Research Proposal Innovation Case - Classifier Calibration

Edward Milsom

1 Executive Summary

Machine Learning (ML) has already found a vast array of applications, including in safety-critical areas such as medical diagnosis[1,2] and autonomous vehicles / driving assistance systems,[3,4,5] and is only expected to grow further, particularly in the healthcare sector.[6] In these domains, there is a serious risk of injury or death if a system makes a mistake. Hence it is crucial that algorithms utilised in these areas are reliable and interpretable. Ensuring classifier uncertainties are calibrated is an important step towards this goal.

A classifier is an algorithm that takes in a data point and assigns it a label. An important example is identifying objects in an image. Many classifiers also give an uncertainty, which should be the probability of the predicted label being correct. However, many classifiers give uncertainties that are **uncalibrated**[7,8] , which essentially means that the uncertainty does not represent the true probability. These uncertainties are therefore less meaningful, less interpretable, and as a result cannot be seen as reliable when used in decision making processes. We can use **calibrators** to obtain more accurate probabilities in many scenarios, but their performance is often unsatisfactory on real world datasets.

We propose a project which will analyse current state-of-the-art calibrators, evaluating their performance on a number of real world datasets, and offer possible improvements to these techniques. More importantly, we will propose entirely new methods of calibration, which after refinement can be compared against existing calibrators. If the project is successful, we will see a marked improvement in classifier calibration using our new techniques, which can easily be implemented in production ML systems to offer instant benefits. This will involve recruiting a small cohort of PhD students, who will each explore different directions as part of their research, as well as a post-doctoral researcher to perform more focused analysis in the most promising directions. We will collaborate with Facebook's AI Research Lab to bring our new calibration techniques to the widely used Machine Learning library PyTorch, therefore fast-tracking their uptake to bring immediate impacts to real world problems.

2 Evaluation of Potential Opportunity

In 2018, Elaine Herzberg became the first person in the world to be killed in a collision with a self-driving car.[9] The incident put a damper on the charge towards bringing autonomous vehicles into widespread use in the near future, but it also brought attention to the risks of using "black-box" Machine Learning models in high-stakes situations. It has been argued that in such critical scenarios, we must move towards using more interpretable models.[10] Machine Learning, and in particular Deep Learning, has become integral to many sectors over the last decade,[11] and is expected to continue to grow in the future.[6] One class of algorithms in particular, called **classifiers**, have improved enormously due to the rise of deep learning. Classifiers have already started to be utilised in safety-critical scenarios like medical diagnosis systems[1,2] and autonomous driving systems.[3,4,5] The potential for injury, death, or damage to property is significant should these systems perform suboptimally.

A key problem with some of the cutting-edge models, particularly within deep learning, is that they provide no uncertainty for their estimates. Without uncertainties, there is no way of knowing how confident a model is in its predictions, and thus it can be difficult to trust them for important decisions. There has been a push to ensure models are more probabilistic in the future.[12]

Some models, however, do already provide uncertainties. This raises a key question: how do we know these

scores (we often call the outputted probability for a class the score) represent the true probability of the prediction being correct? One notion of correctness is to require that outputted uncertainties are **calibrated**. This means that, for any particular score, the true proportion of instances with this score that actually belong to the predicted class matches the score. For example, if there are a set of images that have all been given the score 70% by a classifier that identifies cats, then exactly 70% of those images should actually contain a cat. Though calibration is not a perfect measure of accurate probabilities (you can construct examples where a classifier is technically calibrated, even though it just outputs the same score for any input), in normal circumstances it is a very desirable property.

Many models do not output calibrated probabilities.[7] In these cases, we can use post-hoc **calibrators**, which try to learn a mapping between the scores and the true probabilities. A few different calibrators have been proposed, including Platt scaling, isotonic calibration, and more recently, Beta and Dirichlet calibration.[7] All have their own shortcomings, and we are still a long way from having a one-size-fits-all solution to the problem.

