Uncertainty Quantification for Remotely-Sensed Datasets

Abstract

Deep learning (DL) models are extensively used to analyze and monitor the Earth's surface due to their scalability on large-scale EO datasets and their computational efficiency compared to conventional statistical approaches such as Bayesian analysis. However, they are not capable of explaining their predictions; namely, their outputs, given large-scale datasets as input, are not trustworthy, reliable, and robust which can be measured using uncertainty quantification. In fact, DL models are often considered as uninterpretable black boxes with unknown uncertainties. In contrast, Bayesian analysis is a gold standard technique for uncertainty quantification in order to obtain trustworthy and reliable predictions generated by models fitted small- or moderate-scale datasets (observations) due to its high computational cost. Hence, DL models integrated with Bayesian analysis, that is, Bayesian Neural Networks (BNNs), are slowly gaining great interest, since they allow make their outputs interpretable together with trustworthy and reliable uncertainties. However, BNN inference on large-scale datasets persists high computational cost even on the HPC system, and commonly used methodologies to overcome this challenge are Monte Carlo Markov Chain (MCMC) and variational inference (VI) approaches. Moreover, the VI approach, returning approximate samples, can be scaled on big datasets in contrast to the exact sampling MCMC. Therefore, this study assesses and examines quantum VI paradigm for processing BNN inference on small-scale EO datasets (in our case, hyperspectral images (HSIs)) to improve the sampling power of a conventional VI method. More importantly, we estimate quantum resource required for some example quantum VI models in terms of T-gates but not the implementation of quantum VI models on small-scale HSIs.

14 Introduction

Deep Learning (DL) models are employed for distinct tasks such as recognizing informative patterns in large-scale, high-dimensional datasets (in our case, satellite images or remotely-sensed datasets) and discovering their underlying distributions; here, large-scale, high-dimensional datasets can be denoted by either $\mathcal{S} = \{y_i, \mathbf{x}_i\}_{i=1}^N$ or $\mathcal{S} = \{\mathbf{x}_i\}_{i=1}^N$ depending on whether or not a learning task is to discover underlying distributions generating datasets, where y_i is their true labels, $\mathbf{x}_i$ is their high-dimensional elements, and N refers to their large-scale size. In general, these different tasks can be divided into two categories so-called supervised and unsupervised learning paradigm Murphy [2012]; Goodfellow et al. [2016]. Supervised learning paradigm refers to DL tasks for recognizing informative patterns in big datasets $\mathcal{S} = \{y_i, \mathbf{x}_i\}_{i=1}^N$ in order to predict labels $\hat{y}_i$ with the highest probability $p(\hat{y}_i|\hat{\mathbf{x}}_i, \boldsymbol{\theta})$ given $\hat{\mathbf{x}}_i$ such that the loss function $\mathcal{L}_{\boldsymbol{\theta}}(y, \hat{y})$ between true and predicted labels optimized over trainable parameters $\boldsymbol{\theta}$ is at its minimum value, while unsupervised learning paradigm is to approximately obtain the underlying distribution $p(\mathbf{x})$ of $\mathcal{S} = \{\mathbf{x}_i\}_{i=1}^N$, where labels y_i are not provided as input, such that the approximated distribution $q(\mathbf{x}|\boldsymbol{\theta})$ is closer to the true distribution $p(\mathbf{x})$ measured by optimizing some metrics over trainable parameters $\boldsymbol{\theta}$ such as Kullback-Leibler divergence (KL-divergence). Moreover, DL models are widely used to find solutions to real-world, data-driven and model-driven problems in industry and science $-$ even in Earth observation (EO) Cheng et al. [2020] $-$ due to their scalability on large-scale datasets and their computational efficiency on powerful computing resources (i.e., GPU tensor cores) compared with conventional statistical methods such as intractable Bayesian analysis Pandey et al. [2022]; Willard et al. [2020]. However, DL models are often perceived as uninterpretable black-boxes generating untrustworthy and unreliable predictions, while intractable Bayesian analysis outputs predictions with trustworthy and reliable uncertainties Gal et al. [2022]. Hence, DL models with the help of intractable Bayesian analysis are gaining great attention for designing novel learning models, that is, Bayesian Neural Networks (BNNs). In fact, BNNs combining the best of both DNN and Bayesian analysis can be scaled on large-scale datasets and computed cheaply on GPU tensor cores if an efficient sampling technique for them is available while generating predictions with trustworthy and reliable errors/uncertainty estimates at the same time Jospin et al. [2022]; Olivier et al. [2021].

In contrast to DNNs, BNNs still persist high computational cost for computing a posterior distribution $p(\theta|\mathcal{S})$, directly proportional to the product of a likelihood $p(\mathcal{S}|\theta)$ and a prior $p(\theta)$, and inversely proportional to an evidence $p(\mathcal{S})$, which is a probability distribution function integrating out parameter space Olivier et al. [2021]; Zhang et al. [2017]. Moreover, the evidence is an intractable function due to its high dimensional parameter space, and so the posterior (i.e., BNNs). To weaken this intractable BNN problem, the authors of the articles MacKay [1992]; Neal [1995]; Blei et al. [2017] proposed a machinery so-called variational inference (VI) which approximates the posterior by a tractable parametrized distribution. Another method for tackling an intractable BNN is a Monte Carlo Markov Chain (MCMC) technique Brooks et al. [2011]; Hoffman and Gelman [2011] which, however, does not scale on big datasets as a VI method does. In fact, both MCMC and VI are indispensable tools for generating samples efficiently from intractable posteriors and for quantifying parameter uncertainty for safety-critical and human-centered EO tasks regardless of their respective imperfection, i.e., MCMC does not scale well on large scale datasets but generates exact samples from a posterior, while VI scales on large scale datasets but generates approximate samples from a posterior.

