Recurrent Residual Learning for Sequence Classification

Yiren Wang, University of Illinois at Urbana-Champaign
yiren@illinois.edu

Fei Tian, Microsoft Research
fetia@microsoft.com

1. Motivation

LSTM

Gating Mechanism

- Controls information flow with parametric gating units
- Handles gradient vanish problem in vanilla RNN

Optimization Difficulty

- Needs carefully designed optimization procedure
- Especially for long sequences

ResNet

Residual Connecting Mechanism

- Learns a residual mapping between layers with identity connections
- Enhances information flow with extra skip connections
- Effective for optimizing deep architectures

Can we leverage ResNet to improve LSTM for modeling sequential data?

Sequence Classification Tasks

Model long sequence → Strong info flow over a long time range

New input signals at every time step → Necessary to "forget" the less critical historical information

2. Our Work

Can we incorporate ResNet to improve LSTM, and leverage the strengths of both gating mechanism in LSTM, and residual connecting mechanism in ResNet?

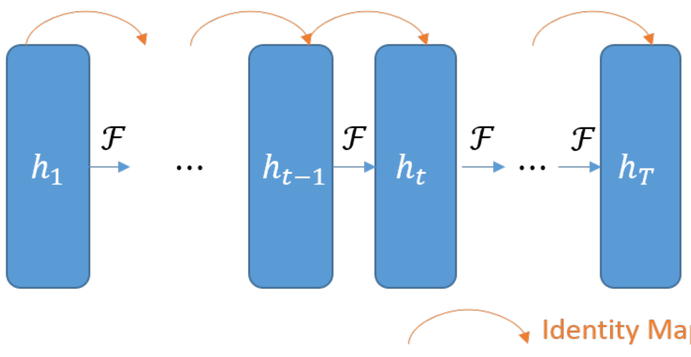
- Introduce the residual connecting mechanism into the recurrent structure, and propose recurrent residual networks for sequence learning.
- Present in-depth analysis of the strengths and limitations of LSTM and ResNet in respect of sequence learning.
- Propose two novel models that incorporate the strengths of both mechanisms.

3. Models

I. Recurrent Residual Network (RRN)

- Leverage residual learning to directly to construct recurrent neural network
- Force a direct information flow in different time steps of RNNs by identity (skip) connections

$$h_t = f(g(h_{t-1}) + \mathcal{F}(h_{t-1}, x_t; W))$$



$\mathcal{F}$: K-deep transformation

$$y_1^t = \sigma(x_t W_1 + h_{t-1} U_1 + b_1)$$
$$y_2^t = \sigma(x_t W_2 + y_1^t U_2 + b_2)$$
$$\dots$$
$$y_K^t = \sigma(x_t W_K + y_{K-1}^t U_K + b_K)$$
$$\mathcal{F}(h_{t-1}, x_t) = y_K^t$$

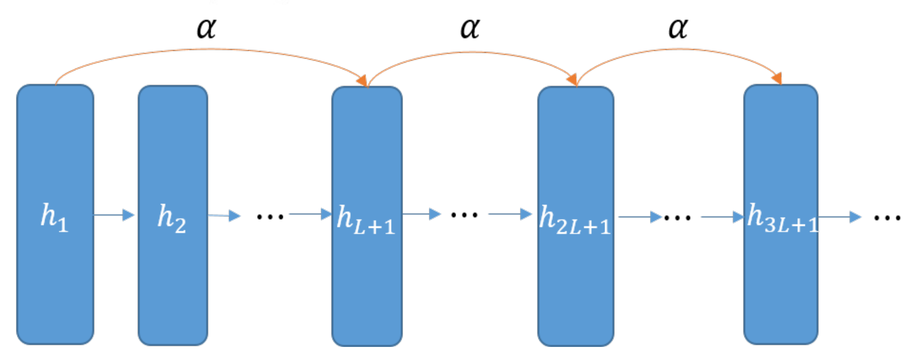
II. Gated Residual RNN

- Incorporate ResNet (residual connections) to improve LSTM
- Combine both gating mechanism and residual connecting mechanism

II – 1. Skip-Connected LSTM (SC-LSTM)

- Introduce identity (SC-LSTM-I) and parametric connections (SC-LSTM-P) into standard LSTM to enhance the information flow
- Add skip connections between two LSTM hidden states with a wide range of distance L

$$h_{t+L} = tanh(c_{t+L}) \otimes o_{t+L} + \alpha h_t$$



* skip connections only exist in time steps { 1,1+L,1+2L,⋯ ,1+⌊T−L−1⌋L }

II – 2. Hybrid Residual LSTM (HRL)

- Combine the two independent signals propagated by LSTM and ResNet respectively to form a signal
- The final hidden state is the mean pooling on the two representations

$$h_T^{HRL} = \frac{1}{2}(h_T^{LSTM} + h_T^{RRN})$$

4. Experiments

I. Experiment Settings

- Datasets

Dataset	Ave. Len	Max Len	#Classes	#Train : #Test
AG's News	34	211	4	120,000 : 7,600
IMDB	281	2,956	2	25,000 : 25,000
20NG	267	11,924	20	11,293 : 7,528
P-MNIST	784	784	10	60,000 : 10,000

- Model Size

	AG's News	*IMDB*	*20NG*	*P-MNIST*
Non-Hybird	256 × 512	256 × 512	500 × 768	1 × 100
Hybrid	256 × 384	256 × 384	256 × 512	1 × 80

II. Experimental Results

Model/Task	*AG's News*	*IMDB*	*20NG*	*P-MNIST*
LSTM	91.76%	88.88%	79.21%	90.64%
RRN	91.19%	89.13%	79.76%	88.63%
SC-LSTM-P	92.01%	90.74%	**82.98%**	94.46%
SC-LSTM-I	**92.05%**	90.67%	81.85%	**94.80%**
LSTM+GRU	91.05%	89.23%	80.12%	90.28%
HRL	91.90%	**90.92%**	81.73%	90.33%

5. Conclusions

- Direct adaptation of ResNet performs well in sequence classification.
- Residual learning significantly improves LSTM's performance, particularly in tasks with fairly long sequences.
- Future work: apply residual learning to other sequential tasks, such as language modeling, and RNN-based neural machine translation.