Comparison of quantum and classical methods for labels and patterns in Restricted Boltzmann Machines

Vivek Dixit$^1$, Yaroslav Koshka$^{1,2}$, Tamer Aldwairi$^3$, M.A. Novotny$^{1,4}$

$^1$CCS, HPCC, Mississippi State University (MSU), Mississippi State, MS 39759, USA
$^2$Department of Electrical and Computer Engineering, MSU, Mississippi State, MS 39762, USA
$^3$Mathematics and Computer Science, Ursinus College, Collegeville, PA 19426, USA
$^4$Department of Physics and Astronomy, MSU, Mississippi State, MS 39762, USA

E-mail: man40@msstate.edu

Abstract. Classification and data reconstruction using a restricted Boltzmann machine (RBM) is presented. RBM is an energy-based model which assigns low energy values to the configurations of interest. It is a generative model, once trained it can be used to produce samples from the target distribution. The D-Wave 2000Q is a quantum computer which has been used to exploit its quantum effect for machine learning. Bars-and-stripes (BAS) and cybersecurity (ISCX) datasets were used to train RBMs. The weights and biases of trained RBMs were used to map onto the D-Wave. Classification as well as image reconstruction were performed. Classification accuracy of both datasets indicates comparable performance using D-Wave’s adiabatic annealing and classical Gibbs’s sampling.

1. Introduction

D-Wave 2000Q is D-Wave’s most advanced quantum machine, and is well suited to discrete optimization problems. It finds solution to a problem by finding ground state of the corresponding classical Ising spin model. For a user defined classical Ising spin glass, it returns a state using quantum adiabatic procedure. The returned state is hopefully a ground state. It is a new tool, and researchers are trying to figure out applications by comparing its performance with other existing approaches and techniques.

Mott et al [1] used D-Wave to classify Higgs-boson-decay signals vs. background. They showed that the quantum and classical annealing-based classifiers perform comparable to the state-of-the-art machine learning methods. Mniszewski et al [2] found that the results for graph partitioning using D-Wave systems were comparable to current state-of-the-art methods. Alexandrov et al [3] used a D-Wave for matrix factorization. Kais et al [4] showed an exact mapping between the electronic structure Hamiltonian and the Ising Hamiltonian. Lidar et al [5] used a D-Wave for classification of DNA sequences according to their binding affinities. Koshka
et al [6, 7, 8, 9] performed classification and image reconstruction of a bars and stripes dataset and explored the energy landscape using a RBM embedded onto the D-Wave.

The objective of this work is to investigate D-Wave’s QPU (Quantum Processing Unit) performance on bars-and-stripes (BAS) and cybersecurity (ISCX) datasets. The BAS dataset was chosen for its simplicity, the correct ground states can be visually identified. Classification and reconstruction on BAS dataset demonstrated in our previous work [6], has been reproduced here in order to investigate how the same approach works for more complex ISCX dataset, which should be expected to result in a more complicated energy landscape. The ISCX dataset was generated by the Canadian Institute for Cybersecurity. It has been recently shown that the attack and benign records of the ISCX dataset can be classified using a RBM [10]. In this work, we compare the classification performance of a regular RBM which uses Gibb’s sampling and a RBM embedded on the D-Wave.

2. Methods

2.1. Restricted Boltzmann Machine (RBM)

A RBM is a generative artificial neural network that can learn a probability distribution over multiple variables. It consists of a layer of observable variables (visible layer) and a layer of latent variables (hidden layer). Connections between the variables in the same layer are not permitted. Let $v$ and $h$ be vectors of variables in the visible and hidden layer, respectively. The energy of the model is given as:

$$E(v, h) = -b^Tv - c^Th - v^TWh,$$

where $b$ and $c$ are bias vectors at the visible and hidden layers, respectively. $W$ is the weight matrix of the model parameters. A RBM is an energy based model, with the probability distribution, $P(v, h)$, a function of the energy of the configuration,

$$P(v, h) = \frac{1}{Z} e^{-E(v, h)}, \quad Z = \sum_{v} \sum_{h} e^{-E(v, h)}. \quad (2)$$

