CS6208 Project
MaxEnt Loss for Calibration of Graph Neural Networks under Out-of-distribution shift

Neo Yuan Rong Dexter
A0214975W

1 Introduction

We present a new, simple and effective loss function for calibrating graph neural networks. Miscalibration is the problem whereby a model's probabilities do not reflect it's correctness, making it difficult for real-world deployment. Our findings show that graph neural networks are not immune to calibration issues and remain miscalibrated on both in-distribution (ID) and out-of-distribution (OOD). We compare our method against other baselines on a novel graph-form of Celeb-A faces dataset and evaluate it on both ID and OOD.

2 Preliminaries

Expected Calibration Error (ECE) Calibration is the measure of how a model's predicted confidence is aligned with its correctness. Both over- and under-confident predictions lead to miscalibration and non-meaningful probabilities. There are many different metrics that can be used to measure calibration, with the most popular metric estimated using ECE [1], which divides the model's probabilities into B bins. The number of samples, average accuracy and confidence for each bin is represented by n_b, *acc* and *conf*. The weighted differences between each *acc* and *conf* bins are measured: ECE $= \sum_{b=1}^{B} \frac{n_b}{N} |acc(B_b) - conf(B_b)|$.

Out-of-distribution Shifts Typically, a classifier is trained with the assumption that the test distribution is aligned with the training set. This assumption would no longer hold true, when inputs deviate greatly from the train set. These shifted samples are considered OOD and can be caused by differences in resolution, lighting, blurs or noise [2]. An illustration of OOD is shown in Figure 1, where for computer vision tasks images can be simply augmented to replicate an OOD shift, for graphs we propose a similar task and use a resolution shift (e.g different number of nodes and edges). OOD samples generally cause a drop in recognition performance and miscalibration.

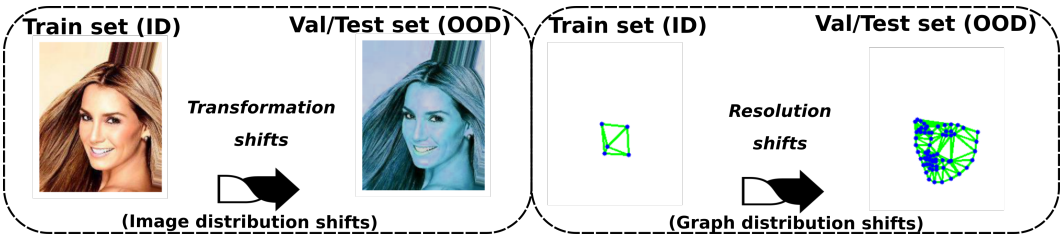

Figure 1: Distribution shift often occur in traditional computer vision tasks (left), for the graph ML tasks, we propose to use a ODD shift in resolution (right)

