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

SCAN for Code

Erik Pautsch[1], John Li[2], Silvio Rizzi[3], George K. Thiruvathukal[1], Maria Pantoja[4]

[1]Loyola University Chicago, [2]University of California San Diego, [3]Argonne National Laboratory, [4]California Polytechnic State University

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.

What is Uncertainty?

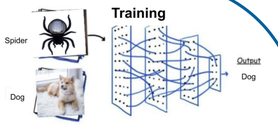
- Uncertainty in Deep Learning (DL) helps determine the reliability of a DL model's output

Uncertainty Due to Noise (Aleatoric)

- Fog in a traffic monitoring system introduces unavoidable noise, no matter the sophistication of the model

Uncertainty Due to a Lack of Data (Epistemic)

- That same system may misclassify a rare type of vehicle it has not been trained on, illustrating uncertainty due to limited data



Inferencing on an image of a cow, a classifier trained on only spiders and dogs is likely to assign a higher softmax probability to the 'dog' class due to the greater similarity between the two mammals.

A graffitied stop sign presents misclassification and detection risks for DL models that make knowing the uncertainty in self-driving systems essential.

- Quantifying uncertainty is crucial for reliability in a variety of applications, including:
  - medical settings
  - self driving cars
  - robotic automation
- Detecting high levels of uncertainty gives systems the option to take action based on pre-set heuristics, or trigger alerts for human intervention

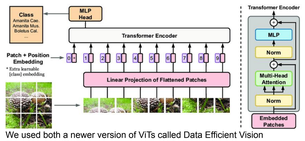
Neural Networks and Datasets Used

Datasets

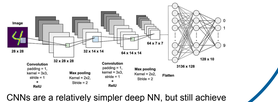
- ImageNet-1K: 1.2 million images labeled with 1,000 object categories
- MNIST: 70k images of handwritten digits

Neural Networks (NN)

- Vision Transformer (ViT)
- Convolutional Neural Network (CNN)
- Multiple versions of each network were examined with parameters ranging from 100k to 86M



We used both a newer version of ViTs called Data Efficient Vision Transformers (DeiT), as well as regular ViTs (seen above).



CNNs are a relatively simpler deep NN, but still achieve accuracies of >98% on the MNIST dataset.

Determining Uncertainty

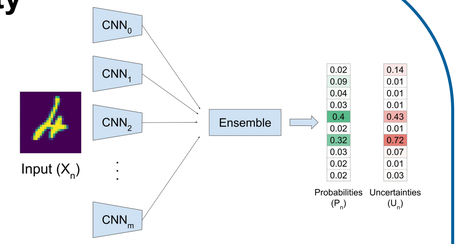
- We employed two primary methods of determining uncertainty on CNNs and ViTs

Hyper-Deep Ensembles

- Combines multiple DL models with different hyperparameters
- Diversity of the ensembles captures different aspects of the data's structure, thereby improving overall uncertainty estimation
- Enhances prediction accuracy, as the ensemble can capture more complexities and variabilities in the data than a single model

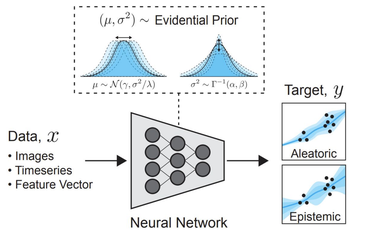
Evidential

- Incorporates uncertainty directly into the learning process, providing speed and efficiency in uncertainty estimation
- Modifies the loss function to incorporate uncertainty
- **Binary Evidential** implements one evidential model per class, improving depth at the cost of computational complexity



Individual Models

Aggregating an ensemble's logits into a single probability distribution is depicted, where the standard deviation of this distribution quantifies the level of uncertainty in the model's predictions.



An input 'x' is introduced into the NN. The network uses an evidential prior to form a posterior distribution. The target prediction is derived from this distribution. Uncertainty is quantified as the number of classes divided by the sum of the evidence.

System Architectures

The Polaris supercomputer at Argonne National Laboratory.

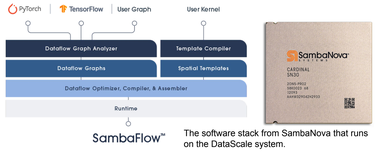
- Three distinct High-Performance Computing (HPC) architectures were used

Polaris - Based on NVIDIA A100

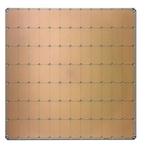
- Advanced hardware architecture provides massive computational power, enabling efficient training and inferencing of complex DL models
- Leveraged MPI4PY, a Python interface to MPI, to enable distributed uncertainty calculations, significantly enhancing the efficiency and scalability

SambaNova DataScale

- 64 Cardinal SN30 Reconfigurable Data Units (RDUs) with 64 TB of total memory
- Changes to the model's parameters or operations require recompilation, which can have a significant time cost
- Capable of training and inferencing in a distributed, parallel fashion using MPI



The software stack from SambaNova that runs on the DataScale system.



