A Tensorized Transformer for Language Modeling

Tianjin University

Xindian Ma, Peng Zhang, Shuai Zhang, Nan Duan,
Yuexian Hou, Dawei Song, Ming Zhou

NeurIPS 2019

Background

- Transformer has led to breakthroughs in natural language processing tasks.
- Transformer, and its variant model BERT, limit the effective deployment of the model to limited resource setting.

- Some compression methods have been proved.
  - TT-embedding
  - BTRNN
  - Tensorizing Neural Networks

Some Compression Methods

- TT-Embedding [1]
  - Tensor-Train decomposition is used to compress the embedding layer (look-up table).
- BTRNN [2]
  - Block-term tensor decomposition is used to compress the input layers in LSTM
- Tensorizing Neural Networks [3]
  - Tensor Train format is used to compress the fully-connected layers.

[1] Valentin Khrulkov, Oleksii Hrinchuk, Leyla Mirvakhabova, and Ivan Oseledets. Tensorized embedding layers for efficient model compression. arXiv preprint arXiv:1901.10787, 2019
[2] Jinmian Ye, Linnan Wang, Guangxi Li, Di Chen, Shandian Zhe, Xinqi Chu, and Zenglin Xu. Learning compact recurrent neural networks with block-term tensor decomposition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 9378–9387, 2018.
[3] Novikov A, Podoprikhin D, Osokin A, et al. Tensorizing neural networks[C]//Advances in neural information processing systems. 2015: 442-450.

Compressed Transformer

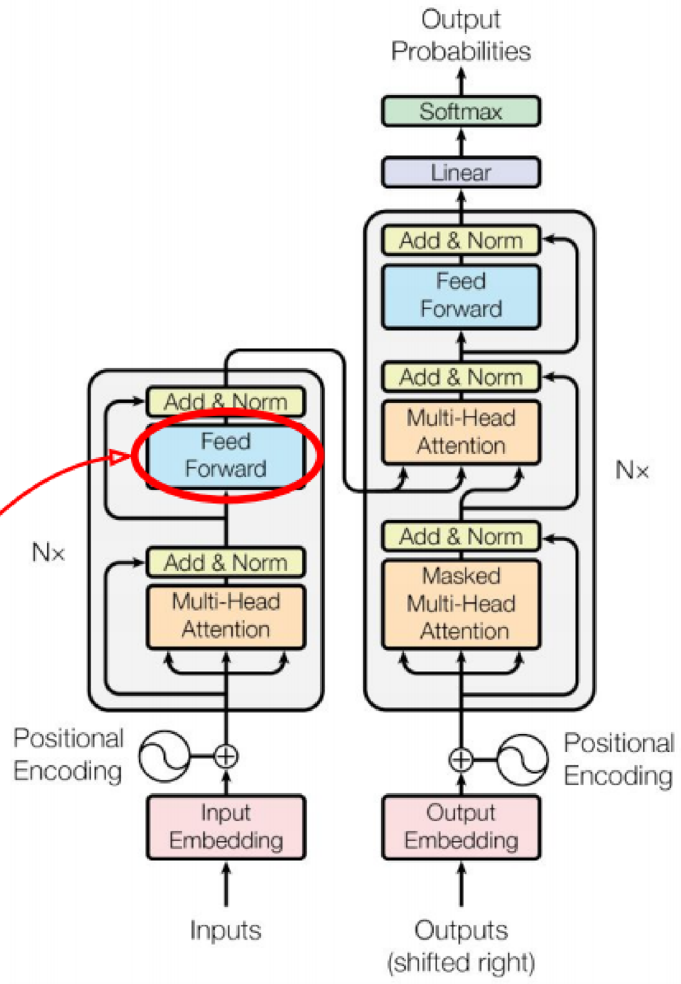
Transformer

model for **state-of-the-art** results in NLP:

- neural machine translation
- Q&A
- NER
- POS-tagging

Up to **213×10^6** parameters

$2^{21} \approx 2\times10^6$ parameters in **one** Feed Forward block!



