11-2023

Optimized Uncertainty Estimation for Vision Transformers: Enhancing Adversarial Robustness and Performance Using Selective Classification

Erik Pautsch
Loyola University Chicago, epautsch@luc.edu

John Li
University of California - San Diego

Silvio Rizzi
Argonne National Laboratory, srizzi@anl.gov

George K. Thiruvathukal
Loyola University Chicago, gkt@cs.luc.edu

Maria Pantoja
California Polytechnic State University, San Luis Obispo

Recommended Citation

Pautsch, Erik; Li, John; Rizzi, Silvio; Thiruvathukal, George K.; Pantoja, Maria (2023). Optimized Uncertainty Estimation for Vision Transformers: Enhancing Adversarial Robustness and Performance Using Selective Classification. In proceedings of First Workshop on Enabling Predictive Science with Optimization and Uncertainty Quantification in HPC (EPSOUQ-HPC 2023), https://doi.org/10.6084/m9.figshare.24397216

Optimized Uncertainty Estimation for Vision Transformers: Enhancing Adversarial Robustness and Performance Using Selective Classification

Erik Pautsch
Loyola University Chicago
Chicago, USA
epautsch@luc.edu

John Li
UCSD
San Diego, USA
jzl011@ucsd.edu

Silvio Rizzi
Argonne National Lab
Chicago, USA
srizzi@anl.gov

George K. Thiruvathukal
Loyola University Chicago
Chicago, USA
gthiruvathukal@luc.edu

Maria Pantoja
CalPoly
San Luis Obispo, USA
mpanto01@calpoly.edu

ABSTRACT

Deep Learning models often exhibit undue confidence when encountering out-of-distribution (OOD) inputs, misclassifying with high confidence. The ideal outcome, in these cases, would be an "I do not know" verdict. We enhance the trustworthiness of our models through selective classification, allowing the model to abstain from making predictions when facing uncertainty. Rather than a singular prediction, the model offers a prediction distribution, enabling users to gauge the model's trustworthiness and determine the need for human intervention. We assess uncertainty in two baseline models: a Convolutional Neural Network (CNN) and a Vision Transformer (ViT). By leveraging these uncertainty values, we minimize errors by refraining from predictions during high uncertainty. Additionally, we evaluate these models across various distributed architectures, including new AI architectures, Cerebras CS-2, and SambaNova SN30.

KEYWORDS

Uncertainty, Deep Learning, Ensembles, Evidential Learning, Selective Classification

ACM Reference Format:
Erik Pautsch, John Li, Silvio Rizzi, George K. Thiruvathukal, and Maria Pantoja. 2023. Optimized Uncertainty Estimation for Vision Transformers: Enhancing Adversarial Robustness and Performance Using Selective Classification. In *Proceedings of SuperComputer Conference (SC '23).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Selective classification in deep learning (DL) [6] addresses the challenge of overconfident misclassifications by enabling models to abstain from predictions during uncertain scenarios. Uncertainty evaluation is critical for some DL tasks, for example, for pedestrian detection in self-driving cars [2] where an AI model may not recognize a human since he is partially covered by a bicycle or any new (for training) situation like this. In [13], researchers train a path-planning robot to ask for help when it does not understand the command by measuring the uncertainty of Large Language Models (LLM). In [12], the authors benchmarked one pointwise and three approximate Bayesian DL models to predict uncertainty in the classification of cancer of unknown primary, using three RNA-seq datasets. Evaluation of uncertainty is crucial, and being able to output "I do not know" and ask for help is as important as having high accuracy.

Numerous techniques have been developed to quantify uncertainties in DL. For instance, Temperature Scaling [7] is a post-processing technique to calibrate neural networks. A limitation of this method is the need for the scaling parameter to be learned during DL training. Bayesian Neural Networks (BNN) [5], which transform neuron weights into probability distributions, can sometimes yield subpar results, especially when employing Monte Carlo dropout [5]. In [11], researchers present Peer Loss Functions; this approach enables models to learn even from noisy labels.

This paper focuses on deep ensembles and evidential learning for uncertainty quantification. While ensembles are recognized for their comprehensive uncertainty data, evidential learning stands out for its computational efficiency and lack of need for extra training stages. The **main contributions** of our study are the introduction of uncertainty evaluations to ViTs, a detailed comparison of select uncertainty estimation techniques tailored to our context, and a cursory examination of optimizing uncertainty calculation with High-Performance Computing (HPC) paradigms. Figure 1 highlights an example where a ViT-based classifier misclassifies under adversarial conditions. While the mean cams for both images are very similar, there is a significant discrepancy in the uncertainty. In the adversarial case, the high uncertainty indicates potential misclassification, allowing for intervention or rejection of the result.

