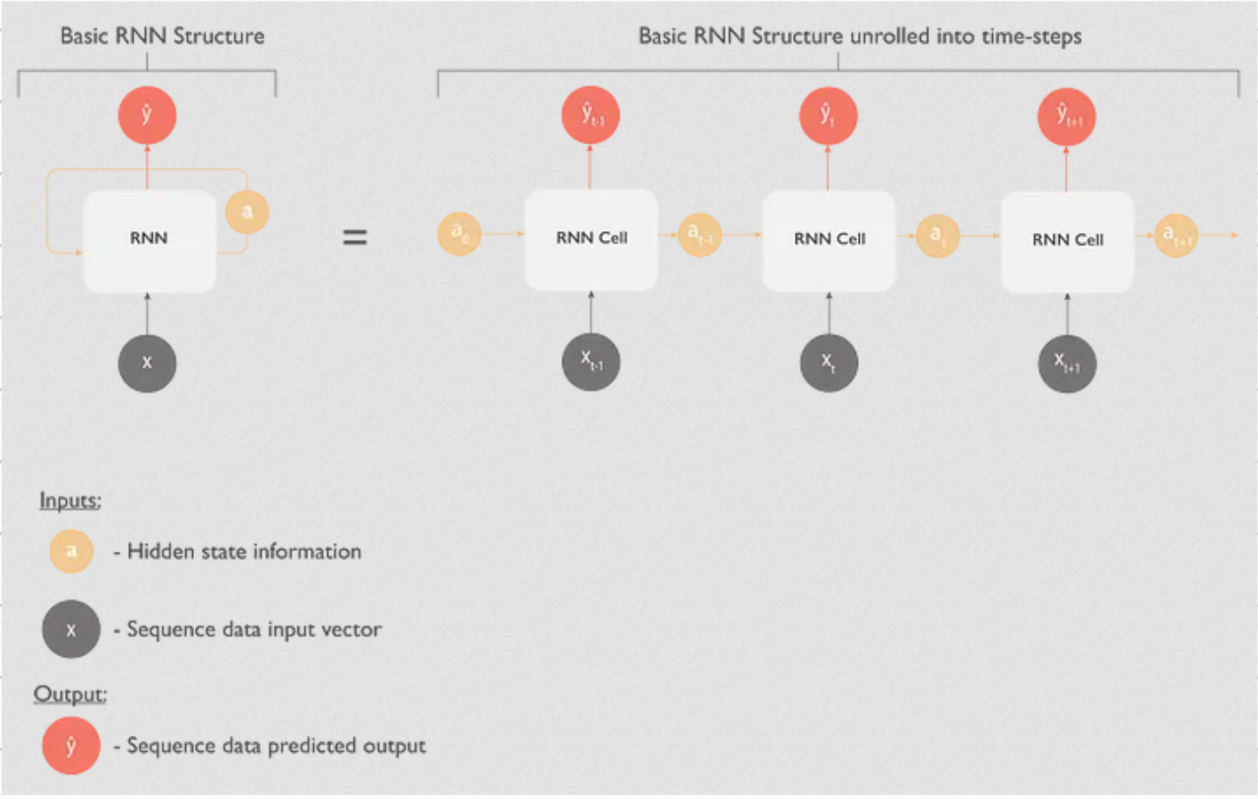
Basic Architecture



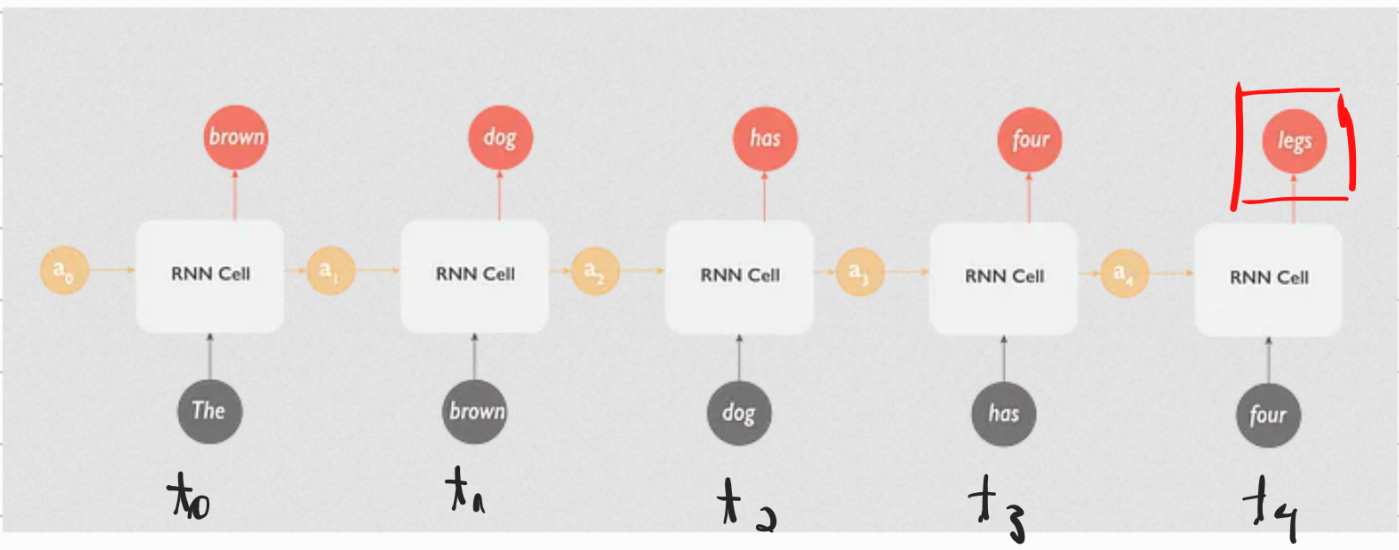
We unroll RNN basic structure (remove cycles) into repetitive chain of cells to correspond to the lenght of data

The input and output of RNN structure can vary

RNN types

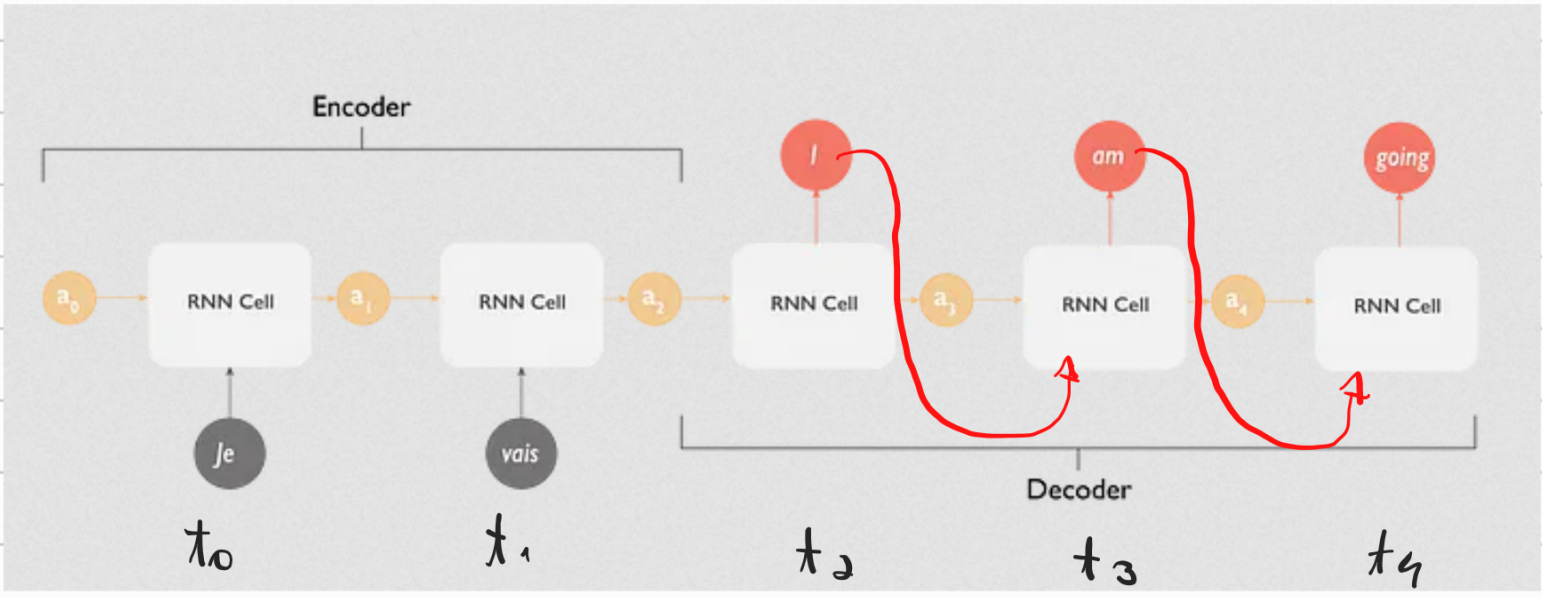
Many-to-Many (Same length)

Input number = Output number every time-step



Many-to-Many (Different length)

Input number ≠ Output number

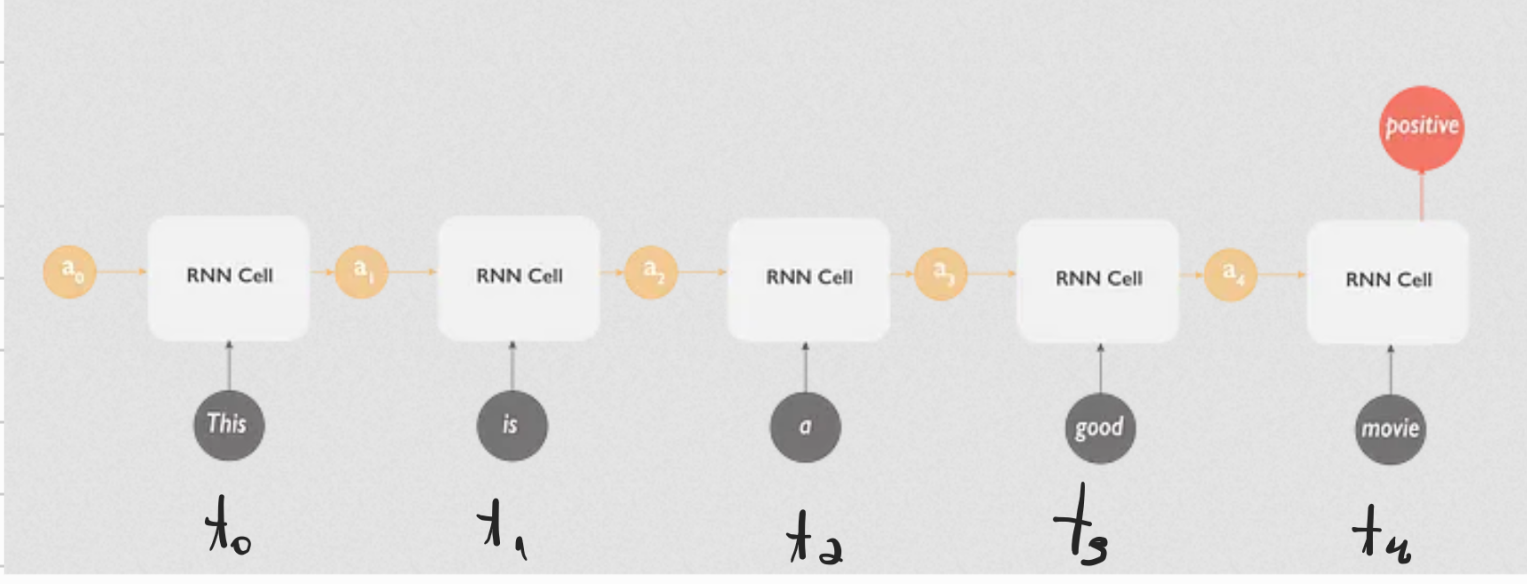


Encoder → Takes an input seq and maps to some internal state

Decoder → Takes internal state and generate some output

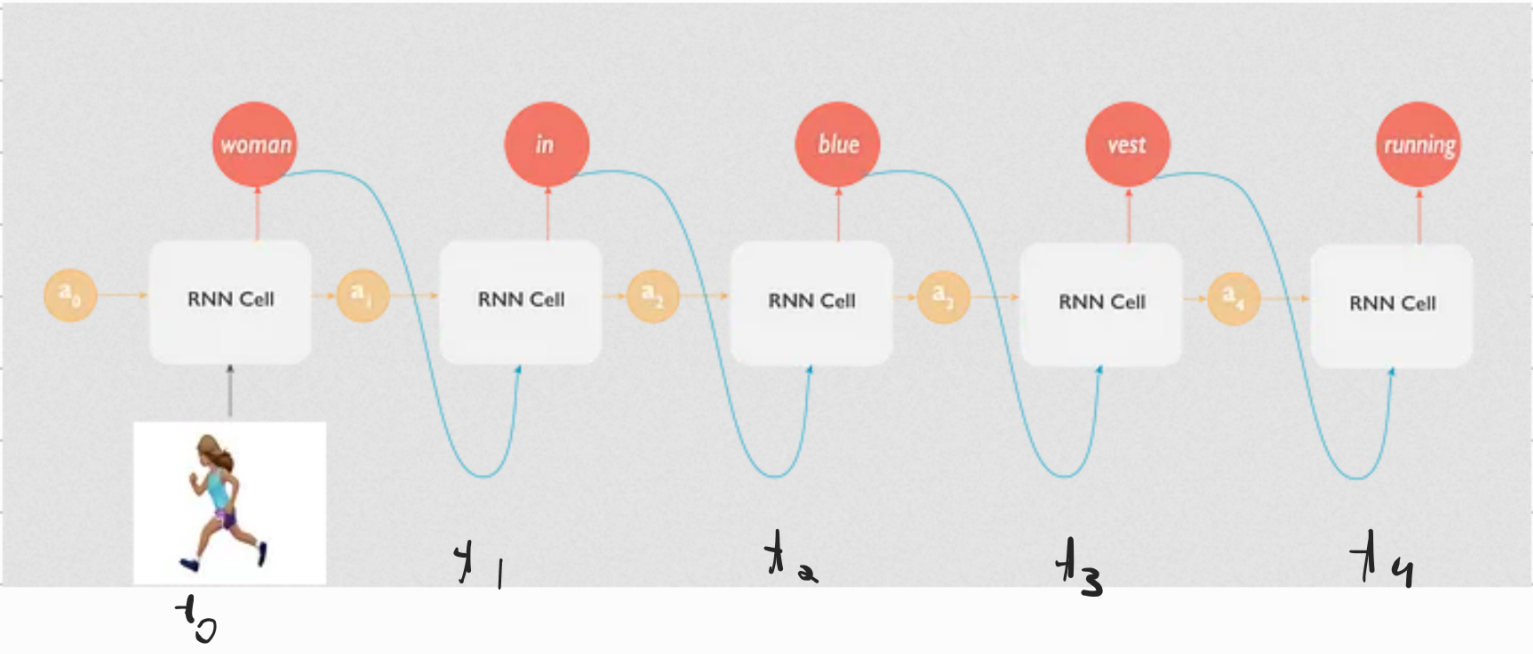
Many-to-One

Has many inputs at each time step but outputs just a single value in last time step

