Efficient Sequence Learning with Group Recurrent Networks

[1,2]Fei Gao, [3]Lijun Wu, [2]Li Zhao, [2]Tao Qin, [1]Xueqi Cheng and [2]Tie-Yan Liu

[1]Chinese Academy of Sciences [2]Microsoft Research [3]Sun Yat-sen University

NAACL HLT
New Orleans, Louisiana | June 1-June 6, 2018

1. Motivation

Recurrent Neural Networks have been widely used for sequence learning, like LM, NMT.
However, successful RNN models are usually big, which are time-consuming and memory-costly.
Our work aims to design an **efficient RNN architecture** for sequence learning with **better parameter efficiency**.
Main approach (an divide-and-conquer approach):

- **1.** Divide the hidden states into different disjoint groups to learn **intra-group** temporal correlation efficiently
- **2.** Rearrange the representation to recover the **inter-group** temporal correlation with almost no cost

2. Architecture

a. Group recurrent layer for ***intra-group*** correlation

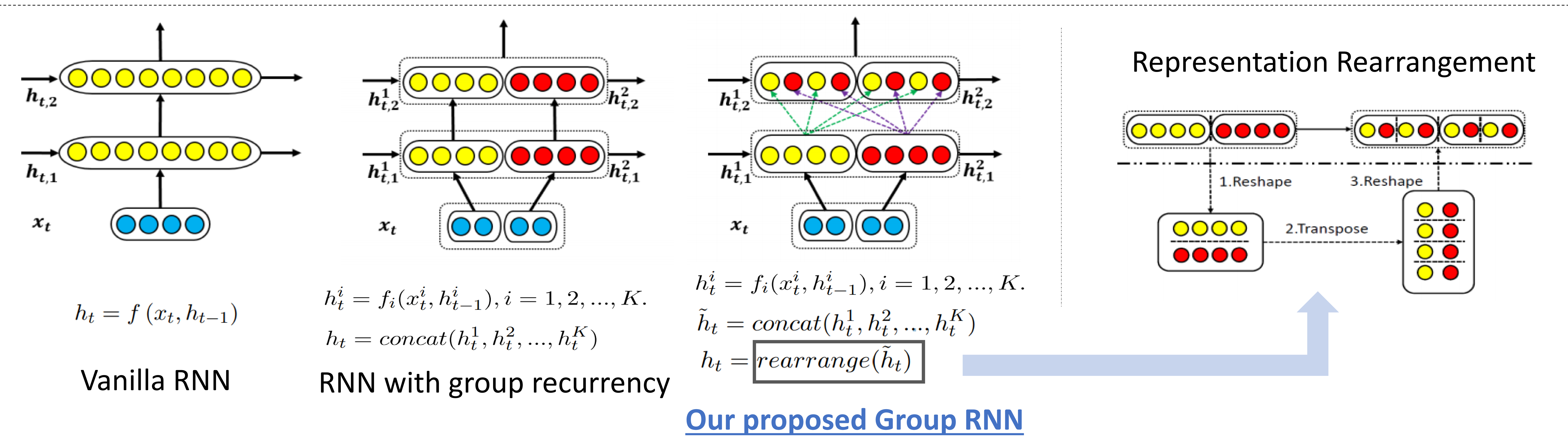
* Split input and hidden states in to K disjoint groups
* Each group performs recurrent learning independently

👍 K times more efficient w.r.t parameter and computation

👎 Representation ability drops since failing to capture the inter-group correlation

b. Representation rearrangement layer for ***inter-group*** correlation

* rearrange the hidden representation, to make sure the subsequent layers, or time steps, can see features from all input groups
* The operation is parameter-free and simple



3. Experiments

Language model: PTB dataset

Model	Parameters	Validation Set	Test Set
LSTM (Zaremba et al., 2014)	66M	82.2	78.4
2 Group LSTM	**48M**	**82.0**	**78.6**
LSTM + BD (Gal and Ghahramani, 2016)	66M	77.9	75.2
2 Group LSTM + BD	**48M**	**79.9**	**75.8**
LSTM + WT (Press and Wolf, 2017)	51M	77.4	74.3
2 Group LSTM + WT	**33M**	**76.8**	**73.3**
LSTM + BD + WT (Press and Wolf, 2017)	51M	75.8	73.2
2 Group LSTM + BD + WT	**33M**	**75.6**	**71.8**

The effect of representation rearrangement (on PTB dataset)