The sum over all possible $v$ and $h$ vectors makes the partition function $Z$ very difficult to evaluate. The joint probability, $P(v, h)$, being a function of $Z$ is also hard. Due to the bipartite graph structure of the RBM, the conditional distributions $P(h|v)$ and $P(v|h)$ are simple to compute,

$$P(h|v) = \prod_{i=1}^{m} P(h_i|v), \quad P(v|h) = \prod_{j=1}^{n} P(v_j|h). \quad (3)$$

where ‘$m$’ and ‘$n$’ are the number of units in the hidden and visible layers. The individual activation probabilities are given by:

$$P(h_j = 1|v) = \sigma \left( c_j + \sum_{i=1}^{n} w_{ij} v_i \right), \quad P(v_i = 1|h) = \sigma \left( b_i + \sum_{j=1}^{m} w_{ij} h_j \right) \quad (4)$$

where $\sigma$ is the sigmoid function. A RBM is trained by maximizing the probability distribution over the training data. The log likehood is given by:
\[ l(W, b, c) = \sum_{t=1}^{n} \log P(v^{(t)}) = \sum_{t=1}^{n} \log \sum_{h} P(v^{(t)}, h) \]  

where \( v^{(t)} \) is a sample in the training dataset and

\[ l(W, b, c) = \sum_{t=1}^{n} \log \sum_{h} e^{-E(v^{(t)}, h)} - n \cdot \log \sum_{v,h} e^{-E(v,h)}. \]  

Denote \( \theta = \{W, b, c\} \). The gradient of the log-likelihood is given by:

\[ \nabla_{\theta} l(\theta) = \sum_{t=1}^{n} \frac{\sum_{h} e^{-E(v^{(t)}, h)} \nabla_{\theta} (-E(v^{(t)}, h))}{\sum_{h} e^{-E(v^{(t)}, h)}} - n \cdot \frac{\sum_{v,h} e^{-E(v,h)} \nabla_{\theta} (-E(v, h))}{\sum_{v,h} e^{-E(v,h)}} \]  

\[ \nabla_{\theta} l(\theta) = \sum_{t=1}^{n} \mathbb{E}_{P(h|v^{(t)})}[\nabla_{\theta} (-E(v^{(t)}, h))] - n \cdot \mathbb{E}_{P(h|v)}[\nabla_{\theta} (-E(v, h))]. \]  

Once we have an equation for the gradient of the log-likelihood, weights and biases can be estimated using gradient accent optimization: \( \theta_{\text{new}} = \theta_{\text{old}} + \epsilon \cdot \nabla_{\theta} l(\theta) \), where \( \epsilon \) is the learning rate.

### 2.2. D-Wave System

The Hamiltonian for a D-Wave system of qubits can be represented as:

\[ H_{\text{Ising}} = -\frac{A(s)}{2} \left( \sum_i \sigma_i^x \right) + \frac{B(s)}{2} \left( \sum_i h_i \sigma_i^z + \sum_{(i>j)} J_{ij} \sigma_i^z \sigma_j^z \right) \]  

where \( \sigma_i^{x,z} \) are Pauli matrices operating on qubit \( q_i \), and \( h_i \) and \( J_{i,j} \) are the qubit biases and coupling strengths. \( s \) is called the anneal fraction. \( A(s) \) and \( B(s) \) are known as anneal functions. At \( s = 0 \), \( A(s) \gg B(s) \), while \( A(s) \ll B(s) \) for \( s = 1 \). As we increase \( s \) from 0 to 1, anneal functions change gradually to meet these boundary conditions. In the standard quantum annealing protocol, \( s \) changes from 0 to 1. The network of qubits starts in a global superposition over all possible classical states and after \( s = 1 \), the system is measured in a single classical state.

