Effects of Calibration on Neural Network Conformal Prediction Accuracy

Ryan Lee

Received November 26, 2024
Accepted March 19, 2025
Electronic access April 15, 2025

Conformal prediction is a machine learning framework that allows a machine learning model to output a set of predictions containing multiple classes, generally with a coverage guarantee that ensures the output will contain the desired class with a certain probability. This achieves two things: (1) it increases the chance that the model will output the correct class, even if it is not what the model believes is most likely or is not very confident in any of its answers, and (2) it quantifies the model's uncertainty for different classes. Proposed algorithms for conformal prediction can be affected by different calibration techniques and result in changes in conformal accuracy. Therefore, the present work analyzed the effects of common calibration techniques on a convolutional neural network (ResNet-34) trained on the computer vision benchmark CIFAR10 dataset with conformal algorithms and evaluated the accuracy of different combinations of algorithms. The results showed that using the regularized adaptive predictive sets algorithm (RAPS) along with label smoothing, tends to serve conformal prediction purposes best. Additionally, the findings showed that Platt scaling is a useful tool for eliminating overconfidence, although it doesn't help with creating reasonable set sizes. These findings may help create more accurate conformal machine learning models.

Introduction

Making prediction models more trustworthy at test time is of critical importance, especially when using them in high-stakes settings, such as providing reliable medical diagnoses, or other use cases where a safety net may be desired (e.g., word prediction algorithms). Numerous techniques have been invented to do this, broadly falling into the categories of variance reduction[1–4], which attempts to prevent model overfitting, calibration[5,6], which tries to make outputted probabilities closer to their true values, and conformal prediction[7,8], which allows multiple outputs to create consistency. This paper focused on the latter two and how they interact with each other either beneficially or detrimentally.

Conformal prediction is a technique that allows a model to return multiple prediction results to meet a coverage guarantee α, which is the probability that the model will output the correct class in its output set[8]. This is helpful in situations where either quantifying the likelihood of potential classifications is useful, such as in medical diagnoses, or when returning more results based on uncertainty is beneficial, for example in prediction algorithms. Accurate probabilities for each class must be determined, as inaccurate probabilities would effectively create overconfident or underconfident models that would not work well within the conformal prediction paradigm. Conformal prediction has been suggested to be effective on many real-world examples, including for detection of hardware trojans[9]and drug discovery[10].

This study analyzed the effect of commonly used algorithms for CNN calibration, such as Platt scaling and label smoothing, on conformal prediction as opposed to on commonly used single-output methods. While the algorithms used in this paper have been proven to be effective in their purposes of calibration or prediction[6,11] , there is not much literature on the effects of combining different techniques, as much of the literature focuses on improvements to conformal algorithms directly. Therefore, this paper aimed to help bridge the gap by analyzing the effectiveness of different traditional calibration techniques in the conformal prediction framework.

Literature Review

Conformal prediction has always been closely tied to calibration, as its main goal is to provide statistical guarantees, which is not possible without accurate probabilities, as demonstrated in the original conformal prediction paper by Gammerman et al.[7] Modern day conformal prediction includes a myriad of techniques meant to build on this framework. For example, Bastani et al. give a conformal prediction algorithm that achieves coverage guarantees even against adversarial data and distribution shift[12]. Similar ideas are given by Cauchois et al., with the goal of calibrating for difficult data points, for which coverage might not occur[13]. These ideas can be further extended for image recognition in the case of obfuscated imagery. Similarly, Vovk et al. created an algorithm that takes an existing predictive system and calibrates it, while also remaining computationally

efficient, unlike full conformal prediction, which refits the conformal predictor many times [14]. Other forms of calibration have taken the form of directly changing the way a model is trained before conformal prediction techniques are applied in order to make a model more receptive to the paradigm, such as work done by Stutz et al. [15] In summary, many special techniques have been derived to improve the conformal prediction paradigm itself. But, using classical methods to calibrate a model intended for conformal prediction has received limited attention, which is the intent of this paper.

