



AI and Deep Learning

5-2 Recurrent Neural Network II

Zhonglei Wang

WISE, SOE, CHOW

XMU

(Version v1)

Contents

1. GRU

2. LSTM

GRU

1. GRU is short for a gated recurrent unit (Cho et al., 2014)
2. It can capture **long dependence** by a **reset gate** and an **update gate**
 - It introduces a new candidate state
 - Reset gate: Controls how much past information we use to generate the candidate state
 - Update gate: Balances candidate and past hidden states when generating the new hidden state

[Cho, K., van Merriënboer, B., Bahdanau, D. and Bengio, Y.(2014). On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259.]

Reset gate

1. Reset gate controls how much we forget the past

2. Reset gate consists of

- $\Gamma_r^{<i>} = \sigma \left\{ \mathbf{W}_r \cdot \left(\mathbf{x}^{<i>T}, \mathbf{a}^{<i-1>T} \right)^T + \mathbf{b}_r \right\}$ of the same dimension as $\mathbf{a}^{<i-1>}$
- $\tilde{\mathbf{a}}^{<i>} = \tanh \left\{ \mathbf{W} \cdot \left(\mathbf{x}^{<i>T}, (\Gamma_r^{<i>} \circ \mathbf{a}^{<i-1>})^T \right)^T + \mathbf{b} \right\}$
- $\Gamma_r^{<i>}$ controls how much we “forget” the past when obtaining $\tilde{\mathbf{a}}^{<i>}$

3. Model parameters:

- $\mathbf{W}_r, \mathbf{b}_r$ for the reset gate $\Gamma_r^{<i>}$
- $\mathbf{W}, \mathbf{b}$ for obtaining $\tilde{\mathbf{a}}^{<i>}$