The arrangement of qubits on the D-Wave chip forms a C16 Chimera graph: 2048 qubits are mapped into a 16 \times 16 matrix of unit cells each of 8 qubits. Figure 1(a) shows a C3 Chimera graph with 3 \times 3 unit cells. Within each unit cell there are two set of 4 qubits which are connected to each other in a bipartite fashion. Each qubit in a unit cell connects to two qubits of the other unit cells. Thus, each qubit connects to a maximum of 6 other qubits. This connectivity can be enhanced by forming ferromagnetic bonds between the qubits as shown in the Figure 1. We follow the embedding approach that was successfully utilized in our previous work [6, 7, 8, 9]. One visible (hidden) unit of a RBM is formed by connecting 16 vertical (horizontal) qubits. There are 64 sets of vertical and horizontal qubits on a C16 Chimera graph. Thus, we get a RBM of 64 visible and hidden units by forming chains in this fashion. Of course care must be taken for any inaccessible qubits to form a further restricted RBM.
3. Results and Discussion

Two datasets (BAS and ISCX) were used for the RBM training and successively embedding onto the D-Wave. BAS is a binary dataset with each record of 64 bit length. The last two bits represent the label of the record: 01 for a bar pattern and 10 for a stripe (Figure 4). Our BAS dataset was comprised of 500 unique records. The number of unique samples used for training was 300, with the remaining 200 samples held for testing the classification accuracy. The ISCX dataset was generated from a bigger dataset obtained from the ‘Canadian Institute for Cybersecurity’ using feature selection [10]. It has 9800 unique records, each record was design to be of length 64 bits. The last two bits represent the label of the record: normal network traffic (01) and malicious network traffic or attacks (10). 7000 records were used for training and the remaining records for the testing.

First, a RBM was trained on the BAS dataset. Weights and biases of the trained RBM were used to embed it onto the D-Wave chip. 95.5% of bars and 94.5% stripes patterns are correctly predicted. Image reconstruction and data generation were also performed using the trained RBM. For image reconstruction, first 25 bits of a record was corrupted, then fed to the RBM framework. After a few Gibb’s cycles, the output was sampled from the visible units. It was found that in almost all the cases full image reconstruction was achieved. The corrupted and reconstructed images are shown in figure 2 (a). For data generation, a random input vector of 0s and 1s was fed to the RBM. The output was a pattern similar to a pattern in the BAS dataset. An example of random input and sampled output is shown in Figure 2(b). Next, the cybersecurity, ISCX, dataset was used to train a RBM. ISCX test dataset had 1401 benign and 1399 attack records. 84.4% of benign and 83.3% of attack records were correctly predicted. Weights and biases obtained from the trained models were then used for D-Wave embedding. Weights were used as coupling strengths $J_{i,j}$ and biases specified the local fields $h_i$ of the qubits, thus forming the objective function. As described earlier, qubits were combined to form RBM units. The embedded RBM had 64 visible and 64 hidden units. We refer to this embedded RBM as a quantum RBM, while the regular RBM as a classical RBM. Once the embedding was
complete, a bias vector $h$ was applied to the visible units of the quantum RBM. The components of this bias vector were proportional to the input sample data. At this step noise was introduced to the bias vector in order to hide the classifying labels or to corrupt the image. A forward anneal was performed and the final state of the qubits were obtained. The final state of the corrupted unit gave the classification/reconstruction result. This procedure of image/data reconstruction is the same as used in the earlier work [6]. In the case of the BAS dataset image reconstruction was performed using the quantum RBM. In most cases reconstruction was easily performed and the output was identical to the uncorrupted image, however, in a few cases a similar but non-identical pattern was obtained as shown in Figure 3(a). Classification was performed on the cybersecurity (ISCX) dataset. Figure 3(c) and 3(d) show two examples where classification of the ISCX samples were performed using the quantum RBM.