Methods

Model Architecture and Metaparameters

The model used for the experiments is a ResNet-34 [11], built using the PyTorch library [16]. The loss function used in all parts of the model was binary cross-entropy loss with no modifications. The optimizer for the main training loop was Adam [17]. However, during Platt scaling [18](where applicable), the low-memory BFGS algorithm (LBFGS) was used for optimization as it is more efficient and suitable to the relatively simple nature of the required optimization. There were many metaparameters utilized in different trials of the experiment, including many for the RAPS conformal algorithm. These include $k_{reg} = 1$, $\alpha = 0.15$, and $\lambda = 0.2$. For the naive algorithm, a softmax with varying temperature values was used; the exact values are detailed in the results. In addition, the use of label smoothing necessitated a value $\varepsilon = 0.1$, for all trials where it was relevant.

Data Preparation

The dataset of choice was CIFAR10, comprised of a total of 60,000 images split evenly into 6,000 images of each of ten classes, depicting common objects (ex. planes, cars, horses) [19]. These images are 32 by 32 pixels, and are in full color. For the purposes of some of the tested calibration methods, a calibration set was required. To accomplish this, another split was created. First, the original training set of 50,000 images (already partitioned off when CIFAR10 was imported) was split into a 45,000 image training set and a 5,000 image validation set. The validation set was created using PyTorch's random

Algorithms

A metric of conformal accuracy was defined as the percentage of images in the testing set that the model outputs a set containing the correct answer for, which will be referred to later. One calibration method tested was label smoothing [6], which prevents models from becoming overconfident via applying the formula

$$y' = (1-\varepsilon)\cdot y + \frac{\varepsilon}{K}$$

where y is the array of outputs after a softmax is applied, K is the number of classes, and ε is a small constant chosen as a metaparameter [6]. In this way the softmax score of a class is shrunk if it is large and slightly increased if it is small. In other words, it forces all values to hold a significant nonzero value, introducing some noise into the outputs, so that the model does not become overconfident. This was especially relevant for the purpose of the paper as overconfident models may exaggerate the larger probabilities and shrink the smaller ones, resulting in an inaccurate representation of confidence for each class and therefore causing unwanted set size reduction by excluding probabilities that could have been significant. The value of $\varepsilon = 0.1$ in this study was chosen so that the highest possible value of a class probability would be smaller than 0.9, which helps to facilitate the conformal algorithm in producing more than one output before reaching the desired threshold of .85, unless the model is very confident on a class. Another method employed was Platt scaling/temperature scaling, which is a method that tries to optimize a factor T which all logits are then divided by before a softmax is applied [18]. This serves to calibrate the model and acquire a T that most accurately represents the actual confidence of the model in general. Mathematically, it attempts to minimize the negative log-likelihood of the calibration dataset. Two conformal prediction output algorithms were used, the most intuitive being the aforementioned naive algorithm, which directly takes values from Platt/temperature scaling and accumulates the corresponding classes of the largest probabilities until the desired coverage threshold is reached. There are two glaring flaws with this algorithm, the first being that Platt scaled values are not necessarily accurate and therefore are not truly representative of the coverage, and the second being that this algorithm tends to create massive sets for images it is less certain about and very small sets for images it is overconfident on [8]. To fix the issues above, the RAPS algorithm was used [8]. It first takes a calibration set of data given metaparameters k_{reg}, λ, α and calculates the following value for each item in the calibration set:

$$E(j) = \sum_i p_i + \lambda_1[i > k_{\text{reg}}]$$

where the summation is taken over all predictions up to and including the correct one (from highest probability to lowest), and pıis the outputted Platt scaled probability associated with the i"th" image8. In short, for each image, E(j) penalizes the outputted probabilities that are beyond the first k_{reg} most likely by adding λ to them. Then the sum of all probabilities greater than or equal to the probability of the correct class for said image is calculated. This is done for every image. Finally, a value $\hat{\tau}_{ccal}$ representing a 1-α quantile cutoff based on all E(j)s is returned. To actually predict the classes for a new image, a similar algorithm is applied. First, all values past the first k_{reg} are penalized with an addition of λ to its value. Then, these

Fig. 1 The model, in the case that only a softmax is used during training along with the naive algorithm, rarely outputs more than one value, even when said value is incorrect.

