5-2 More RNNs

Zhonglei Wang

WISE and SOE, XMU, 2025

Contents

GRU

1. It is short for a gated recurrent unit (Cho et al., 2014)
2. It can capture dependence of various time scales by a reset gate and an update gate

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

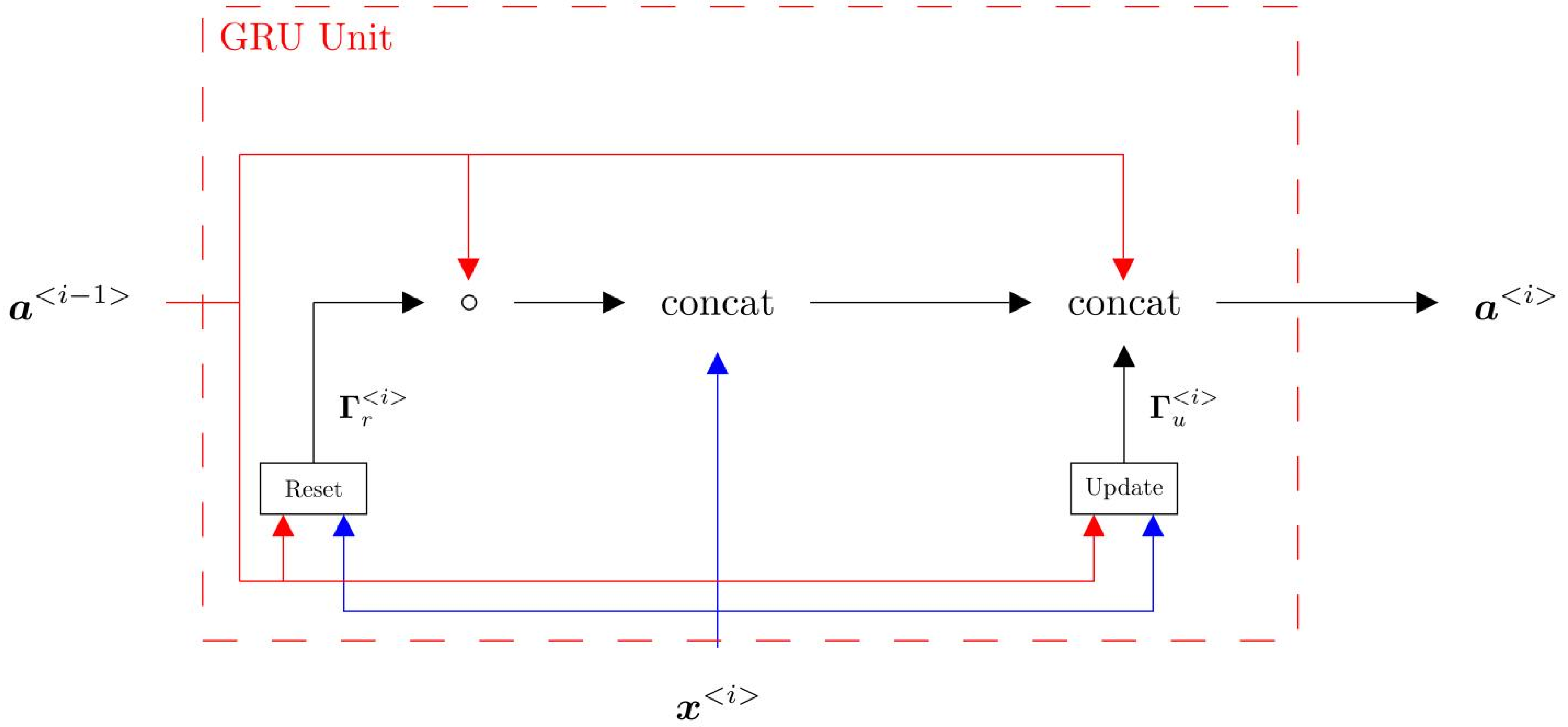
Reset gate

1. Reset gate controls how much we forget the past
2. Reset gate consists of
   - $\boldsymbol{\Gamma}_r^{<i>} = \sigma \left\{ \boldsymbol{W}_r \left(\boldsymbol{x}^{<i>\mathrm{T}}, \boldsymbol{a}^{<i-1>\mathrm{T}} \right)^{\mathrm{T}} + \boldsymbol{b}_r \right\}$: of the same dimension as $\boldsymbol{a}^{<i-1>}$
   - $\tilde{\boldsymbol{a}}^{<i>} = \tanh \left\{ \boldsymbol{W} \left(\boldsymbol{x}^{<i>\mathrm{T}}, (\boldsymbol{\Gamma}_r^{<i>} \circ \boldsymbol{a}^{<i-1>})^{\mathrm{T}} \right)^{\mathrm{T}} + \boldsymbol{b} \right\}$
   - $\boldsymbol{\Gamma}_r^{<i>}$ controls how much we “forget” the past when obtaining $\tilde{\boldsymbol{a}}^{<i>}$
3. Model parameters:
   - $\boldsymbol{W}_r, \boldsymbol{b}_r$ for the reset gate $\boldsymbol{\Gamma}_r^{<i>}$
   - $\boldsymbol{W}, \boldsymbol{b}$ for obtaining $\tilde{\boldsymbol{a}}^{<i>}$

