Model Evaluation & Calibration

Analysis of ResNet-18 & DenseNet-121 on CIFAR Datasets

EMRE SAYGIN - 120200069

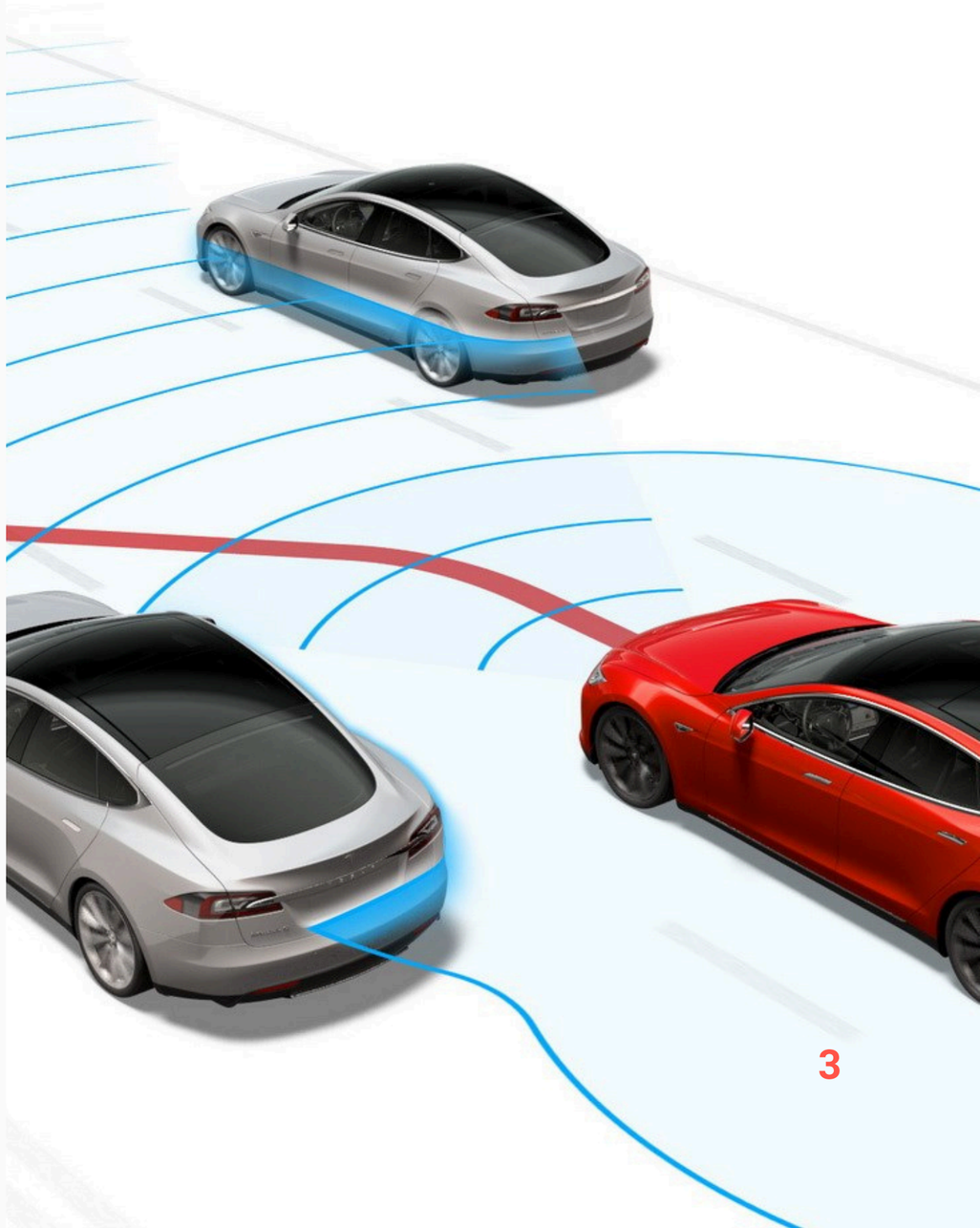
Table of Contents

The Problem & Motivation

HIGH ACCURACY ≠ HIGH RELIABILITY

The Risk: Overconfident predictions in safety-critical systems (e.g., Autonomous Driving, Healthcare).

Goal: Measure and improve confidence using Calibration.

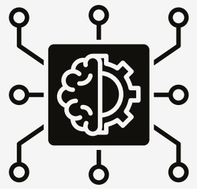


Methodology

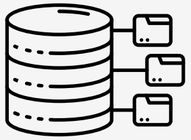
Experimental Framework: All models were implemented using PyTorch and trained from scratch on a local NVIDIA GTX 1650 Ti GPU to ensure a controlled environment.