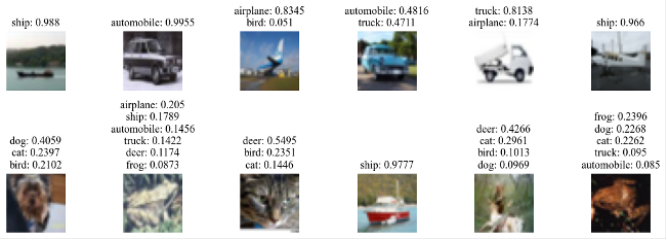

Fig. 2 Platt scaling on the naive model will often output one or two values, but hard images may cause it to output large amounts of values that make the set somewhat meaningless.

probabilities are summed from largest to smallest until $\hat{\tau}_{ccal}$is exceeded.

The metaparameter k_{reg} penalizes larger sets by adding an extra λ of value to a set for each extra element in the set after the first k_{reg} values, making it reach the threshold faster. It serves as a way of keeping set sizes small, as on difficult images, large sets with many miniscule probabilities on the tail end of the set could be outputted8. In this study, $k_{reg} = 1$ was used, as the total number of classes is relatively small, at 10 total, and therefore quickly penalizing additions is essential to set size minimization. Similarly, this study used $\lambda = .2$, so that in the very worst case scenario, the algorithm would only output 5 potential classes maximum.

Results

The data collected during testing is presented in Table 1. Trials were done five times for each category to ensure validity and consistency; relevant standard deviation statistics for accuracy over the five trials are given. The Colab Notebook with the relevant code can be found on GitHub . For specific information on the algorithms and architecture, consult the methods section.

Firstly, it is notable that naive models with a softmax have extremely low set sizes that imply overconfidence, confirming the idea that softmax probabilities aren't representative of true probabilities (see Figure 1). Also, the completely uncalibrated baseline, where outputs were only softmaxed for viewing, performed better in terms of the usual accuracy metric than the RAPS models, but did significantly worse in terms of conformal accuracy once label smoothing was applied. Indeed, this also implies that the RAPS models are more reasonably confident.

Platt scaling on naive models somewhat appears to solve this issue, with a higher average size, but due to a lack of regularization (which can be found in RAPS), difficult sets still output too many results to be meaningful (see Figure 2).

Using label smoothing to fix the overconfident images in the above case causes an exacerbation of the other issue, causing so many sets to be output on average for all images that the model ceases to provide meaningful information (see Figure 3).

[i] https://github.com/sav716/ResNet-Calibration-Colab-Notebook

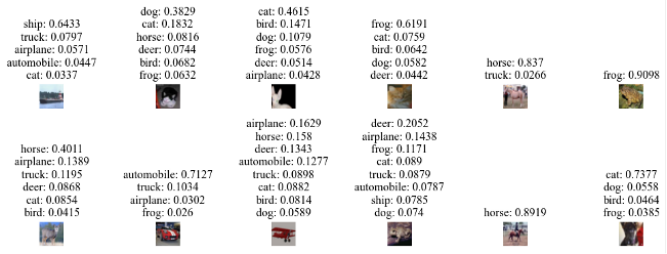

Fig. 3 With the application of label smoothing along with Platt scaling, the underconfident naive model outputs an egregious number of classes, no better than blind guesses.

RAPS seemed to significantly improve set size by creating neither massive sets nor sets with one value, best fitting the use of conformal prediction. Label smoothing also seemed to help increase conformal accuracy, though it had negligible effect on standard accuracy, if any at all. Utilizing label smoothing and Platt scaling, along with the RAPS algorithm, created the model which had the best results of all trials (see Figure 4).

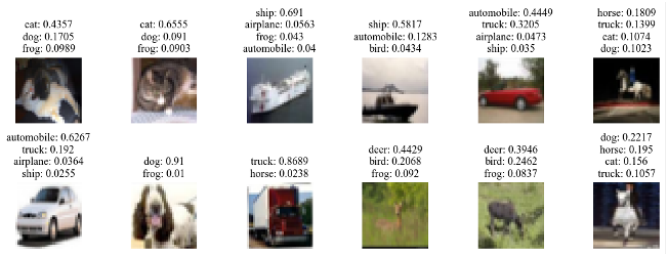

Fig. 4 Using RAPS with both Platt scaling and label smoothing allows the model to provide multiple results that best represent how sure it is of each image.

Discussion

In this section, the differences of the two conformal algorithms studied are discussed, particularly whether combining calibration and conformal prediction is effective. Special focus is put on specifically label smoothing and its differing impact on RAPS and the naive algorithm. Also analyzed are the effects of tem-

Table 1 Combinations of calibration and conformal output methods, with their respective average accuracies set sizes. The best performers for each accuracy metric have been bolded.

