Quantifying Differences Between Normally and Conformally-Trained Models for Classification Tasks

Cameron Husselbury*, Fergus Imrie**
*University of Manchester, **University of Oxford

1. Introduction

- Typically machine learning models are used to make point predictions.
- Conformal prediction is a framework for creating statistically valid prediction regions based on the assumption of exchangeability of data.
- It is applied post hoc to any underlying black box point predictor.
- Conformal training trains the model to be more optimised for conformal prediction. This allows for smaller confidence sets.
- This project seeks to quantify the differences in what conformally-trained models use to make predictions in comparison to normally-trained models.

2. Methods

Conformal Prediction

Conformal prediction takes a trained model and produces a prediction set with a user-specified probability of the true point or class being contained within it. This is done by by taking unseen calibration data and using this to construct a set of possible labels $\mathcal{C}(X_{test})$ such that

$$\mathbb{P}(Y_{test} \in \mathcal{C}(X_{test})) \geq 1 - \alpha,$$

where (X_{test}, Y_{test}) is a test point from the same distribution, and α is a user-specified error rate. We can assume the data is exchangeable because any new data point is equally likely to be ranked in any position relative to the existing data.

Since Y_{test} is exchangeable with the values $Y_1, \ldots, Y_n$ in the calibration set, we can say that [1]

$$\mathbb{P}(Y_{test} \text{ is among the } \lceil (1-\alpha)(n+1) \rceil \text{ smallest of } Y_1, \ldots, Y_n) \geq 1 - \alpha.$$

Now define

$$\hat{q} = \lceil (1-\alpha)(n+1) \rceil \text{ smallest of } Y_1, \ldots, Y_n,$$

so that

$$\mathbb{P}(Y_{test} \leq \hat{q}) \geq 1 - \alpha,$$

from which we can obtain the prediction interval $\mathcal{C} = (-\infty, \hat{q}]$.

Conformal Training

Conformal training was introduced in a paper by Stutz et al [2], and seeks to improve conformal prediction by training the model with this as the end goal. It achieves this by splitting each mini-batch during the training of the model into two parts, B_{cal} and B_{pred}. The first is for calibration of the cutoff threshold τ, and the second is for prediction and loss computation, where the threshold τ is used to generate confidence sets. A loss function aimed at minimising the size of these confidence sets is then used to update the model's weights. The below figure is taken from [2] and serves as a useful visualisation of the process.

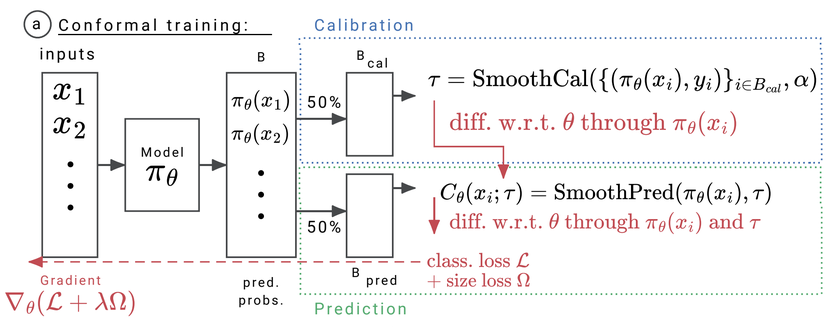


Figure 1:
Illustration of conformal training. Conformal training calibrates on the first half of each mini-batch and predicts confidence sets on the other half.