Data Splitting Strategy: A crucial 90/10 split was applied. 10% of the training data was isolated as a Validation Set strictly to learn the calibration parameter (T), preventing data leakage.

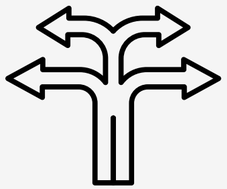
Standardization: To ensure a fair comparison, both architectures shared identical hyperparameters (10 Epochs, Batch Size 64, SGD Optimizer).



Models:
ResNet-18 vs.
DenseNet-121



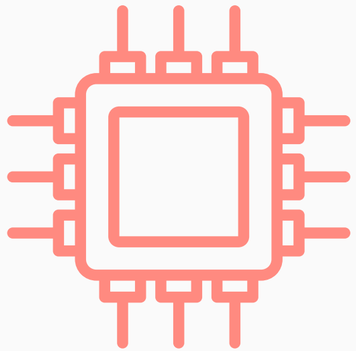
Dataset:
CIFAR-10 and
CIFAR-100



Technique:
Temperature Scaling
(Post-processing)

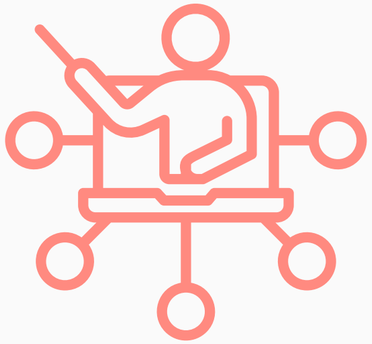
Metrics:
Accuracy, ECE, NLL

Experimental Setup



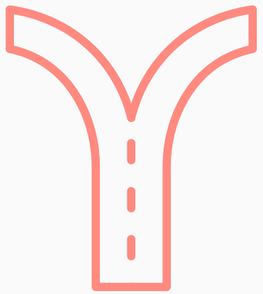
HARDWARE

Local Machine
(NVIDIA GTX 1650 TI)