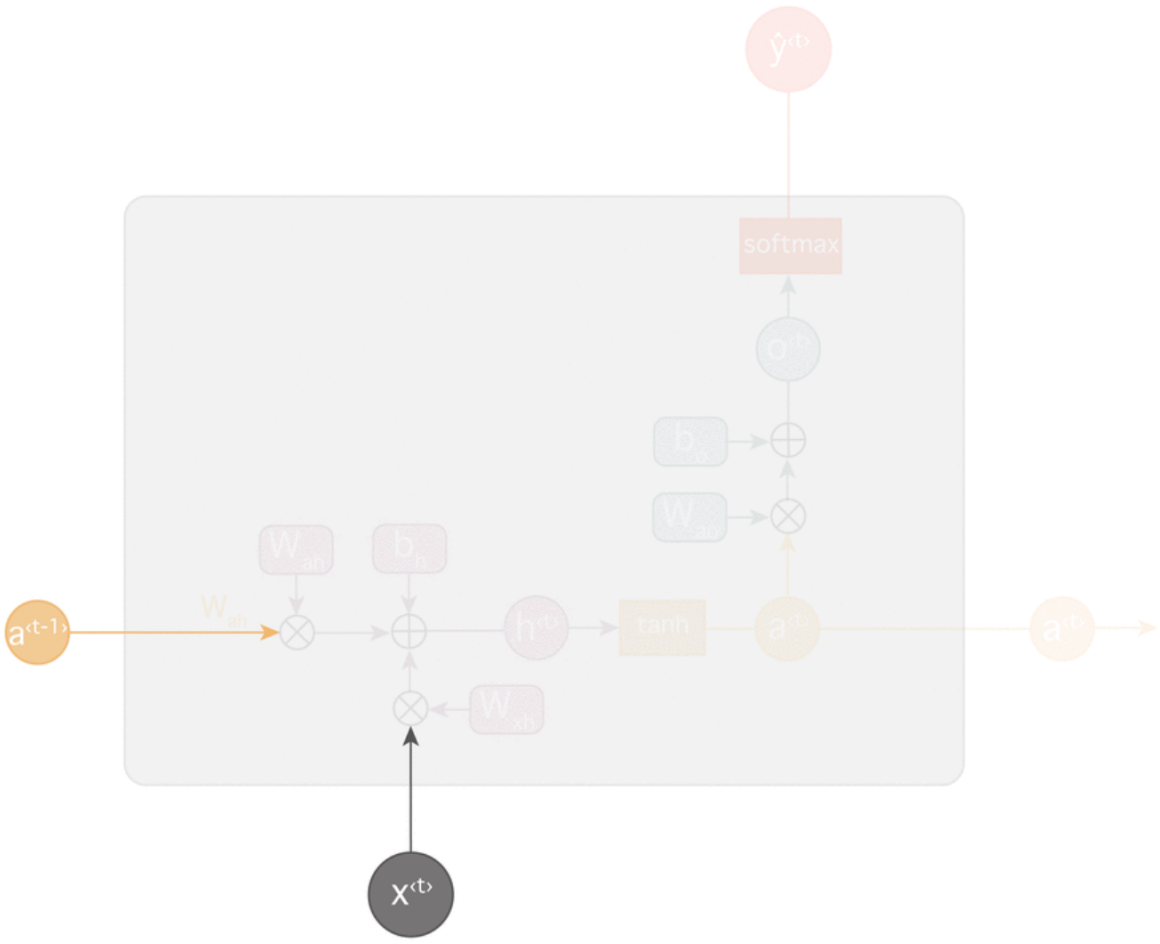


One-to-Many

Takes a single input at the first time step and outputs a sequence of values in remaining time steps



Forward RNN



Step 1) The cell receives 2 inputs $x^{<t>}$ and $a^{<t-1>}$

$x^{<t>}$: Input cell t

$a^{<t-1>}$: Activation cell t-1

Forward Propagation Equations

1. $h^{<t>} = (W_{xh} * x^{<t>}) + (W_{ah} * a^{<t-1>}) + b_h$
2. $a^{<t>} = \tanh(h^{<t>})$
3. $o^{<t>} = (W_{ao} * a^{<t>}) + b_o$
5. $\hat{y}^{<t>} = \text{softmax}(o^{<t>})$

Inputs:

- $a^{<t-1>}$ - hidden state activation at previous time-step <t-1>
- $x^{<t>}$ - input at time-step <t>

Vector Operators:

- $\otimes$ - matrix multiplication
- $\oplus$ - matrix addition

Activation Functions:

- $\tanh$ - hyperbolic tangent function
- softmax - softmax function

Weight Matrices & Bias Vectors:

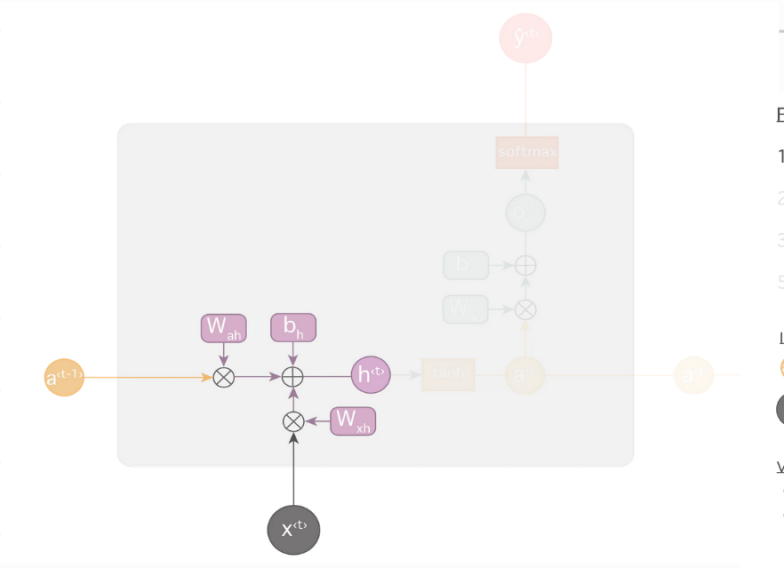
- W_{ah} - hidden-to-hidden weight matrix
- W_{xh} - input-to-hidden weight matrix
- b_h - hidden state bias vector
- W_{ao} - hidden-to-output weight matrix
- b_o - output bias vector

Outputs:

- $h^{<t>}$ - hidden state at time-step <t>
- $a^{<t>}$ - hidden state activation at time-step <t>
- $o^{<t>}$ - Predicted output at time-step <t>
- $\hat{y}^{<t>}$ - Predicted output activation at time-step <t>

Step 2) Do matrix multiplication between W_{xh} & $x^{<t>}$ and between W_{ah} & $a^{<t-1>}$. Next, it computes $h^{<t>}$ by adding the above together with bias term

(i) $h^{<t>} = W_{xh} * X^{<t>} + W_{ah} * a^{<t-1>} + b_h$



Forward Propagation Equations

1. $h^{<t>} = (W_{xh} * x^{<t>}) + (W_{ah} * a^{<t-1>}) + b_h$
2. $a^{<t>} = \tanh(h^{<t>})$
3. $o^{<t>} = (W_{ao} * a^{<t>}) + b_o$
5. $\hat{y}^{<t>} = \text{softmax}(o^{<t>})$

Inputs:

- $a^{<t-1>}$ - hidden state activation at previous time-step <t-1>
- $x^{<t>}$ - input at time-step <t>

Vector Operators:

- $\otimes$ - matrix multiplication
- $\oplus$ - matrix addition

Activation Functions:

- $\tanh$ - hyperbolic tangent function
- softmax - softmax function

Weight Matrices & Bias Vectors:

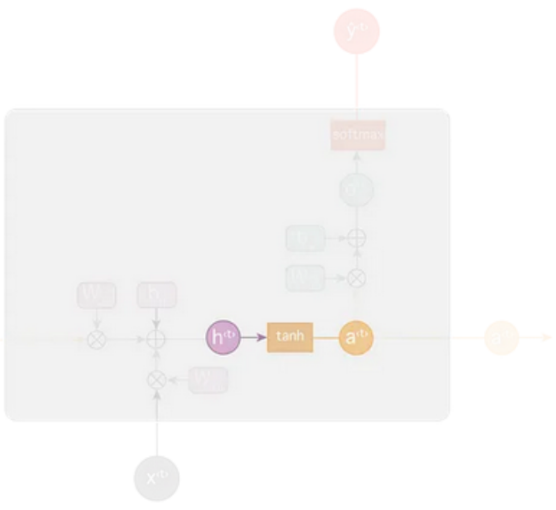
- W_{ah} - hidden-to-hidden weight matrix
- W_{xh} - input-to-hidden weight matrix
- b_h - hidden state bias vector
- W_{ao} - hidden-to-output weight matrix
- b_o - output bias vector

Outputs:

- $h^{<t>}$ - hidden state at time-step <t>
- $a^{<t>}$ - hidden state activation at time-step <t>
- $o^{<t>}$ - Predicted output at time-step <t>
- $\hat{y}^{<t>}$ - Predicted output activation at time-step <t>

Step 3) Calculate $a^{<t>}$ by passing $h^{<t>}$ through an activation function f (tanh or ReLU). I will use tanh as example

(ii) $a^{<t>} = f(h^{<t>})$



Forward Propagation Equations

1. $h^{<t>} = (W_{ah} * x^{<t>}) + (W_{ah} * a^{<t-1>}) + b_h$
2. $a^{<t>} = \tanh(h^{<t>})$
3. $o^{<t>} = (W_{ao} * a^{<t>}) + b_o$
5. $\hat{y}^{<t>} = \text{softmax}(o^{<t>})$

Inputs:
- $a^{<t-1>}$ - hidden state activation at previous time-step ‹t-1›
- $x^{<t>}$ - input at time-step ‹t›

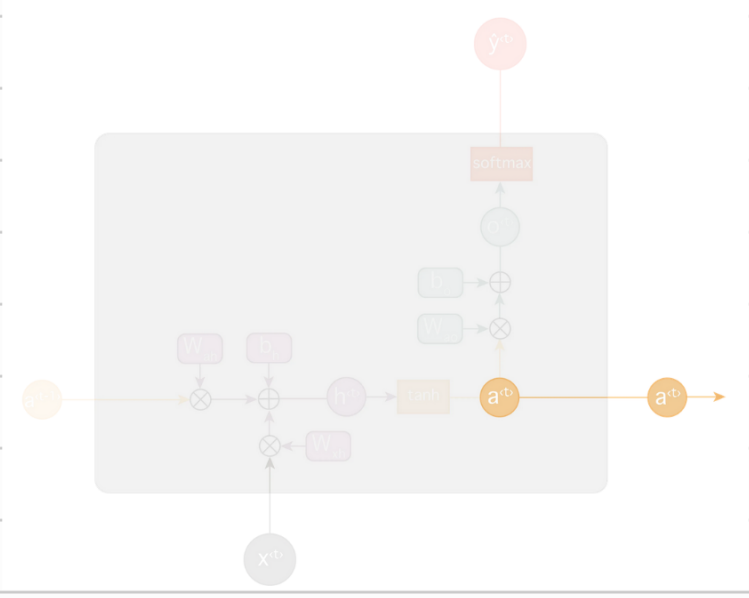
Vector Operators:
- ⊗ - matrix multiplication
- ⊕ - matrix addition

Activation Functions:
- tanh - hyperbolic tangent function
- softmax - softmax function

Weight Matrices & Bias Vectors:
- W_{ah} - hidden-to-hidden weight matrix
- W_{ax} - input-to-hidden weight matrix
- b_h - hidden state bias vector
- W_{ao} - hidden-to-output weight matrix
- b_o - output bias vector

Outputs:
- $h^{<t>}$ - hidden state at time-step ‹t›
- $a^{<t>}$ - hidden state activation at time-step ‹t›
- $o^{<t>}$ - Predicted output at time-step ‹t›
- $\hat{y}^{<t>}$ - Predicted output activation at time-step ‹t›

Step 4) The value $a^{<t>}$ is passed to the next cell



RNN Cell Equations

1. $h^{<t>} = (W_{ah} * x^{<t>}) + (W_{ah} * a^{<t-1>}) + b_h$
2. $a^{<t>} = \tanh(h^{<t>})$
3. $o^{<t>} = (W_{ao} * a^{<t>}) + b_o$
5. $\hat{y}^{<t>} = \text{softmax}(o^{<t>})$

Inputs:
- $a^{<t-1>}$ - hidden state activation at previous time-step ‹t-1›
- $x^{<t>}$ - input at time-step ‹t›

Vector Operators:
- ⊗ - matrix multiplication
- ⊕ - matrix addition

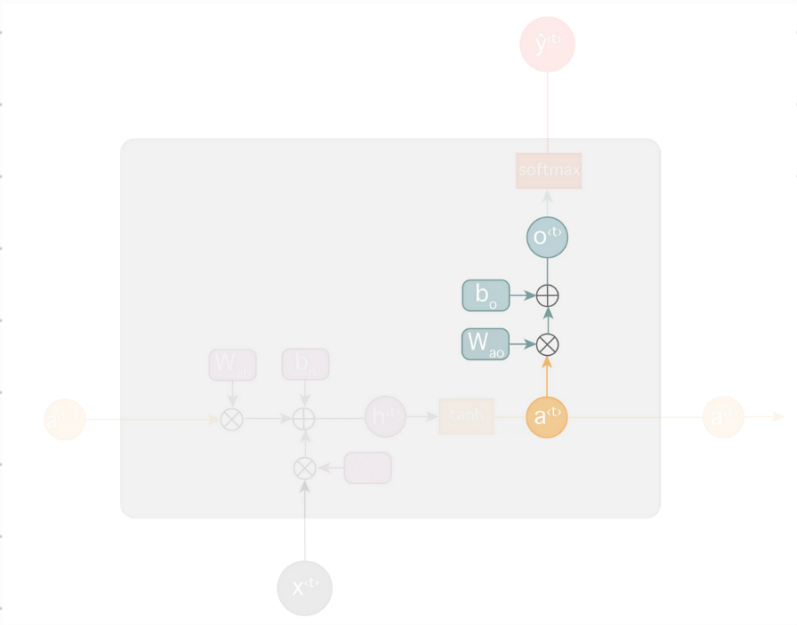
Activation Functions:
- tanh - hyperbolic tangent function
- softmax - softmax function

Weight Matrices & Bias Vectors:
- W_{ah} - hidden-to-hidden weight matrix
- W_{ax} - input-to-hidden weight matrix
- b_h - hidden state bias vector
- W_{ao} - hidden-to-output weight matrix
- b_o - output bias vector

Outputs:
- $h^{<t>}$ - hidden state at time-step ‹t›
- $a^{<t>}$ - hidden state activation at time-step ‹t›
- $o^{<t>}$ - Predicted output at time-step ‹t›
- $\hat{y}^{<t>}$ - Predicted output activation at time-step ‹t›

