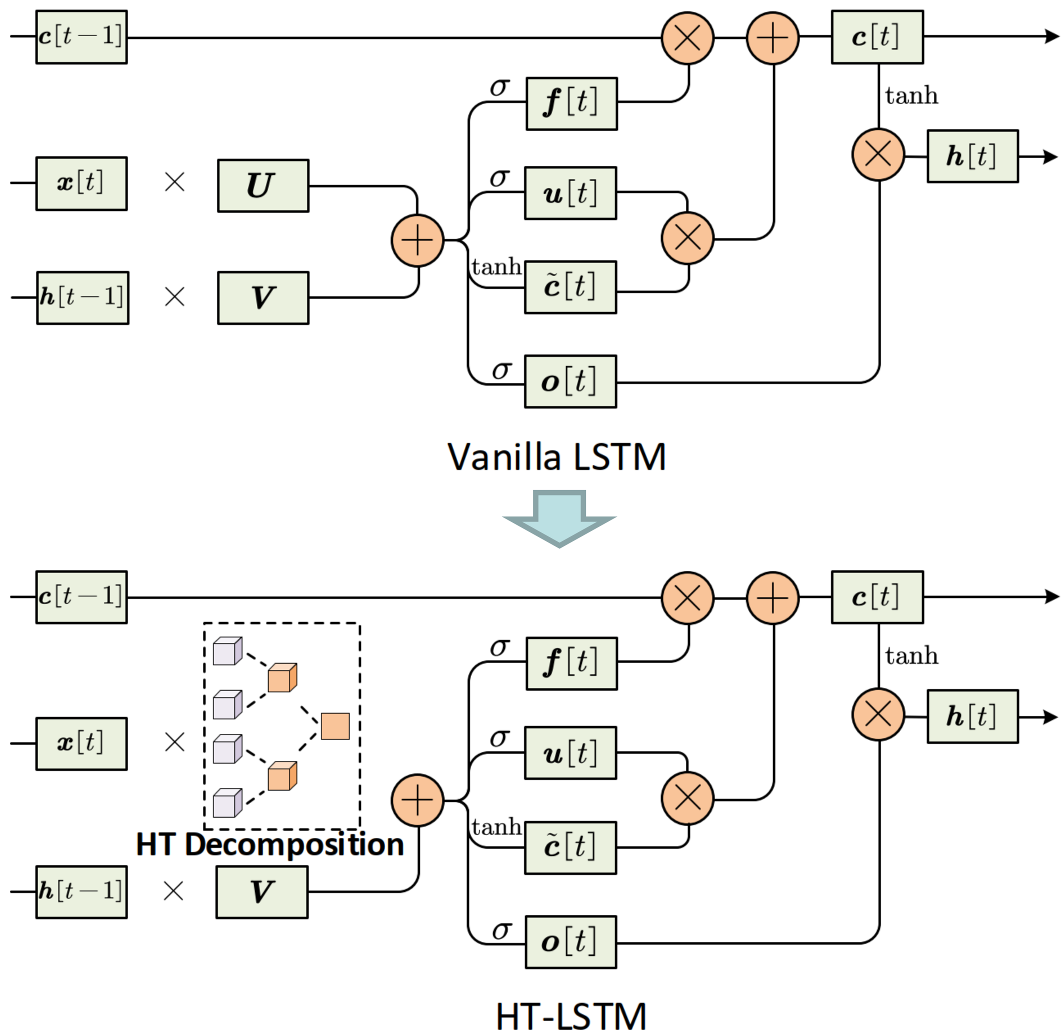
Compressing Recurrent Neural Networks Using Hierarchical Tucker Tensor Decomposition