TRAINING

10 EPOCHS
Batch Size: 64
Optimizer: SGD



SPLIT

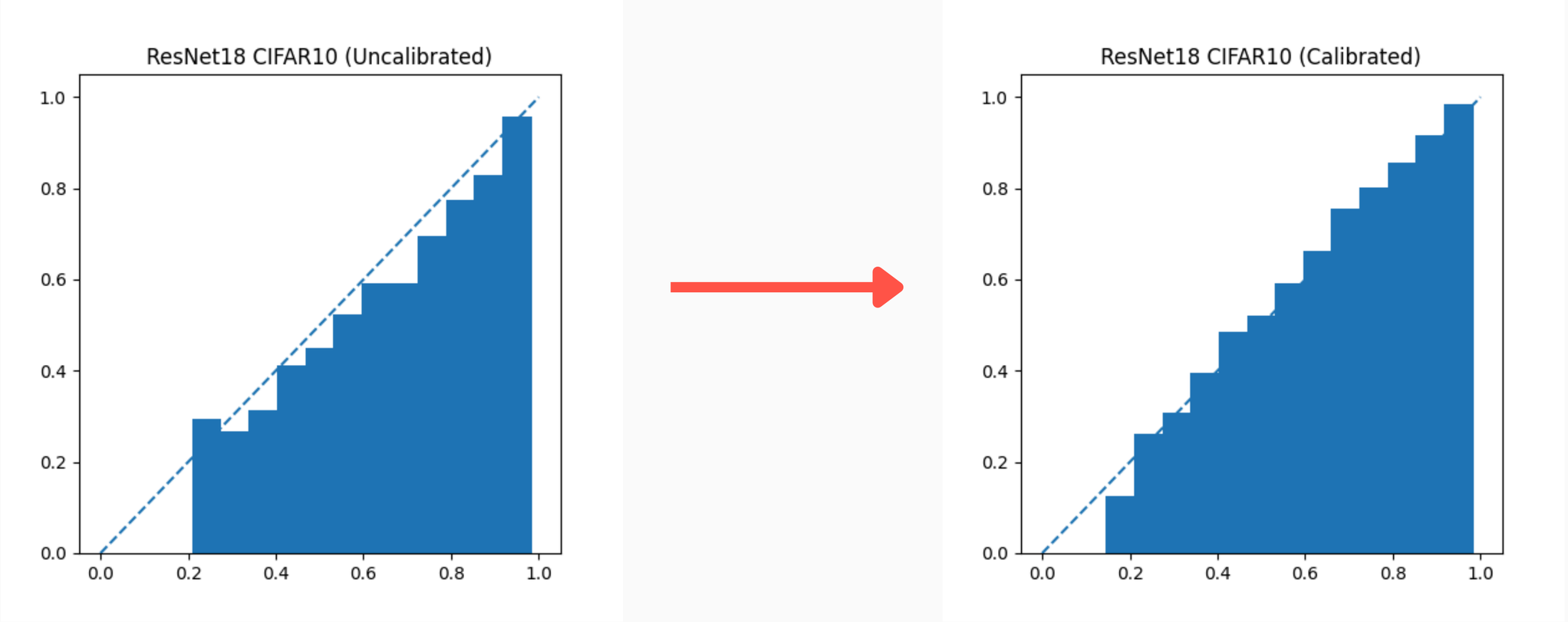
90% Train
10% Validation

Quantitative Results

The following table presents a comparative analysis of calibration performance. We focus on the reduction of Expected Calibration Error (ECE) and Negative Log Likelihood (NLL) after applying Temperature Scaling. The Best T column indicates the learned temperature parameter, where $T > 1$ confirms the presence of initial overconfidence.

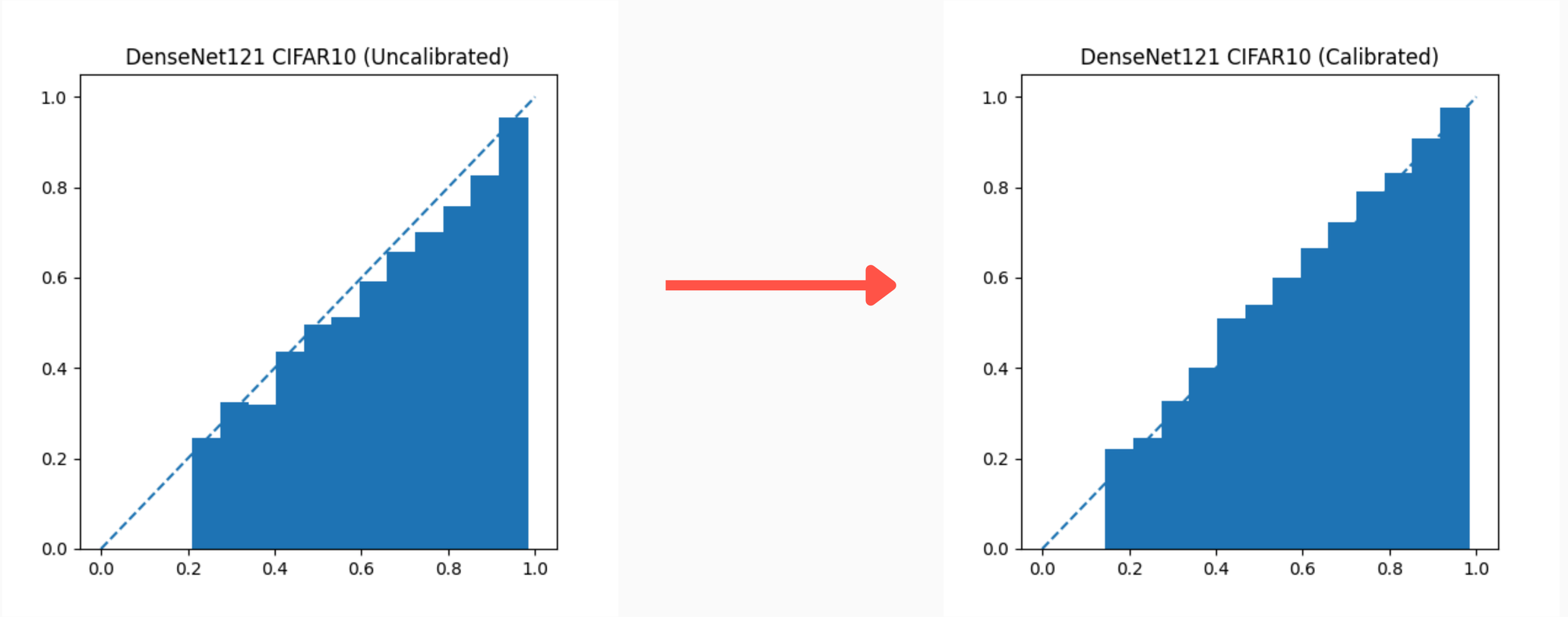
Model	Dataset	ECE Pre	ECE Post	NLL Pre	NLL Post	Best T
ResNet-18	CIFAR-10	0.0403	**0.0400**	0.7779	**0.7708**	1.3134
ResNet-18	CIFAR-100	**0.0494**	0.0580	**2.1925**	2.1956	1.2807
DenseNet-121	CIFAR-10	**0.0334**	0.0466	0.7210	**0.7133**	1.2735
DenseNet-121	CIFAR-100	**0.0406**	0.0605	**2.0357**	2.0451	1.2607

Visual Analysis



The transition demonstrates how Temperature Scaling softens the initial mild overconfidence, aligning the confidence bars more closely with the ideal diagonal line.

