Comparing deep learning with quantum inference on the D-Wave 2X

Nga T. T. Nguyen and Garrett T. Kenyon

Abstract—We used a quantum annealing D-Wave 2X computer to obtain solutions to NP-hard sparse coding problems for inferring representation of reduced dimensional MNIST images. For comparison, we implemented two deep neural network architectures. The first (AlexNet-like) approximately matched the architecture of the sparse coding model. The second was state-of-the-art (RESNET). Classification based on the D-Wave 2X was superior to matching pursuit and AlexNet and nearly equivalent to RESNET.

1 INTRODUCTION

Deep learning has yielded impressive advances across a variety of machine learning tasks such as Alpha Go Zero [1] and the ImageNet challenge [2]. However, deep neural networks can be spoofed by adversarial examples [3], possibly due to an underlying reliance on low-level image statistics [4]. Moreover, the ability of deep neural networks to learn by analogy [5] as well as the importance of network depth in constructing abstract semantically-meaningful representations [6] has been called into question. Unsupervised learning, on the other hand, particularly Boltzmann Machines [7] and sparse coding paradigms [8], seek to learn joint distributions or causes directly from unlabeled data and thus may be less prone to some of the issues that have plagued task-specific deep learning approaches. For example, features optimized in an entirely unsupervised manner for sparse coding can nonetheless support performance on image classification tasks that is only slightly below that achieved by standard deep neural network architectures such as AlexNet [9]. Sparse inference in these examples, however, relies on a convex approximation to the desired solution. Optimal sparse inference, especially when derived from a highly overcomplete dictionary, is NP-hard, making it difficult to fully assess the potential of such approaches with existing algorithms. Quantum computers offer a possible strategy for rapidly obtaining good solutions to NP-hard sparse inference problems. Previous research [10] demonstrated that models of sparse inference based on lateral inhibition between binary neurons can be directly implemented on the D-Wave 2X Quantum Annealing Computer. Here, we show that on a suitably low-dimensional problem that fits on available quantum annealing architectures, a single layer of features optimized in an unsupervised manner for sparse inference followed by a linear support vector machine (SVM) can exceed the classification performance of a standard AlexNet-like deep neural network trained in a fully supervised manner for the specific task in question.

2 SPARSE-CODING

The hypothesis that neurons encode stimuli by inferring sparse representations explains many of the response properties of simple cells in the mammalian primary visual cortex [8], [11]. Given an overcomplete, non-orthonormal basis $\{\phi_i\}$, inferring a sparse representation involves finding the minimal set of non-zero activation coefficients $\boldsymbol{a}$ that accurately reconstruct a given input signal $\boldsymbol{X}$, corresponding to a minimum of the following energy function:

$$E(\boldsymbol{X}, \boldsymbol{\phi}, \boldsymbol{a}) = \min_{\{\boldsymbol{a}\}}[\,\frac{1}{2}||\boldsymbol{X} - \boldsymbol{\phi}\boldsymbol{a}||^2 + \lambda||\boldsymbol{a}||_0\,] \quad (1)$$

where λ is a trade-off parameter that determines the balance between reconstruction error of the original input image $\boldsymbol{X}$ and the number of non-zero (sparse) activation coefficients. A larger λ encourages sparser solutions. We define $\gamma = \frac{\text{rank}(\boldsymbol{a})}{\text{rank}(\boldsymbol{X})}$ as the overcompleteness of the basis $\{\phi_i\}$. The energy function Eq. (1) is non-convex and contains multiple local minima, so that finding a sparse representation falls into an NP-hard complexity class of decision problems [12].

3 QUANTUM INFERENCE

The D-Wave 2X [13] consists of 1152 quantum bits (qubits) arranged into 12x12 unit cells, forming a Chimera structure with dimensions 12x12x8. Sparse interactions between qubits are restricted to the 16 connections within a unit cell and the 16 connections between nearest-neighboring unit cells [13]. The D-Wave 2X [13] finds optimal solutions to a (discrete) Ising system consisting of N_q binary variables via quantum annealing. Such N_q-body systems can be described by the following classical Hamiltonian:

$$H(\boldsymbol{h}, \boldsymbol{Q}, \boldsymbol{a}) = \sum_{i}^{N_q} h_i a_i + \sum_{i<j}^{N_q} Q_{ij} a_i a_j \quad (2)$$

with binary activation coefficients $a_i = \{0, 1\}\, \forall i \in (1, 2, 3, ..., N_q)$. This objective function defines a QUBO problem. We cast our sparse coding problem, Eq. (1), into QUBO form, Eq. (2), by the transformations [14]:

$$h_i = (-\boldsymbol{\phi}^T \boldsymbol{X} + (\lambda + \frac{1}{2}))_i, \quad Q_{ij} = (\boldsymbol{\phi}^T \boldsymbol{\phi})_{ij}. \quad (3)$$