Miao Yin, Siyu Liao, Xiao-Yang Liu,
Xiaodong Wang and Bo Yuan

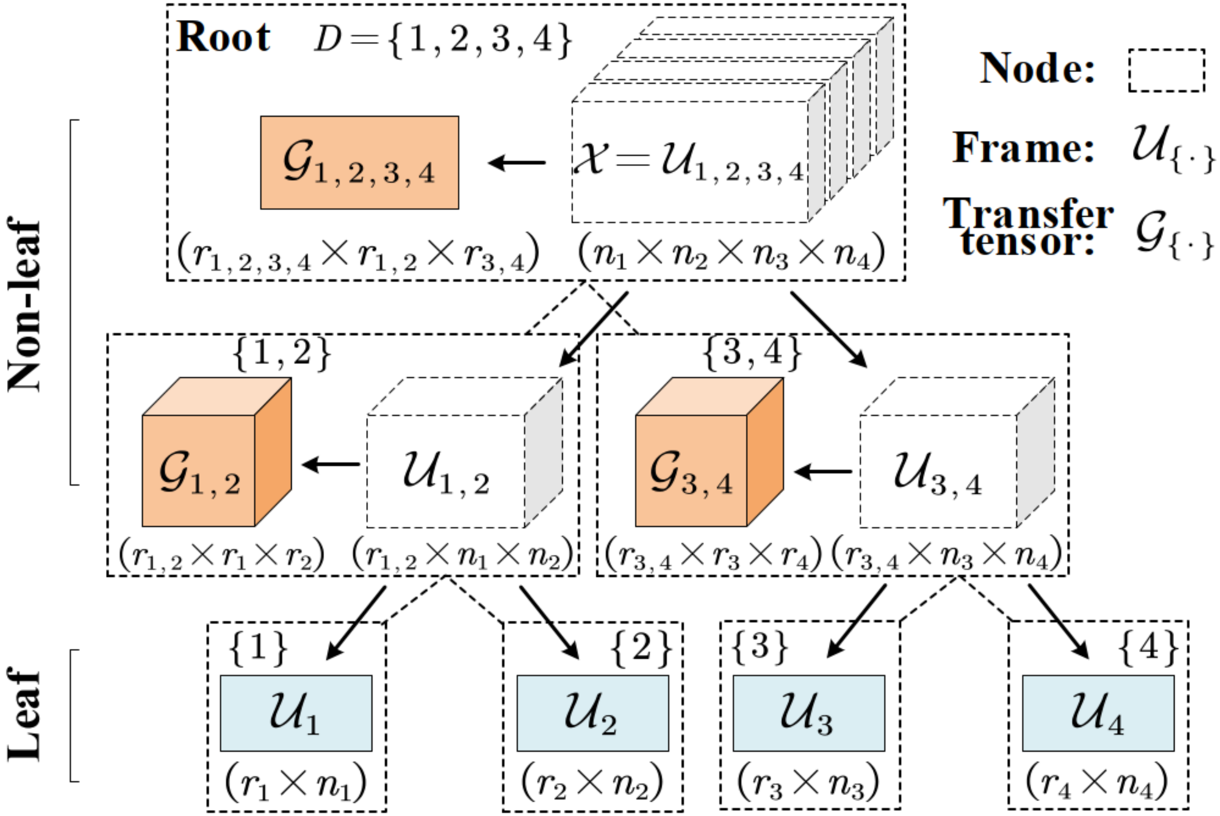
Architecture



Benefits of HT-RNN

- **Parameter Sharing.** The decomposed parameters in the same level can perform parameter sharing via transfer tensors.
- **Stronger Representation Power.** the hierarchical structure imposed on the input-to-hidden layers makes RNNs can exploit and extract the important representation and pattern from high-dimensional data in a much more hierarchical way, thereby significantly improving RNN models' representation capability.
- **lower storage and computational costs.** HT decomposition inherently provides higher complexity reduction on the same-size tensor data with the same selected rank.

