Preview of Award 2211841 - Annual Project Report

Cover

Federal Agency and Organization Element to Which Report is Submitted:	4900
Federal Award or Other Identifying Number Assigned by Agency:	2211841
Project Title:	FET: Medium: A Quantum Computing Based Approach to Undirected Generative Machine Learning Models
PD/PI Name:	Samee U Khan, Principal Investigator Joseph Picone, Co-Principal Investigator
Recipient Organization:	Mississippi State University
Project/Grant Period:	10/01/2022 - 09/30/2025
Reporting Period:	10/01/2022 - 09/30/2023
Submitting Official (if other than PD\PI):	N/A
Submission Date:	N/A
Signature of Submitting Official (signature shall be submitted in accordance with agency specific instructions)	N/A

Accomplishments

*** What are the major goals of the project?**

Two overarching goals were listed in the proposal: (1) explore the fundamental nature of the claimed quantum annealing behavior and confirm that a wide range of configuration options (Markov Random Fields) trained with a variety of application-relevant data) have this property of finding the so-called ‘difficult to find’ local valleys (LVs); (2) apply this behavior to Deep Graphical Models to achieve superior classification and pattern reconstruction accuracy.

There were three specific tasks outlined in the research plan: (1) confirm the LV-related QAC behavior for more complex graphs and datasets, (2) develop a hybrid classical quantum (HCQ) sampling algorithm, and (3) apply HCQ to contemporary machine learning problems.

We also proposed to focus on two extremely challenging machine learning (ML) tasks that have extensive experience with, and which are representative of “needle in the haystack” problems: (1) seizure detection on EEG signals, which involves learning effective temporal and spatial dependencies, and (2) automatic interpretation of digital pathology images, which involves analyzing extremely large high resolution images (50K x 50K pixels) by integrating local and global dependencies.

*** What was accomplished under these goals and objectives (you must provide information for at least one of the 4 categories below)?**

Major Activities:

In the first year of this three-year project, we focused on Objective 1 and Tasks T1 and T2 of Objective 3 in the attached project timeline. Objective 1, which was executed by the Mississippi State University (MS State) team, focused on understanding the behavior of quantum annealing computers (QAC). Our primary target architecture was the D-Wave quantum computer. Specific activities included (1) classical training of RBMs, (2) classical search for local minima in the RBM energy function, (3) optimization of RBM embedding into the D-Wave hardware, and (4) RBM-based search for the ground state and the local minima.

MS State and Temple collaborated on Objective 2 and focused on (1) establishing reliable benchmarks using existing algorithms, and (2) running a series of diagnostic experiments.

Specific Objectives:

With respect to Objective 1, which was to confirm the behavior of the D-Wave quantum computer for graphs and datasets that are more complex than those explored in the pre-proposal work, we established these objectives: Task 1 – use D-Wave to find unfamiliar local valleys missed by even a prohibitively long classical search, Task 2 – separate the most promising of those local valleys by statistical analysis using a set of criteria, and Task 3 – statistically compare basins of attraction for otherwise similar local valleys concerning the size of their basins of attraction.

We also began work on Task 1 of Objective 2 – develop a hybrid classical-quantum (HCQ) sampling algorithm for RBM training. The specific objective here was to develop an approach for embedding large-size graphs in the QAC hardware.

With respect to Objective 3, our goal (T1) was to define an experimental paradigm that was sufficiently rich to measure statistically significant differences in performance between algorithms, yet also something that can be reasonably computed on a quantum computer. We also established a family of diagnostic experiments (T2) that explored approaches to quantize floating-point data without a significant loss in precision.

Significant Results:

Objective 1:

Our investigation confirmed observations from the latest publications in the field – the D-Wave quantum annealer improves the time-to-solution when doing its primary job – searching for the ground state. However, we are confirming our preliminary observation that this also means a deteriorated potential for using the D-Wave to obtain a representative, diverse sample from probability distributions. The preliminary criterion we used for the sample quality was the number of different local valleys found by the D-Wave compared to classical MCMC methods. Compared to our pre-proposal experiments on the Chimera D-Wave hardware, the Pegasus hardware consistently provides fewer local valleys compared to the classical search.

