Calibration techniques for node classification using graph neural networks on medical image data

Author name(s) withheld EMAIL(S) WITHHELD
Address withheld

Editors: Under Review for MIDL 2023

Abstract

Miscalibration of deep neural networks (DNNs) can lead to unreliable predictions and hinder their use in clinical decision-making. This miscalibration is often caused by overconfident probability estimates. Calibration techniques such as model ensembles, regularization terms, and post-hoc scaling of the predictions can be employed to improve the calibration performance of DNNs. In contrast to DNNs, graph neural networks (GNNs) tend to exhibit underconfidence. In this study, we investigate the efficacy of calibration techniques developed for DNNs when applied to GNNs trained on medical image data, and compare the calibration performance of binary and multiclass node classification on a benchmark dataset and a medical image dataset. We find that post-hoc calibration using temperature scaling is the most effective method for improving the calibration performance, followed by the addition of a regularization term to the loss function during training. Our results further indicate that these calibration techniques are more effective for multiclass classification tasks compared to binary classification tasks.

Keywords: Calibration, graph neural networks, node classification, medical imaging

1. Introduction

The trustworthiness of deep neural networks (DNNs) has become an active topic of research over the past years. As DNNs are increasingly being used for clinical decision-making, it is important that their predicted probabilities accurately reflect the true underlying probabilities. Studies have shown that most DNNs are prone to miscalibration, often because of an overconfidence in their probability estimates (Guo et al., 2017). Methods to improve the calibration performance include averaging the predictions of multiple models, adding a regularization term during training, or post-hoc scaling of the predictions.

In contrast to DNNs, poor calibration observed in graph neural networks (GNNs) is usually attributed to underconfident estimates (Wang et al., 2021; Liu et al., 2022). Accordingly, calibration improvement techniques should decrease the entropy (thereby increasing the confidence) of probability estimates for correct predictions.

In this work, we investigate if calibration improvement methods developed for DNNs are generalizable to GNNs trained on medical image data. We compare calibration performances for both binary-and multiclass classification using GNNs trained on a benchmark dataset (i.e. Cora) and a dataset comprised of 150 3 D time-of-flight Magnetic Resonance Angiography (TOF-MRA) images visualizing the intracranial arterial vasculature.

© 2023 Anonymous.

2. Materials and methods

2.1. Data

2.1.1. Cora dataset

We included the Cora dataset as benchmark, which is a popular citation network dataset and generally considered a benchmark for node classification tasks (Sen et al., 2008). The Cora dataset consists of a single graph with 2,708 nodes and 5,429 edges, divided into 7 classes. Each node has 1,433 features and relates to a paper from a research area or class. We selected the majority class to train a binary classifier, which accounted for 30.2% of the class instances. For multiclass classification, we included all classes. The number of training samples per class was set at 20.

2.1.2. Medical image dataset

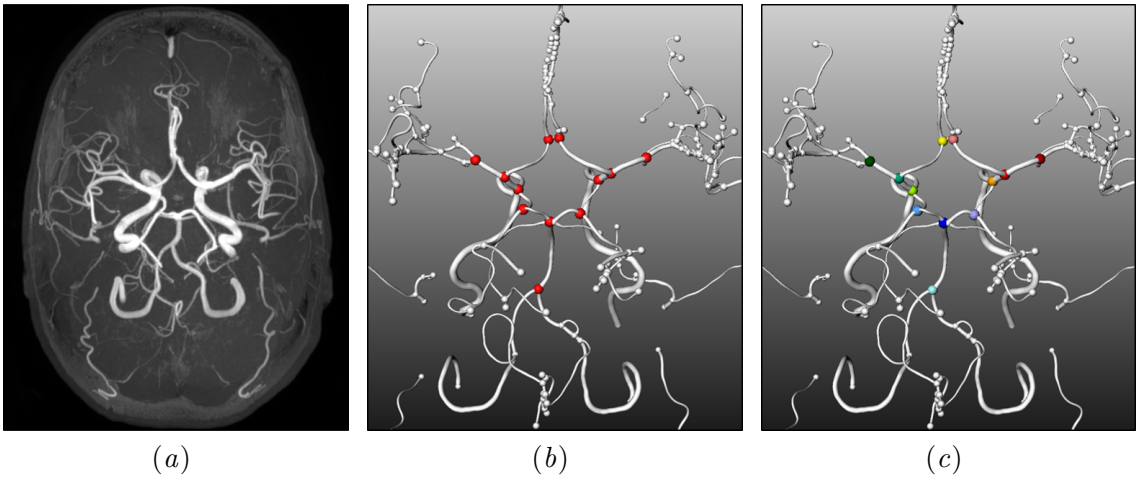
To obtain graphs from medical image data, we used 150 time-of-flight Magnetic Resonance Angiography (TOF-MRA) images from healthy subjects who were scanned on a 3 T Philips MRI system with a voxel size of 0.357 x 0.357 x 1.0 mm. We obtained segmentations of the intracranial arteries using a validated 3 D U-Net (Anonymous, 2021). Skeletonization by successive erosion of border voxels was performed to identify artery centerlines. We converted these skeletons into graph representations with nodes relating to artery bifurcations or endpoints and edges relating to connected arteries.

