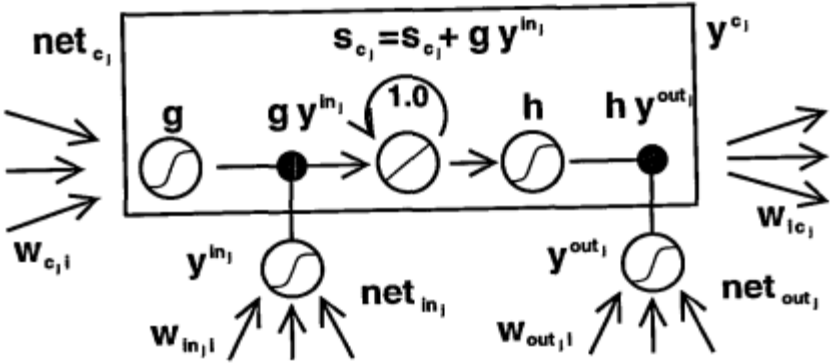
Long Short Term Memory Networks

Brian Cheung

LSTM

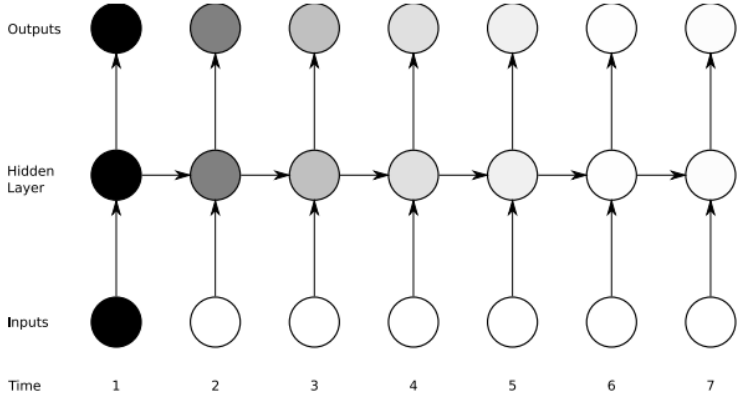


Hochreiter & Schmidhuber 1997

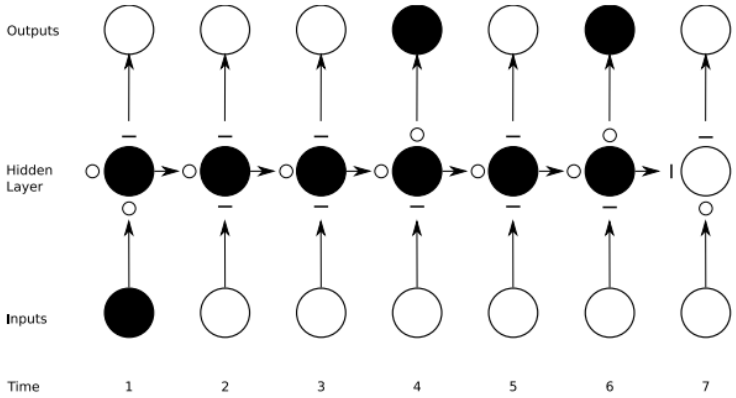
- Proposed by Sepp Hochreiter in 1997
- Originally used approximate error gradient with Real Time Recurrent Learning and truncated backpropagation through time
- Used for processing long range contextual information

LSTMs reduce vanishing gradient problem

Standard Recurrent Network



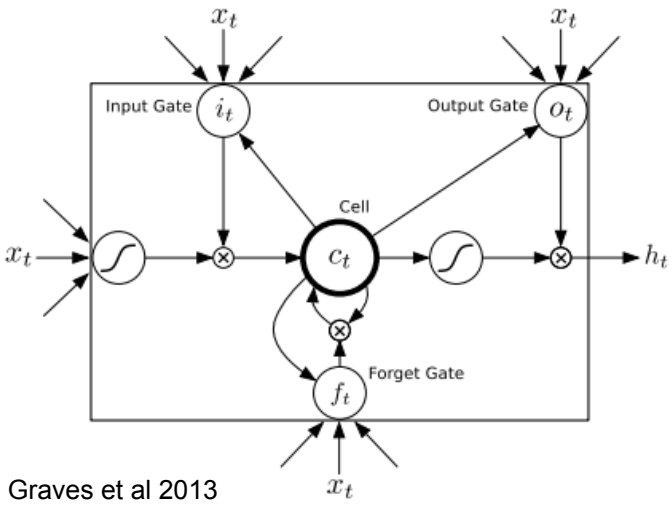
LSTM Network



Graves et al 2013

- The darker the shade, the greater the sensitivity
- The sensitivity decays exponentially over time as new inputs overwrite the activation of hidden unit and the network ‘forgets’ the first input

LSTMs reduce vanishing gradient problem



Graves et al 2013

- Memory cells and gating units allow information to be stored for long periods of time.
- Memory cells are additive in time
  - Gradients also additive in time which alleviates vanishing gradient

Backpropagation through time

- Derivative of objective function, O, with respect to linear activations, a

$$a_j = \sum_{i=1} w_{ij} b_i$$

$$\delta_j^t \stackrel{\text{def}}{=} \frac{\partial O}{\partial a_j^t}$$

- Gradient descent weight update

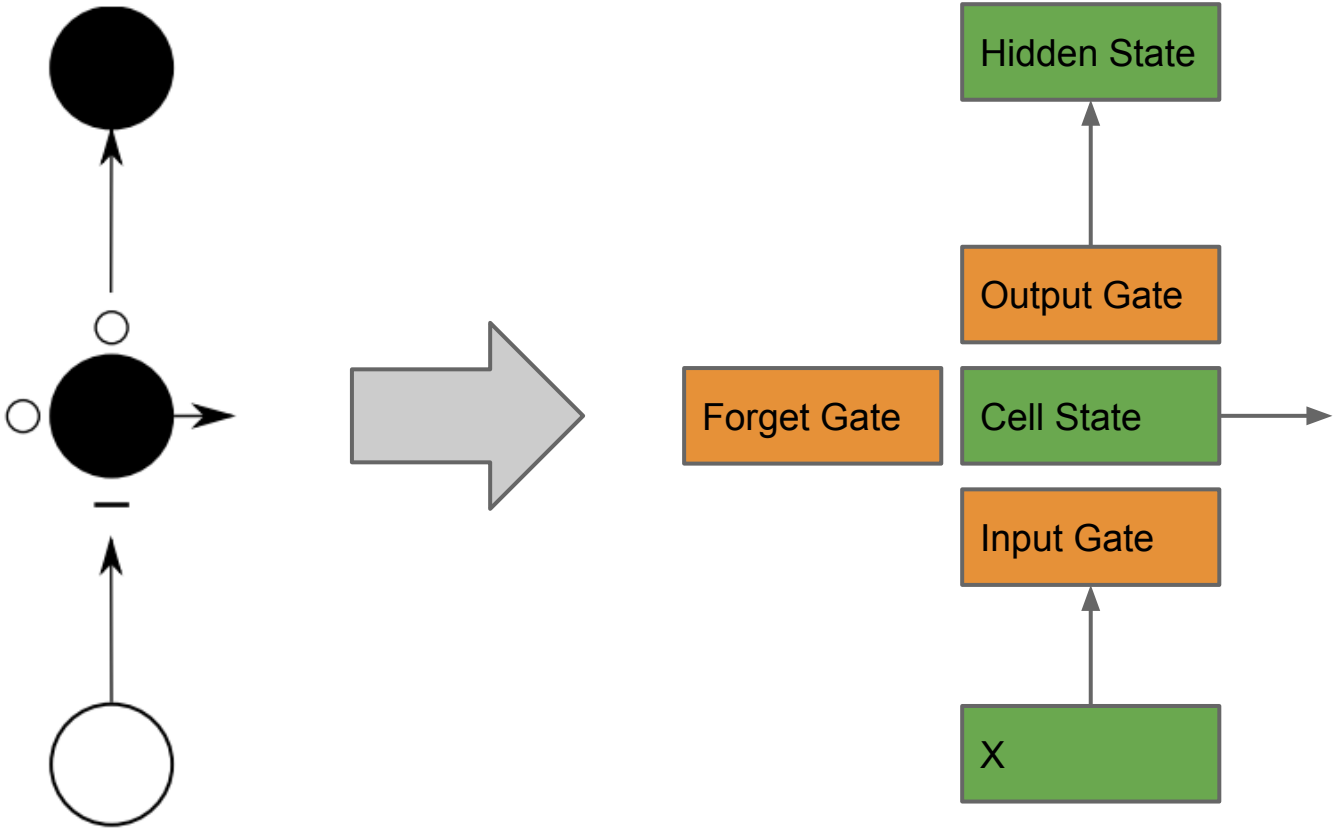
$$\frac{\partial O}{\partial w_{ij}} = \sum_{t=1}^{T} \frac{\partial O}{\partial a_j^t} \frac{\partial a_j^t}{\partial w_{ij}} = \sum_{t=1}^{T} \delta_j^t b_i^t$$

