Residual Tensor Train: A Quantum-Inspired Approach for Learning Multiple Multilinear Correlations

Yiwei Chen, Yu Pan, Senior Member, IEEE, and Daoyi Dong, Senior Member, IEEE

Abstract—States of quantum many-body systems are defined in a high-dimensional Hilbert space, where rich and complex interactions among subsystems can be modeled. In machine learning, complex multiple multilinear correlations may also exist within input features. In this article, we present a quantum-inspired multilinear model, named residual tensor train (ResTT), to capture the multiple multilinear correlations of features, from low to high orders, within a single model. ResTT is able to build a robust decision boundary in a high-dimensional space for solving fitting and classification tasks. In particular, we prove that the fully connected layer and the Volterra series can be taken as special cases of ResTT. Furthermore, we derive the rule for weight initialization that stabilizes the training of ResTT based on a mean-field analysis. We prove that such a rule is much more relaxed than that of tensor train (TT), which means that ResTT can easily address the vanishing and exploding gradient problem that exists in the existing TT models. Numerical experiments demonstrate that ResTT outperforms the state-of-the-art tensor network and benchmark deep learning models on MNIST and Fashion-MNIST datasets. Moreover, ResTT achieves better performance than other statistical methods on two practical examples with limited data, which are known to have complex feature interactions.

Impact Statement—Optimal learning may require considering high-order multilinear feature interactions. Recently, quantum-inspired tensor networks have been proposed to exploit high-order correlations. However, the vanilla tensor networks are specialized in capturing only one type of multilinear feature interaction. This article proposes a novel network design which is able to describe generic feature interactions of different orders, within one model. Due to the residual connections, the training stability has also been promoted to an unprecedented level when compared to previous tensor network methods. Our methodology offers an alternative way of building a robust and efficient model for data deficient tasks which may have complex feature interactions.

Index Terms—Artificial neural networks, classification and regression, machine learning.

I. INTRODUCTION

MULTILINEAR feature correlations play an important role in machine learning [1]–[3]. Linear methods, such as linear regression (LR) [4], leverage linear correlation to solve simple machine learning problems. However, a single linear correlation may underfit the data that contain complex underlying patterns. In [5], a bilinear model has been utilized to separate the style and content of handwritten digits. Since then, low-order feature interactions, including bilinear and trilinear correlations, have been widely adopted to improve the capability of modeling the multilinearity hidden in data [6]–[9]. For high-order feature interactions, due to the dimensionality issue that leads to expensive computational and memory costs, the modeling of multilinear correlations is still a challenging problem.

Recently, quantum-inspired machine learning has drawn much attention and demonstrated potential applicability in computer vision [10]–[12], natural language processing [13]–[15], and recommendation system [16]. Increasing efforts [10], [17]–[20] have been made to build multilinear model based on tools that were originally developed for studying many-body quantum systems, such as tensor train (TT) [21], [22]. In the TT model, input features are embedded into a high-dimensional Hilbert space by tensor product, in which a simple decision boundary may correspond to a highly complex one in the original low-dimensional feature space. Thus, by defining a trainable high-order weight tensor, multilinear interactions can be modeled by simply multiplying the embedded input with this tensor [see Fig. 1(a)]. In order to overcome the dimensionality issue, a one-dimensional chain of low-order tensors, namely TT, can be used to approximate the high-order weight tensor [see Fig. 1(b)] to significantly reduce the number of trainable parameters. Particularly, for \( N \) input features, the vanilla TT model captures \( N \)-order multilinear correlations. Nevertheless, linear, bilinear, and higher order correlations may exist simultaneously among features, and thus, a more generalized model needs to be built to cover these scenarios.

In this article, we propose a residual tensor train (ResTT), which is capable of modeling multiple multilinear feature interactions within a single model. The main innovation of ResTT is
to incorporate skip connections [23] into the present TT model. By designing the network topology of skip connections, the ResTT can flexibly model diverse types of feature interactions. The main advantages of ResTT and contributions of this article are summarized as follows.

1) A novel multilinear model is proposed. ResTT is able to learn multiple multilinear feature correlations in a data-driven way. Besides, ResTT can flexibly model a diverse set of feature interactions by simply modifying the skip connections. In particular, by designing the network topology of skip connections, ResTT can be reduced to the fully connected layer [24] and Volterra series [25]. In other words, the fully connected layer and Volterra series can be looked as special cases of ResTT.

2) The mean-field analysis [26] is extended to TT-based models for deriving the rules of initialization that stabilize the training process. In particular, ResTT is demonstrated to be immune to the gradient vanishing and exploding problem with a much more relaxed training stability condition than that of the previous TT [16]. We find that the variance of input features will disturb the training of vanilla TT, and a strict initial condition has to be satisfied to ensure its stability. In contrast, the stability condition for ResTT is significantly relaxed, and there is no need to impose strict restrictions on the statistics of the input features.

3) Comprehensive experiments are conducted on the synthetic dataset, the image classification tasks, including MNIST [27] and Fashion-MNIST [28], and two practical tasks. The results on MNIST and Fashion-MNIST demonstrate that ResTT consistently outperforms the state-of-the-art tensor network models. In particular, the convergence of ResTT is faster and more stable than the plain TT. Inspired by the experimental setup in [3], ResTT is also compared with the benchmark deep learning models on limited training data (1%, 5%, and 10%) and has demonstrated superior performance in both prediction accuracy and training robustness. We focus on this comparison since limited training data is reported as a major issue in many practical tasks, in which data acquisition is expensive and the lack of training samples may easily cause overfitting [29]. Besides, ResTT achieves the best performance on Boston Housing [30] and average localization error (ALE) [31] datasets as compared to other statistical algorithms, which further demonstrates its advantage in modeling complex feature interactions with limited training data.

The rest of this article is organized as follows. Section II provides a brief introduction to the background and related works. In Section III, we introduce ResTT in detail. Section IV presents the mean-field analysis for TT and ResTT. Section V presents the experimental results on the synthetic and real datasets. Finally, Section VI concludes this article.

II. PRELIMINARIES

In this section, we first introduce the notation and definitions used in this article, and then, we review the TT approach for supervised learning tasks [10]. Finally, the related works are discussed.

A. Notation and Definition

Definition 1 (Tensor): Tensor, also known as multidimensional or N-mode array, is a generalization of vector (one index) and matrix (two indices) to an arbitrary number of indices. In this article, we denote tensors by calligraphic capitals. An N-order tensor is denoted by $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$, in which $I_n$ ($1 \leq n \leq N$, $n \in \mathbb{N}^+$) is the dimension of the $n$th index. The element of $\mathcal{A}$ is denoted by $a_{i_1, i_2, \cdots, i_N}$ ($1 \leq i_n \leq I_n, I_n \in \mathbb{N}^+$).

The diagrammatic notation for a tensor is drawn as a circle with edges, in which the circle represents the elements of the tensor and the edges represent its each individual index. For example, the graphics shown in Fig. 2(a) represent vector (first-order tensor), matrix (second-order tensor), third-order tensor and $N$th-order tensor, respectively.

Definition 2 (Tensor contraction): Tensor contraction is an operation that combines two or more tensors into a new one. In this article, we denote tensor contraction by $C_{I_k} [\cdot]$, where $I_k$ is the index to be contracted. For example, a matrix $\mathcal{M} = \{ m_{i_1, i_2} \} \in \mathbb{R}^{I_1 \times I_2}$ and a third-order tensor $\mathcal{T} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ can...
be contracted along the index $I_1$ as
\[ C = C_{I_1} [M, T] \tag{1} \]
with $C \in \mathbb{R}^{I_2 \times J_1 \times I_2}$ being the new third-order tensor. The elements of $C$ are given by
\[ c_{i_2 j_1 i_2} = \sum_{i_1} m_{i_1 i_2} t_{j_1 i_2 i_1}. \tag{2} \]
Note that the tensors to be contracted must have one or more compatible indices of the same dimension, such as the index $I_1$ in $M$ and $T$.

