Thesis Proposal
Self-calibrating Convolutional Neural Networks for Image Classification

Duy Khoi Tran
Matrikel Nr.: 9030205

Master of Science in Autonomous Systems

University of Applied Sciences Bonn-Rhein-Sieg
Advisor: Prof. Dr. Paul G. Plöger
Second Advisor: ???
Third Advisor: ???

January 7, 2019

Abstract

While training image classifiers, besides accuracy, the confidence of the network in their choices or predictions is also important. However, the representation of output uncertainties in modern convolutional neural networks, commonly Softmax, does not reflect actual reality and tends to be over-confident. Thus, they it needs calibrationg. The aim of this project is to explore and establish the sufficient solutions for producing self-calibrating convolutional image classifiers, which calibrates its confidence in training phase and requires as little post-processing as possible.

1 Introduction

1.1 Background

Modern Neural Networks are among the state-of-the-art solutions for many real-world complicated tasks. They have achieved outstanding performance in computer vision[NMK+14], natural language processing [LK17] and many other fields.

Apart from the overall prediction accuracy of a network, the confidence of the predictions should be taken into consideration. This information can help determine if the predictions should be trusted. This is of great relevance especially for high-stake applications such as medical diagnosis or self-driving car. A popular mechanism to quantify such information in classification task is the utilization of Softmax activation at the ouput of a network. This Softmax value of each class aims to convey directly the probabilities of the choices being correct. For example, if we have ten samples classified as a class with softmax value of 0.9, exactly nine out of ten samples should be correctly classified. If the predictions of a neural network are able to match its uncertainties to realitiy and estimate sufficiently the probability distribution of the training data, they are deemed well-calibrated [GPSW17, Kae18]. In an early research of binary classification, deep neural networks were thought to give well-calibrated results [NMC05]. However, recent works from Guo et al. shows that, with Softmax alone, modern deep neural networks, as highly accurate as they may be, are not well-calibrated [GPSW17]. This is attributed to popular deep learning practices such as depth increase, batch normalization and weight decay. In light of this fact, more researchers are starting to look for calibration methods to remedy this problem.

1.2 Calibration methods

Calibration methods are those that attempt to derive the accurate estimation of the class distribution from a network in arbitrary manner. However, they generally fall in to two categorizations: post-processing calibration and training phase calibration.

Post-processing calibration methods generally produce a mapping from the uncalibrated output of a pretrained network to calibrated probabilities. Methods in this type impose new learning tasks for the parameters of the mapping and, thus, require a dedicated calibration dataset that reflects the or the calibration can suffer from overfitting. Popular post-processing methods involve either putting the uncalibrated output into bins of calibrated probabilities [ZE01, ZE02, PNFCH15] or a variant of Platt scaling, Temperature Scaling [Pla99, GPSW17, KFE18] that learns a regression model based on various types of loss [MGJG18a].Post-processing caibration methods are relatively well-studied and intensively evaluated [Kae18]. Temperature scaling has been established as the de facto method and acts as a baseline for both types of methods.

Training phase calibration is not as mature but is garnering a lot of research attention due to the elimination of output adjustment after training. This method type dominated by methods with Bayesian formalism [LPB17], in which prior distribution of weights are involved and the posterior distribution can be approximated using various approaches such as Laplace approximation [MBC+99], Markov chain Monte Carlo [Nea96] and variational Bayesian methods [BCKW15, Gra11, LW16]. Other diverse popular directions involve ensemble of models [GG16, LPB17], modification of losses or activation [MGJG18b, MSK18], similarity metrics [PM18] or self-learning post-processing parameters [NZV18].

Although the research volume for training phase calibration is quite substantial, there are some problems. The research works draw vague connections to each other. Their evaluation often use either different dataset and/or metrics depending aiming for potentially different aspects (out-of-distribution detection, robustness against adversarial attacks, etc.) Thus, a sufficient comparative evaluation can be challenging to produce from the literature.