To get an idea of what calibration techniques are currently in wide use, we can look at some popular machine learning libraries. Scikit-learn[13] is a Python library that mainly focuses on traditional machine learning techniques (i.e. not deep neural networks). We see from the documentation[14] that Scikit-learn supports logistic calibration (Platt scaling) and isotonic calibration. These are both calibration techniques for binary classifiers (classifiers that can only identify one class, e.g. "Cat" vs. "Not Cat"). Binary calibrators do not generalise well to the multiclass case[7] (classifiers that can identify several classes, e.g. "Cat", "Dog", "Horse"). Furthermore, these two techniques are now quite old, and Platt scaling in particular has recently been generalised to less restrictive assumptions using a technique called Beta calibration.[15]

PyTorch[16] is a popular deep learning library. Most of the time, deep neural networks are utilised as multiclass classifiers. Though the scores outputted by neural networks are unnormalised, a technique called softmax is commonly used to produce pseudo-probabilities. However, these pseudo-probabilities are usually uncalibrated,[17] and PyTorch provides no native implementation for multiclass calibration. Techniques specifically for neural networks, in particular temperature scaling,[17] have grown in popularity since they have been proposed, though in the case of PyTorch users must still implement them manually.

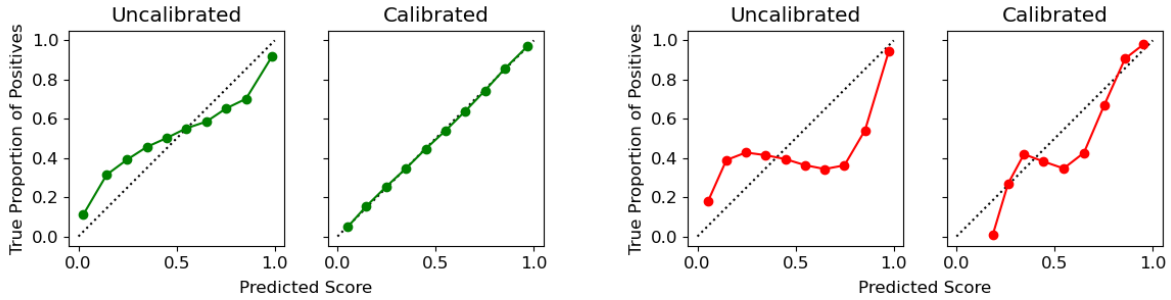
As previously noted, ML is only expected to become more integral to our lives in the future, so while calibration may not be glamorous, it is vital that we are using models that give us accurate probabilities when making critical decisions. The potential benefits to society are enormous, but the work needs to start now to ensure wide adoption.

3 Value Proposition

Dirichlet calibration was proposed in a recent paper.[7] It is a multiclass generalisation of Beta calibration, and can also be viewed as a less restricted version of temperature scaling. In the paper, the authors demonstrate that Dirichlet calibration produces state-of-the-art performance using various evaluation metrics. However, they themselves acknowledge that Dirichlet calibration is just a starting point, and there is much more research to be done in producing new calibrators that are based on different probability distributions from the exponential family (Dirichlet calibration itself is based on the Dirichlet distribution, a member of the exponential family), or even based on mixtures of distributions.

Figure 1 shows reliability diagrams for two datasets: one where Dirichlet calibration performs well, and one where it performs poorly. A reliability diagram simply shows how predicted scores relate to the true proportion of positives. A diagonal straight line going upwards from corner-to-corner is ideal, since then our predicted scores match the true probabilities. We can see in Figure 1a that on this dataset, the calibration

curve becomes a straight line, meaning this classifier is now calibrated, and its uncertainties are therefore reliable. However in Figure 1b, we see that even after applying Dirichlet calibration on this other dataset, the classifier is still very poorly calibrated. This means we cannot rely on its uncertainties, and therefore we should not be using it in safety-critical situations. Clearly Dirichlet calibration is not a perfect solution in all situations.



(a) Good calibration is achieved here

(b) Calibration is poor here

Figure 1: Reliability diagrams showing Dirichlet calibration performing differently on certain datasets. The straight diagonal is the optimal calibration curve.