The paper is organized as follows. In Section 2, we describe the datasets, uncertainty evaluation methods, and accelerations we performed; in Section 3, we present a summary of our results; and in Sections 4 and 5 we present our conclusion and future work respectively.

SC '23, November 12-17, 2023, Denver, CO

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 METHODS

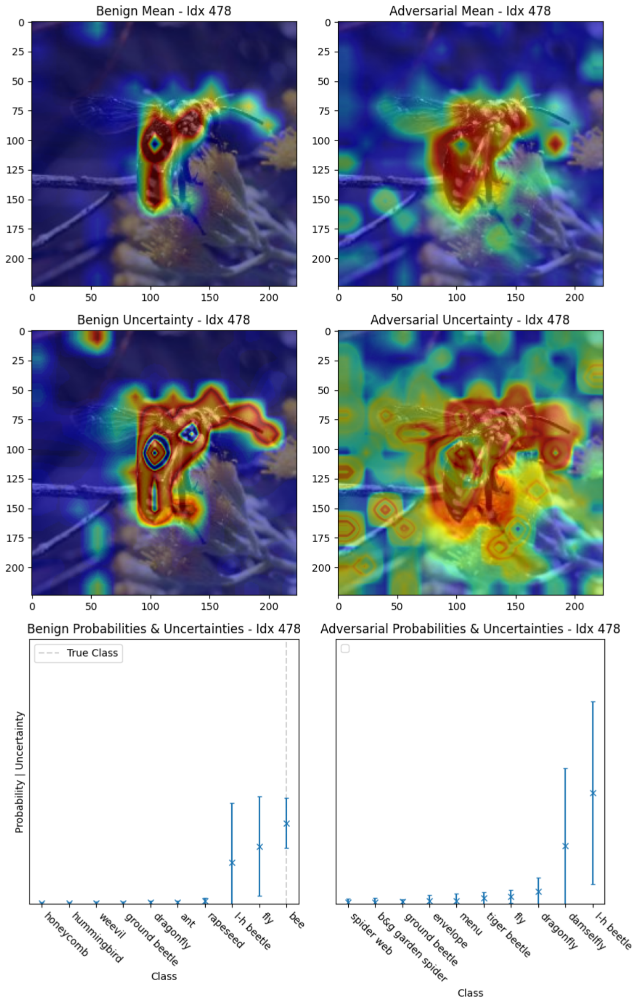

Figure 1: Class Activation Maps (CAM), first four images, for ViT Classifier (axis represents pixel values). The uncertainty of the adversarial cam is high despite the mean being similar to the benign one (last two images). The classifier misclassifies under Adversarial attack, but the uncertainty is high and is rejected by the classifier.

We explored multiple uncertainty estimation methods, with ensembles and evidential learning emerging as the most promising due to their reliability and efficiency, respectively.

2.1 Ensembles

Ensembles [15] aggregate outputs from several models trained independently on identical data. These models arise from Hyper-Parameter Optimization (HPO), thus negating extra computational time. By aggregating the M best models, we yield optimal hyperparameters and uncertainty estimations. With input x and label y, a neural network (NN) predicts the distribution $p_\theta(y \mid x)$ where θ represents the NN's parameters. We harness an ensemble of M models and their hyperparameters, $(\theta_m)_{m=1}^{M}$, training them in parallel on the same dataset. The combined prediction is: $p(y \mid x) = M^{-1} \sum_{m=1}^{M} p_{\theta_m}(y \mid x, \theta_m)$. Uncertainties are calculated from the standard deviation of predicted probabilities across the ensemble.

