RNN Notation

$\boldsymbol{X}^{(i)\langle t\rangle}$: the *t*-th element of input sequence in the *i*-th training example
$\boldsymbol{T}_{\boldsymbol{X}}^{(i)}$: the input sequence length of the *i*-th training example
$\boldsymbol{Y}^{(i)\langle t\rangle}$: the *t*-th element of output sequence in the *i*-th training example
$\boldsymbol{T}_{\boldsymbol{Y}}^{(i)}$: the output sequence length of the *i*-th training example

RNN Forward Propagation

$\boldsymbol{w}_{\boldsymbol{ax}}$: the weight parameter from $x^{<t>}$ to the hidden layer, shared for every time step
$\boldsymbol{w}_{\boldsymbol{aa}}$: the weight parameter of the horizontal connection, shared for every time step
$\boldsymbol{w}_{\boldsymbol{ya}}$: the weight parameter from the hidden layer to $y^{<t>}$, shared for every time step

Hidden Layer Activation:

$$a^{\langle t\rangle} = \mathrm{g}_1\left(w_{aa}a^{\langle t-1\rangle} + w_{ax}x^{\langle t\rangle} + b_a\right) \leftarrow \underline{\tanh}/\mathrm{ReLU}$$

Simplified Hidden Layer Activation:

$$a^{\langle t\rangle} = \mathrm{g}_1\left(w_a\left[a^{\langle t-1\rangle}, x^{\langle t\rangle}\right] + b_a\right)$$
$$w_a = [w_{aa} \vdots w_{ax}],\ \ \text{shape} = (100, 100 + 10000)$$
$$\left[a^{\langle t-1\rangle}, x^{\langle t\rangle}\right] = \begin{bmatrix} a^{\langle t-1\rangle} \\ \cdots \\ x^{\langle t\rangle} \end{bmatrix},\ \ \text{shape} = (100 + 10000, 1)$$

Output Layer Activation:

$$\hat{y}^{\langle t\rangle} = \mathrm{g}_2\left(w_{ya}a^{\langle t\rangle} + b_y\right) \leftarrow \text{sigmoid/softmax}$$

RNN Backpropagation through Time

Loss Function:

$$\mathcal{L}^{\langle t\rangle}\left(\hat{y}^{\langle t\rangle}, y^{\langle t\rangle}\right) = -y^{\langle t\rangle}\log\left(\hat{y}^{\langle t\rangle}\right) - \left(1 - y^{\langle t\rangle}\right)\log\left(1 - \hat{y}^{\langle t\rangle}\right)$$
$$\mathcal{L}(\hat{y}, y) = \sum_{t=1}^{T_y} \mathcal{L}^{\langle t\rangle}\left(\hat{y}^{\langle t\rangle}, y^{\langle t\rangle}\right)$$

Different Types of RNN

- Many-to-many architecture
  - Name entity recognition
  - Machine translation (T_X != T_Y)
- Many-to-one architecture
  - Sentiment classification
- One-to-one architecture
- One-to-many architecture
  - music generation

Gated Recurrent Unit (GRU)

Candidate Cell:

$$\tilde{c}^{<t>} = \tanh\left(w_c\left[\Gamma_r \circ c^{\langle t-1\rangle}, x^{\langle t\rangle}\right] + b_c\right)$$

Update Gate:

$$\Gamma_u = \text{sigmoid}\left(w_u\left[c^{\langle t-1\rangle}, x^{\langle t\rangle}\right] + b_u\right)$$

Relevance Gate

$$\Gamma_r = \text{sigmoid}\left(w_r\left[c^{\langle t-1\rangle}, x^{\langle t\rangle}\right] + b_r\right)$$

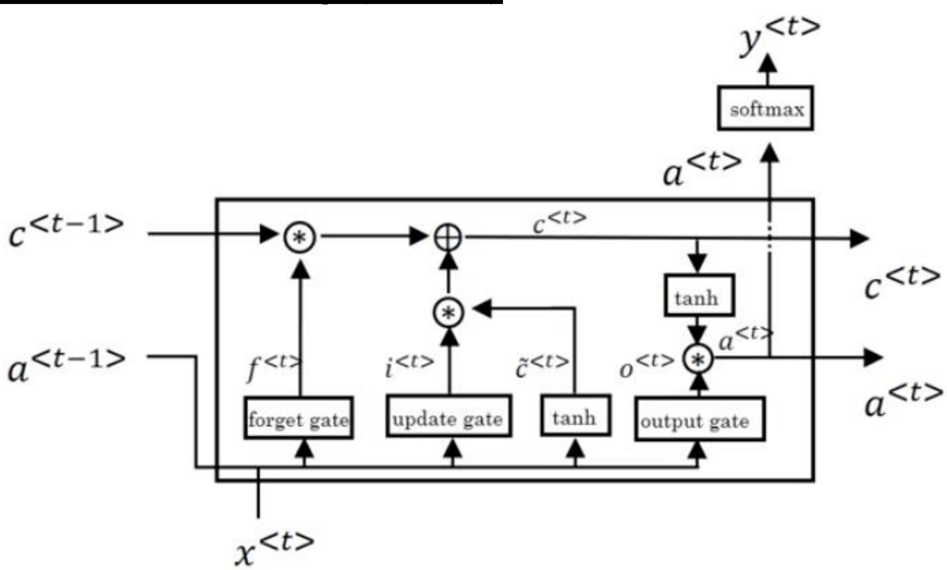
Memory Cell:

$$c^{\langle t\rangle} = \Gamma_u \circ \tilde{c}^{\langle t\rangle} + (1 - \Gamma_u) \circ c^{\langle t-1\rangle}$$

Unit Activation:

$$a^{\langle t\rangle} = c^{\langle t\rangle}$$

Long Short-Term Memory (LSTM)



Candidate Cell:

$$\tilde{c}^{<t>} = \tanh\left(w_c\left[a^{\langle t-1\rangle}, x^{\langle t\rangle}\right] + b_c\right)$$

Update Gate:

$$\Gamma_u = \text{sigmoid}\left(w_u\left[c^{\langle t-1\rangle}, x^{\langle t\rangle}\right] + b_u\right)$$

Forget Gate:

$$\Gamma_f = \text{sigmoid}\left(w_f\left[c^{\langle t-1\rangle}, x^{\langle t\rangle}\right] + b_f\right)$$

Output Gate:

$$\Gamma_o = \text{sigmoid}\left(w_o\left[c^{\langle t-1\rangle}, x^{\langle t\rangle}\right] + b_o\right)$$

Memory Cell:

$$c^{\langle t \rangle} = \Gamma_u \circ \tilde{c}^{\langle t \rangle} + \Gamma_f \circ c^{\langle t-1 \rangle}$$

Unit Activation:

$$a^{\langle t \rangle} = \Gamma_o \circ c^{\langle t \rangle}$$

Deep RNN

$\boldsymbol{a^{[l]\langle t \rangle}}$: the activation value of the *l*-th hidden layer for the *t*-th element

Deep RNN Activation:

$$a^{[l]\langle t \rangle} = \mathrm{g}\left(w_a^{[l]}\left[a^{[l]\langle t-1 \rangle}, a^{[l-1]\langle t \rangle}\right] + b_a^{[l]}\right)$$
$$a^{[0]\langle t \rangle} = x^{\langle t \rangle}$$
$$\hat{y}^{\langle t \rangle} = \mathrm{g}\left(w_{ya} a^{[L]\langle t \rangle} + b_y\right)$$