Appendix

Anonymous ACL-IJCNLP submission

We offer some proof as supplementary materials to help authors better understand our model. The appendix includes 3 pages and is organized into sections:

- Tensor and Matrix Product Operators
- Theorem
- Experiment

A Tensor and Matrix Product Operators

As introduced in (Cichocki et al., 2009), the concept of tensor is specified as:

Definition1
(Tensor). Let $D_1, D_2..., D_N \in N$ denote index upper bounds. A tensor $\mathcal{T} \in \mathbb{R}^{D_1,...,D_n}$ of order N is an N-way array where elements $\mathcal{T}_{d_1,d_2,...,d_n}$ are indexed by $d_n \in \{1, 2, ..., D_n\}$ for $1 \leq n \leq N$

Definition2
(Matrix product operator). We can reshape a matrix to high order tensor, denote as:

$$\mathbf{M}_{x \times y} = \mathbf{M}_{i_1 i_2 \ldots i_n, j_1 j_2 \ldots j_n} \tag{1}$$

Here, the one-dimensional coordinate x of the input signal $\mathbf{x}$ with dimension N_x is reshaped into a coordinate in a n-dimensional space, labelled by $(i_1 i_2 \cdots i_n)$. Hence, there is a one-to-one mapping between x and $(i_1 i_2 \cdots i_n)$. Similarly, the one-dimensional coordinate y of the output signal $\mathbf{y}$ with dimension N_y is also reshaped into a coordinate in a n-dimensional space, and there is a one-to-one correspondence between y and $(j_1 j_2 \cdots j_n)$. If I_k and J_k are the dimensions of i_k and j_k, respectively, then

$$\prod_{k=1}^{n} I_k = N_x, \quad \prod_{k=1}^{n} J_k = N_y. \tag{2}$$

The MPO representation of M is obtained by factorizing it into a product of n local tensors

$$M_{i_1 \cdots i_n, j_1 \cdots j_n} = \mathcal{T}^{(1)}[i_1, j_1] \cdots \mathcal{T}^{(n)}[i_n, j_n] \tag{3}$$

where $\mathcal{T}^{(k)}[j_k, i_k]$ is a $D_{k-1} \times D_k$ matrix with D_k the virtual basis dimension on the bond linking $\mathcal{T}^{(k)}$ and $\mathcal{T}^{(k+1)}$ with $D_0 = D_n = 1$.

B Theorem

Theorem 1. *Suppose that the tensor $\mathbf{W}^{(k)}$ of matrix W that is satisfy*

$$\begin{aligned} &\mathbf{W} = \mathbf{W}^{(k)} + \mathbf{E}^{(k)}, D(\mathbf{W}^{(k)}) = d_k, \\ &where \quad ||\mathbf{E}^{(k)}||_F^2 = \epsilon_k^2, k = 1, ..., d-1. \end{aligned} \tag{4}$$

Then MPO $(\mathbf{W})$ with the k-th bond dimension d_k upper bound of truncation error satisfy:

$$||\mathbf{W} - MPO\ (\mathbf{W})||_F \leq \sqrt{\sum_{k=1}^{d-1} \epsilon_k^2} \tag{5}$$

$Proof$. The proof is by induction. For $n = 2$ the statement follows from the properites of the SVD. Consider an arbitrary $n > 2$. Then the first unfolding $\mathbf{W}^{(1)}$ is decomposed as

$$\mathbf{W}^{(1)} = \mathbf{U}_1 \lambda_1 \mathbf{V}_1 + \mathbf{E}^{(1)} = \mathbf{U}_1 \mathbf{B}^{(1)} + \mathbf{E}^{(1)} \tag{6}$$

where $\mathbf{U}_1$ is of size $r_1 \times i_1 \times j_1$ and $||\mathbf{E}^{(1)}||_F^2 = \epsilon_1^2$. The matrix $\mathbf{B}_1$ is naturally associated with a $(n-1)$-dimensional tensor $\mathcal{B}^{(1)}$ with elements $\mathcal{B}^{(1)}(\alpha, i_2, j_2, ..., i_n, j_n)$, which will be decomposed further. This means that $\mathbf{B}_1$ will be approximated by some other matrix $\hat{\mathbf{B}}_1$. From the properties of the SVD it follows that $\mathbf{U}_1^T \mathbf{E}^{(1)} = 0$, and thus

$$\begin{aligned} &||\mathbf{W} - \mathcal{B}^{(1)}||_F^2 \\ &= ||\mathbf{W}_1 - \mathbf{U}_1 \hat{\mathbf{B}}_1||_F^2 \\ &= ||\mathbf{W}_1 - \mathbf{U}_1(\hat{\mathbf{B}}_1 + \mathbf{B}_1 - \mathbf{B}_1)||_F^2 \\ &= ||\mathbf{W}_1 - \mathbf{U}_1 \mathbf{B}_1||_F^2 + ||\mathbf{U}_1(\hat{\mathbf{B}}_1 - \mathbf{B}_1)||_F^2 \end{aligned} \tag{7}$$

and since $\mathbf{U}_1$ has orthonormal columns,

$$||\mathbf{W} - \mathcal{B}^{(1)}||_F^2 \leq \epsilon_1^2 + ||\mathbf{B}_1 - \hat{\mathbf{B}}_1||_F^2. \quad (8)$$

and thus it is not difficult to see from the orthonormality of columns of $\mathbf{U}_1$ that the distance of the k-th unfolding $(k = 2, ..., d_k - 1)$ of the $(d-1)$-dimensional tensor $\mathcal{B}^{(1)}$ to the d_k-th rank matrix cannot be larger then ϵ_k. Proceeding by induction, we have

$$||\mathbf{B}_1 - \hat{\mathbf{B}}_1||_F^2 \leq \sum_{k=2}^{d-1} \epsilon_k^2, \quad (9)$$

combine with Eq. (8), this completes the proof.