We considered a new hypothesis that a reduced annealing time while complicating the discovery of the ground state, should produce a more diverse sample. Experiments with the annealing time remain inconclusive; however, no apparent benefits have been observed. A more statistically reliable investigation will be required.

Concerning the statistical analysis of the relative "importance" of the local valleys found by the classical vs. the D-Wave search, the results were qualitatively similar to what had been observed in the pre-proposal work on much smaller graphs. The good news was that most of those higher numbers of the local valleys found by the classical search but missed by the D-Wave are higher-energy, lower-probability (i.e., not representative) states. Concerning the states more critical for a high-quality sample, the D-Wave and the classical search miss many local valleys found by the other. This observation preliminary confirmed our central hypothesis explaining that the lack of significant improvements in the D-Wave-based training compared to the classical training, as reported by many research groups, preliminary supported the merits of our main focus on the hybrid classical-quantum training.

Objective 2:

We developed procedures to use graph sizes of significantly higher dimensionality than in the pre-proposal work. The Pegasus architecture of the D-Wave allowed us to move from 64 RBM visible units and incomplete RBM connectivity to (currently) 144 visible units (pixels of the image) and complete RBM connectivity.

Our approach to testing the adequacy of embedding was based on using classically trained (no use of QC) RBMs and applying the D-Wave to reconstruct the classification label by finding the ground state of a model with qubits corresponding to visible RBM units clamped to each of the training patterns. With proper optimization, we observed that the classification error was similar to that from the classical testing (without the D-Wave), which indicated sufficiently precise embedding. This embedding is ready for future experiments with hybrid classical-quantum RBM training algorithms.

Objective 3:

There are three interesting results regarding Objective 3. To test the generalization abilities of our deep learning models, we used three synthetic balanced datasets of varying difficulty levels. These are summarized in the second attached image (data.pdf). These data sets were designed to test the generalization capability of a machine learning algorithm. Overtraining will result in poor performance on these sets.

We also demonstrated that linear 16-bit quantization of the floating-point values in these data sets does not reduce performance. Since a QC currently can only handle binary inputs, this allows us to map real data sets into a format conducive to QC experimentation.

Third, and most interestingly, our quantum-based model is able to capture the underlying structure of data more efficiently with a small amount of training data. By using only 2,000 data points for training, the QRBM model achieved an error rate of 21% on set no. 10, which is about 50% lower than the error rate of competing models. Compared to the alternatives, QRBM shows superior generalization capability.

In order to test the generalization capabilities of machine learning systems, three synthetic datasets were created using the Python-based open-source tool IMLD that we developed. In dataset number 8, the model is asked to identify an optimal non-linear decision surface based on three classes with complex distributions. Five tight equally spaced Gaussian distributions were used to generate dataset number 9. Gaussian noise was added to dataset #10 after the samples were manually created using IMLD.

We used the 16-bit binary encoding or representation mode for data. Every feature is represented using 16-bit binary data, with each bit representing a binary value (0 or 1). To obtain data in the linear 16-bit binary mode first need to determine the range of values that you want to encode. For example, all the features in set 10 are in the range of (-1.4, 1.4). This is because we need to ensure that your data falls within this range. The second step is dividing the range into 2^16 (65536) equal intervals. Each interval will represent a unique 16-bit binary value. Then map each feature to the nearest interval value and lastly convert the mapped values to their binary representation using 16 bits.

In general, Restricted Boltzmann Machines are not designed to do classification tasks. Hence, a clamp decoding approach is used for this purpose. In this method, the labels are first replaced with a 0 or 1, then a chain is constructed by decoding a number of times to converge to the correct label. Each time, the generated features of the data are replaced with the actual features, but the predicted label remains intact. For each data point, we capture the predicted labels after a few jumps and vote by the majority on all generated labels. In general, RBMs are used as classifiers in this manner. In addition, all other models, especially the QRBM model, were only given one chance to decode the eval set.