In the graphical notation, a contraction is represented by a sharing edge, which indicates that the two tensor nodes are contracted along this particular index.

**Definition 3 (Tensor train):** TT is a set of tensors to be contracted in a one-dimensional mode. Specifically, $N$ tensors denoted by \{ $W^{(1)} \in \mathbb{R}^{I_1 \times V_1}, W^{(2)} \in \mathbb{R}^{V_1 \times I_2 \times V_2}, \ldots, W^{(N)} \in \mathbb{R}^{V_{N-1} \times I_N}$ \} constitute a TT if contracted as
\[ W = C_{V_1, \ldots, V_N} [W^{(1)}, W^{(2)}, \ldots, W^{(N)}] \tag{3} \]
where $W \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ is the generated $N$th-order tensor and $V_1, V_2, \ldots, V_{N-1}$ are called virtual bond dimensions.

The graphical representation of the contraction process of a four-node TT is drawn in Fig. 2(b). It is worth mentioning that the virtual bond dimensions could be used to control the total number of parameters.

**Definition 4 (Feature interaction):** In statistics, feature interaction [32] describes a situation in which the effect of one feature on the outcome depends on the state of other features. It is common to use products of features, also known as multilinear feature correlations, to represent different types of feature interaction. The number of features in the product term is defined as the order of interaction. For example, a bilinear model with an output $y$ and two input features $x_1$ and $x_2$ can be formulated as
\[ y = c \cdot x_1 x_2 \tag{4} \]
where $c \cdot x_1 x_2$ describes the second-order interaction between $x_1$ and $x_2$.

### B. Multilinear Learning Approaches

Linear methods have a wide range of applications [33]. However, a linear model may not be sufficient to describe the correlation between features for complicated tasks, and thus, multilinear methods have attracted increasingly attention in recent years. Early studies of multilinear models mainly focused on bilinear [5] and trilinear correlations [34]. Recently, low-order feature interactions have been adopted into deep learning models. The work in [35] utilized the bilinear correlation to construct a novel neural network layer, which provides a powerful way to describe relational information than the standard fully connected layer. Combined with convolutional layers, such a layer has shown to produce the output tensor $Y^{(1)}$ in $\mathbb{R}^{V_1}$. The second weight tensor $W^{(2)}$ and the second input vector $X^{(2)}$ are contracted to generate an intermediate matrix $A^{(2)} \in \mathbb{R}^{V_1 \times V_2}$, which is then contracted with $Y^{(1)} \in \mathbb{R}^{V_1}$ to generate the second output tensor.

**C. Tensor Train Approach**

TT approach [10] provides an effective way to model high-order feature interactions. Given a set of feature vectors \{ $X^{(1)} \in \mathbb{R}^{I_1}, X^{(2)} \in \mathbb{R}^{I_2}, \ldots, X^{(N)} \in \mathbb{R}^{I_N}$ \}, the network model is defined as
\[ Y = C_{I_1, \ldots, I_N} [W, X^{(1)}, X^{(2)}, \ldots, X^{(N)}] \tag{5} \]
where $W \in \mathbb{R}^{I_1 \times \cdots \times I_N \times O}$ is a general $N$th-order weight tensor. Note that the number of trainable parameters scales exponentially as $O(\prod_{n=1}^{N} I_n)$. To reduce the computation cost, $W$ can be approximated by a set of low-order tensors, which is further contracted according to (3). The contraction of TT is equivalent to the feedforward process in neural networks. For example, the red arrows in Fig. 3 indicate a forward direction of TT. The input vector $X^{(1)}$ is contracted with the weight tensor $W^{(1)}$ to produce the output tensor $Y^{(1)} \in \mathbb{R}^{V_1}$. The second weight tensor $W^{(2)}$ and the second input vector $X^{(2)}$ are contracted to generate an intermediate matrix $A^{(2)} \in \mathbb{R}^{V_1 \times V_2}$, which is then contracted with $Y^{(1)} \in \mathbb{R}^{V_1}$ to generate the second output tensor.

![Fig. 3. TT model that is contracted from left to right. The outputs after the contractions are denoted in a blue font.](image-url)
Fig. 4. Graphical representation of ResTT. (a) Pipeline of the general ResTT. At each summing junction, two shortcuts (in red line) are added to the output. The arrows indicate the contraction direction. (b) and (c) ResTT forms of an $N$-node fully connected layer and the discrete Volterra series up to the $N$th order, respectively. (d) Example of ResTT with three inputs and its equivalent form.

$$\mathbf{y}^{(2)} \in \mathbb{R}^{V_2}$$. The final output is denoted by $\mathbf{y}^{(N)} \in \mathbb{R}^{O}$, whose elements can be written as

$$y^{(N)}_o = \sum_{v_1 \cdots v_{N-1} v_N} w^{(1)}_{v_1 v_1} x^{(1)}_{v_1} \cdots w^{(N)}_{v_N v_N} x^{(N)}_{v_N}$$

which clearly encodes the $N$th-order interactions among the input features, without low-order terms.

TT can be optimized by gradient descent with respect to the mean squared error. In [10], two nearby tensors are contracted to form a new high-order tensor at each step, and then, the gradient is calculated by taking the derivative of the cost function with respect to this tensor. After updating the parameters, a singular value decomposition is performed to restore the updated tensor into two low-order tensors. This method sweeps back and forth along the TT and iteratively minimizes the cost function. Notably, at each step, all of the tensor nodes must be contracted, which may involve hundreds of multiplications, causing the gradient vanishing or exploding problem. Several procedures must be taken to counteract this effect, e.g., normalizing the input and output at each step, dividing the bond tensor with its largest element, etc. This algorithm has been implemented in the TensorNetwork library [19], which uses TensorFlow as the back end for speeding up the tensor computations.

III. RESIDUAL TENSOR TRAIN

Residual structure is widely used in the design of deep neural networks. In ResNet, the identity mapping enabled by skip connections allows the input signal to propagate directly from one layer to any other layer, which solves the problem of network degradation to a certain extent [41]. Besides, the residual network can be regarded as an ensemble model that consists of a series of path sets [42], thereby improving the robustness of the model. Similar to ResNet, the ResTT is proposed to solve the following issues.

1) The plain TT only models the $N$th-order interaction between all the input features and ignores the low-order interactions between a subset of features, such as linear or bilinear correlations. More importantly, the plain TT lacks the flexibility in designing the combination of feature interactions with the required orders.

2) The current TT models are difficult to train with a large number of tensor nodes. As residual connection has been proven to be effective in stabilizing the training process for deep neural networks, it can also be added to the TT network to improve its training stability.