Visual Analysis



DenseNet-121 exhibits an inherently stable confidence profile, where Temperature Scaling provides a subtle fine-tuning effect to further improve alignment with the diagonal.

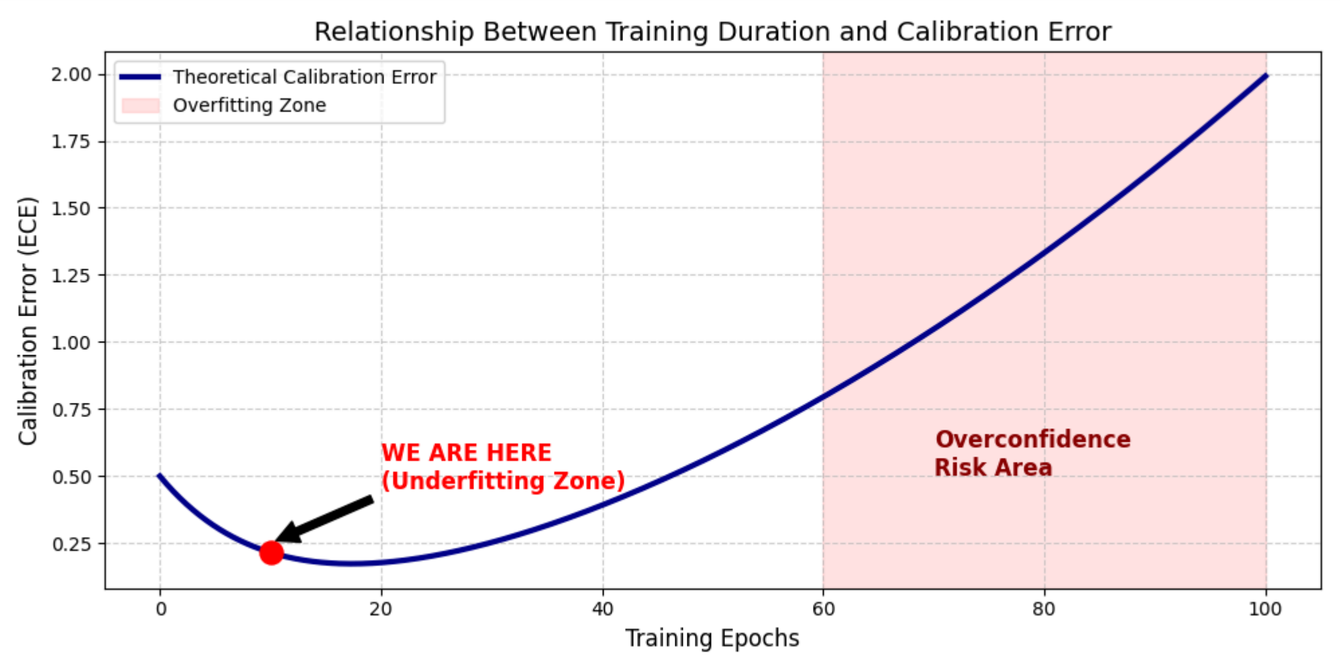
Critical Analysis

Why only marginal improvement?

- **Hypothesis:** Underfitting / Early Learning Phase

Insight: Overconfidence typically appears after overfitting (prolonged training).

Conclusion: Models trained for 10 epochs are not “arrogant” enough to need heavy calibration.



Conclusion

ARCHITECTURE

Comparing the architectures, DenseNet-121 consistently achieved lower NLL scores than ResNet-18. This suggests that its dense connectivity structure may inherently produce more robust probability estimates, especially in the early stages of training.

TAKEAWAY

Our primary finding is that calibration techniques are context-dependent. While Temperature Scaling is a powerful tool, it yields the most significant improvements on fully converged models where overfitting has led to severe overconfidence, rather than on early-stage models.

The End

Thank You!

EMRE SAYGIN