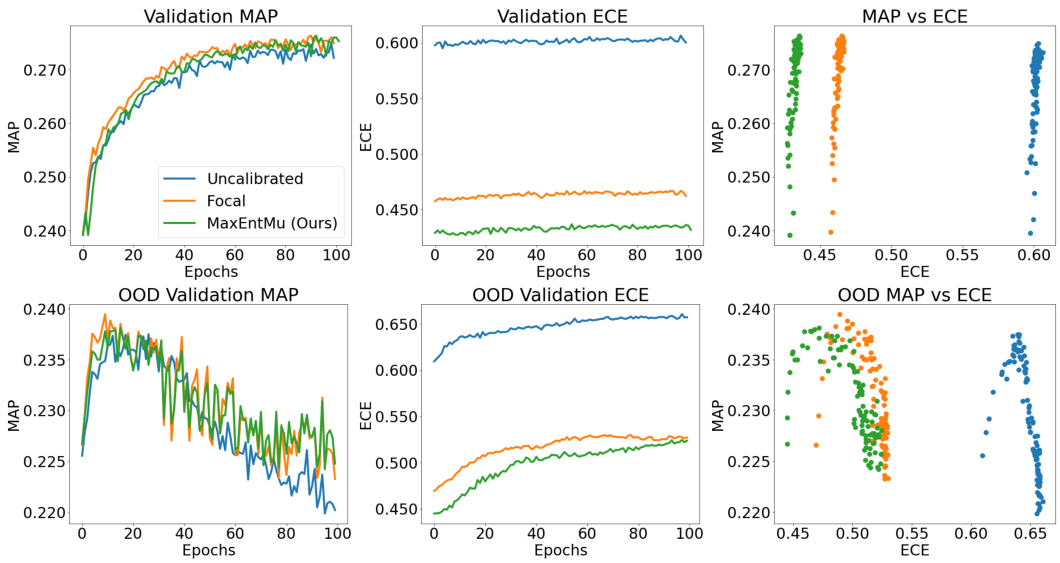

Figure 2: Learning curves comparing the performances of different loss functions on ID (top) and OOD (bottom) samples.When evaluated OOD, all methods have a decrease in performance, however MaxEnt Loss still delivers the best calibration performance overall.

3 MaxEnt Loss

For classification tasks, our models are typically trained using the ground-truth onehot labels y_k, where the cross-entropy loss between the predictions p_k is minimized across all classes K and $\mathcal{Y} = \{1, ..., K\}$ is a vector containing each random variable. The binary form of the cross entropy loss is written as $\mathcal{L}_{CE} = -\sum_k y_k \log(p_k) - \sum_k (1 - y_k) \log(1 - p_k)$. Originally proposed for object detection tasks, Focal loss [3] has also been show to improve model calibration [4]. The hyperparameter γ is used to control the order of the polynomial focal term and the binary form of the Focal loss is given as: $\mathcal{L}_F = -\sum_k (1 - p_k)^\gamma y_k \log(p_k) - \sum_k p_k^\gamma (1 - y_k) \log(1 - p_k)$.

Based on the Principle of Maximum Entropy [5], we propose MaxEnt loss by adding the following constraint to the Focal loss. Where the scalar μ is a global constraint computed from the ratio for each class. Specifically, the prior distribution $\frac{N_k}{N}$ is used to compute the expectation where $\mathbb{E}[\mathcal{Y}] = \sum_k \mathcal{Y} \frac{N_k}{N} = \mu$. Subjected to the given constraint, the Lagrange multiplier λ_1 can thereafter be computed for each attribute numerically using Newton Raphson. The MaxEnt Loss is given by:

$$\mathcal{L}_{ME}^{M} = \underbrace{\mathcal{L}_F}_{\text{Focal Loss}} + \underbrace{\lambda_1 \left(\sum_k \mathcal{Y} p_k - \mu \right)}_{\text{Mean constraint}} \tag{1}$$

4 Evaluation

For our experiments, we use Pytorch's implementation of GCNConv and train uncalibrated models as a baseline using cross entropy loss and compare it with Focal and MaxEnt loss. In Figure 2, we can see that all loss functions converge to similar MAP, however cross entropy loss results in the worst calibration performance. Similar to standard deep learning tasks, Focal loss improves the calibration performance greatly with an overall improvement of roughly 10%. Our method achieves the best performance, with an additional improvement of 5%. As expected, we observe a drop in MAP and ECE for all methods when evaluated OOD, but MaxEnt Loss is still able to deliver the best calibration performance OOD.

References

[1] Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *Proc. 29th AAAI Conference on Artificial Intelligence*, page 2901–2907, 2015.

[2] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv*, abs/1610.02136, 2017.

[3] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *arXiv*, abs/1708.02002, 2017.

[4] Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip HS Torr, and Puneet K Dokania. Calibrating deep neural networks using focal loss. In *Advances in Neural Information Processing Systems*, 2020.

[5] E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106:620–630, May 1957.