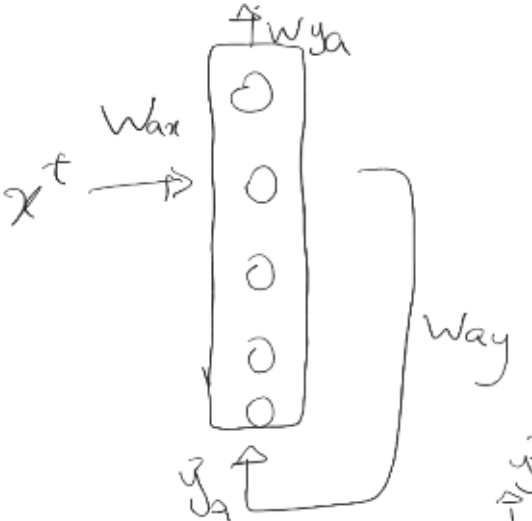
1) what a standard NN does.

for sequences:

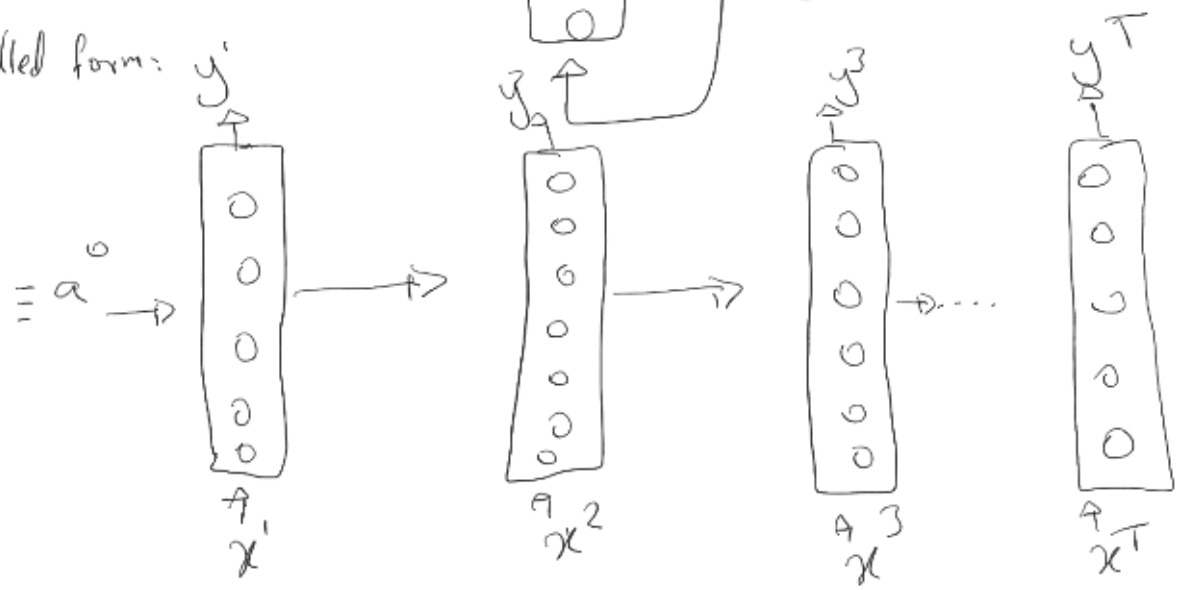
1- input length

2- feature sharing.

2) RNN



Unrolled form:



activation func

$$a^{\langle t \rangle} = f\left(W_{aa} a^{\langle t-1 \rangle} + W_{ax} x^{\langle t \rangle} + b_a\right)$$

