Last-layer committee machines for uncertainty quantification

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce here a form of committee machines that gives good predictions of classification confidence, while being computationally efficient. The initial development of this method was motivated by our work on benthic mapping based on a large dataset of ocean floor images. These wild type images vary dramatically in terms of their classification difficulty and often result in low inter-rater agreement. We show that our method is able to identify difficult to classify images using model uncertainty, consistent with Bayesian neural networks and Monte Carlo sampling. However, our method drastically reduces the computational requirements and offers a more efficient strategy. This enables us to provide these uncertain predictions to a human specialist and offers a form of active learning to enhance the classification accuracy of the dataset. We provide both a benchmark study to demonstrate this approach and first results of the BenthicNet dataset.

1 Introduction

Uncertainty quantification of model predictions plays an important role for predictions of high-risk applications (Gawlikowski et al. 2023; Hüllermeier and Waegeman 2021; Huang et al. 2024). However, evaluating uncertainties of model predictions are typically computationally expensive or difficult to implement. This is in part, due to applications of large networks and datasets that are commonly used today (*e.g.*, large language models, transformers, ResNets, etc.). Bayesian and their approximating methods, such as Monte Carol sampling and ensembles require sampling multiple models or subsets of the training dataset. While, recent approaches have been proposed to address these concerns (Harrison et al. 2024; Lee et al. 2015; Lakshminarayanan et al. 2017), there continues a need for more efficient and accessible techniques.

The motivation for the development of this work is related to the advancement of benthic habitat mapping using underwater seafloor images from the hierarchical BenthicNet dataset (Lowe et al. 2024; Misiuk et al. 2024). This dataset is a collection of global seafloor images with corresponding annotation files which are labelled according to the CATAMI classification scheme (Althaus, Hill, Edwards, et al. 2014; Althaus, Hill, Ferrari, et al. 2015). We selected two sub-datasets: German Bank 2010 and Substrate (depth 2). These datasets are available as one-hot classifications and represents images which are known to be difficult because of low inter-class heterogeneity represented across the different samples (Xu et al. 2024; Humblot-Renaux et al. 2024).

In this study, we propose last-layer committee machines (LLCMs) as an ensemble method using shared network parameters for a classifier consisting of M committee machines (*i.e.*, linear layers). We demonstrate the requisite network diversity of the penultimate layer is facilitated by random weight initialization and can be increased by enabling different hyperparameters during training such as logit normalization (Wei et al. 2022) or label smoothing (Szegedy et al. 2015). We intro-

Submitted to Workshop on Bayesian Decision-making and Uncertainty, 38th Conference on Neural Information Processing Systems (BDU at NeurIPS 2024). Do not distribute.

duce and demonstrate LLCMs on the MNIST dataset before evaluating on the substantially more difficult subsets of the BenthicNet dataset. LLCMs offers a comparable (or competitive) strategy for Bayesian approximations, where uncertainty quantifications are obtained in a single forward-pass during inference for both the dataset and single predictions.

2 Methods

2.1 Last-layer committee machines

A committee machine (or committee method) is a form of an ensemble learner used to boost model performance by averaging over multiple models, often used with classification and regression trees such as random forests. In the case of neural networks, average probability distributions can be computed as,

$$\bar{p}(y|\mathbf{x};\mathbf{w}) = \frac{1}{M}\sum_{m=1}^{M} p(y|\mathbf{x};\mathbf{w}^m) \tag{1}$$

where M represents the number of models and individual probability distributions are calculated from output logits from using network weights $\mathbf{w}^m$ and the softmax function. These deep ensembles often require substantial computational effort and memory usage, which is largely the result of averaging large networks and/or datasets. Rather than using the full network as ensembles, we propose to use a list of linear layers as a last-layer committee machine as part of the network architecture. Training such a network involves backpropagating the mean loss of the committee machine, which effectively results with gradient updates from each committee machine member (Figure 1).