A. General Model

The diagrammatic representation of a general ResTT model is shown in Fig. 4(a), in which the contractions are conducted from
left to right. First, $\mathcal{X}^{(1)}$ is contracted with $\mathcal{W}^{(1,1)}$ to generate the output $\mathcal{Y}^{(1)} \in \mathbb{R}^{V}$ by

$$
\mathcal{Y}^{(1)} = C_{I_x}[\mathcal{X}^{(1)}, \mathcal{W}^{(1,1)}].
$$
(7)

Then, $\mathcal{X}^{(2)}$ is contracted with $\mathcal{Y}^{1}$ and $\mathcal{W}^{(2,1)}$ to generate the normal output $\mathcal{Y}^{(2,1)} \in \mathbb{R}^{V_2}$ by

$$
\mathcal{Y}^{(2,1)} = C_{V_1, I_2}[\mathcal{Y}^{(1)}, \mathcal{W}^{(2,1)}, \mathcal{X}^{(2)}].
$$
(8)

The virtual bond dimension of the TT is assumed to be the same, i.e., $V_1 = V_2 = \cdots = V_{N-1} = r$, and thus, we can directly add the skip connection $\mathcal{Y}^{(1)}$ to the output $\mathcal{Y}^{(2,1)}$. Besides, $\mathcal{X}^{(2)}$ is contracted with an additional node $\mathcal{W}^{(2,2)} \in \mathbb{R}^{I_2 \times V_2}$ to generate the low-order interaction terms that include the features in the input $\mathcal{X}^{(2)}$, according to the following formula:

$$
\mathcal{Y}^{(2,2)} = C_{I_2}[\mathcal{X}^{(2)}, \mathcal{W}^{(2,2)}].
$$
(9)

The output of the second layer of ResTT is given by

$$
\mathcal{Y}^{(2)} = \mathcal{Y}^{(2,1)} + \mathcal{Y}^{(1)} + \mathcal{Y}^{(2,2)}.
$$
(10)

Note that the skip connection $\mathcal{Y}^{(1)}$ produces the low-order terms that include features in the input $\mathcal{X}^{(1)}$. By repeating this process, we obtain

$$
\mathcal{Y}^{(n,1)} = C_{V_{N-1}, I_n} [\mathcal{Y}^{(n-1)}, \mathcal{W}^{(n,1)}, \mathcal{X}^{(n)}]
$$
$$
\mathcal{Y}^{(n,2)} = C_{I_n} [\mathcal{X}^{(n)}, \mathcal{W}^{(n,2)}]
$$
$$
\mathcal{Y}^{(n)} = \mathcal{Y}^{(n,1)} + \mathcal{Y}^{(n-1)} + \mathcal{Y}^{(n,2)}.
$$
(11)

For the last node, a linear mapping is defined by

$$
\mathcal{Y}^{(N,3)} = C_{V_{N-1}} [\mathcal{W}^{(N,3)}, \mathcal{Y}^{(N-1)}]
$$
(12)

with $\mathcal{W}^{(N,3)} \in \mathbb{R}^{V_{N-1} \times O}$, and the final output of ResTT can be written as

$$
\mathcal{Y}^{(N)} = \mathcal{Y}^{(N,1)} + \mathcal{Y}^{(N,3)} + \mathcal{Y}^{(N,2)}.
$$
(13)

The detailed expression of $\mathcal{Y}^{(N)}$ is given in Appendix A. From the detailed expression, we can see that the output of a general ResTT contains the interaction terms from first order (linear) to $N$th order ($N$-linear).

### B. Special Cases

It is clear from (11) that different connections at each layer generate interaction terms with different orders of feature interactions. Therefore, it is convenient to modify the combination of multilinear terms contained in the final output by adding or deleting the specific connections between the tensor nodes. In particular, many famous models are found to be special cases of the general ResTT. For example, if only the residual connections shown in Fig. 4(b) are kept, ResTT will degrade to an $N$-node fully connected layer as

$$
\mathcal{Y}^{(N)} = \sum_{n=1}^{N} C_{I_n} [\mathcal{X}^{(n)}, \mathcal{W}^{(n,2)}].
$$
(14)

In this case, the model captures only the linear correlations.

Meanwhile, if we assume that the input and output are scalars, i.e., $\mathcal{Y}^{(N)} \in \mathbb{R}$, the general ResTT will degrade to the famous discrete Volterra series [25] when the residual connections in the upper half of Fig. 4(a) are deleted [see Fig. 4(c)]. The resulting Volterra series is given by

$$
y = \sum_{i_N} h_{i_N}^{(N)} x_{i_N}^{(N)} + \sum_{i_N \neq i_{N-1}} h_{i_N, i_{N-1}}^{(N)} x_{i_N}^{(N)} x_{i_{N-1}}^{(N-1)} + \cdots
$$
$$
+ \sum_{i_1, \ldots, i_N} h_{i_1, \ldots, i_N}^{(1)} \prod_{j=1}^{N} x_{i_j}^{(j)}
$$
(15)

with

$$
h_{i_1}^{(N)} = w_{i_1}^{(N,2)}
$$
$$
h_{i_1, i_2}^{(N-1)} = \sum_{i_3, \ldots, i_{N-1}} w_{i_1, i_2, i_3, \ldots, i_{N-1}}^{(N-1)} w_{i_3, \ldots, i_{N-1}}^{(N-1)}
$$
$$
\vdots
$$
$$
\sum_{i_1, \ldots, i_N} h_{i_1, \ldots, i_N}^{(1)} = \sum_{i_1, \ldots, i_N} w_{i_1}^{(1,1)} w_{i_2}^{(2,1)} \cdots w_{i_N}^{(N,1)}
$$
(16)

being the tensorization form of the elements of Volterra kernels denoted by $\mathcal{H}^{(N)} \in \mathbb{R}^{L_N}$, $\mathcal{H}^{(N-1)} \in \mathbb{R}^{L_N \times L_{N-1}}$, $\ldots$, $\mathcal{H}^{(1)} \in \mathbb{R}^{L_1 \times L_2 \times \cdots \times L_N}$, where the $k$th-order tensor $\mathcal{H}^{(k)}$ is called the $k$th Volterra kernel.

In the last, we provide a simple example to further illustrate the flexibility of ResTT. As shown in Fig. 4(d), the final output of ResTT can be written as

$$
y = w^{(1)} x^{(1)} + w^{(2,2)} x^{(2)} + w^{(3,2)} x^{(3)}
$$
$$
+ w^{(1)} w^{(2,1)} x^{(1)} x^{(2)} + w^{(1)} w^{(3,1)} x^{(1)} x^{(3)}
$$
$$
+ w^{(2,2)} w^{(3,1)} x^{(2)} x^{(3)} + w^{(1)} w^{(2,1)} w^{(3,1)} x^{(1)} x^{(2)} x^{(3)}.
$$
(17)

Suppose that the skip connection of $y^{(1)}$ [the orange line in Fig. 4(d)] has been removed. Then, the output of ResTT will change to

$$
y = w^{(1)} x^{(1)} + w^{(2,2)} x^{(2)} + w^{(3,2)} x^{(3)}
$$
$$
+ w^{(1)} w^{(2,1)} x^{(1)} x^{(2)} + w^{(2,2)} w^{(3,1)} x^{(2)} x^{(3)}
$$
$$
+ w^{(1)} w^{(2,1)} w^{(3,1)} x^{(1)} x^{(2)} x^{(3)}
$$
(18)

in which the second-order interaction between the input features $x^{(1)}$ and $x^{(3)}$ is discarded.

### C. Optimization

Since the operation of tensor contraction is similar to the linear transformation layer of conventional neural networks, we can adopt the standard backpropagation algorithm [43] to optimize ResTT. Here, we briefly show the gradient calculation of the weight tensor $\mathcal{W}^{(k,1)}$ as follows:

$$
\Delta \mathcal{W}^{(k,1)} = - \frac{\partial c}{\partial \mathcal{W}^{(k,1)}}
$$
$$
= - \frac{\partial c}{\partial \mathcal{Y}^{(N)}} \frac{\partial \mathcal{Y}^{(N)}}{\partial \mathcal{Y}^{(N-1)}} \frac{\partial \mathcal{Y}^{(N-1)}}{\partial \mathcal{Y}^{(N-2)}} \cdots \frac{\partial \mathcal{Y}^{(k)}}{\partial \mathcal{W}^{(k,1)}}
$$

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Algorithm 1: ResTT for Supervised Learning.

