Universidad Carlos III de Madrid
Bachelor's in Data Science and Engineering

Calibration in CNN's

Neural Networks

Name:
Carlos Casanova Romero
Ricardo Vazquez Alvarez

Professor:
Pablo Martinez Olmos

April 9, 2024

Understanding Calibration in CNN's

1.1 Introduction

Calibration in Convolutional Neural Networks is very important for ensuring reliable predictions by aligning predicted probabilities with confidences. Understanding calibration enhances the reliability and trustworthiness of CNN models in their applications. In this project we will investigate the calibration of two CNNs for classifying birds and cats in the CIFAR-10 dataset. Based on the paper *"On calibration of model neural networks"*.

The analysis of calibration will be done with the reliability diagram which plots the expected sample accuracy as a function of confidence. If the model is perfectly calibrated the diagram should plot a perfect diagonal, with any deviation representing miscalibration. Furthermore, as a quantitative measure we have the ECE or Expected Calibration Error which measures the discrepancy between predicted probabilities and accuracies across different confidence levels.

To further reduce and perfect calibration there are many methods such as Histogram binning, Isotonic regression, ... In this project we will delve into **temperature scaling** or simple Platt's scaling which is a parametric approach that given the logit vector zi, the new confidence probabilities are $p(a \cdot z_i)$, being a the temperature parameter. With a = 1 recovering the original probabilities. Due to the fact that temperature scaling does not change the maximum of the softmax function, it does not affect the model's accuracy, only its calibration.

We then first prepare and download the CIFAR-10 dataset by filtering the indexes of the classes 2 and 3 which are cats and birds. Obtaining the train, validation and test loaders. To further study the calibration of the models, in this assignment we will start studying compact LeNet-5 and then delve into VGG model as a much larger convolutional NN.

1.2 LeNet-5

We start by defining the structure and model of a LeNet-5 where on the forward method we do not return the probabilities using a log_softmax function but the logits obtained from the classifier. This is due to the future temperature scaling to apply on the probabilities to reduced the Expected Calibration Error.

When training and validating the LeNet-5 model we achieve a validation loss of 0.48 and a accuracy of a 78%.

To compute the reliability diagram and ECE score, we first extract the confidence scores and true labels from the model's evaluation performance method. The confidence scores are obtained by exponentiating the log probabilities, as we are using a log_softmax function. These scores represent the model's confidence in its predictions. The true labels are the actual labels corresponding to the input data.

For the reliability diagram, we utilize the calibration_curve function from sklearn. This function takes the confidence scores and true labels as inputs and returns the true probabilities and predicted probabilities for a specified number of bins. These probabilities serve as the axes for the reliability diagram.

To compute the ECE score, we define a function that calculates the absolute difference between the predicted and true probabilities, and then takes the weighted average of these differences.

The results for the LeNet-5 model are the following: Accuracy of 78% and ECE of 0.0236 which is very very small. Sort of a perfect calibrated model.

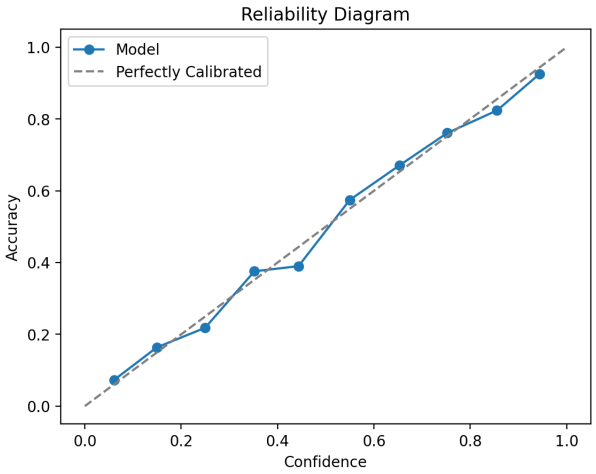

Figure 1.1: LeNet-5 Reliability Diagram and ECE score

1.2.1 Temperature Scaling

We implement a basic temperature scaling (Platt's Scaling), where the output probability $p(z_i)$ is modified by $p(a \cdot z_i)$, where a is the temperature constant chosen to smooth the output probabilities. In the evaluation performance method, we calculate the probabilities of the scaled logits.