Step 5) Calculate the unnormalized log probability of each possible value of true output $o^{<t>}$

$$(iii)\quad o^{<t>} = W_{ao} * a^{<t>} + b_o$$



Forward Propagation Equations

1. $h^{<t>} = (W_{ah} * x^{<t>}) + (W_{ah} * a^{<t-1>}) + b_h$
2. $a^{<t>} = \tanh(h^{<t>})$
3. $o^{<t>} = (W_{ao} * a^{<t>}) + b_o$
5. $\hat{y}^{<t>} = \text{softmax}(o^{<t>})$

Inputs:
- $a^{<t-1>}$ - hidden state activation at previous time-step <t-1>
- $x^{<t>}$ - input at time-step <t>

Vector Operators:
- ⊗ - matrix multiplication
- ⊕ - matrix addition

Activation Functions:
- tanh - hyperbolic tangent function
- softmax - softmax function

Weight Matrices & Bias Vectors:
- W_{hh} - hidden-to-hidden weight matrix
- W_{ah} - input-to-hidden weight matrix
- b_h - hidden state bias vector
- W_{ao} - hidden-to-output weight matrix
- b_o - output bias vector

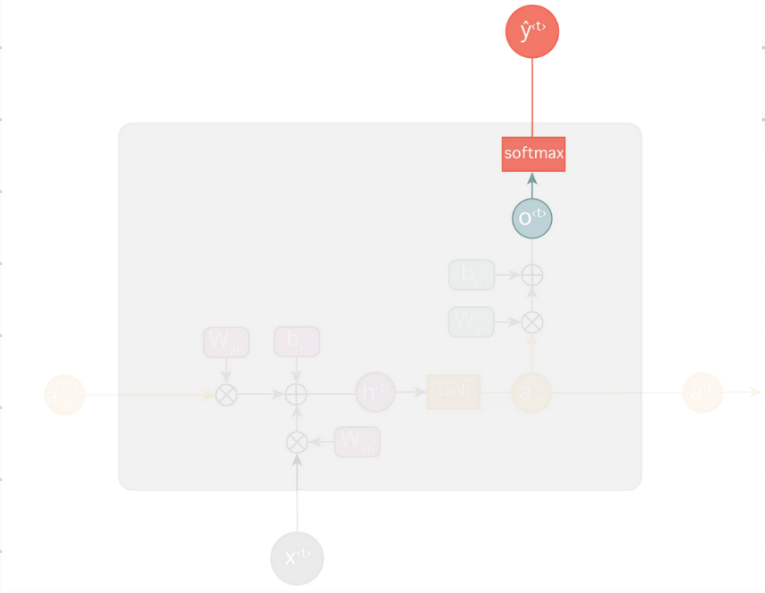
Outputs:
- $h^{<t>}$ - hidden state at time-step <t>
- $a^{<t>}$ - hidden state activation at time-step <t>
- $o^{<t>}$ - Predicted output at time-step <t>
- $\hat{y}^{<t>}$ - Predicted output activation at time-step <t>

Step 6) Calculate the vector of normalized probabilities $\hat{y}^{<t>}$ by passing $o^{<t>}$ through an activation function g (sigmoid or softmax)

The choice of g depends on the expected output, for example, for binary class output use sigmoid (remember logistic regression), but for multi-class output use softmax

$$(iv)\quad \hat{y}^{<t>} = g(o^{<t>})$$

Note: In examples I'll use softmax



Forward Propagation Equations

1. $h^{\langle t \rangle} = (W_{xh} * x^{\langle t \rangle}) + (W_{ah} * a^{\langle t-1 \rangle}) + b_h$
2. $a^{\langle t \rangle} = \tanh(h^{\langle t \rangle})$
3. $o^{\langle t \rangle} = (W_{ao} * a^{\langle t \rangle}) + b_o$
5. $\hat{y}^{\langle t \rangle} = \text{softmax}(o^{\langle t \rangle})$

Inputs:
- $a^{\langle t-1 \rangle}$ - hidden state activation at previous time-step ‹t-1›
- $x^{\langle t \rangle}$ - input at time-step ‹t›

Vector Operators:
- ⊗ - matrix multiplication
- ⊕ - matrix addition

Activation Functions:
- tanh - hyperbolic tangent function
- softmax - softmax function

Weight Matrices & Bias Vectors:
- W_{ah} - hidden-to-hidden weight matrix
- W_{xh} - input-to-hidden weight matrix
- b_h - hidden state bias vector
- W_{ao} - hidden-to-output weight matrix
- b_o - output bias vector

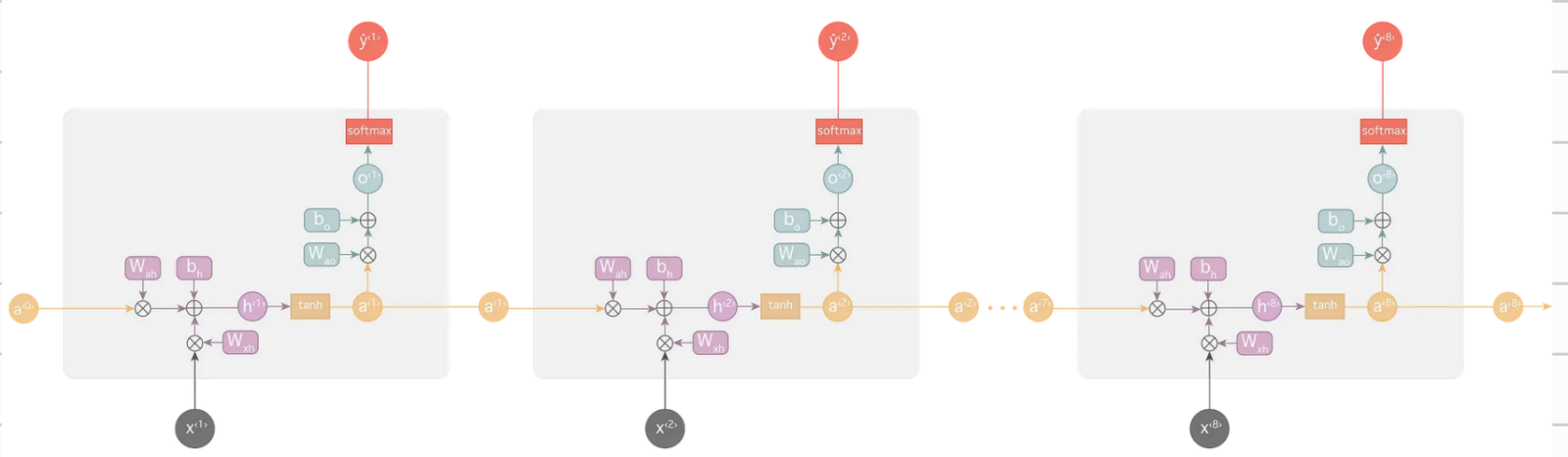
Outputs:
- $h^{\langle t \rangle}$ - hidden state at time-step ‹t›
- $a^{\langle t \rangle}$ - hidden state activation at time-step ‹t›
- $o^{\langle t \rangle}$ - Predicted output at time-step ‹t›
- $\hat{y}^{\langle t \rangle}$ - Predicted output activation at time-step ‹t›

Forward propagation through time

1) Initialize a hidden state $a^{\langle 0 \rangle}$ and W_{xh}, W_{ah}, W_{ao}, b_h, b_o

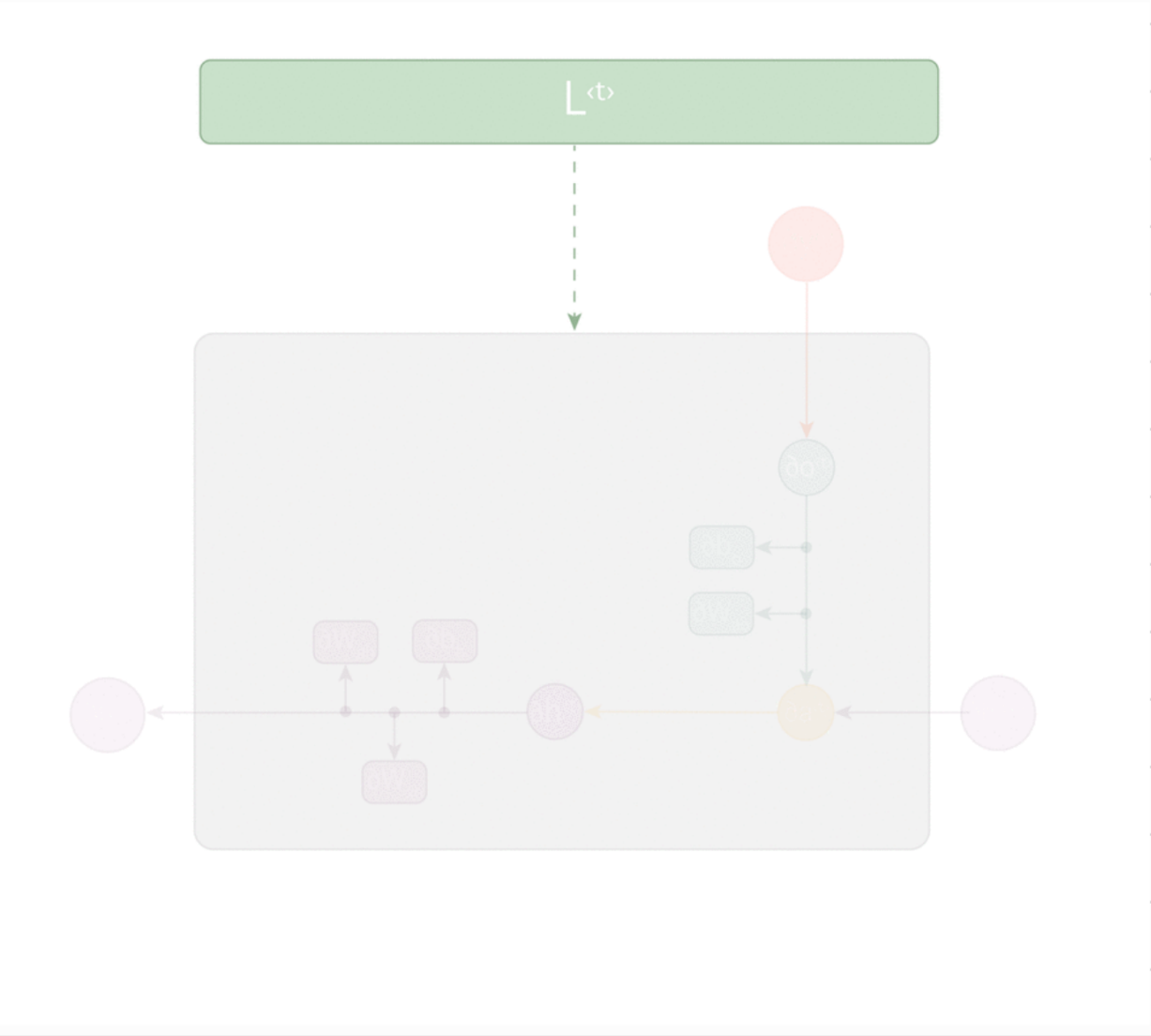
There will be shared across all time steps $t = 1$ to T

2) Repeat individual steps above for each cell in time



Backward RNN

Goal → Compute $\frac{\partial L^{<t>}}{\partial W_{ih}}, \frac{\partial L^{<t>}}{\partial W_{ah}}, \frac{\partial L^{<t>}}{\partial W_{ao}}, \frac{\partial L^{<t>}}{\partial b_h}, \frac{\partial L^{<t>}}{\partial b_o}$



Step 1) Choose some loss function based on your task. As example I'll use Cross-Entropy

$$L^{<t>}(y^{<t>}, \hat{y}^{<t>}) = -\sum_{i}^{C} y_i^{<t>} \log \hat{y}_i^{<t>} \quad (1)$$

Note: For any loss $\mathcal{L}$, $\mathcal{L} \in \mathbb{R}$

Removing i from summatory

$$L(y^{<t>}, \hat{y}^{<t>}) = -y_i^{<t>} \log \hat{y}_i^{<t>} + \sum_{i \neq k}^{C} -y_k^{<t>} \log \hat{y}_k^{<t>} \quad (2)$$

$L^{<t>}(y^{<t>}, \hat{y}^{<t>})$ = Cross-entropy loss function for multi-class outputs at time t

C = number of classes
$y_i^{<t>}$ = True output for class i at time t
$\hat{y}_i^{<t>}$ = Predicted output for class i at time t
$y_k^{<t>}$ = True output for other class k at time t
$\hat{y}_k^{<t>}$ = Predicted output for other class k at time t

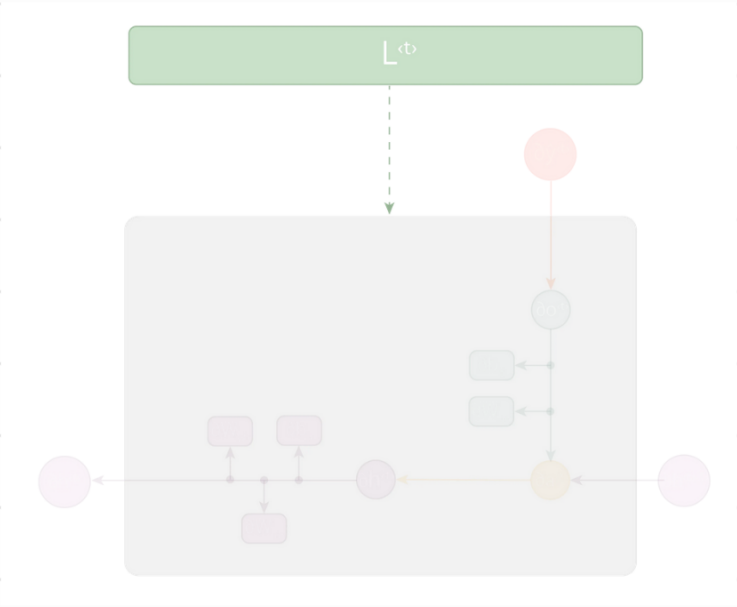
$y_i^{<t>}, y_k^{<t>} \in y^{<t>}$ and $\hat{y}_i^{<t>}, \hat{y}_k^{<t>} \in \hat{y}^{<t>}$

Note

(a) $y^{<t>}$ is one-hot encoded. Therefore, $y_i^{<t>}$ is 0 or 1

eg.: $y^{<t>} = \begin{pmatrix} y_0^{<t>} \\ y_1^{<t>} \\ y_2 \\ \vdots \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \end{pmatrix} \begin{matrix} \rightarrow \text{Class 0} \\ \rightarrow \text{Class 1} \\ \rightarrow \text{Class 2} \end{matrix}$ $\therefore y^{<t>}$ is class t

(b) $\hat{y}^{<t>}$ is normalized. Therefore, $\hat{y}_i^{<t>} \in [0, 1]$ and $y_i^{<t>} \in \mathbb{R}$



