11-2023

Optimizing Uncertainty Quantification of Vision Transformers in Deep Learning on Novel AI Architectures

Erik Pautsch
Loyola University Chicago, epautsch@luc.edu

John Li
University of California - San Diego

Silvio Rizzi
Argonne National Laboratory

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

Maria Pantoja
California Polytechnic State University, San Luis Obispo

Optimizing Uncertainty Quantification of Vision Transformers in Deep Learning on Novel AI Architectures

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 (DL) methods have shown substantial efficacy in computer vision (CV) and natural language processing (NLP). Despite their proficiency, the inconsistency in input data distributions can compromise prediction reliability. This study mitigates this issue by introducing uncertainty evaluations in DL models, thereby enhancing dependability through a distribution of predictions. Our focus lies on the Vision Transformer (ViT), a DL model that harmonizes both local and global behavior. We conduct extensive experiments on the ImageNet-1K dataset, a vast resource with over a million images across 1,000 categories. ViTs, while competitive, are vulnerable to adversarial attacks, making uncertainty estimation crucial for robust predictions.

Our research advances the field by integrating uncertainty evaluations into ViTs, comparing two significant uncertainty estimation methodologies, and expediting uncertainty computations on high-performance computing (HPC) architectures, such as the Cerebras CS-2, SambaNova DataScale, and the Polaris supercomputer, utilizing the MPI4PY package for efficient distributed training.

KEYWORDS

Uncertainty, Deep Learning, Ensembles, Evidential Learning

ACM Reference Format:
Erik Pautsch, John Li, Silvio Rizzi, George K. Thiruvathukal, and Maria Pantoja. 2023. Optimizing Uncertainty Quantification of Vision Transformers in Deep Learning on Novel AI Architectures. In *Proceedings of SuperComputer Conference (SC '23).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 PREVIOUS WORK

Numerous techniques have been developed to quantify uncertainties in DL. For instance, Temperature Scaling [5] 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) [3], which transform neuron weights into probability distributions, can sometimes yield subpar results, especially when employing Monte Carlo dropout [4]. In this paper, we focus 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. In [8] researchers present Peer Loss Functions, in this approach enables models to learn even from noisy labels.

2 METHODS

Among various uncertainty estimation techniques, we identified ensembles [9] and evidential learning [1] as the most promising.

2.1 Ensembles

Ensembles aggregate outputs from several independently trained models on the same data. These intermediate models, optimized using Hyper-parameter Optimization (HPO), alleviate the computational burden of uncertainty evaluation. Given a classification task with input x and label y, the ensemble employs the predictive distribution $p_\theta(y \mid x)$, with θ denoting the neural network's (NN) parameters. Predictions are then aggregated as $p(y \mid x) = M^{-1} \sum_{m=1}^{M} p_{\theta_m}(y \mid x, \theta_m)$.

2.2 Evidential Learning

In **Evidential learning** the softmax layer is substituted with an activation layer, generating an evidence vector that informs the predicted Dirichlet distribution. Given the evidence vector as $f(x_i \mid \theta)$, the Dirichlet distribution's parameters are $\alpha_i = f(x_i \mid \theta) + 1$, which approximate the class probabilities.

2.3 Binary Evidential Learning

Additionally, our **Binary evidential learning** method employs n binary classifiers for individual class probability and uncertainty estimation. We favored ensembles for their robust and comprehensive uncertainty assessment, whereas evidential learning stands out for its efficiency and bypassing of redundant training/inference cycles.

Our experiments used the ImageNet-1K dataset and a Projected Gradient Descent (PGD) adversarial attack. ImageNet-1K offers

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

diverse data for extensive uncertainty evaluation, while PGD, a common adversarial attack in DL, perturbs input data to mislead the model's predictions, testing the uncertainty methods' resilience under hostile conditions.

Our goal is to integrate these uncertainty estimation methods with the ViT, expose it to PGD attacks, and thoroughly evaluate the model's uncertainty estimation ability, interpretability, robustness, and real-world application utility.

3 RESULTS

Our comprehensive experiments on the MNIST and ImageNet-1K datasets with Convolutional Neural Networks (CNN) and ViTs, evaluated on three different system architectures (Polaris supercomputer [6], SambaNova DataScale [2], and Cerebras CS-2[7]), have yielded significant insights into the quantification of uncertainty in DL models.

Ensembles and evidential learning demonstrated distinct advantages and limitations. Ensembles, leveraging multiple independently trained models, provided robust insights into uncertainty but required up to n times more resources with an ensemble of size n when compared to evidential learning. Evidential learning, despite being computationally efficient and requiring fewer resources, demonstrated a lack of robustness, leading to significantly decreased robustness compared to ensembles. We theorize that this is due to evidential learning's high correlation between probability and uncertainty. Further, we applied adversarial perturbations, using a PGD attack, to the ImageNet-1K data, revealing that such attacks significantly influence the uncertainty in DL models. We found an approximate 30% increase in uncertainty and a 31.8% decrease in accuracy. Despite a significant increase in the uncertainty of the class activation maps (CAM) compared to benign images, this increase in uncertainty was not clearly reflected in the probability/uncertainty distribution. Contrary to expectations, this distribution displayed higher accuracy for images with high uncertainty, and lower accuracy for lower uncertainty.

The testing of the system architectures revealed important distinctions. Established distributed supercomputers like Polaris provided scalable and well-understood paradigms for GPU use in training and inference, making them reliable choices for uncertainty quantification in DL. In contrast, emerging systems like SambaNova DataScale and Cerebras CS-2 presented certain challenges, including software bugs and limited parallelization capabilities. Despite these challenges, they promise potential efficiency gains in scaling uncertainty computations.

Our results underline the practical value of evaluating uncertainty in DL models. They highlight the trade-off between the detailed insights provided by ensemble methods and the computational efficiency of evidential learning, while also addressing the opportunities and challenges presented by different HPC architectures.

4 CONCLUSION

This study gauges the robustness of ensemble and evidential uncertainty evaluation methods on ViTs with adversarial attacks. Ensembles delivered broad uncertainty insights, unaffected by class probabilities. We found evidential learning was fast yet less specific with a clear uncertainty-probability correlation, while binary evidential was infeasible for datasets with many classes like ImageNet-1K. We also highlighted the vulnerability of ViTs: adversarial images had higher uncertainty but misclassified ones showed lower values than correctly labeled counterparts. Distinct CAM uncertainty patterns emerged between benign and adversarial images.

5 FUTURE WORK

Future work may explore Multi-Input Multi-Output (MIMO) ensembles, enhanced classification via aggregated CAMs, and uncertainty in AI-powered differential solvers. The value of uncertainty analysis in DL is underscored. Access our experimental code at https://github.com/epautsch/UncertaintyANL.

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.

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] 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

[3] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of The 33rd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 48)*, Maria Florina Balcan and Kilian Q. Weinberger (Eds.). PMLR, New York, New York, USA, 1050–1059. https://proceedings.mlr.press/v48/gal16.html

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

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

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

[7] 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

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

[9] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc., USA. https://proceedings.neurips.cc/paper_files/paper/2019/file/8558cb408c1d76621371888657d2eb1d-Paper.pdf