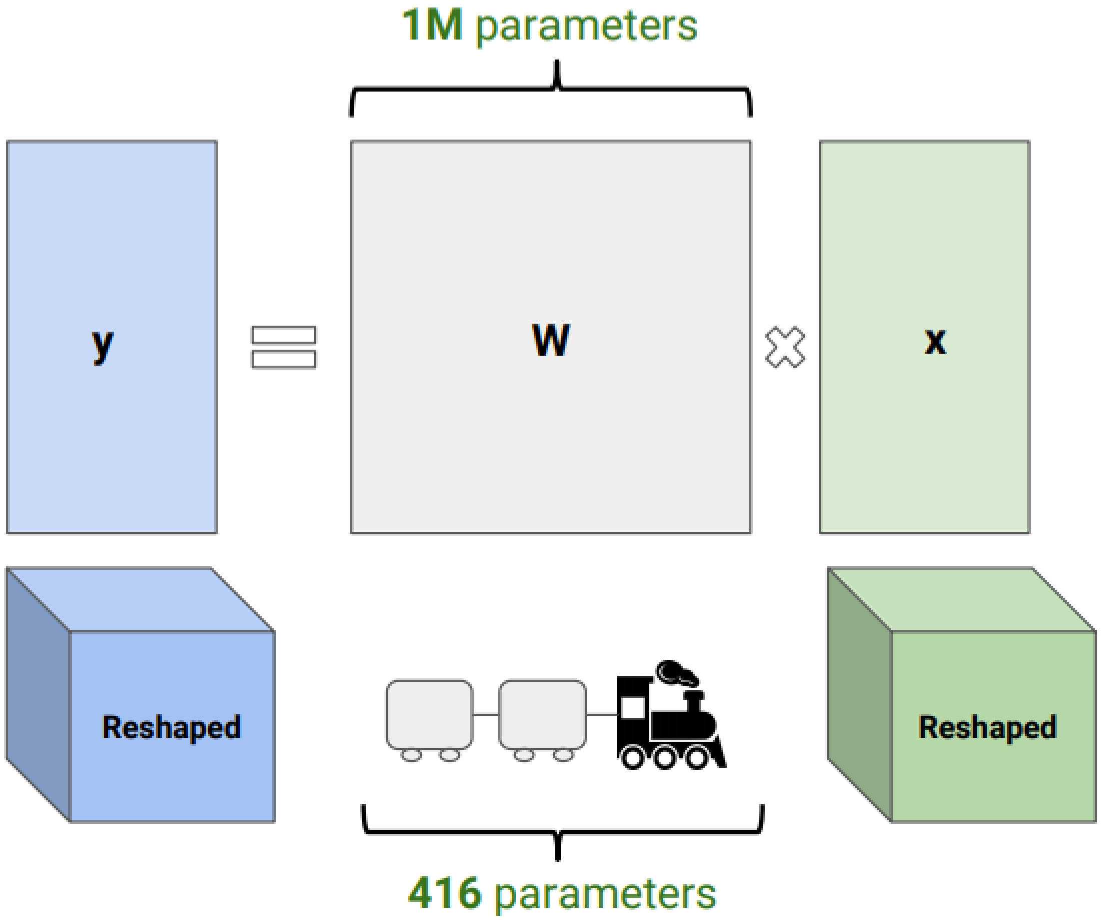
[1] http://nla.skoltech.ru/projects/files/presentations/team_21.pdf

Methods

Explicit structure:

Tensor- train

SVD + finetune



Problem Formulation

- The goals are:
  - To linearly represent a self-attention by a group of basic vectors
  - To compress multi-head attention in Transformer
  - After compressing, it can be directly integrated into the encoder and decoder framework of Transformer

Methods

Basic Ideas

- Low-rank decomposition
- Parameters sharing

Using Tucker decomposition formulation is to construct Single-block attention

Using Block-term decomposition + Parameters sharing formulation is to construct multi-head mechanisms(Multi-linear attention)

