicates as mix of Over & Under confidence

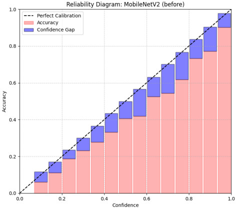
After Temperature Scaling
(WideResNet-28-10, CIFAR-100)



Better Diagonal Alignment

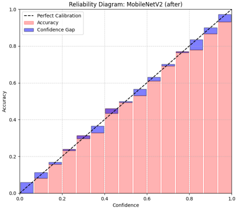
Visual Analysis: Reliability Diagrams

Before Calibration
(MobileNetV2, Birds-400)



Gap indicates Overconfidence

After Temperature Scaling
(MobileNetV2, Birds-400)



Better Diagonal Alignment

Key Analysis

1. **Hypothesis Confirmed:** Deep models (ResNet-164) are indeed highly miscalibrated (ECE $\approx 15\%$).
2. **Temperature Scaling Performance:** TS matched or neared the performance of non-parametric methods (Histogram Binning) without the complexity of bin selection.
3. **Vector Scaling Fails:** Consistently worse than TS. The extra parameters ($2K$) led to overfitting on the validation set.
4. **The InceptionV3 Exception:** InceptionV3 on Birds was already calibrated (ECE 0.50%). Post-processing methods actually degraded it slightly.

Extra Work: Label Smoothing (LS)

Hypothesis: Can we fix calibration during training instead of post-processing?

Method:

- Replaced one-hot targets with soft targets (mass α distributed to other classes).
- Fine-tuned ResNet-56 on CIFAR-10.

Results (Negative):

- **Accuracy Drop:** 94.12% $\rightarrow$ 92.90%.
- **Calibration Degraded:** ECE increased 3.96% $\rightarrow$ 6.05%.

Conclusion: Re-training with LS was destructive compared to the safe, post-processing nature of Temperature Scaling.

Conclusion

- **Reproduction Successful:** We verified that increased depth/width leads to miscalibration in modern DNNs.
- **Temperature Scaling is Robust:** Effective across diverse architectures (ResNets, DenseNets, MobileNets).
- **Simplicity $>$ Complexity:** Simple scalar scaling outperformed vector scaling and complex training modifications (Label Smoothing).

Thank You!