1.3 Goal of the thesis

The aim of this thesis project is to produce the aforementioned comparative evaluation of methods for self-calibrating convolutional neural networks for Image Classification. This work is inspired by the Master's thesis of Markus Kängsepp [Kae18] that focuses solely on post-processing. The evaluation will be done accross multiple metrics, which include novel ones proposed alongside a few methods [JKGG18, MW17], and determine or develop the most suitable calibration method(s) in term of robustness, accuracy cost, computational cost, etc. to be applied in practical deep learning project. This project is based on the Master's Thesis of Markus Kängsepp [Kae18] focusing on post-processing convolutional neural networks in the context of image classification. Attempts at novelty will also be taken in the form of experimenting self-learning variants of existing post-processing methods similar to the works of [NZV18]. The product of this thesis should assist in the creation of Aptiv's deep learning framework for generating highly accurate and trustworthy intelligent systems.

2 Methodology

In order to fulfill the goals of the project, the prominent training-phase calibration methods will be chosen and implemented in Tensorflow Python. The implementations will be applied in training identical popular convolutional neural networks with common image datasets using the same configurations, unless otherwise specified. The performance and other relevant aspects of the trained networks will be evaluated accross multiple metrics, which include but not limited to Expected Calibration Error, Maximum Calibration Error, Negative Log Likelihood, Brier Score [Bri50] and Trust Score [JKGG18].

For methods with Bayesian formalism, Tensorflow Probability might be used. Tensorflow Probability [Ten] is a Tensorflow extension that allows distribution definition and sampling process in computational graph. From these capabilities, weights and other parameters can be represented as samples of a distributions and facilitate Bayesian modelling.

For experimentation of self-learning variants of post-processing methods, possible ideas are as below:

- **Histogram Binning** [ZE01]: This class of non-parametric methods divide uncalibrated predictions into mutually exclusive intervals with a confidence score. With a sufficiently big and reflective calibration set, the bounds of these intervals can be calculated. In this thesis, experiementation will be done to learn these bounds during training providing the minibatch is sufficiently big and the number of bins is relatively low.
- **Parametric methods**: Attempts to make the neural networks learn the calibrating parameters during training will be made similar to the works of [NZV18].

There are other classes of post-processing methods that might be further examined and applied directly in training.

3 Project Plan

3.1 Work packages

In conducting the project, the following work packages will be handled:

- WP1: Literature research.
- WP2: Methods implementation and model training.
- WP3: Adjusting, testing of models.
- WP4: Results extraction and Thesis report writing.

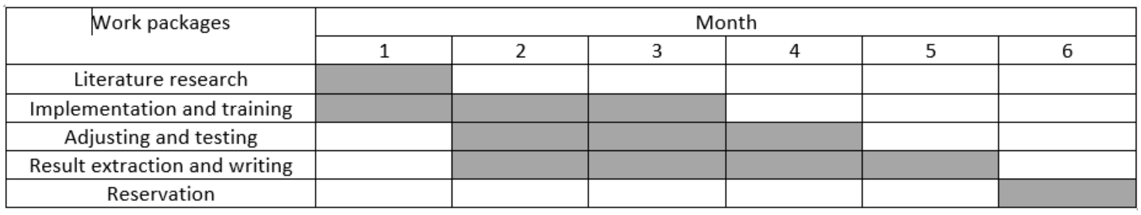
All of these workpackages will be somewhat carried out in parallel.

3.2 Schedule

An estimated time line for the execution of the project can be seen in Figure 1

3.3 Milestones

- Literature research leading to a survey of convey the landscape of self-learning convolutional neural network calibration
- Complete Python implementation of the calibration and the creation of neural network architecture with the methods applied.
- Sufficient empirical results that can reach to conclusive statements
- Establishment of the most suitable calibration methods



Work packages	Month					
	1	2	3	4	5	6
Literature research	▨					
Implementation and training	▨	▨	▨			
Adjusting and testing		▨	▨	▨		
Result extraction and writing		▨	▨	▨	▨	
Reservation						▨

Figure 1: Estimated schedule

4 Deliverables

Minimum Deliveries:

- Minimum:
  - A survey involving the analysis of the state-of-the-art in neural network calibration.
  - A theoretical comparative evaluation based on information from the researched literature.
  - Establishment of the most suitable calibration methods based on this evaluation.
- Expected:
  - Complete Python implementation of the calibration and the creation of neural network architecture with the methods applied.
  - Sufficient empirical results and derication of relevant observations.
  - Establishment of the most suitable calibration methods from the empirical data.