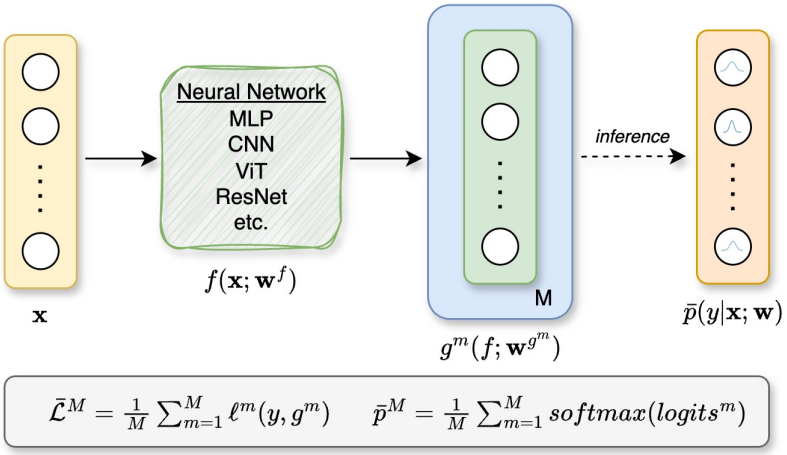


Figure 1: Network architecture of a last-layer committee machine. Input features are obtained from a network which are subsequently passed to the LLCM module. For each forward pass, mean loss is computed using M committee machine members which is then backpropagated during training. At inference time, a distribution of softmax distributions is obtained by averaging committee machine members.

For every forward-pass during training, each member of the committee machine receives identical feature representations; therefore, network diversity or loss exploration during backpropagation is dependent on the initialization of each member. That is, if committee machine members are identically initialized, they will all have the same weights during and after training. If they are non-identically initialized, each committee machine member will explore its own loss landscape. We show that including techniques such as, logit normalization, label smoothing, and/or class weighting during training can influence the degree of network diversity of model weights of the LLCM module.

These simple modifications result with a network architecture that can provide model uncertainty for both the dataset and single samples at inference. It drastically reduces computational and memory intensive requirements and can be parallelized and/or scaled to multiple devices. In addition, LLCMs can be used with several feature extracting networks, such as MLPs, CNNs, ViT, ResNets, etc.

2.2 Parametric confidence score metric

To evaluate performance and model uncertainty, a parametric confidence score metric was developed that uses the top-2 mean probabilities and their corresponding standard deviations from predicted mean softmax distributions (Figure 2).

This approach provided a metric whereby scores can be modulated by parameter k and standard deviations of the top-2 mean softmax probabilities. This effectively results with predictions based on model uncertainty, reflective on the factor of the standard deviations. We used multiple k values and results were reported for accuracy, F1-score, and confusion matrices. We plotted accuracy with respect to the fraction of remaining of samples after applying each k-value. Correct and incorrect model predictions were computed from confusion matrices and also plotted against remaining samples. Together, these plots provide valuable insights on model performance, uncertainty, and a visual representation to compare different models and hyperparameters.

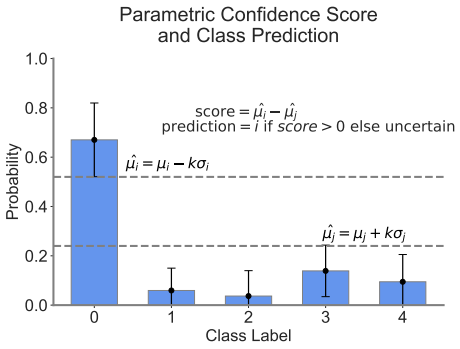


Figure 2: Calculation of confidence score and class prediction. For each sample mean softmax distribution $\mathcal{P}$, let i (first argmax) $= \text{argmax}\,\mathcal{P}$, j (second argmax) $= \text{argmax}^2\,\mathcal{P}$, and k a parameter for scaling standard deviations. Class label i is only returned if confidence score > 0 otherwise, the sample is referred to a domain expert for review.

3 Experiments

3.1 Network diversity of LLCMs

Our initial concern with this approach was that averaging committee members would decrease network diversity or will would converge during training. To investigate this, we created a 2-block CNN (Conv2d-ReLU-MaxPool2d) with a LLCM module consisting of 10 committee members with 2048 input features and 10 output features (classes). After training the MNIST dataset, each committee member was flattened and the coefficient of variations (CV) for the learned weights were computed (*i.e.*, 2048×10 weights per member). Committee members with similar CV values would equate to being similar; therefore, we report the standard deviations of CV values across all committee machine members (Table 1).

Table 1: Model performances using a 10-member LLCM and the MNIST dataset.