This is where our proposed research project comes in. We aim to investigate new directions such as those suggested by the authors of the Dirichlet calibration paper, who we will be collaborating with. Specifically, we will identify datasets for which Dirichlet calibration performs poorly (it is impossible to provide a precise definition of "good" or "poor" performance, but there are an array of evaluation metrics that can be used to compare calibrators, such as Brier score, Confidence Expected Calibration Error (ECE), Classwise ECE, Maximum Calibration Error etc.). Once this has been established, we can begin investigating alternative calibration maps. Inspired by Dirichlet calibration, we already plan to investigate calibration maps derived from both exponential family probability distributions such as the Gamma and Wishart distribution, as well as mixture distributions.

We will also investigate a recently proposed probability distribution known as the Metalog distribution.[18] If its use in calibration is deemed practical, the Metalog distribution could be very promising due to its high flexibility, allowing it to model a wider range of score distributions, which will result in better calibration maps, and therefore better calibrated probabilities. Furthermore, the Metalog distribution can be fitted using Linear Least Squares, which is an extremely efficient algorithm compared to descent-based methods that are used for many other distributions, meaning that fewer computational resources are needed. This has the added benefit of reducing our impact on the environment by using less energy.

In discussions with Facebook's AI Research Lab (FAIR), who develop PyTorch, they expressed their desire to extend the capability of their library by including calibration techniques, though they felt the current options were limited. In response to this, they would be willing to partially fund our research project so that they have better techniques to implement. They do not feel it is necessary for them to acquire IP rights, as long as the resulting work is in the public domain, to prevent any competitors from developing similar proprietary systems. Due to calibration being quite a niche area of study, it is much easier and cheaper for them to fund our project, which already involves multiple experts, rather than go through the expensive process of hiring new researchers and assembling their own project . They also hope that the partnership will allow them to recruit at least one of our PhD students, who by the end of the project will also be experts. PyTorch is widely used within Deep Learning and so native calibration implementations will benefit all those working in the field, and will encourage more people to consider the concept of calibration when they are developing systems for real-world use.

4 Impact Plan

The project will be undertaken in collaboration with the authors of the Dirichlet calibration paper.[7] We are already in discussions with them and we are all eager to make the project a reality. In addition, we aim to recruit 3-4 new PhD students to carry out the main body of research, and 1 postdoctoral researcher. Facebook have already committed £500,000 to the project, and so we will require a grant of £500,000 to cover the salaries of staff members, as well as the stipends for the PhD students. The grant will also cover equipment costs, which due to the nature of the project will be low and mainly limited to personal computers for the researchers, in addition to standard maintenance costs the university must cover. The duration of the project will be 4 years.

By the end of the project, we will have completed 3 phases / goals. First, the strengths and weaknesses of existing calibration techniques will have been analysed in depth, with our findings published via scientific papers at well-established conferences, such as NeurIPS and ICML. Second, a variety of novel calibration techniques will have been derived, implemented, and tested in an experimental setting. The most promising of these approaches will proceed to the third phase, where they will be developed into a publicly available open source library called "calibPy", as a demonstration of the techniques. After that, Facebook intends to incorporate these techniques into PyTorch. The details of these three phases are laid out in the following paragraphs.

In the first phase, dubbed the "Assessment" phase, researchers will consider many existing calibrators, including Platt scaling, Isotonic calibration, Beta calibration, and Empirical Binning for binary classifiers, and Temperature scaling, Vector scaling, Matrix scaling, and Dirichlet calibration for multiclass classifiers. They will be evaluated on many of the most common datasets, e.g. MNIST, CIFAR-10, Wine Quality, and Pima Indians Diabetes, using a wide range of evaluation metrics and proper scoring rules, including Binary ECE, Binary MCE, Confidence ECE, Classwise ECE, Brier Score, and Log-Loss. The findings will be disseminated via a scientific paper published at a well-established conference.

In the second phase, dubbed the "Exploration" phase, we will explore the avenues discussed in the previous section, namely the use of different exponential family probability distributions, the Metalog distribution, as well as an investigation into how mixture Distributions can be applied to the calibration problem. These investigations will take place in an informal manner, with the research group meeting on a regular basis to share findings and identify which directions have the most potential.