The emergence of quantum algorithms for accelerating some conventional algorithms attracts engineers and scientists alike who persistently attempt to find solutions to intractable problems efficiently by inventing and designing classical algorithms An et al. [2021]; Harrow et al. [2009]. Moreover, there exists the quantum versions of MCMC and VI methods (for short, quantum MCMC and VI) which promise theoretical quantum advantage for some computational problems over their classical counterparts due to the inherent probabilistic nature of quantum machines, that is, a quantum annealer, a quantum simulator, or a universal quantum computer Montanaro [2015]; Layden et al. [2022]; Benedetti et al. [2021]. However, no quantum advantage is demonstrated for finding solutions to practically relevant problems since currently existing quantum machines, that is, noisy intermediate-scale quantum (NISQ) computers, comprise a limited number of error-prone quantum bits (qubits) ≤ 100 and quantum gates Preskill [2018a], while there is a theoretical guarantee to build fault-tolerant quantum (FTQ) computers having error-free qubits > 100 and quantum gates for demonstrating quantum advantage for real-world problems Preskill [1997].

Therefore, in this use-case study, we survey and examine theoretically a quantum/classical VI method due to its scalability on large-scale datasets as opposed to a MCMC technique from the perspective of computational complexity theoretic conjectures. More importantly, a quantum/classical VI method returns solutions to BNNs with trustworthy and reliable uncertainty estimates more efficient than their classical counterparts, while we apply BNNs to safety-critical and human-centered EO tasks specifically involving small-scale real-world datasets, i.e., EO Use-Cases using small-scale hyperspectral image datasets DLR. In addition, we provide the quantum resources required for computing BNNs on proof-of-concept-small (e.g., small-scale) and operational-size-big (e.g., large-scale) EO datasets as we critically stick to the scalability and development roadmap of quantum machines provided by industry and academia.

15 Problem Definition: Earth Observation Use-Case

Sensors on Earth observation satellites detect spectral signals reflected on natural and human-made objects on Earth's surface, and huge amounts of spectral signals in distinct wavelength ranges (Terabytes of data per day) are archived in data storage devices day and night ESA. A hyperspectral imaging satellite, e.g., an EnMAP satellite DLR, is imaging sensors mounted on a satellite for sensing spectral wavelengths in ranges of 420 nm to 1000 nm (VNIR) and from 900 nm to 2450 nm (SWIR). Its mission is to collect hyperspectral imaging data in order to provide vital information for scientific inquiries, societal grand challenges, and key stakeholders and decision makers relating to DLR

- climate change impact and interventions,
- hazard and risk assessment,
- biodiversity and ecosystem processes, and
- land cover changes and surface processes.

DNNs for data-driven tasks require big labeled datasets (i.e., a data-driven approach), while hyperspectral images (HSIs) for e.g., hazard and risk assessment, are images with limited label information, namely, the limited availability of training (benchmark) HSIs, compared to conventional benchmark remote-sensing datasets like multispectral images Cheng et al. [2017]; Paoletti et al. [2019]. There is also the commonly known limitation that DNNs do not yield their confidence level for making high stake decisions for hazard and risk assessment. Hence, the persisting challenge is to invent and design inherently interpretable data-driven models for HSIs together with error/uncertainty estimates due to both uninterpretable black-box DNN models and (almost) lack of benchmark labeled-HSI datasets, since the answers to the above-questions are already utilized to make high stake decisions − safety-critical and human-centered EO decisions Rudin [2018].

BNNs combining both DNN and Bayesian model are widely believed to be inherently interpretable data-driven models for both small- and large-scale datasets only if the efficient sampling algorithms from them are available due to their infeasible evidence Jospin et al. [2022]; Olivier et al. [2021]. More importantly, BNNs are data-efficient models which can be trained on limited label datasets, since they provide uncertainty information in their predictions and weights. The authors of the articles Alcolea and Resano [2022]; Joshaghani et al. [2022] utilized and assessed BNNs for limited benchmark labeled-HSI datasets to generate predictions with trustworthy and reliable uncertainties as they used classical MCMC and VI techniques to generate samples from the posterior of BNNs. Both MCMC and VI sampling tools are far from perfect, and their imperfection is inspected and analyzed numerically on limited labeled-HSIs, e.g., their scalability and precision, by the authors of the article Ries et al. [2022]. Regardless of their imperfection, these sampling methods are base algorithms to invent and benchmark novel sampling methods like Generative Quantum Machine Learning Zoufal [2021], Evidential Deep Learning Sensoy et al. [2018], and Dempster–Shafer Theory of Evidence Deng [2015].

Furthermore, the authors of the articles Layden et al. [2022]; Montanaro [2015]; An et al. [2021] conducted a research on quantum MCMC and quantum-enhanced MCMC methods through the lens of theoretical computational complexity in order to speed-up the conventional MCMC algorithm, while the authors of the article Benedetti et al. [2021] focused on a quantum VI method to show quantum advantage over its classical counterpart because of classically hard-to-simulate quantum circuits such as Instantaneous Quantum Polynomial (IQP) circuits Bremner et al. [2010] and Quantum Approximate Optimization Algorithm (QAOA) sampling Farhi and Harrow [2016]. IQP circuits are quantum circuits equivalent to a so-called partition function, not efficiently simulable on conventional computers. More importantly, the impact of these quantum algorithms will be enormous for processing BNNs on limited benchmark labeled-HSI datasets for making high stake decisions − safety-critical and human-centered EO decisions when we have an access to reasonable noisy intermediate-scale and fault-tolerant quantum computers (QCs) integrated with supercomputers, high performance computing (HPC): That is, HPC+QCs for computational problems of practical significance.

This Earth Observation Use-Case (EO UC) study surveys and examines a quantum VI tool together with a hybrid approach (i.e., HPC+QCs) for the limited HSIs while assessing distinct quantum computers including a quantum annealer, a quantum simulator, or a universal quantum computer by critically sticking to their scalability and development roadmap provided by industry and academia. In addition, we provide our pseudo-algorithms for processing BNNs via the quantum VI technique on high stake problems listed above.

16 Classical Bayesian Neural Networks

Classical Bayesian Neural Networks, for short, Bayesian Neural Networks (BNNs), are referred to as stochastic Deep Neural networks (DNNs) trained using Bayesian analysis on datasets. BNNs generating probability distributions of predictions and parameters (i.e., weights) are natural data-efficient and inherently interpretable models thanks to their respective uncertainties, that is, uncertainties in predictions and weights Jospin et al. [2022]; Koller et al. [2022]. In contrast, conventional DNNs considered as uninterpretable black-box models require big labeled datasets, and even they are needed to be trained and tested on sub-datasets including training, test, and validation sets, while one does not need to divide datasets into training, test, and validation sets for training BNNs. For limited labeled datasets, this division of a dataset raises a challenge for training DNNs but not for BNNs Olivier et al. [2021]. Moreover, DNNs yield also point estimates of predictions with point weights lacking their uncertainty, i.e., lacking explainability due to the uninterpretable black-box paradigm Rudin [2018].

