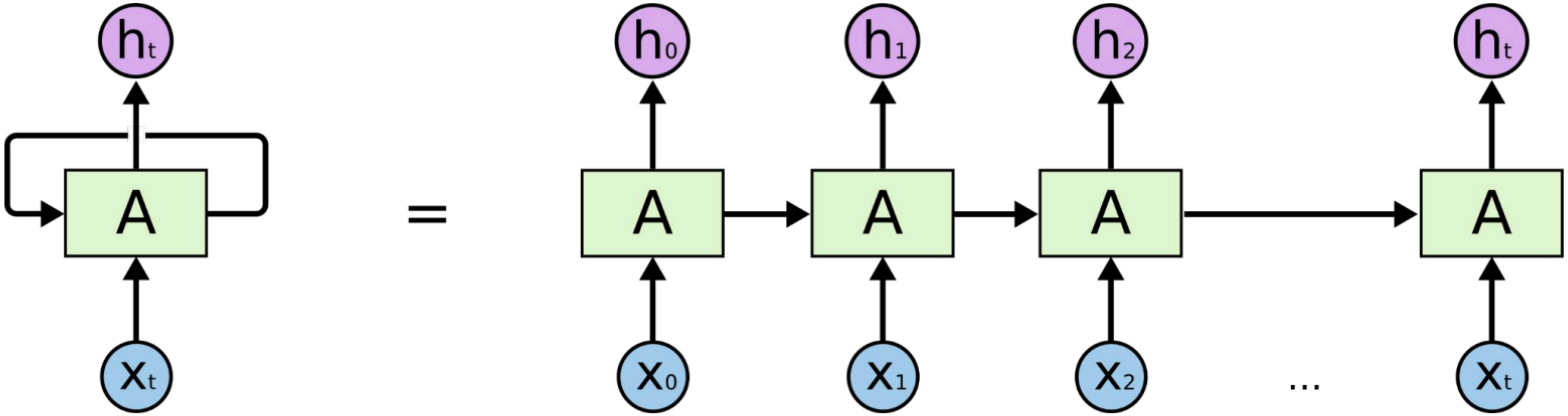
Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation