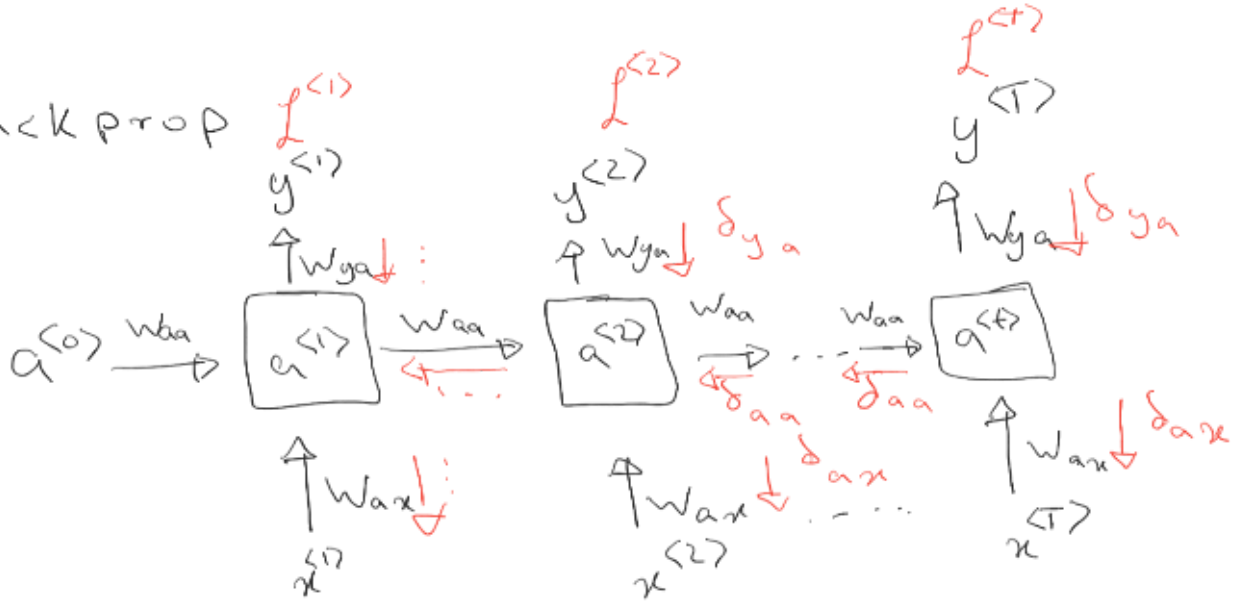
$$y^{\langle t \rangle} = g\left(W_{ya} a^{\langle t \rangle} + b_y\right)$$

activation fun

$$a^{\langle t \rangle} = f\left(W_a \begin{bmatrix} a^{\langle t-1 \rangle} \\ x^{\langle t \rangle} \end{bmatrix} + b_a\right)$$

where $W_a = \begin{bmatrix} W_{aa} \vdots W_{ax} \end{bmatrix}$

Backprop



$$\mathcal{L}^{\langle t \rangle} = -y^{\langle t \rangle} \log \hat{y}^{\langle t \rangle} - (1 - y^{\langle t \rangle}) \log (1 - \hat{y}^{\langle t \rangle})$$

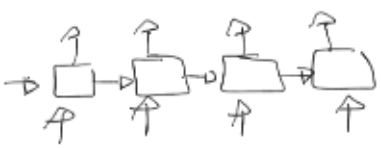
$$\mathcal{L} = \sum_{t=1}^{T} \mathcal{L}^{\langle t \rangle}$$

RNN types from input/outpt lengths

$T_y = T_x = T$

many-to-many

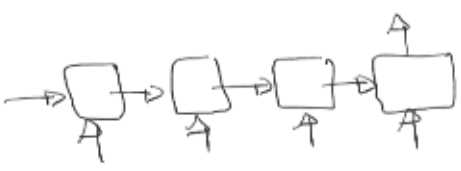
"NER, translation?"



$T_y = 1, T_x = T$

many-to-one

"Text classification"



$T_y = T, T_x = 1$

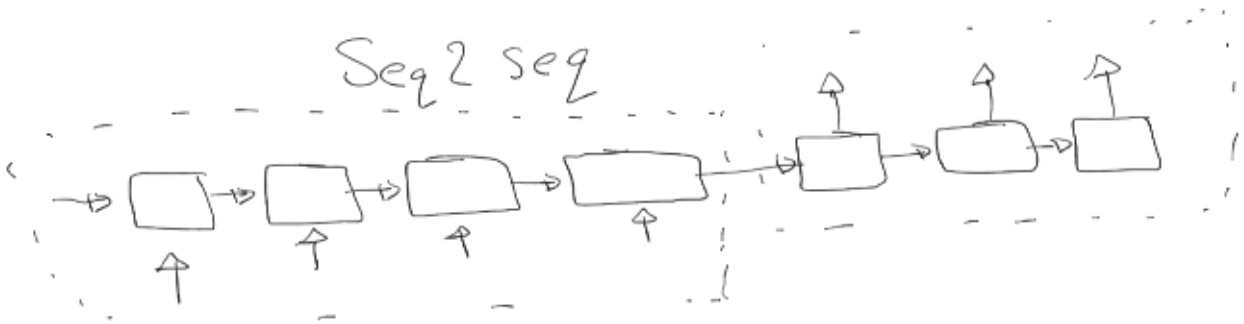
$T_x = T_y = 1$

"Music generation"

