Machine Learning by Two-Dimensional Hierarchical Tensor Networks: A Quantum Information Theoretic Perspective on Deep Architectures

Ding Liu,[1,2] Shi-Ju Ran,[2,*] Peter Wittek,[2,†] Cheng Peng,[3]
Raul Blázquez García,[2] Gang Su,[3,4] and Maciej Lewenstein[2,5]

[1] *Department of Computer Science and Technology, School of Computer Science & Software Engineering, Tianjin Polytechnic University, Tianjin 300387, China*
[2] *ICFO-Institut de Ciencies Fotoniques, The Barcelona Institute of Science and Technology, 08860 Castelldefels (Barcelona), Spain*
[3] *Theoretical Condensed Matter Physics and Computational Materials Physics Laboratory, School of Physical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China*
[4] *Kavli Institute for Theoretical Sciences, University of Chinese Academy of Sciences, Beijing 100190, China*
[5] *ICREA, Passeig Lluis Companys 23, 08010 Barcelona, Spain*

The resemblance between the methods used in studying quantum-many body physics and in machine learning has drawn considerable attention. In particular, tensor networks (TNs) and deep learning architectures bear striking similarities to the extent that TNs can be used for machine learning. Previous results used one-dimensional TNs in image recognition, showing limited scalability and a high bond dimension. In this work, we train two-dimensional hierarchical TNs to solve image recognition problems, using a training algorithm derived from the multipartite entanglement renormalization ansatz (MERA). This approach overcomes scalability issues and implies novel mathematical connections among quantum many-body physics, quantum information theory, and machine learning. While keeping the TN unitary in the training phase, TN states can be defined, which optimally encodes each class of the images into a quantum many-body state. We study the quantum features of the TN states, including quantum entanglement and fidelity. We suggest these quantities could be novel properties that characterize the image classes, as well as the machine learning tasks. Our work could be further applied to identifying possible quantum properties of certain artificial intelligence methods.

I. INTRODUCTION

Over the past years, we have witnessed a booming progress in applying quantum theories and technologies to realistic problems. Paradigmatic examples include quantum simulators [1] and quantum computers [2–4] aimed at tackling challenging problems that are beyond the capability of classical digital computations. The power of these methods stems from the properties quantum many-body systems.

Tensor networks (TNs) belong to the most powerful numerical tools for studying quantum many-body systems [5–7]. The main challenge lies in the exponential growth of the Hilbert space with the system size, making exact descriptions of such quantum states impossible even for systems as small as $\mathcal{O}(10^2)$ electrons. To break the "exponential wall", TNs were suggested as an efficient ansatz that lowers the computational cost to a polynomial dependence on the system size. Astonishing achievements have been made in studying, e.g. spins, bosons, fermions, anyons, gauge fields, and so on [6–9]. TNs are also exploited to predict interactions that are used to design quantum simulators [10].

As TNs allowed the numerical treatment of difficult physical systems by providing layers of abstraction, deep learning achieved similar striking advances in automated feature extraction and pattern recognition [11]. The resemblance between the two approaches is beyond superficial. At at theoretical level, there is a mapping between deep learning and the renormalization group [12–14], which in turn connects holography and deep learning [15, 16], and also allows studying network design from the perspective of quantum entanglement [17]. In turn, neural networks can represent quantum states [18–21].

Most recently, TNs have been applied to solve machine learning problems such as dimensionality reduction [22, 23], handwriting recognition [24, 25], and linguistic applications [26]. Through a feature mapping, an image described as classical information is transferred into a product state defined in a Hilbert space. Then these states are acted onto a TN, giving an output vector that determines the classification of the images into a predefined number of classes. Going further with this clue, it can be seen that when using a vector space for solving image recognition problems, one faces a similar "exponential wall" as in quantum many-body systems. For recognizing an object in the real world, there exist infinite possibilities since the shapes and colors change, in principle, continuously. An image or a gray-scale photo provides an approximation, where the total number of possibilities is lowered to 256^N per channel, with N describing the number of pixels, and it is assumed to be fixed for simplicity. Similar to the applications in quantum physics, TNs show a promising way to lower such an

* Corresponding author. Email: shi-ju.ran@icfo.eu
† Corresponding author. Email: peter.wittek@icfo.eu

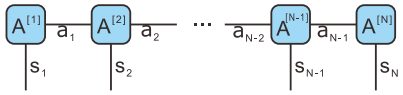

FIG. 1. (Color online) Schematic diagram of a matrix product state.

exponentially large space to a polynomial one.

This work contributes in two aspects. Firstly, we derive an efficient quantum-inspired learning algorithm based on a hierarchical representation that is known as tree TN (TTN) [27–39]. Compared with Refs. [24, 25] where a one-dimensional (1D) TN (called matrix product state (MPS) [40]) is used, TTN suits more the two-dimensional (2D) nature of images. The algorithm is inspired by the multipartite entanglement renormalization ansatz (MERA) approach [41–44], where the tensors in the TN are kept to be unitary during the training. We test the algorithm on both the MNIST (handwriting recognition with binary images) and CIFAR (recognition of color images) databases and obtain accuracies comparable to the performance of convolutional neural networks. More importantly, the TN states can then be defined that optimally encodes each class of images as a quantum many-body state, which is akin to the study of a duality between probabilistic graphical models and TNs [45]. We contrast the bond dimension and model complexity, with results indicating that a growing bond dimension overfits the data. we study the representation in the different layers in the hierarchical TN with t-SNE [46], and find that the level of abstraction changes the same way as in a deep convolutional neural network [47] or a deep belief network [48], and the highest level of the hierarchy allows for a clear separation of the classes. Finally, we show that the fidelities between each two TN states from the two different image classes are low, and we calculate the entanglement entropy of each TN state, which gives an indication of the difficulty of each class.