Furthermore, BNNs combine DNNs and Bayesian analysis in order to quantify uncertainties in their predictions and weights, since they better utilize the available dataset, either small or big datasets. Namely, they are DNN models analyzed using Bayesian analysis while their weights and predictions follow certain probability distributions. To design BNNs, we first choose an appropriate DNN model $F_{\boldsymbol{\theta}} = F_{\boldsymbol{\theta}}(\cdot)$ for a given dataset $\mathcal{S} = \{y_i, \mathbf{x}_i\}_{i=1}^N$. Secondly, its weights and predictions are needed to be defined according to some prior $p(\boldsymbol{\theta})$ and likelihood $p(\mathcal{S}|F_{\boldsymbol{\theta}})$ distributions:

$$\begin{aligned} &\boldsymbol{\theta} \sim p(\boldsymbol{\theta}) = \mathcal{N}(0, \sigma^2\mathbf{I}), \\ &p(\mathcal{S}|F_{\boldsymbol{\theta}}) = p(\mathcal{S}_y|\mathcal{S}_x, F_{\boldsymbol{\theta}}) = \mathcal{N}(\mathcal{S}_y; F_{\boldsymbol{\theta}}(\mathcal{S}_x), \sigma^2\mathbf{I}); \end{aligned} \tag{1}$$

where weights $\boldsymbol{\theta}$ are sampled from a normal distribution $\mathcal{N}(0, \sigma^2)$ with zero mean and known uncertainty σ^2. $\mathcal{S}_y$ and $\mathcal{S}_x$ denote labels $\{y_i\}_{i=1}^N$ and input data points $\{\mathbf{x}_i\}_{i=1}^N$ for BNNs, e.g., $F_{\boldsymbol{\theta}}(\mathcal{S}_x)$. We note that one can represent a prior and likelihood by any probability distribution function instead of a normal distribution. For simplicity, we utilized a normal distribution $\mathcal{N}(\cdot)$. To quantify uncertainties in predictions and weights, BNNs utilize the Bayes' theorem:

$$p(F_{\boldsymbol{\theta}}|\mathcal{S}) = \frac{p(\mathcal{S}|F_{\boldsymbol{\theta}})p(\boldsymbol{\theta})}{p(\mathcal{S})} \quad \longleftrightarrow \quad p(\boldsymbol{\theta}|\mathcal{S}) = \frac{p(\mathcal{S}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{S})}, \quad \textit{given} \quad p(\mathcal{S}) = \int_{\Omega_\theta} p(\mathcal{S}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}; \tag{2}$$

here $p(\boldsymbol{\theta}|\mathcal{S})$ is the posterior, and $p(\mathcal{S})$ is the evidence integrating over parameter space Ω_θ. Finally, after computing the posterior distribution expressed by Eq. (2), we can calculate a probability to predict a label $\hat{y}$ given a test data point $\hat{\mathbf{x}}$ and dataset $\mathcal{S}$, that is, a predictive posterior:

$$p(\hat{y}|\hat{\mathbf{x}}, \mathcal{S}) = \int_{\Omega_{\boldsymbol{\theta}}} p(\hat{y}|\hat{\mathbf{x}}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{S})d\boldsymbol{\theta}. \tag{3}$$

In particular, the posterior $p(\boldsymbol{\theta}|\mathcal{S})$ gives uncertainties in weights $-$ this uncertainty is called an epistemic uncertainty, while the predictive likelihood $p(\hat{y}|\hat{\mathbf{x}}, \boldsymbol{\theta})$ yields uncertainties in predictions $-$ this uncertainty is called an aleatoric uncertainty. Therefore, the predictive posterior $p(\hat{y}|\hat{\mathbf{x}}, \mathcal{S})$ generates total uncertainties in predictions by leveraging both epistemic and aleatoric uncertainties Hüllermeier and Waegeman [2021]; Gawlikowski et al. [2022]. By convention, the epistemic uncertainty related to the random noise (randomness) in a dataset can be reduced by increasing the size of a dataset, while the aleatoric uncertainty associated with a lack of knowledge in a model $\boldsymbol{\theta}$ is an irreducible uncertainty even by increasing the size of a dataset.

The parameter space $\Omega_{\boldsymbol{\theta}}$ of modern DNNs includes several thousands to millions of tuneable parameters $\boldsymbol{\theta}$. This high dimensional space of parameters raises a challenge to integrate the evidence $p(\mathcal{S})$ as well as predictive posterior $p(\hat{y}|\hat{\mathbf{x}}, \mathcal{S})$ over Ω_θ; namely, computing the evidence and predictive posterior is an intractable problem Arora and Barak [2009]. Thus, the posterior $p(\boldsymbol{\theta}|\mathcal{S})$ is a hard-to-compute function on conventional computers due to the intractable evidence.

In order to tackle these intractability challenges, studies proposed several different techniques including so-called variational inference (VI) MacKay [1992]; Neal [1995]; Blei et al. [2017] and Monte Carlo Markov Chain (MCMC) Brooks et al. [2011]; Hoffman and Gelman [2011]. VI is a method to approximate the posterior $p(\boldsymbol{\theta}|\mathcal{S})$ by a tractable variational distribution $q_c(\boldsymbol{\theta};\boldsymbol{\lambda})$ via some distance metric over a variational parameter $\boldsymbol{\lambda}$, and to collect samples from this tractable distribution, while MCMC returns (almost) exact samples directly from the posterior $p(\boldsymbol{\theta}|\mathcal{S})$ via like the No-U-Turn Sampler Hoffman and Gelman [2011]. Note that these techniques have their own advantages and imperfections for approximate sampling and scalability on big datasets Ries et al. [2022]. In particular, the VI method returns approximate samples and scales well on big datasets (i.e., computationally cheap), while the MCMC generates almost exact samples and poorly scales on big datasets (i.e., computationally expensive). Thus, it is of great importance to design and assess the quantum VI instead of the quantum MCMC due to its scalability on big datasets in order to make it better on approximate samples.