$$\Delta \mathbf{w}(n) = -\alpha \frac{\partial O}{\partial \mathbf{w}(n)}$$

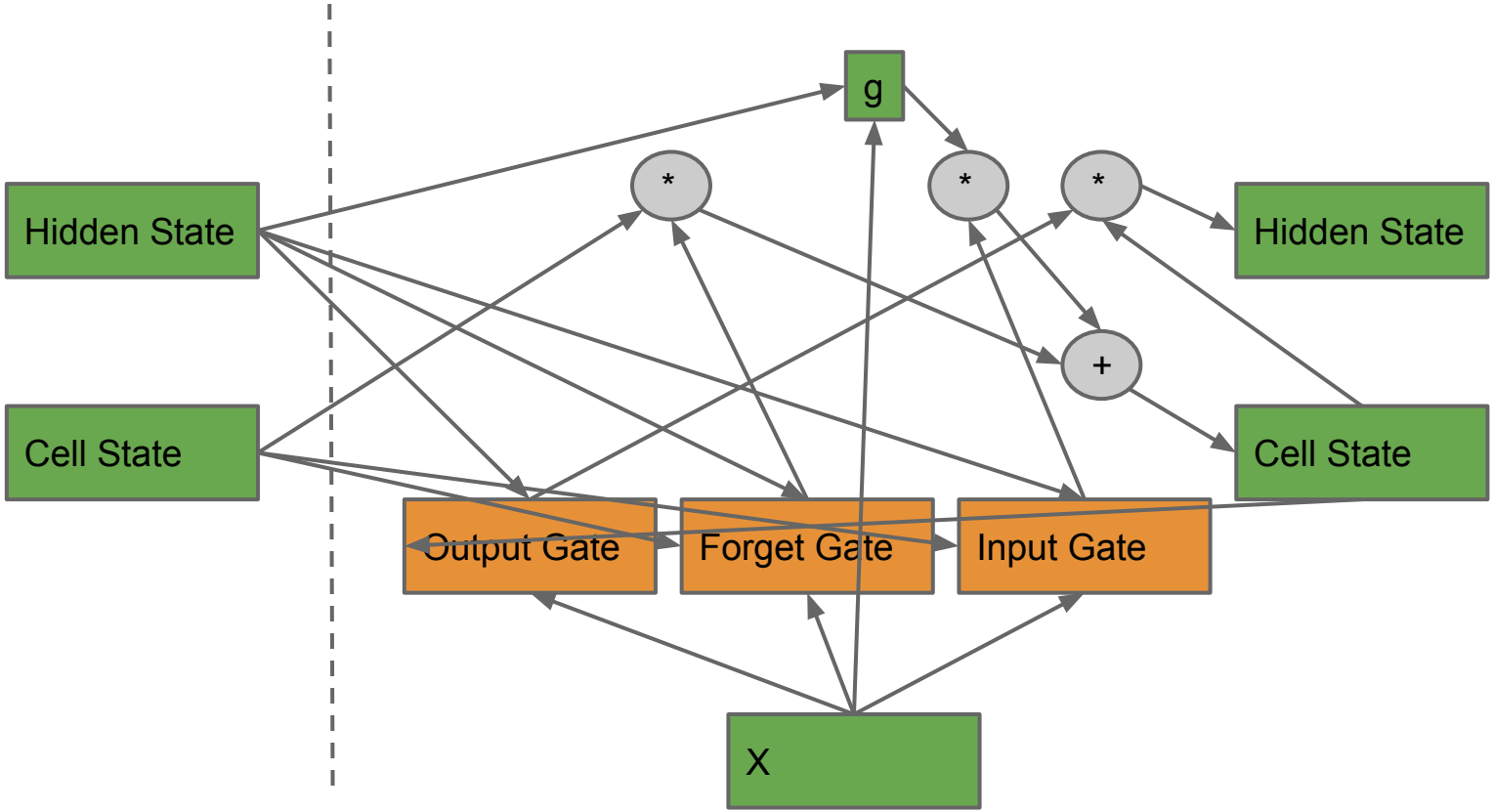
Anatomy of an LSTM node

x	X	Input to the LSTM node at a particular time point (character, phoneme, word)
ι	Input Gate	Determines whether the input will be stored in the memory cell
φ	Forget Gate	Determines whether current contents of memory will be forgotten (erased)
ω	Output Gate	Determines if current memory contents will be output
c	Cell State	Memory Cell
h	Hidden State	Output of LSTM node
g	g	Input nonlinearity of Memory Cell (i.e. tanh, sigmoid)