II. PRELIMINARIES OF TENSOR NETWORK AND MACHINE LEARNING

A TN is defined as a group of tensors whose indexes are shared and contracted in a specific way. TN can represent the partition function of a classical system, and also of a quantum many-body state. For the latter, one famous example is the MPS (Fig. 1), with the following mathematical representation:

$$\Psi_{s_1 s_2 \cdots s_{N-1} s_N} = \sum_{\alpha_1 \cdots \alpha_{N-1}} A^{[1]}_{s_1 \alpha_1} A^{[2]}_{s_2 \alpha_1 \alpha_2} \cdots A^{[N-1]}_{s_{N-1} \alpha_{N-2} \alpha_{N-1}} A^{[N]}_{s_N \alpha_{N-1}}. \quad (1)$$

When describing a physical state, the indexes $\{s_n\}$ are called physical bonds that represent the physical Hilbert space, and dummy indexes $\{\alpha_m\}$ are called virtual bonds

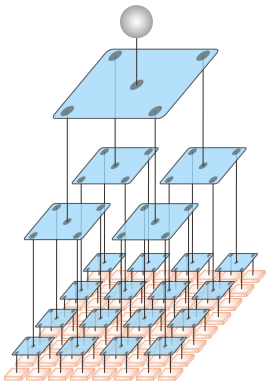

FIG. 2. (Color online) The illustration of a TTN. The squares at the bottom represent the vectors obtained from the pixels of one image through the feature map. The sphere at the top represents the label.

that carry the quantum entanglement. MPS is essentially a 1D state representation. When applied to 2D systems, MPS suffers severe restrictions since one has to choose a snake-like 1D path that covers the 2D manifold. This issue is known as the area law of entanglement entropy [49–51].

A TTN (Fig. 2) provides a natural expression for 2D states, which we can write as a hierarchical structure of K layers:

$$\Psi_{\alpha_{1,1} \cdots \alpha_{N_1,4}} = \sum_{\{\alpha\}} \prod_{k=1}^{K} \prod_{n=1}^{N_k} T^{[k,n]}_{\alpha_{k+1,n}, \alpha_{k,n,1} \alpha_{k,n,2} \alpha_{k,n,3} \alpha_{k,n,4}}, \quad (2)$$

where N_k is the number of tensors in the k-th layer. For simplicity, we ignore the indexes by using bold capitals, and write Eq. (2) as

$$\mathbf{\Psi} = \prod_{k=1}^{K} \prod_{n=1}^{N_k} \mathbf{T}^{[k,n]}, \quad (3)$$

as long as no confusion is caused. The summation signs are also omitted by providing that all indexes that are shared by two tensors will be contracted. Meanwhile, all vectors are assumed to be column. We will follow these conventions throughout this paper.

In a TTN, each local tensor is chosen to have one upward index and four downward indexes. For representing a pure state, the tensor on the top only has four downward indexes. All the indexes except the downward ones of the tensors in the first layer are dummy and will be contracted. In our work, the TTN is slightly different from the pure state representation, by adding an upward index to the top tensor (Fig. 2). This added index corresponds to the labels in the supervised machine learning.

Before training, we need to prepare the data with a feature function that maps N scalars (N is the dimension of the images) to the space of N vectors. The choice of the feature function is arbitrary: we chose the one used

in Ref. [24], where the dimension of each vector (denoted by d) can be controlled. Then, the space is transformed from that of N scalars to a d^N-dimensional Hilbert space.

After "vectorizing" the j-th image in the dataset, the output for classification is a $\tilde{d}$-dimensional vector obtained by contracting the vectors with the TTN, which reads as

$$\tilde{\mathbf{L}}^{[j]} = \mathbf{\Psi} \prod_{n=1}^{N} \mathbf{v}^{[j,n]}, \tag{4}$$

where $\{\mathbf{v}^{[j,n]}\}$ denotes the n-th vector given by the j-th sample. One can see that $\tilde{d}$ is the dimension of the upward index of the top tensor, and should equal to the number of the classes. We use the convention that the position of the maximum value gives the classification of the image predicted by the TTN, akin to a softmax layer in a deep learning network.

For training, the cost function to be minimized is the square error, which is defined as

$$f = \sum_{j=1}^{J} |\tilde{\mathbf{L}}^{[j]} - \mathbf{L}^{[j]}|^2, \tag{5}$$

where J is the number of training samples. $\mathbf{L}^{[j]}$ is a $\tilde{d}$-dimensional vector corresponding to the j-th label. For example, if the j-th sample belongs to the p-th class, $\mathbf{L}^{[j]}$ is defined as

$$L_{\alpha}^{[j]} = \begin{cases} 1, & \text{if } \alpha = p \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

III. MERA-INSPIRED TRAINING ALGORITHM

We use MERA to derive a highly efficient training algorithm. To proceed, let us rewrite the cost function in the following form

$$f = \sum_{j=1}^{J} (\prod_{nn'} \mathbf{v}^{[j,n']\dagger} \mathbf{\Psi}^\dagger \mathbf{\Psi} \mathbf{v}^{[j,n]} - 2 \prod_{n} \mathbf{L}^{[j]\dagger} \mathbf{\Psi} \mathbf{v}^{[j,n]} + 1). \tag{7}$$

Here, the third term comes from the normalization of $\mathbf{L}^{[j]}$, and we assume that the second term is always real.

The dominant cost comes from the first term. We borrow the idea from the MERA approach to reduce this cost. The central idea of MERA is the renormalization group of the entanglement [41]. The renormalization group flows are implemented by tensors that satisfy orthogonal conditions. More specifically, the indexes of one tensor are grouped into to two kinds, say upward and downward indexes. By contracting all the downward indexes of a tensor with its conjugate, one gets an identity (Fig. 2), i.e., $\mathbf{T}\mathbf{T}^\dagger = \mathbf{I}$. On one hand, the orthogonality makes the state remain normalized, a basic requirement of quantum states. On the other hand, the renormalization group flows can be considered as the compressions of the Hilbert space (from the downward to upward indexes). The orthogonality ensure that such compressions are unbiased with $\mathbf{T}^\dagger\mathbf{T} \simeq \mathbf{I}$ in the subspace. The difference from the identity characterizes the errors caused by the compressions.