16.1 Classical Variational Inference

Variational inference (VI) is a machinery to approximate the posterior written in Eq. (2) by some easy-to-sample distribution $q_c(\boldsymbol{\theta};\boldsymbol{\lambda})$. To define a easy-to-sample distribution, we optimize a reverse Kullback-Leibler divergence (KL-divergence) as done for training conventional DNN models Jospin et al. [2022]:

$$\begin{aligned}&\mathrm{argmin}_{\boldsymbol{\lambda}} KL(q_c(\boldsymbol{\theta};\boldsymbol{\lambda})||p(\boldsymbol{\theta}|\mathcal{S}))\\ &\qquad = \sum_{\boldsymbol{\lambda}} q_c(\boldsymbol{\theta};\boldsymbol{\lambda}) \log\left(\frac{q_c(\boldsymbol{\theta};\boldsymbol{\lambda})}{p(\boldsymbol{\theta}|\mathcal{S})}\right) = \mathbb{E}_{q_c(\boldsymbol{\theta};\boldsymbol{\lambda})}\left[\log\left(\frac{q_c(\boldsymbol{\theta};\boldsymbol{\lambda})}{p(\boldsymbol{\theta}|\mathcal{S})}\right)\right],\end{aligned} \tag{4}$$

where $KL(q_c(\boldsymbol{\theta};\boldsymbol{\lambda})||p(\boldsymbol{\theta}|\mathcal{S}))$ is equal to zero or minimized if and only if $q_c(\boldsymbol{\theta};\boldsymbol{\lambda}) \approx p(\boldsymbol{\theta}|\mathcal{S})$. If we expand the above KL-divergence by using the posterior expressed in Eq. (2) and rearrange it then we have:

$$KL(q_c(\boldsymbol{\theta};\boldsymbol{\lambda})||p(\boldsymbol{\theta}|\mathcal{S})) = -\left(\mathbb{E}_{q_c(\boldsymbol{\theta};\boldsymbol{\lambda})}\left[\log(p(\mathcal{S}|\boldsymbol{\theta})p(\boldsymbol{\theta}))\right] - \mathbb{E}_{q_c(\boldsymbol{\theta};\boldsymbol{\lambda})}\left[\log q_c(\boldsymbol{\theta};\boldsymbol{\lambda})\right]\right) + \mathbb{E}_{q_c(\boldsymbol{\theta};\boldsymbol{\lambda})}\left[\log p(\mathcal{S})\right]. \tag{5}$$

We easily notice that the KL-divergence is still a hard-to-optimize function, since it includes a log evidence $\log p(\mathcal{S})$ where the evidence $p(\mathcal{S})$ is an intractable distribution function. To overcome the hardness of computing the KL-divergence, we utilize the fact that $KL(q_c(\boldsymbol{\theta};\boldsymbol{\lambda})||p(\boldsymbol{\theta}|\mathcal{S})) \geq 0$, and so:

$$\mathbb{E}_{q_c(\boldsymbol{\theta};\boldsymbol{\lambda})}\left[\log p(\mathcal{S})\right] \geq \mathbb{E}_{q_c(\boldsymbol{\theta};\boldsymbol{\lambda})}\left[\log(p(\mathcal{S}|\boldsymbol{\theta})p(\boldsymbol{\theta}))\right] - \mathbb{E}_{q_c(\boldsymbol{\theta};\boldsymbol{\lambda})}\left[\log q_c(\boldsymbol{\theta};\boldsymbol{\lambda})\right] \tag{6}$$

where the expression on right hand side is called an evidence lower bound (in short, ELBO):

$$\begin{aligned}ELBO(p(\mathcal{S},\boldsymbol{\theta})||q_c(\boldsymbol{\theta};\boldsymbol{\lambda})) &= \mathbb{E}_{q_c(\boldsymbol{\theta};\boldsymbol{\lambda})}\left[\log(p(\mathcal{S}|\boldsymbol{\theta})p(\boldsymbol{\theta}))\right] - \mathbb{E}_{q_c(\boldsymbol{\theta};\boldsymbol{\lambda})}\left[\log q_c(\boldsymbol{\theta};\boldsymbol{\lambda})\right]\\ &= \mathbb{E}_{q_c(\boldsymbol{\theta};\boldsymbol{\lambda})}\left[\log\frac{p(\mathcal{S},\boldsymbol{\theta})}{q_c(\boldsymbol{\theta};\boldsymbol{\lambda})}\right] = \mathcal{L}(\boldsymbol{\lambda},\boldsymbol{\theta};\mathcal{S}).\end{aligned} \tag{7}$$

More importantly, the ELBO is a tractable metric function compared with the KL-divergence, and the following condition is satisfied:

$$\mathrm{argmax}_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda},\boldsymbol{\theta};\mathcal{S}) = \mathrm{argmin}_{\boldsymbol{\lambda}} KL(q_c(\boldsymbol{\theta};\boldsymbol{\lambda})||p(\boldsymbol{\theta}|\mathcal{S})). \tag{8}$$

For scaling and optimizing the ELBO on small- and large-scale datasets $\mathcal{S} = \{y_i, \mathbf{x}_i\}_{i=1}^{N}$, stochastic variational inference (SVI) is extensively utilized for scaling BNNs on datasets, and it is a machinery to randomly generate M mini-batches and to optimize the ELBO on those batches:

$$\mathcal{L}(\boldsymbol{\lambda}) = \frac{N}{M} \sum_{i=1}^{M} \mathcal{L}(\lambda_i, \boldsymbol{\theta}; \mathcal{X}_i), \quad \text{argmax}_{\boldsymbol{\theta}, \boldsymbol{\lambda}} \, \mathcal{L}(\boldsymbol{\lambda}). \tag{9}$$

We refer to the article Blei et al. [2017] for the interested readers for detailed discussions on the SVI optimizing algorithm. In general, the SVI algorithm returns approximate samples of variational parameters $\{\lambda_i\}_{i=1}^{N}$ and model weights $\boldsymbol{\theta}$.

17 Quantum Bayesian Neural Networks

Quantum Bayesian Neural Networks (QBNNs) are BNNs boosted by quantum algorithms which are designed to solve efficiently some hard computational problems on quantum computers Benedetti et al. [2021]. Moreover, they promise to generate solutions to a class of computational problems much faster than conventional computing resources, and quantum computers (i.e., quantum circuits) are even able to represent classically intractable probability distributions due to their inherently probabilistic nature and non-classical correlation property, that is, quantum circuits with large entanglement Bremner et al. [2010].