Group	without	with (improvement)
2	82.5	**78.6 (+4.7%)**
4	86.6	**82.6 (+4.6%)**

NMT: De-En dataset

Model	Params	BLEU
NPMT (Huang et al., 2017)	Unclear	30.08
RNNSearch	6.0M	31.03
2 Group RNNSearch	**4.3M**	**31.08**
4 Group RNNSearch	3.4M	30.96
8 Group RNNSearch	3.0M	30.73
16 Group RNNSearch	2.7M	30.35

NMT: En-De dataset

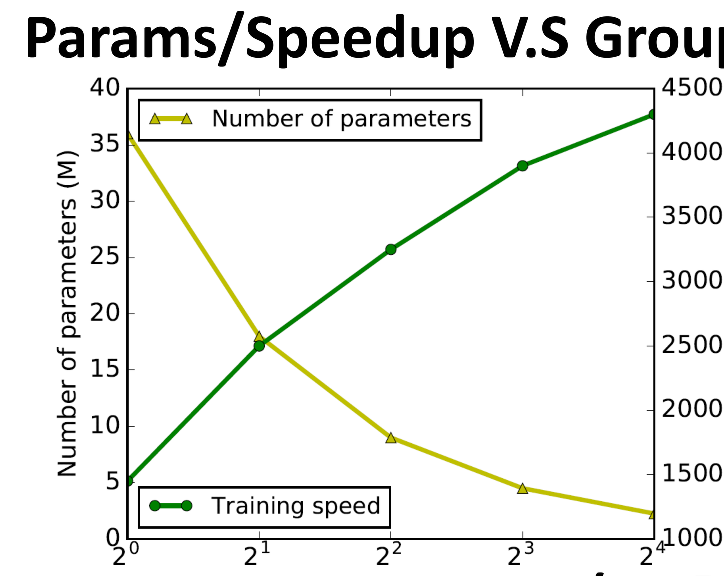
Model	Params	BLEU
DeepLAU (Wang et al., 2017)	Unclear	23.80
GNMT (Wu et al., 2016)	160M‡	24.61
2 Group RNNSearch	111M	23.93
4 Group RNNSearch	78M	23.61

We conduct extensive experiments on language modeling, neural machine translations and abstractive summarization, showing that our method achieves competing performance with much less computing resource.

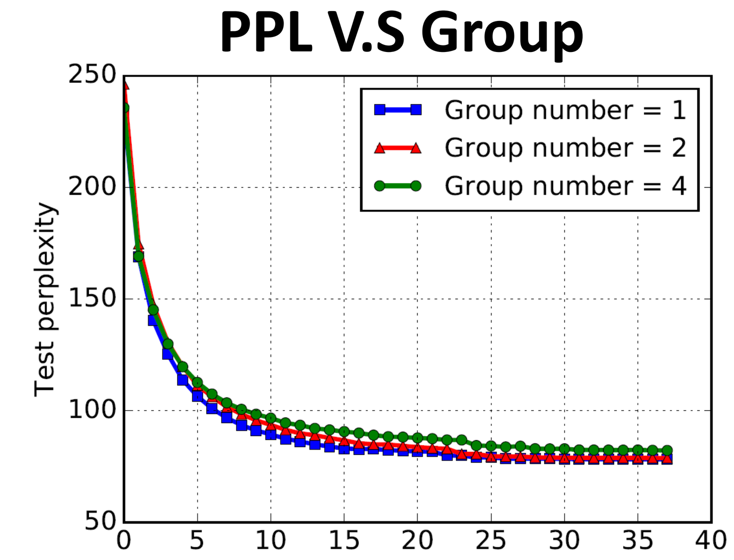
Abstractive summarization: Gigaword dataset

Model	Params	R-1	R-2	R-L
(Rush et al., 2015)	Unclear	29.8	11.9	26.9
(Luong et al., 2015)	Unclear	33.1	14.4	30.7
(Chopra et al., 2016)	Unclear	33.8	15.9	31.1
RNNSearch	24.1M	34.4	15.8	31.8
2 Group RNNSearch	**17.0M**	**34.8**	**15.9**	**32.1**
4 Group RNNSearch	13.5M	34.3	15.7	31.6
8 Group RNNSearch	11.8M	34.3	15.6	31.6
16 Group RNNSearch	10.9M	33.8	15.3	31.2

Params/Speedup V.S Group



PPL V.S Group



(on PTB dataset)

Contact
feiga@microsoft.com