In the last attached figure (qrbm_training.pdf), we demonstrate the generalization ability of the three models based on the number of data points used during the training process. Each point shows the performance of the models on evaluation set based on a balanced subset of training points. Since the data points are randomly selected and stochastic in nature RBM and QRBM the reported results are the average of multiple runs. Using only a small portion of the training data, QRBM achieves better performance and demonstrates the ability of the model to generalize. The model is more efficient in terms of using less than 10,000 data points as compared to the other two models. It is evident from the study that quantum computing-inspired models have the potential to revolutionize machine learning, particularly in situations where limited data availability poses a challenge.

Key outcomes or Other achievements:

We have integrated our deep RBM model into our machine learning visualization tool, IMLD. This makes it very easy to do back-to-back comparisons with other conventional algorithms.

We have developed the computing infrastructure, in Python, necessary to run local experiments and connect with the D-Wave cloud. This will allow us to consider much more complex algorithms and understand ways we can overcome the limitations in qubits that these machines have.

*** What opportunities for training and professional development has the project provided?**

Two graduate students and three undergraduate students have learned how to program a quantum computer. These students are also participating in the 2023 IEEE Signal Processing in Medicine and Biology Symposium (IEEE SPMB 2023), a virtual conference held on December 2, 2023.

*** Have the results been disseminated to communities of interest? If so, please provide details.**

The most intereresting QRBM result needs further validation. A publication is under development and will be presented at IEEE SPMB 2024. We will also produce a book chapter as part of this conference participation that will document how to map conventional machine learning problems onto a quantum computer.

*** What do you plan to do during the next reporting period to accomplish the goals?**

Finish Objective 1 – Confirm that the previously observed behavior of the D-Wave for graphs and datasets that are more complex than those explored in the pre-proposal work. Specifically, concerning finishing Task 1 - conduct a more statistically reliable investigation of a possible benefit of a shorter annealing time on the D-Wave's ability to find a larger than otherwise number of distinct high probability local valleys.

Finish Task 3 – statistically compare basins of attraction for otherwise similar local valleys concerning the size of their basins of attraction. Submit a journal manuscript to Quantum Information Processing.

Continue working on Objective 2 - develop a hybrid classical-quantum (HCQ) sampling algorithm for RBM training. Specifically, finish Task 1 – develop an approach for embedding larger-size graphs in the QAC hardware. We have reasons to deprioritize Task 2 until the third year. Instead, we will prioritize work on Task 3 – develop training algorithms for RBM, based on a hybrid classical-quantum sampling algorithm to achieve improvements beyond the mere speed up of nondirected generative model training.

Regarding Objective 3, we have two main goals. First, we want to further understand and validate our primary result that the QRBM system requires less training data to find a good solution. We plan to re-run these experiments on more realistic data sets involving EEG and digital pathology (DPATH) data. We also need to look at how performance varies as a function of the number of training epochs.

Second, we need to implement more contemporary machine learning models that incorporate attention (e.g., transformers). Mapping these computationally intensive algorithms onto a quantum computer will be a challenge. We will start by working with our simulator, and then progressively migrate stages to the QC.

In terms of new tasks for Year 2, we will develop a set of baseline experiments on actual EEG and digital pathology data sets. Both of these data sets represent specific but different machine learning challenges. EEG data requires extremely long-term temporal context. DPATH data requires spatial context. We are investigating a number of recent machine learning models that efficiently encode long-term context. In the first quarter of the second year, we plan to complete a baseline implementation of these algorithms in Python, and then we will investigate how to map them into a QC. In the process, we will investigate optimization of the training process.

Supporting Files

Filename	Description	Uploaded By	Uploaded On
timeline.pdf	The original timeline for the project.	Joseph Picone	09/17/2023
data.pdf	An overview of the three data sets used for diagnostic testing.	Joseph Picone	09/17/2023
table.pdf	A summary of some preliminary results for the baseline RBM system.	Joseph Picone	09/17/2023
qrbm_training.pdf	A demonstration that the quantum RBM system (QRBM) is finding good solutions using a small number of data points.	Joseph Picone	09/17/2023

Products

Books

Book Chapters

Inventions

Journals or Juried Conference Papers
View all journal publications currently available in the NSF Public Access Repository for this award.

The results in the NSF Public Access Repository will include a comprehensive listing of all journal publications recorded to date that are associated with this award.