Model[1]	**Accuracy**	**F1-score**	**CS**[2]	**BS**[3]	**ECE**[4]	**CV**$_\sigma$[5]
A: $-/-/-$	0.991	0.991	0.979	0.003	0.004	9.537
B: $-/+/-$	0.990	0.991	0.776	0.072	0.181	906.146
C: $+/+/-$	0.989	0.989	0.771	0.074	0.185	8372.234
D: $+/-/-$	0.991	0.991	0.979	0.003	0.003	9.414
E: $+/-/+$	0.992	0.992	0.841	0.026	0.115	25.502
F: $-/-/+$	0.992	0.992	0.842	0.026	0.115	23.713
A:[6] $-/-/-$	0.991	0.991	0.980	0.003	0.001	**0.0**

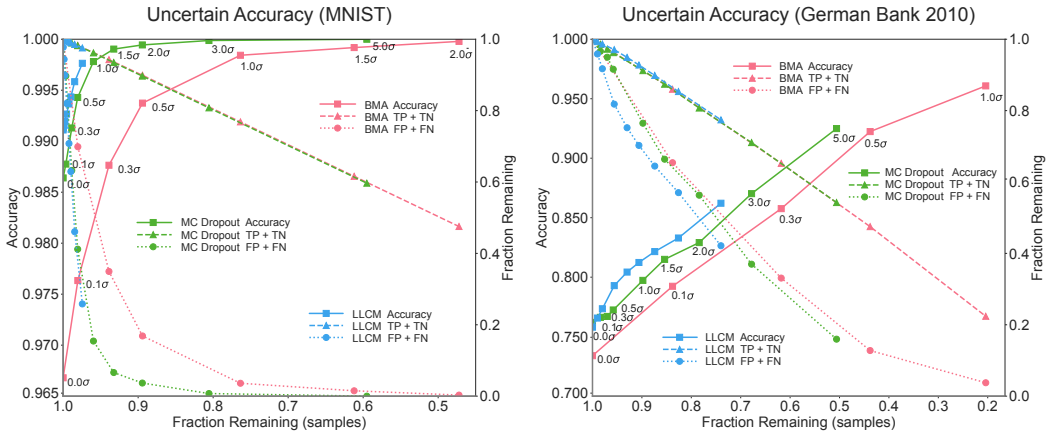
[1] Models A-F are defined based on training hyperparameters using class weights / logit normalization / label smoothing (amount of smoothing, 0.1). These hyperparameters are either applied (denoted by $+$) or omitted (denoted by $-$). [2] Confidence score (CS) is defined as the difference between the top-2 mean softmax probabilities. [3] Brier score (BS). [4] Expected calibration error (ECE). [5] Refer to text. [6] All committee machine members weights were initialized with ones.

The last entry of Table 1 shows the result where all committee machine members had weights initialized to ones. This resulted with a CV$_\sigma$ of 0.0, which indicates that all members are identical. However, by using random initialization of weights with/without different training hyperparame-

ters, a diverse range of committee machine members can be obtained. We do observe from the BS and ECE, that models may require additional calibration (*e.g.*, temperature scaling).

3.2 Comparison of model uncertainty using LLCMs, BNNs, and Monte Carlo Dropouts

We compared model uncertainty of the LLCM method to Bayesian model averaging (BMA) and Monte Carlo (MC) dropout using the parametric confidence score metric (Figure 2). We converted the CNN described above by replacing the LLCM with a single classifier and then used the utilities from the Pyro framework (Bingham et al. 2019) to convert the network to a Bayesian model. For the MC dropout experiments, we added a dropout layer at the end of each block before sampling with a rate of 0.1. In the case of the LLMC, we used a 100-member committee machine. Figure 3a show the results for the MNIST dataset.



(a) (b)

Figure 3: Parametric sigma plots for BMA, MC dropout, and LLCM using the (a) MNIST and (b) German Bank 2010 datasets. For both the MC dropout and LLCM, results are reported using class weights and logit normalization. BMA and MC dropouts were performed using 100 sampling of weights.

Model uncertainty can be realized by noticing the increase in accuracy as the $k \times \sigma$ increases, which removes uncertain model predictions. An important feature of this plot is the differential decay rates for the correct model predictions (dashed line) and the incorrect model predictions (dotted lines) for each model. This preferred removal of samples offers a utility to: 1) identify uncertain samples, and 2) boost overall performance and model confidence.

We next investigated the more challenging BenthicNet dataset. As an example, we show the results from the German Bank 2010 dataset (Figure 3b). In this case, we used a BenthicNet pre-trained ResNet-50 network (Xu et al. 2024) to create a last-layer BNN, LLCM, and added dropout layers ($p = 0.01$) after each ReLU activation for the MC dropout experiments. Immediately, we see the effects of using a challenging wild type dataset. However, even in this case we can identify uncertain samples that require external review. All models used for both datasets were capable of evaluating model uncertainty and resulted with uncertain samples with LLCMs being comparable (or competitive) with current approaches.