Anatomy of an LSTM node



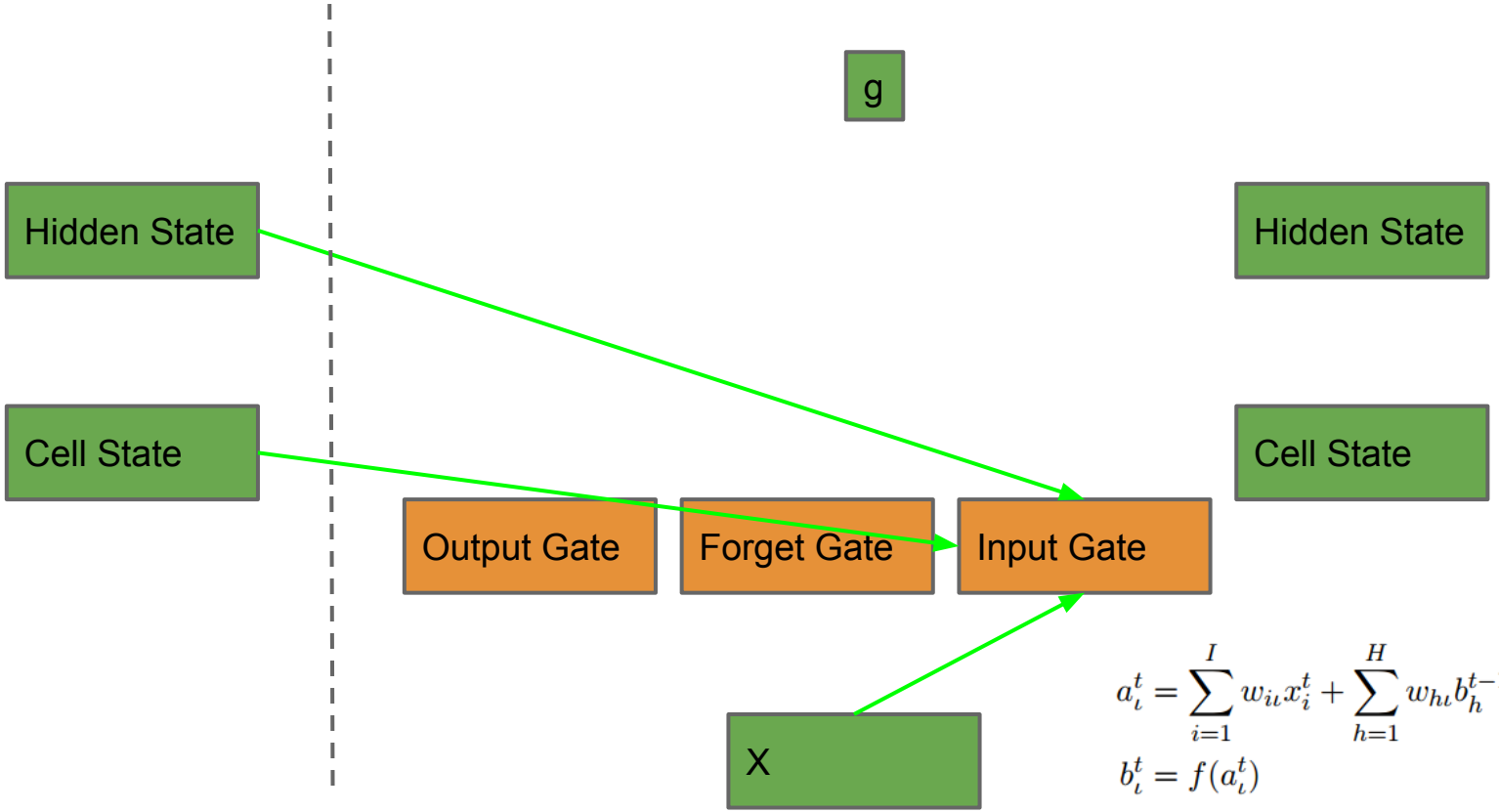
LSTM Forward Propagation



Time: t-1

t

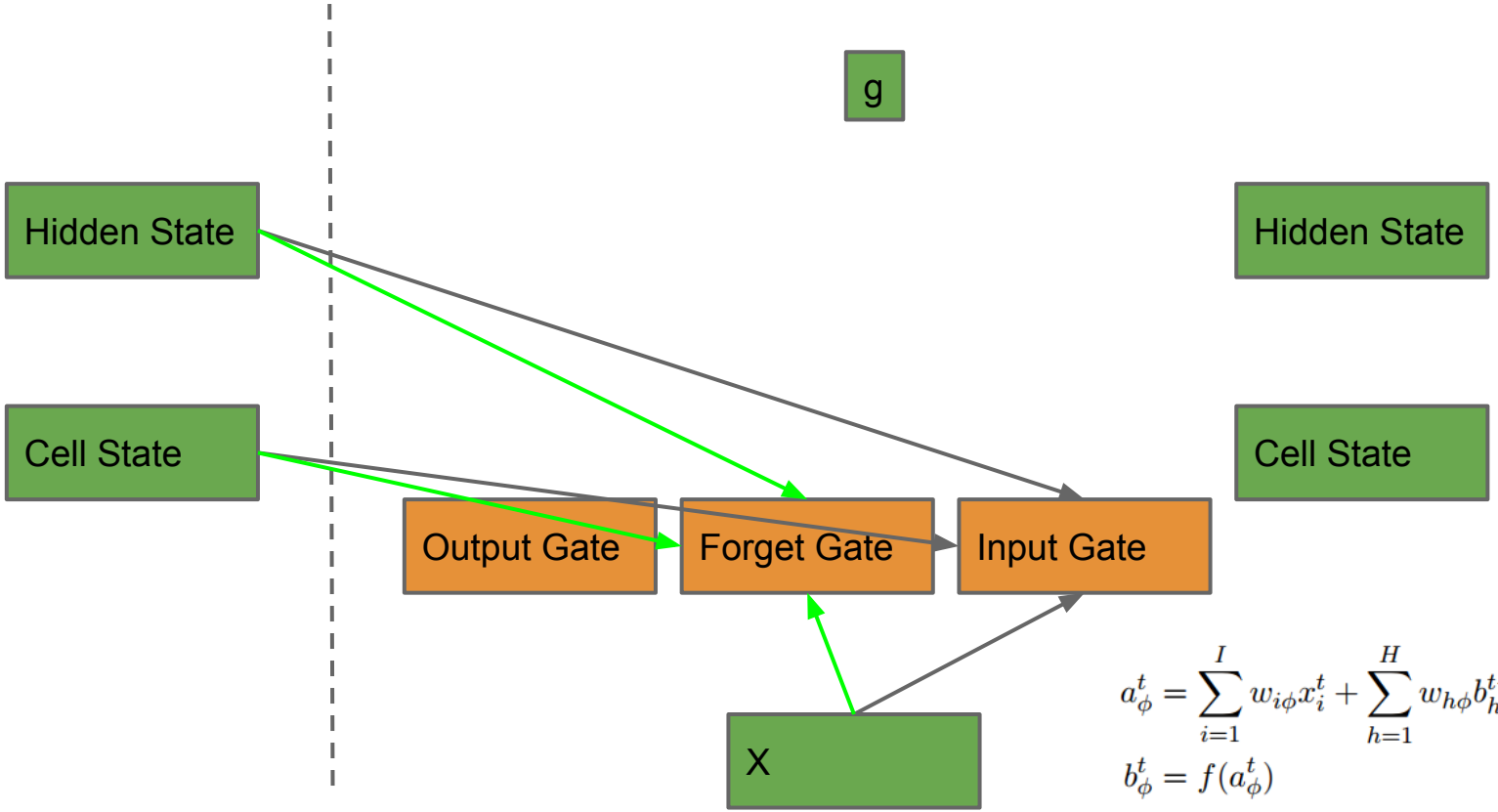
1. Input Gate



$$a_\iota^t = \sum_{i=1}^{I} w_{i\iota} x_i^t + \sum_{h=1}^{H} w_{h\iota} b_h^{t-1} + \sum_{c=1}^{C} w_{c\iota} s_c^{t-1}$$

$$b_\iota^t = f(a_\iota^t)$$

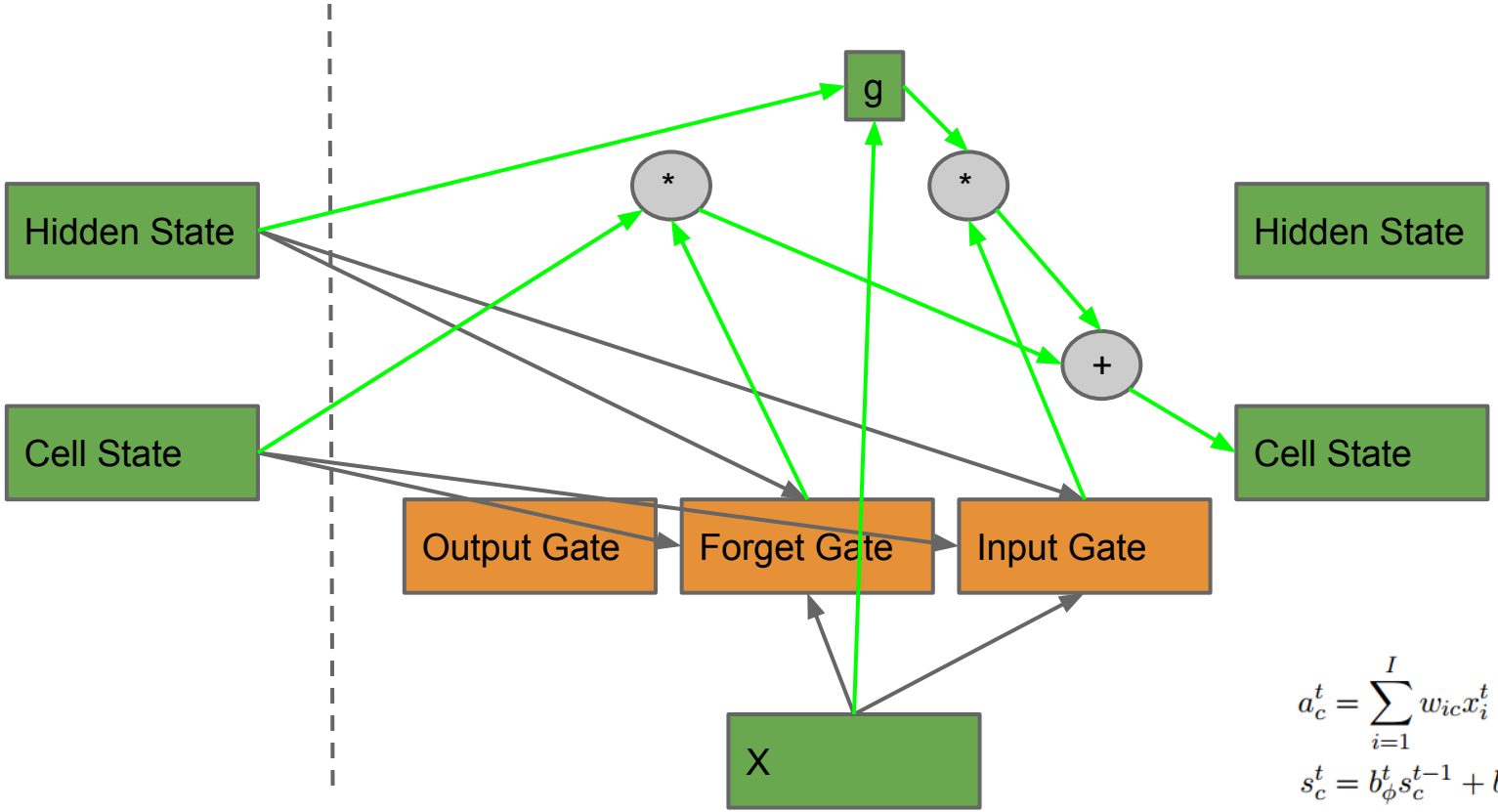
2. Forget Gate



$$a_\phi^t = \sum_{i=1}^{I} w_{i\phi} x_i^t + \sum_{h=1}^{H} w_{h\phi} b_h^{t-1} + \sum_{c=1}^{C} w_{c\phi} s_c^{t-1}$$