The node classification task focused on predicting the nodes forming the circle of Willis (CoW), an anastomosis of arteries located at the base of the brain. To train a binary classifier, we assigned positive class labels to all nodes or bifurcations forming the CoW. For multiclass classification, we assigned each CoW node a separate class (see Figure 1). All remaining nodes were regarded as background and assigned another class. For each node, the coordinate of its position, the radius, the number of connected edges and the logarithm of the sum of its connected edge lengths were included as features.

2.2. Calibration

We compared node classification results using five methods aimed at improving calibration, which are further described in this section.

- Temperature Scaling (TS): TS is a popular post-processing calibration method because of its simplicity and effectiveness (Guo et al., 2017). It uses a scalar parameter to rescale the logits of a network before the softmax layer.
- Monte-Carlo dropout (MC-dropout): this is a technique that can be used to approximate Bayesian uncertainty in DNNs (Gal and Ghahramani, 2016). During inference, we created a model ensemble by running a model multiple times (N=100) using an edge-level dropout layer before the final convolutional layer (λ=0.5). The average model estimates were used for evaluation.
- Model ensemble: we introduced another source of randomization by training multiple models (N=10) with different seeds and shuffling the training indices. Since the Cora dataset consists of a single graph, only the training indices for the medical image



(a) (b) (c)

Figure 1: Labelling of nodes. (a) Intracranial arteries on a 3 D TOF-MRA scan. (b) Node labelling for binary classification. (c) Node labelling for multiclass classification. Each color represents a specific target class (background nodes shown in white).

dataset were shuffled. We again averaged the model estimates to evaluate model performances.

- Accuracy-versus-uncertainty (AvU): we evaluated the use of a regularization term adapted from the study by Mody et al. (2022). The proposed loss function captures the relation between accuracy and uncertainty, and reaches a minimum when all accurate voxels are certain and all inaccurate voxels are uncertain. We follow their approach of computing an AvU loss for a range of certainty thresholds. The final loss function is the sum of the cross entropy loss and α*AvU loss, with $\alpha = 0.1$ or 1.0.

- Graph calibration loss (GCL): the study by Wang et al. (2022) introduced a modified cross-entropy loss to address the underconfidence of GNNs. We implemented their proposed loss function and optimized the calibration hyperparameter to $\gamma = 1.5$.

2.3. Experiments

Node classification was performed using PyTorch Geometric (version 1.12.1). We evaluated the classification performance using balanced accuracy to account for class imbalance. Miscalibration was assessed with reliability diagrams and the expected calibration error (ECE, M=10), for which we used the open-source Netcal calibration framework (Küppers et al., 2020).

2.3.1. Cora dataset

The dataset included masks for training (140 nodes), validation (500 nodes) and test (1000 nodes). We trained a two-layer graph convolutional network (GCN) (Kipf and Welling,

2016) using a hidden layer dimension of 16. The learning rate was set to 0.01, the weight-decay to 5e-3 and the number of epochs to 200. The Adam optimizer was used for parameter optimization. The training process was stopped if the validation loss had not decreased for 10 successive iterations.

2.3.2. Medical image dataset

The graphs from the medical image dataset were split into a training (N=80), validation (N=20), and test set (N=50). A two-layer GCN was trained with a hidden layer dimension of 16, a dropout rate of 0.1, and a batch size of 25. We used a learning rate of 0.01, weight decay of 5e-4, and the Adam optimizer. Early stopping was applied with a window size of 20.

Because of the large amount of background nodes, with only 4.97% of the nodes constituting to target classes, the ECE was evaluated for smaller, local regions in addition to the graph on a global-level (i.e. including all nodes). These local regions were defined by including a maximum of three neighbouring nodes for each target node (see an example visualization in Appendix A). The balanced accuracy and ECE were calculated on a subject-level and tested for statistically significant differences between methods using a Wilcoxon signed-rank test.