"Simpl NN

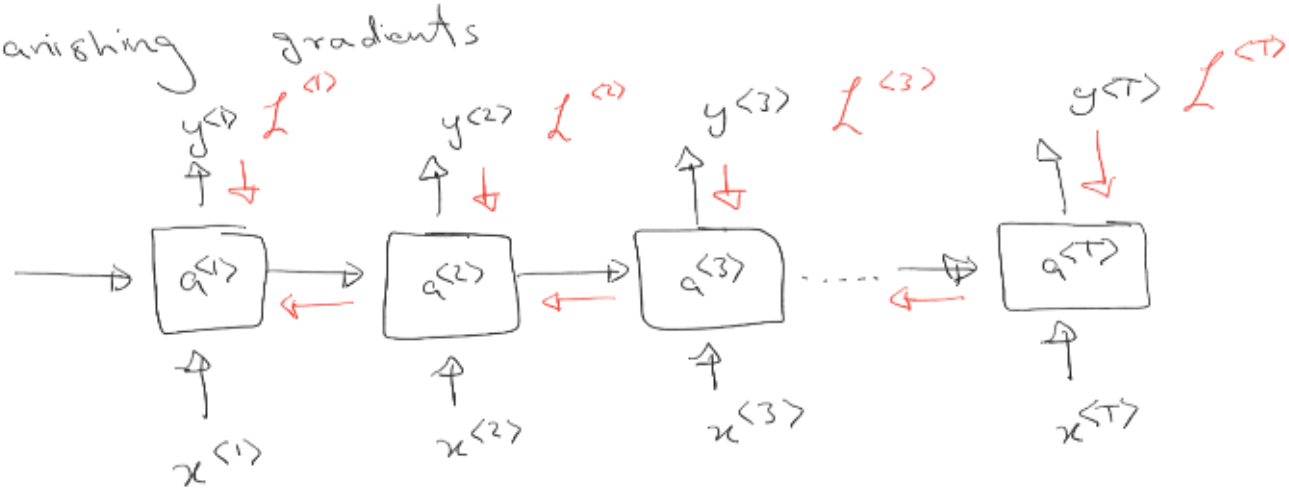
$T_y = T_1$, $T_x = T_2$

Seq2 seq



"Machine translation"

Vanishing gradients

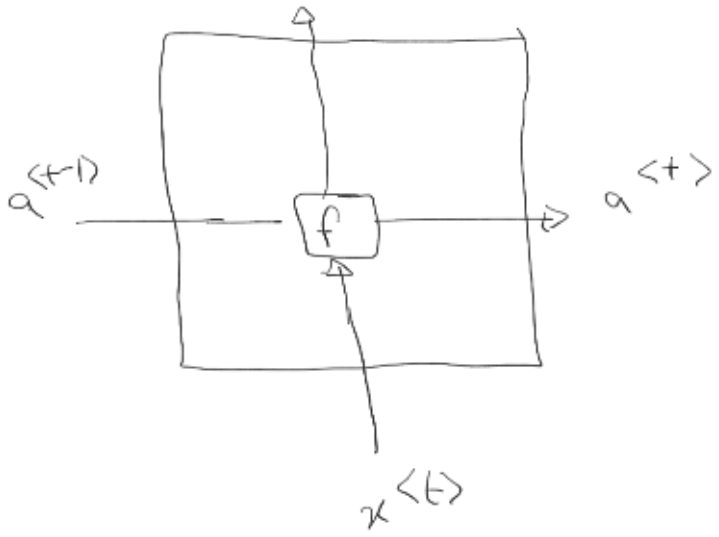


gradient computed at later time steps will have a problem to effect computations in earlier time steps.

GRU :

for RNN : $a^{<t>} = f\left(W_a \begin{bmatrix} a^{<t-1>} \\ x^{<t>} \end{bmatrix} + b_a\right)$