$$b_\phi^t = f(a_\phi^t)$$

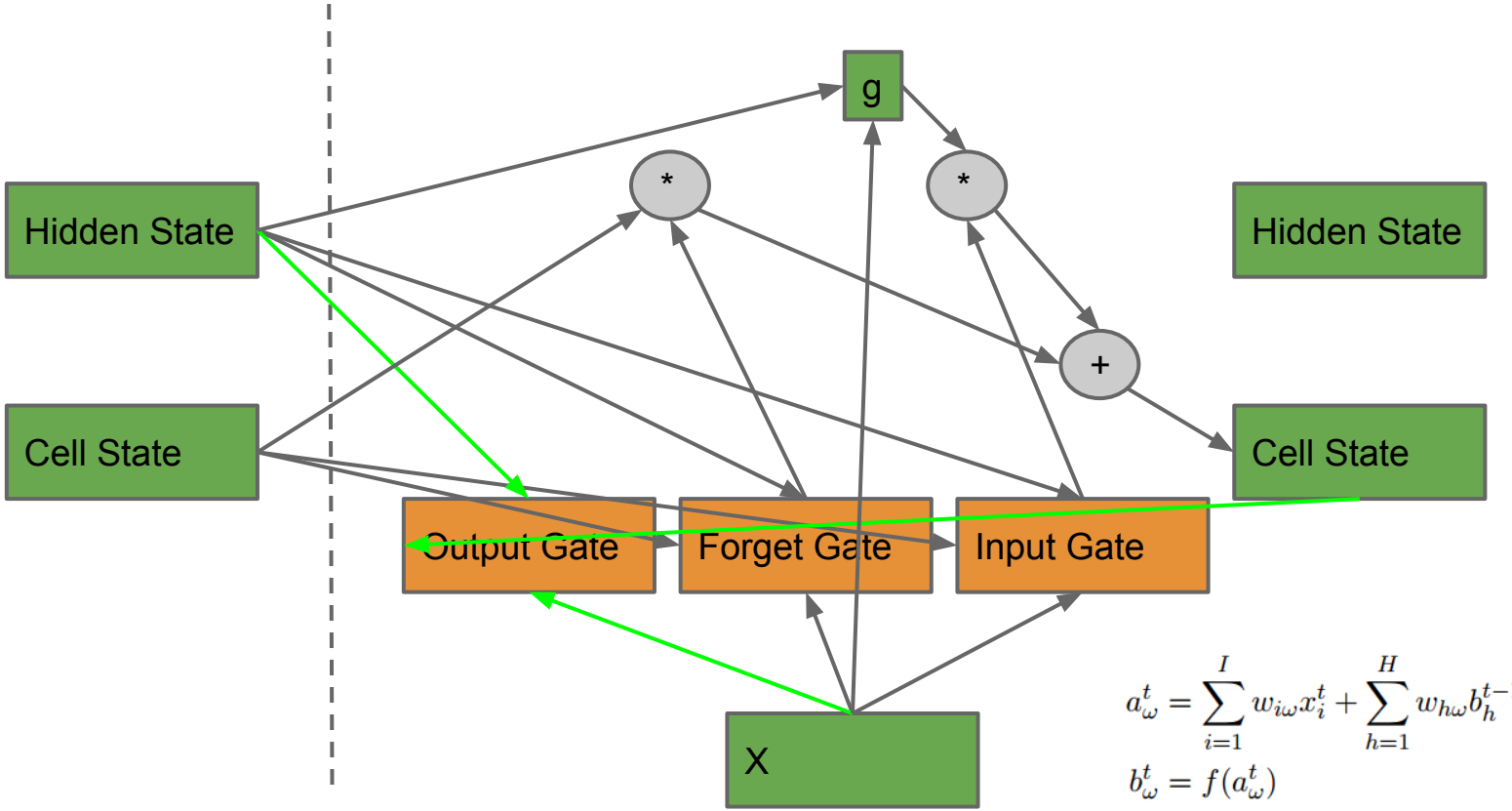
3. Cell State



$$a_c^t = \sum_{i=1}^{I} w_{ic} x_i^t + \sum_{h=1}^{H} w_{hc} b_h^{t-1}$$

$$s_c^t = b_\phi^t s_c^{t-1} + b_\iota^t g(a_c^t)$$

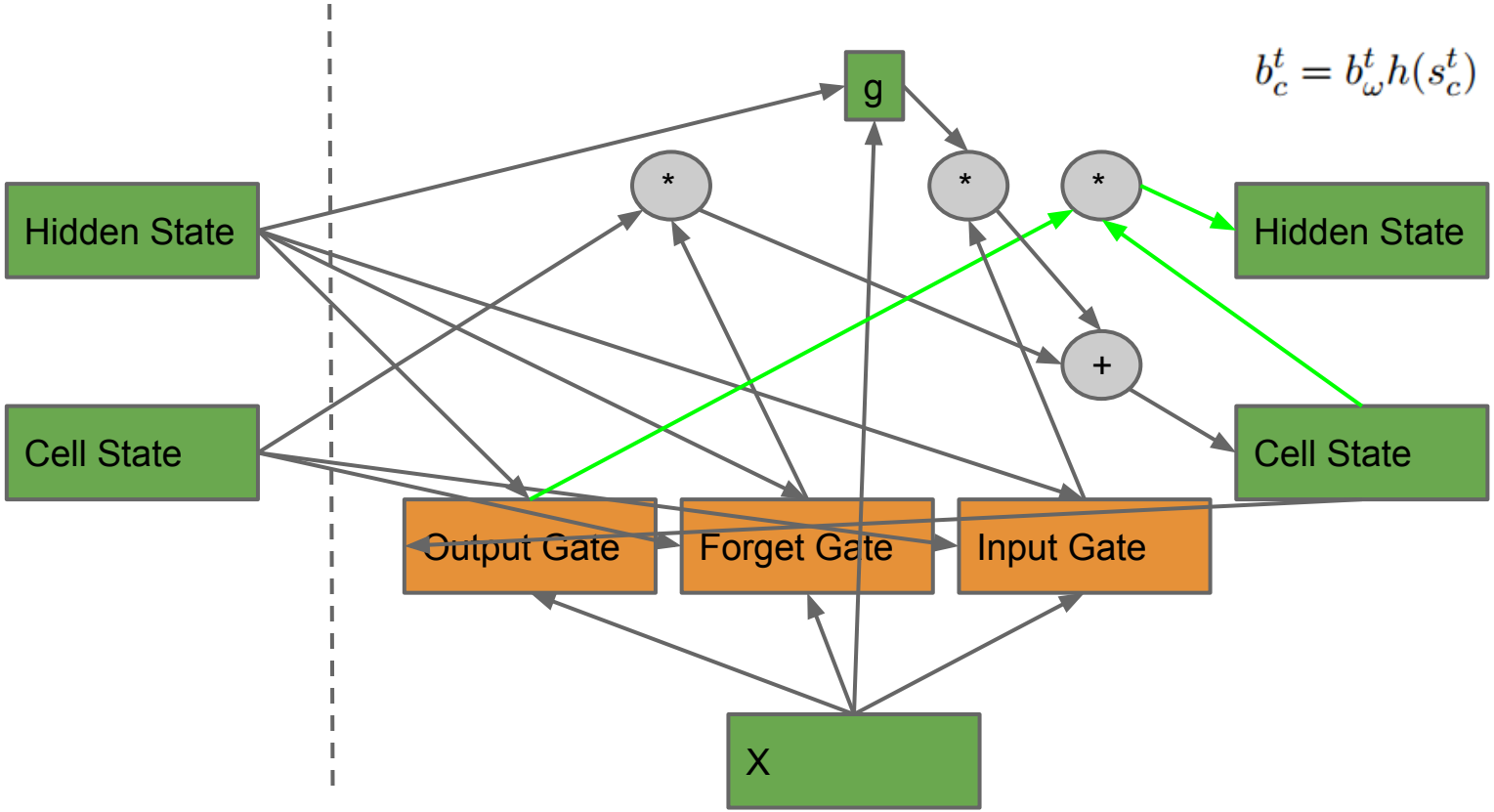
4. Output Gate



$$a_\omega^t = \sum_{i=1}^{I} w_{i\omega} x_i^t + \sum_{h=1}^{H} w_{h\omega} b_h^{t-1} + \sum_{c=1}^{C} w_{c\omega} s_c^t$$