Update gate

1. Update gate aggregate information to obtain $\mathbf{a}^{<i>}$

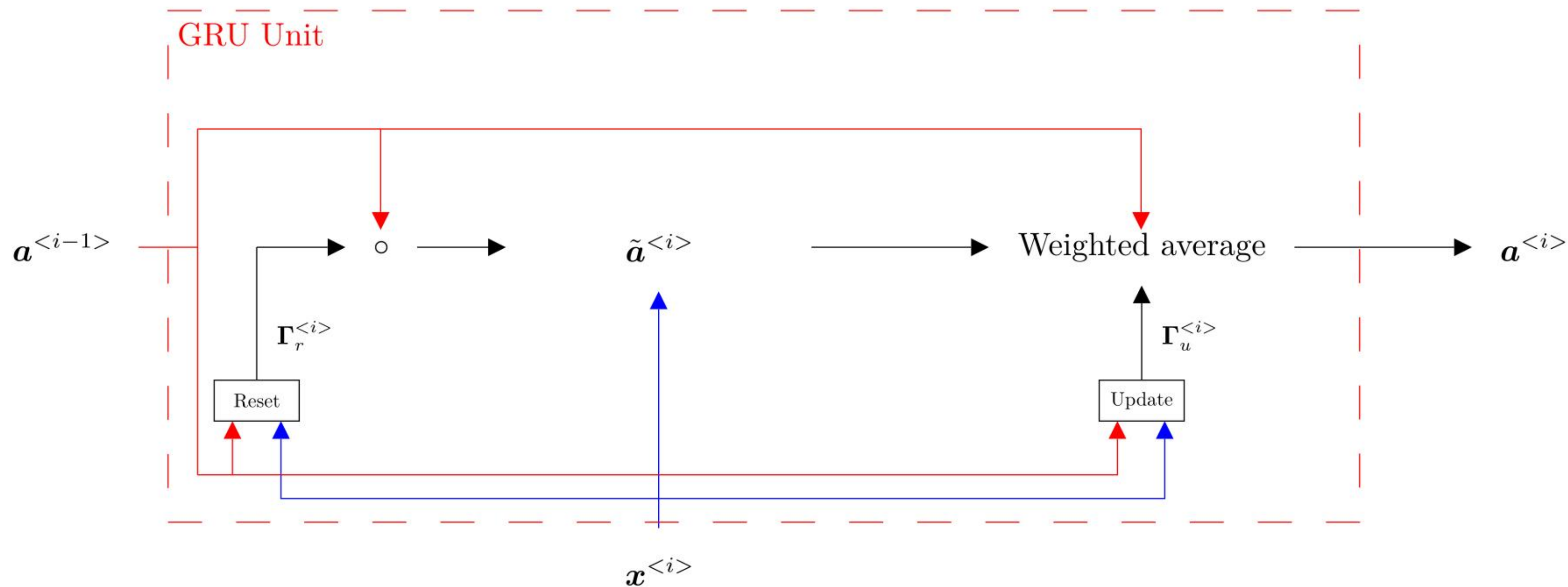
2. Reset gate consists of

- $\Gamma_u^{<i>} = \sigma \left\{ \mathbf{W}_u \cdot \left(\mathbf{x}^{<i>\text{T}}, \mathbf{a}^{<i-1>\text{T}} \right)^{\text{T}} + \mathbf{b}_u \right\}$ of the same dimension as $\mathbf{a}^{<i-1>}$
- $\mathbf{a}^{<i>} = (1 - \Gamma_u^{<i>}) \circ \mathbf{a}^{<i-1>} + \Gamma_u^{<i>} \circ \tilde{\mathbf{a}}^{<i>}$

3. Model parameters:

- $\mathbf{W}_u, \mathbf{b}_u$ for the update gate $\Gamma_u^{<i>}$

Flowchart



LSTM

1. LSTM is short for a long short-term memory unit
2. It allows for capturing dependence of various time scales using an **update gate**, a **forget gate** and an **output gate**
3. First generate a candidate memory

$$\tilde{\mathbf{c}}^{<i>} = \tanh \left\{ \mathbf{W}_c \cdot \left(\mathbf{x}^{<i>^T}, \mathbf{a}^{<i-1>^T} \right)^T + \mathbf{b}_c \right\}$$

- It has the same length with

[部分内容来自于吴恩达教授在 Coursera 上的课程 Deep Learning Specialization]

LSTM

1. Input gate is

$$\mathbf{\Gamma}_i^{<i>} = \sigma \left\{ \mathbf{W}_i \cdot \left(\mathbf{x}^{<i>T}, \mathbf{a}^{<i-1>T} \right)^T + \mathbf{b}_i \right\} \text{ of the same dimension as } \mathbf{c}^{<i-1>}$$

2. Forget gate is

$$\mathbf{\Gamma}_f^{<i>} = \sigma \left\{ \mathbf{W}_f \cdot \left(\mathbf{x}^{<i>T}, \mathbf{a}^{<i-1>T} \right)^T + \mathbf{b}_f \right\} \text{ of the same dimension as } \mathbf{c}^{<i-1>}$$

3. Both gates are used to obtain a final memory

$$\mathbf{c}^{<i>} = \mathbf{\Gamma}_i^{<i>} \circ \tilde{\mathbf{c}}^{<i>} + \mathbf{\Gamma}_f^{<i>} \circ \mathbf{c}^{<i-1>}$$

4. Thus, the memory can be “erased” by the forget gate $\mathbf{\Gamma}_f^{<i>}$

LSTM

1. Output gate is

$$\mathbf{\Gamma}_o^{<i>} = \sigma \left\{ \mathbf{W}_o \cdot \left(\mathbf{x}^{<i>T}, \mathbf{a}^{<i-1>T} \right)^T + \mathbf{b}_o \right\} \text{ of the same dimension as } \mathbf{c}^{<i-1>}$$