There are some proposals to design QBNNs based on quantum DL techniques Allcock et al. [2020] combined with BNNs Berner et al. [2021], and computing the VI approach on a quantum computer Benedetti et al. [2021]. Namely, it is a machinery to approximate the posterior by a tractable distribution $q_c(\boldsymbol{\theta}; \boldsymbol{\lambda})$ by optimizing the ELBO measure expressed in Eq. (7). Indeed, this approximation makes BNNs scalable on large-scale datasets but generates approximate samples $-$ not exact samples. Therefore, we propose to utilize a parametrized quantum circuit (PQC) which represents a family of probability distributions $q_Q(\boldsymbol{\theta}; \boldsymbol{\lambda})$ in order to make the VI technique better on generating good approximate samples. Some PQCs used to generate samples even are known to be not simulable on a conventional computer Bremner et al. [2010]; Farhi and Harrow [2016]. Hence, we approximate the posterior by the quantum variational distribution $q_Q(\boldsymbol{\theta}; \boldsymbol{\lambda})$ due to its representational power over the classical variational distribution $q_c(\boldsymbol{\theta}; \boldsymbol{\lambda})$ Benedetti et al. [2021]. This quantum approximation can be dubbed "quantum BNNs" or "quantum variational inference".

17.1 Quantum Variational Inference

Quantum variational inference (QVI) is referred to as representing a variational distribution $q_Q(\boldsymbol{\theta}; \boldsymbol{\lambda})$ and sampling it on quantum machines:

$$q_Q(\boldsymbol{\theta}; \boldsymbol{\lambda}) = |\langle \boldsymbol{\theta} | \psi(\boldsymbol{\lambda}, \mathbf{x}_i) \rangle|^2, \tag{10}$$

where $\mathbf{x}_i$ and $\boldsymbol{\lambda}$ denote an input data point and trainable parameters, respectively. Namely, given a quantum learning model, $\hat{\mathcal{O}}(\mathbf{x}_i) H^{\otimes n} \hat{\mathcal{O}}(\boldsymbol{\lambda})$, where data encoding quantum gate $\hat{\mathcal{O}}(\mathbf{x}_i)$ and an initial quantum state $|0\rangle^{\otimes n}$, the final quantum state is prepared in the state $|\psi(\boldsymbol{\lambda}, \mathbf{x}_i)\rangle$ by an evolution $\hat{\mathcal{O}}(\boldsymbol{\lambda}) H^{\otimes n} \hat{\mathcal{O}}(\mathbf{x}_i) \, |0\rangle^{\otimes n}$, and then we measure it in the basis $|\boldsymbol{\theta}\rangle$; here, $\hat{\mathcal{O}}(\boldsymbol{\lambda})$ is a PQC model, and $H^{\otimes n}$ are Hadamard gates. In fact, this basis measurement outputs $\boldsymbol{\theta}$ samples with their corresponding probabilities $q_Q(\boldsymbol{\theta}; \boldsymbol{\lambda})$ Bremner et al. [2010]; Benedetti et al. [2021].

To obtain the variational parameters $\boldsymbol{\lambda}$ of the PQC, we can optimize, e.g., the ELBO expressed in Eq. (7) while replacing its classical variational distribution $q_c(\boldsymbol{\theta}; \boldsymbol{\lambda})$ by the quantum variational distribution $q_Q(\boldsymbol{\theta}; \boldsymbol{\lambda})$:

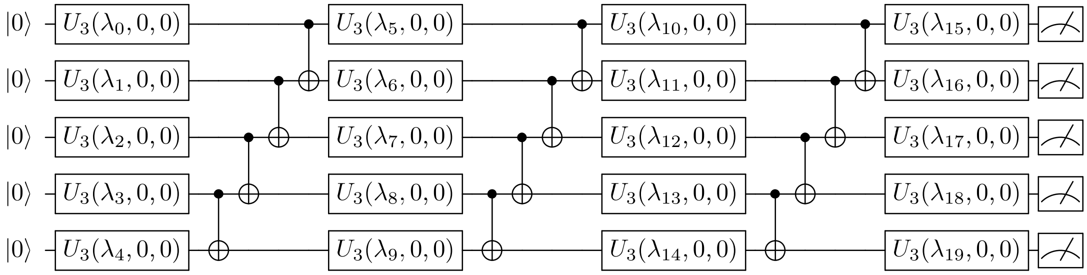

Figure 10: A real-amplitude quantum circuit having depth-one is transpiled into the Clifford+T gate set. It is used to demonstrate the power of QML models by the authros of the article Abbas et al. [2021].

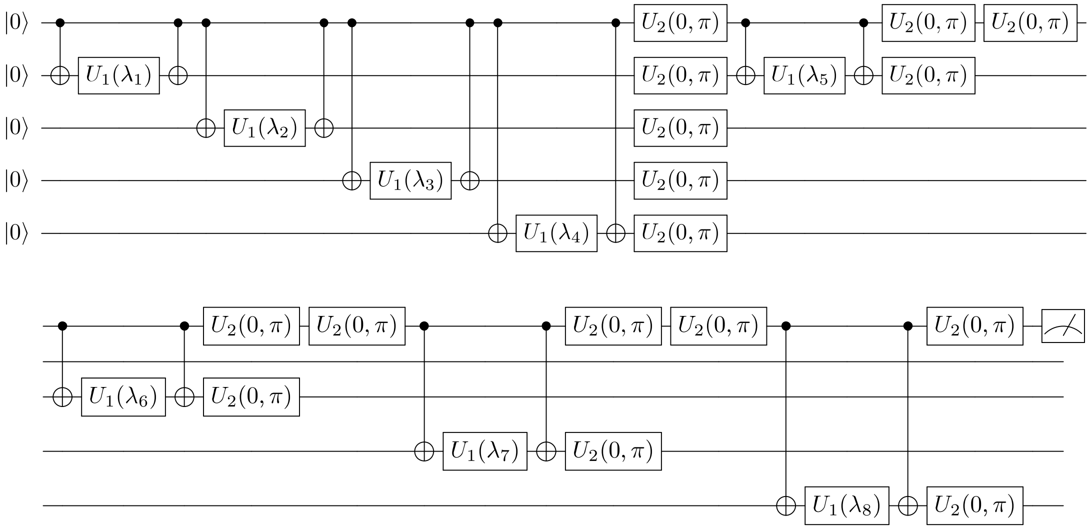

Figure 11: An energy-based quantum circuit having depth-one is transpiled into the Clifford+T gate set. This QML model is proposed for the NISQ device by the authors of the article Farhi and Neven [2018].