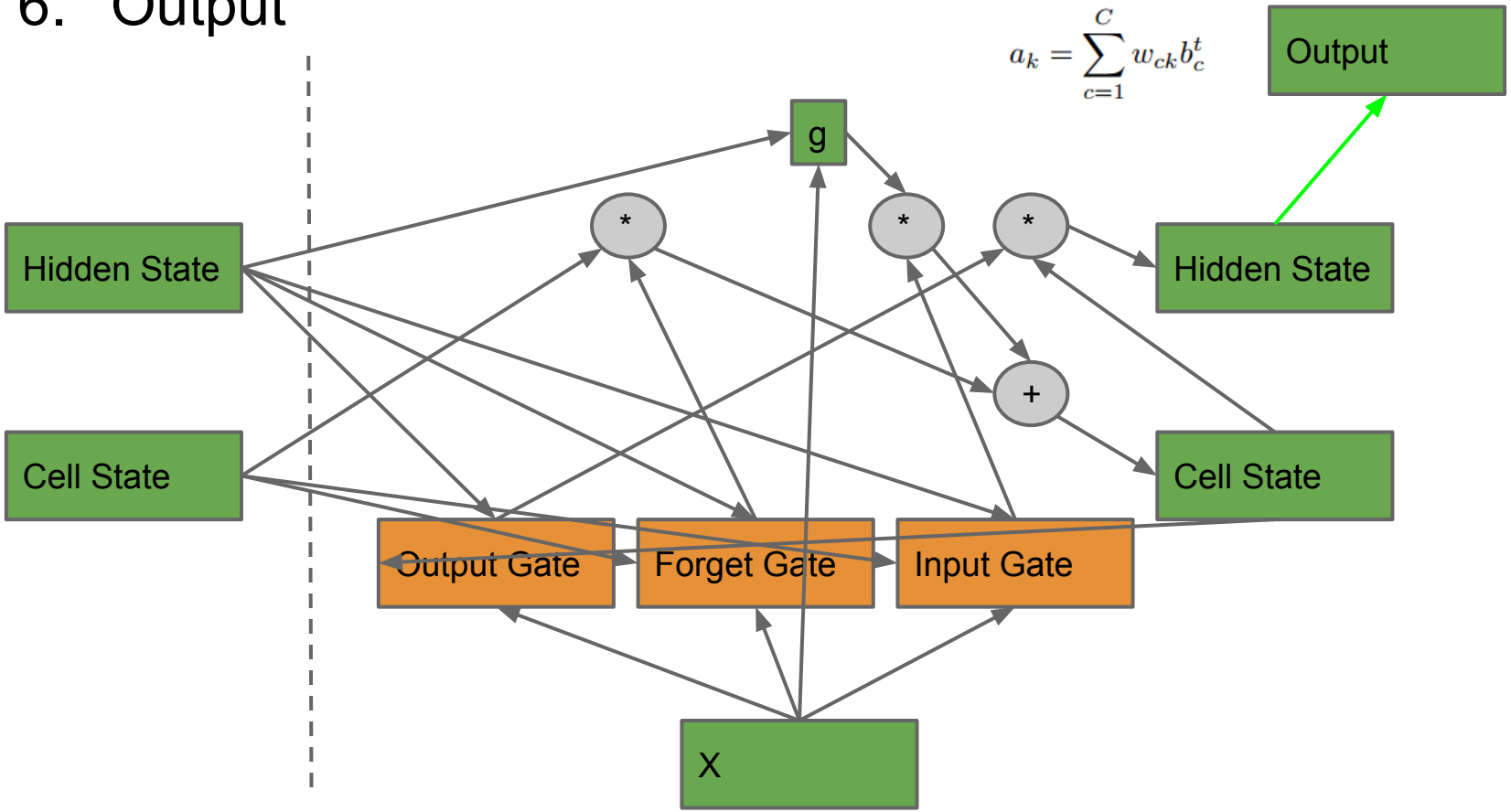
$$b_\omega^t = f(a_\omega^t)$$

5. Hidden State



6. Output

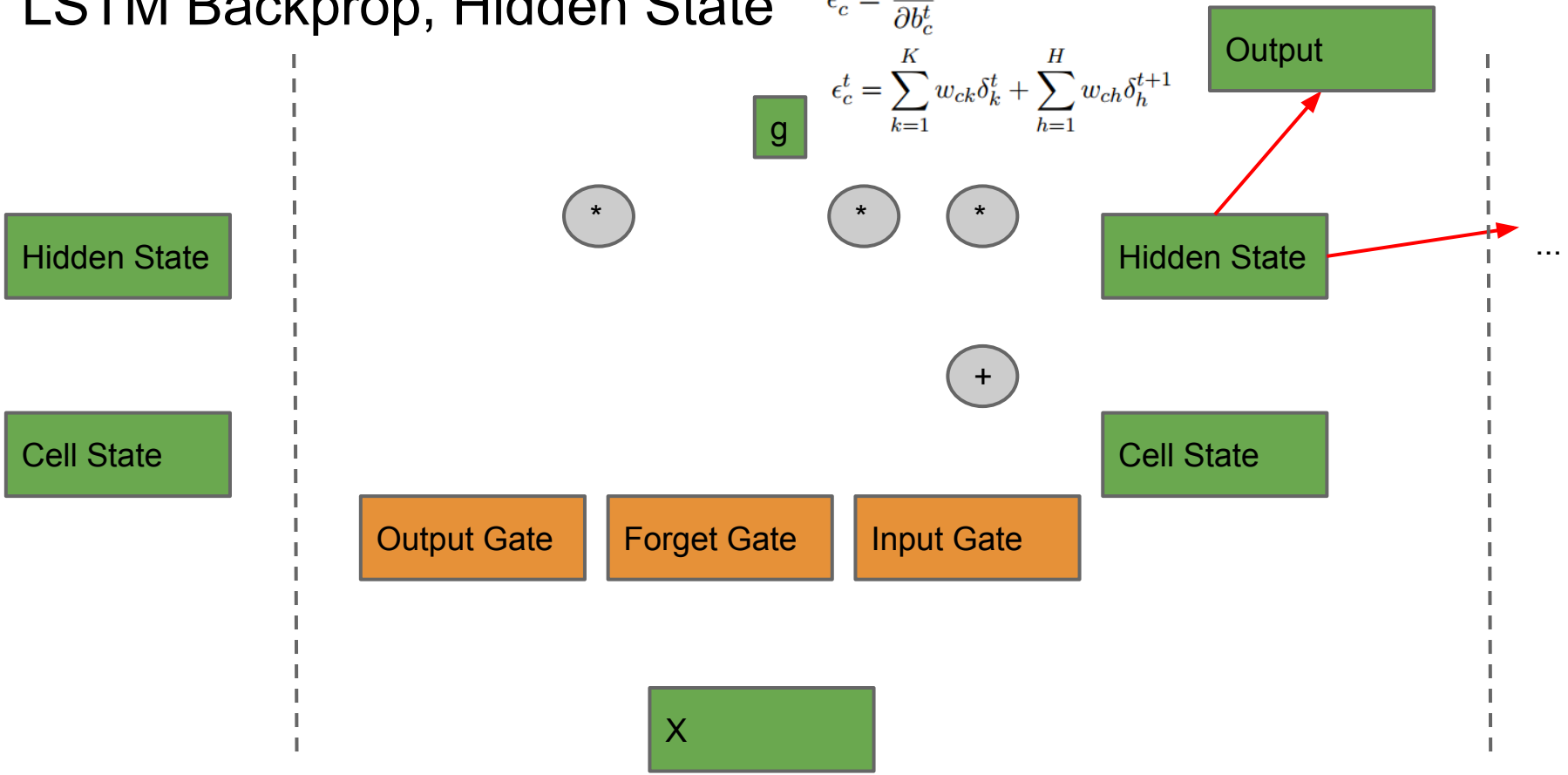
$$a_k = \sum_{c=1}^{C} w_{ck} b_c^t$$



LSTM Backprop, Hidden State

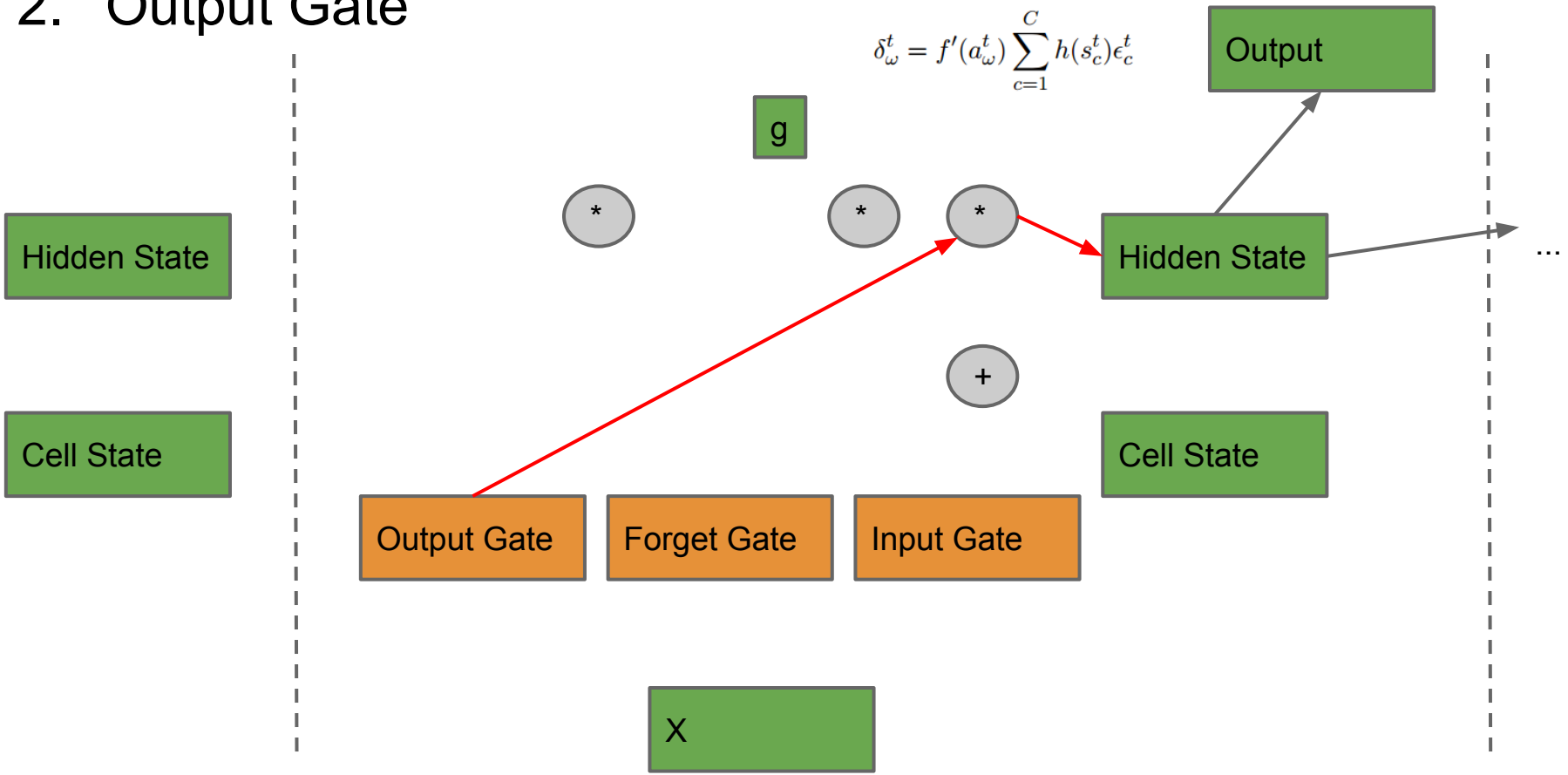
$$\epsilon_c^t \stackrel{\text{def}}{=} \frac{\partial O}{\partial b_c^t}$$

$$\epsilon_c^t = \sum_{k=1}^{K} w_{ck}\delta_k^t + \sum_{h=1}^{H} w_{ch}\delta_h^{t+1}$$