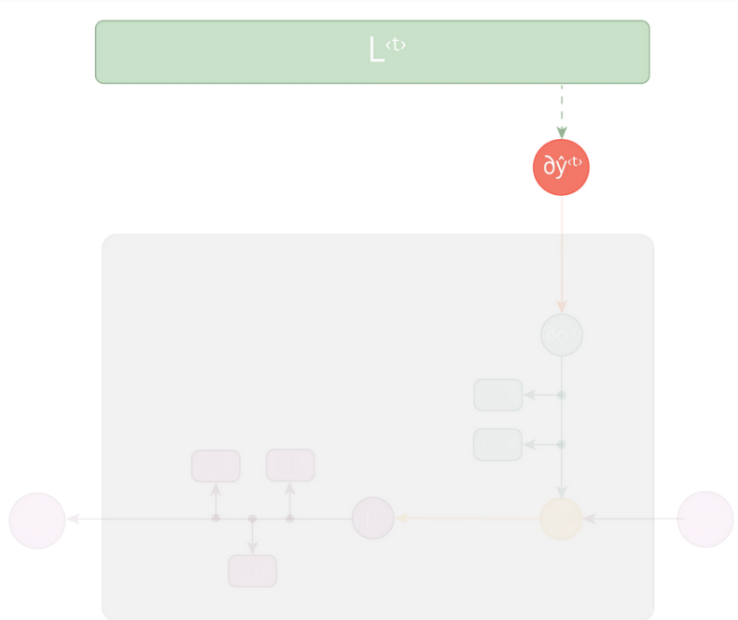
Step 2) Calculate $\frac{\partial L^{<t>}}{\partial \hat{y}^{<t>}} = \left(\frac{\partial L^{<t>}}{\partial \hat{y}_1^{<t>}}, \frac{\partial L^{<t>}}{\partial \hat{y}_2^{<t>}}, \dots, \frac{\partial L^{<t>}}{\partial \hat{y}_C^{<t>}} \right)^T$

$$\frac{\partial L^{<t>}}{\partial y_i^{<t>}} = \frac{\partial}{\partial y_i^{<t>}} \left(-y_i^{<t>} \log \hat{y}_i^{<t>} + \sum_{i \neq k} -y_k^{<t>} \log \hat{y}_k^{<t>} \right)$$ (the $\log \hat{y}_k^{<t>}$ term → 0)

$$\frac{\partial L}{\partial y_i^{<t>}} = \frac{-y_i^{<t>}}{\hat{y}_i^{<t>}} \quad (3)$$

$$\frac{\partial L^{<t>}}{\partial y_k^{<t>}} = \frac{\partial}{\partial y_k^{<t>}} \left(-y_i^{<t>} \log \hat{y}_i^{<t>} + \sum_{i \neq k} -y_k^{<t>} \log \hat{y}_k^{<t>} \right)$$ (the $\log \hat{y}_i^{<t>}$ term → 0)

$$\frac{\partial L^{<t>}}{\partial y_k^{<t>}} = \sum_{i \neq k} \frac{-y_k^{<t>}}{\hat{y}_k^{<t>}} \quad (4)$$

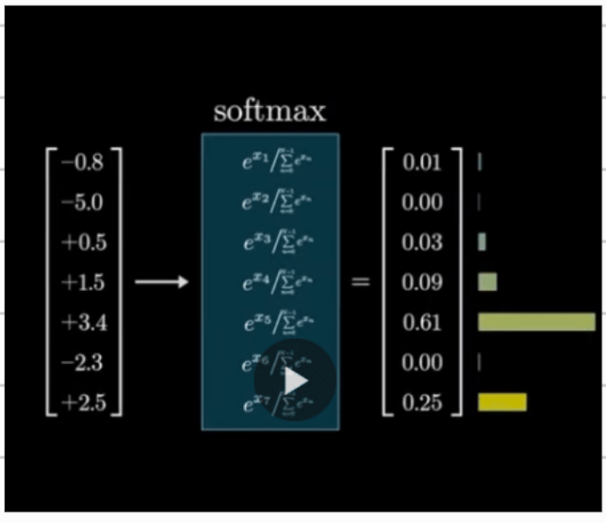


Step 3) Calculate $\frac{\partial L^{<t>}}{\partial o^{<t>}} = \left(\frac{\partial L^{<t>}}{\partial o_1^{<t>}}, \frac{\partial L^{<t>}}{\partial o_2^{<t>}}, \dots, \frac{\partial L^{<t>}}{\partial o_C^{<t>}}\right)^T$

Remember the property of softmax that if you change any $o_u^{<t>}$ ALL $\hat{y}_u^{<t>}$ where $u \in [1, C]$ (u includes i and k) will also change

Softmax

$$\hat{y}_u^{<t>} = \frac{e^{o_u^{<t>}}}{\sum_j^C e^{o_j^{<t>}}}$$



Therefore, the chain rule is $L^{<t>} \to \hat{y}_i^{<t>} \to o_i^{<t>}$ and $L^{<t>} \to \hat{y}_k^{<t>} \to o_i^{<t>}$

$$\frac{\partial L^{<t>}}{\partial o_i^{<t>}} = \frac{\partial L^{<t>}}{\partial \hat{y}_i^{<t>}} \cdot \frac{\partial \hat{y}_i^{<t>}}{\partial o_i^{<t>}} + \frac{\partial L^{<t>}}{\partial \hat{y}_k^{<t>}} \cdot \frac{\partial \hat{y}_k^{<t>}}{\partial o_i^{<t>}}$$

Softmax derivative

$$\hat{y}_u^{<t>} = \frac{e^{o_u^{<t>}}}{\sum_j^C e^{a_j}}, \quad \text{let } S = \sum_j^C e^{a_j}$$

$$\hat{y}_u^{<t>} = \frac{e^{o_u^{<t>}}}{S}$$

Case 1

$$\frac{\partial \hat{y}_u^{<t>}}{\partial o_u^{<t>}} = \frac{e^{o_u^{<t>}} \cdot S - e^{o_u^{<t>}} \cdot e^{o_u^{<t>}}}{S^2}$$

$$= \frac{e^{o_u^{<t>}} \left(S - e^{o_u^{<t>}}\right)}{S^2}$$

$$= \frac{e^{o_u^{<t>}}}{S} \cdot \left(1 - \frac{e^{o_u^{<t>}}}{S}\right)$$

$$\therefore \boxed{\frac{\partial \hat{y}_u^{<t>}}{\partial o_u^{<t>}} = \hat{y}_u^{<t>}\left(1 - \hat{y}_u^{<t>}\right)} \quad \square$$

Case 2

$\frac{\partial \hat{y}_v^{<t>}}{\partial o_u^{<t>}}$ where $v \neq u$

$$\frac{\partial \hat{y}_v^{<t>}}{\partial o_u^{<t>}} = \frac{0 \cdot S - e^{o_v^{<t>}} \cdot e^{o_u^{<t>}}}{S^2}$$

$$= -\frac{e^{o_v^{<t>}}}{S} \cdot \frac{e^{o_u^{<t>}}}{S}$$

$$\therefore \boxed{\frac{\partial \hat{y}_v^{<t>}}{\partial o_u^{<t>}} = -\hat{y}_v^{<t>} \cdot \hat{y}_u^{<t>}} \quad \square$$

From equations $\frac{\partial L}{\partial y_i^{<t>}} = \frac{-y_i^{<t>}}{\hat{y}_i^{<t>}}$ (3) &

$\frac{\partial L^{<t>}}{\partial y_k^{<t>}} = \sum_{i \neq k} \frac{-y_k^{<t>}}{\hat{y}_k^{<t>}}$ (4) and from softmax derivatives $\frac{\partial \hat{y}_i^{<t>}}{\partial o_i^{<t>}} = \hat{y}_i^{<t>}\left(1 - \hat{y}_i^{<t>}\right)$ &

$$\frac{\partial \hat{y}_k^{<t>}}{\partial o_i^{<t>}} = -\hat{y}_k^{<t>} \cdot \hat{y}_i^{<t>}$$

$$\frac{\partial L^{<t>}}{\partial o_i^{<t>}} = \frac{\partial L^{<t>}}{\partial \hat{y}_i^{<t>}} \cdot \frac{\partial \hat{y}_i^{<t>}}{\partial o_i^{<t>}} + \frac{\partial L^{<t>}}{\partial \hat{y}_k^{<t>}} \cdot \frac{\partial \hat{y}_k^{<t>}}{\partial o_i^{<t>}}$$

$$= \left(\frac{-y_i^{<t>}}{\hat{y}_i^{<t>}} \cdot \hat{y}_i^{<t>} \left(1 - \hat{y}_i^{<t>}\right) \right) + \left(\sum_{i \neq k}^{C} \left(\frac{-y_k^{<t>}}{\hat{y}_k^{<t>}} \right) \cdot -\hat{y}_k^{<t>} \cdot \hat{y}_i^{<t>} \right)$$

$$= -y_i^{<t>}\left(1 - \hat{y}_i^{<t>}\right) + \left(\sum_{i \neq k}^{C} y_k^{<t>} \cdot \hat{y}_i^{<t>} \right)$$

$$= -y_i^{<t>} + y_i^{<t>} \hat{y}_i^{<t>} + \hat{y}_i^{<t>} \sum_{i \neq k}^{C} y_k^{<t>}$$

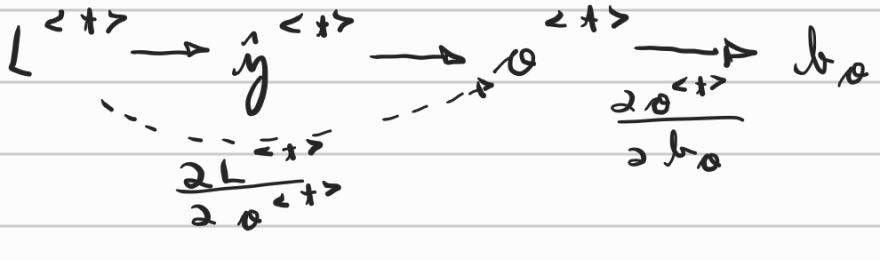
$$= -y_i^{<t>} + \hat{y}_i^{<t>} \left(y_i^{<t>} + \sum_{i \neq k}^{C} y_k^{<t>} \right)$$

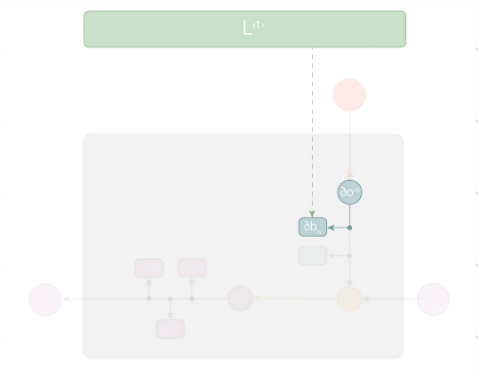
$= 1$ because $y^{<t>}$ is a one-hot encoded vector

$$\therefore \boxed{\frac{\partial L^{<t>}}{\partial o_i^{<t>}} = \hat{y}_i^{<t>} - y_i^{<t>}} \quad (5)$$

Step 4) Calculate $\frac{\partial L^{<t>}}{\partial b_o} = \left(\frac{\partial L^{<t>}}{\partial b_{o_1}}, \frac{\partial L^{<t>}}{\partial b_{o_2}}, \ldots, \frac{\partial L^{<t>}}{\partial b_{o_C}} \right)^T$

Chain-Rule





$$\frac{\partial L^{<t>}}{\partial b_o} = \underbrace{\frac{\partial L^{<t>}}{\partial \hat{y}^{<t>}} \cdot \frac{\partial \hat{y}^{<t>}}{\partial o^{<t>}}}_{\frac{\partial L}{\partial o^{<t>}}} \cdot \frac{\partial o^{<t>}}{\partial b_o} = \frac{\partial L^{<t>}}{\partial o^{<t>}} \cdot \frac{\partial o^{<t>}}{\partial b_o}$$

*From $o^{<t>} = (W_{ao} \cdot a^{<t>}) + b_o$ we want $\frac{\partial o^{<t>}}{\partial b_o}$

$$W_{ao} = \begin{pmatrix} w^{11}_{ao} & w^{12}_{ao} & \dots & w^{1m}_{ao} \\ w^{21}_{ao} & w^{22}_{ao} & \dots & w^{2m}_{ao} \\ \vdots & \vdots & \ddots & \vdots \\ w^{c1}_{ao} & w^{c2}_{ao} & \dots & w^{cm}_{ao} \end{pmatrix}_{c \times m}$$

$$a^{<t>} = \begin{pmatrix} a_1^{<t>} \\ a_2^{<t>} \\ a_3^{<t>} \\ \vdots \\ a_m^{<t>} \end{pmatrix}_{m \times 1}, \quad b_o = \begin{pmatrix} b_{o_1} \\ b_{o_2} \\ b_{o_3} \\ \vdots \\ b_{o_c} \end{pmatrix}_{c \times 1}, \quad o^{<t>} = \begin{pmatrix} o_1^{<t>} \\ o_2^{<t>} \\ o_3^{<t>} \\ \vdots \\ o_c^{<t>} \end{pmatrix}_{c \times 1}$$

$$\begin{cases} o_1^{<t>} = W^{11}_{ao} \cdot a_1^{<t>} + W^{12}_{ao} \cdot a_2^{<t>} + \dots + W^{1m}_{ao} a_m^{<t>} + b_{o_1} \\ o_2^{<t>} = W^{21}_{ao} \cdot a_1^{<t>} + W^{22}_{ao} \cdot a_2^{<t>} + \dots + W^{2m}_{ao} a_m^{<t>} + b_{o_2} \\ \vdots \\ o_c^{<t>} = W^{c1}_{ao} \cdot a_1^{<t>} + W^{c2}_{ao} \cdot a_2^{<t>} + \dots + W^{cm}_{ao} a_m^{<t>} + b_{o_c} \end{cases}$$

Let $u \in \mathbb{N}$ and $u \in [1, c]$

$$\frac{\partial o_u^{<t>}}{\partial b_u^{<t>}} = \frac{\partial}{\partial b_u^{<t>}} \left(\underbrace{W^{u1}_{ao} \cdot a_1^{<t>}}_{0} + \underbrace{W^{u2}_{ao} \cdot a_2^{<t>}}_{0} + \dots + \underbrace{W^{uu}_{ao} \cdot a_u}_{1} + \dots + \underbrace{W^{um}_{ao} a_m}_{0} + \underbrace{b_{o_u}}_{1} \right)$$