$\hat{y}^{<t>} = g(W_{ya}\, a^{<t>} + b_y)$

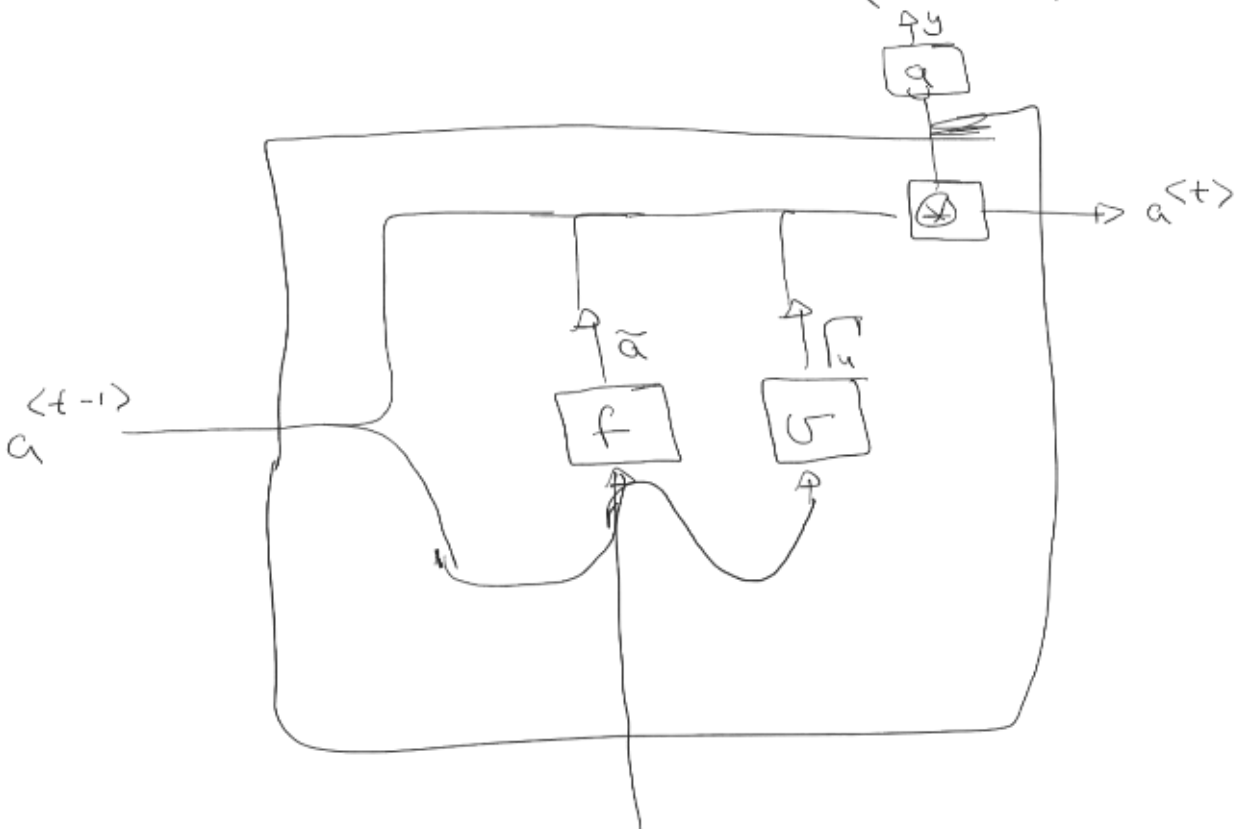


for GRU we have

$$\tilde{a}^{\langle t \rangle} = f\left(W_a \begin{bmatrix} a^{\langle t-1 \rangle} \\ x^{\langle t \rangle} \end{bmatrix} + b_a \right)$$

$$\Gamma_u = \overset{\text{sigmoid}}{\sigma}\left(W_u \begin{bmatrix} a^{\langle t-1 \rangle} \\ x^{\langle t \rangle} \end{bmatrix} + b_u \right)$$

$$a^{\langle t \rangle} = \Gamma_u \, \tilde{a}^{\langle t \rangle} + \left(1 - \Gamma_u\right) a^{\langle t-1 \rangle} \quad (*)$$



$x^{\langle t \rangle}$

Standard GRU:

relevance gate:

$$\Gamma_r = \sigma\left(W_r \begin{bmatrix} a^{\langle t-1 \rangle} \\ x^{\langle t \rangle} \end{bmatrix} + b_r\right)$$

$$\tilde{a}^{\langle t \rangle} = f\left(W_a \begin{bmatrix} \Gamma_r * a^{\langle t-1 \rangle} \\ x^{\langle t \rangle} \end{bmatrix} + b_a\right)$$

$$\Gamma_u = \sigma\left(W_u \begin{bmatrix} a^{\langle t-1 \rangle} \\ x^{\langle t \rangle} \end{bmatrix} + b_u\right)$$

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

LSTM:

$$\tilde{c}^{\langle t \rangle} = f\left(W_c \begin{bmatrix} a^{\langle t-1 \rangle} \\ x^{\langle t \rangle} \end{bmatrix} + b_c\right)$$