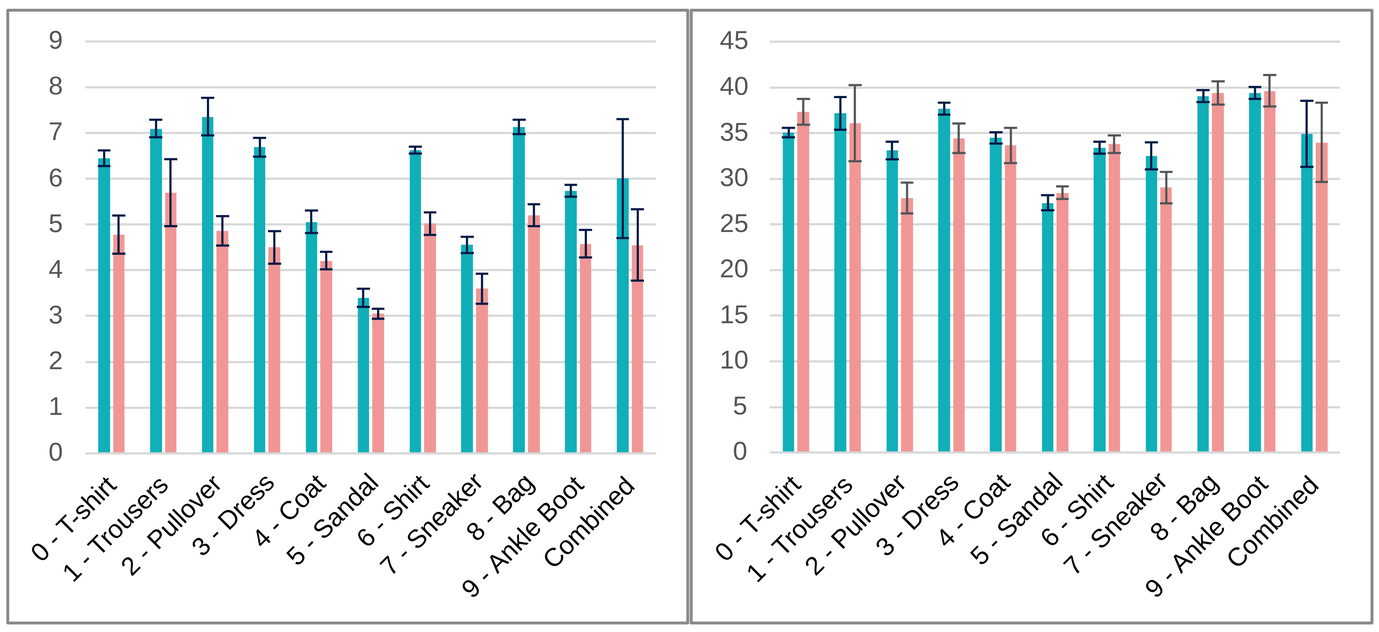
3. Experimental Setup

- **Datasets:** Our models were trained on MNIST, Fashion-MNIST, and CIFAR-10. We show results only for the Fashion-MNIST dataset.
- **Model and hyperparameters:** We used a 2-layer, 128 unit Multi-layer Perceptron model, trained over 150 epochs on 55000 images, with a learning rate of 0.01, and a batch size of 100. One was trained normally, and the other conformally, but besides this they both used the same settings. These settings were chosen in order to replicate those used in Stutz et al.'s. For the full settings see [2].
- **Metrics and analysis:** Our primary analysis technique was occlusion sensitivity, which is measured by placing a black 2 x 2 pixel square over part of the image (occluding it) and translating it pixel by pixel until all pixels of the image have been covered at some point. [continued top right]

Upon each step in the process, the logits (and probabilities) for the modified image are computed using the model. The sensitivity is then computed by subtracting the occluded confidence from the confidence of the original image (confidence being of the model in its prediction). Each pixel's impact is measured as the average of the image's changes in sensitivity whilst that individual pixel is occluded. We also performed an integrated gradients analysis (omitted from this poster). We class a significant region as any region consisting of sensitivities greater than 1.96σ above the mean sensitivity.

4. Results

The charts below show the per-class and combined results of the occlusion sensitivity test on both the normally-trained and conformally-trained models. The results show that the average number of distinct regions per image across all classes is 32.1% greater for the normally-trained model versus the conformally-trained model (6.012 v. 4.55).



Charts:
Occlusion sensitivity test results for the normally and conformally-trained models on the Fashion-MNIST dataset.
Left: Bar chart comparing the number of distinct regions between the normally and conformally-trained models.
Right: Bar chart comparing the mean areas between the normally and conformally-trained models when each of the distinct regions are combined.