In our case with the TTN, each tensor has one upward and four downward indexes, which gives a non-square orthogonal matrix by grouping the downward indexes into a large one. Such tensors are called isometries. When all the tensors are isometries, the TTN gives a unitary transformation that compresses a d^N-dimensional space to a $\tilde{d}$-dimensional one. One will approximately have $\mathbf{\Psi}\mathbf{\Psi}^\dagger \simeq \mathbf{I}$ in the subspace that optimizes the classification. For this reason, the first term can be considered as a constant with the orthogonality, and the cost function becomes

$$f = -\sum_{j=1}^{J} \prod_{n} \mathbf{L}^{[j]\dagger} \mathbf{\Psi} \mathbf{v}^{[j,n]}. \tag{8}$$

Each term in f is simply the contraction of one TN, which can be efficiently computed. We stress that independent of Eq. (5), Eq. (8) can be directly used as the cost function. This will lead to a more interesting picture connected to the quantum information theory. The details are given in Sec. V.

The tensors in the TTN are updated alternatively to minimize Eq. (8). To update the tensor $\mathbf{T}^{[k,n]}$ for instance, we assume other tensors are fixed and define the *environment tensor* $\mathbf{E}^{[k,n]}$, which is calculated by contracting everything in Eq. (8) after taking out $\mathbf{T}^{[k,n]}$ (Fig. 3) [44]. Then the cost function becomes $f = -\mathrm{Tr}(\mathbf{T}^{[k,n]}\mathbf{E}^{[k,n]})$. Under the constraint that $\mathbf{T}^{[k,n]}$ is an isometry, the solution of the optimal point is given by $\mathbf{T}^{[k,n]} = \mathbf{V}\mathbf{U}^\dagger$ where $\mathbf{V}$ and $\mathbf{U}$ are calculated from the singular value decomposition $\mathbf{E}^{[k,n]} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\dagger$. At this point, we have $f = -\sum_a \Lambda_a$.

The update of one tensor becomes the calculation of the environment tensor and its singular value decomposition. In the alternating process for updating all the tensors, some tricks are used to accelerate the computations. The idea is to save some intermediate results to avoid repetitive calculations by taking advantage of the tree structure. Another trick is to normalize the vector each time obtained by contracting four vectors with a tensor.

The strategy for the training is the one-against-all classification scheme in machine learning (here dubbed as Strategy-I). For each class, we train one TTN so that it recognizes whether an image belongs to this class. The output of Eq. (4) is a two-dimensional vector. We fix the label for a *yes* answer as $\mathbf{L}^{yes} = [1, 0]$. For P classes, we will accordingly have P TTNs, denoted by $\{\mathbf{\Psi}^{(p)}\}$. Then for recognizing an image (vectorized to $\{\mathbf{v}^{[n]}\}$), we define a P-dimensional vector $\mathbf{F}$ as

$$F_p = \mathbf{L}^{yes\dagger} \mathbf{\Psi}^{(p)} \prod_{n} \mathbf{v}^{[n]}. \tag{9}$$

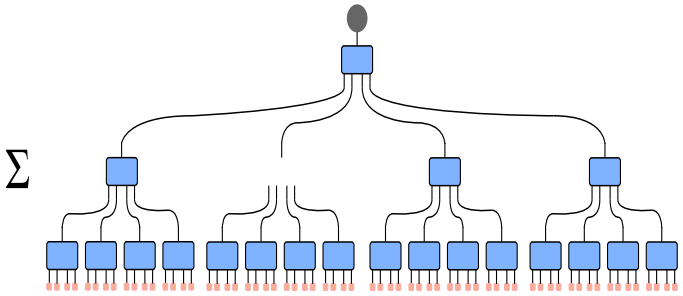

FIG. 3. Illustration of the environment tensor.

The position of its maximal element gives which class the image belongs to. For comparison, we also directly train the TTN by the samples labeled in P classes (dubbed as Strategy-II). In this case, the output [Eq. (4)] is P-dimensional, where the position of its maximal element is expected to give the correct class.

The scaling of both time complexity and space complexity is $O((b_v{}^5 + b_i^4)dN_T)$, where d is the dimension of input vector; b_v the dimension of virtual bond; b_i the dimension of input bond; N_T the number of training inputs.

IV. EXPERIMENTS ON IMAGE RECOGNITION

Our approach to classify image data begins by mapping each pixel x_j to a d-component vector $\phi^{s_j}(x_j)$. This feature map was introduced by [24]) and defined as Eq. (10)

$$\phi^{s_j}(x_j) = \sqrt{\binom{d-1}{s_j-1}}(cos(\frac{\pi}{2}x_j))^{d-s_j}(sin(\frac{\pi}{2}x_j))^{s_j-1} \tag{10}$$

where s_j runs from 1 to d. By using a larger d, the TTN has the potential to approximate a richer class of functions. Our implementation is available under an open source license [52]

A. Benchmark on CIFAR-10

To verify the representation power of TTNs, we used the CIFAR-10 dataset [53]. The dataset consists of 60,000 32×32 RGB images in 10 classes, with 6,000 instances per class. There are 50,000 training images and 10,000 test images. Each RGB image was originally 32×32 pixels: we transformed them to grayscale. Working with gray-scale images reduced the complexity of training, with the trade-off being that less information was available for learning.

We built a TTN with five layers and used the MERA-like algorithm (Section III) to train the model. Specifically, we built a binary classification model to investigate key machine learning and quantum features, instead of constructing a complex multiclass model. We found both the input bond (physical indexes) and the virtual bond (geometrical indexes) had a great impact on the representation power of TTNs, as showed in Fig. 4. This indicates that the limitation of representation power (learnability) of the TTNs is related to the input bond. The same way, the virtual bond determine how accurately the TTNs approximate this limitation.

From the perspective of tensor algebra, the representation power of TTNs depends on the tensor contracted from the entire TTN. Thus the limitation of this relies on the input bond. Furthermore, the TTNs could be considered as a decomposition of this complete contraction, and the virtual bond determine how well the TTNs approximate this. Moreover, this phenomenon could be interpreted from the perspective of quantum many-body theory: the higher entanglement in a quantum many-body system, the more representation power this quantum system has.