Input: Features $X^{(1)} \in \mathbb{R}^{I_1}, X^{(2)} \in \mathbb{R}^{I_2}, \ldots, X^{(N)} \in \mathbb{R}^{I_N}$; Label $Y \in \mathbb{R}^{O}$.

Initialize: The number of epochs $K$; Weight tensors $W^{(1)} = \mathbb{R}^{I_1 \times V_1}$, $W^{(2)} = \mathbb{R}^{V_1 \times I_2 \times V_2}$, $W^{(3)} = \mathbb{R}^{I_2 \times V_2}$, \ldots, $W^{(N-1)} = \mathbb{R}^{V_{N-1} \times I_N \times O}$, $W^{(N-2)} = \mathbb{R}^{I_N \times O}$, $W^{(N)} = \mathbb{R}^{V_{N-1} \times O}$; $W^{(N)}(1) = C_{I_1} [\chi^{(1)}, W^{(1)}]$.

1: for $k = 1, \ldots, K$ do
2: \quad $Y^{(1)} = C_{I_1} (X^{(1)}, W^{(1)})$.
3: for $n = 2, \ldots, N$ do
4: \quad $Y^{(n)} = C_{V_{n-1} \times I_n \times V_n} [Y^{(n-1)}, W^{(n-1)}, X^{(n)}]$.
5: \quad $Y^{(n)}(2) = C_{V_{n-1} \times I_n} [\chi^{(n)}, W^{(n-2)}]$.
6: if $n \neq N$ then
7: \quad $Y^{(n)}(3) = C_{V_{n-2} \times I_n \times O} [Y^{(n-1)}(2), W^{(n)}]$. 
8: end if
9: end for
10: $Y^{(N)}(3) = C_{V_{N-1} \times O} [Y^{(N-1)}(2), Y^{(N-2)}(1)]$.
11: $Y^{(N)} = Y^{(N)}(3) + Y^{(N-2)}(1) + Y^{(N)}(2)$.
12: Calculate the cost function: $c(Y^{(1)}, Y^{(N)})$.
13: Update the weight tensors by backpropagation.
14: end for

\[ = - \frac{\partial c}{\partial \chi^{(N)}} \cdot (I + A^{(N)}) \cdots \frac{\partial Y^{(k)}}{\partial W^{(k-1)}} \]

\[ = - \frac{\partial c}{\partial \chi^{(N)}} \cdot (I + A^{(N)}) \cdots (I + A^{(k+1)}) \]

\[ \otimes \chi^{(k)} \otimes Y^{(k-1)} \]

D. Complexity Analysis

Here, we compare the memory complexity and time complexity of three multilinear models, including the general tensorized model in (5), the plain TT in (3), and ResTT. The memory complexity of (5) is $O(\prod_{n=1}^{N} I_k)$. In contrast, the memory complexities of TT and ResTT are $O(\sum_{n=1}^{N} I_n r^2)$ and $O(\sum_{n=1}^{N} (I_n r^2 + I_n r^3))$, respectively. The time complexities for TT and ResTT are $O(\sum_{n=1}^{N} I_n r^3)$ and $O(\sum_{n=1}^{N} (I_n r^3 + I_n r))$, respectively, while (5) has a complexity of $O(\prod_{n=1}^{N} I_k)$.

IV. MEAN-FIELD ANALYSIS

Mean-field analysis [26], [44] provides an efficient way to assess the training stability of a deep neural network. In this article, we extend this analysis to study the characteristics of signal propagations in TT and ResTT. First, we assume that all the weights in TT and ResTT are initialized from the same Gaussian distribution $N(0, \sigma_w^2/r)$; then, we calculate the evolution of the variances of the outputs $\{Y^{(k)}\}$ and the average magnitude of the gradients. The detailed derivations are given in Appendix B.

A. Tensor Train

As all the weights are independent identically distributed with zero mean, the output $Y^{(k)}$ satisfies

\[ \mathbb{E}(y^{(k)}_{w_{ik}}) = 0 \]  

and the variance of the output, denoted by $q^{(k)}$, is calculated as

\[ q^{(k)} = \mathbb{E} \left( \sum_{i_k} (x^{(k)}_{i_k})^2 \right) \sigma_w^2 q^{(k-1)} = s^{(k)} q^{(k-1)} \]

with $s^{(k)} = \mathbb{E} \left( \sum_{i_k} (x^{(k)}_{i_k})^2 \right) \sigma_w^2$. Hence, the relation between the input and output variances at each layer is characterized by the slope factor $s^{(k)}$. As can be seen from (23), the slope factor is influenced by the variance of the input data, which is significantly different from the conventional deep learning model where the evolution of the output at each layer can only be influenced by the variance of the weights. In addition, the backward propagation in TT is studied by calculating the input–output Jacobian, which relates the gradients to the weight tensor at a given position. The Jacobian for the input $Y^{(k-1)}$ is denoted by $J^{(k)} \in \mathbb{R}^{V_{k-1} \times I_k \times V_k}$ and calculated as

\[ J^{(k)} = \frac{\partial Y^{(N)}}{\partial Y^{(N-1)}} \cdots \frac{\partial Y^{(k)}}{\partial Y^{(k-1)}} = A^{(N)} \cdots A^{(k+1)} \otimes \chi^{(k)} \]

with the intermediate matrix $A^{(k)}$ defined in (20). The quadratic mean $\chi^{(k)}$ of the elements in the Jacobian characterizes the magnitude of the gradients in backpropagation, which is given by

\[ \chi^{(k)} = s^{(k)} \chi^{(k+1)} \].

According to (23) and (24), the signals in the forward and backward propagations will explode in a long TT when $s^{(k)} > 1$. In contrast, if $s^{(k)} < 1$, the signals will tend to vanish. Therefore, the condition of stability for the training is given by the following
formula:

\[ s(k) = 1, \quad k \geq 1. \]  \hspace{1cm} (25)

Since \( s(k) \) is determined by the input data, we have to preprocess the input in order to stabilize the training. For example, by employing the trigonometric mapping \([45]\) to the input and transforming the feature vector \( \mathcal{X}^{(k)} \) as

\[ \phi(\mathcal{X}^{(k)}) = \frac{1}{\sqrt{I_k}} [\cos \left( \frac{\pi}{2} \mathcal{X}^{(k)} \right), \sin \left( \frac{\pi}{2} \mathcal{X}^{(k)} \right)] \]  \hspace{1cm} (26)

we have \( \mathbb{E}(\sum w (x_i^{(k)})^2) = 1 \). Then, (25) can be satisfied by letting \( \sigma_w^2 = 1 \). In this case, we get

\[ q^{(k)} = (s^{(k-1)}), \quad \chi^{(k)} = \chi^{(k+1)} \]  \hspace{1cm} (27)

which is the critical condition for the stable training of TT.

B. Residual Tensor Train

In ResTT, the \( k \)th output vector also satisfies \( \mathbb{E}(Y^{(k)}_{vk}) = 0 \), and the variance of this vector is calculated as

\[ q^{(k)} = (s^{(k)} + 1)q^{(k-1)} + s^{(k)}. \]  \hspace{1cm} (28)

In backward propagation, the quadratic mean of the input–output Jacobian for the input \( Y^{(k-1)} \) is calculated as

\[ \chi^{(k)} = (s^{(k-1)})\chi^{(k+1)} \]  \hspace{1cm} (29)

which shares the same slope factor as the forward propagation. Hence, \((s^{(k)} + 1)\) determines the evolution of the signals in the training of ResTT. Note that for any choice of \( \sigma_w^2 \), \( (s^{(k)} + 1) \) is always larger than 1, which means that the signals in the forward and backward propagation will never vanish. Meanwhile, in order to prevent the signals from exploding, the following condition has to be satisfied:

\[ s^{(k)} \ll 1. \]  \hspace{1cm} (30)

According to (23), as long as \( \sigma_w^2 \ll 1 \), the condition of stability (30) will be satisfied given that the variance of the input data is bounded. Compared to TT, which requires \( \mathbb{E}(\sum w (x_i^{(k)})^2)\sigma_w^2 = 1 \), the condition \( \sigma_w^2 \ll 1 \) is significantly relaxed.