Authors: Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, Yoshua Bengio

Dayeon Kim

Background - RNN

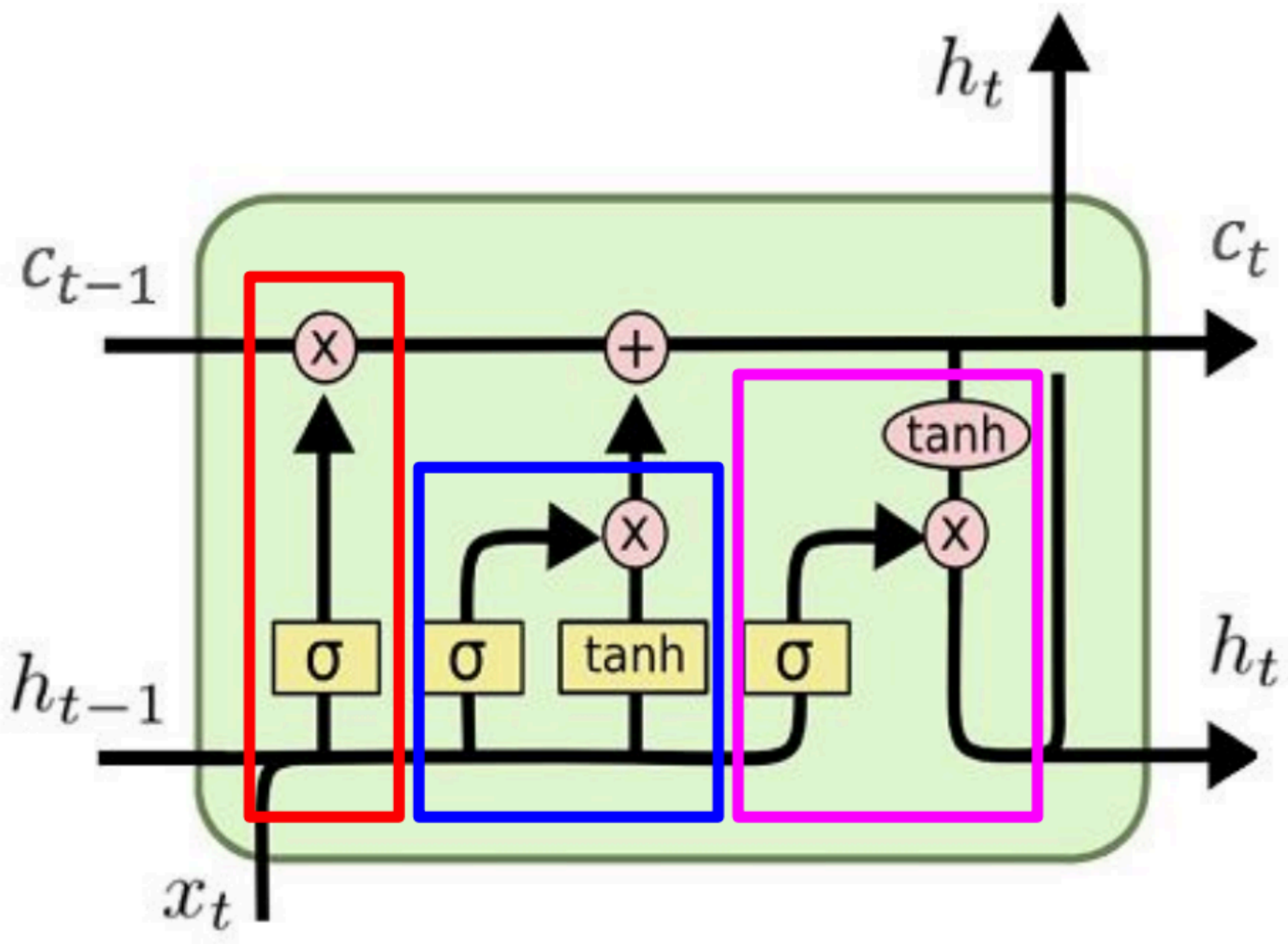


designed to process sequential data by maintaining an internal memory state

feedback connection: hidden state at each time step used as an another input

vanishing/exploding gradient problem

Background - LSTM



- **Forget Gate**
  - how much **old** information to **remember**
- **Input Gate**
  - how much **new** information to **add**
- **Output Gate**
  - how much information to **expose** to next **output** or **hidden state**

Problem Description - SMT(statistical machine translation)