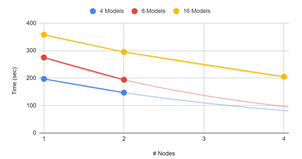
The Cerebras 2nd generation Wafer-Scale Engine (WSE-2) powers the CS-2 system's computations.

Cerebras CS-2

- 850,000 processing cores with 40 GB of on-chip memory
- Memory and fabric bandwidth of 20 PB/sec and 200 PB/sec, respectively
- Parallel and distributed model usage is not currently officially supported
  - We expect to be able to implement this ourselves in future work with the CS-2

Results

- Our experiments provided us with insightful results on a variety of fronts
- We focus on the quantitative and qualitative performance of both the scaling of uncertainty calculations on the HPC systems and our two methods of uncertainty quantification



As the number of nodes increases, computation time decreases, demonstrating the scalability of our approach. Different colors represent experiments with varying ensemble sizes, with larger ensembles requiring more time for the same number of nodes.

HPC Systems - Performance and Comparison

- Established distributed supercomputers (Polaris) provide scalable solutions and well-understood paradigms for GPU utilization in training and inference
- The SambaNova and Cerebras systems hold the promise of significant efficiency gains, yet also present challenges, including software bugs and restricted parallelization capabilities

	Polaris	SambaNova	Cerebras
MNIST	00:01.04	00:00.69	00:00.09
ImageNet-1K	36:40.00	424:07.00	N/A

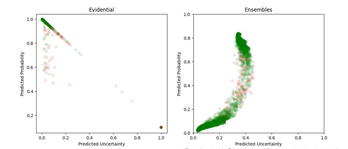
Comparison of average epoch completion times (in MM:SS.ss format) across different HPC architectures and datasets. Note: Cerebras CS-2 currently lacks support for VIsion Transformers, indicated by 'N/A'.

	Polaris	SambaNova	Cerebras
Sample/Sec. (thousands)	67.3	101.4	777.8

Comparison of MNIST epoch processing rates (in thousands of samples per second) for single models across each HPC architecture.

Ensembles vs Evidential

- Ensembles provide a comprehensive view at the cost of higher resource usage, with resource requirements scaling with the number of ensemble models
- Evidential learning, while less robust, offers a more computationally efficient approach, requiring fewer resources and enabling faster computations

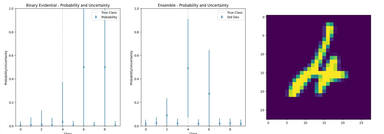


Higher concentration of correct predictions at low uncertainty with significant overlap of incorrect predictions, indicating less consistent uncertainty estimation.

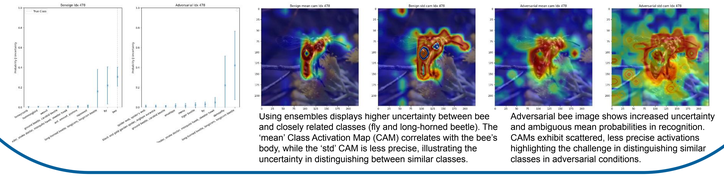
'Backward C' shape illustrates nuanced uncertainty representation, with higher certainty at low and high predicted probabilities. The gap with mixed predictions shows the model's ability to recognize when it is unsure.

CNNs - Integration of ensembles and evidential learning was efficient, facilitated by the relatively simpler architecture of CNNs

ViTs - Ensemble implementation was accomplished within ViTs; however, evidential learning adaptation remains an ongoing exploration



Interpretation of a challenging MNIST '4' sample: Binary Evidential and Ensemble methods provide different uncertainty and probability profiles for each class, highlighting their varied approaches to dealing with ambiguous inputs.



Using ensembles displays higher uncertainty between bee and closely related classes (fly and long-horned beetle). The 'mean' Class Activation Map (CAM) correlates with the bee's body, while the 'std' CAM is less precise, illustrating the uncertainty in distinguishing between similar classes.

Adversarial bee image shows increased uncertainty and ambiguous mean probabilities in recognition. CAMs exhibit scattered, less precise activations highlighting the challenge in distinguishing similar classes in adversarial conditions.

Conclusion

- Quantifying uncertainty in DL models offers key insights for certainty-critical applications, revealing how models handle ambiguous and adversarial samples
- Ensemble methods provide a comprehensive view of uncertainty but require more resources, while evidential learning is more computationally efficient yet less robust
- Established supercomputers provide scalability, while emerging systems like SambaNova and Cerebras present potential for efficiency gains but also challenges

Future Work

- Further investigation of methods to enhance the effectiveness of evidential learning
- Devising custom solutions in order to leverage the full potential of novel HPC architectures
- Analyzing and mitigating the impact of adversarial attacks on the uncertainty quantification process

U.S. DEPARTMENT OF ENERGY Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.

Part of this research was supported by Sustainable Horizons Institute which is part of the Exascale Computing Project 7-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.

Argonne NATIONAL LABORATORY | Argonne Leadership Computing Facility