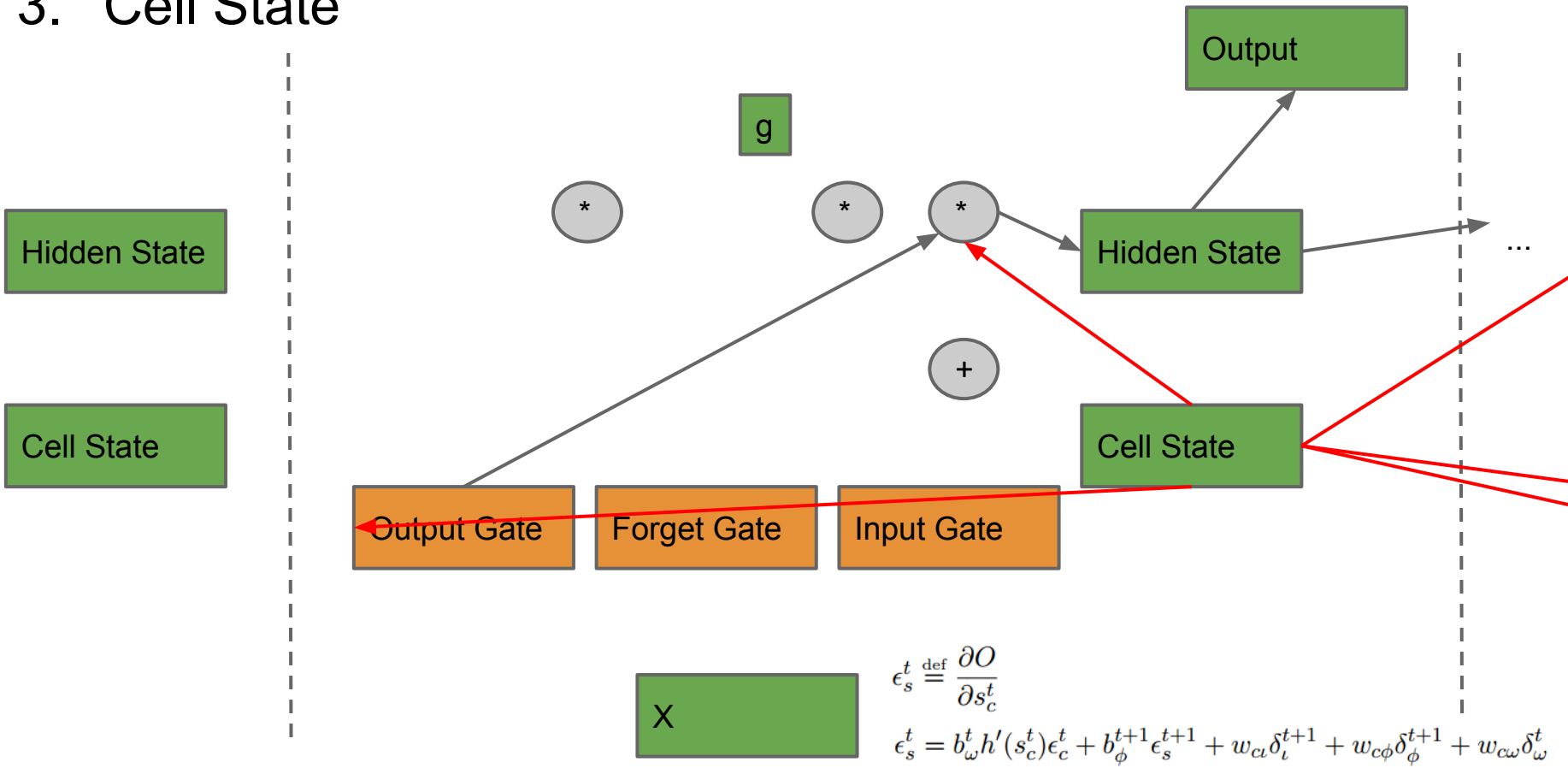


2. Output Gate

$$\delta_\omega^t = f'(a_\omega^t) \sum_{c=1}^{C} h(s_c^t)\epsilon_c^t$$



3. Cell State

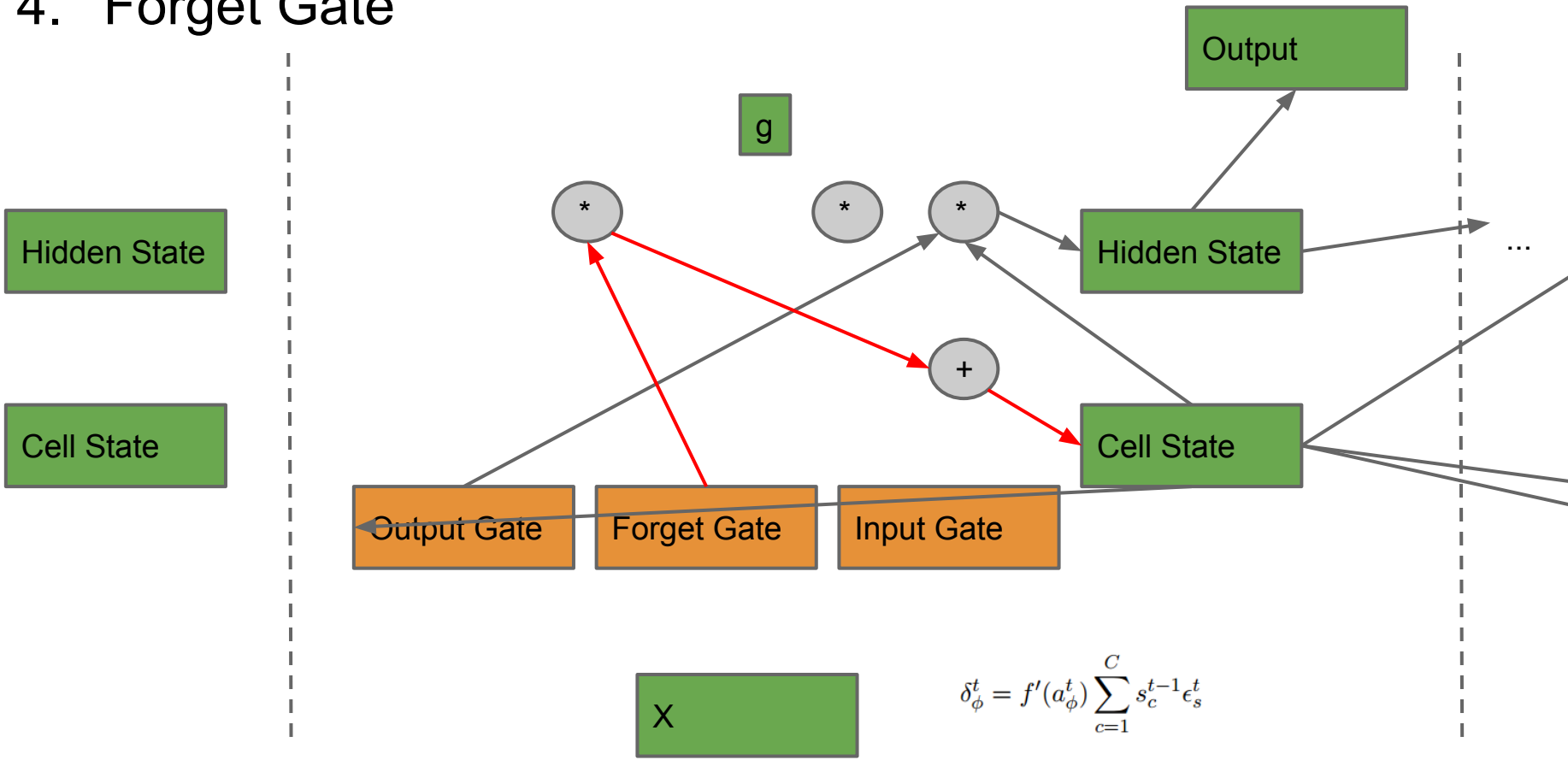


$$\epsilon_s^t \stackrel{\text{def}}{=} \frac{\partial O}{\partial s_c^t}$$