The sequence of convolutional and pooling layers in the feature extraction part of a deep learning network is known to arrive at higher and higher levels of abstractions that helps separating the classes in a discriminative learner [11]. This is often visualized by embedding the representation in two dimensions by t-SNE [46], and by coloring the instances according to their classes. If the classes clearly separate in this embedding, the subsequent classifier will have an easy task performing classification at a high accuracy. We plotted this embedding for each layer in the TN in Fig. 5. We observe the same pattern as in deep learning, having a clear separation in the highest level of abstraction.

B. Benchmark on MNIST

To test the generalization of TTNs on a benchmark dataset, we used the MNIST collection, which is widely used in handwritten digit recognition. The training set consists of 60,000 examples, and the test set of 10,000 examples. Each gray-scale image of MNIST was originally 28×28 pixels, and we rescaled them to 16×16 pixels for building TTNs with four layers on this scale. The MERA-like algorithm was used to train the model.

Similar to the last experiment, we built a binary model to show the performance of generalization. With the increase of bond dimension (both of the input bond and virtual bond), we found an apparent rise of training accuracy, which is consistent with the results in Fig. 6. At the same time, we observed the decline of testing accuracy. The increase of bond dimension leads to a sharp increase of the number of parameters and, as a result, it will give rise to overfitting and lower the performance of generalization. Therefore, one must pay attention to finding the optimal bond dimension – we can think of this as a hyperparameter controlling model complexity.

For multi-class learning, we choose the one-against-all strategy to build a 10-class model, which classify an in-

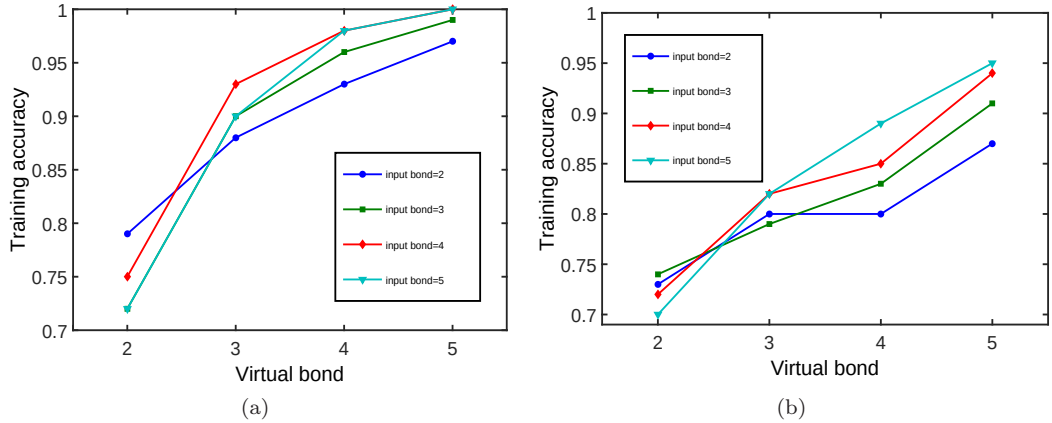

FIG. 4. Binary classification accuracy on CIFAR-10. (a) Number of training samples=200; (b) Number of training samples=600.

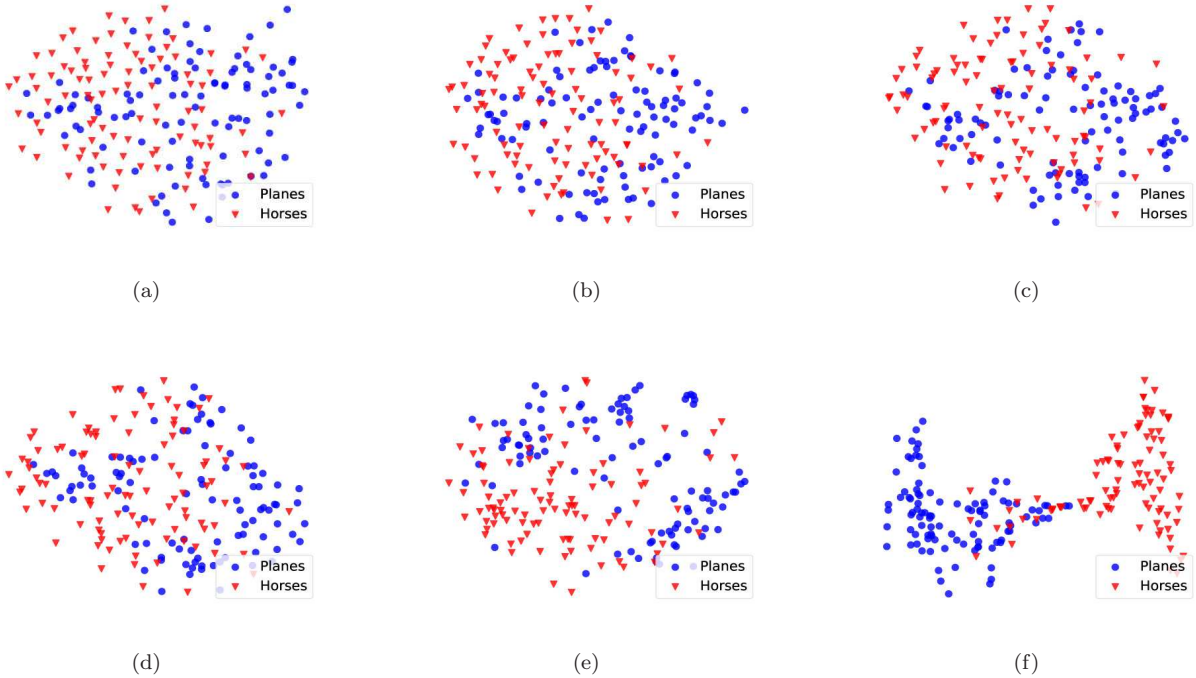

FIG. 5. Embedding of data instances of CIFAR-10 by t-SNE corresponding to each layer in the TN. (a) Original data distribution; (b) 1st layer; (c) 2nd layer; (d) 3rd layer; (e) 4th layer; (f) 5th layer.

put image by choosing the label for which the output is largest. To avoid overfitting and lower computing cost, we apply different feature map to each individual model. The parameters configuration and testing results are in Table I. We repeated the t-SNE visualization on MNIST, but since the classes separate well even in the raw data, we did not include the corresponding figures here.