$$argmax_y P(x|y)P(y)$$

maximizes the conditional probability given a source sequence

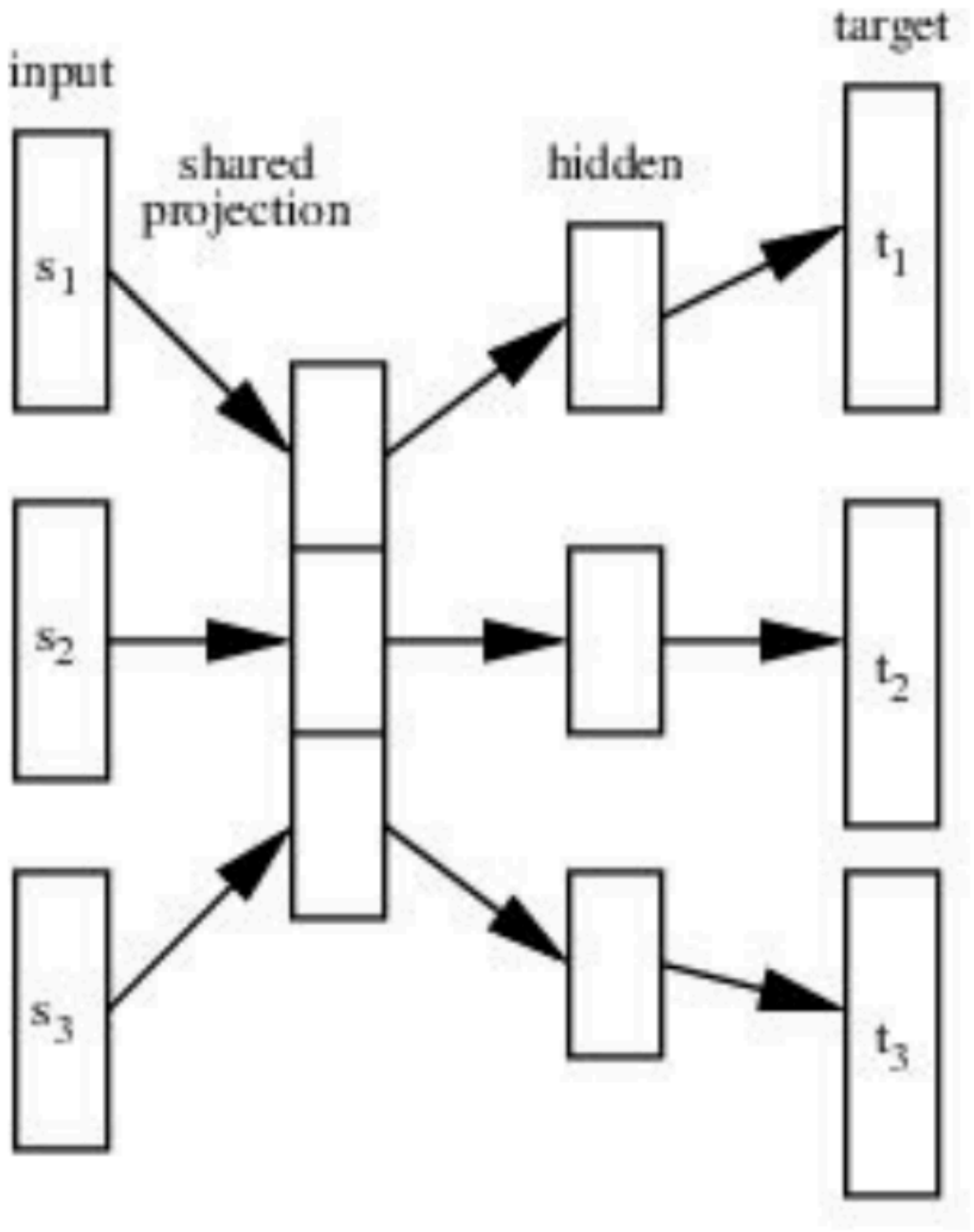
maximize

$$\log p(\mathbf{f} \mid \mathbf{e}) = \sum_{n=1}^{N} w_n f_n(\mathbf{f}, \mathbf{e})$$

$$\text{BLEU} = \min\left(1, \frac{\text{output-length}}{\text{reference-length}}\right)\left(\prod_{i=1}^{4} \text{precision}_i\right)^{\frac{1}{4}}$$

log-linear model with features and weights is optimized to maximize the BLEU score

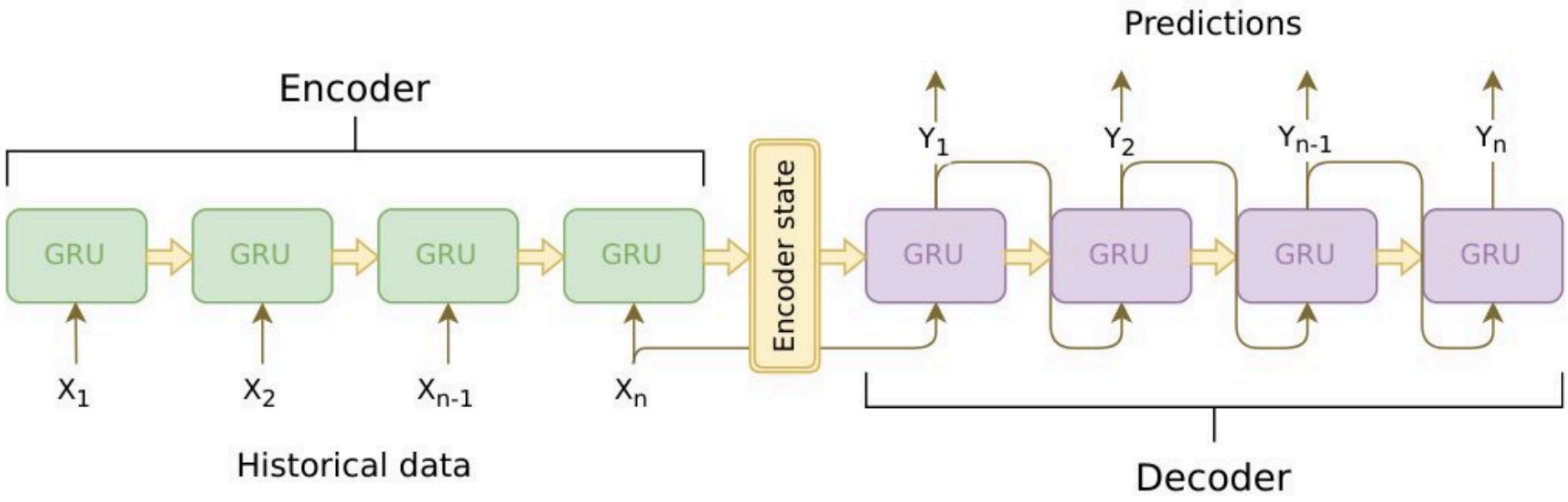
Related works - neural networks in machine translation