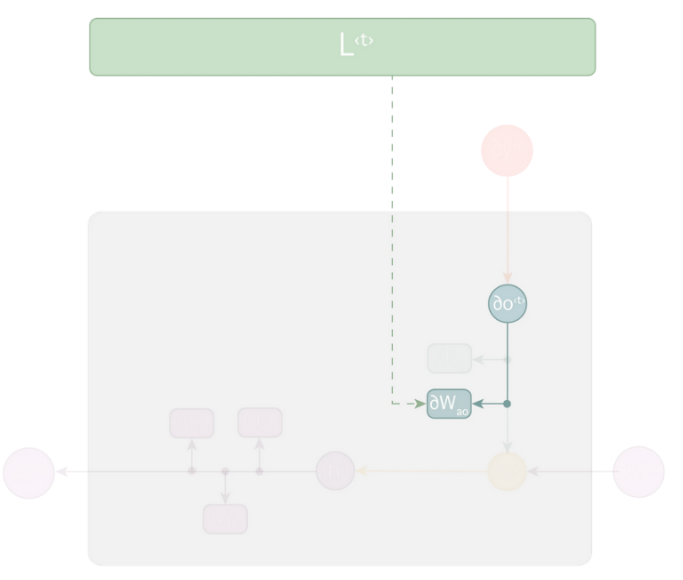
$$= 1$$

WLOG $\frac{\partial o^{<t>}}{\partial b_M^{<t>}} = 1$ (unitary vector) //

$$\rightarrow \frac{\partial L^{<t>}}{\partial b_o} = \frac{\partial L^{<t>}}{\partial o^{<t>}} \cdot \underbrace{\frac{\partial o^{<t>}}{\partial b_o}}_{1} \quad \therefore \quad \boxed{\frac{\partial L^{<t>}}{\partial b_o} = \frac{\partial L^{<t>}}{\partial o^{<t>}}} \quad (6)$$

Step 5) Calculate $\frac{\partial L^{<t>}}{\partial W_{ao}} =$

$$\begin{pmatrix} \frac{\partial L^{<t>}}{\partial w_{ao}^{11}} & \frac{\partial L^{<t>}}{\partial w_{ao}^{12}} & \cdots & \frac{\partial L^{<t>}}{\partial w_{ao}^{1m}} \\ \frac{\partial L^{<t>}}{\partial w_{ao}^{21}} & \frac{\partial L^{<t>}}{\partial w_{ao}^{22}} & \cdots & \frac{\partial L^{<t>}}{\partial w_{ao}^{2m}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial L^{<t>}}{\partial w_{ao}^{c1}} & \frac{\partial L^{<t>}}{\partial w_{ao}^{c2}} & \cdots & \frac{\partial L^{<t>}}{\partial w_{ao}^{cm}} \end{pmatrix}_{c \times m}$$



Chain Rule

$$L^{<t>} \rightarrow \hat{y}^{<t>} \rightarrow o^{<t>} \rightarrow W_{ao}$$

$\frac{\partial L^{<t>}}{\partial o^{<t>}}$ $\quad$ $\frac{\partial o^{<t>}}{\partial W_{ao}}$

$$\therefore \frac{\partial L^{<t>}}{\partial W_{ao}} = \frac{\partial L^{<t>}}{\partial o^{<t>}} \cdot \frac{\partial o^{<t>}}{\partial W_{ao}} \; //$$

*From $o^{<t>} = (W_{ao} \cdot a^{<t>}) + b_o$ we want $\frac{\partial o^{<t>}}{\partial W_{ao}}$

$$W_{ao} = \begin{pmatrix} w^{11}_{ao} & w^{12}_{ao} & \dots & w^{1m}_{ao} \\ w^{21}_{ao} & w^{22}_{ao} & \dots & w^{2m}_{ao} \\ \vdots & \vdots & \ddots & \vdots \\ w^{c1}_{ao} & w^{c2}_{ao} & \dots & w^{cm}_{ao} \end{pmatrix}_{c \times m}$$

$$a^{<t>} = \begin{pmatrix} a_1^{<t>} \\ a_2^{<t>} \\ a_3^{<t>} \\ \vdots \\ a_m^{<t>} \end{pmatrix}_{m \times 1}, \quad b_o = \begin{pmatrix} b_{o_1} \\ b_{o_2} \\ b_{o_3} \\ \vdots \\ b_{o_c} \end{pmatrix}_{c \times 1}, \quad o^{<t>} = \begin{pmatrix} o_1^{<t>} \\ o_2^{<t>} \\ o_3^{<t>} \\ \vdots \\ o_c^{<t>} \end{pmatrix}_{c \times 1}$$

$$\begin{cases} o_1^{<t>} = W^{11}_{ao} \cdot a_1^{<t>} + W^{12}_{ao} \cdot a_2^{<t>} + \dots + W^{1m}_{ao} a_m^{<t>} + b_{o_1} \\ o_2^{<t>} = W^{21}_{ao} \cdot a_1^{<t>} + W^{22}_{ao} \cdot a_2^{<t>} + \dots + W^{2m}_{ao} a_m^{<t>} + b_{o_2} \\ \vdots \\ o_c^{<t>} = W^{c1}_{ao} \cdot a_1^{<t>} + W^{c2}_{ao} \cdot a_2^{<t>} + \dots + W^{cm}_{ao} a_m^{<t>} + b_{o_c} \end{cases}$$

Let $u, v \in \mathbb{N}$ where $u \in [1, c]$ and $v \in [1, m]$

$$\frac{\partial o_u^{<t>}}{\partial W^{uv}_{ao}} = \frac{\partial}{\partial W^{uv}_{ao}} \left(W^{u1}_{ao} \cdot a_1^{<t>} + W^{u2}_{ao} \cdot a_2^{<t>} + \dots + W^{uv}_{ao} \cdot a_v^{<t>} + \dots + W^{um}_{ao} a_m + b_{o_u} \right)$$

(the $W^{u1}_{ao} \cdot a_1^{<t>}$, $W^{u2}_{ao} \cdot a_2^{<t>}$, $W^{um}_{ao} a_m$ and b_{o_u} terms are 0; $W^{uv}_{ao} \cdot a_v^{<t>}$ is $a_v^{<t>}$)

$$= a_v^{<t>}$$

WLOG $\frac{\partial o^{<t>}}{\partial W_{ao}} = a^{<t>}$ □

$$\frac{\partial L^{<t>}}{\partial W_{ao}} = \frac{\partial L^{<t>}}{\partial o^{<t>}} \cdot \frac{\partial o^{<t>}}{\partial W_{ao}} = \underbrace{\frac{\partial L^{<t>}}{\partial o^{<t>}}}_{C \times 1} \cdot \underbrace{\left(a^{<t>}\right)^T}_{m \times 1 \quad 1 \times m}$$

$$\therefore \boxed{\frac{\partial L^{<t>}}{\partial W_{ao}} = \frac{\partial L}{\partial o^{<t>}} \cdot a^{<t>T}} \quad (7)$$

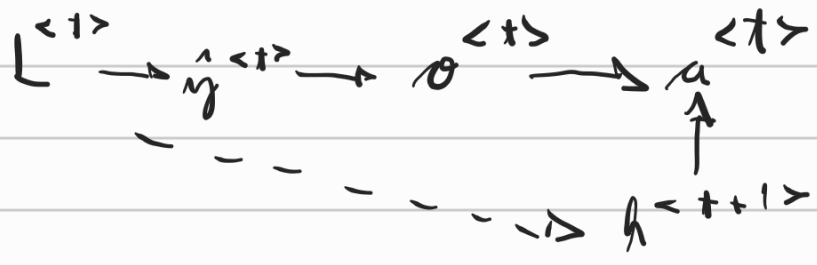
Note: $\frac{\partial L}{\partial o^{<t>}} \cdot a^{<t>T} = \begin{pmatrix} \frac{\partial L}{\partial o_1^{<t>}} \\ \frac{\partial L}{\partial o_2^{<t>}} \\ \vdots \\ \frac{\partial L}{\partial o_C^{<t>}} \end{pmatrix}_{C \times 1} \cdot \underbrace{\left(a_1^{<t>}, a_2^{<t>}, \dots, a_m^{<t>}\right)}_{1 \times m}$

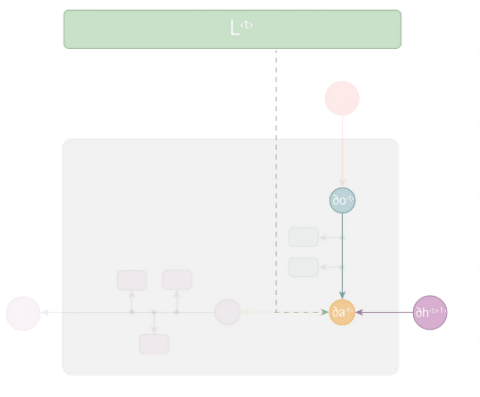
Eg.: $\frac{\partial L^{<t>}}{\partial W_{ao}^{11}} = \frac{\partial L}{\partial o_1^{<t>}} \cdot a_1^{<t>}$

Step 6) Calculate $\frac{\partial L^{<t>}}{\partial a^{<t>}} = \left(\frac{\partial L^{<t>}}{\partial a_1^{<t>}} \; \frac{\partial L^{<t>}}{\partial a_2^{<t>}} \; \dots \; \frac{\partial L^{<t>}}{\partial a_m^{<t>}}\right)^T$

Chain Rule

Remember that $a^{<t>}$ is used by $o^{<t>}$ and $h^{<t+1>}$ during forward propagation





Note

$h^{<t+1>}$ contributes to $a^{<t>}$ and $a^{<t>}$ contributes to $L^{<t>}$, so we can derive $L^{<t>}$ in respect to $h^{<t+1>}$ (this explain the dotted line)

$$\frac{\partial L^{<t>}}{\partial a^{<t>}} = \frac{\partial L^{<t>}}{\partial o^{<t>}} \cdot \frac{\partial o^{<t>}}{a^{<t>}} + \frac{\partial L^{<t>}}{\partial h^{<t+1>}} \cdot \frac{\partial h^{<t+1>}}{\partial a^{<t>}}$$

*From $o^{<t>} = (W_{ao} \cdot a^{<t>}) + b_o$ we want $\frac{\partial o^{<t>}}{\partial a^{<t>}}$

$$W_{ao} = \begin{pmatrix} W_{ao}^{11} & W_{ao}^{12} & \dots & W_{ao}^{1m} \\ W_{ao}^{21} & W_{ao}^{22} & \dots & W_{ao}^{2m} \\ \vdots & \vdots & \ddots & \vdots \\ W_{ao}^{C1} & W_{ao}^{C2} & \dots & W_{ao}^{Cm} \end{pmatrix}_{C \times m}$$

$$a^{<t>} = \begin{pmatrix} a_1^{<t>} \\ a_2^{<t>} \\ a_3^{<t>} \\ \vdots \\ a_m^{<t>} \end{pmatrix}_{m \times 1} \quad b_o = \begin{pmatrix} b_{o_1} \\ b_{o_2} \\ b_{o_3} \\ \vdots \\ b_{o_C} \end{pmatrix}_{C \times 1} \quad o^{<t>} = \begin{pmatrix} o_1^{<t>} \\ o_2^{<t>} \\ o_3^{<t>} \\ \vdots \\ o_C^{<t>} \end{pmatrix}_{C \times 1}$$

$$\begin{cases} o_1^{<t>} = W_{ao}^{11} \cdot a_1^{<t>} + W_{ao}^{12} \cdot a_2^{<t>} + \dots + W_{ao}^{1m} a_m^{<t>} + b_{o_1} \\ o_2^{<t>} = W_{ao}^{21} \cdot a_1^{<t>} + W_{ao}^{22} \cdot a_2^{<t>} + \dots + W_{ao}^{2m} a_m^{<t>} + b_{o_2} \\ \vdots \\ o_C^{<t>} = W_{ao}^{C1} \cdot a_1^{<t>} + W_{ao}^{C2} \cdot a_2^{<t>} + \dots + W_{ao}^{Cm} a_m^{<t>} + b_{o_C} \end{cases}$$

Let $u, v \in \mathbb{N}$ where $u \in [1, C]$ and $v \in [1, m]$

$$\frac{\partial o_u^{\langle t \rangle}}{\partial a_v^{\langle t \rangle}} = \frac{\partial}{\partial a_v^{\langle t \rangle}} \left(w_{ao}^{u1} a_1^{\langle t \rangle} + w_{ao}^{u2} a_2^{\langle t \rangle} + \dots + w_{ao}^{uv} a_v^{\langle t \rangle} + \dots + w_{ao}^{um} a_m^{\langle t \rangle} + b_{o_u}^{\langle t \rangle} \right)$$

(a_1, a_2, a_m → 0; $w_{ao}^{uv} a_v$ → w_{ao}^{uv})

$$= w_{ao}^{uv}$$

$$\therefore \text{WLOG} \quad \boxed{\frac{\partial o^{\langle t \rangle}}{\partial a^{\langle t \rangle}} = W_{ao}} \quad \square$$

$\star$ From $h^{\langle t \rangle} = W_{xh} \cdot x^{\langle t \rangle} + W_{ah} * a^{\langle t-1 \rangle} + b_h$

We want $\frac{\partial h^{\langle t \rangle}}{\partial a^{\langle t-1 \rangle}}$

$$W_{xh} = \begin{pmatrix} w_{xh}^{11} & w_{xh}^{12} & \dots & w_{xh}^{1n} \\ w_{xh}^{21} & w_{xh}^{22} & \dots & w_{xh}^{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{xh}^{m2} & w_{xh}^{m2} & \dots & w_{xh}^{mn} \end{pmatrix}_{m \times n}, \quad x^{\langle t \rangle} = \begin{pmatrix} x_1^{\langle t \rangle} \\ x_2^{\langle t \rangle} \\ x_3^{\langle t \rangle} \\ \vdots \\ x_n^{\langle t \rangle} \end{pmatrix}_{n \times 1}$$

$$W_{ah} = \begin{pmatrix} w_{ah}^{11} & w_{ah}^{12} & \dots & w_{ah}^{1m} \\ w_{ah}^{21} & w_{ah}^{22} & \dots & w_{ah}^{2m} \\ \vdots & \vdots & \ddots & \vdots \\ w_{ah}^{m1} & w_{ah}^{m2} & \dots & w_{ah}^{mm} \end{pmatrix}_{m \times m}, \quad a^{\langle t-1 \rangle} = \begin{pmatrix} a_1^{\langle t-1 \rangle} \\ a_2^{\langle t-1 \rangle} \\ a_3^{\langle t-1 \rangle} \\ \vdots \\ a_m^{\langle t-1 \rangle} \end{pmatrix}_{m \times 1}$$

$$h^{\langle t \rangle} = \begin{pmatrix} h_1^{\langle t \rangle} \\ h_2^{\langle t \rangle} \\ \vdots \\ h_m^{\langle t \rangle} \end{pmatrix}_{m \times 1}, \quad b_h = \begin{pmatrix} b_{h_1} \\ b_{h_2} \\ \vdots \\ b_{h_m} \end{pmatrix}_{m \times 1}$$

$$\begin{cases} h_1^{<t>} = \left(w_{xh}^{11} \cdot x_1^{<t>} + w_{xh}^{12} x_2^{<t>} + \dots + w_{xh}^{1m} x_n^{<t>}\right) + \left(w_{ah}^{11} \cdot a_1^{<t-1>} + w_{ah}^{12} a_2^{<t-1>} + \dots + w_{ah}^{1m} a_m^{<t-1>}\right) + b_{h_1} \\ h_2^{<t>} = \left(w_{xh}^{21} \cdot x_1^{<t>} + w_{xh}^{22} x_2^{<t>} + \dots + w_{xh}^{2m} x_n^{<t>}\right) + \left(w_{ah}^{21} \cdot a_1^{<t-1>} + w_{ah}^{22} a_2^{<t-1>} + \dots + w_{ah}^{2m} a_m^{<t-1>}\right) + b_{h_2} \\ \vdots \\ h_m^{<t>} = \left(w_{xh}^{m1} \cdot x_1^{<t>} + w_{xh}^{m2} x_2^{<t>} + \dots + w_{xh}^{mm} x_n^{<t>}\right) + \left(w_{ah}^{m1} \cdot a_1^{<t-1>} + w_{ah}^{m2} a_2^{<t-1>} + \dots + w_{ah}^{mm} a_m^{<t-1>}\right) + b_{h_m} \end{cases}$$