Licenses

Other Conference Presentations / Papers
M. Babaei, P. Meng, S. S. Shalamzari, and J. Picone (2023). *Accelerating Generalization: Unveiling the Power of QRBM for Efficient Data Modeling with Limited Training Data*. IEEE Signal Processing in Medicine and Biology Symposium. Philadelphia, Pennsylvania, USA. Status = OTHER; Acknowledgement of Federal Support = Yes

K. Ellenberger, D. Couch, J. Greer, N. Gregory, L. Sanchez, K. Love, Y. Koshka, and S. U. Khan (2023). *Quantum Task Mapping for Distributed Heterogeneous Computing Systems*. IEEE Quantum Week. Bellevue, WA, USA. Status = AWAITING_PUBLICATION; Acknowledgement of Federal Support = Yes

K. Ellenberger, D. Couch, J. Greer, N. Gregory, L. Sanchez, K. Love, Y. Koshka, and S. U. Khan (2024). *Quantum task mapping for large-scale heterogeneous computing systems*. SPIE Quantum West. San Francisco, CA. Status = SUBMITTED; Acknowledgement of Federal Support = Yes

Other Products

Other Publications

Patent Applications

Technologies or Techniques

Thesis/Dissertations

Websites or Other Internet Sites
IMLD: The ISIP Machine Learning Demo
https://isip.piconepress.com/projects/imld/

This interactive demo is used to demonstrate machine learning concepts. It has been evolving since the 1990's. We have added implementations of the core algorithms used in this project (Restricted Boltzmann Machines) to the demo to make it easy for users to replicate our results. We use this tool in our graduate level machine learning course and have distributed it as open source software for many years now.

Supporting Files

Filename	Description	Uploaded By	Uploaded On
project_website.pdf	The project web site.	Joseph Picone	09/18/2023
imld_website.pdf	The IMLD web site.	Joseph Picone	09/18/2023

Participants/Organizations

What individuals have worked on the project?

Name	Most Senior Project Role	Nearest Person Month Worked
Khan, Samee	PD/PI	1

Name	Most Senior Project Role	Nearest Person Month Worked
Picone, Joseph	Co PD/PI	1
Babaei, Mahdi	Graduate Student (research assistant)	7
El Yazizi, Abdelmoula	Graduate Student (research assistant)	12
Ellenberger, Kenzie	Graduate Student (research assistant)	12
Meng, Phuykong	Undergraduate Student	2

Full details of individuals who have worked on the project:

Samee U Khan
Email: skhan@ece.msstate.edu
Most Senior Project Role: PD/PI
Nearest Person Month Worked: 1

Contribution to the Project: Project management, graduate student supervision, quantum computing theory and implementation

Funding Support: None

Change in active other support: No

International Collaboration: No
International Travel: No

Joseph Picone
Email: joseph.picone@gmail.com
Most Senior Project Role: Co PD/PI
Nearest Person Month Worked: 1

Contribution to the Project: Development of machine learning algorithms and optimization for quantum computing

Funding Support: None

Change in active other support: No

International Collaboration: No
International Travel: No

Mahdi Babaei
Email: mahdi.babaei@temple.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 7

Contribution to the Project: Development of machine learning algorithms that are conducive to implementation on a quantum computer; benchmarking these algorithms on standard tasks.

Funding Support: None

International Collaboration: No
International Travel: No

Abdelmoula El Yazizi
Email: ae897@msstate.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 12

Contribution to the Project: quantum computing implementations of machine learning

Funding Support: None.

International Collaboration: No
International Travel: No

Kenzie Ellenberger
Email: comets45@hotmail.com
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 12

Contribution to the Project: software development to support quantum computing implementations

Funding Support: None

International Collaboration: No
International Travel: No

Phuykong Meng
Email: phuykong.meng@temple.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 2

Contribution to the Project: Software developer responsible for learning how to run jobs and program the DWAVE and IBM quantum computers.

Funding Support: None

International Collaboration: No
International Travel: No

What other organizations have been involved as partners?
Nothing to report.