V. EXPERIMENTS

We have conducted comprehensive experiments on the synthetic dataset, two image classification datasets, and two practical examples with limited data. In terms of the experimental setup, we optimize the models by the Adam optimizer \([46]\). The learning rate is chosen from \( \{1 \times 10^{-2}, 1 \times 10^{-3}, 1 \times 10^{-4}\} \). All the models are trained for 100 epochs. We fix the mini-batch size as 512 and set the weight decay as \( 1 \times 10^{-6} \) for all tasks. The implementations are based on Pytorch \([47]\), and the models are trained on a single NVIDIA GTX 1080Ti.

A. Synthetic Dataset

In order to compare the capability of ResTT in modeling multiple multilinear feature correlations, we generate a synthetic dataset by

\[ \hat{y} = \sum_{i=1}^{d} w_i^{(1)} x_i^{(1)} + \sum_{i,j=1}^{d} w_{ij}^{(2)} x_i^{(1)} x_j^{(2)} \]

\[ + \sum_{i,j,k=1}^{d} w_{ijk}^{(3)} x_i^{(1)} x_j^{(2)} x_k^{(3)} \]  \hspace{1cm} (31)

where the elements of \( \mathcal{W}^{(1)} \in \mathbb{R}^d, \mathcal{W}^{(2)} \in \mathbb{R}^{d \times d} \), and \( \mathcal{W}^{(3)} \in \mathbb{R}^{d \times d \times d} \) are randomly sampled from \( \mathcal{N}(0, 0.1) \), and \( \mathcal{X}^{(1)}, \mathcal{X}^{(2)}, \mathcal{X}^{(3)} \in \mathbb{R}^d \) are randomly sampled from \( \mathcal{N}(0, 0.5) \). The training sizes are 10 000 and 50 000, while the test size is 10 000 for all the cases. The feature dimension \( d \) is selected from \( \{1, 10, 20, 30, 40\} \). We conduct experiments on TT in (6) with \( r = 20 \), ResTT with \( r = 20 \), and the LR model. Each model has been run five times to average the effect of random initialization.

We use the root-mean-square error (RMSE) as the evaluation metric. The results are shown in Fig. 5. It is clear that TT performs better than LR, with its ability to model high-order feature interaction instead of the low-order ones. Meanwhile, the performance of TT is worse than that of ResTT, since ResTT can model a comprehensive set of feature interactions. Fig. 5 also shows that the advantage of ResTT over TT and LR becomes more dominant with the increase of feature dimension \( d \).

B. Image Classification

In line with the previous works \([10], [19], [40]\), we first use the MNIST and Fashion-MNIST datasets to evaluate the classification performance of ResTT. However, it should be noted that the intended application of the current ResTT is not for large-scale image datasets. On the one hand, images have to be flattened into a list as the input to the current ResTT, which makes it difficult for ResTT to outperform convolutional neural networks (CNNs) that take advantage of spatial correlations. On the other hand, the current ResTT is an extended linear model without using any advanced network components, such as nonlinear activation, which makes it a nonfair competition.
between ResTT and other deep learning models on large-scale datasets.

1) Datasets:
   a) **MNIST**: The MNIST is a dataset of handwritten digits from 0 to 9, which contains 60,000 training samples and 10,000 testing samples with $28 \times 28$ grayscale pixels. Each pixel takes integer in the range [0, 255]. Following [10], each pixel value of the image is normalized to [0, 1] and mapped into a 2-D space by a trigonometric function defined in (26). Hence, each feature of this task is a 2-D vector. Each image is scaled down to $14 \times 14$ by averaging the clusters of four adjacent pixels and then flattened into a list to adapt to the structure of TT.
   b) **Fashion-MNIST**: The Fashion-MNIST is a dataset of Zalando’s article images, which consists of a training set of 60,000 samples and a test set of 10,000 samples. Each image is a $28 \times 28$ grayscale image, associated with a label from ten classes. Similar to MNIST, each pixel value is normalized to [0, 1] and mapped to a 2-D space by the trigonometric function. Again, each image is scaled down to $14 \times 14$ by averaging the clusters of four adjacent pixels and then flattened into a list to adapt to the structure of TT.

2) Models: The following models based on tensor network are tested in the experiments:
   a) **MPS** [10]: MPS is a special TT model in which all the tensor nodes are kept unitary and trained by the DMRG-like optimization algorithm.
   b) **TTN** [18]: TTN is a two-layer tensor network model that supports hierarchical feature extraction, which uses a training algorithm derived from the multipartite entanglement renormalization ansatz.
   c) **EPS-SBS** [40]: Entangled plaquette state (EPS) is defined as a product of tensors on overlapping clusters of variables and string-bond state (SBS) is defined by placing MPS over strings on a graph, which needs not to be a one-dimensional lattice. EPS-SBS is a tensor network that consists of an EPS followed by an SBS and optimized by stochastic gradient descent.
   d) **Snake-SBS** [40]: Snake-SBS consists of four overlapping SBS structures in a snake pattern and is optimized by stochastic gradient descent.
   e) **TT-BN**: TT with batch normalization (BN). In this model, a BN layer [48] is added after each contraction. TT-BN is trained by backpropagation.
   f) **MeanTT**: Mean-field TT is the vanilla TT model, which is initialized using the stability condition derived in Section IV-A. MeanTT is trained by backpropagation.

3) Results: Table I shows that the additional connections in the ResTT do not introduce a significant amount of parameters, which means that TT and ResTT can be compared with the same virtual bond dimension $r$. This is because most of the parameters are contained in the high-order tensors generated by \{V^{(n,1)}\}, while the additional parameters are contained in \{W^{(n,2)}\}, which only generate low-order tensors. The test accuracies on MNIST and Fashion-MNIST datasets are presented in Table II. It can be observed that ResTT outperforms all the state-of-the-art tensor network models on both MNIST and Fashion-MNIST datasets. In particular, MeanTT achieves better test accuracy than the regularized MPS and TT-BN, which indicates that proper initialization of the weights can lead to improved performance without any regularization techniques. The inference and convergence time for ResTT with different virtual bond dimensions on MNIST dataset are given in Table III. Both IT and BT are less than 1 s. The increase in $r$ leads to an increase in IT and BT. Meanwhile, CT decreases as $r$ increases, unless $r$ is too large.