In Eq. (3), the bias term $\boldsymbol{h}$ (elements h_i) is proportional to the weighted input $\boldsymbol{\phi}^T \boldsymbol{X}$ while the coupling term $\boldsymbol{Q}$ (elements Q_{ij}) corresponds to lateral competition (see also [15]) between qubits given by the interaction matrix $\boldsymbol{\phi}^T \boldsymbol{\phi}$ (with self-interaction excluded and $\boldsymbol{Q}$ being symmetric i.e. $Q_{ij} = Q_{ji}\ \forall i \neq j$). Thus, we infer a sparse representation via Eq. (1) on D-Wave 2X hardware by associating each neuron with a single

- *Nga T. T. Nguyen is with Los Alamos National Laboratory, Los Alamos, NM 87545, USA.*
- *Garrett T. Kenyon is with Los Alamos National Laboratory & New Mexico Consortium, Los Alamos, NM 87545, USA.*
  E-mail: gkenyon@lanl.gov

PMES Workshop, Dallas, 11 Nov 2018. `http://j.mp/pmes18`

Fig. 1. Top: Original MNIST images downsampled to 12x12, Middle: Reconstructed images from bottleneck autoencoder, Bottom: D-Wave reconstruction from randomly selected imprinted features.

feature ϕ_i, represented as a binary logical qubit, with logical qubits embedded on the D-Wave physical chimera.

Despite the sparsity of physical connections on the D-Wave, it is nonetheless possible to construct graphs with arbitrarily dense connectivity by employing "embedding" techniques. Embedding works by chaining together physical qubits so as to extend the effective connectivity but at the cost of reducing the total number of available logical qubits. The D-Wave API provides a heuristic algorithm that searches for an optimal embedding that minimizes the number of physical qubits that are chained together. The exact mapping of a spin glass problem onto the physical D-Wave 2X chimera, including defects, can typically contain approximately $N_q \sim 1000$ spins (qubits) with > 3000 local spin-spin interactions. In contrast, embedding an arbitrary QUBO problem onto the same 2X chimera typically allows no more than $N_q \sim 47$ nodes (logical qubits) but these nodes may be fully connected. Thus, embedding effectively trades qubits for connectivity.

4 Results

D-Wave inputs: reduced dimensional images We first downsampled original 28x28 MNIST images to 12x12 squares and trained a bottleneck autoencoder on TensorFlow that consisted of 47 neurons at its narrowest point. The resulting autoencoder reconstructed images shown in Fig. 1 constitute a database of images with greatly reduced dimensionality, suitable for analysis by the D-Wave 2X. To construct an overcomplete dictionary ($\gamma > 1$), we represent each image as a set of overlapping patches. Sliding a 6x6 patch throughout the entire image of total size 12x12 with a stride $s = 2$, we obtain a set of 4x4 (= 16) patches for each image. The dictionary, which now consists of 47 randomly selected 6x6 patches, has overcompleteness of $\gamma = 47/36 = 1.31$. Consequently, the feature map becomes 4x4x47.

Comparisons The standard deep learning architecture we used, namely AlexNet-like, as our benchmark classifier started with an initial convolution layer with a stride of 2 and 12 features, yielding a 4x4x12 feature map, the same feature map generated by the D-Wave 2X. A subsequent feature map in the deep neural network classifier was 5x5x16, representing an addition convolutional layer. The deep neural network classifier employed a final all-to-all layer using 47 features, which in turn fed into an SoftMAX classifier. Fig. 2 shows that quantum inference enabled superior classification performance compared to an AlexNet-like deep learning architecture. The same deep convolutional neural network (DCNN) produced near-state-of-the-art results of $\sim 98.9\%$ on the original MNIST 28x28 pixel dataset (not plotted), affirming the validity of our DCNN implementation. When applied to the reduced dimensional MNIST dataset, the classification accuracy of the DCNN declined to $94.54 \pm 0.7\%$ (green connected circles with error bars). The D-Wave 2X followed by an Liblinear SVM classifier [16], on the other hand, produces a slightly higher classification score, 95.68% (purple circles) and exhibited almost no variation

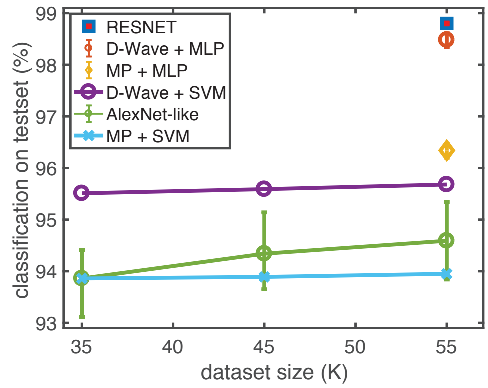

Fig. 2. Image classification on the D-Wave 2X, near state-of-the-art DCNN (AlexNet-like) built with TensorFlow, and matching pursuit for our reduced dimensional 12x12 autoencoded MNIST dataset plotted as a function of dataset size. Training and test sets in each case are in units of 1,000 images and divided in a ratio of 5/1, respectively. We also show classification results using state-of-the-art RESNET for our customized MNIST images.