Were other collaborators or contacts involved? If so, please provide details.
Nothing to report

Impacts

What is the impact on the development of the principal discipline(s) of the project?

The principal disciplines involved in this project are quantum computing and machine learning. The quantum computing component of the project has focused on quantum annealers while the machine learning component has focused on how to map machine learning problems onto a quantum architecture.

With respect to quantum computing, specifically quantum annealing, we achieved a new understanding of the limitations of quantum annealers (specifically for the D-Wave architecture) for sampling from complex probability distributions. The main new observation concerning that is that new generations of quantum annealing hardware, while showing well-documented in the literature improvements in the main task – finding the ground state – are likely to not improve concerning the sampling applications. This may justify hybrid approaches to sampling.

With respect to the machine learning component of the problem, we demonstrated that mapping a continuous-valued problem onto a discrete-valued problem can be done in a straightforward manner using linear quantization without a loss in performance. We also produced preliminary results that seem to suggest we can find good solutions to deep learning systems with less data and faster computation. This seems consistent with recent literature, though these results need to be more thoroughly validated on larger, more significant experiments.

What is the impact on other disciplines?

The ability to find good solutions for complex systems trained on small amounts of data will impact many applications of machine learning in the health sciences, were training data is often difficult to come by. It will enable many applications where large data sets do not exist or the cost of collecting data is prohibitive.

What is the impact on the development of human resources?

Two graduate students and three undergraduate students have been trained on how to write software and run experiments on the two leading cloud-based quantum computing solutions – D-Wave and IBM. Students have also learned how to run quantum computing simulations of a Linux cluster, and how to implement sampling-based approaches to machine learning (e.g., Boltzmann machines). Students have also been studying contemporary algorithms based on transformers and attention mechanisms.

What was the impact on teaching and educational experiences?

The core architecture we have developed, a restricted Boltzmann machine, has been added to our open-source machine learning demonstration platform known as the ISIP Machine Learning Demo (IMLD). This is being used in our graduate-level machine learning course. We have also introduced lectures on quantum computing into this course. The course materials are also available as open source materials.

What is the impact on physical resources that form infrastructure?
Nothing to report.

What is the impact on institutional resources that form infrastructure?

The Neuronix Linux cluster, which is our primary computing resource at Temple University, can now run a quantum computing simulation, and can be used to connect to cloud-based quantum computers to run experiments. We have gained a much better understanding of the challenges in integrating our vast archives of machine learning data into a quantum computing framework.

What is the impact on information resources that form infrastructure?

We have created a project web site that is used to disseminate information about the project and share our open source resources.

What is the impact on technology transfer?

We have produced reference implementations of our quantum algorithms that researchers in industry and academia can use to guide their research. The demonstration tool, IMLD, is a valuable resource for creating reference baselines doing quick comparisons of algorithms.

What is the impact on society beyond science and technology?
Nothing to report.

What percentage of the award's budget was spent in a foreign country?

No budget has been spent in a foreign country.

Changes/Problems

Changes in approach and reason for change

As we gain a better understanding of how to run complex systems on the DWAVE quantum computer, we have had to reevaluate the cost of computing time and how to best optimize our systems to minimize computing expenses and memory utilization. This is an ongoing activity since our understanding of the best ways to map our algorithms onto these systems, and how to utilize their tools, is evolving.

Actual or Anticipated problems or delays and actions or plans to resolve them

The graduate student identified by Temple University and recruited for the project was terminated effective July 31, 2024 due to a personal conduct issue. This was a significant and unexpected loss. We have recruited and are training a replacement, but that student will not be available to work on the project until January 1, 2024. In the meantime, we are increasing the effort of our undergraduates assigned to the project.

Changes that have a significant impact on expenditures

As mentioned above, the sudden department of a graduate student has delayed expenditures. We expect to get back on track by Fall 2024.

Significant changes in use or care of human subjects
Nothing to report.

Significant changes in use or care of vertebrate animals
Nothing to report.

Significant changes in use or care of biohazards
Nothing to report.

Change in primary performance site location
Nothing to report.

Special Requirements

Responses to any special reporting requirements specified in the award terms and conditions, as well as any award specific reporting requirements.
Nothing to report.