4 Conclusions

In this study, we proposed the LLCMs as a method to evaluate model uncertainty and identify uncertain samples for review by a domain expert. For areas such as health, autonomous driving, ocean management, and other high-risk areas deploying machine learning, having a human-in-loop during decision-making can be beneficial, if not essential. The LLCM method presented here scales well both on model sizes and datasets, offering an efficient approach for Bayesian approximations and a strategy to identify uncertain model predictions.

References

Althaus, F., Hill, N., Edwards, L., and Ferrari, R. (2014). *CATAMI classification scheme for scoring marine biota and substrata in underwater imagery: a pictorial guide to the collaborative and annotation tools foranalysis of marine imagery and video (CATAMI) classification scheme*. Version 1.4. CATAMI.org, pp. 1–102. URL: https://catami.org/wp-content/uploads/sites/2/2023/03/CATAMI_class_PDFGuide_V4_20141218.pdf (Retrieved February 19, 2024).

Althaus, F., Hill, N., Ferrari, R., Edwards, L., Przeslawski, R., Schönberg, C. H. L., Stuart-Smith, R., Barrett, N., Edgar, G., Colquhoun, J., Tran, M., Jordan, A., Rees, T., and Gowlett-Holmes, K. (2015). "A standardised vocabulary for identifying benthic biota and substrata from underwater imagery: the CATAMI classification scheme". In: *PLOS ONE* **10**(10), pp. 1–18. DOI: 10.1371/journaltitle.pone.0141039.

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., and Goodman, N. D. (2019). "Pyro: deep universal probabilistic programming". In: *Journal of Machine Learning Research* **20**(28), pp. 1–6. URL: https://jmlr.org/papers/volume20/18-403/18-403.pdf (Retrieved February 7, 2024).

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., and Zhu, X. X. (2023). "A survey of uncertainty in deep neural networks". In: *Artificial Intelligence Review* **56**, S1513–S1589. DOI: 10.1007/s10462-023-10562-9.

Harrison, J., Willes, J., and Snoek, J. (2024). "Variational Bayesian last layers". In: *International Conference on Learning Representations (ICLR)*. Vienna, Austria. DOI: 10.48550/arxiv.2404.11599. arXiv: 2404.11599 [cs.LG].

Huang, L., Ruan, S., Xing, Y., and Feng, M. (2024). "A review of uncertainty quantification in medical image analysis: probabilistic and non-probabilistic methods". In: *Medical Image Analysis* **97**(103223), pp. 1–27. DOI: 10.1016/j.media.2024.103223.

Hüllermeier, E. and Waegeman, W. (2021). "Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods". In: *Machine Learning*. Vol. 110. 3, pp. 457–506. DOI: 10.1007/s10994-021-05946-3.

Humblot-Renaux, G., Johansen, A. S., Schmidt, J. E., Irlind, A. F., Madsen, N., Moeslund, T. B., and Pedersen, M. (2024). "Underwater uncertainty: a multi-annotator image dataset for benthic habitat classification". In: *European Conference on Computer Vision (ECCV)*. URL: https://vbn.aau.dk/en/publications/underwater-uncertainty-a-multi-annotator-image-dataset-for-benthi (Retrieved September 1, 2024).

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). "Simple and scalable predictive uncertainty estimation using deep ensembles". In: *Conference on Neural Information Processing Systems*. Long Beach, CA. DOI: 10.48550/arxiv.1612.01474. arXiv: 1612.01474 [stat.ML].

Lee, S., Purushwalkam, S., Cogswell, M., Crandall, D., and Batra, D. (2015). *Why M heads are better than one: training a diverse ensemble of deep networks*. DOI: 10.48550/arxiv.1511.06314. arXiv: 1511.06314 [cs.CV].

Lowe, S. C. et al. (2024). *Benthicnet: a global compilation of seafloor images for deep learning applications*. arXiv: 2405.05241 [cs.CV].

Misiuk, B., Lowe, S., and Xu, I. (2024). *Benthicnet*. Federated Research Data Repository. DOI: 10.20383/103.0614.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2015). *Rethinking the Inception architecture for computer vision*. arXiv: 1512.00567 [cs.CV].

Wei, H., Xie, R., Cheng, H., Feng, L., An, B., and Li, Y. (2022). *Mitigating neural network overconfidence with logit normalization*. arXiv: 2205.09310 [cs.LG].

Xu, I., Misiuk, B., Lowe, S. C., Gillis, M., Brown, C., and Trappenberg, T. (2024). "Hierarchical multi-label classification with missing information for benthic habitat imagery". In: *International Joint Conference on Neural Networks (IJCNN)*. Yokohama, Japan.