Next, we conduct multiple sets of experiments by choosing $r$ in \{20, 50, 100\} and varying the fraction of data for training by \{1\%, 5\%, 10\%, 20\%, 50\%, 100\%\}. Each model has been run ten times to average the effect of random initialization. The mean and standard deviation of the test accuracy are reported in Table IV. When the fraction is 1\%, that is, only 600 randomly selected samples are used for training, ResTT can achieve an error rate of 10.18\% on the test set, which contains 10,000 test samples. Moreover, it is worth mentioning that ResTT converges much faster than MeanTT. As can be seen from Fig. 6, the training of MeanTT with $r = 20$ is highly unstable on a fraction.
TABLE IV

| Dataset       | Method | $r$ | Parm. | Training 1% | Training 5% | Training 10% | Training 20% | Training 50% | Training 100% |
|---------------|--------|-----|-------|-------------|-------------|--------------|---------------|--------------|--------------|
| MNIST         | MeanTT | 20  | 0.15M | 30.16±14.59 | 37.03±14.95 | 48.53±26.75 | 68.14±20.70  | 70.01±33.17  | 38.54±33.14  |
|               | MeanTT | 100 | 3.88M | 85.87±0.6054 | 94.57±0.1331 | 96.20±0.1762 | 97.27±0.0721 | 98.36±0.0412 | 98.70±0.0474 |
|               | ResTT  | 20  | 0.16M | 89.52±0.2917 | 95.19±0.1088 | 96.49±0.0889 | 97.41±0.0902 | 98.34±0.0306 | 98.53±0.0397 |
|               | ResTT  | 100 | 3.92M | 89.82±0.2755 | 95.46±0.0768 | 96.74±0.0706 | 97.68±0.0687 | 98.66±0.0329 | 98.96±0.0304 |
| Fashion-MNIST | MeanTT | 20  | 0.15M | 22.77±0.071  | 30.72±0.5816 | 53.11±0.1682 | 64.67±0.1979 | 66.29±0.0804 | 37.44±0.1035 |
|               | MeanTT | 100 | 3.88M | 77.07±0.4257 | 82.72±0.2014 | 84.39±0.1861 | 85.92±0.1282 | 87.66±0.0986 | 88.79±0.0841 |
|               | ResTT  | 20  | 0.16M | 75.48±0.4445 | 82.11±0.2052 | 83.95±0.0925 | 85.45±0.0731 | 86.95±0.0695 | 87.95±0.0913 |
|               | ResTT  | 100 | 3.92M | 78.86±0.3120 | 84.18±0.1554 | 85.69±0.1116 | 87.16±0.1363 | 88.93±0.1132 | 89.85±0.1064 |

Fig. 6. Test accuracy of ResTT (blue curve) and MeanTT (orange curve) versus the training epochs on MNIST. (a) and (b) are the model performances with $r = 20$ and $r = 100$, respectively.

of data, while ResTT can converge to an approximately optimal point within just 20 epochs. In Fig. 6(b), ResTT with $r = 100$ takes only about five epochs to complete the training. According to the numerical results, ResTT consistently outperforms MeanTT in terms of the training stability and test accuracy, even if we have shown that MeanTT is already more stable in training than other tensor network models such as regularized MPS and TT-BN. Therefore, it is clear that the residual connections, with proper initialization guided by the mean-field analysis, can indeed improve the effectiveness and training stability of the model.

C. Image Classification With Limited Data

To further study the performance of ResTT and benchmark deep learning models in scenarios with limited data, we conduct experiments using a small amount of training data from MNIST and Fashion-MNIST.

1) Datasets: The fraction of training data for MNIST and Fashion-MNIST are selected from $\{1\%, 5\%, 10\%\}$. The training data are randomly sampled from the original training set. All the experiments are repeated ten times at each fraction to alleviate the effect of randomness.

2) Models: Models for comparison are listed as follows.
   a) LR: LR is a statistical model that captures the linear correlation of the input features.
   b) Multilayer perceptron (MLP): MLP is the feedforward neural network that consists of multiple fully connected layers (with threshold activation). With a layer-by-layer structure, MLP is able to distinguish data that are not linearly separable. The MLP here is adopted from [28], which is composed of two layers with 100 hidden neurons and the rectified linear unit (ReLU) activation.
   c) LeNet-5: LeNet-5 is the lightweight CNN that consists of two sets of convolutional and average pooling layers, followed by two fully connected layers and finally a Softmax layer.
   d) ResNet18: ResNet [23] is composed of multiple residual blocks and is one of the most influential CNN structure in deep learning literature. The ResNet18 refers to the 18-layer deep ResNet.
TABLE V
Comparisons of the Classification Accuracy (%) with Different Fractions of Training Data on the Test Sets of MNIST and Fashion-MNIST

| Fraction | Model  | MNIST     | Fashion-MNIST |
|----------|--------|-----------|---------------|
| 1%       | LR     | 9.8±0     | 10±0          |
|          | MLP    | 67.28±8.896| 10±0          |
|          | LeNet-5| 84.59±6.781| 75.90±0.8228 |
|          | ResNet18| 75.44±1.011| 75.66±0.3875 |
|          | ResTT  | 89.82±0.2755| 78.86±0.3120 |

The virtual bond dimension for ResTT is 100

TABLE VI
Comparisons of R² Score and RMSE on the Test Sets of Boston Housing and ALE

| Model      | Boston Housing | ALE          |
|------------|----------------|--------------|
|            | R² score | RMSE | R² score | RMSE |
| LR         | 0.654    | 4.61 | 0.715    | 0.181 |
| PR         | 0.779    | 3.68 | 0.591    | 0.218 |
| RR         | 0.664    | 4.55 | 0.719    | 0.183 |
| PRR        | 0.831    | 3.23 | 0.593    | 0.216 |
| LaR        | 0.673    | 4.48 | 0.721    | 0.182 |
| PLaR       | 0.848    | 3.05 | 0.619    | 0.209 |
| SVR        | 0.846    | 3.07 | 0.644    | 0.202 |
| MLP        | 0.829    | 3.47 | 0.704    | 0.201 |
| ResTT      | 0.864    | 2.96 | 0.736    | 0.179 |

The virtual bond dimension for ResTT is 50

3) Results: Results with various fractions of training data are shown in Table V. For LR and MLP, since the models are overly simple, limited training data may result in underfitting and nonconvergence. It is clear that the performance of ResTT is significantly better than LeNet-5 and ResNet18 in terms of prediction accuracy and training robustness for different fractions. Particularly, for the fraction of 1%, ResTT also beats the multitask learning methods [3] on MNIST, including the deep multitask representation learning model based on Tensor Train (DMTRL-TT) (∼87%), DMTRL-Tucker (∼83%) and DMTRL-LAF (∼85%).

D. Practical Examples With Limited Data

In this subsection, we demonstrate the application of ResTT on real tasks with limited and sequential data, which are known to have complex feature interactions. Particularly, the experiments are conducted on Boston Housing and ALE datasets, where ResTT is compared with other popular methods.

1) Datasets:
   a) Boston Housing: The Boston Housing dataset is collected by the U.S. Census Service concerning housing in the area of Boston Mass. The task is to use 13 variables, such as crime rate and tax rate, to predict Boston housing prices. The dataset is small with only 506 samples. We use 354 samples (70%) for training and the rest (152 samples) for testing.
   b) ALE: The dataset is to predict the ALE with applications to wireless sensor networks, which consists of four features, namely, anchor ratio, transmission range of a sensor, node density, and iteration count. The dataset is small with 107 samples. We use 74 samples (70%) for training and the rest (33 samples) for testing.

2) Models: Models for comparison are listed as follows.

a) LR: LR is a statistical model that captures the linear correlation of the input features.

b) Polynomial regression (PR): PR is a statistical model that captures the polynomial correlations of the input features. We use the quadratic model, which exploits quadratic interaction terms of the input features for modeling.

c) Ridge regression (RR): RR is a special LR model constrained by L2 regularization.

d) Polynomial ridge regression (PRR): PRR is a polynomial RR model. Here, we use the quadratic model.

e) Lasso regression (LaR): LaR is a special LR model constrained by L1 regularization.

f) Polynomial Lasso regression (PLaR): PLaR is a polynomial LaR model. Here, we use the quadratic model.

g) Support vector regression (SVR): SVR is a kernel-based regression method, which aims to learn a regression hyperplane in high-dimensional feature space [49]. Here, we adopt the radial basis function kernel.

h) MLP: The MLP here is composed of two layers with 100 hidden neurons and the ReLU activation.