$$\epsilon_s^t = b_\omega^t h'(s_c^t)\epsilon_c^t + b_\phi^{t+1}\epsilon_s^{t+1} + w_{c\iota}\delta_\iota^{t+1} + w_{c\phi}\delta_\phi^{t+1} + w_{c\omega}\delta_\omega^t$$

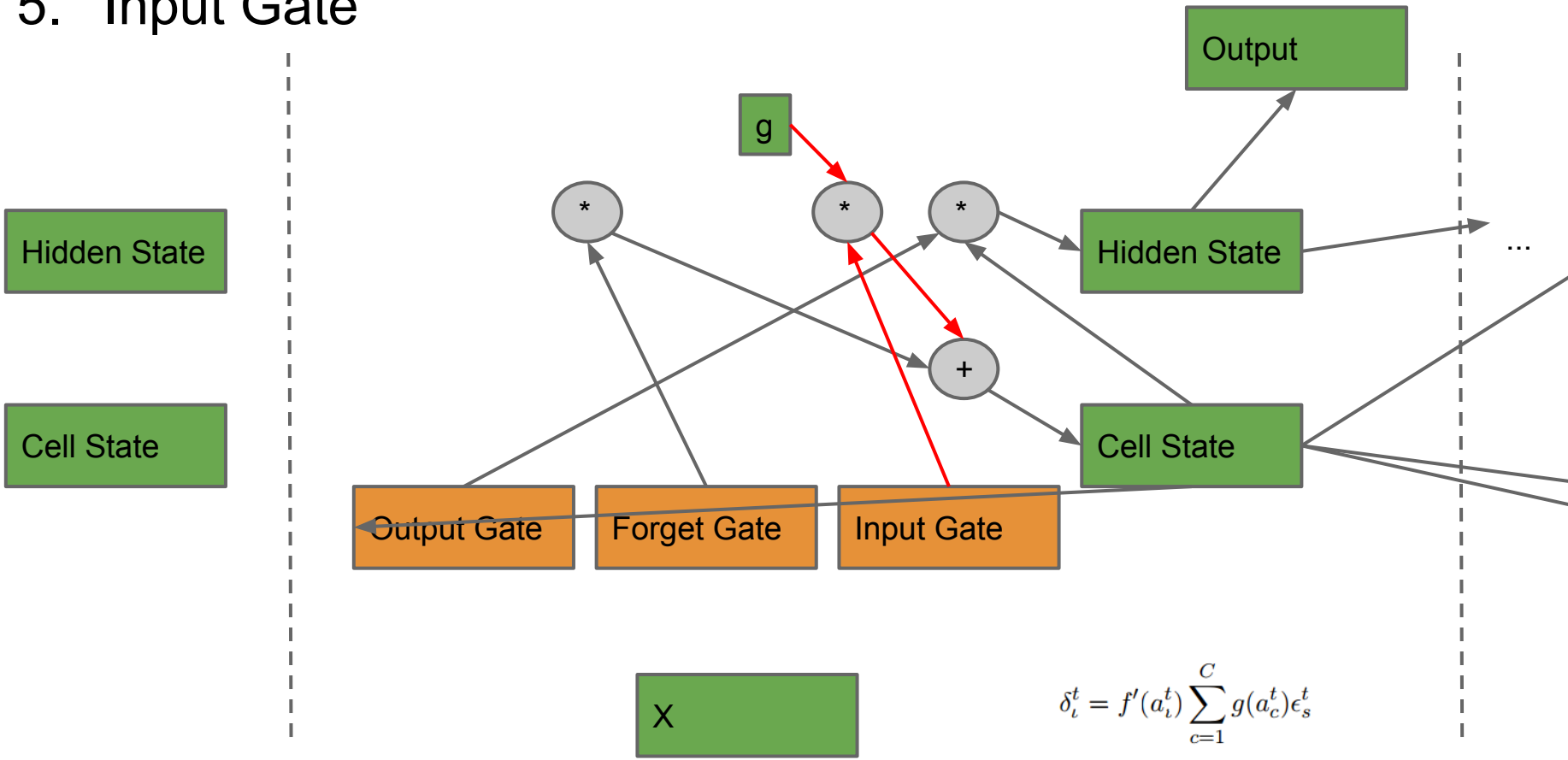
$$\delta_c^t = b_\iota^t g'(a_c^t)\epsilon_s^t$$