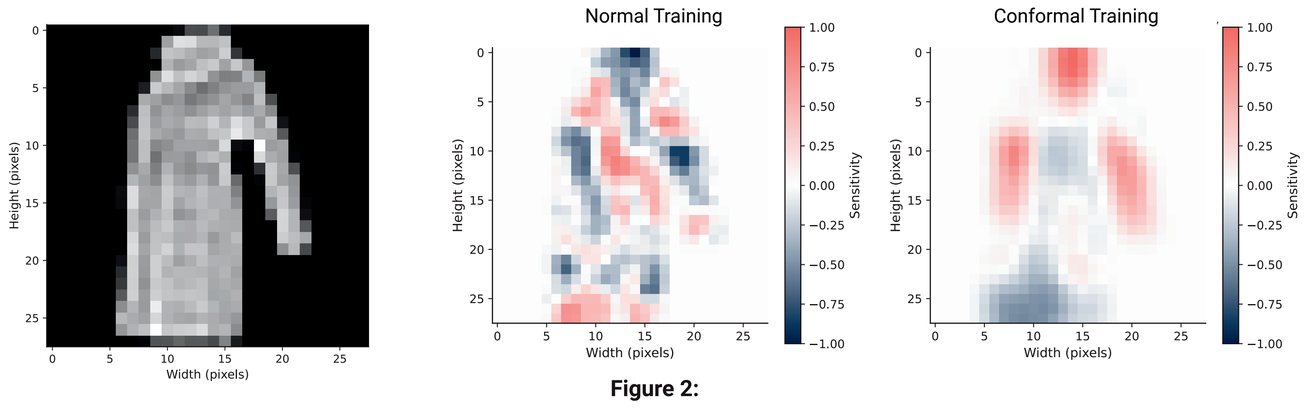


Figure 2:
Left: Original image from the Fashion-MNIST dataset. **Middle:** Occlusion sensitivity map for the normally-trained model on the same image. **Right:** Occlusion sensitivity map for the conformally-trained model on the same image.

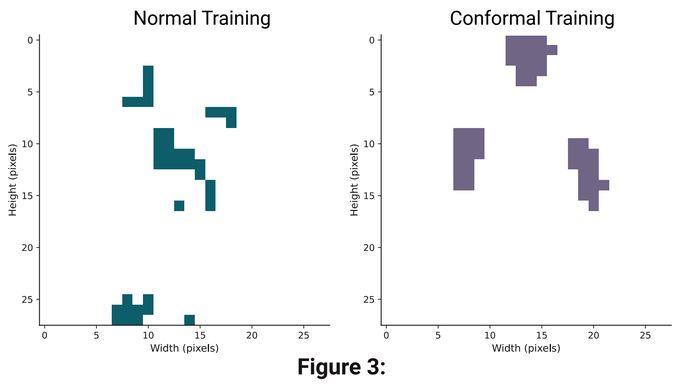


Figure 3:
Left: Significant regions from the occlusion sensitivity map for the normally-trained model. **Right:** Significant regions from the occlusion sensitivity map for the conformally-trained model.

Figure **2** contains visual representations of the relative occlusion sensitivities of predictions made by the normally and conformally-trained models, and Figure **3** shows the significant regions of these occlusion maps.

- Across all classes the total areas of the occlusion sensitivity maps were similar, indicating that both models are taking the same amount of the image into account, however the conformally-trained model appears to consider fewer, larger individual areas than the normally-trained model.
- Figure **2** shows that the conformally-trained model appears to focus more on distinct features of the clothing, for example a sleeve.
- The areas in blue on the occlusion sensitivity map are areas which negatively impact the model's ability to distinguish the object in the image, From Figure **2** we can see that in this case the body of the clothing does not contribute positively to making the prediction - This could be due to multiple different types of clothing having this same feature.
- Similar results were seen on MNIST and CIFAR-10 however we chose to display the Fashion-MNIST results for clarity.

References

[1] Tibshirani, R - Conformal Prediction, Advanced Topics in Statistical Learning, Spring 2023
[2] Stutz, D et al. - *Learning Optimal Conformal Classifiers*, arXiv:2110.09192