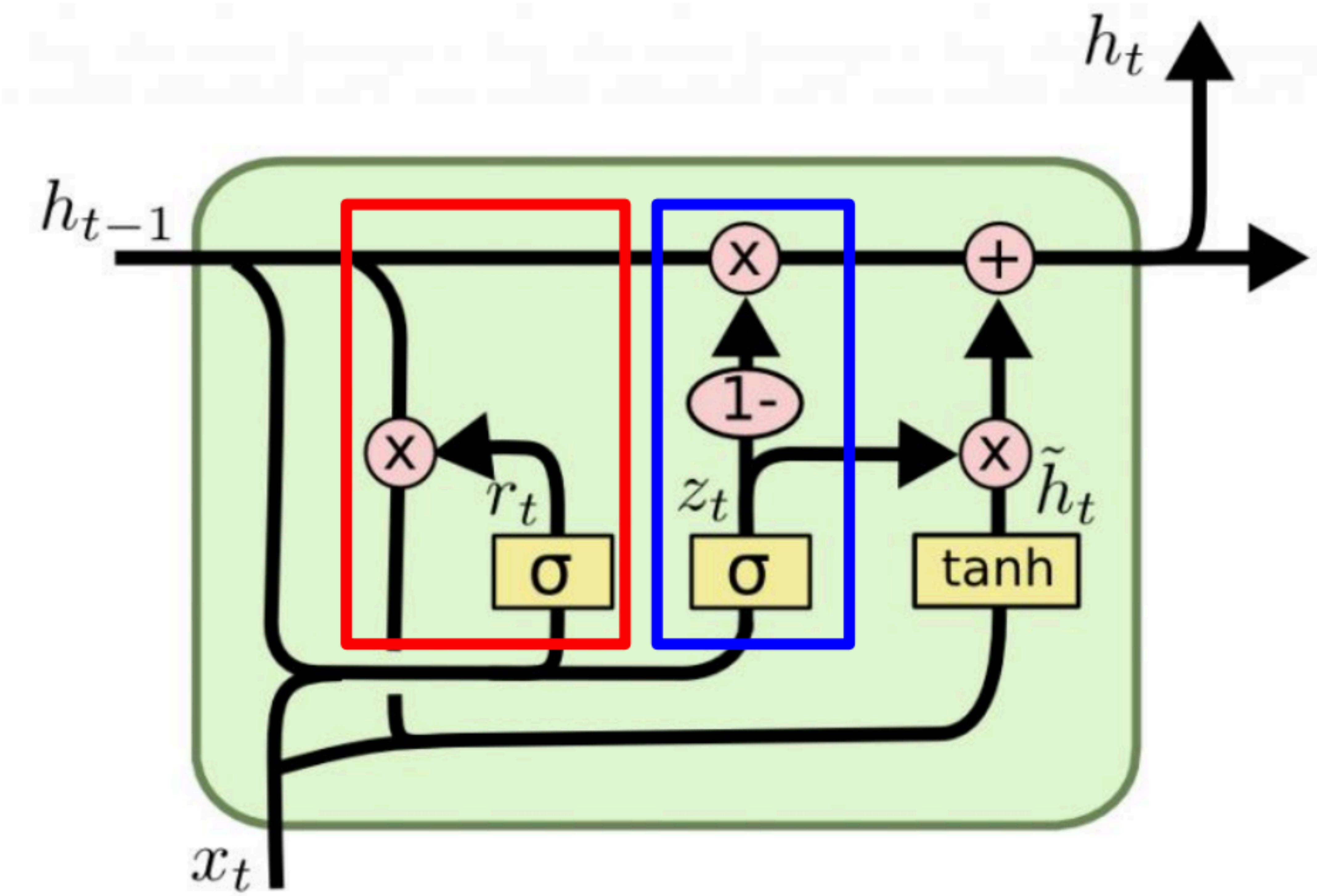
(Schwenk, 2012)