Description	Conformal Output Algorithm	Average Accuracy	Accuracy Standard Deviation	Average Conformal Accuracy	Conformal Accuracy Standard Deviation	Average Set Size
No calibration	Naive	**74.837%**	1.510%	75.096%	1.370%	1.0371
Platt Scaling	Naive	73.786%	4.028%	92.920%	1.080%	2.2197
Softmax in training	Naive	57.735%	2.264%	61.176%	1.846%	1.1324
Label smoothing	Naive	58.041%	2.040%	62.144%	2.076%	1.1550
Label smoothing with Platt scaling	Naive	55.521%	1.013%	94.708%	1.584%	5.5511
Softmax in training	RAPS	58.455%	2.160%	78.108%	1.145%	2.2293
Label Smoothinga	RAPS	56.747%	1.243%	**88.240%**[b]	1.508%	3.6141
Softmax Scalingc with Label Smoothing	RAPS	57.440%	1.508%	88.712%	1.331%	4.4865

[a] Note that the RAPS Algorithm inherently comes with Platt scaling.

[b] Although the conformal accuracy of the label smoothing and Platt scaling model was higher, it output too many values on average to be considered well-performing.

[c] In this version the Platt scaling step in RAPS was replaced by a softmax. This was done to demonstrate that Platt scaling is a better choice (see the discussion).

perature/Platt scaling and whether this effect of their inaccurate probability outputs is apparent here.

General Comparisons between RAPS and the Naive Algorithm

Without any form of calibration, it seems that the ResNet easily becomes overconfident in its predictions as can been seen by extremely high values after applying a softmax, even if the class is incorrect (Figure 1). With the naive algorithm, set sizes of course are small as the cutoff of .85 is very hard to not reach when most of the first classes in each image have corresponding softmax values extremely close to 1. Platt scaling does help to negate this effect and there is an increase in set size as seen in Table 1, but the naive algorithm is still holding down accuracy as using RAPS with the ResNet immediately increases conformal accuracy by a wide margin.

Effects of Label Smoothing with RAPS

First, it is imperative thatε is not overly large, as this could cause even images that the model is confidently correct about to output many results, due to the smoothing causing the large correct probability to simply not be big enough to meet the coverage requirement, which causes all the other probabilities (which are potentially all small and approximately equal) to accumulate until the threshold is met. This effect can be seen in the label smoothing and Platt scaling case of a naive model: the high conformal accuracy is meaningless because the sets contain half the classes on average (Figure 3). As both Platt scaling and label smoothing serve to decrease overconfidence, the effect goes too far in this case; the model is extremely underconfident in its predictions. RAPS moderates this effect by prevent too many tiny probabilities from being output, decreasing set size: comparing Figure 3 and 4 shows that the model is not necessarily much more confident in Figure 4, but outputs less classes. It is significant that although conformal accuracy barely increased, set size went down drastically. Thus, it can be concluded that these two strategies complement each other very well because Platt scaling can help the model be more moderate in its confidence. As label smoothing forces probabilities into a narrower and more moderate range during training, RAPS is able to be more effective since probabilities on the low end after label smoothing remain low enough that if they are included in the set, they definitely have met the condition for λ to be added, which prevents too many of them from being added, an issue seen in the naive algorithm (see the below section).

Platt Scaling vs Temperature Scaling

As mentioned before, temperature scaling and Platt scaling are ways to compress values between 0 and 1, but this does not make them accurate probabilities. Though the set size may seem somewhat acceptable in the naive version, it contains many outlier sets for hard images that are not useful (Figure 2), which can be solved by the regularization RAPS does. Therefore, while Platt scaling certainly eliminates model overconfidence in the naive algorithm, it doesn't serve conformal prediction's purpose, which requires more accurate probabilities to provide an accurate set that is neither full of negligible probabilities or dominated by a single one. This idea is further corroborated by the increase in set size with no corresponding increase in conformal accuracy if the Platt scaling step of the RAPS algorithm is replaced by a normal softmax (Table 1), as Platt scaling is a calibrated softmax [18].