2. The output gate is used to obtain

$$\mathbf{a}^{<i>} = \mathbf{\Gamma}_o^{<i>} \circ \mathbf{c}^{<i>}$$

3. Thus, the memory can be further “controlled” by the output gate $\mathbf{\Gamma}_o^{<i>}$

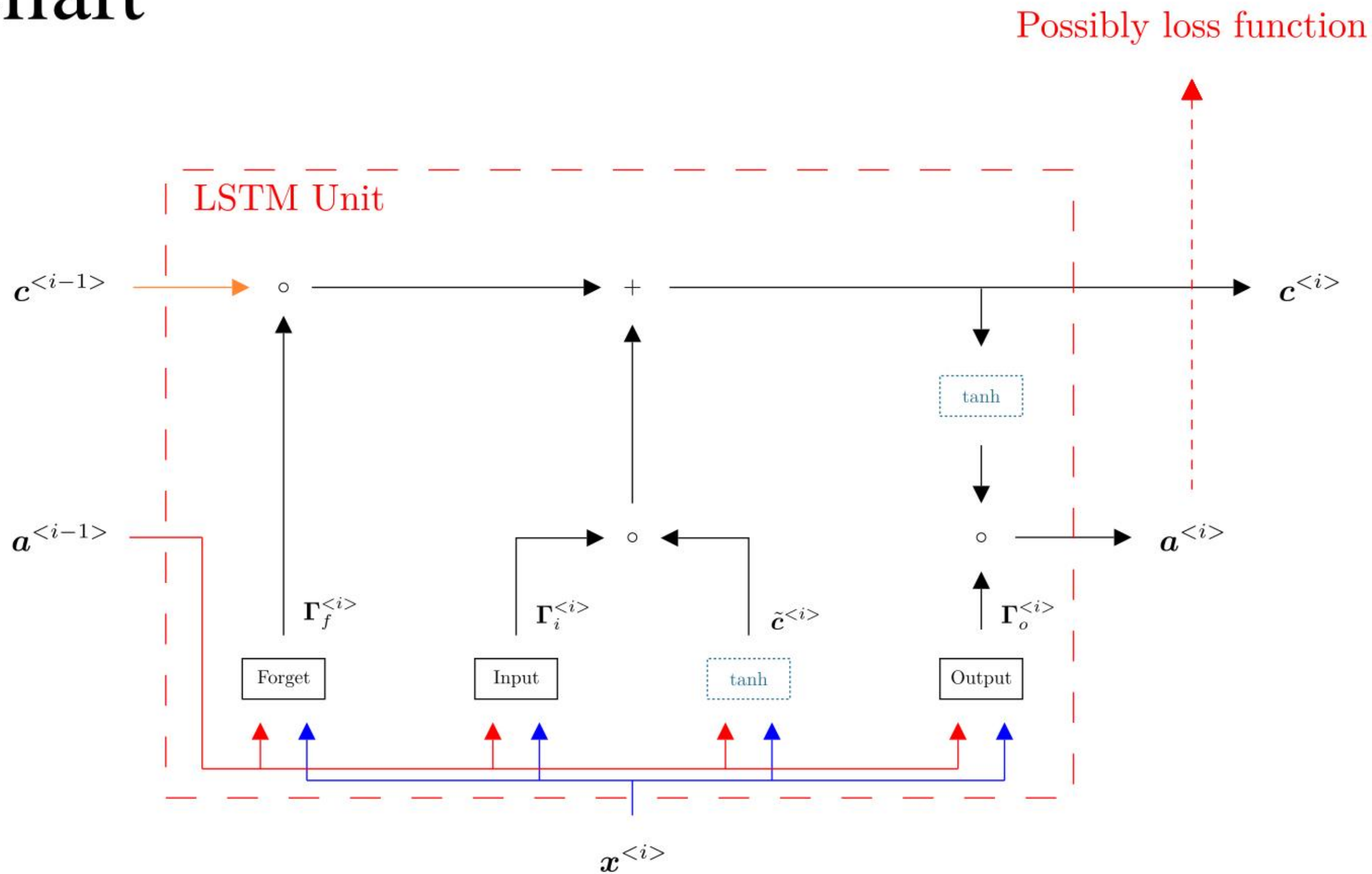
4. Initialize $\mathbf{a}^{<0>} = \mathbf{c}^{<0>} = 0$

LSTM

1. Model parameters are

- $\mathbf{W}_c, \mathbf{b}_c$ for the candidate memory
- $\mathbf{W}_i, \mathbf{b}_i$ for the input gate
- $\mathbf{W}_f, \mathbf{b}_f$ for the forget gate
- $\mathbf{W}_o, \mathbf{b}_o$ for the output gate

Flowchart



Deep RNN

1. We can stack the structure vertically to get a deeper RNN
2. Two dimensions for RNN
 - Time: Recurrent units within the same layer are applied repeatedly across different time steps
 - Layer: Within the same time step, the output of the first-layer recurrent unit is fed into the second layer.