Update gate

1. Update gate aggregate information to obtain $\boldsymbol{a}^{<i>}$
2. Reset gate consists of
   - $\boldsymbol{\Gamma}_u^{<i>} = \sigma\left\{\boldsymbol{W}_u\left(\boldsymbol{x}^{<i>\mathrm{T}}, \boldsymbol{a}^{<i-1>\mathrm{T}}\right)^{\mathrm{T}} + \boldsymbol{b}_u\right\}$: of the same dimension as $\boldsymbol{a}^{<i-1>}$
   - $\boldsymbol{a}^{<i>} = (1 - \boldsymbol{\Gamma}_u^{<i>}) \circ \boldsymbol{a}^{<i-1>} + \boldsymbol{\Gamma}_u^{<i>} \circ \tilde{\boldsymbol{a}}^{<i>}$
   - $\boldsymbol{A} \circ \boldsymbol{B}$: Hadamard matrix production of two matrices $\boldsymbol{A}$ and $\boldsymbol{B} of the same dimension$
3. Model parameters:
   - $\boldsymbol{W}_u, \boldsymbol{b}_u$ for the update gate $\boldsymbol{\Gamma}_u^{<i>}$

Flowchart


GRU Unit
$a^{<i-1>}$
$\circ$
concat
concat
$a^{<i>}$
$\Gamma_r^{<i>}$
$\Gamma_u^{<i>}$
Reset
Update
$x^{<i>}$

LSTM

1. It 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 activation

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

[Partially based on materials used by Dr. Ng in the Deep Learning Specialization course on Coursera DeepLearning.AI]

LSTM

1. An input gate is
$$\boldsymbol{\Gamma}_i^{<i>} = \sigma\left\{\boldsymbol{W}_i\left({\boldsymbol{x}^{<i>}}^{\mathrm{T}}, {\boldsymbol{a}^{<i-1>}}^{\mathrm{T}}\right)^{\mathrm{T}} + \boldsymbol{b}_i\right\} \text{ of the same dimension as } \boldsymbol{c}^{<i-1>}$$
2. A forget gate is
$$\boldsymbol{\Gamma}_f^{<i>} = \sigma\left\{\boldsymbol{W}_f\left({\boldsymbol{x}^{<i>}}^{\mathrm{T}}, {\boldsymbol{a}^{<i-1>}}^{\mathrm{T}}\right)^{\mathrm{T}} + \boldsymbol{b}_f\right\} \text{ of the same dimension as } \boldsymbol{c}^{<i-1>}$$
3. Both gates are used to obtain an activation
$$\boldsymbol{c}^{<i>} = \boldsymbol{\Gamma}_i^{<i>} \circ \tilde{\boldsymbol{c}}^{<i>} + \boldsymbol{\Gamma}_f^{<i>} \circ \tilde{\boldsymbol{c}}^{<i-1>}$$
4. Thus, the memory can be "erased" by the forget gate $\boldsymbol{\Gamma}_f^{<i>}$

LSTM

1. An output gate is

$$\boldsymbol{\Gamma}_o^{<i>} = \sigma\left\{W_o\left(\boldsymbol{x}^{<i>^{\mathrm{T}}}, \boldsymbol{a}^{<i-1>^{\mathrm{T}}}\right)^{\mathrm{T}} + b_o\right\} \text{ of the same dimension as } \boldsymbol{c}^{<i-1>}$$