- Fixed input/output length
- Prediction of single word at a time
- Bag-of-words: absence of order information

Method -RNN Encoder & Decoder

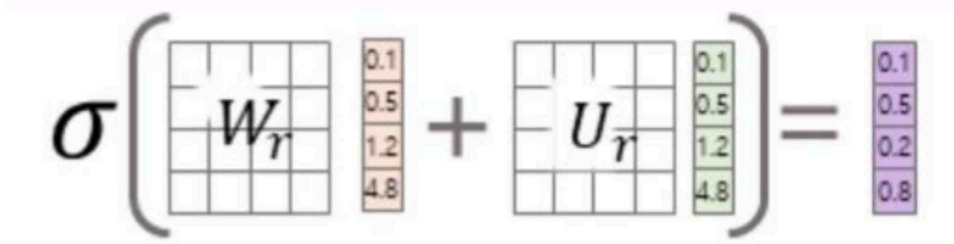


Method -GRU(Gate Recurrent Unit)



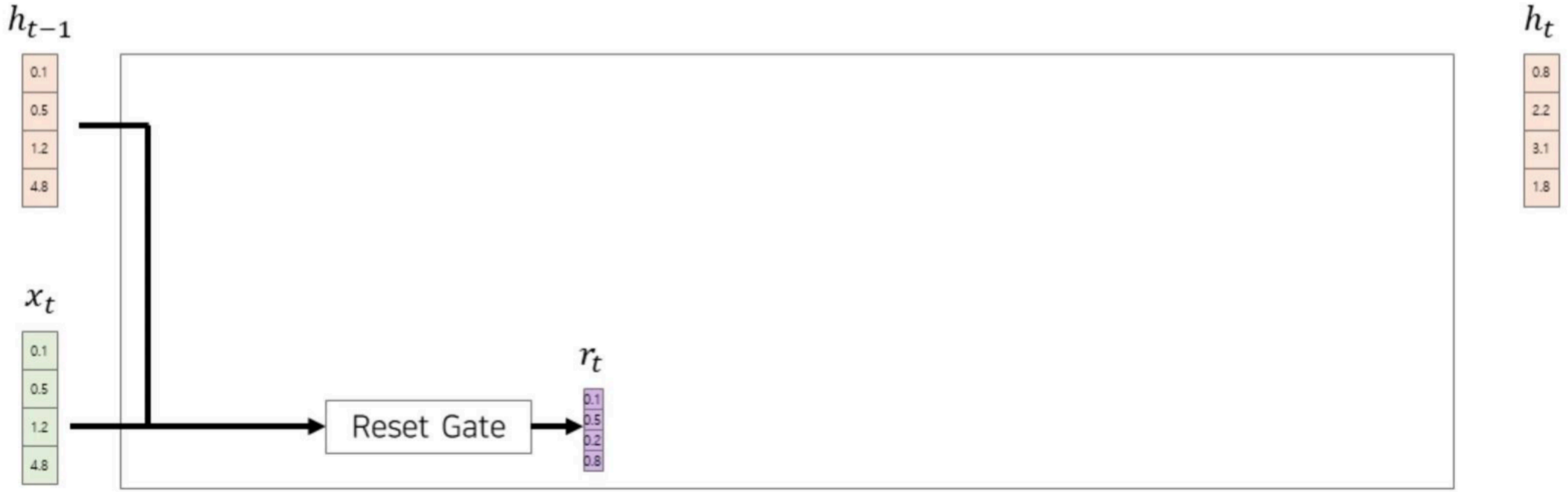
- Reset Gate
  - how much **previous** information to **forget**
- Update Gate
  - how much previous information from **reset gate** to **carry**