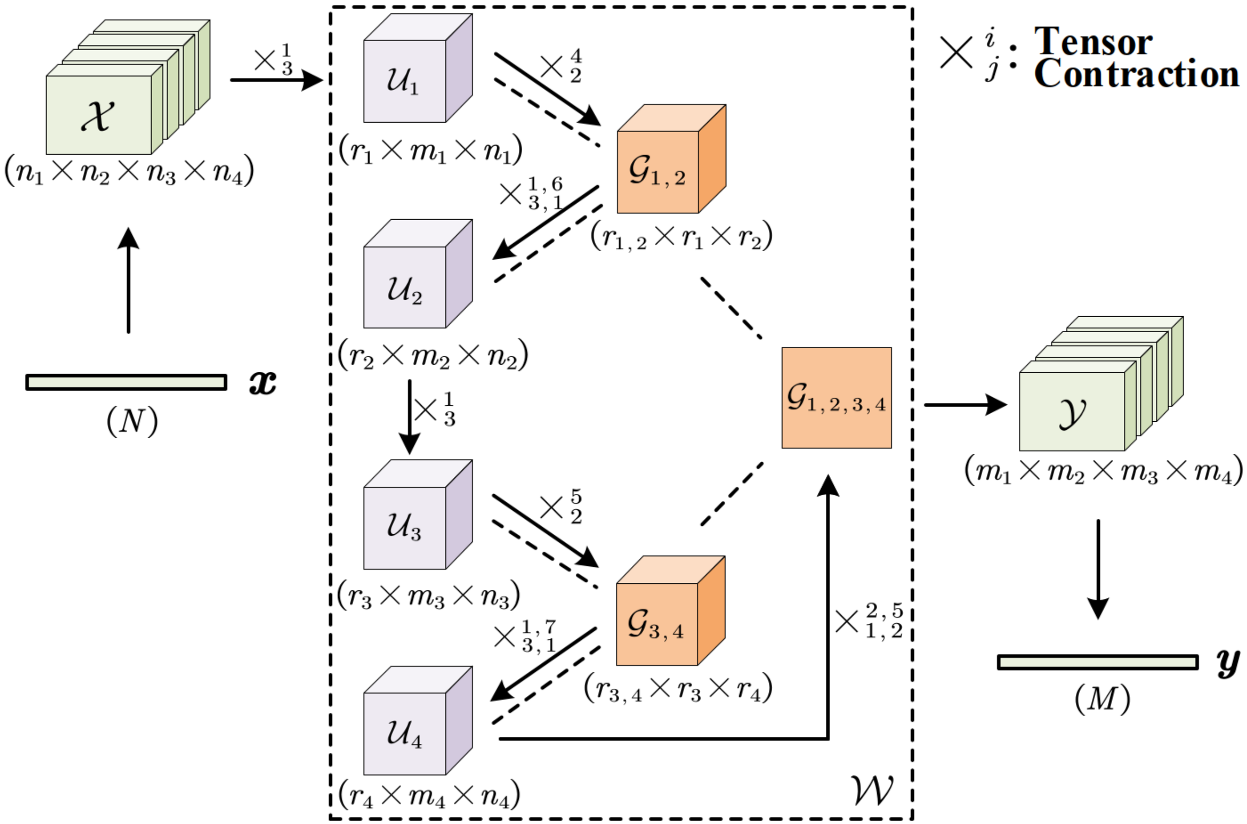
HT Decomposition



$$\mathcal{U}_s = \mathcal{G}_s \times_1^2 \mathcal{U}_{s1} \times_1^2 \mathcal{U}_{s2}$$

Rutgers

Forward Computation in a HT-Layer



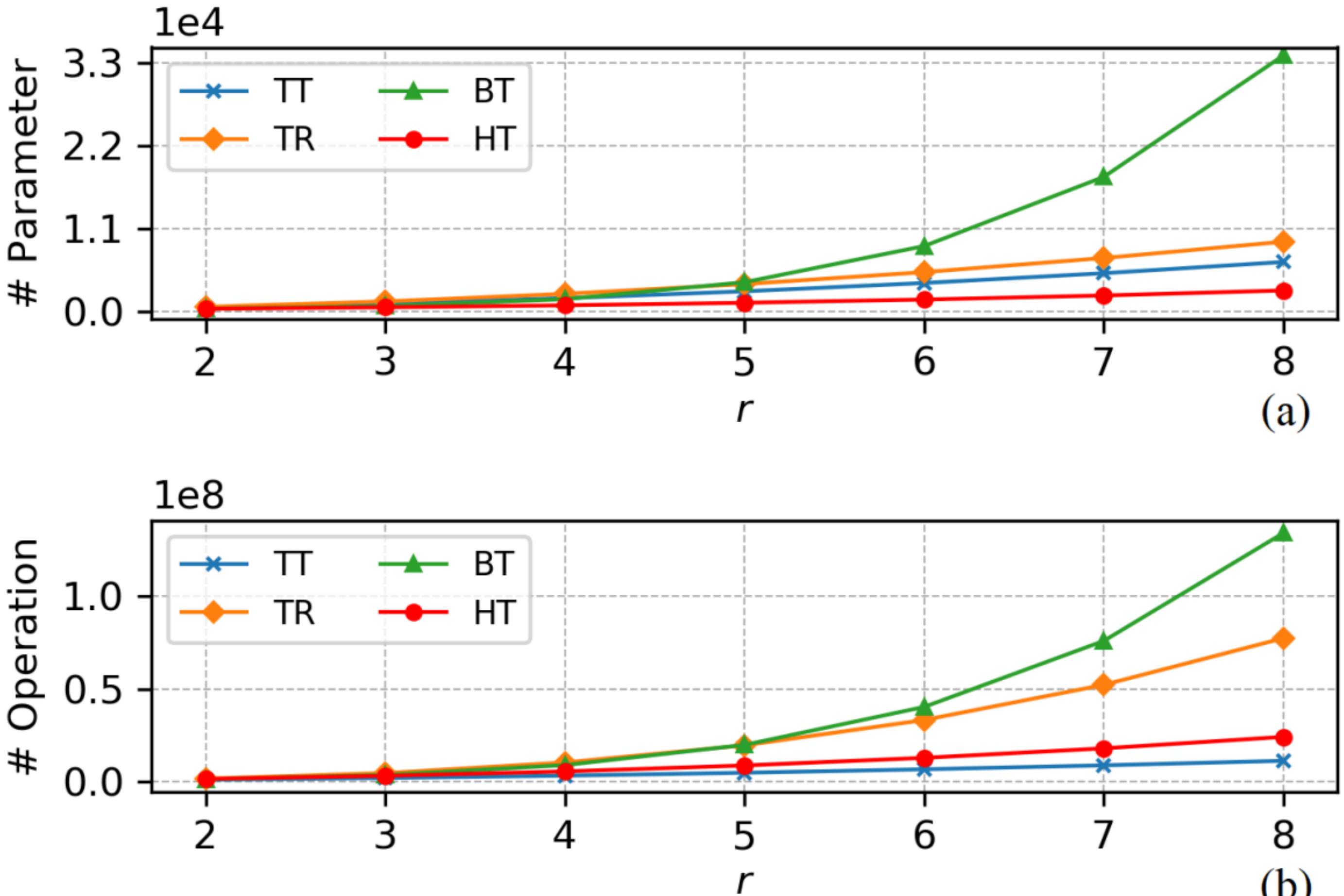
$$\mathcal{Y}_{(\boldsymbol{i})} = \sum_{\boldsymbol{j}} \sum_{k=1}^{r_D} \sum_{p=1}^{r_{D_1}} \sum_{q=1}^{r_{D_2}} (\mathcal{G}_D)_{(k,p,q)}$$