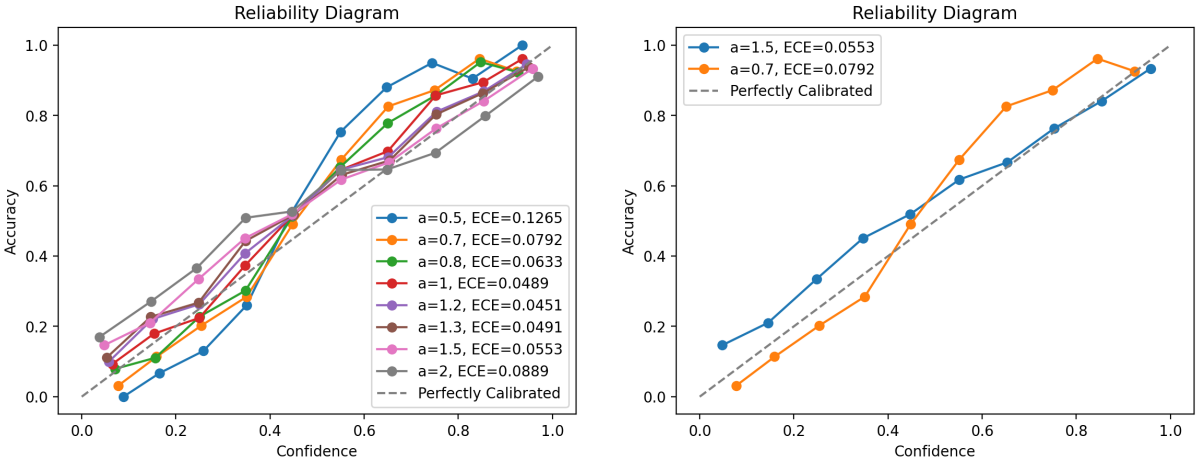

Figure 1.2: LeNet-5 Reliability Diagrams showcasing differences on temperature parameter

- Small Values of a (a $<$ 1):

  The calibrated probabilities become more confident, but this results in overconfidence. It compresses the range of the logits, making the softmax probabilities sharper.

  The reliability diagram shows underestimation of uncertainty.

- Large Values of a (a $>$ 1):

  Larger values of a yield smoother probability curves. Spreads out the logits, resulting in smoother probability curves after the softmax, reducing the sharpness of the predicted probabilities.

  The calibrated probabilities become more conservative, reducing overconfidence.

1.2.2 Conclusions

The LeNet-5 model achieved a very low ECE score, indicating nearly perfect calibration. Consequently, it is challenging to further reduce the ECE score through temperature scaling. However, by applying a temperature parameter of 1.2, we managed to decrease the ECE score by 0.002.

When showcasing significant differences using different parameter values, we observe that setting a = 0.7 results in greater overconfidence in the predictions, while a = 1.5 produces a smoother curve in the reliability diagram, thereby reducing overconfidence for true values.

1.3 VGG net

In this part we apply the same calibration investigation to the task, but using a much larger model, namely the Visual Geometry Group network consisting of 16 layers, which has already been pre-trained.

To begin, we modify the Classifier layer to fulfill our task, ensuring it returns the necessary logits. We then train the model and apply temperature scaling to obtain its reliability diagram and ECE score.

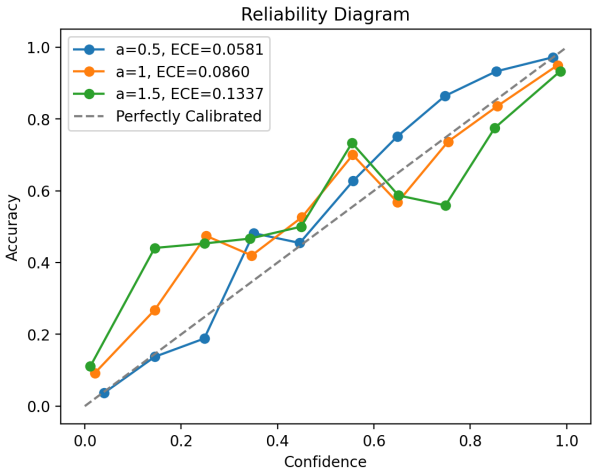

Figure 1.3: VGG Reliability Diagram and ECE scores

It achieves a total accuracy of 84.4%, which is higher than that of the LeNet-5 model. Additionally, it has an ECE of 0.086, which is remarkably good for a model of this size and accuracy. When applying temperature scaling, we observe that the temperature parameter of 0.5 increases overconfidence in some parts of the curve, resulting in a reduction of 0.03 in the ECE, indicating an improvement.