2.2 Evidential Learning

Evidential Learning [1, 14] differentiates itself by replacing a classical NN's softmax layer with an evidence vector, predicting the Dirichlet distribution parameters. For a given sample x_i for a K-class classification problem, the cross entropy loss to be minimized for learning evidence e_i is $\mathcal{L}(i_i, e_i, \theta) = \sum_{k=1}^{K} y_i^k (logS_i - log(e_i^k + 1))$; where y_i is the one-hot k-dimmensional label, S is the total strength of the Dirichlet function $DIR(p \mid \alpha)$ which is parametrized by $\alpha \in \mathbb{R}$ and $S = \sum_{k=1}^{k} \alpha_k$, $e_i = g(f(x_i; \theta))$ with f as the output of the NN parametrized by θ and g as the evidence function to keep e_k non-negative. Based on [14] $\alpha_k = e_k + 1$. For the inference step, the probabilities for each class k are then given by $p_k = \frac{\alpha_k}{S}$. The uncertainty for each sample is $u = \frac{K}{S}$. For more information [1, 14] and the code provided in Appendix A.

Obtaining an overall uncertainty instead of per-class values was a significant limitation. To address this, we introduced in this paper what we call *Binary Evidential Learning*: a method that trains k binary classifiers, with calculations resembling standard evidential learning, but for $k = 2$ classes. For example, if classifying digit 1, one class will be "yes" if the image contains one and "no" for any other digit.

Selective Classification enhances accuracy in scenarios demanding robustness by allowing experts to review uncertain predictions. We propose a selective classification algorithm that delegates predictions to experts when uncertainty exceeds a predetermined threshold. Incorrect predictions corrected post-review are labeled correct, while unnecessary deferrals are deemed incorrect. This strategy discourages excessive deferrals, promoting a balance between human expertise and automated efficiency.

2.3 Models and DataSets

In our experimentation, we employed the MNIST dataset [9] with Convolutional Neural Networks (CNN) and the ImageNet1K Dataset with Visual Transformers (ViT) [3] under a Projected Gradient Descent (PGD) adversarial attack. Model specifics are elaborated on in Appendix A (code files).

3 RESULTS

Our research emphasizes the efficiency and accuracy of various uncertainty quantification methods implemented across diverse AI platforms. We analyzed their performance on Polaris, Cerebras CS-2, and SambaNova DataScale. Polaris is a traditional supercomputer using AMD EPYC processors and NVIDIA A100 GPUs, whereas Cerebras CS-2 and SambaNova Datascale are AI accelerators featuring wafer-scale architecture and reconfigurable dataflow, respectively. We recorded the accuracy enhancements these methods furnished.

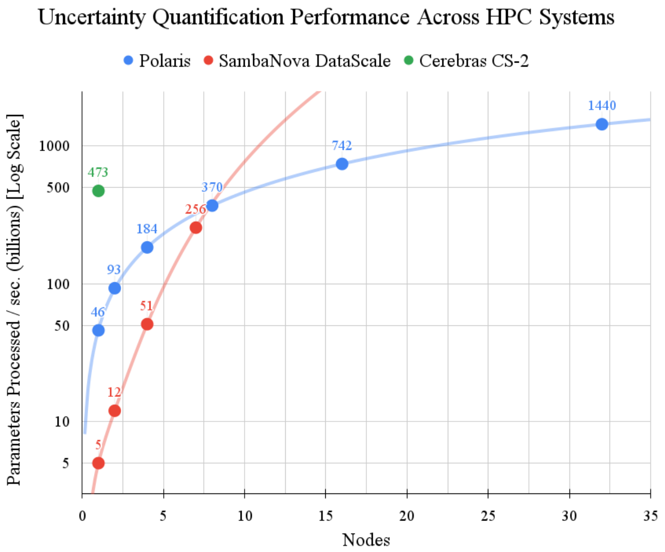

Figure 2: Uncertainty Quantification Performance Across HPC Systems. Parameters processed per second (PPS) at different node counts for Polaris, SambaNova DataScale, and Cerebras CS-2. Polaris scales linearly, SambaNova delivers rapid initial growth, while Cerebras achieves remarkable PPS on a single node.

3.1 Platform-Specific Performance

For specifics regarding the platforms we employed, readers are invited to look into the works on Polaris [8], Cerebras [10], and SambaNova [4]. Figure 2 compares each system's performance when processing one epoch of ImageNet-1K's training dataset, with uncertainty calculations included. Ensemble learning shines with its parallel scaling capabilities on distributed systems like Polaris, as seen in Table 1. Polaris has 4 GPUs per node; we have to make sure we run the script for the GPU affinity that will assign to each MPI rank node a GPU, and we use several different nodes and GPU combinations to show we can scale. However, its scalability is tested on platforms like Cerebras and SambaNova, which are optimized for extensive models rather than managing multiple processes simultaneously. Conversely, since it only trains once, evidential learning runs on average ten times faster than ensembles. However, it can only use one node and does not provide a robust uncertainty outcome.