In the final phase, dubbed the "Refinement" phase, we will take the best ideas from the Exploration phase and test them more thoroughly, using a similar methodology to the first phase. Those calibration methods that are deemed state-of-the-art will be developed into an open source library "calibPy", and scientific papers will be written detailing their function and performance. Both the papers and library will enable other researchers to evaluate our novel approaches, with the hope that they will succeed existing calibration techniques and therefore bring value to the field of Machine Learning.

Following this, Facebook plan to implement these algorithms into PyTorch, which will provide fast, easy-to-use, robust tools to anyone already using their library, and will dramatically increase the visibility and uptake of calibration techniques. PyTorch is amongst the most popular of Machine Learning libraries[19] at the time of writing, meaning the output of this research project will have large and wide ranging impacts.

As a result of this project being carried out, we will finally have models which provide uncertainties that we can really trust, and thus the current wild-west approach to machine learning will be pushed further towards a safer relationship with technologies in the real world.

References

[1] K. Suzuki, "Machine learning in computer-aided diagnosis of the thorax and colon in ct: A survey," *IEICE transactions on information and systems*, vol. E96-D, pp. 772–783, Apr 2013. 24174708[pmid].

[2] K. Murphy, B. van Ginneken, A. Schilham, B. de Hoop, H. Gietema, and M. Prokop, "A large-scale evaluation of automatic pulmonary nodule detection in chest ct using local image features and k-nearest-neighbour classification," *Medical Image Analysis*, vol. 13, no. 5, pp. 757–770, 2009. Includes Special Section on the 12th International Conference on Medical Imaging and Computer Assisted Intervention.

[3] P. Koopman and M. Wagner, "Challenges in autonomous vehicle testing and validation," *SAE International Journal of Transportation Safety*, vol. 4, pp. 15–24, 04 2016.

[4] B. Spanfelner, D. Richter, S. Ebel, U. Wilhelm, and C. Patz, "Challenges in applying the ISO 26262 for driver assistance systems," 2012.

[5] M. Gharib, P. Lollini, M. Botta, E. Amparore, S. Donatelli, and A. Bondavalli, "On the safety of automotive systems incorporating machine learning based components: A position paper," 06 2018.

[6] E. Sparks, "Four predictions for artificial intelligence and machine learning in 2021." `https://www.forbes.com/sites/evansparks/2021/02/10/four-predictions-for-artificial-intelligence-and-machine-learning-in-2021/`, 2021. Accessed: 2021-03-02.

[7] M. Kull, M. Perelló-Nieto, M. Kängsepp, T. de Menezes e Silva Filho, H. Song, and P. A. Flach, "Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with dirichlet calibration," *CoRR*, vol. abs/1910.12656, 2019.

[8] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," pp. 625–632, 01 2005.

[9] "Uber's self-driving operator charged over fatal crash." `https://www.bbc.co.uk/news/technology-54175359`, 2020. Accessed: 2021-03-02.

[10] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, pp. 206–215, May 2019.

[11] R. Metz, "How AI came to rule our lives over the last decade." `https://edition.cnn.com/2019/12/21/tech/artificial-intelligence-decade/`, 2019. Accessed: 2021-03-02.

[12] Z. Ghahramani, "Probabilistic machine learning and artificial intelligence," *Nature*, vol. 521, pp. 452–459, May 2015. On Probabilistic models.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[14] "Scikit-learn User Guide: 1.16: Probability Calibration." `https://scikit-learn.org/stable/modules/calibration.html`. Accessed: 2021-03-03.

[15] M. Kull, T. S. Filho, and P. Flach, "Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* (A. Singh and J. Zhu, eds.), vol. 54 of *Proceedings of Machine Learning Research*, (Fort Lauderdale, FL, USA), pp. 623–631, PMLR, 20–22 Apr 2017.

[16] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019.

[17] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," *CoRR*, vol. abs/1706.04599, 2017.

[18] T. W. Keelin, "The metalog distributions," *Decision Analysis*, vol. 13, no. 4, pp. 243–277, 2016.

[19] "State of Machine Learning and Data Science 2020." `https://www.kaggle.com/kaggle-survey-2020`.