Let $u, v \in \mathbb{N}$, $u \in [1, m]$ and $v \in [1, m]$

$$\frac{\partial h_u^{<t>}}{\partial a_v^{<t-1>}} = \frac{\partial}{\partial a_v^{<t-1>}} \Bigg(\underbrace{\left(w_{xh}^{u1} \cdot x_1^{<t>} + w_{xh}^{u2} x_2^{<t>} + \dots + w_{xh}^{uv} x_v^{<t>} + \dots + w_{xh}^{un} x_n^{<t>}\right)}_{\to 0}$$

$$+ \left(\underbrace{w_{ah}^{u1} \cdot a_1^{<t-1>}}_{0} + \underbrace{w_{ah}^{u2} a_2^{<t-1>}}_{0} + \dots + \underbrace{w_{ah}^{uv} a_v^{<t-1>}}_{w_{ah}^{uv}} + \dots + \underbrace{w_{ah}^{um} a_m^{<t-1>}}_{0}\right) + \underbrace{b_{h_u}}_{0} \Bigg)$$

$$= w_{ah}^{uv}$$

$$\therefore \text{WLOG} \quad \frac{\partial h^{<t>}}{\partial a^{<t-1>}} = W_{ah} \quad \square$$

$$\overset{t+1}{\longleftrightarrow} \quad \frac{\partial h^{<t+1>}}{\partial a^{<t>}} = W_{ah} \quad \square$$

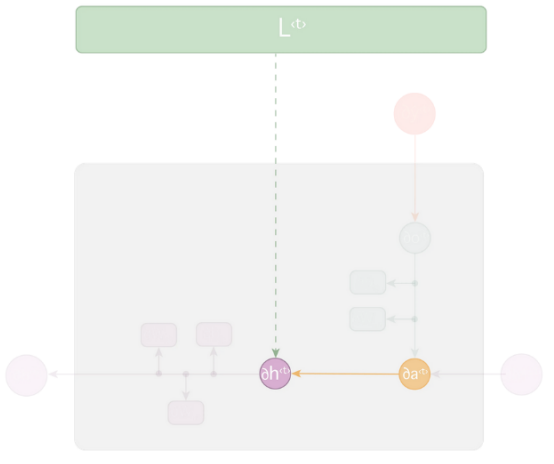
$$\frac{\partial L^{<t>}}{\partial a^{<t>}} = \frac{\partial L^{<t>}}{\partial o^{<t>}} \cdot \frac{\partial o^{<t>}}{a^{<t>}} + \frac{\partial L^{<t>}}{\partial h^{<t+1>}} \cdot \frac{\partial h^{<t+1>}}{\partial a^{<t>}} \quad \therefore$$

$$\boxed{\frac{\partial L^{<t>}}{\partial a^{<t>}} = \underbrace{W_{ao}^{T}}_{m \times c} \cdot \underbrace{\frac{\partial L^{<t>}}{\partial o^{<t>}}}_{c \times 1} + \underbrace{W_{ah}^{T}}_{m \times m} \cdot \underbrace{\frac{\partial L^{<t>}}{\partial h^{<t+1>}}}_{m \times 1}} \quad (8)$$

Step 7) Calculate $\frac{\partial L^{<t>}}{\partial h^{<t>}} = \left(\frac{\partial L^{<t>}}{\partial h_1^{<t>}}, \frac{\partial L^{<t>}}{\partial h_2^{<t>}}, \dots, \frac{\partial L^{<t>}}{\partial h_m^{<t>}}\right)^T$ $1 \times m$

Chain - Rule

$L^{<t>} \rightarrow \hat{y}^{<t>} \rightarrow o^{<t>} \rightarrow a^{<t>} \rightarrow h^{<t>}$



$$\frac{\partial L^{<t>}}{\partial h^{<t>}} = \frac{\partial L^{<t>}}{\partial o^{<t>}} \cdot \frac{\partial a^{<t>}}{\partial h^{<t>}}$$

From $a^{<t>} = \tanh(h^{<t>})$ we want $\frac{\partial a^{<t>}}{\partial h^{<t>}}$

$$a^{<t>} = \begin{pmatrix} a_1^{<t>} \\ a_2^{<t>} \\ a_3^{<t>} \\ \vdots \\ a_m^{<t>} \end{pmatrix}_{m \times 1}, \quad h^{<t>} = \begin{pmatrix} h_1^{<t>} \\ h_2^{<t>} \\ \vdots \\ h_m^{<t>} \end{pmatrix}_{m \times 1}$$

$$\begin{cases} a_1^{<t>} = \tanh(h_1^{<t>}) \\ a_2^{<t>} = \tanh(h_2^{<t>}) \\ \vdots \\ a_m^{<t>} = \tanh(h_m^{<t>}) \end{cases}$$

Let $u \in \mathbb{N}$ where $u \in [1, m]$

$$\frac{\partial a_u^{<t>}}{\partial h_u^{<t>}} = \frac{\partial}{\partial} \left(\tanh(h_u^{<t>})\right)$$

$$= 1 - \left(a_u^{<t>}\right)^2$$

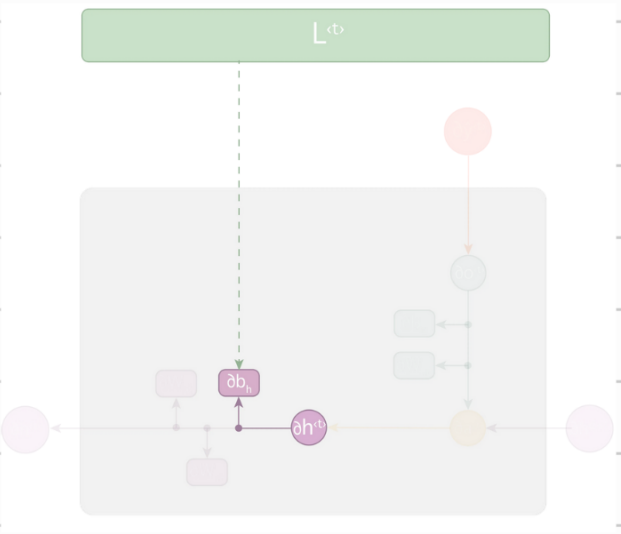
$\therefore$ WLOG $\frac{\partial a^{<t>}}{\partial h^{<t>}} = 1 - \left(a^{<t>}\right)^2$

$$\boxed{\frac{\partial L^{<t>}}{\partial h^{<t>}} = \frac{\partial L^{<t>}}{\partial a^{<t>}} \cdot \left(1 - \left(a^{<t>}\right)^2\right)} \quad (9)$$

Step 8) Calculate $\frac{\partial L^{<t>}}{\partial b_h} = \left(\frac{\partial L^{<t>}}{\partial b_{h_1}}, \frac{\partial L^{<t>}}{\partial b_{h_2}}, \dots, \frac{\partial L^{<t>}}{\partial b_{h_m}} \right)$

Chain rule

$L^{<t>} \rightarrow \hat{y}^{<t>} \rightarrow o^{<t>} \rightarrow a^{<t>} \rightarrow h^{<t>} \rightarrow b_h$



$$\frac{\partial L^{<t>}}{\partial b_h} = \frac{\partial L^{<t>}}{\partial h^{<t>}} \cdot \frac{\partial h^{<t>}}{\partial b_h}$$

* From $h^{<t>} = W_{xh} \cdot x^{<t>} + W_{ah} * a^{<t-1>} + b_h$

We want $\frac{\partial h^{<t>}}{\partial b_h}$

$$W_{xh} = \begin{pmatrix} w^{11}_{xh} & w^{12}_{xh} & \dots & w^{1n}_{xh} \\ w^{21}_{xh} & w^{22}_{xh} & \dots & w^{2n}_{xh} \\ \vdots & \vdots & \ddots & \vdots \\ w^{m2}_{xh} & w^{m2}_{xh} & \dots & w^{mn}_{xh} \end{pmatrix}_{m \times n}, \quad x^{<t>} = \begin{pmatrix} x_1^{<t>} \\ x_2^{<t>} \\ x_3^{<t>} \\ \vdots \\ x_n^{<t>} \end{pmatrix}_{n \times 1}$$

$$W_{ah} = \begin{pmatrix} w^{11}_{ah} & w^{12}_{ah} & \dots & w^{1m}_{ah} \\ w^{21}_{ah} & w^{22}_{ah} & \dots & w^{2m}_{ah} \\ \vdots & \vdots & \ddots & \vdots \\ w^{m1}_{ah} & w^{m2}_{ah} & \dots & w^{mm}_{ah} \end{pmatrix}_{m \times m}, \quad a^{<t-1>} = \begin{pmatrix} a_1^{<t-1>} \\ a_2^{<t-1>} \\ a_3^{<t-1>} \\ \vdots \\ a_m^{<t-1>} \end{pmatrix}_{m \times 1}$$

$$h^{<t>} = \begin{pmatrix} h_1^{<t>} \\ h_2^{<t>} \\ \vdots \\ h_m^{<t>} \end{pmatrix}_{m \times 1}, \quad b_h = \begin{pmatrix} b_{h_1} \\ b_{h_2} \\ \vdots \\ b_{h_m} \end{pmatrix}_{m \times 1}$$

$$\begin{cases} h_1^{<t>} = \left(w_{xh}^{11} \cdot x_1^{<t>} + w_{xh}^{12} x_2^{<t>} + \dots + w_{xh}^{1m} x_m^{<t>}\right) + \left(w_{ah}^{11} \cdot a_1^{<t-1>} + w_{ah}^{12} a_2^{<t-1>} + \dots + w_{ah}^{1m} a_m^{<t-1>}\right) + b_{h_1} \\ h_2^{<t>} = \left(w_{xh}^{21} \cdot x_1^{<t>} + w_{xh}^{22} x_2^{<t>} + \dots + w_{xh}^{2m} x_m^{<t>}\right) + \left(w_{ah}^{21} \cdot a_1^{<t-1>} + w_{ah}^{22} a_2^{<t-1>} + \dots + w_{ah}^{2m} a_m^{<t-1>}\right) + b_{h_2} \\ \vdots \\ h_m^{<t>} = \left(w_{xh}^{m1} \cdot x_1^{<t>} + w_{xh}^{m2} x_2^{<t>} + \dots + w_{xh}^{mm} x_m^{<t>}\right) + \left(w_{ah}^{m1} \cdot a_1^{<t-1>} + w_{ah}^{m2} a_2^{<t-1>} + \dots + w_{ah}^{mm} a_m^{<t-1>}\right) + b_{h_m} \end{cases}$$

Let $u \in \mathbb{N}$ where $u \in [1, m]$

$$\frac{\partial h_u^{<t>}}{\partial b_{h_u}} = \frac{\partial}{\partial b_{h_u}}\Bigg(\cancelto{0}{\left(w_{xh}^{u1} \cdot x_1^{<t>} + w_{xh}^{u2} x_2^{<t>} + \dots + w_{xh}^{um} x_m^{<t>}\right)}$$

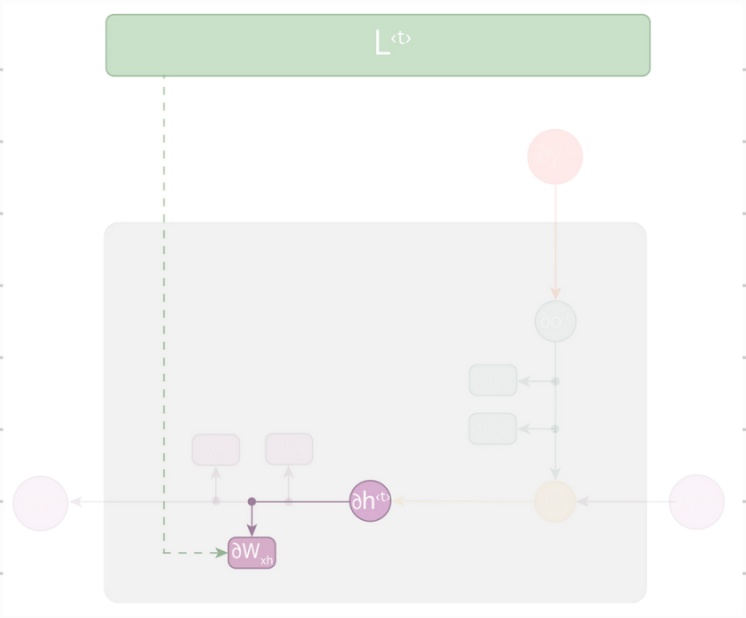
$$+ \cancelto{0}{\left(w_{ah}^{u1} \cdot a_1^{<t-1>} + w_{ah}^{u2} a_2^{<t-1>} + \dots + w_{ah}^{um} a_m^{<t-1>}\right)} + \underbrace{b_{h_u}}_{1}\Bigg) = 1$$

$$\therefore \text{WLOG } \frac{\partial h^{<t>}}{\partial b_h} = 1 \text{ (unitary vector)}$$

$$\longrightarrow \boxed{\frac{\partial L^{<t>}}{\partial b_h} = \frac{\partial L^{<t>}}{\partial h^{<t>}}} \quad (10)$$

Step 9) Compute $\frac{\partial L^{<t>}}{\partial W_{xh}}$

$$= \begin{pmatrix} \frac{\partial L^{<t>}}{\partial w_{xh}^{11}} & \frac{\partial L^{<t>}}{\partial w_{xh}^{12}} & \cdots & \frac{\partial L^{<t>}}{\partial w_{xh}^{1n}} \\ \frac{\partial L^{<t>}}{\partial w_{xh}^{21}} & \frac{\partial L^{<t>}}{\partial w_{xh}^{22}} & \cdots & \frac{\partial L^{<t>}}{\partial w_{xh}^{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial L^{<t>}}{\partial w_{xh}^{m1}} & \frac{\partial L^{<t>}}{\partial w_{xh}^{m2}} & \cdots & \frac{\partial L^{<t>}}{\partial w_{xh}^{mn}} \end{pmatrix}_{m \times n}$$



Chain Rule

$$L^{<t>} \rightarrow \hat{y}^{<t>} \rightarrow o^{<t>} \rightarrow a^{<t>} \rightarrow h^{<t>} \rightarrow W_{xh}$$

$$\frac{\partial L^{<t>}}{\partial W_{xh}} = \frac{\partial L^{<t>}}{\partial h^{<t>}} \cdot \frac{\partial h^{<t>}}{\partial W_{xh}}$$