Transformer Language Modeling

- Scaled Dot Production Attention

- $Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V$

- Multi-head Attention

- $MultiHeadAttention(Q, K, V) = Concat(head_1, \cdots, head_k)W^o$

$$where\ head_i = Attention\boxed{(QW_i^Q, KW_i^K, VW_i^V)}$$

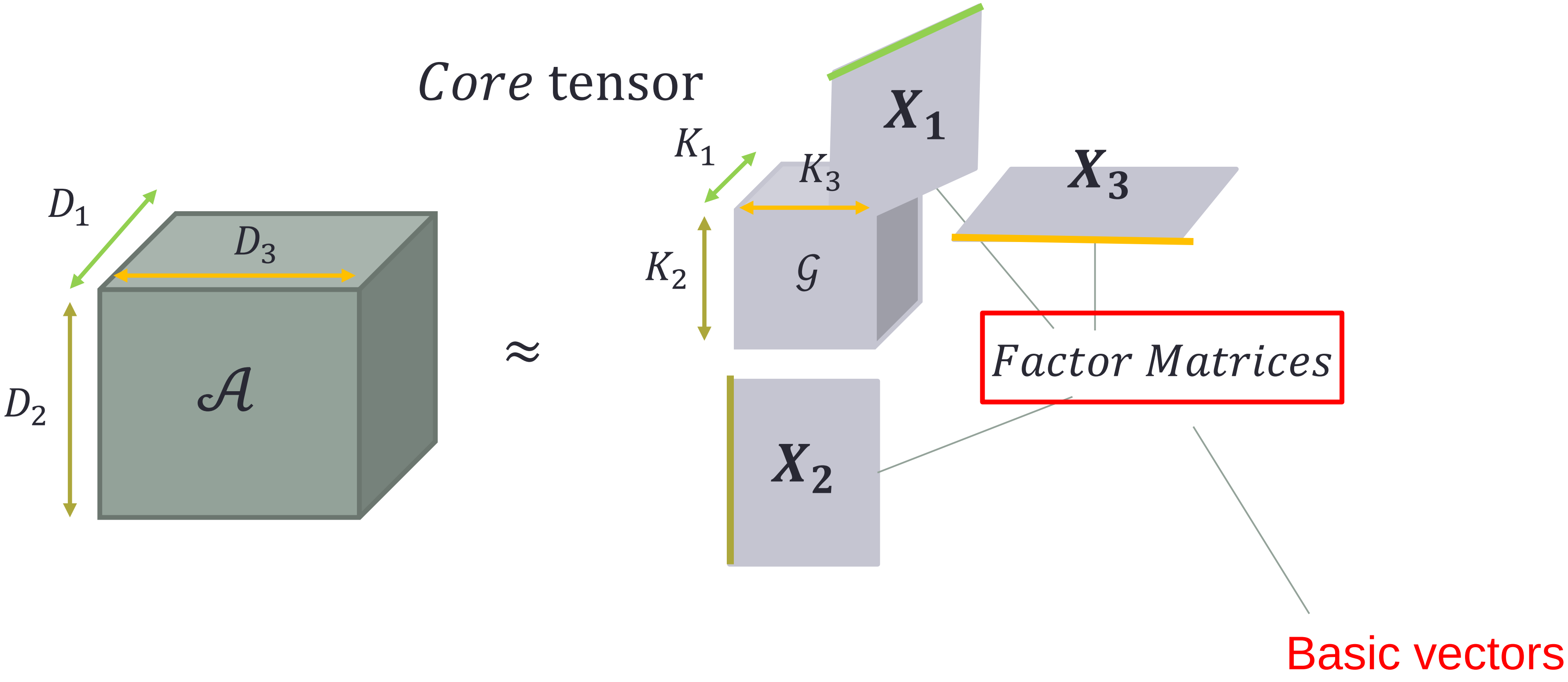
Multi-group parameters

Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.