V. ENCODING IMAGES IN QUANTUM STATES: FIDELITY AND ENTANGLEMENT

Taking a TTN $\mathbf{\Psi}$ trained with Strategy-II, we define P TN state as

$$\mathbf{\Phi}^{[p]} = \mathbf{L}^{[p]\dagger}\mathbf{\Psi}. \tag{11}$$

In $\mathbf{\Phi}^{[p]}$, the upward index of the top tensor is contracted with the label ($\mathbf{L}^{[p]}$), giving a TN state that represents a pure quantum state.

The quantum state representations allow us to use quantum theories to study images and the related issues.

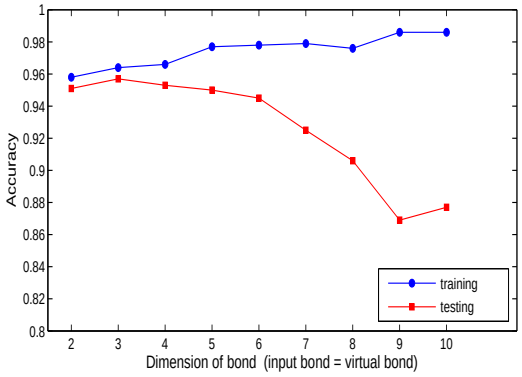

FIG. 6. Training and test accuracy as the function of the dimension of indexes on the MNIST dataset. The number of training samples is 1000 for each pair of classes.

.

TABLE I. 10-class classification on MNIST

model	Training accuracy(%)	Testing accuracy(%)	Input bond	Virtual bond
0	96	97	3	3
1	97	97	3	3
2	96	95	3	4
3	94	93	4	4
4	96	95	2	3
5	94	95	6	6
6	97	96	2	3
7	94	94	6	6
8	93	93	6	6
9	94	93	4	6
10-class		92	/	/

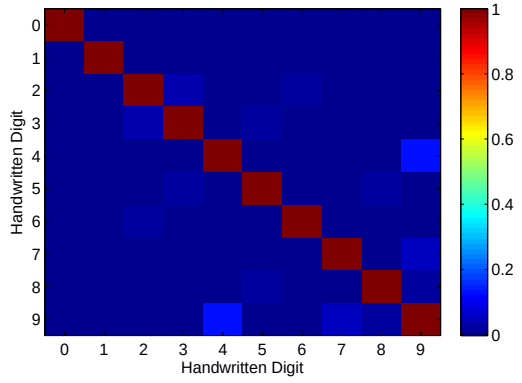

FIG. 7. Fidelity $\mathcal{F}_{p'p}$ between each two handwritten digits, which ranges from -0.0032 to 1. The diagonal terms $\mathcal{F}_{pp} = 1$ because the quantum states are normalized.

Let us begin with the cost function. In Section III, we started from a frequently used cost function in Eq. (5), and derived a cost function in Eq. (8). In the following, we show that such a cost function can be understood by the notion of fidelity. With Eq. (11), the cost function in Eq. (8) can be rewritten as

$$f = -\sum_j \mathbf{\Phi}^{[p]T} \prod_n \mathbf{v}^{[j,n]}. \tag{12}$$

Knowing that the fidelity between two states is defined as their inner product, each term in the summation is simply the fidelity [2, 54] between a vectorized image and the corresponding state $\mathbf{\Phi}^{[p]}$. Considering the fidelity measure as the distance between two states, $\{\mathbf{\Phi}^{[p]}\}$ are the P states, where the distance between each $\mathbf{\Phi}^{[p]}$ and the corresponding vectorized images are minimized. In other words, the cost function is in fact the total fidelity, and $\mathbf{\Phi}^{[p]}$ is the quantum state that optimally encodes the p-th class of images.

Note that due to the orthogonality of the top tensor, such P states are orthogonal to each other, i.e., $\mathbf{\Phi}^{[p']\dagger}\mathbf{\Phi}^{[p]} = I_{p'p}$. This might trap us to a bad local minimum. For this reason, we propose Strategy-I. For each class, we train a TTN to give *yes-or-no* answers. Each TTN gives two TN states labeled *yes* and *no*, respectively. Then we will have $2P$ TN states. $\{\mathbf{\Phi}^{[p]}\}$ are then defined by taking the P *yes*-labeled TN states. The elements of $\mathbf{F}$ in Eq. (9) are defined by the summation of the fidelity between $\mathbf{\Phi}^{[p]}$ and the class of vectorized images. In this scenario, the classification is decided by finding the $\mathbf{\Phi}^{[p]}$ that gives the maximal fidelity with the input image, while the orthogonal conditions among $\{\mathbf{\Phi}^{[p]}\}$ no longer exist.

Besides the algorithmic interpretation, fidelity may imply more intrinsic information. Without the orthogonality of $\{\mathbf{\Phi}^{[p]}\}$, the fidelity $\mathcal{F}_{p'p} = \mathbf{\Phi}^{[p']\dagger}\mathbf{\Phi}^{[p']}$ (Fig. 8 (a)) describes the differences between the quantum states that encode different classes of images. As shown in Fig. 7, $\mathcal{F}_{p'p}$ remains quite small in most cases, indicating that the orthogonality still approximately holds using Strategy-I. Still, some results are still relatively large, e.g., $\mathcal{F}_{4,9} = [\textit{add the number}]$. We speculate this is closely related to the ways how the data are fed and processed in the TN. In our case, two image classes that have similar shapes will result in a larger fidelity, because the TTN essentially provides a *real-space renormalization flow*. In other words, the input vectors are still initially arranged and renormalized layer by layer according to their spatial locations in the image; each tensor renormalizes four nearest-neighboring vectors into one vector. Fidelity can be potentially applied to building a network, where the nodes are classes of images and the weights of the connections are given by the $\mathcal{F}_{p'p}$. This might provide a mathematical model on how different classes of images are associated to each other. We leave these questions for future investigations.