N.Ensembles	N.Nodes	N.GPUs	Time(sec)
4	1	4	3min17sec
4	2	8	2min27sec
8	1	4	4min35sec
8	2	8	3min14sec
16	1	4	5min58sec
16	2	8	4min55sec
16	4	16	3min25sec

Table 1: MNIST Dataset Timings on Polaris Using Ensembles

3.2 Accuracy

Each uncertainty methodology – ensembles, evidential learning, and binary evidential – has advantages and limitations. Ensembles, benefiting from the insights of multiple models, are reliable in uncertainty metrics but are computationally intensive. Evidential learning stands out for its efficiency but often correlates probability directly with uncertainty, lacking class-specific details. Binary evidential rectifies this by offering class-specific uncertainty but increases training overhead, especially as class numbers grow. When we integrated these uncertainty estimates into our selective classification framework, we observed a spectrum of accuracy levels juxtaposed against the base models. This variation emerges primarily because of the deferment strategy applied to uncertain predictions. Our method accentuates the correlation between the level of uncertainty and the accuracy of the prediction. Our selective classification often mirrored or bettered the benchmark accuracy, as represented in Figure 3. It is worth noting that evidential learning, despite its occasional lower accuracy metric, has the potential to match or even surpass ensemble methods in specific contexts. The improvements in accuracy by binary evidential on MNIST and ensembles on ImageNet were noteworthy, standing at 12.78% and 2.73%, respectively.

4 CONCLUSION

In this work, we delved into the characterization of uncertainty in DL, emphasizing the two predominant uncertainty evaluations (ensembles and evidential learning). Our research demonstrates the potential acceleration benefits of employing HPC and distributed systems for uncertainty evaluations. These evaluations, as shown, can be essential for DL models, offering avenues for distribution if computational constraints arise. For scenarios with limited HPC

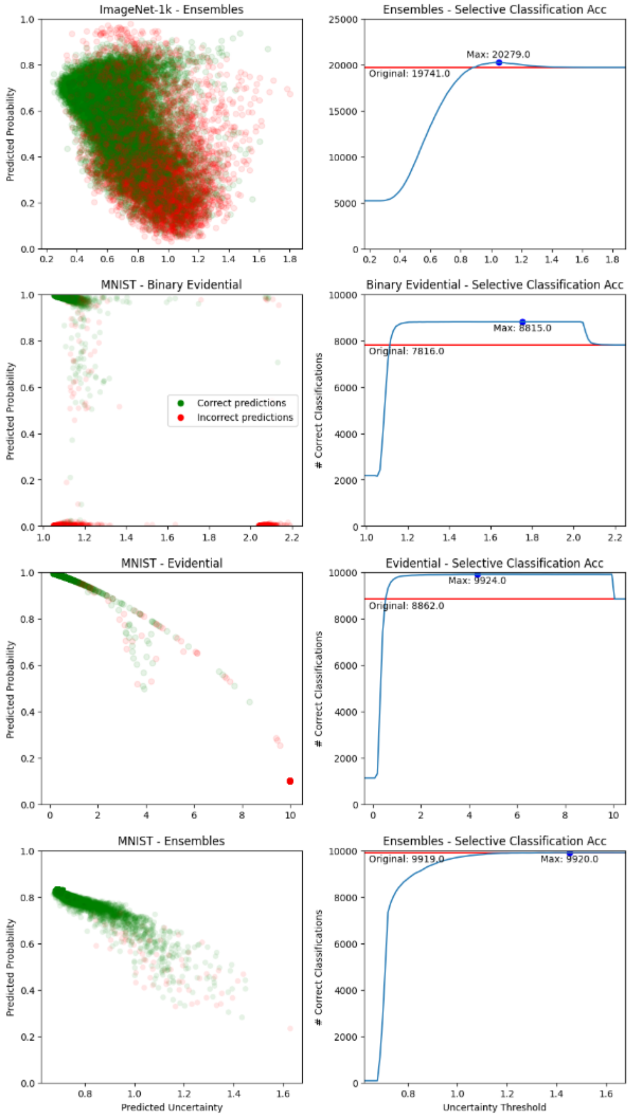

Figure 3: Model uncertainty/probability distributions and their respective selective classification accuracies. First column: Comparison of Binary-Evidential, Evidential, and Ensemble probability/uncertainty distributions. Correct predictions are green, and incorrect are red. Second Column: Selective Classification accuracies over a range of uncertainty thresholds (blue) vs. the original model accuracy (red)

resources, evidential learning emerges as a viable alternative, granting uncertainty insights with minimal overhead. We demonstrated how valuable uncertainty evaluations can be in the particular case where a ViT-based classifier misclassifies under adversarial conditions, but when using uncertainty evaluation, the high uncertainty obtained on the perturbed images indicates potential misclassification, allowing for intervention or rejection of the result (Figure 1).