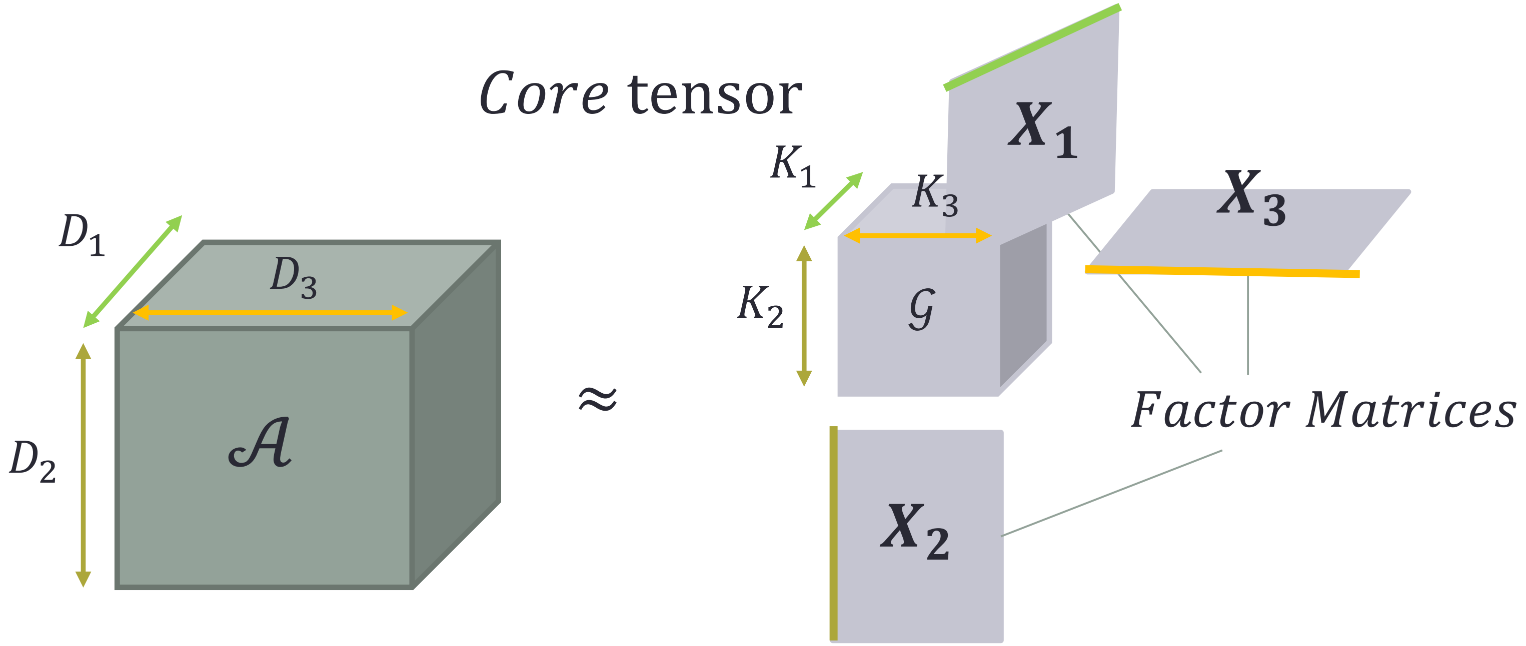
Linear Representation

- **Theorem:** "Scaled Dot Production Attention" can be represented by a linear combination of a set of basis vectors.
- $$Attention(Q, K, V) = (e_1, \cdots, e_n)M$$
- $where\ M \in \mathbb{R}^{n \times d}\ is\ a\ coeffcient\ matrix.$
-

Tucker Decomposition



Single-Block Attention

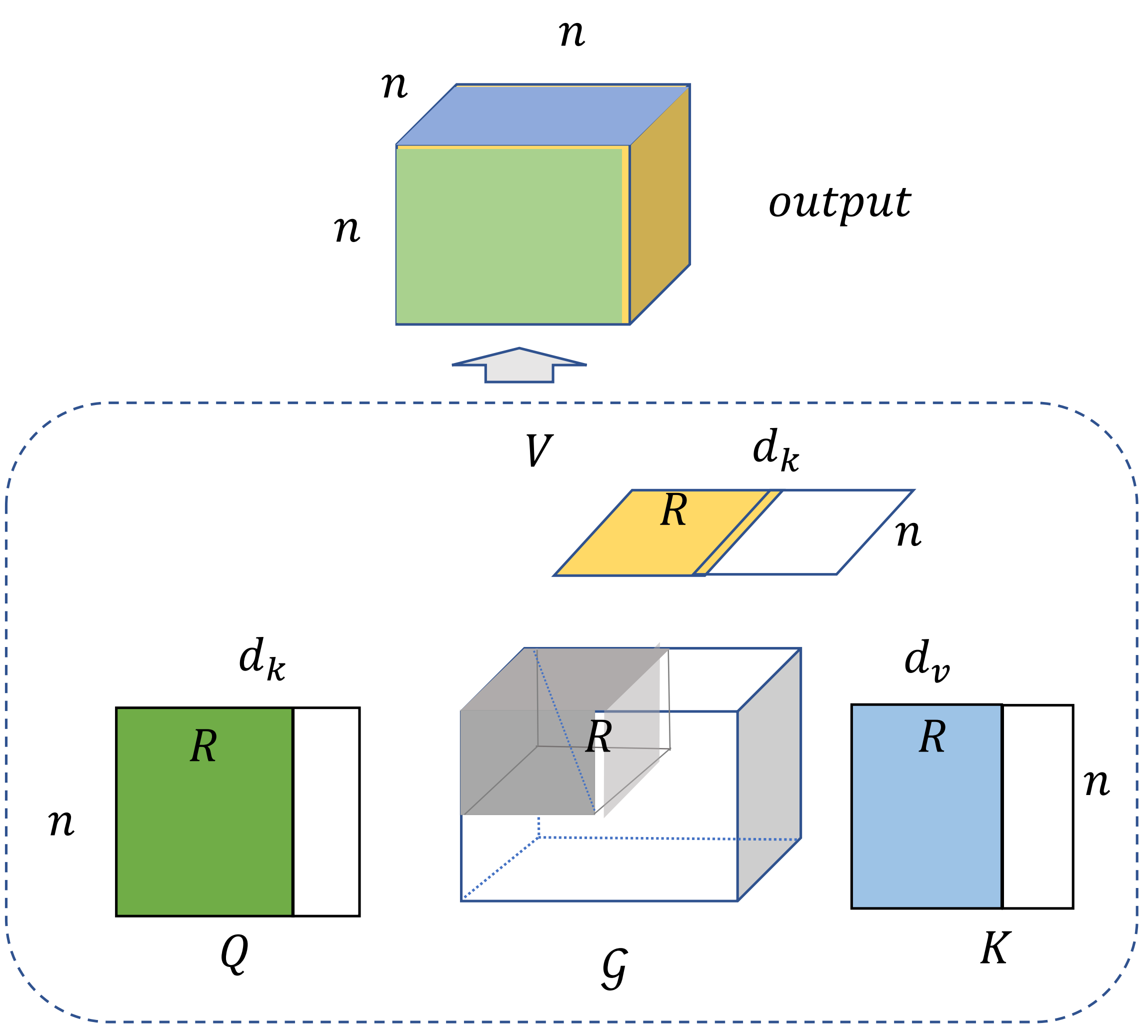


$\circ$ *is the outer product.*

$$Attention_{\mathrm{TD}}(\mathcal{G}; Q, K, V) = \mathcal{G} \cdot_1 Q \cdot_2 K \cdot_3 V$$

$$= \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{m=1}^{M} \mathcal{G}_{ijm}\, Q_i \circ K_j \circ V_m$$