Another important concept of quantum mechanics is

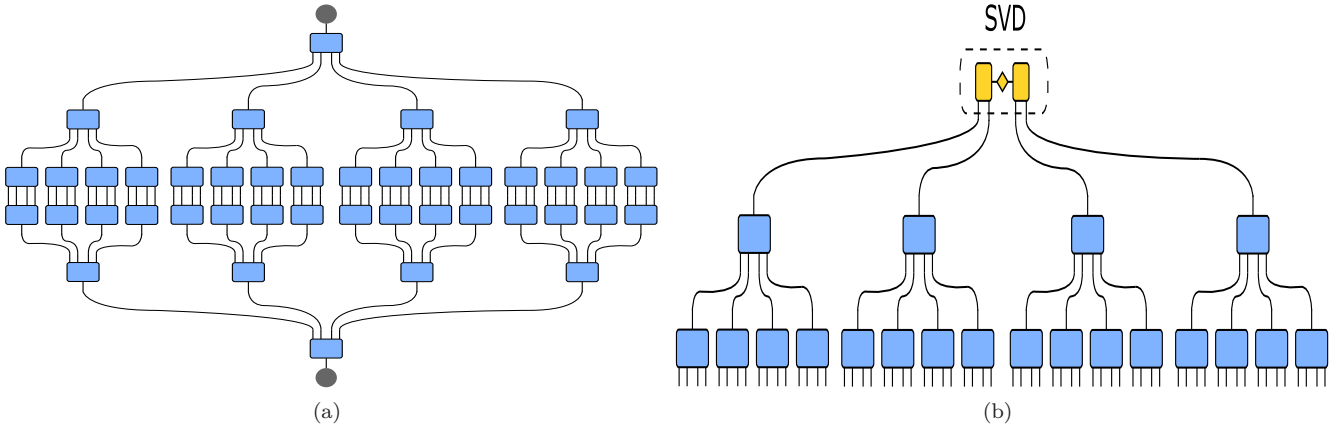

FIG. 8. Schematic diagram of fidelity and entanglement entropy calculation. (a) Fidelity; (b) Entanglement entropy.

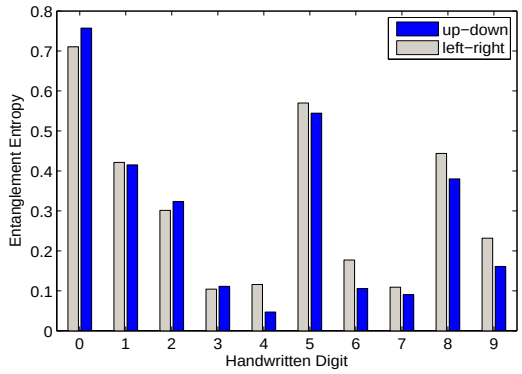

FIG. 9. Entanglement entropy corresponding to each handwritten digit.

(bipartite) entanglement, a quantum version of correlations [54, 55]. It is one of the key characters that distinguishes the quantum states from classical ones. Entanglement is usually given by a normalized positive-defined vector called entanglement spectrum (denoted as $\mathbf{\Lambda}$), and is measured by the entanglement entropy $S = -\sum_a \Lambda_a^2 \ln \Lambda_a^2$. Having two subsystems, entanglement entropy measures the amount of information of one subsystem that can be gained by measuring the other subsystem. In the framework of TN, entanglement entropy determines the minimal dimensions of the dummy indexes needed for reaching a certain precision.

In our image recognition, entanglement entropy characterizes how much information of one part of the image we can gain by knowing the rest part of the image. In other words, if we only know a part of an image and want to predict the rest according to the trained TTN (the quantum state that encodes the corresponding class), the entanglement entropy measures how accurately this can be done. Here, an important analog is between knowing a part of the image and measuring the corresponding subsystem of the quantum state. Thus, the trained TTN might be used on image processing, e.g., to recover an image from a damaged or compressed lower-resolution version.

Fig. 9 shows the entanglement entropy for each class in the MNIST dataset. With the TTN, the entanglement spectrum is simply the singular values of the matrix $\mathbf{M} = \mathbf{L}^\dagger \mathbf{T}^{[K,1]}$ with $\mathbf{L}$ the label and $\mathbf{T}^{[K,1]}$ the top tensor (Fig. 8 (b)). This is because the all the tensors in the TTN are orthogonal. Note that $\mathbf{M}$ has four indexes, of which each represents the effective space renormalized from one quarter of the vectorized image. Thus, the bipartition of the entanglement determines how the four indexes of $\mathbf{M}$ are grouped into two bigger indexes before calculating the SVD. We compute two kinds of entanglement entropy by cutting the system in the middle along the x or y direction. Our results suggest that the images of "0" and "4" are the easiest and hardest, respectively, to predict one part of the image by knowing the other part.

VI. CONCLUSION AND OUTLOOK

We continued the forays into using tensor networks for machine learning, focusing on hierarchical, two-dimensional tree tensor networks that we found a natural fit for image recognition problems. This proved a scalable approach that had a high precision, and we can conclude the following observations:

- The limitation of representation power (learnability) of the TTNs model strongly depends on the open indexes (physical indexes). And, the inner indexes (geometrical indexes) determine how well the TTNs approximate this limitation.

- A hierarchical tensor network exhibits the same increase level of abstraction as a deep convolutional neural network or a deep belief network.

- Fidelity can give us an insight how difficult it is to tell two classes apart.

- Entanglement entropy can characterize the difficulty of representing a class of problems.

In future work, we plan to use fidelity-based training in an unsupervised setting and applying the trained TTN to recover damaged or compressed images and using entanglement entropy to characterize the accuracy.