3. Results

Reliability diagrams are shown in Figure 2 for the Cora dataset and in Figure 3 for the medical image dataset. Table 1 and 2 summarize the results in terms of balanced accuracy and ECE per method. For all uncalibrated models, we observed an underconfidence in the predicted probabilities. ECE scores for multiclass classification models were lower than for binary classification models for both datasets.

3.1. Cora dataset

Uncalibrated node classification on the binary Cora dataset resulted in a balanced accuracy of 0.76 and an ECE of 0.124. Post-hoc calibration using TS resulted in slightly better calibrated probabilities for low confident estimates. Calibration techniques using average estimates from multiple models (MC- dropout and model ensemble) showed no improvement in model calibration. The AvU loss resulted in a slightly better calibrated model and higher accuracy. The model trained using GCL showed no improvements. Multiclass node classifications yielded a balanced accuracy of 0.83 and an ECE of 0.089. Similar to binary classification, TS improved model calibration whereas MC-dropout and the model ensemble did not affect calibration performances. Both the AvU loss and GCL resulted in better calibrated models, with similar ECE scores.

3.2. Medical image dataset

The mean number of nodes per graph was 171 $\pm$ 30. The local graph regions (based on a maximum of three target node neighbours) included a mean number of 29 $\pm$ 6 nodes. A mean balanced accuracy of 0.58, mean local ECE of 0.236 and mean global ECE of 0.030 was achieved for uncalibrated, binary node classification on the MI dataset. ECE scores

after TS showed a calibration improvement only locally. For MC-dropout and the model ensemble, no improvements in calibration were observed. Notably, the AvU loss and GCL resulted in worse discrimination and calibration performances than the uncalibrated model.

For multiclass classification, we obtained a mean balanced accuracy of 0.69, mean local ECE score of 0.233 and mean global ECE score of 0.049. In line with the previous results, TS improved calibration performances on a local- and global-level. MC-dropout showed small improvements in ECE scores at the cost of accuracy. The calibration performance of the model ensemble did not improve compared to the uncalibrated model. Both the AvU loss and GCL resulted in better calibrated models.

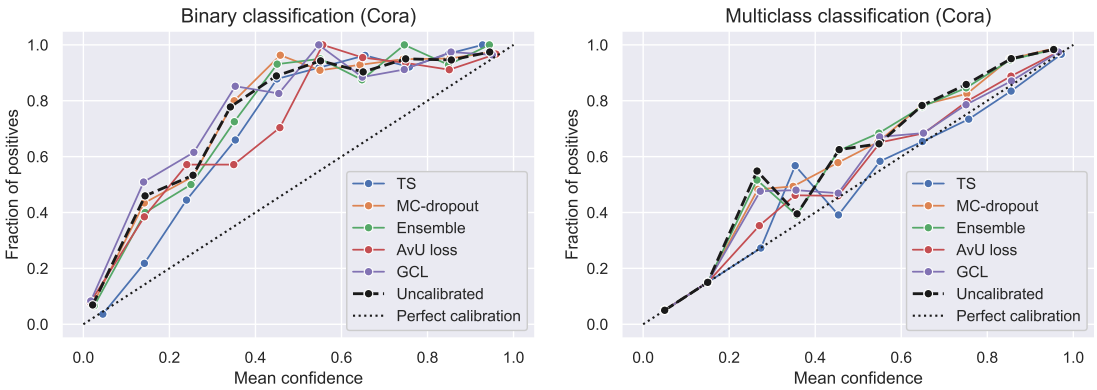


Figure 2: Reliability diagrams for different calibration techniques on the Cora dataset. Left: binary node classification. Right: multiclass node classification.

Table 1: Results for different calibration techniques on the Cora dataset. BA = balanced accuracy; ECE = expected calibration error; TS = Temperature Scaling; MC-dropout = Monte-Carlo dropout; AvU = accuracy-versus-uncertainty; GCL = graph calibration loss.