$$\begin{aligned} ELBO(p(\mathcal{S},\boldsymbol{\theta})||q_Q(\boldsymbol{\theta};\boldsymbol{\lambda})) &= \mathbb{E}_{q_Q(\boldsymbol{\theta};\boldsymbol{\lambda})}\left[\log\frac{p(\mathcal{S},\boldsymbol{\theta})}{q_Q(\boldsymbol{\theta};\boldsymbol{\lambda})}\right] \\ &= \mathcal{L}_Q(\boldsymbol{\lambda},\boldsymbol{\theta};\mathcal{S}). \end{aligned} \tag{11}$$

The ELBO function $\mathcal{L}_Q(\boldsymbol{\lambda},\boldsymbol{\theta};\mathcal{S})$ can be maximized on quantum computers thanks to quantum differentiable programming paradigm identical to classical differentiable programming one, namely, automatic differentiation for AI models Bergholm et al. [2022].

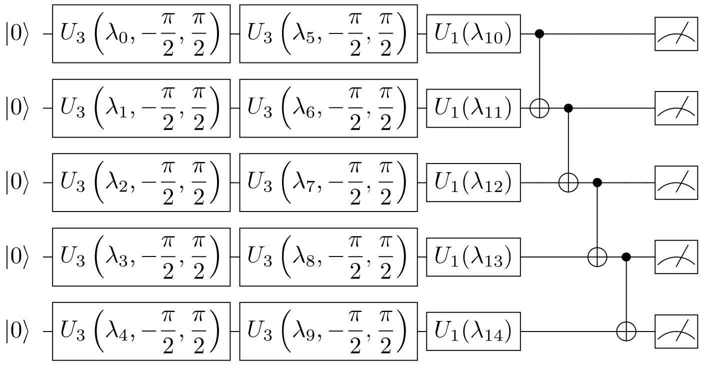
|0⟩
$U_3\left(\lambda_0, -\frac{\pi}{2}, \frac{\pi}{2}\right)$
$U_3\left(\lambda_5, -\frac{\pi}{2}, \frac{\pi}{2}\right)$
$U_1(\lambda_{10})$
|0⟩
$U_3\left(\lambda_1, -\frac{\pi}{2}, \frac{\pi}{2}\right)$
$U_3\left(\lambda_6, -\frac{\pi}{2}, \frac{\pi}{2}\right)$
$U_1(\lambda_{11})$
|0⟩
$U_3\left(\lambda_2, -\frac{\pi}{2}, \frac{\pi}{2}\right)$
$U_3\left(\lambda_7, -\frac{\pi}{2}, \frac{\pi}{2}\right)$
$U_1(\lambda_{12})$
|0⟩
$U_3\left(\lambda_3, -\frac{\pi}{2}, \frac{\pi}{2}\right)$
$U_3\left(\lambda_8, -\frac{\pi}{2}, \frac{\pi}{2}\right)$
$U_1(\lambda_{13})$
|0⟩
$U_3\left(\lambda_4, -\frac{\pi}{2}, \frac{\pi}{2}\right)$
$U_3\left(\lambda_9, -\frac{\pi}{2}, \frac{\pi}{2}\right)$
$U_1(\lambda_{14})$

Figure 12: A strongly-entangling quantum circuit having depth-one is transpiled into the Clifford+T gate set. This QML model is proposed to build a powerful quantum learning model in the article Schuld et al. [2020].

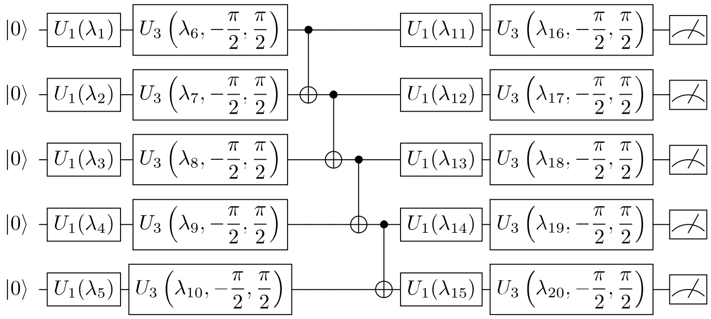
|0⟩
$U_1(\lambda_1)$
$U_3\left(\lambda_6, -\frac{\pi}{2}, \frac{\pi}{2}\right)$
$U_1(\lambda_{11})$
$U_3\left(\lambda_{16}, -\frac{\pi}{2}, \frac{\pi}{2}\right)$
|0⟩
$U_1(\lambda_2)$
$U_3\left(\lambda_7, -\frac{\pi}{2}, \frac{\pi}{2}\right)$
$U_1(\lambda_{12})$
$U_3\left(\lambda_{17}, -\frac{\pi}{2}, \frac{\pi}{2}\right)$
|0⟩
$U_1(\lambda_3)$
$U_3\left(\lambda_8, -\frac{\pi}{2}, \frac{\pi}{2}\right)$
$U_1(\lambda_{13})$
$U_3\left(\lambda_{18}, -\frac{\pi}{2}, \frac{\pi}{2}\right)$
|0⟩
$U_1(\lambda_4)$
$U_3\left(\lambda_9, -\frac{\pi}{2}, \frac{\pi}{2}\right)$
$U_1(\lambda_{14})$
$U_3\left(\lambda_{19}, -\frac{\pi}{2}, \frac{\pi}{2}\right)$
|0⟩
$U_1(\lambda_5)$
$U_3\left(\lambda_{10}, -\frac{\pi}{2}, \frac{\pi}{2}\right)$
$U_1(\lambda_{15})$
$U_3\left(\lambda_{20}, -\frac{\pi}{2}, \frac{\pi}{2}\right)$

Figure 13: A hardware efficient quantum circuit having depth-one is transpiled into the Clifford+T gate set. This PQC is used for quantum variational inference in the article Benedetti et al. [2021].