ACKNOWLEDGMENTS

SJR is grateful to Ivan Glasser and Nicola Pancotti for stimulating discussions. DL was supported by the China Scholarship Council (201609345008) and the National Natural Science Foundation in China (61771340). SJR, PW, and ML acknowledge support the Spanish Ministry of Economy and Competitiveness (Severo Ochoa Programme for Centres of Excellence in R&D SEV-2015-0522), Fundació Privada Cellex, and Generalitat de Catalunya CERCA Programme. SJR and ML were further supported by ERC AdG OSYRIS (ERC-2013-AdG Grant No. 339106), the Spanish MINECO grants FOQUS (FIS2013-46768-P), FISICATEAMO (FIS2016-79508-P), Catalan AGAUR SGR 874, EU FETPRO QUIC, EQuaM (FP7/2007-2013 Grant No. 323714), and Fundació Catalunya - La Pedrera · Ignacio Cirac Program Chair. PW acknowledges financial support from the ERC (Consolidator Grant QITBOX) and QIBEQI FIS2016-80773-P), and a hardware donation by Nvidia Corporation.

[1] Andreas Trabesinger, J. Ignacio Cirac, Peter Zoller, Immanuel Bloch, Jean Dalibard, and Sylvain Nascimbéne, "Nature Physics Insight - Quantum Simulation," Nature Physics **8** (2012).
[2] Andrew Steane, "Quantum computing," Reports on Progress in Physics **61**, 117–173 (1998), arXiv:quant-ph/9708022.
[3] Emanuel Knill, "Physics: quantum computing," Nature **463**, 441–443 (2010).
[4] Iulia Buluta, Sahel Ashhab, and Franco Nori, "Natural and artificial atoms for quantum computation," Reports on Progress in Physics **74**, 104401 (2011), arXiv:1002.1871.
[5] Román Orús, "A practical introduction to tensor networks: Matrix product states and projected entangled pair states," Annals of Physics **349**, 117 (2014), arXiv:1306.2164.
[6] Román Orús, "Advances on tensor network theory: symmetries, fermions, entanglement, and holography," The European Physical Journal B **87**, 280 (2014), arXiv:1407.6552.
[7] Shi-Ju Ran, Emanuele Tirrito, Cheng Peng, Xi Chen, Gang Su, and Maciej Lewenstein, "Review of tensor network contraction approaches," (2017), arXiv:1708.09213.
[8] Frank Verstraete, Valentin Murg, and J. Ignacio Cirac, "Matrix product states, projected entangled pair states, and variational renormalization group methods for quantum spin systems," Advances in Physics **57**, 143–224 (2008), arXiv:0907.2796.
[9] J. Ignacio. Cirac and Frank Verstraete, "Renormalization and tensor product states in spin chains and lattices," Journal of Physics A: Mathematical and Theoretical **42**, 504004 (2009), 0910.1130:0910.1130.
[10] Shi-Ju Ran, Angelo Piga, Cheng Peng, Gang Su, and Maciej Lewenstein, "Few-body systems capture many-body physics: Tensor network approach," (2017), arXiv:1703.09814.
[11] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning," Nature **521**, 436–444 (2015).
[12] Pankaj Mehta and David J. Schwab, "An exact mapping between the variational renormalization group and deep learning," (2014), arXiv:1410.3831.
[13] Yoav Levine, David Yakira, Nadav Cohen, and Amnon Shashua, "Deep learning and quantum physics: A fundamental bridge," (2017), arXiv:1704.01552.
[14] Maciej Koch-Janusz and Zohar Ringel, "Mutual information, neural networks and the renormalization group," (2017), arXiv:1704.06279.
[15] Yi-Zhuang You, Zhao Yang, and Xiao-Liang Qi, "Machine learning spatial geometry from entanglement features," (2017), arXiv:1709.01223.
[16] Wen-Cong Gan and Fu-Wen Shu, "Holography as deep learning," (2017), arXiv:1705.05750.
[17] Yoav Levine, David Yakira, Nadav Cohen, and Amnon Shashua, "Deep Learning and Quantum Entanglement: Fundamental Connections with Implications to Network Design." (2017), arXiv:1704.01552.
[18] Giuseppe Carleo and Matthias Troyer, "Solving the quantum many-body problem with artificial neural networks," Science **355**, 602–606 (2017), arXiv:1606.02318.
[19] Jing Chen, Song Cheng, Haidong Xie, Lei Wang, and Tao Xiang, "On the equivalence of restricted Boltzmann machines and tensor network states," (2017), arXiv:1701.04831.
[20] Yichen Huang and Joel E. Moore, "Neural network representation of tensor network and chiral states," (2017), arXiv:1701.06246.
[21] Ivan Glasser, Nicola Pancotti, Moritz August, Ivan D. Rodriguez, and J. Ignacio Cirac, "Neural networks quantum states, string-bond states and chiral topological states," (2017), arXiv:1710.04045.
[22] Andrzej Cichocki, Namgil Lee, Ivan Oseledets, Anh-Huy Phan, Qibin Zhao, Danilo P. Mandic, and Others, "Tensor networks for dimensionality reduction and large-scale optimization: Part 1 low-rank tensor decompositions," Foundations and Trends® in Machine Learning **9**, 249–429 (2016).
[23] Andrzej Cichocki, Anh-Huy Phan, Qibin Zhao, Namgil Lee, Ivan Oseledets, Masashi Sugiyama, Danilo P. Mandic, and Others, "Tensor networks for dimensionality reduction and large-scale optimization: Part 2 applications and future perspectives," Foundations and Trends® in Machine Learning **9**, 431–673 (2017).
[24] E. Miles Stoudenmire and David J. Schwab, "Supervised learning with tensor networks," Advances in Neural