5 FUTURE WORK

To better exploit the capabilities of new AI Architectures, we would like to explore the transformation of ensembles into network fusion and multi-input multi-output to improve ensemble performance and accuracy. We would also like to extend the uncertainty study to partial differential equation acceleration. Finally, we would like to explore how to evaluate uncertainty when we cannot access the model.

ACKNOWLEDGMENTS

Part of this research was supported by Sustainable Horizons Institute which is part of the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U.S. Department of Energy Office of Science and the National Nuclear Security Administration and by Argonne National Laboratory. This research used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357.

Appendix A AVAILABILITY OF DATA

Access our experimental code at :
https://github.com/epautsch/UncertaintyANL.

REFERENCES

[1] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. 2019. Deep Evidential Regression. *CoRR* abs/1910.02600 (2019), x–x. arXiv:1910.02600 http://arxiv.org/abs/1910.02600

[2] Mariusz Bojarski, Philip Yeres, Anna Choromanska, Krzysztof Choromanski, Bernhard Firner, Lawrence Jackel, and Urs Muller. 2017. Explaining How a Deep Neural Network Trained with End-to-End Learning Steers a Car. arXiv:1704.07911 [cs.CV]

[3] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. 2019. On the Relationship between Self-Attention and Convolutional Layers. *CoRR* abs/1911.03584 (2019), x–x. arXiv:1911.03584 http://arxiv.org/abs/1911.03584

[4] M. Emani, V. Vishwanath, C. Adams, M. E. Papka, R. Stevens, L. Florescu, S. Jairath, W. Liu, T. Nama, and A. Sujeeth. 2021. Accelerating Scientific Applications With SambaNova Reconfigurable Dataflow Architecture. *Computing in Science and; Engineering* 23, 02 (mar 2021), 114–119. https://doi.org/10.1109/MCSE.2021.3057203

[5] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. arXiv:1506.02142 [stat.ML]

[6] Yonatan Geifman and Ran El-Yaniv. 2017. Selective Classification for Deep Neural Networks. arXiv:1705.08500 [cs.LG]

[7] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. *CoRR* abs/1706.04599 (2017), x. arXiv:1706.04599 http://arxiv.org/abs/1706.04599

[8] Argonne National Lab. 2023. *Polaris*. Argonne National Lab. Retrieved July 22, 2023 from https://www.alcf.anl.gov/polaris

[9] Yann LeCun et al. 1995. Learning algorithms for classification: A comparison on handwritten digit recognition. *Neural networks: the statistical mechanics perspective* 261, 276 (1995), 2.

[10] Sean Lie. 2022. Cerebras Architecture Deep Dive: First Look Inside the HW/SW Co-Design for Deep Learning : Cerebras Systems. In *2022 IEEE Hot Chips 34 Symposium (HCS)*. Cerebras, Sunnyvale,CA,USA, 1–34. https://doi.org/10.1109/HCS55958.2022.9895479

[11] Yang Liu and Hongyi Guo. 2020. Peer Loss Functions: Learning from Noisy Labels without Knowing Noise Rates. arXiv:1910.03231 [cs.LG]

[12] Samual MacDonald et al. 2023. Generalising uncertainty improves accuracy and safety of deep learning analytics applied to oncology. nature:10.1038/s41598-023-31126-5

[13] Allen Z. Ren et al. 2023. Robots That Ask For Help: Uncertainty Alignment for Large Language Model Planners. arXiv:2307.01928 [cs.RO]

[14] Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential Deep Learning to Quantify Classification Uncertainty. arXiv:1806.01768 [cs.LG]

[15] Florian Wenzel, Jasper Snoek, Dustin Tran, and Rodolphe Jenatton. 2020. Hyperparameter Ensembles for Robustness and Uncertainty Quantification. arXiv:arXiv:2006.13570