3) Results: We adopt the RMSE and R² score [4] as the evaluation metrics. Experimental results are shown in Table VI. We see that ResTT achieves the best performance among all the competitors. For the Boston Housing dataset, the performance of polynomial models is significantly better than that of linear models, whereas the linear models outperform the polynomial models on the ALE dataset. Note that ResTT achieves stable performance on both datasets, which demonstrates its effectiveness for modeling multiple multilinear correlations.

VI. CONCLUSION

In this article, we proposed a general framework that combines residual connections with TT to model feature interactions, from first order to Nth order, within one model. Mean-field analysis was extended to TT and ResTT to derive the condition of stability for the initialization of deep tensor networks. In particular, the stability condition of ResTT was significantly relaxed due to the residual connections. ResTT demonstrated...
significant improvement in training stability and convergence speed over the plain TT both theoretically and experimentally.

Numerical experiments on the synthetic and image classification datasets and two practical tasks showed the robust performance of ResTT. Particularly, ResTT outperformed the benchmark deep learning models in the case of limited training data.

The future direction would be to evaluate these tools on more complex problems, in which high-order multimodal correlations naturally exist, such as interpreting electroencephalography signals [50]. For example, it would be interesting to explore the application of ResTT in multitask learning and continual learning models, for which the high-order feature interaction is often adopted. As ResTT can be seen as an enhanced version of Volterra series, we also anticipate its applications in the learning with sequential data [50]. Moreover, considering the excellent performance of ResTT on limited training data, it is promising to apply ResTT to tasks for which training data acquisition is expensive [51].

APPENDIX A

DETAILED EXPRESSION OF THE OUTPUT OF A GENERAL RESIDUAL
TT

The elements of the final output of the ResTT with $N$ input features are calculated in the following, ranked from high order to low order. First, the $N$-th order term is calculated to be

$$
\sum_{v_1, \ldots, v_{N-1}, i_1, \ldots, i_N} w^{(1)}_{i_1 v_1} w^{(2)}_{i_1 i_2 v_2} \cdots w^{(N,1)}_{v_{N-1} i_N} x^{(1)}_{i_1} \cdots x^{(N)}_{i_N}.
$$

(32)

This term is the output of a plain TT network. Second, the $(N-1)$th-order terms are calculated to be

$$
\sum_{v_2, \ldots, v_{N-1}, i_2, \ldots, i_N} w^{(2)}_{i_2 v_2} \cdots w^{(N,1)}_{v_{N-1} i_N} x^{(2)}_{i_2} \cdots x^{(N)}_{i_N}.
$$

\ldots

(33)

Similarly, the $(N-2)$th-order terms can be calculated as

$$
\sum_{v_3, \ldots, v_{N-1}, i_3, \ldots, i_N} w^{(3,2)}_{i_3 v_3} w^{(4,1)}_{i_3 i_4 v_4} \cdots w^{(N)}_{v_{N-1} i_N} x^{(3)}_{i_3} \cdots x^{(N)}_{i_N}.
$$

\ldots

The rest of the interaction terms can be obtained in the same manner. Finally, we have the following first-order terms (linear correlations):

$$
\sum_{v_1, \ldots, v_{N-2}, i_1, \ldots, i_{N-2}} w^{(1,1)}_{i_1 v_1} \cdots w^{(N,3)}_{v_{N-2} i_{N-2}} x^{(1)}_{i_1} \cdots x^{(N-2)}_{i_{N-2}}.
$$

(34)

The final output of ResTT is the summation of all terms, from first order (linear) to $N$th order (Nonlinear).

APPENDIX B

MEAN-FIELD ANALYSIS OF TT AND RESIDUAL TT

Recall that all the weights are initialized from the same Gaussian distribution $\mathcal{N}(0, \sigma_w^2/\ell)$. We have

$$
\mathbb{E}(y^{(k)}_{vk}) = 0.
$$

(36)

Then, the variance of the $k$th output vector is

$$
q^{(k)} = \mathbb{E} \left[ \sum_{v_{k-1}, i_k} \left( y^{(k-1)}_{v_{k-1} i_k} x^{(k)}_{i_k} \right)^2 \right].
$$

(37)

Since $y^{(k-1)}_{v_{k-1}}, w_{v_{k-1} i_k v_k}$, and $x^{(k)}_{i_k}$ are independent, we have

$$
\mathbb{E}(y^{(k-1)}_{v_{k-1}} w^{(k)}_{v_{k-1} i_k v_k} x^{(k)}_{i_k}) = 0.
$$

(38)

Hence, (37) can be simplified as

$$
q^{(k)} = \mathbb{E} \left[ \sum_{v_{k-1}, i_k} (y^{(k-1)}_{v_{k-1}})^2 (w^{(k)}_{v_{k-1} i_k v_k})^2 (x^{(k)}_{i_k})^2 \right]
$$

$$
= \mathbb{E} \left[ \sum_{i_k} (x^{(k)}_{i_k})^2 \right] \sigma_w^2 q^{(k-1)}.
$$

(39)

Similarly, the derivatives of quadratic mean $\chi^{(k)}$ of the elements in $\mathcal{J}^{(k)}$ are given by

$$
\chi^{(k)} = \frac{1}{r} \mathbb{E} \left( \sum_{v_{k-1}} (j^{(k)}_{v_{k-1}})^2 \right) = \frac{1}{r} \mathbb{E} \left( \sum_{v_{k-1}} (j^{(k+1)} \cdot \alpha^{(k)}_{v_{k-1}})^2 \right)
$$

$$
= \frac{1}{r} \mathbb{E} \left( \sum_{v_{k-1}} (j^{(k+1)}_{v_{k-1}})^2 \alpha^{(k)}_{v_{k-1}}^2 \right).
$$
\[
\frac{1}{r} E \left[ \sum_{v_k, v_{k-1}, v_k} (j_{v_k}^{(k+1)})^2 (u_{v_{k-1}} v_k v_{k-1})^2 (x_{i_k}^{(k)})^2 \right] \\
= E \left( \sum_{i_k} (x_{i_k}^{(k)})^2 \right) \sigma^2 w \chi^{(k+1)}, \tag{40} 
\]

The evolution of signals in ResTT is slightly different due to the residual terms. Given that
\[
\gamma^{(k)} = \gamma^{(k+1)} + \gamma^{(k+2)} + \gamma^{(k-1)} 
\]
we have
\[
q^{(k)} = \frac{1}{r} E \left( \sum_{v_k, v_{k-1}, v_k} (y_{v_{k-1}}^{(k-1)})^2 (u_{v_{k-1}} v_k v_{k-1})^2 (x_{i_k}^{(k)})^2 \right) + \frac{1}{r} E \left( \sum_{i_k} (y_{v_{k-1}}^{(k-1)})^2 \right) \\
= \left( \sigma^2 w \sum_{i_k} (x_{i_k}^{(k)})^2 \right) + 1)q^{(k-1)} + \sigma_n^2 E \left( \sum_{i_k} (x_{i_k}^{(k)})^2 \right) \\
= (s^{(k)} + 1)q^{(k-1)} + s^{(k)} \tag{41} 
\]

and
\[
\chi^{(k)} = \frac{1}{r} E \left( \sum_{v_k, v_{k-1}} (j_{v_k}^{(k)})^2 (u_{v_{k-1}} v_k v_{k-1})^2 + 2 \right) \\
= \frac{1}{r} E \left( \sum_{v_k, v_{k-1}} (j_{v_k}^{(k+1)})^2 (u_{v_{k-1}} v_k v_{k-1})^2 (x_{i_k}^{(k)})^2 + 1 \right) \\
= (s^{(k)} + 1)\chi^{(k+1)}. \tag{42} 
\]

REFERENCES

[1] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, “MPCA: Multilinear principal component analysis of tensor objects,” IEEE Trans. Neural Netw., vol. 19, no. 1, pp. 18–39, Jan. 2008.