18 Sizing Quantum Machines

For training PQC models on limited benchmark labeled-HSI datasets, we utilize a classical layer for reducing the dimensionality of the features of the HSI datasets due to a limited number of input qubits. However, how much one needs to reduce the dimensionality of the given HSI dataset depends on quantum computers utilizing, that is, whether we have an access to a NISQ device having error-prone qubits ≤ 100 or a fault-tolerant quantum (FTQ) computer having error-free qubits > 100. In particular, the classical machine plays a less role for pre-processing the HSI dataset, and we can feed many informative features to quantum computers (less dimensionality-reduction) as the number of the error-free qubits of quantum machines increases. In particular, we assume that we use EnMAP HSIs with 230 spectral bands and 145×145 spatial dimensions, that is, a size of the dataset. Moreover, EnMAP HSIs having $21,205$ data points and 230 features are a small-scale image dataset compared with conventional multispectral images for training DL models. To execute the PQC model on NISQ machines having ≤ 100 input qubits, we either reduce the spectral bands of the EnMAP HSIs from 230 to at most 100 or select the most of informative 100 bands to be compatible with the input qubits by utilizing a classical machine. Instead, for FTQ machines having more than 100 input qubits, we persevere more spectral bands of EnMAP HSIs when performing the dimensionality-reduction or feature selection technique in their spectral bands by using a classical machine.

Towards quantum resource estimation, we assess four different PQC models expressed by the Clifford+T gate set (see Figures 10-13). The Clifford-T gate set is defined by U_1, U_2, U_3, and CNOT gates:

$$\begin{aligned} U_1(\lambda) &= \begin{pmatrix} 1 & 0 \\ 0 & e^{i\lambda} \end{pmatrix}, \quad U_2(\lambda, \phi) = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -e^{i\phi} \\ e^{i\lambda} & e^{i(\lambda+\phi)} \end{pmatrix}, \\ U_3(\lambda, \phi, \gamma) &= \begin{pmatrix} \cos(\lambda/2) & -e^{i\gamma}\sin(\lambda/2) \\ -e^{i\phi}\sin(\lambda/2) & e^{i(\phi+\gamma)}\cos(\lambda/2) \end{pmatrix}, \end{aligned} \tag{12}$$

where, for example, $U_1(\pi/4) = T$, $U_1(\pi/2) = S$, $U_2(0, \pi) = H$. Hence, the Clifford-T gate set is $\{U_1(\pi/2)$, $U_2(0, \pi)$, CNOT, $U_1(\pi/4)\}$. Given a HPC+QC system, the four PQC models shown in the Figure 10-13 comprise several parametrized $U_1(\lambda)$ gates. We can execute them on the HPC instead of QCs, and quantum resource required for executing them on QCs is then $\mathcal{O}(1)$ (constant time) if there is either no sign of T-gates or a small number of T-gates. In particular, we deploy them on either HPC or quantum computers depending on the existence and a number of T-gates in their configuration during the training phase via stochastic variational inference (SVI) expressed by the equation (8). Furthermore, a number of T-gates defines quantum resource required for deploying QML models on NISQ and FTQ computers. To determine the number of T gates, we use the concept of symmetry breaking of conventional neural networks Fok et al. [2017]. We strongly emphasize that QML models also breaks the symmetry in their weights in order to decrease their redundant parametrized quantum gates and to generalize better on unseen data points than conventional neural networks. In particular, each weights within a quantum layer must have different digital values for capturing particular features. Hence, we assume that each layer of QML models must have at most a single T-gate at each learning iteration. Hence, our QML models having depth-one can have only one T-gate. Towards quantum resource required for executing them on digital quantum computers Fowler and Gidney [2019]:

1. If our PQCs have 10^8 T-gates and five logical qubits then we need $158,431$ physical qubits (i.e., $9,375$ state distillation qubits and $149,056$ physical qubits) on the surface code distance of $d = 25$, and our QML models then take around 5 hours.
2. If our PQCs have three T-gates and five logical qubits then we need $50,700$ physical qubits (i.e., $14,400$ state distillation qubits and $36,300$ physical qubits) on the surface code distance of $d = 11$, and our QML models then take around 8 hours.
3. If our PQCs have one T-gates and five logical qubits then we need $15,135$ physical qubits (i.e., $14,400$ state distillation qubits and 735 physical qubits) on the surface code distance of $d = 7$, and our QML models then take around 2 hours.

Quantum resource estimation demonstrates that some QML models can not be simulated on the HPC system if a number of T-gates is sufficiently high at the quantum ISA level, and otherwise, we deploy them on quantum computers Beverland et al. [2022]; Reiher et al. [2017]. In addition, we present the scaling of physical qubits and surface (code) distance with respect to the gate error rate in the Figure 14, since our PQC models require the logical error rate denoted by P_L of around 10^{-15} and the gate error p of 10^{-3} given the threshold error rate p_th of 0.57 (the green line in the Figure 14). See also the chapter 3 for the detailed discussion on the assessment and quantum development roadmap of quantum machines.

18.1 Present Day

Superconducting-based quantum machines in the current market comprise around 100 error-prone qubits and depth-5 faulty quantum gates, while quantum learning models require more than depth-5 quantum gates. Hence, quantum variational models can be only implemented as a proof-of-the-concept, when the elements in HSIs have no more than 5-10 percent overlap.