2. The output gate is used to obtain

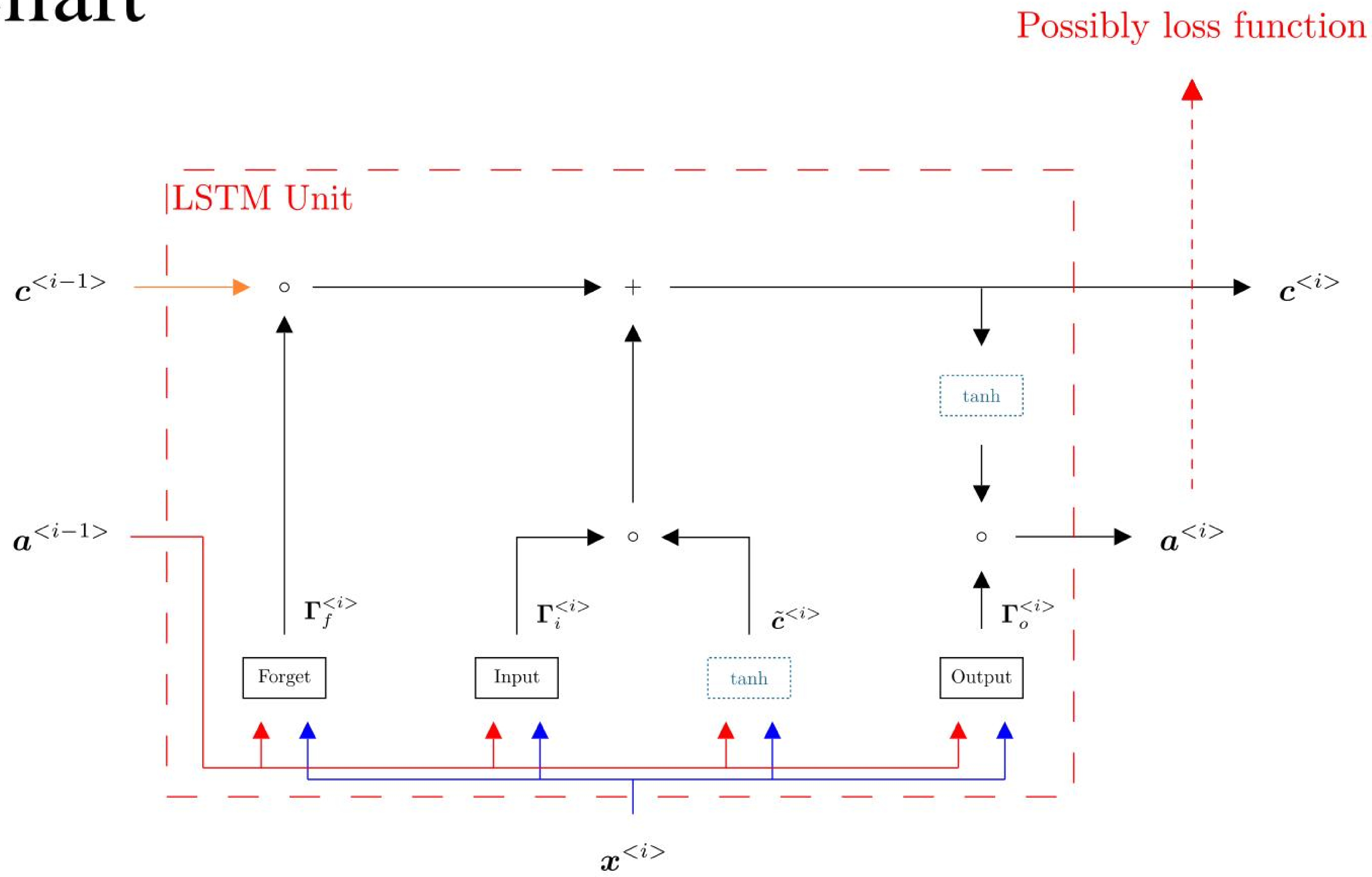
$$\boldsymbol{a}^{<i>} = \boldsymbol{\Gamma}_o^{<i>} \circ \boldsymbol{c}^{<i>}$$

3. Thus, the memory can be further "controlled" by the output gate $\boldsymbol{\Gamma}_o^{<i>}$
4. Initialize $\boldsymbol{a}^{<0>} = \boldsymbol{c}^{<0>} = 0$

LSTM

1. Model parameters are
   - $\boldsymbol{W}_c, \boldsymbol{b}_c$ for the candidate activation
   - $\boldsymbol{W}_i, \boldsymbol{b}_i$ for the input gate
   - $\boldsymbol{W}_f, \boldsymbol{b}_f$ for the forget gate
   - $\boldsymbol{W}_o, \boldsymbol{b}_o$ for the output gate

Flowchart



Possibly loss function
LSTM Unit
$\boldsymbol{c}^{<i-1>}$
$\circ$
$+$
$\boldsymbol{c}^{<i>}$
tanh
$\boldsymbol{a}^{<i-1>}$
$\circ$
$\circ$
$\boldsymbol{a}^{<i>}$
$\boldsymbol{\Gamma}_f^{<i>}$
$\boldsymbol{\Gamma}_i^{<i>}$
$\tilde{\boldsymbol{c}}^{<i>}$
$\boldsymbol{\Gamma}_o^{<i>}$
Forget
Input
tanh
Output
$\boldsymbol{x}^{<i>}$

Deep RNN

1. We can stack the structure vertically to get a deeper RNN