CORA	Binary classification		Multiclass classification	
	BA	ECE (x10^{-2})	BA	ECE (x10^{-2})
Uncalibrated	0.76	12.4	0.83	8.9
TS	0.76	10.7	0.83	2.3
MC-dropout	0.76	12.4	0.83	8.9
Ensemble	0.76	12.3	0.83	8.9
AvU loss	0.80	10.0	0.82	3.3
GCL	0.77	12.8	0.83	3.4

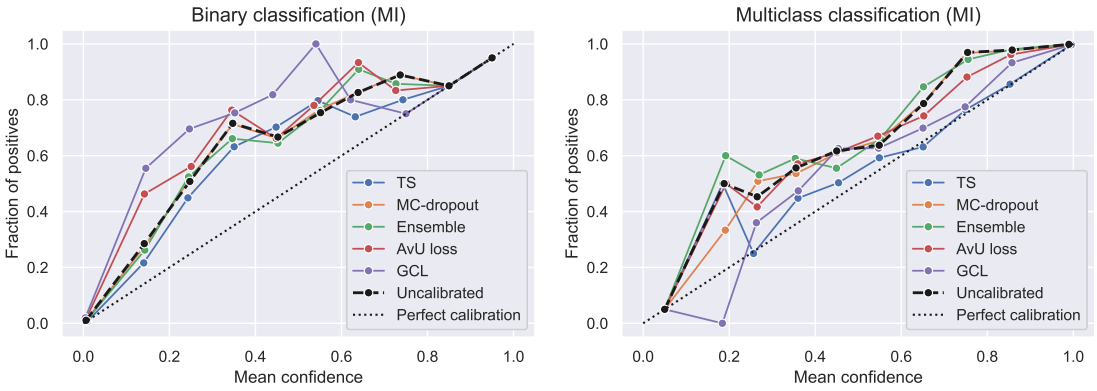


Figure 3: Reliability diagrams for different calibration techniques on the medical image (MI) dataset. Left: binary node classification. Right: multiclass node classification.

Table 2: Results for different calibration techniques on the medical image (MI) dataset. BA = balanced accuracy; ECE = expected calibration error; TS = Temperature Scaling; MC-dropout = Monte-Carlo dropout; AvU = accuracy-versus-uncertainty; GCL = graph calibration loss. The mean and standard deviation are calculated across all subjects. * represents a statistical significant difference (p<0.05) when compared to uncalibrated models, bold values denote significant improvements.

MI	Binary classification		
	BA	ECE local ($x10^{-2}$)	ECE global ($x10^{-2}$)
Uncalibrated	0.58 ± 0.07	23.6 ± 4.7	3.0 ± 1.0
TS	0.58 ± 0.07	**21.9 ± 4.5***	**2.7 ± 0.9***
MC-dropout	0.58 ± 0.08	23.1 ± 5.4	3.0 ± 1.0
Ensemble	0.58 ± 0.08	**22.8 ± 5.1***	3.0 ± 1.0
AvU loss	0.56 ± 0.06*	25.1 ± 4.5*	3.3 ± 1.1*
GCL	0.51 ± 0.03*	27.8 ± 4.5*	3.7 ± 1.1*
	Multiclass classification		
	BA	ECE local ($x10^{-2}$)	ECE global ($x10^{-2}$)
Uncalibrated	0.69 ± 0.19	23.3 ± 6.1	4.9 ± 1.4
TS	0.69 ± 0.19	**18.9 ± 5.2***	**2.8 ± 1.0***
MC-dropout	0.67 ± 0.17*	23.6 ± 5.5	4.9 ± 1.2
Ensemble	0.65 ± 0.18*	23.8 ± 5.9	5.0 ± 1.4
AvU loss	0.67 ± 0.19	22.7 ± 6.1	**3.9 ± 1.1***
GCL	0.64 ± 0.19*	**20.0 ± 6.0***	**3.2 ± 1.2***

4. Discussion and conclusion

In this work, we compared the performance of several calibration techniques for binary and multiclass node classification on the Cora dataset and a medical image dataset. Our results showed that multiclass node classification yielded better calibrated and more discriminative models than binary classification. Overall, best calibration performances were achieved using post-hoc TS, followed by the addition of a regularization term to the loss function during training for multiclass classification.

Calibration techniques such as MC-dropout and a model ensemble, which are based on averaging the predictions of multiple models, did not generally improve calibration performances for both datasets. Although previous studies have shown that ensembles trained with random initializations perform well (Fort et al., 2019), the underperformance observed in this study is likely attributed to the two-layer GCN architectures and their lack of diversity.

