Can You Trust Your Deep Neural Network?

UPPSALA UNIVERSITET

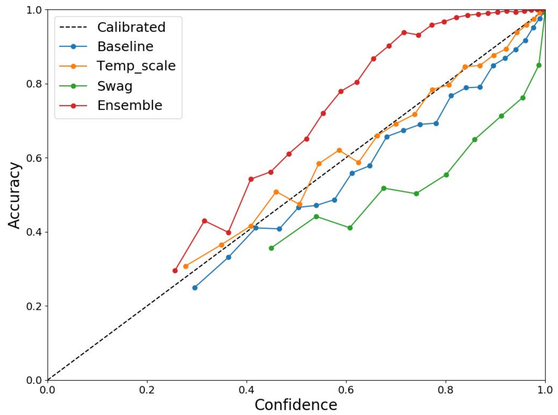
Authors:
Andreas Isaac
andreas.isaac.0280@student.uu.se
Melanie Andersson
melanie.andersson.8478@student.uu.se
Sara Hedar
sara.hedar.7006@student.uu.se

Supervisors:
Joakim Lindblad
Nataša Sladoje

Project in Computational Science 2019

Problem Formulation

Modern neural networks tend to be miscalibrated. In other words the networks are not aware of what they know and what they do not know. In this study three methods for producing more well-calibrated models are evaluated. The network structures LeNet-5, VGG-16 and ResNet-50 were applied to the MNIST dataset and evaluated using five certainty metrics.



Reliability diagram for VGG-16 using adaptive bins. Temperature scaling results in the most well-calibrated model as it is closest to the perfectly calibrated line. Ensemble methods produce overconfident models while SWAG produces underconfident models.

Methods

Temperature Scaling [1]**:** Rescales the input to the softmax layer. The scaling factor is calculated by minimising the Negative Log Likelihood.

Ensemble Methods [2]**:** Combines multiple models to obtain a more well-calibrated model.

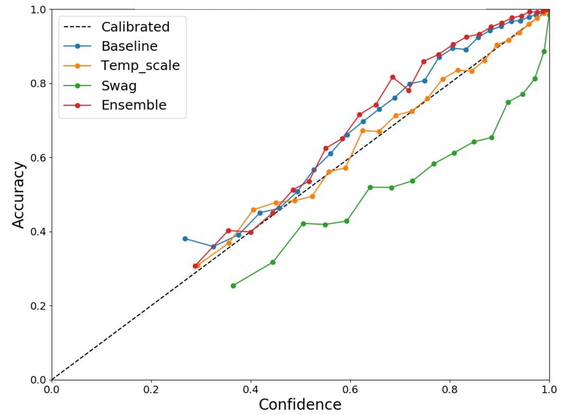
Stochastic Weight Averaging Gaussian (SWAG) [3]**:** An approximate Bayesian inference technique which extends SWA, where neural network weights are averaged.

Conclusion

Temperature scaling is the best method for achieving well-calibrated models. The basis of this conclusion is that the binning based metrics give more consistent results, the generated reliability diagrams and the simplicity of the implementation.

Certainty Metrics

To evaluate how well-calibrated the networks are the following certainty metrics were implemented: Negative Log Likelihood, Brier Score, Expected Calibration Error [4], Adaptive Expected Calibration Error [5] and Adaptive Maximum Expected Calibration Error [5]. The last three metrics are based on grouping the data points in bins.



Reliability diagram for LeNet-5 using adaptive bins. Temperature scaling is the most well-calibrated model while SWAG is furthest away from the perfectly calibrated line.

Result and Discussion

Temperature Scaling

+ Produces well-calibrated models.
+ Simple implementation.
+ Short training/test time.

Ensemble Methods

+ Promising since it allows a wide range of ensemble member types.
+ Short training/test time.
+ Produced models are not overconfident.

- Does not work well with deep convolutional neural networks.

SWAG

- Produces overconfident models.
- Long training/test time.
- Difficult to train models. → Needed to change hyperparameters.
- Difficult to compare with other methods.

Sources

[1] Chuan Guo et al. “On Calibration of Modern Neural Networks”. In: CoRR abs/1706.04599 (2017). arXiv:1706.04599.url: http://arxiv.org/abs/1706.04599.
[2] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. “Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles”. In: Advances in Neural Information Processing Systems 30. Ed. by I. Guyon e al. Curran Associates, Inc., 2017, pp. 6402–6413.
[3] Wesley Maddox et al. “A Simple Baseline for Bayesian Uncertainty in Deep Learning”. In: arXiv preprint arXiv:1902.02476 (2019).
[4] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. “Obtaining well calibrated probabilities using bayesian binning”. In:Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015.
[5] Yukun Ding et al. “Evaluation of Neural Network Uncertainty Estimation with Application to Resource-Constrained Platforms”. In: (2019). arXiv:1903.02050.