C Experiment

C.1 Additional Details of MPO

In this paper, the MPOP is proposed for compressing pre-trained Language Models. In order to show that the process of incorporating several MPO sturctures into ALBERT-based and BERT-based pre-trained language models respectively. We introduce MPO decomposition in ALBERT and BERT details as follows:

Layers	Matrix shape	MPO shape $[d_{k-1}, i_k, j_k, d_k]$
AlbertEmbeddings	30000×128	$\mathcal{A}_1$:$[1, 5, 2, 10]$ $\mathcal{A}_2$:$[10, 10, 2, 200]$ $\mathcal{C}$:$[200, 10, 4, 480]$ $\mathcal{A}_3$:$[480, 10, 4, 12]$ $\mathcal{A}_4$:$[12, 6, 2, 1]$
AlbertLayer	768×3072	$\mathcal{A}_1$:$[1, 3, 4, 12]$ $\mathcal{A}_2$:$[12, 4, 4, 192]$ $\mathcal{C}$:$[192, 4, 8, 384]$ $\mathcal{A}_3$:$[384, 4, 6, 16]$ $\mathcal{A}_4$:$[16, 4, 4, 1]$
	3072×768	$\mathcal{A}_1$:$[1, 4, 3, 12]$ $\mathcal{A}_2$:$[12, 4, 4, 192]$ $\mathcal{C}$:$[192, 8, 4, 384]$ $\mathcal{A}_3$:$[384, 6, 4, 16]$ $\mathcal{A}_4$:$[16, 4, 4, 1]$
AlbertAttention (query/key/value/ output)	768×768	$\mathcal{A}_1$:$[1, 3, 4, 12]$ $\mathcal{A}_2$:$[12, 4, 4, 192]$ $\mathcal{C}$:$[192, 4, 4, 192]$ $\mathcal{A}_3$:$[192, 4, 4, 12]$ $\mathcal{A}_4$:$[12, 4, 3, 1]$

Table 1: ALBERT MPO Decomposition Shape

There is some slight difference of MPO structure between ALBERT and BERT. In word embedding layer, we use MPO to decompose a matrix of shape [30720,768] rather than [30522,768], for "30522" can not be reshaped to dimensions of i_k as introduced in Eq. (2). Specifically, We get [30720,768] by zero padding first, then we apply MPO decomposition, at last, we clip the paddings before computing with input tokens. In intermediate and output layers, BERT and ALBERT share all of the shape of matrix.

Layers	Matrix shape	MPO shape $[d_{k-1}, i_k, j_k, d_k]$
BertEmbeddings	30720×768	$\mathcal{A}_1 : [1, 5, 2, 10]$ $\mathcal{A}_2 : [10, 10, 2, 200]$ $\mathcal{C} : [200, 10, 4, 480]$ $\mathcal{A}_3 : [480, 10, 4, 12]$ $\mathcal{A}_4 : [12, 6, 2, 1]$
BertIntermediate	768×3072	$\mathcal{A}_1 : [1, 3, 4, 12]$ $\mathcal{A}_1 : [12, 4, 4, 192]$ $\mathcal{C}$:$[192, 4, 8, 384]$ $\mathcal{A}_3$:$[384, 4, 6, 16]$ $\mathcal{A}_4$:$[16, 4, 4, 1]$
Bertoutput	3072×768	$\mathcal{A}_1$:$[1, 4, 3, 12]$ $\mathcal{A}_2$:$[12, 4, 4, 192]$ $\mathcal{C}$:$[192, 8, 4, 384]$ $\mathcal{A}_3$:$[384, 6, 4, 16]$ $\mathcal{A}_4$:$[16, 4, 4, 1]$
BertAttention (query/key/value/ output)	768×768	$\mathcal{A}_1$:$[1, 3, 4, 12]$ $\mathcal{A}_2$:$[12, 4, 4, 192]$ $\mathcal{C}$:$[192, 4, 4, 192]$ $\mathcal{A}_3$:$[192, 4, 4, 12]$ $\mathcal{A}_4$:$[12, 4, 3, 1]$

Table 2: BERT MPO Decomposition Shape

C.2 Experimental Details in Pre-trained Language Modeling

Now, we report some details of experiments as a relevant supplementary material. Firstly, we expand all the matrices $\{\mathbf{M}_k\}_{k=1}^N$ in ALBERT into MPO structure with $\left\{\{\mathcal{A}_1, \mathcal{A}_2, \mathcal{C}, \mathcal{A}_3, \mathcal{A}_4\}_k\right\}_{k=1}^N$. Specific details in C.1. In the experimental of main text, "$\text{MPOP}_{\text{full}}$" means that we fine-tune all these tensors compare with "$\text{MPOP}_{\text{full+LFA}}$"denotes that we fine-tune these tensors with central tensor fixed. Then, we can further compressing the MPO structure by truncating $\{d_k\}$ to $\{d'_k\}$ as described in the main text. At the same time, Dimension-Squeezing method can also be used for compression and fine-tuning.

Hardware We trained our model on one machine with 4 NVIDIA Titan V GPUs. For our base models, we adopt all these models released by Huggingface [1].

Optimizer We used the Adam optimizer and vary the learning rate over the course of training. The

[1]https://huggingface.co/

vary formula (Vaswani et al., 2017) is follows in our work. We also used the $warmup\_steps = 1000$.

References

Andrzej Cichocki, Rafal Zdunek, Anh Huy Phan, and Shun-ichi Amari. 2009. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.