Method -GRU(Gate Recurrent Unit)



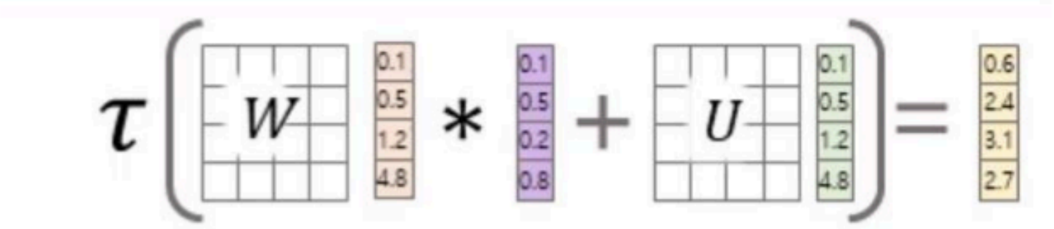
Reset Gate

$$r_j = \sigma\left(\left[\mathbf{W}_r\mathbf{x}\right]_j + \left[\mathbf{U}_r\mathbf{h}_{\langle t-1\rangle}\right]_j\right)$$

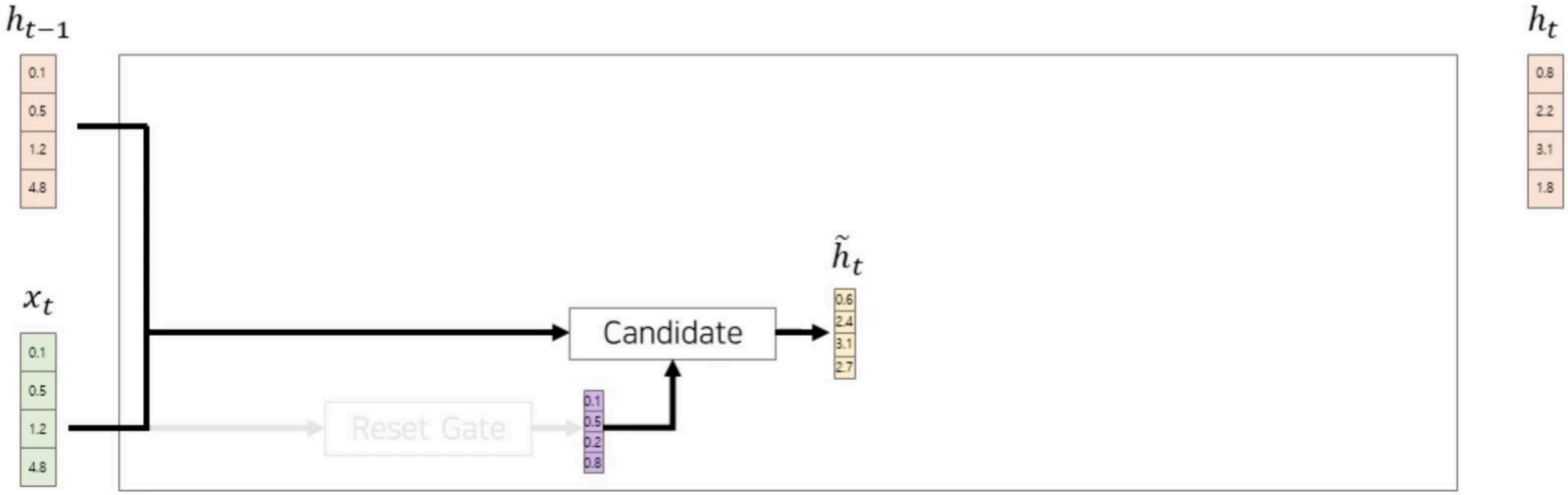


Method -GRU(Gate Recurrent Unit)