Single-Block Attention in Transformer



Lower-Rank Decomposition

- The Core-tensor $\mathcal{G}$ is defined as follows.
  - $\mathcal{G}_{ijm} = \begin{cases} rand(0,1)\,, & i = j = m \\ 0 & otherwise \end{cases}$
  - In this case , it can be set as $I = J = M = R$
  - R is the Rank.

The time complexity of Single-block attention is $\mathcal{O}(N^3)$.

The time complexity of Scaled dot production attention is $\mathcal{O}(N^2 d)$.

Reconstruction for Scaled dot product attention

- **Corollary:** Scaled dot product attention can be reconstructed by Single block attention
  - $Attention(Q,K,V)_{i,m} = \sum_{j=1}^{J} Attention_{TD}(\mathcal{G};Q,K,V)_{i,j,m}$
  - $where\ i,j\ and\ m\ are\ the\ indices\ of\ the\ single-block\ attention's\ ouput.$

Graphing of Reconstruction

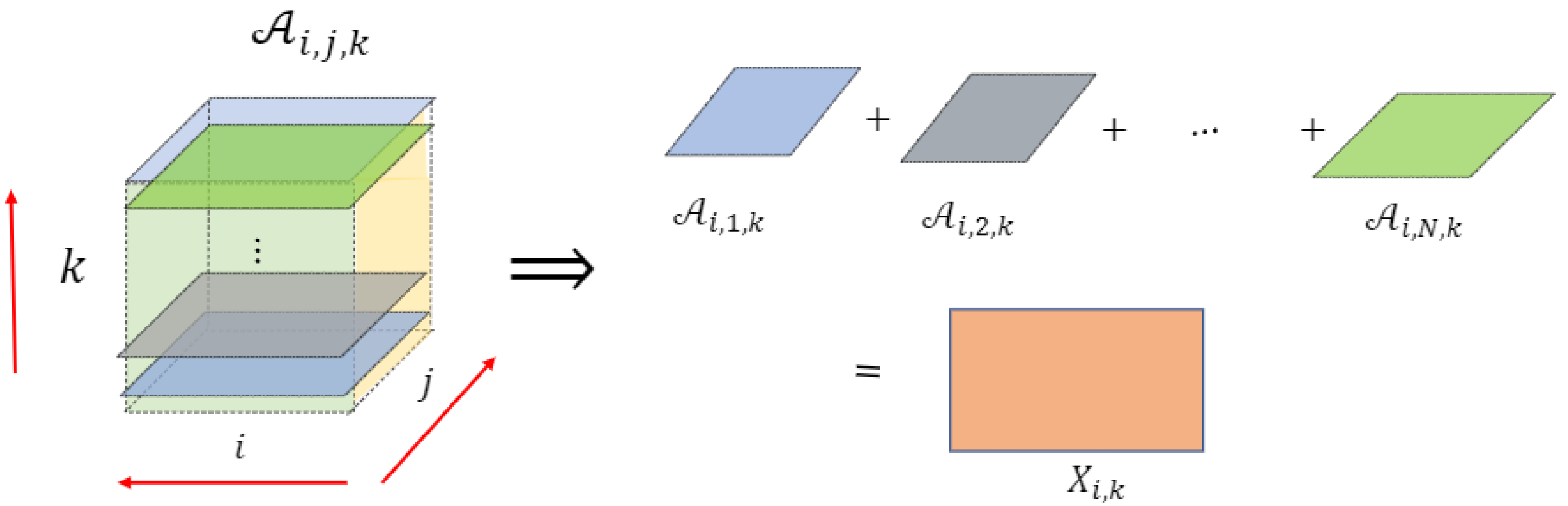
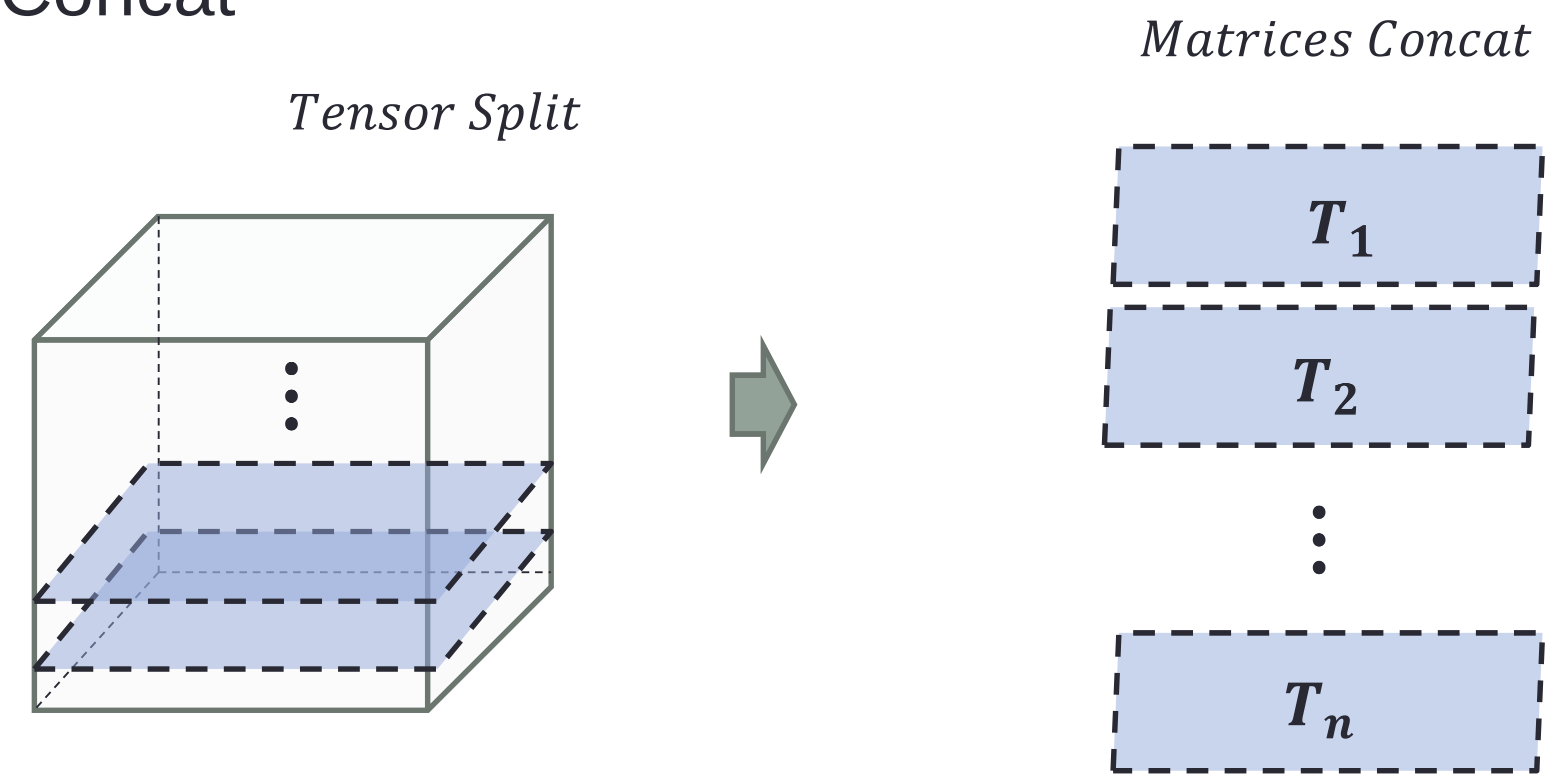