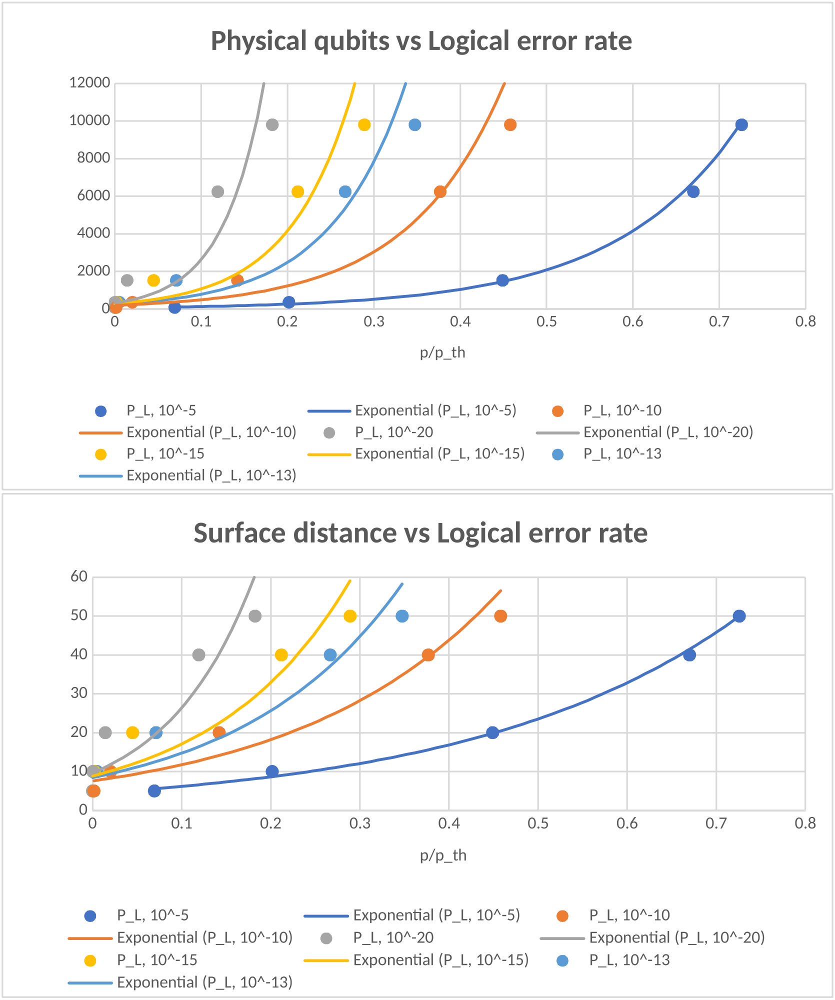

Figure 14: Quantum resource estimation for different logical error rates P_L.

18.2 3-5 years

In 3-5 years, quantum machines begin to have 100 error-prone input qubits and depth-100 faulty quantum gates. Quantum variational inference models can be executed on those quantum machines, while the elements in HSIs could be overlapping up to 10-30 percentage.

18.3 15 years

By this time, quantum machines will have around thousands of error-corrected input qubits and more than depth-100 quantum gates. Quantum variational inference models then can be implemented for operational-sized HSIs having more than 30 percentage overlap in their elements on fault-tolerant quantum machines.

19 SWOT analysis

19.1 Strengths

- Quantum machines could be applied to generate data samples from classically difficult distributions Coyle et al. [2020].
- Proved exponential speed-up in at least one scenario Liu et al. [2021].

19.2 Weaknesses

- Data loading is a major obstacle for achieving exponential speed-up of some QML algorithms Tang [2021b].
- Measurement error mitigation is limited very strongly by the number of qubits and the circuit depth. Quek et al. [2022].
- Quantum machines can be difficult to train due to the error-correction scheme Stilck França and García-Patrón [2021].

19.3 Opportunities

- Major shift in the quality of quantum computers. NISQ machines may be available with less 100 high-quality error-prone qubits in the reference time-frame of 3-5 years, and FTQ machines available with more than 100 fully error corrected qubits in the reference time-frame of 15 years.
- New applications of classical machine learning for quantum computing: compiling, mapping, control, error correction.

19.4 Threats

- Fundamental lack of ability to control, mitigate and correct sources of noise in the quantum machines.
- Novel classical algorithms inspired by quantum computing may outperform some pure quantum algorithms.

20 Conclusion

Deep Learning (DL) models are extensively applied to process big EO datasets due to the powerful computing machines like GPU tensor cores and availability of benchmark labeled-datasets. They are often considered as uninterpretable black-box models with dubious uncertainties: their outputs are not trustworthy and reliable estimates for making high stake decisions involving EO datasets. As opposed to DL models, classical Bayesian statistical approaches are inherently interpretable models generating trustworthy and reliable predictions with error/uncertainty estimates but there is the challenge that they do not scale well as the size of datasets increases or computationally expensive. This challenge can be tackled by combining the best of both DL model and Bayesian analysis, that is, Bayesian Neural Networks (BNNs); namely, DL models scale well on increasing the size of benchmark labeled-datasets, while Bayesian approaches generate the trustworthy and reliable predictions with their confidence level. However, BNNs are still computationally expensive due to their intractable posterior distributions. To weaken BNNs, variational inference (VI) paradigm approximates the intractable posterior by a tractable variational distribution function by optimizing the ELBO metric. Hence, BNNs become scalable interpretable models as the size of benchmark label-datasets increases. More importantly, they generalize well on small-scale datasets compared with DL models. There persists, however, the imperfection that the tractable variational distribution returns approximate samples for uncertainty quantification − not exact samples.

Hyperspectral image (HSI) datasets obtained by hyperspectral imagery satellites are used to make safety-critical and human-centered EO decisions such as hazard and risk assessment (that is, EO Use-Case). In contrast to conventional benchmark labeled-multispectral satellite images, there is the limited availability of benchmark HSIs, namely, there is either the lack of labeled-HSI datasets or small-scale benchmark labeled-HSIs, while DL models require large-scale datasets as input. Hence, we propose to apply BNNs to small-scale benchmark labeled-HSIs for making high stake decisions, since they provide the confidence in their predictions measured by error/uncertainty estimates. In addition, to estimate their uncertainties with high precision, we utilize a quantum variational inference instead of its classical counterpart. For quantum variational inference paradigm, we replace a classical variational distribution function by a parameterized quantum circuit (PQC). According to computational complexity theoretic conjectures, PQCs can not be sampled on a conventional computer. This fact proves that quantum variational inference exhibits so-called quantum advantage over its classical counterpart. The quantum variational distribution approximates the intractable posterior better and generates more superior samples for uncertainty quantification than ones generated by the classical variational distribution − closer to exact samples. In particular, the PQCs can be executed on *superconducting- and photonic-technology machines* integrated with a classical HPC workflow; HPC+QCs paradigm. The classical part selects informative features from limited labeled-HSI images and performs the dimensionality-reduction on them depending on NISQ machine (3-5 years) or FTQ machine (15 years). The larger and more error-free the qubits, the less the classical resource usage for pre-processing HSI datasets.