Information Processing Systems **29**, 4799–4807 (2016), 1605.05775.
[25] Zhao-Yu Han, Jun Wang, Heng Fan, Lei Wang, and Pan Zhang, "Unsupervised generative modeling using matrix product states," (2017), arXiv:1709.01662.
[26] Angel J. Gallego and Román Orús, "The physical structure of grammatical correlations: equivalences, formalizations and consequences," (2017), arXiv:1708.01525.
[27] Mark Fannes, Bruno Nachtergaele, and Reinhard F. Werner, "Ground states of VBS models on Cayley trees," Journal of Statistical Physics **66**, 939 (1992).
[28] Harald Niggemann, Andreas Klümper, and Johannes Zittartz, "Quantum phase transition in spin-3/2 systems on the hexagonal lattice–optimum ground state approach," Zeitschrift für Physik B Condensed Matter **104**, 103–110 (1997), arXiv:cond-mat/9702178.
[29] Barry Friedman, "A density matrix renormalization group approach to interacting quantum systems on Cayley trees," Journal of Physics: Condensed Matter **9**, 9021 (1997).
[30] Marie-Bernadette Lepetit, Maixent Cousy, and Gustavo M. Pastor, "Density-matrix renormalization study of the Hubbard model on a Bethe lattice," European Physical Journal B: Condensed Matter Physics **13**, 421–427 (2000), arXiv:cond-mat/9801228.
[31] Miguel Angel Martín-Delgado, Javier Rodriguez-Laguna, and Germán Sierra, "Density-matrix renormalization-group study of excitons in dendrimers," Physical Review B **65**, 155116 (2002), arXiv:cond-mat/0012382.
[32] Yaoyun Shi, Luming Duan, and Guifre Vidal, "Classical simulation of quantum many-body systems with a tree tensor network," Physical Review A **74**, 022320 (2006), arXiv:quant-ph/0511070.
[33] Daniel Nagaj, Edward Farhi, Jeffrey Goldstone, Peter Shor, and Igor Sylvester, "Quantum transverse-field Ising model on an infinite tree from matrix product states," Physical Review B **77**, 214431 (2008), arXiv:0712.1806.
[34] Luca Tagliacozzo, Glen Evenbly, and Guifre Vidal, "Simulation of two-dimensional quantum systems using a tree tensor network that exploits the entropic area law," Physical Review B **80**, 235127 (2009), arXiv:0903.5017.
[35] Valentin Murg, Frank Verstraete, Örs Legeza, and Reinhard M. Noack, "Simulating strongly correlated quantum systems with tree tensor networks," Physical Review B **82**, 205105 (2010), arXiv:1006.3095.
[36] Wei Li, Jan von Delft, and Tao Xiang, "Efficient simulation of infinite tree tensor network states on the Bethe lattice," Physical Review B **86**, 195137 (2012), arXiv:1209.2387.
[37] Naoi Nakatani and Garnet Kin-Lic Chan, "Efficient tree tensor network states (TTNS) for quantum chemistry: Generalizations of the density matrix renormalization group algorithm," Journal of Chemical Physics **138**, 134113 (2013), 1302.2298:1302.2298.
[38] Iztok Pižorn, Frank Verstraete, and Robert M. Konik, "Tree tensor networks and entanglement spectra," Physical Review B **88**, 195102 (2013), arXiv:1309.2255.
[39] Valentin Murg, Frank Verstraete, Reinhold Schneider, Péter R. Nagy, and Örs Legeza, "Tree tensor network state with variable tensor order: An efficient multireference method for strongly correlated systems," Journal of Chemical Theory and Computation **11**, 1027–1036 (2015).
[40] Stellan Östlund and Stefan Rommer, "Thermodynamic limit of density matrix renormalization," Physical Review Letters **75**, 3537 (1995), arXiv:cond-mat/9503107.
[41] Guifre Vidal, "Entanglement renormalization," Physical Review Letters **99**, 220405 (2007), arXiv:cond-mat/0512165.
[42] Guifre Vidal, "Class of quantum many-body states that can be efficiently simulated," Physical Review Letters **101**, 110501 (2008), arXiv:quant-ph/0610099.
[43] Lukasz Cincio, Jacek Dziarmaga, and Marek M. Rams, "Multiscale entanglement renormalization ansatz in two dimensions: quantum Ising model," Physical Review Letters **100**, 240603 (2008), arXiv:0710.3829.
[44] Glen Evenbly and Guifre Vidal, "Algorithms for entanglement renormalization," Physical Review B **79**, 144108 (2009), arXiv:0707.1454.
[45] Elina Robeva and Anna Seigal, "Duality of graphical models and tensor networks," (2017), arXiv:1710.01437.
[46] Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-SNE," Journal of Machine Learning Research **9**, 2579–2605 (2008).
[47] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, Vol. 25 (2012) pp. 1097–1105.
[48] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh, "A fast learning algorithm for deep belief nets," Neural Computation **18**, 1527–1554 (2006).
[49] Frank Verstraete and J. Ignacio Cirac, "Matrix product states represent ground states faithfully," Physical Review B **73**, 094423 (2006), arXiv:cond-mat/0505140.
[50] Matthew B. Hastings, "An area law for one-dimensional quantum systems," Journal of Statistical Mechanics: Theory and Experiment **2007** (2007), 10.1088/1742-5468/2007/08/P08024, arXiv:0705.2024.
[51] Norbert Schuch, Michael M. Wolf, Frank Verstraete, and J. Ignacio Cirac, "Entropy scaling and simulability by matrix product states," Physical Review Letters **100**, 030504 (2008), arXiv:0705.0292.
[52] The code of the implementation is available at `https://github.com/dingliu0305/Tree-Tensor-Networks-in-Machine-Learning`.
[53] Alex Krizhevsky and Geoffrey Hinton, *Learning multiple layers of features from tiny images*, Tech. Rep. (2009).
[54] Charles H. Bennett and David P. DiVincenzo, "Quantum information and computation," Nature **404**, 247–255 (2000).
[55] Ryszard Horodecki, Paweł Horodecki, Michał Horodecki, and Karol Horodecki, "Quantum entanglement," Reviews of Modern Physics **81**, 865–942 (2009), arXiv:quant-ph/0702225.