Candidate Gate

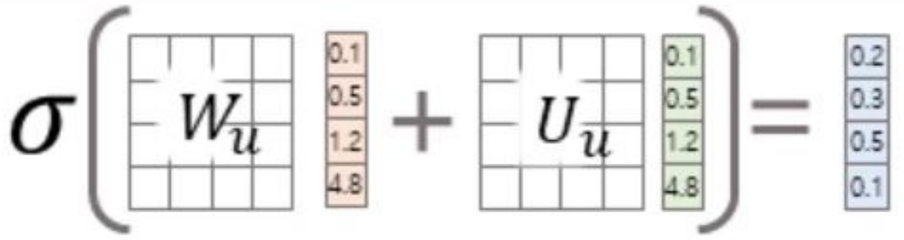


$$\tilde{h}_j^{\langle t\rangle} = \phi\left([\mathbf{W}\mathbf{x}]_j + [\mathbf{U}(\mathbf{r}\odot\mathbf{h}_{\langle t-1\rangle})]_j\right)$$

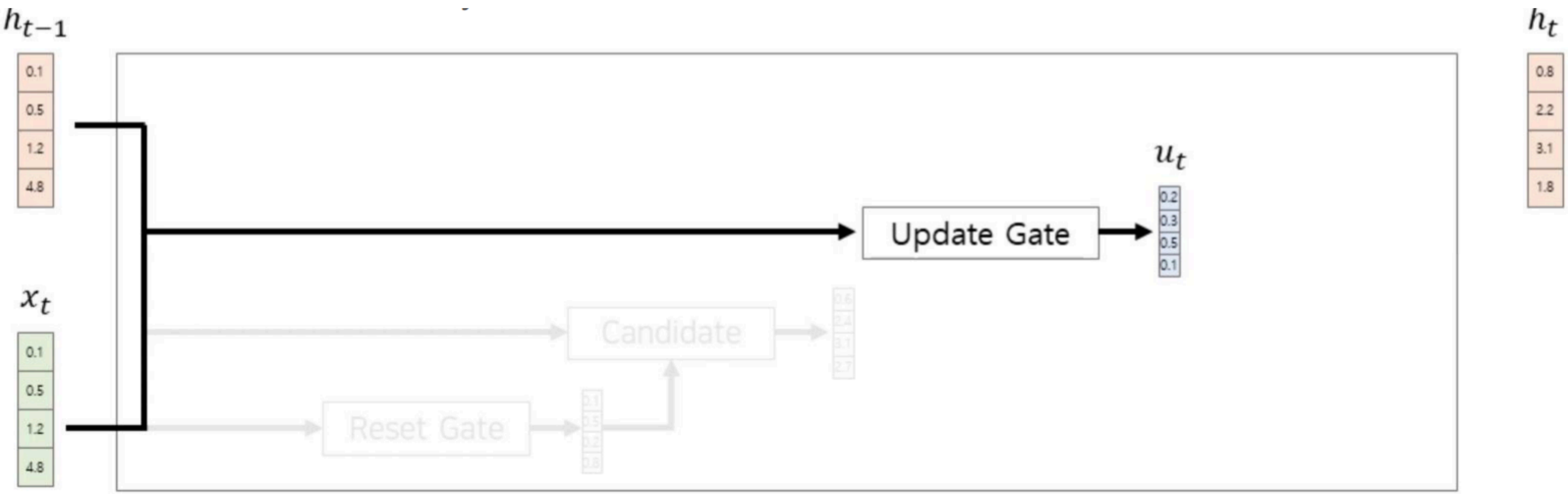


Method -GRU(Gate Recurrent Unit)

Update Gate



$$z_j = \sigma\left([\mathbf{W}_z\mathbf{x}]_j + \left[\mathbf{U}_z\mathbf{h}_{\langle t-1\rangle}\right]_j\right)$$



Method -GRU(Gate Recurrent Unit)

Hidden State

$$\left(\begin{bmatrix}1.0\\1.0\\1.0\\1.0\end{bmatrix} - \begin{bmatrix}0.2\\0.3\\0.5\\0.1\end{bmatrix}\right) * \begin{bmatrix}0.1\\0.5\\1.2\\4.8\end{bmatrix} + \begin{bmatrix}0.2\\0.3\\0.5\\0.1\end{bmatrix} * \begin{bmatrix}0.6\\2.4\\3.1\\2.7\end{bmatrix} = \begin{bmatrix}0.8\\2.2\\3.1\\1.8\end{bmatrix}$$

$$h_j^{\langle t \rangle} = z_j h_j^{\langle t-1 \rangle} + (1 - z_j)\tilde{h}_j^{\langle t \rangle}$$