Furthermore, adding a regularization term during training yielded worse performances than the uncalibrated model for binary classification on the medical image dataset. This may be caused by a lack of model capacity, as balanced accuracy scores were relatively low. Other work has shown that the performance of node classification significantly improves using higher-order convolutional operators (Vos et al., 2022). Here, we implemented the GCN layer proposed by Kipf and Welling (2016) to compare the results to the Cora dataset. Although the model calibration was relatively good, the discriminative power was low. Combining a calibration technique with another graph operator may yield superior results in deriving a model that has both good calibration and discrimination.

We found lower mean ECE scores on a global-level than a local-level. The study of Hsu et al. (2022) discussed several factors affecting GNN calibration, among which the neighbourhood similarity of nodes. In other words, nodes that have similar label predictions as their neighbours tend to have lower calibration errors. This also provides an explanation for the higher ECE scores of local nodes: the nodes surrounding and representing the CoW have different label predictions and may be inclined to have similar predictions as their neighbours.

The balanced accuracy obtained for multiclass classification on the Cora dataset was comparable to the accuracy scores reported in literature (range 0.814-0.816) (Wang et al., 2021; Liu et al., 2022). Our ECE scores for the uncalibrated model and TS model were slightly lower than the scores reported in Wang et al. (2022). This may be caused by minor differences in model architecture (e.g. no dropout) or feature normalization. We did not perform node feature normalization because this did not prove to be beneficial on the MI dataset, where we included the node coordinates as features. Regardless, we obtained a similar ECE score on the Cora dataset using GCL during training.

In Mody et al. (2022), the authors propose the use of an additional loss term that penalizes high probability of uncertainty in large accurate regions (e.g. at the core of a segmented organ). Such implementations or further optimization of the loss function may also improve the current calibration performance.

Our results show that calibration techniques developed for DNNs to mitigate overconfidence can improve calibration for underconfident GCN models trained on medical image data. Post-hoc scaling using TS yielded the best calibration performance. Taking into

account that this method is also accuracy preserving, we would recommend the use of TS on medical image data. Future work will evaluate calibration performances for other convolutional operators.

References

Anonymous. Details omitted for double-blind reviewing. 2021.

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

Hans Hao-Hsun Hsu, Yuesong Shen, Christian Tomani, and Daniel Cremers. What makes graph neural networks miscalibrated? *arXiv preprint arXiv:2210.06391*, 2022.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Fabian Küppers, Jan Kronenberger, Amirhossein Shantia, and Anselm Haselhoff. Multivariate confidence calibration for object detection. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

Tong Liu, Yushan Liu, Marcel Hildebrandt, Mitchell Joblin, Hang Li, and Volker Tresp. On calibration of graph neural networks for node classification. *arXiv preprint arXiv:2206.01570*, 2022.

Prerak Mody, Nicolas F Chaves-de Plaza, Klaus Hildebrandt, and Marius Staring. Improving error detection in deep learning based radiotherapy autocontouring using bayesian uncertainty. In *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, pages 70–79. Springer, 2022.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. Collective classification in network data. *AI magazine*, 29(3):93–93, 2008.

IN Vos, YM Ruigrok, KM Timmins, BK Velthuis, and HJ Kuijf. Improving automated intracranial artery labeling using atlas-based features in graph convolutional nets. In *Medical Imaging 2022: Image Processing*, volume 12032, pages 314–320. SPIE, 2022.

Min Wang, Hao Yang, and Qing Cheng. Gcl: Graph calibration loss for trustworthy graph neural network. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 988–996, 2022.

Xiao Wang, Hongrui Liu, Chuan Shi, and Cheng Yang. Be confident! towards trustworthy graph neural networks via confidence calibration. *Advances in Neural Information Processing Systems*, 34:23768–23779, 2021.

Appendix A. Expected calibration error (ECE)

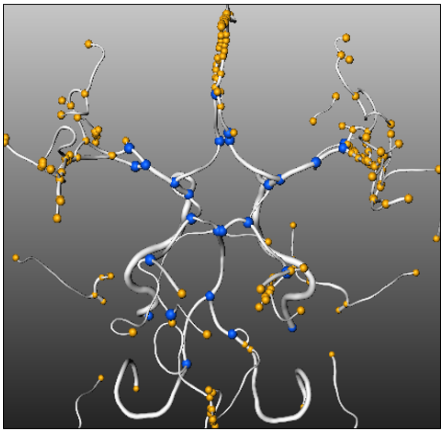


Figure 4: Visualization of nodes considered for calculation of the expected calibration error (ECE) on the medical image dataset. Nodes marked with a blue color were used to calculate local ECE scores. Nodes marked with an orange color represent the remaining nodes. For global ECE scores, all nodes were included.