[2] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, “A survey of multilinear subspace learning for tensor data,” Pattern Recognit., vol. 44, no. 7, pp. 1540–1551, 2011.

[3] Y. Yang and T. Hospedales, “Deep multi-task representation learning: A tensor factorisation approach,” in Proc. Int. Conf. Learn. Represent., 2017.

[4] D. A. Freedman, Statistical Models: Theory and Practice. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[5] J. B. Tenenbaum and W. T. Freeman, “Separating style and content in nonlinear dynamics of learning in deep linear neural networks,” in Proc. Int. Conf. Learn. Represent., 2018.

[6] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

[7] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[8] S. Boyd and L. Chua, “Fading memory and the problem of approximating nonlinear operators with Volterra series,” IEEE Trans. Circuits Syst., vol. CAS-32, no. 11, pp. 1150–1161, Nov. 1985.

[9] M. A. Saxe, J. L. McClelland, and S. Ganguli, “Exact solutions to the nonlinear dynamics of learning in deep linear neural networks,” in Proc. Int. Conf. Learn. Represent., 2014.

[10] Y. LeCun and C. Cortes, “MNIST handwritten digit database,” 2010. [Online]. Available: http://yann.lecun.com/exdb/mnist/

[11] H. Xia, K. Rasul, and R. Vollgraf, “Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms,” 2017. arXiv:1708.07747.

[12] S. Boyd and L. Chua, “Fading memory and the problem of approximating nonlinear operators with Volterra series,” IEEE Trans. Circuits Syst., vol. CAS-32, no. 11, pp. 1150–1161, Nov. 1985.

[13] A. Sordoni, J.-Y. Nie, and Y. Bengio, “Modeling term dependencies with bilinear CNN models for fine-grained visual recognition,” in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 1449–1457.

[14] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, “Compact bilinear pooling,” in Proc. IEEE Int. Conf. Comput. Vis., 2015, 2015, pp. 2019–2027.

[15] A. Sordoni, J.-Y. Nie, and Y. Bengio, “Modeling term dependencies with bilinear CNN models for fine-grained visual recognition,” in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 1449–1457.

[16] H. Zheng, J. Fu, Z.-J. Zha, and J. Luo, “Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 5007–5016.

[17] T. Do, T.-T. Do, H. Tran, E. Tijputra, and Q. D. Tran, “Compact trilinear interaction for visual question answering,” in Proc. Int. Conf. Comput. Vis., 2019, pp. 392–401.

[18] E. Stoudenmire and D. J. Schwab, “Supervised learning with tensor networks,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2016, pp. 4799–4807.
[39] M. Blondel, M. Ishihata, A. Fujino, and N. Ueda, “Polynomial networks and factorization machines: New insights and efficient training algorithms,” in Proc. Int. Conf. Mach. Learn., 2016, pp. 850–858.

[40] I. Glasser, N. Pancotti, and J. I. Cirac, “Supervised learning with generalized tensor networks,” 2018. [Online]. Available: https://arxiv.org/abs/1806.05964

[41] K. He, X. Zhang, S. Ren, and J. Sun, “Identity mappings in deep residual networks,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 630–645.

[42] A. Veit, M. J. Wilber, and S. Belongie, “Residual networks behave like ensembles of relatively shallow networks,” in Proc. Int. Conf. Neural Inf. Process. Syst., Dec. 2016, pp. 550–558.

[43] Y. LeCun et al., “Backpropagation applied to handwritten zip code recognition,” Neural Comput., vol. 1, no. 4, pp. 541–551, 1989.

[44] S. S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein, “Deep information propagation,” in Proc. Int. Conf. Learn. Represent., 2017.

[45] V. Vapnik, The Nature of Statistical Learning Theory. New York, NY, USA: Springer, 2013.

[46] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. Int. Conf. Learn. Representation, 2015.

[47] A. Paszke et al., “PyTorch: An imperative style, high-performance deep learning library,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2019, pp. 8026–8037.

[48] S. Joffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in Proc. Int. Conf. Mach. Learn., 2015, pp. 448–456.

[49] A. J. Smola and B. Schölkopf, “A tutorial on support vector regression,” Statist. Comput., vol. 14, no. 3, pp. 199–222, 2004.

[50] D. Zhang, L. Yao, K. Chen, S. Wang, X. Chang, and Y. Liu, “Making sense of spatio-temporal preserving representations for EEG-based human intention recognition,” IEEE Trans. Cybern., vol. 50, no. 7, pp. 3033–3044, Jul. 2020.

[51] K. Chen, L. Yao, D. Zhang, X. Wang, X. Chang, and F. Nie, “A semisupervised recurrent convolutional attention model for human activity recognition,” IEEE Trans. Neural Netw. Learn. Syst., vol. 31, no. 5, pp. 1747–1756, May 2020.

Yu Pan (Senior Member, IEEE) received the bachelor’s degree in mathematics and applied mathematics from Sichuan University, Chengdu, China, in 2007, and the Ph.D. degree in systems science from the Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China, in 2012. From 2012 to 2015, he was a Research Fellow with the College of Engineering, Australian National University, Canberra, ACT, Australia. He was a Research Associate with the School of Engineering and Information Technology, University of New South Wales, Canberra, from 2015 to 2016, and a Postdoctoral Fellow with Hong Kong Polytechnic University, Hong Kong, in 2016. Since July 2016, he has been a Research Professor with the Institute of Cyber-Systems and Control, College of Control Science and Engineering, Zhejiang University, Hangzhou, China. His research interests include quantum technology, single-photon imaging, and computational intelligence.

Dr. Pan is the Co-Chair of the Technical Committee on Quantum Cybernetics of the IEEE Systems, Man, and Cybernetics Society.

Daoyi Dong (Senior Member, IEEE) received the B.E. degree in automatic control and the Ph.D. degree in engineering from the University of Science and Technology of China, Hefei, China, in 2001 and 2006, respectively.

He is currently a Scientia Associate Professor with the University of New South Wales, Canberra, ACT, Australia. He was with the Institute of Systems Science, Chinese Academy of Sciences, Beijing, China, and with Zhejiang University, Hangzhou, China. He had visiting positions with Princeton University, Princeton, NJ, USA; RIKEN, Wako-Shi, Japan; University of Duisburg-Essen, Duisburg, Germany; and the University of Hong Kong, Hong Kong. His research interests include quantum control, machine learning, and renewable energy.

Dr. Dong was the recipient of an ACA Temasek Young Educator Award by the Asian Control Association, an International Collaboration Award and an Australian Postdoctoral Fellowship from the Australian Research Council, and a Humboldt Research Fellowship from the Alexander von Humboldt Foundation of Germany. He is a Member-at-Large on the Board of Governors and was the Associate Vice-President for conferences and meetings of the IEEE Systems, Man and Cybernetics Society. He was an Associate Editor for IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS from 2015 to 2021. He is an Associate Editor for IEEE TRANSACTIONS ON CYBERNETICS and a Technical Editor for IEEE/ASME TRANSACTIONS ON MECHATRONICS.

Yiwei Chen received the B.E. degree in control science and engineering, in 2018, from Zhejiang University, Hangzhou, China, where he is currently toward the Ph.D. degree with the Institute of Cyber-Systems and Control.

His research interests include computational imaging, quantum machine learning, and its applications.