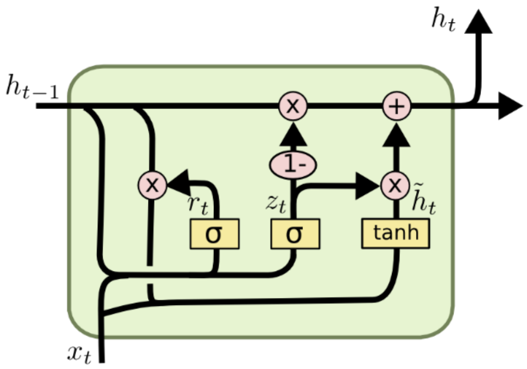
Gated Recurrent Units (GRUs)

- The Gated Recurrent Unit (GRU) can be viewed as a simplification of the LSTM, which does not use explicit cell states.
- Another difference is that the LSTM directly controls the amount of information changed in the hidden state using separate forget and output gates. On the other hand, a GRU uses a single reset gate to achieve the same goal.
- However, the basic idea in the GRU is quite similar to that of an LSTM, in terms of how it partially resets the hidden states.



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

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

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

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

First, let's introduce the notations:

- Just as the LSTM uses input, output, and forget gates to decide how much of the information from the previous time-stamp to carry over to the next step, the GRU uses the update and the reset gates.
- The reset gate r decides how much of the hidden state to carry over from the previous time-stamp for a matrix-based update (like a recurrent neural network).
- The update gate z decides the relative strength of the contributions of this matrix-based update and a more direct contribution from the hidden vector h $_{t-1}$ at the previous time-stamp.
- The update gate of the GRU simultaneously performs the role of the input and forget gates in the LSTM in the form of z and 1 − z, respectively.

How GRU Works

The first step is to generate what is known as the **candidate hidden state**.

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

- It takes in the input and the hidden state from the previous timestamp t-1 which is multiplied by the reset gate output rt.

- Later passed this entire information to the tanh function, the resultant value is the candidate's hidden state.
- The most important part of this equation is how we are using the value of the reset gate to control how much influence the previous hidden state can have on the candidate state.
- If the value of rt is equal to 1 then it means the entire information from the previous hidden state Ht-1 is being considered.
- Likewise, if the value of rt is 0 then that means the information from the previous hidden state is completely ignored.

Hidden state

- Once we have the candidate state, it is used to generate the current hidden state H_t.
- It is where the Update gate comes into the picture. Now, this is a very interesting equation, instead of using a separate gate like in LSTM in GRU we use a single update gate to control both the historical information which is H_{t-1} as well as the new information which comes from the candidate state.

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

- Similarly, if the value of $(1-z_t)$ is on the second term will become entirely 0 and the current hidden state will entirely depend on the first term i.e the information from the hidden state at the previous timestamp t-1.
-