4. Forget Gate



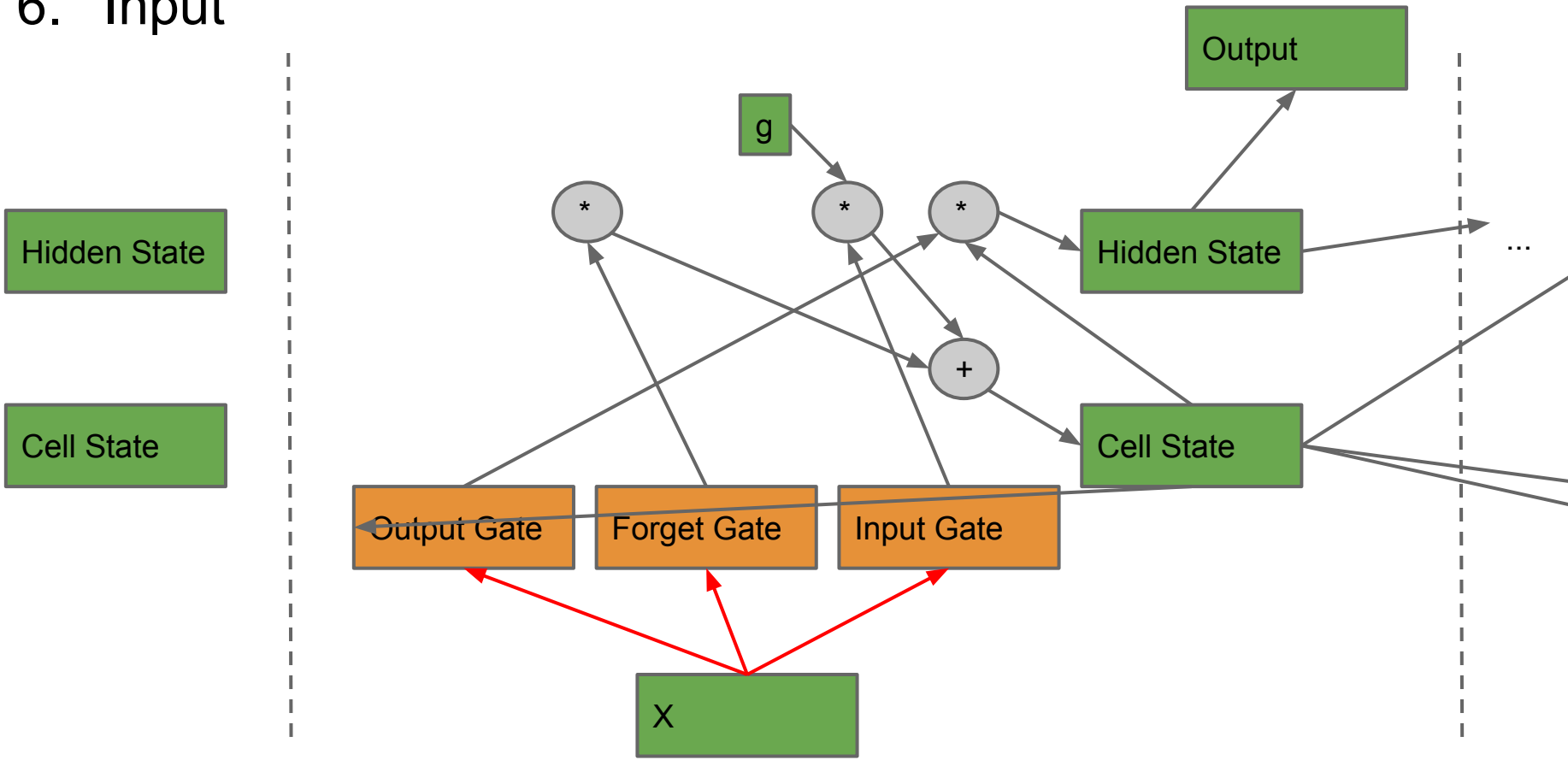
$$\delta_\phi^t = f'(a_\phi^t) \sum_{c=1}^{C} s_c^{t-1} \epsilon_s^t$$

5. Input Gate



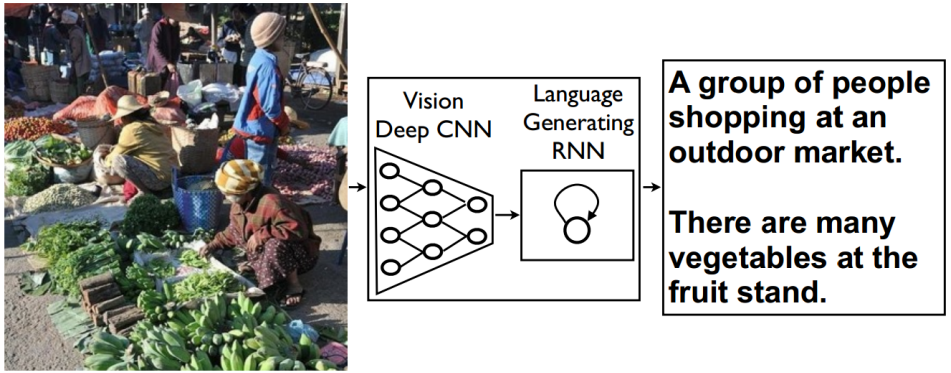
$$\delta_\iota^t = f'(a_\iota^t) \sum_{c=1}^{C} g(a_c^t)\epsilon_s^t$$

6. Input

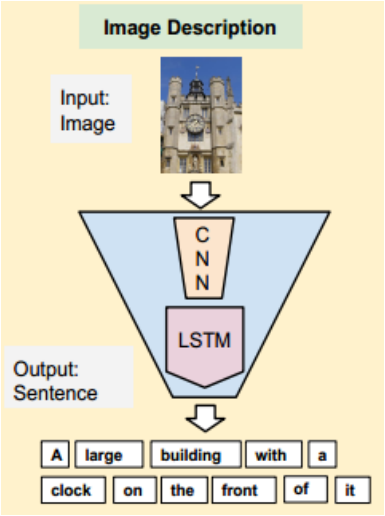

Output
g
*
*
*
Hidden State
...
Hidden State
+
Cell State
Cell State
Output Gate
Forget Gate
Input Gate
X

LSTM Applications

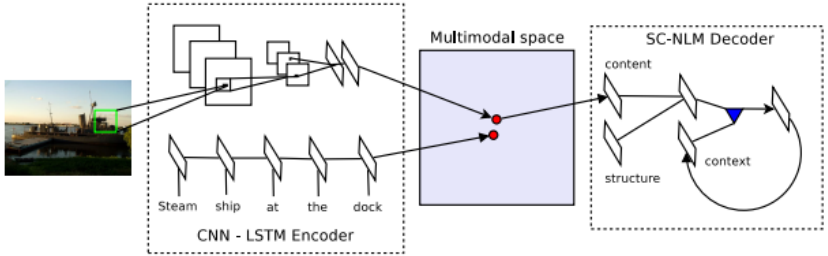
Image Captioning



Vinyals et al 2014



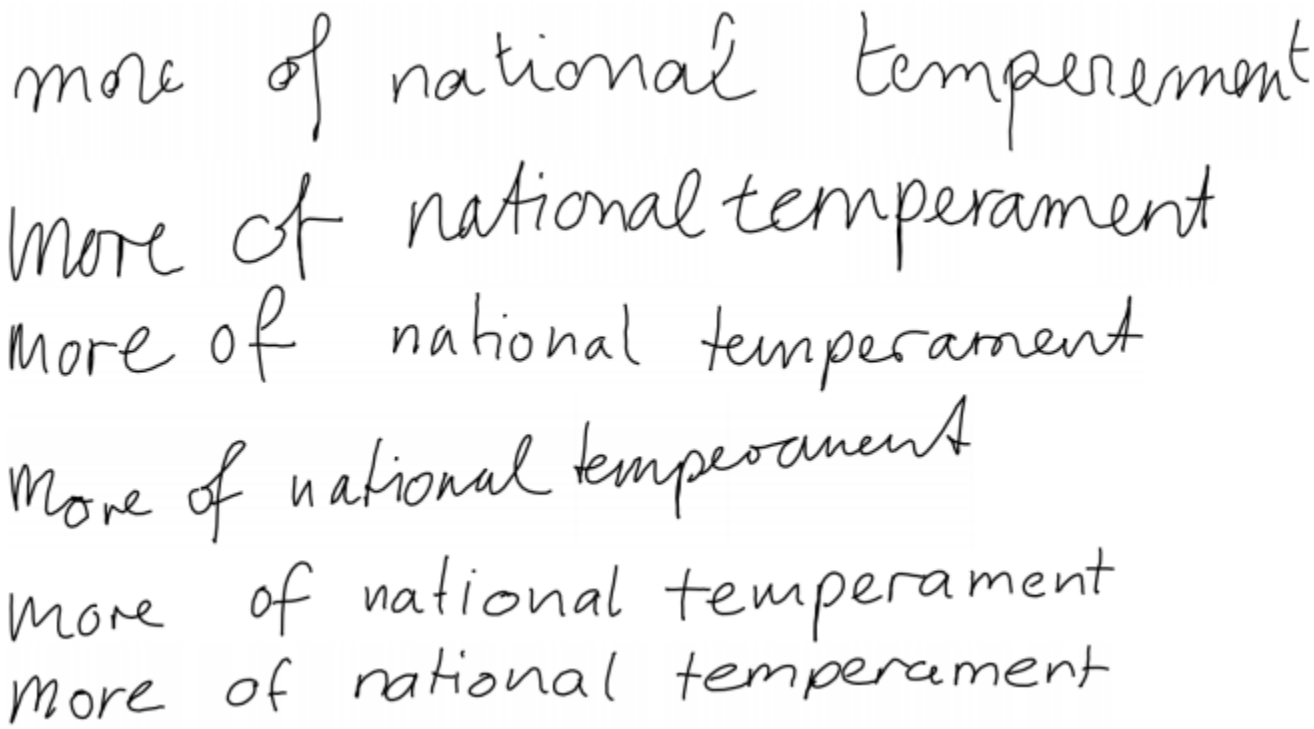
Donahue et al 2014



Kiros et al 2014

LSTM Applications

Handwriting Synthesis



Graves 2014

LSTM Applications

- Speech recognition (Graves et al 2013)
- Neural Machine Translation (Sutskever et al 2014)
- Neural Turing Machine (Graves et al 2014)