Limitations of This Study

A significant limitation of this study was its relatively simple architecture and simple benchmark dataset of small images and few categories. It would be worth investigating this issue with more complex convolutional neural networks (e.g., a ResNet with more layers), and a benchmark dataset that better represents the real world, such as ImageNet.

Future Directions

The scope of this study was, as mentioned above, limited to one architecture and a simple benchmark dataset. It would be worth corroborating the results found in this paper with more complex models and datasets. In addition, the number of algorithms tested for both conformal prediction and model calibration was relatively limited in scope. Thus, a direction worth investigating would be how other algorithms for conformal prediction and calibration function in when applied together, as it may be possible to draw more generalized conclusions from algorithms.

Conclusion

This study looked for benefits from applying calibration techniques to conformal output algorithms. A combination of techniques creates optimal results in terms of conformal accuracy. Indeed, optimal results were achieved with RAPS (which includes Platt scaling) and the usage of label smoothing in this study. Using set size as a metric, it also seems that in terms of benefits reaped from utilizing conformal prediction, there is a good middle ground between large sets that are meaningless and overconfident small sets that can be achieved with these algorithms. This study also found that RAPS is almost essential to better set sizes that on average contain more than one element, as it could cull set sizes for images that naive algorithms might output too many classes for. Also notable was the tendency of the naive algorithm to start having values lose their significance as more regularization techniques are applied (due to large sets), which RAPS was able to avoid. In summary, conformal prediction algorithms should use RAPS with label smoothing for optimal results.

Acknowledgements

The author would like to thank their mentor, Mariel Werner, for their invaluable advice and insight in helping to complete this paper.

References

1 A. Konstantinov and L. Utkin, *Knowledge-Based Systems*, 2021, **222**, 106993.

2 O. Bohdal, Y. Yang and T. Hospedales, *Transactions on Machine Learning Research*, 2023, **2**, 1–21.

3 R. Roelofs, N. Cain, J. Shlens and M. C. Mozer, International Conference on Artificial Intelligence and Statistics, 2022, pp. 4036–4054.

4 S. Ioffe, Proceedings of Machine Learning Research, 2015, pp. 448–456.

5 K. R. M. Fernando and C. P. Tsokos, *IEEE Transactions on Neural Networks and Learning Systems*, 2022, **33**, 2940–2951.

6 M. Minderer, J. Djolonga, R. Romijnders, F. Hubis, X. Zhai, N. Houlsby, D. Tran and M. Lucic, Advances in Neural Information Processing Systems, 2021, pp. 15682–15694.

7 A. Gammerman, V. Vovk and V. Vapnik, 14th Conference on Uncertainty in Artificial Intelligence, 1998, pp. 148–155.

8 A. Angelopoulos, S. Bates, J. Malik and M. Jordan, International Conference on Learning Representations, 2021, pp. 1–17.

9 R. Vishwakarma and A. Rezaei, 2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD), 2023, pp. 1–9.

10 J. Alvarsson, S. A. McShane, U. Norinder and O. Spjuth, *Journal of Pharmaceutical Sciences*, 2021, **110**, 42–49.

11 K. He, X. Zhang, S. Ren and J. Sun, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, **29**, 770–778.

12 O. Bastani, V. Gupta, C. Jung, G. Noarov, R. Ramalingam and A. Roth, *Advances in Neural Information Processing Systems*, 2022, **35**, 29362–29373.

13 M. Cauchois, S. Gupta and J. C. Duchi, *Journal of Machine Learning Research*, 2021, **22**, 1–42.

14 V. Vovk, I. Petej, P. Toccaceli, A. Gammerman, E. Ahlberg and L. Carlsson, *Proceedings of Machine Learning Research*, 2020, **128**, 84–99.

15 D. Stutz, K. Dvijotham, A. T. Cemgil and A. Doucet, *International Conference on Learning Representations*, 2022, **10**, 1–27.

16 A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga and A. Desmaison, *Advances in Neural Information Processing Systems*, 2019, **32**, 8026–8037.

17 D. P. Kingma and J. Ba, *International Conference on Learning Representations*, 2015, **3**, 1–15.

18 J. Platt, *Advances in Large Margin Classifiers*, 2000, vol. 10, pp. 61–74.

19 A. Krizhevsky and G. Hinton, *Learning multiple layers of features from tiny images*, 2009.