Figure 1: Tensor $\mathcal{A}$ is a 3-order tensor, which represents the Single-block attention in the left. $\mathcal{A}_{i,j,k}$ is the entry of the tensor $\mathcal{A}$. In the right, the graph represents that the summing of tensor slices which is from the tensor splitting in index j.

How to Get Richer Representation

- Tensor Split
- Matrices Concat

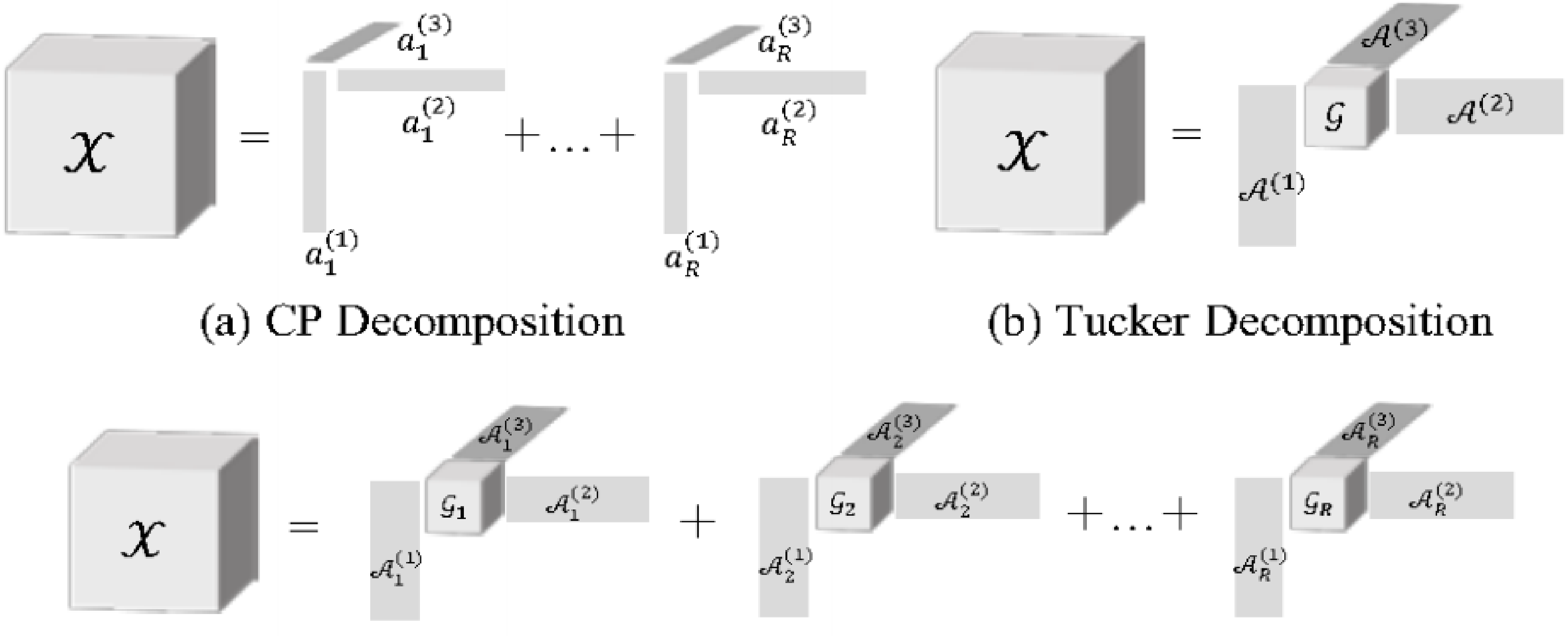


Multi-linear Attention by Block-term Decomposition