$$\cdot (\mathcal{U}_{D_1})_{(p,\varphi_{D_1}(\boldsymbol{i},\boldsymbol{j}))} (\mathcal{U}_{D_2})_{(q,\varphi_{D_2}(\boldsymbol{i},\boldsymbol{j}))} \mathcal{X}_{(\boldsymbol{j})}$$

Complexity Analysis

Model	Space	Time
RNN FP RNN BP	$\mathcal{O}(NM)$	$\mathcal{O}(NM)$ $\mathcal{O}(NM)$
TT-RNN FP TT-RNN BP	$\mathcal{O}(dmnr^2)$	$\mathcal{O}(dmr^2N)$ $\mathcal{O}(d^2mr^4N)$
TR-RNN FP TR-RNN BP	$\mathcal{O}(dmnr^2)$	$\mathcal{O}(dr^3N + dr^3M)$ $\mathcal{O}(d^2r^5N + nd^2r^5M)$
BT-RNN FP BT-RNN BP	$\mathcal{O}(dmnr + r^d)$	$\mathcal{O}(dmr^dNC)$ $\mathcal{O}(d^2mr^dNC)$
HT-RNN FP HT-RNN BP	$\mathcal{O}(dmnr + dr^3)$	$\mathcal{O}(dmr^2N + dr^3N)$ $\mathcal{O}(d^2mr^5N + d^2r^6N)$

Complexity Analysis



(a)

(b)

Experiments (End-to-End)

- UCF11

Model	CR	# Param.	Accuracy (%)
LSTM	1	59M	69.7
TT-LSTM (ICML-17)	17,554×	3,360	79.6
BT-LSTM (CVPR-18)	17,414×	3,387	85.3
TR-LSTM (AAAI-19)	34,193×	1,725	86.9
HT-LSTM (Ours)	**47,375**×	**1,245**	**87.2**

- Youtube Celebrities

Model	CR	# Param.	Accuracy (%)
LSTM	1×	59M	33.2
TT-GRU	17,723×	3,328	80.0
TT-LSTM	17,388×	3,392	75.5
HT-LSTM (Ours)	**72,818**×	**810**	**88.1**

Experiments (with Pretrained CNN)

- UCF11

Model	**Accuracy** (%)
[Wang *et al.*, 2015]	84.2
[Sharma *et al.*, 2015]	86.0
[Cho *et al.*, 2014]	88.0
[Gammulle *et al.*, 2017]	94.6
CNN + LSTM [Pan *et al.*, 2019]	92.3
CNN + TR-LSTM [Pan *et al.*, 2019]	93.8
CNN + HT-LSTM (Ours)	**98.1**

- HMDB51

Model	**Accuracy** (%)
[Wang *et al.*, 2015]	63.2
[Feichtenhofer *et al.*, 2016]	56.8
[Carreira and Zisserman, 2017]	RGB + Flow: **66.4** RGB: 49.8
CNN + LSTM [Pan *et al.*, 2019]	62.9
CNN + TR-LSTM [Pan *et al.*, 2019]	63.8
CNN + HT-LSTM (Ours)	64.2

Conclusion

- A new RNN compression approach using Hierarchical Tucker (HT) decomposition
- The HT-based RNN models exhibit strong hierarchical structure
- The storage and computational costs are lower than SOTAs
- experiments on different datasets show that, our proposed HT-LSTM models significantly outperform the state-of-the-art compressed RNN models in terms of both compression ratio and test accuracy.