* From $h^{<t>} = W_{xh} \cdot x^{<t>} + W_{ah} * a^{<t-1>} + b_h$

We want $\frac{\partial h^{<t>}}{\partial W_{xh}}$

$$W_{xh} = \begin{pmatrix} w_{xh}^{11} & w_{xh}^{12} & \cdots & w_{xh}^{1n} \\ w_{xh}^{21} & w_{xh}^{22} & \cdots & w_{xh}^{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{xh}^{m2} & w_{xh}^{m2} & \cdots & w_{xh}^{mn} \end{pmatrix}_{m \times n} \quad x^{<t>} = \begin{pmatrix} x_1^{<t>} \\ x_2^{<t>} \\ x_3^{<t>} \\ \vdots \\ x_n^{<t>} \end{pmatrix}_{n \times 1}$$

$$W_{ah} = \begin{pmatrix} w_{ah}^{11} & w_{ah}^{12} & \dots & w_{ah}^{1m} \\ w_{ah}^{21} & w_{ah}^{22} & \dots & w_{ah}^{2m} \\ \vdots & \vdots & \ddots & \vdots \\ w_{ah}^{m1} & w_{ah}^{m2} & \dots & w_{ah}^{mm} \end{pmatrix}_{m \times m}, \quad a^{\langle t-1 \rangle} = \begin{pmatrix} a_1^{\langle t-1 \rangle} \\ a_2^{\langle t-1 \rangle} \\ a_3^{\langle t-1 \rangle} \\ \vdots \\ a_m^{\langle t-1 \rangle} \end{pmatrix}_{m \times 1}$$

$$h^{\langle t \rangle} = \begin{pmatrix} h_1^{\langle t \rangle} \\ h_2^{\langle t \rangle} \\ \vdots \\ h_m^{\langle t \rangle} \end{pmatrix}_{m \times 1}, \quad b_h = \begin{pmatrix} b_{h_1} \\ b_{h_2} \\ \vdots \\ b_{h_m} \end{pmatrix}_{m \times 1}$$

$$\begin{cases} h_1^{\langle t \rangle} = \left(w_{xh}^{11} \cdot x_1^{\langle t \rangle} + w_{xh}^{12} x_2^{\langle t \rangle} + \dots + w_{xh}^{1m} x_m^{\langle t \rangle}\right) + \left(w_{ah}^{11} \cdot a_1^{\langle t-1 \rangle} + w_{ah}^{12} a_2^{\langle t-1 \rangle} + \dots + w_{ah}^{1m} a_m^{\langle t-1 \rangle}\right) + b_{h_1} \\ h_2^{\langle t \rangle} = \left(w_{xh}^{21} \cdot x_1^{\langle t \rangle} + w_{xh}^{22} x_2^{\langle t \rangle} + \dots + w_{xh}^{2m} x_m^{\langle t \rangle}\right) + \left(w_{ah}^{21} \cdot a_1^{\langle t-1 \rangle} + w_{ah}^{22} a_2^{\langle t-1 \rangle} + \dots + w_{ah}^{2m} a_m^{\langle t-1 \rangle}\right) + b_{h_2} \\ \vdots \\ h_m^{\langle t \rangle} = \left(w_{xh}^{m1} \cdot x_1^{\langle t \rangle} + w_{xh}^{m2} x_2^{\langle t \rangle} + \dots + w_{xh}^{mm} x_m^{\langle t \rangle}\right) + \left(w_{ah}^{m1} \cdot a_1^{\langle t-1 \rangle} + w_{ah}^{m2} a_2^{\langle t-1 \rangle} + \dots + w_{ah}^{mm} a_m^{\langle t-1 \rangle}\right) + b_{h_m} \end{cases}$$

Let $u, v \in \mathbb{N}$ where $u \in [1, m]$ and $v \in [1, n]$

$$\frac{\partial h_u^{\langle t \rangle}}{\partial w_{xh}^{uv}} = \frac{\partial}{\partial w_{xh}^{uv}} \left(\left(w_{xh}^{u1} \cdot x_1^{\langle t \rangle} + w_{xh}^{u2} x_2^{\langle t \rangle} + \dots + w_{xh}^{uv} x_v^{\langle t \rangle} + \dots + w_{xh}^{um} x_m^{\langle t \rangle}\right) + \left(w_{ah}^{u1} \cdot a_1^{\langle t-1 \rangle} + w_{ah}^{u2} a_2^{\langle t-1 \rangle} + \dots + w_{ah}^{um} a_m^{\langle t-1 \rangle}\right) + b_{h_u} \right)$$

$$= x_v^{\langle t \rangle}$$

$$\therefore \text{WLOG } \frac{\partial h^{\langle t \rangle}}{\partial W_{xh}} = x^{\langle t \rangle}$$

$$\frac{\partial L^{<t>}}{\partial W_{xh}} = \frac{\partial h^{<t>}}{\partial W_{xh}} \cdot x^{<t>T} \quad (11)$$

Note :

$$\underbrace{\begin{pmatrix} \frac{\partial L^{<t>}}{\partial h_1^{<t>}} \\ \frac{\partial L^{<t>}}{\partial h_2^{<t>}} \\ \vdots \\ \frac{\partial L^{<t>}}{\partial h_m^{<t>}} \end{pmatrix}}_{m \times 1} \cdot \underbrace{\begin{pmatrix} x_1^{<t>} & x_2^{<t>} & \cdots & x_m^{<t>} \end{pmatrix}}_{1 \times m} = \underbrace{\frac{\partial L^{<t>}}{\partial W_{xh}}}_{m \times m}$$

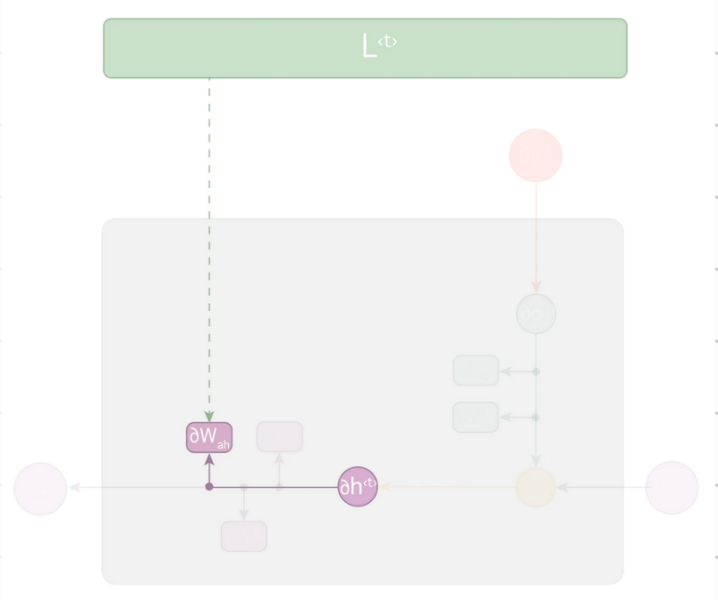
$x^{<t>T}$

$$\frac{\partial L^{<t>}}{\partial h^{<t>}}$$

Step 10) Compute $\frac{\partial L^{<t>}}{\partial W_{ah}} =$

$$\begin{pmatrix} \frac{\partial L^{<t>}}{\partial w_{ah}^{11}} & \frac{\partial L^{<t>}}{\partial w_{ah}^{12}} & \cdots & \frac{\partial L^{<t>}}{\partial w_{ah}^{1m}} \\ \frac{\partial L^{<t>}}{\partial w_{ah}^{21}} & \frac{\partial L^{<t>}}{\partial w_{ah}^{22}} & \cdots & \frac{\partial L^{<t>}}{\partial w_{ah}^{2m}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial L^{<t>}}{\partial w_{ah}^{m1}} & \frac{\partial L^{<t>}}{\partial w_{ah}^{m2}} & \cdots & \frac{\partial L^{<t>}}{\partial w_{ah}^{mm}} \end{pmatrix}$$

$m \times m$



Chain Rule

$$L^{<t>} \rightarrow \hat{y}^{<t>} \rightarrow a^{<t>} \rightarrow h^{<t>} \rightarrow W_{ah}$$

$$\frac{\partial L^{<t>}}{\partial W_{ah}} = \frac{\partial L^{<t>}}{\partial h^{<t>}} \cdot \frac{\partial h^{<t>}}{\partial W_{ah}}$$

* From $h^{<t>} = W_{xh} \cdot x^{<t>} + W_{ah} * a^{<t-1>} + b_h$

We want $\frac{\partial h^{<t>}}{\partial W_{ah}}$

$$W_{xh} = \begin{pmatrix} W_{xh}^{11} & W_{xh}^{12} & \cdots & W_{xh}^{1m} \\ W_{xh}^{21} & W_{xh}^{22} & \cdots & W_{xh}^{2m} \\ \vdots & \vdots & \ddots & \vdots \\ W_{xh}^{m2} & W_{xh}^{m2} & \cdots & W_{xh}^{mm} \end{pmatrix}_{m \times m}, \quad x^{<t>} = \begin{pmatrix} x_1^{<t>} \\ x_2^{<t>} \\ x_3^{<t>} \\ \vdots \\ x_n^{<t>} \end{pmatrix}_{n \times 1}$$

$$W_{ah} = \begin{pmatrix} W_{ah}^{11} & W_{ah}^{12} & \cdots & W_{ah}^{1m} \\ W_{ah}^{21} & W_{ah}^{22} & \cdots & W_{ah}^{2m} \\ \vdots & \vdots & \ddots & \vdots \\ W_{ah}^{m1} & W_{ah}^{m2} & \cdots & W_{ah}^{mm} \end{pmatrix}_{m \times m}, \quad a^{<t-1>} = \begin{pmatrix} a_1^{<t-1>} \\ a_2^{<t-1>} \\ a_3^{<t-1>} \\ \vdots \\ a_m^{<t-1>} \end{pmatrix}_{m \times 1}$$

$$h^{<t>} = \begin{pmatrix} h_1^{<t>} \\ h_2^{<t>} \\ \vdots \\ h_m^{<t>} \end{pmatrix}_{m \times 1}, \quad b_h = \begin{pmatrix} b_{h_1} \\ b_{h_2} \\ \vdots \\ b_{h_m} \end{pmatrix}_{m \times 1}$$

$$\begin{cases} h_1^{<t>} = \left(w_{xh}^{11} \cdot x_1^{<t>} + w_{xh}^{12} x_2^{<t>} + \ldots + w_{xh}^{1m} x_m^{<t>}\right) + \left(w_{ah}^{11} \cdot a_1^{<t-1>} + w_{ah}^{12} a_2^{<t-1>} + \ldots + w_{ah}^{1m} a_m^{<t-1>}\right) + b_{h_1} \\ h_2^{<t>} = \left(w_{xh}^{21} \cdot x_1^{<t>} + w_{xh}^{22} x_2^{<t>} + \ldots + w_{xh}^{2m} x_m^{<t>}\right) + \left(w_{ah}^{21} \cdot a_1^{<t-1>} + w_{ah}^{22} a_2^{<t-1>} + \ldots + w_{ah}^{2m} a_m^{<t-1>}\right) + b_{h_2} \\ \vdots \\ h_m^{<t>} = \left(w_{xh}^{m1} \cdot x_1^{<t>} + w_{xh}^{m2} x_2^{<t>} + \ldots + w_{xh}^{mm} x_m^{<t>}\right) + \left(w_{ah}^{m1} \cdot a_1^{<t-1>} + w_{ah}^{m2} a_2^{<t-1>} + \ldots + w_{ah}^{mm} a_m^{<t-1>}\right) + b_{h_m} \end{cases}$$

Let $u, v \in \mathbb{N}$ where $u, v \in [1, m]$

$$\frac{\partial h_u^{<t>}}{\partial w_{ah}^{uv}} = \frac{\partial}{\partial w_{ah}^{uv}}\left(\left(w_{xh}^{u1} \cdot x_1^{<t>} + w_{xh}^{u2} x_2^{<t>} + \ldots + w_{xh}^{uv} x_v^{<t>} + \ldots + w_{xh}^{um} x_m^{<t>}\right) + \left(w_{ah}^{u1} \cdot a_1^{<t-1>} + w_{ah}^{u2} a_2^{<t-1>} + \ldots + w_{ah}^{uv} a_v^{<t-1>} + \ldots + w_{ah}^{um} a_m^{<t-1>}\right) + b_{h_u}\right)$$

(the x terms and the $a_1^{<t-1>}, a_2^{<t-1>}, a_m^{<t-1>}$ terms are constants with respect to w_{ah}^{uv}, so they vanish; only $a_v^{<t-1>}$ remains)

$$= a_v^{<t-1>}$$

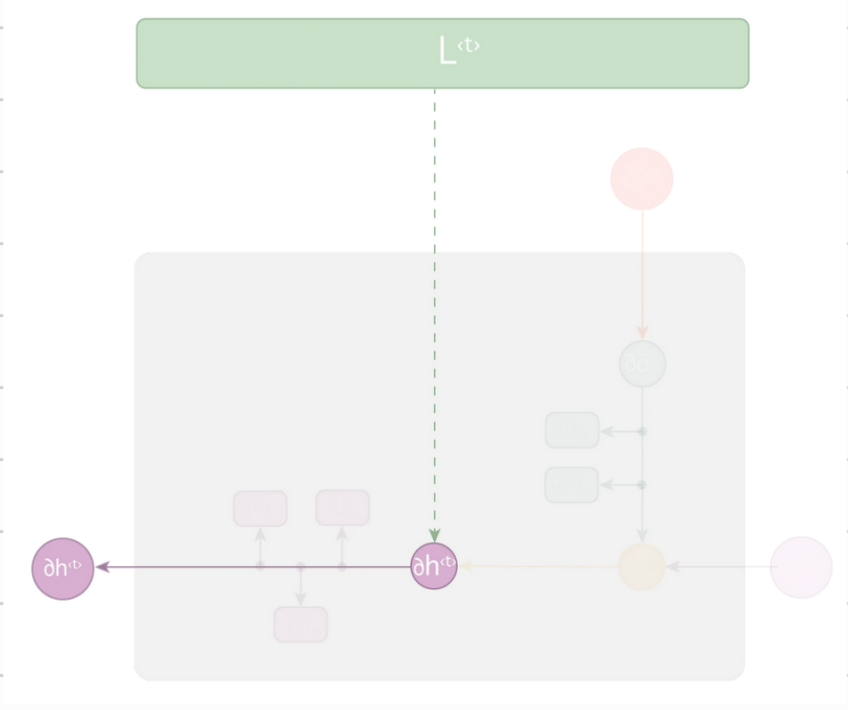
$\therefore$ WLOG $\dfrac{\partial h^{<t>}}{\partial W_{ah}} = a^{<t-1>}$

$$\boxed{\frac{\partial L^{<t>}}{\partial W_{ah}} = \frac{\partial L^{<t>}}{\partial h^{<t>}} \cdot a^{<t-1>^T}} \quad (12)$$