- It is important to constructed the multi-head mechanism for modeling long-range dependency.

- How to design the model with higher compression ratios?
  - 1) Block-term decomposition (method)
  - 2) Parameters sharing (idea)

Block-term Decomposition



(c) Block Term Decomposition

Chen Y, Jin X, Kang B, et al. Sharing Residual Units Through Collective Tensor Factorization To Improve Deep Neural Networks[C]//IJCAI. 2018: 635-641.

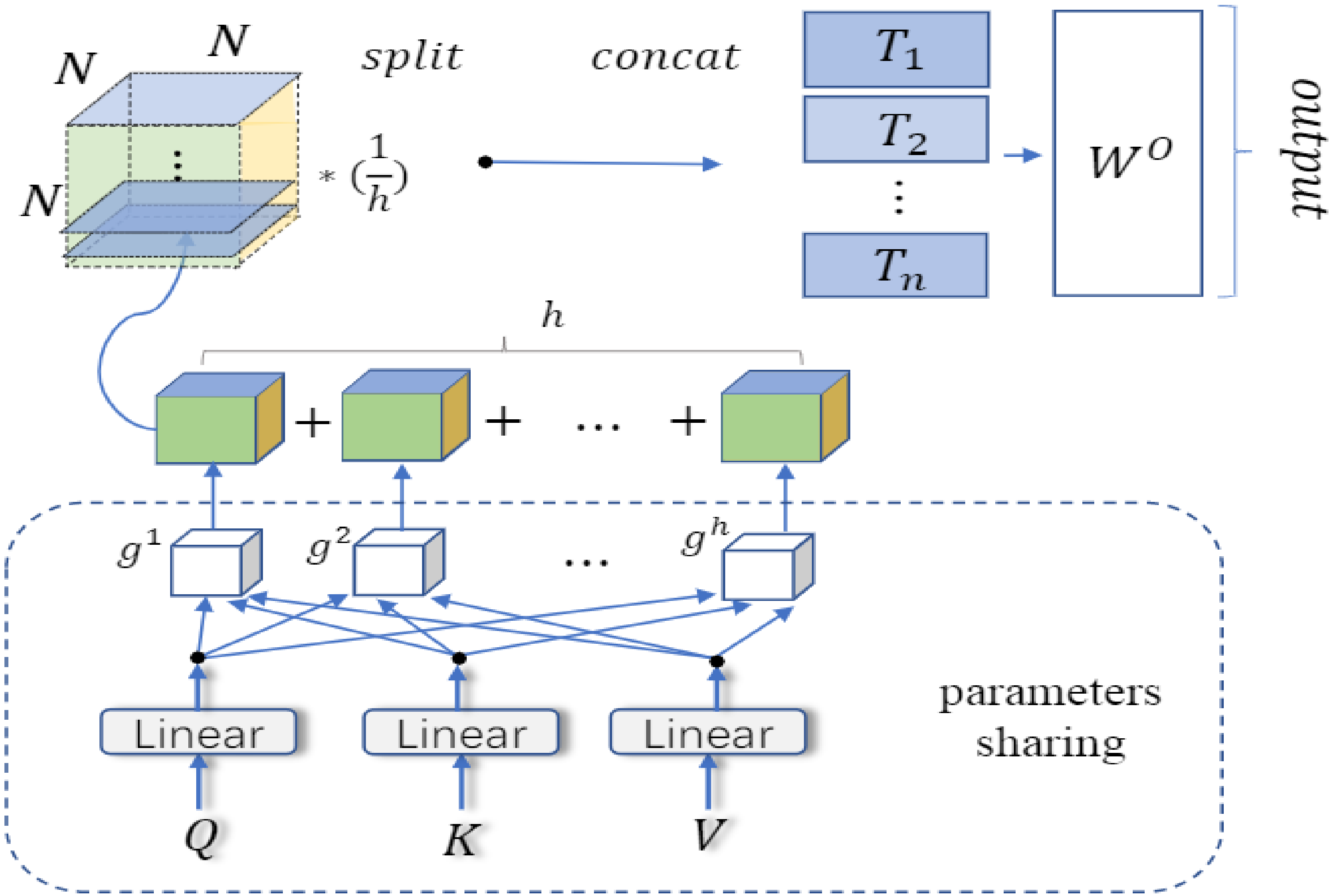
Multi-linear Attention by Block-term Decomposition