References

[BCKW15] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 1613–1622. JMLR.org, 2015.

[Bri50] G. W. Brier. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, 78:1, 1950.

[GG16] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 1050–1059. JMLR.org, 2016.

[GPSW17] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. *CoRR*, abs/1706.04599, 2017.

[Gra11] Alex Graves. Practical variational inference for neural networks. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2348–2356. Curran Associates, Inc., 2011.

[JKGG18] Heinrich Jiang, Been Kim, Melody Guan, and Maya Gupta. To trust or not to trust a classifier. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 5546–5557. Curran Associates, Inc., 2018.

[Kae18] Markus Kaengsepp. Calibration of convolutional neural networks. 2018.

[KFE18] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. *arXiv preprint arXiv:1807.00263*, 2018.

[LK17] Marc Moreno Lopez and Jugal Kalita. Deep learning applied to NLP. *CoRR*, abs/1703.03091, 2017.

[LPB17] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413, 2017.

[LW16] Christos Louizos and Max Welling. Structured and efficient variational deep learning with matrix gaussian posteriors. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1708–1716, New York, New York, USA, 20–22 Jun 2016. PMLR.

[MBC+99] David Mackay, John Bridle, Peter Cheeseman, Sidney Fels, Steve Gull, Andreas Herz, John Hopfield, Doug Kerns, Allen Knutsen, David Koerner, Mike Lewicki, Tom Loredo, Steve Luttrell, Ken Rose, Sibusiso Sibisi, John Skilling, Haim Sompolinsky, and Nick Weir. Bayesian methods for adaptive models. 05 1999.

[MGJG18a] Azadeh Sadat Mozafari, Hugo Siqueira Gomes, Steeven Janny, and Christian Gagné. A new loss function for temperature scaling to have better calibrated deep networks. *arXiv preprint arXiv:1810.11586*, 2018.

[MGJG18b] Azadeh Sadat Mozafari, Hugo Siqueira Gomes, Steeven Janny, and Christian Gagné. A new loss function for temperature scaling to have better calibrated deep networks. *CoRR*, abs/1810.11586, 2018.

[MSK18] Marcin Mozejko, Mateusz Susik, and Rafal Karczewski. Inhibited softmax for uncertainty estimation in neural networks. *CoRR*, abs/1810.01861, 2018.

[MW17] Amit Mandelbaum and Daphna Weinshall. Distance-based confidence score for neural network classifiers. *CoRR*, abs/1709.09844, 2017.

[Nea96] Radford M. Neal. *Bayesian Learning for Neural Networks.* Springer-Verlag, Berlin, Heidelberg, 1996.

[NMC05] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632. ACM, 2005.

[NMK+14] S. S. Nath, G. Mishra, J. Kar, S. Chakraborty, and N. Dey. A survey of image classification methods and techniques. In *2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, pages 554–557, July 2014.

[NZV18] L. Neumann, A. Zisserman, and A. Vedaldi. Relaxed softmax: Efficient confidence auto-calibration for safe pedestrian detection. In *Machine Learning for Intelligent Transportation Systems Workshop, NIPS*, 2018.

[Pla99] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pages 61–74. MIT Press, 1999.

[PM18] Nicolas Papernot and Patrick D. McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *CoRR*, abs/1803.04765, 2018.

[PNFCH15] Mahdi Pakdaman Naeini, Gregory F Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. *Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, 2015:2901–2907, 04 2015.

[Ten] Tensorflow probability. https://www.tensorflow.org/probability/. Accessed: 2019-01-06.

[ZE01] Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 609–616, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[ZE02] Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 694–699, New York, NY, USA, 2002. ACM.