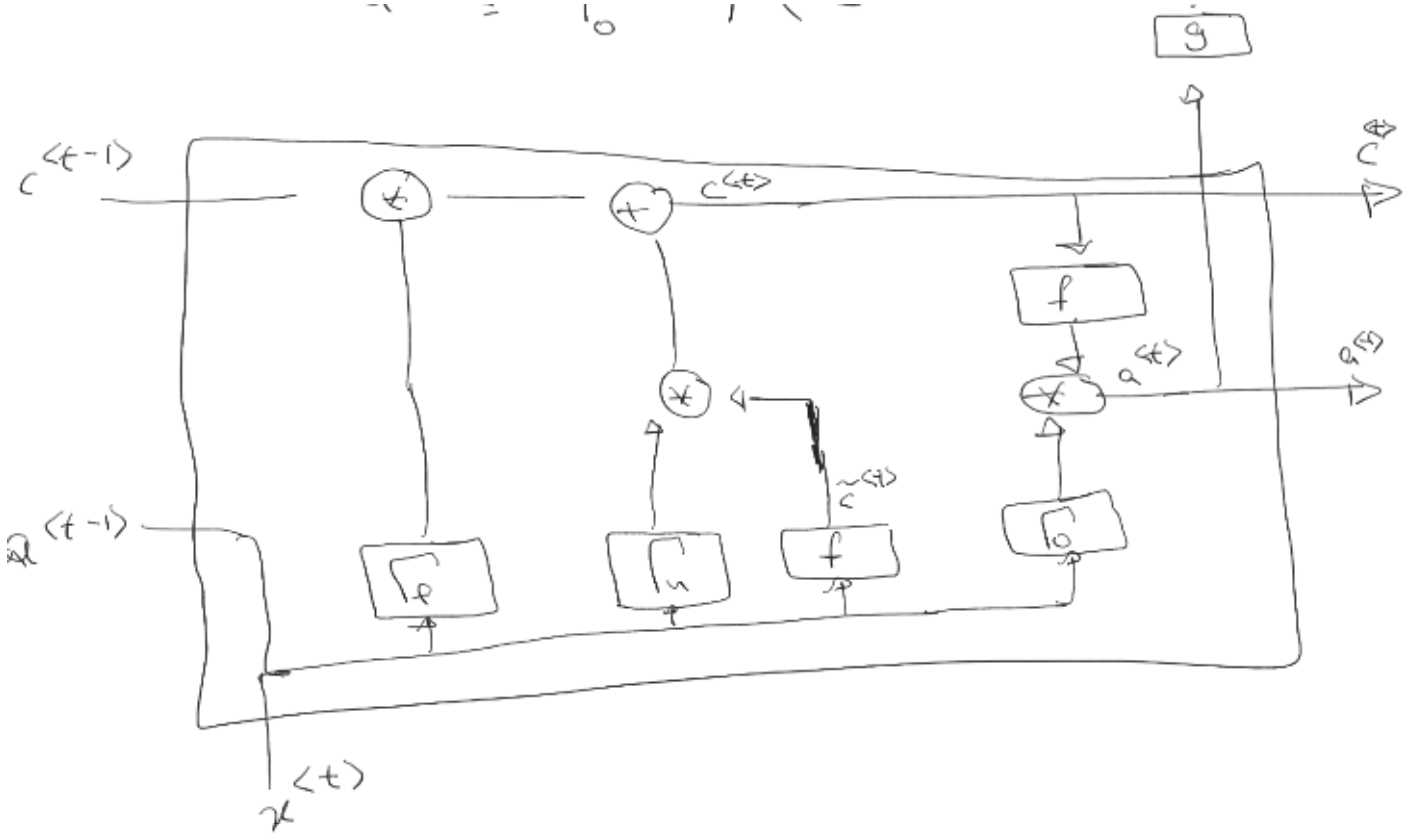
$$\Gamma_u = \sigma\left(W_u \begin{bmatrix} a^{\langle t-1 \rangle} \\ x^{\langle t \rangle} \end{bmatrix} + b_u\right)$$

$$\Gamma_f = \sigma\left(W_f \begin{bmatrix} a^{\langle t-1 \rangle} \\ x^{\langle t \rangle} \end{bmatrix} + b_f\right)$$

$$\Gamma_o = \sigma\left(W_o \begin{bmatrix} a^{\langle t-1 \rangle} \\ x^{\langle t \rangle} \end{bmatrix} + b_o\right)$$

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

$$a^{\langle t \rangle} = \Gamma_o * f(c^{\langle t \rangle})$$


g
c<t-1>
c<t>
f
a<t>
~c<t>
f
a<t-1>
x<t>