GRU

1. GRU

1.1. Mathematical Formulation

GRU is a variant of LSTM, which is more efficient and easier to train.

It has less gates compared to LSTM, which makes it more efficient.

1. Update Gate z_t: Control how much of the new candidate hidden state should be used to update the hidden state and how much of the previous hidden state should be kept.

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z)$$

2. Reset Gate r_t: Control how much of the previous hidden state should be used to update the candidate hidden state.

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r)$$

3. Candidate Hidden State $\tilde{h}_t$: It is obtained by the current input and the result of the hadamard product of the reset gate and the previous hidden state.

$$\tilde{h}_t = \tanh(W_h \cdot [r_t \odot h_{t-1}, x_t] + b_h)$$

4. Hidden State h_t: The hidden state is updated by the update gate and the candidate hidden state.

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$$

For a GRU of $[0, T]$ time steps, the whole forward process can be described as:

Definition 1.1
At timestep t:

1. Gate update:
   - Update Gate: $z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z)$
   - Reset Gate: $r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r)$
2. Candidate Hidden State: $\tilde{h}_t = \tanh(W_h \cdot [r_t \odot h_{t-1}, x_t] + b_h)$
3. Hidden State: $h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$
4. (Optional) Output: $y_t = f\big(W_y h_t + b_y\big)$