To compare performance of the classical RBM and quantum RBM (or D-Wave embedded RBM), we calculated the percentage classification accuracy on both the BAS and ISCX datasets. Accuracy is given by:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \times 100.$$  

T and F stand for true and false; P and N stand for positive and negative. Results are shown in Table 1. Based on these results, we conclude that classification performances of the quantum and classical RBMs are comparable. In other words, minimization of the objective function using adiabatic quantum annealing gave the similar results as Gibb’s sampling of the classical RBM for both BAS and ISCX datasets. This also indicates that the D-Wave optimizer works as well as classical methods for these datasets.

Table 1. Percentage classification accuracy of classical RBM (Gibb’s sampling) and quantum RBM (quantum annealing).

| Dataset | classical RBM | quantum RBM |
|---------|---------------|--------------|
| BAS     | 94.0%         | 94.5%        |
| ISCX    | 83.9%         | 82.2%        |
4. Conclusions

RBM s were trained on bar-and-stripes (BAS) and cybersecurity datasets using contrastive divergence (CD-1). Classification accuracy was measured for the both datasets. Image reconstruction and data generation were performed for the BAS dataset. In addition to the previously demonstrated ability to successfully embed RBM trained on a toy example of BAS dataset, this work shows an effective embedding of a RBM trained on a much more complicated cybersecurity dataset. Image reconstruction was performed and classification accuracy were estimated for the embedded RBM. In the final recognition step, the quantum RBM uses the quantum adiabatic annealing, while Gibb’s sampling is used in the classical RBM. Both the classical and quantum RBM showed comparable classification accuracy. This result is in-line with observations of other researchers [1, 2, 3, 4, 5] where the D-Wave performed in a manner comparable to the existing state-of-the-art methods.

Acknowledgments

The authors would like to acknowledge D-Wave Systems for their use of the Burnaby D-Wave machine. This material is based on research sponsored by the Air Force Research Laboratory (AFRL) under agreement number FA8750-18-1-0096. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory (AFRL) or the U.S. Government.

References

[1] Mott A, Job J, Vlimant J R, Lidar D and Spiropulu M 2017 Nature 550 375–379 URL https://doi.org/10.1038/nature24047
[2] Ushijima-Mwesigwa H, Negre C F A and Mniszewski S M 2017 Proceedings of the Second International Workshop on Post Moore’s Era Supercomputing PMES’17 (New York, NY, USA: ACM) pp 22–29 ISBN 978-1-4503-5126-3 URL http://doi.acm.org/10.1145/3149526.3149531
[3] OMalley D, Vesselinov V V, Alexandrov B S and Alexandrov L B 2018 PLOS ONE 13 1–12 URL https://doi.org/10.1371/journal.pone.0206653
[4] Xia R, Bian T and Kais S 2018 The Journal of Physical Chemistry B 122 3384–3395 pMID: 29099600 (Preprint https://doi.org/10.1021/acs.jpcb.7b10371) URL https://doi.org/10.1021/acs.jpcb.7b10371
[5] YLi R, Felice R D, Rohs R and ALidar D 2018 npj Quantum Information 4 URL https://doi.org/10.1038/s41534-018-0060-8
[6] Koshka Y and Novotny M A 2018 IEEE Transactions on Emerging Topics in Computational Intelligence 1–11 ISSN 2471-285X
[7] Koshka Y and Novotny M A 2018 2018 International Joint Conference on Neural Networks (IJCNN) pp 1–8 ISSN 2161-4407
[8] Koshka Y, Perera D, Hall S and Novotny M A 2017 Neural Computation 29 1815–1837 pMID: 28562219
[9] Koshka Y, Perera D, Hall S and Novotny M A 2016 2016 International Joint Conference on Neural Networks (IJCNN) pp 1948–1954 ISSN 2161-4407
[10] Aldwairi T, Perera D and Novotny M A 2018 Computer Networks 144 111 – 119 ISSN 1389-1286 URL http://www.sciencedirect.com/science/article/pii/S1389128618306005