In order to construct the multi-head mechanism, Multi-linear attention can be formulated as follows:

$$MultiLinearAttention(\mathcal{G}; Q, K, V) = SplitConcat\left(\frac{1}{h} * (T_1 + \cdots + T_h)\right) W^O$$

$$where\ T_j = Attention_{TD}(\mathcal{G}_j; \underbrace{Q\boxed{W^q}, K\boxed{W^K}, V\boxed{W^V}}_{\text{Parameters Sharing}})$$

Multi-Linear Attention



Experimental Results in Language Modeling

One-Billion

Model	Params	Test PPL
RNN-1024+9 Gram [4]	20B	51.3
LSTM-2018-512 [17]	0.83B	43.7
GCNN-14 bottleneck [8]	–	31.9
LSTM-8192-1024+CNN Input [17]	1.04B	30.0
High-Budget MoE [31]	5B	28.0
LSTM+Mos [36]	113M	37.10
Transformer+adaptive input [1]	0.46B	23.7
Transformer-XL Base [7]	0.46B	23.5
Transformer-XL Large [7]	0.8B	21.8
Tensorized Transformer core-1	0.16B	20.5
Tensorized Transformer core-2	0.16B	**19.5**

PTB

WikiText-103

Model	PTB			WikiText-103		
	Params	Val PPL	Test PPL	Params	Val PPL	Test PPL
LSTM+augmented loss [15]	24M	75.7	48.7	–	–	48.7
Variational RHN [38]	23M	67.9	65.4	–	–	45.2
4-layer QRNN [21]	–	–	–	151M	–	33.0
AWD-LSTM-MoS [36]	22M	58.08	55.97	–	29.0	29.2
Transformer+adaptive input [1]	24M	59.1	57	247M	19.8	20.5
Transformer-XL [7]	24M	56.72	54.52	151M	23.1	24.0
Transformer-XL+TT [18]	18 M	57.9*	55.4*	130M	23.61*	25.70*
Tensorized Transformer core-1	12M	60.5	57.9	80.5M	22.7	20.9
Tensorized Transformer core-2	12M	**54.25**	**49.8**	86.5M	**19.7**	**18.9**

Experimental Results in Language Modeling

WMT-16 English-to-German

Model	Params	BLEU
Base-line [30]	–	26.8
Linguistic Input Featurec [29]	–	28.4
Attentional encoder-decoder + BPE [30]	–	34.2
Transformer [34]	52M	34.5*
Tensorized Transformer core-1	21M	34.10
Tensorized Transformer core-2	21.2M	**34.91**

Rico Sennrich, Barry Haddow, and Alexandra Birch. Edinburgh neural machine translation systems for wmt 16.arXiv preprint arXiv:1606.02891, 2016.

Conclusion

- We provided a novel self-attention method, namely Multi-linear attention.
- The Tensorized Transformer model combines two compression ideas, parameters sharing and low-rank decomposition.
- Our methods achieve higher compression ratio and better experimental results in language modeling.
- The Tensorized Transformer model can be implied to more NLP tasks with limited resources through further optimization

Thanks!