between reinitialized runs. Matching pursuit, using the same sparsity as D-Wave ($\sim 12-14\%$), followed by an SVM produces a lowest classification score, $\sim 94\%$ (cyan crosses). If we remove the autoencoder procedure and use the downsampled MNIST images to train on our DCNN model, we obtained an accuracy of 96.45% (not plotted), higher that the D-Wave machine result for the classification, consistent with the expectation that the autoencoder destroys some information. We further examined a transfer learning procedure where the D-Wave sparse representations were fed into a multilayer perceptron (MLP), yielding a classification of $\sim 98.48\%$ (orange circle). Feeding matching pursuit into an MLP yielded $\sim 96.34\%$ classification (yellow diamond). Due to the limited capacity of the D-Wave 2X hardware, our AlexNet-like was tailored to match sparse coding with only 47 fully-interacting neurons. We also explored a state-of-the-art deep learning architecture (RESNET) and obtained a classification of $\sim 98.8\%$ (blue square). Using the tailored dataset that has significant reduction in dimensionality and thus fits on the D-Wave 2X hardware reveals the benefits the quantum inference for this image classification task.

5 Conclusions

We have explored classification performance on dimensionally-reduced binary MNIST images using sparse representations generated by the D-Wave 2X quantum computer. Given the limited number of qubits available on the D-Wave 2X, we first used a bottleneck autoencoder to reduce the intrinsic dimensionality of MNIST images. To obtain features for sparse inference, we imprinted 47 6x6 patches randomly chosen from reduced MNIST images. As a target benchmark, we trained standard deep neural network classifiers, implemented in TensorFlow, on our reduced dimensional database. A size matched deep learning architecture yielded a classification score of $94.54 \pm 0.7\%$ compared to 98.48% obtained by feeding sparse representations inferred on the D-Wave 2X into an MLP. Classification of state-of-the-art model (RESNET) was only slightly higher. The RESNET model however contains orders of magnitude more neurons and free parameters compared to quantum inference. We showed that quantum inference on the D-Wave 2X, using approximately 47 binary neurons and only a single layer of convolutional kernels optimized for sparse reconstruction using an unsupervised training procedure, is compatible with the classification performance obtained by a state-of-the-art deep neural network.

REFERENCES

[1] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis, "Mastering the game of Go without human knowledge," Nature, vol. 550, pp. 354–359, April 2017.

[2] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A, Karpathy, A, Khosla, M. Bernstein, A. C. Berg, and Li Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," arXiv:1409.0575 [cs.CV], September 2014.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Advances in Neural Information Processing Systems 27, NIPS, December 2014.

[4] J. Jo and Y. Bengio, "Measuring the tendency of CNNs to Learn Surface Statistical Regularities," arXiv:1711.11561 [cs.LG], November 2017.

[5] M. Ricci, J. Kim, and T. Serre, "Not-So-CLEVR: Visual Relations Strain Feedforward Neural Networks," arXiv:1802.03390 [cs.CV], February 2018.

[6] R. Shwartz-Ziv and N. Tishby "Opening the Black Box of Deep Neural Networks via Information," arXiv:1703.00810 [cs.LG], April 2017.

[7] G. E. Hinton, "A Practical Guide to Training Restricted Boltzmann Machines," in Neural Networks: Tricks of the Trade. Lecture Notes in Computer Science, vol. 7700. Springer, Berlin, Heidelbergpp. pp. 599-619, 2012.

[8] B. A. Olshausen and D. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," Nature, vol. 381, pp 607–609, June 1996.

[9] A. Coates and A. Y. Ng., "The Importance of Encoding versus Training with Sparse Coding and Vector Quantization," in Proceedings of The 28th International Conference on Machine Learning, ICML, pp. 921-928, June 2011.

[10] N. T. T. Nguyen, A. E. Larson, and G. T. Kenyon, "Generating sparse representations using quantum annealing: Comparison to classical algorithms," in Proceedings of the 2017 IEEE International Conference on Rebooting Computing (ICRC), pp. 146–151, DOI: 10.1109/ICRC.2017.8123653, November 2017.

[11] B. A. Olshausen and D. Field, "Sparse Coding with an Overcomplete Basis Set: A Strategy Employed by V1?" Vision Res., vol. 37, pp. 3311–3325, December 1997.

[12] B. Natarajan, "Sparse approximate solutions to linear systems," SIAM Journal on Computing, vol. 24 , pp. 227–234, April 1995.

[13] D-Wave system: http://www.dwavesys.com/.

[14] N. T. T. Nguyen and G. T. Kenyon, "Solving sparse representation for object classification using quantum D-Wave 2X machine," in the Proceedings of The First International Workshop on Post Moore's Era Supercomputing, pp. 43–44, November 2016.

[15] C. J. Rozell, D. H. Johnson, R. G. Baraniuk, and B. A. Olshausen, "Sparse Coding via Thresholding and Local Competition in Neural Circuits," Neural Computation, vol. 20, pp. 2526–2563, October 2008.

[16] R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang, and C. J. Lin, Software available at http://www.csie.ntu.edu.tw/~cjlin/liblinear, "LIBLINEAR: A Library for Large Linear Classification," Journal of Machine Learning Research, vol. 9, pp. 1871–1874, August 2008.