Step 11) Pass $\dfrac{\partial L^{<t>}}{\partial h^{<t>}}$ to previous cell

Therefore,

$$\boxed{\frac{\partial L^{<t-1>}}{\partial h^{<t>}} = \frac{\partial L^{<t>}}{\partial h^{<t>}}} \quad (13)$$


$L^{\langle t \rangle}$
$\partial h^{\langle t \rangle}$
$\partial h^{\langle t \rangle}$

Equations

Backward

$$\frac{\partial L^{<t>}}{\partial o^{<t>}} = \hat{y}^{<t>} - y^{<t>} \qquad (a)$$

$$\frac{\partial L^{<t>}}{\partial h^{<t+1>}} = \frac{\partial L^{<t+1>}}{\partial h^{<t+1>}} \qquad (b)$$

$$\frac{\partial L^{<t>}}{\partial a^{<t>}} = W_{ao}^{T} \cdot \frac{\partial L^{<t>}}{\partial o^{<t>}} + W_{ah}^{T} \cdot \frac{\partial L^{<t>}}{\partial h^{<t+1>}} \qquad (c)$$

$$\frac{\partial L^{<t>}}{\partial h^{<t>}} = \frac{\partial L^{<t>}}{\partial a^{<t>}} \cdot \left(1 - a^{<t>2}\right) \qquad (d)$$

$$\frac{\partial L}{\partial W_{ao}} = \sum_{t=1}^{T} \frac{\partial L^{<t>}}{\partial o^{<t>}} \cdot \left(a^{<t>}\right)^{T} \qquad (e)$$

$$\frac{\partial L}{\partial b_o} = \sum_{t=1}^{T} \frac{\partial L^{<t>}}{\partial o^{<t>}} \qquad (f)$$

$$\frac{\partial L}{\partial W_{ah}} = \sum_{t=1}^{T} \frac{\partial L^{<t>}}{\partial h^{<t>}} \cdot \left(a^{<t-1>}\right)^{T} \qquad (g)$$

$$\frac{\partial L}{\partial W_{xh}} = \sum_{t=1}^{T} \frac{\partial L^{<t>}}{\partial h^{<t>}} \cdot \left(x^{<t>}\right)^{T} \qquad (h)$$

$$\frac{\partial L}{\partial b_h} = \sum_{t=1}^{T} \frac{\partial L^{<t>}}{\partial h^{<t>}} \qquad (i)$$

$$W_{ao} \leftarrow W_{ao} - \eta \frac{\partial L}{\partial W_{ao}}$$

$$b_o \leftarrow b_o - \eta \frac{\partial L}{\partial b_o}$$

$$W_{ah} \leftarrow W_{ah} - \eta \frac{\partial L}{\partial W_{ah}}$$

$$W_{xh} \leftarrow W_{xh} - \eta \frac{\partial L}{\partial W_{xh}}$$

$$b_h \leftarrow b_h - \eta \frac{\partial L}{\partial b_h}$$

$$L = \sum_{t=0}^{T} L^{<t>}$$

Forward

$$h^{<t>} = (W_{xh} \cdot x^{<t>}) + (W_{ah} * a^{<t-1>}) + b_h \quad (a)$$

$$a^{<t>} = \tanh(h^{<t>}) \quad (b)$$

$$o^{<t>} = (W_{ao} \cdot a^{<t>}) + b_o \quad (c)$$

$$\hat{y}^{<t>} = \text{softmax}(o^{<t>}) \quad (d)$$

$x^{<t>}$ = Input vector in time t

$h^{<t>}$ = Hidden state

W_{xh} = Matrix that is multiplied by x and generates $h^{<t>}$

W_{ah} = Matrix that is multiplied by a and generates h

$a^{<t>}$ = Activation in time t

b^{h} = Hidden bias

$o^{<t>}$ = Non-normalized output in time t

W_{ao} = Matrix that is multiplied by a and generates o

b_o = Output bias

$\hat{y}^{<t>}$ = Normalized output in time t

$$W_{xh} = \begin{pmatrix} w_{xh}^{11} & w_{xh}^{12} & \dots & w_{xh}^{1m} \\ w_{xh}^{21} & w_{xh}^{22} & \dots & w_{xh}^{2m} \\ \vdots & \vdots & \ddots & \vdots \\ w_{xh}^{m2} & w_{xh}^{m2} & \dots & w_{xh}^{mm} \end{pmatrix}_{m \times m}, \quad x^{<t>} = \begin{pmatrix} x_1^{<t>} \\ x_2^{<t>} \\ x_3^{<t>} \\ \vdots \\ x_n^{<t>} \end{pmatrix}_{n \times 1}$$

$$W_{ah} = \begin{pmatrix} w_{ah}^{11} & w_{ah}^{12} & \dots & w_{ah}^{1m} \\ w_{ah}^{21} & w_{ah}^{22} & \dots & w_{ah}^{2m} \\ \vdots & \vdots & \ddots & \vdots \\ w_{ah}^{m1} & w_{ah}^{m2} & \dots & w_{ah}^{mm} \end{pmatrix}_{m \times m}, \quad a^{<t>} = \begin{pmatrix} a_1^{<t>} \\ a_2^{<t>} \\ a_3^{<t>} \\ \vdots \\ a_m^{<t>} \end{pmatrix}_{m \times 1}$$

$$h^{<t>} = \begin{pmatrix} h_1^{<t>} \\ h_2^{<t>} \\ \vdots \\ h_m^{<t>} \end{pmatrix}_{m \times 1}, \quad b_h = \begin{pmatrix} b_{h_1} \\ b_{h_2} \\ \vdots \\ b_{h_m} \end{pmatrix}_{m \times 1}$$

$$W_{out} = \begin{pmatrix} w^{11}_{out} & w^{12}_{out} & \cdots & w^{1m}_{out} \\ w^{21}_{out} & w^{22}_{out} & \cdots & w^{2m}_{out} \\ \vdots & \vdots & \ddots & \vdots \\ w^{C1}_{out} & w^{C2}_{out} & \cdots & w^{Cm}_{out} \end{pmatrix}_{C \times m}$$

$$b_o = \begin{pmatrix} b_{o_1} \\ b_{o_2} \\ b_{o_3} \\ \vdots \\ b_{o_C} \end{pmatrix}_{C \times 1} \quad , \quad o^{<t>} = \begin{pmatrix} o_1^{<t>} \\ o_2^{<t>} \\ o_3^{<t>} \\ \vdots \\ o_C^{<t>} \end{pmatrix}_{C \times 1} \quad | \quad \hat{y}^{<t>} = \begin{pmatrix} \hat{y}_1^{<t>} \\ \hat{y}_2^{<t>} \\ \hat{y}_3^{<t>} \\ \vdots \\ \hat{y}_C^{<t>} \end{pmatrix}_{C \times 1}$$

Exploding & Vanishing gradient

$$\begin{cases} \dfrac{\partial L^{<t>}}{\partial a^{<t>}} = W_{ao}^{T} \cdot \dfrac{\partial L^{<t>}}{\partial o^{<t>}} + W_{ah}^{T} \cdot \dfrac{\partial L^{<t+1>}}{\partial h^{<t+1>}} \\ \dfrac{\partial L^{<t>}}{\partial h^{<t>}} = \dfrac{\partial L^{<t>}}{\partial a^{<t>}} \cdot \left(1 - a^{<t>2}\right) \end{cases}$$

$$\frac{\partial L^{<t>}}{\partial a^{<t>}} = W_{ao}^{T} \cdot \frac{\partial L^{<t>}}{\partial o^{<t>}} + W_{ah}^{T} \left(\frac{\partial L^{<t+1>}}{\partial a^{<t+1>}} \cdot \left(1 - a^{<t+1>2}\right) \right)$$

$$= W_{ao}^{T} \cdot \frac{\partial L^{<t>}}{\partial o^{<t>}} + W_{ah}^{T} \left(1 - a^{<t+1>2}\right) \cdot \frac{\partial L^{<t+1>}}{\partial a^{<t+1>}}$$

$$= W_{ao}^{T} \cdot \frac{\partial L^{<t>}}{\partial o^{<t>}} + W_{ah}^{T} \left(1 - a^{<t+1>2}\right) \cdot \left(W_{ao}^{T} \cdot \frac{\partial L^{<t+1>}}{\partial o^{<t+1>}} + W_{ah}^{T} \cdot \frac{\partial L^{<t+2>}}{\partial h^{<t+2>}} \right)$$

$$= W_{ao}^{T} \cdot \frac{\partial L^{<t>}}{\partial o^{<t>}} + W_{ao}^{T} \cdot \frac{\partial L^{<t+1>}}{\partial o^{<t+1>}} \left[W_{ah}^{T} \left(1 - a^{<t+1>2}\right) \right] + \left(W_{ah}^{T}\right)^{2} \left(1 - a^{<t+1>2}\right) \cdot \frac{\partial L^{<t+2>}}{\partial h^{<t+2>}}$$

$$= W_{ao}^{T} \cdot \frac{\partial L^{<t>}}{\partial o^{<t>}} + W_{ao}^{T} \cdot \frac{\partial L^{<t+1>}}{\partial o^{<t+1>}} \left[W_{ah}^{T} \left(1 - a^{<t+1>2}\right) \right] + \left(W_{ah}^{T}\right)^{2} \left(1 - a^{<t+1>2}\right) \left(1 - a^{<t+2>2}\right) \cdot \frac{\partial L^{<t+3>}}{\partial a^{<t+3>}}$$

$$= W_{ao}^{T} \cdot \frac{\partial L^{<t>}}{\partial o^{<t>}} + W_{ao}^{T} \cdot \frac{\partial L^{<t+1>}}{\partial o^{<t+1>}} \left[W_{ah}^{T} \left(1 - a^{<t+1>2}\right) \right] + \left(W_{ah}^{T}\right)^{2} \left(1 - a^{<t+1>2}\right) \left(1 - a^{<t+2>2}\right) \cdot$$

$$\left(W_{ao}^{T} \cdot \frac{\partial L^{<t+2>}}{\partial o^{<t+2>}} + W_{ah}^{T} \cdot \frac{\partial L^{<t+3>}}{\partial h^{<t+3>}} \right)$$

$$= W_{ao}^T \cdot \frac{\partial L^{<t>}}{\partial o^{<t>}} + W_{ao}^T \cdot \frac{\partial L^{<t+1>}}{\partial o^{<t+1>}} \left[W_{ah}^T \left(1 - a^{<t+1>2}\right) \right] + W_{ao}^T \cdot \frac{\partial L^{<t+2>}}{\partial o^{<t+2>}} \left[\left(W_{ah}^T\right)^2 \left(1 - a^{<t+1>2}\right)\left(1 - a^{<t+2>2}\right) \right]$$

$$+ \left(W_{ah}^T\right)^3 \left(1 - a^{<t+1>2}\right)\left(1 - a^{<t+2>2}\right)\left(1 - a^{<t+3>2}\right) \cdot \frac{\partial L^{<t+3>}}{\partial a^{<t+3>}}$$

$$\vdots$$

$$\frac{\partial L^{<t>}}{\partial a^{<t>}} = W_{ao}^T \left[\frac{\partial L^{<t>}}{\partial o^{<t>}} + \sum_{i=1}^{T'} \frac{\partial L^{<t+i>}}{\partial o^{<t+i>}} \cdot \left(W_{ah}^T\right)^i \cdot \prod_{j=1}^{i+j-1} 1 - \left(a^{<t+j>}\right)^2 \right]$$

$$+ \left(W_{ah}^T\right)^{T'} \cdot \prod_{i=1}^{T'} 1 - \left(a^{<t+i>}\right)^2$$

Where T' is the last timestamp

* Note that if T' is too big and:

1) W_{ah} and W_{ao} have large values, so $\frac{\partial L^{<t>}}{\partial a^{<t>}}$ will be high and the gradient will overflow (explode)

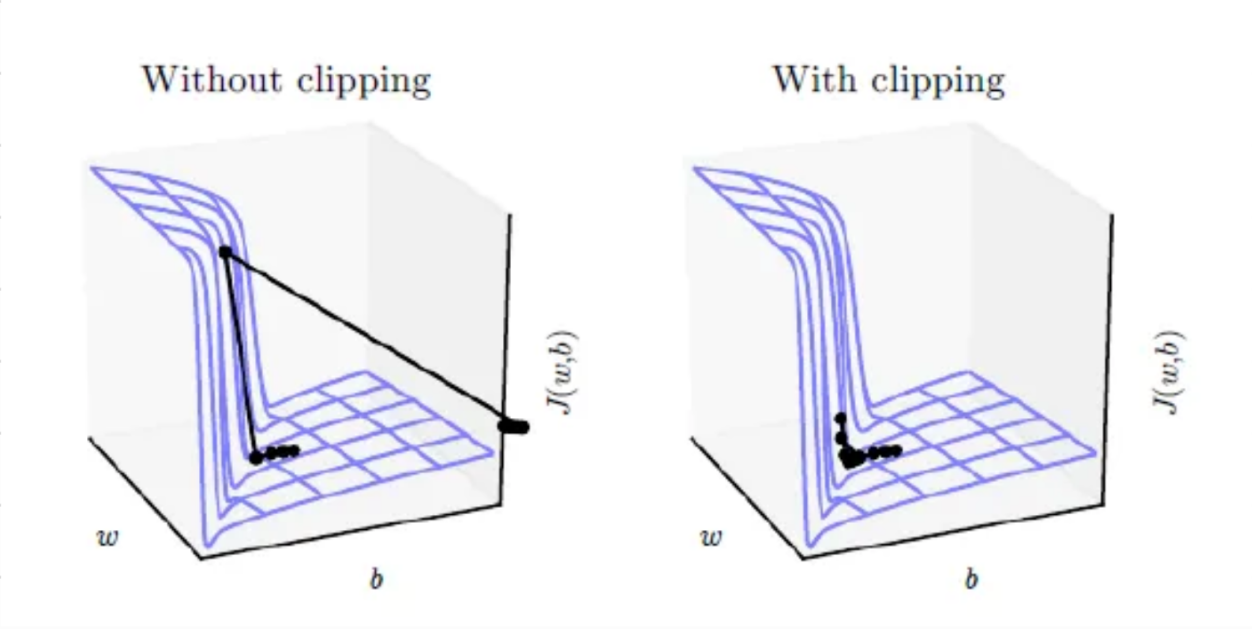
2) W_{ah} and W_{ao} have small values, so $\frac{\partial L^{<t>}}{\partial a^{<t>}}$ will be high and the gradient will underflow (vanish)

Therefore, if a RNN has much cells, it will be hard to learn due to the exploding and vanishing gradient problem

Gradient Clipping: Places a predefined threshold on gradients to prevent them from getting to large. This approach does not change the direction of the gradient, only change it's length

$$\text{if } \|g\| > \text{threshold}:$$
$$g \leftarrow \frac{\text{threshold} \cdot g}{\|g\|}$$

Where g is the gradient and $\|g\|$ is the norm of the gradient



Identity RNN: All weights are initialized to identity matrix and and all activation functions are set to ReLU. This forces the network to stay close to identity function and the problem of vanish/explode gradient is solved