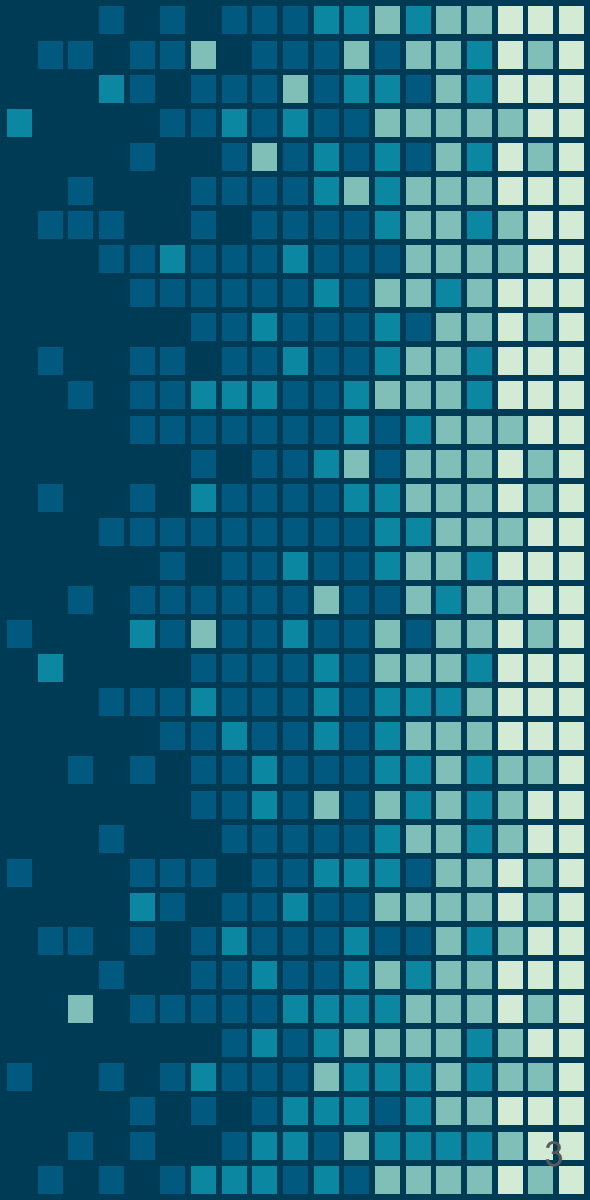

HEIDELBERG AI

We'd like to thank our sponsor

the Medical Image Computing Group @ DKFZ

What are differentiable neural computers?



Overview

- Introduction
- Recurrent neural network in general (RNN)
- Long short-term memory (LSTM)
- Differentiable neural computer (DNC)
- Applications of DNC

About Me

Jörg Franke

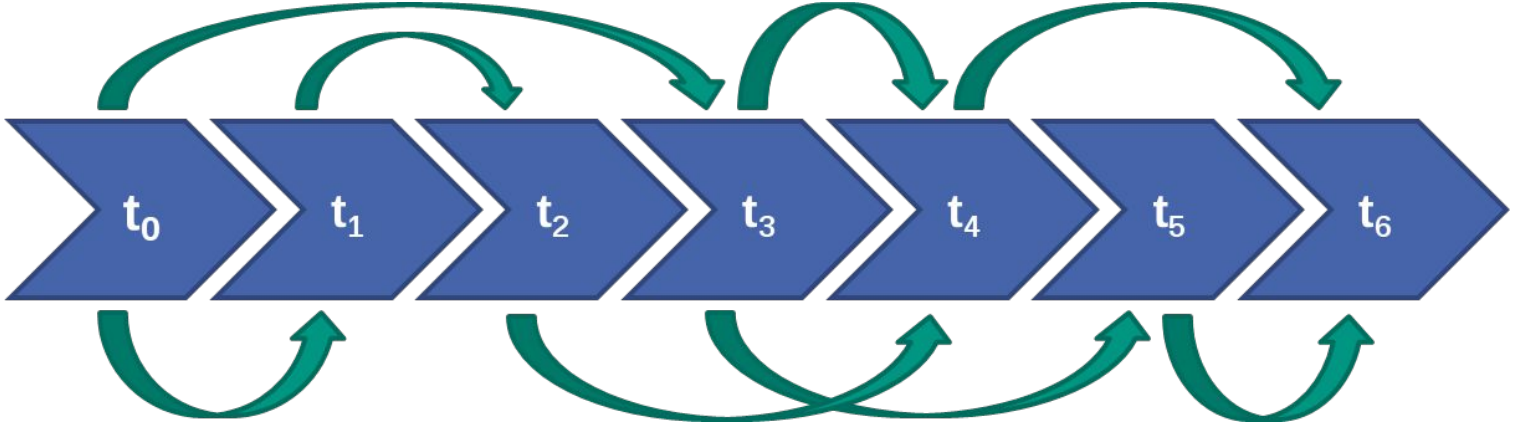
Currently:

Master Student	@ KIT (Information Engineering)
Master Thesis	@ Interactive Systems Lab
Student Research Assistant	@ RG Multilingual Speech Recognition
Deep Learning Developer	@ UnderstandAI GmbH

Introduction

Recognition in sequences is important

- Plenty real world data is interdependent
  - Speech → Automatic speech recognition
  - Languages → Machine translation
  - → Answering Questions
  - DNA → Segmentation of DNA
  - Handwriting → Handwriting recognition
  - Scenes → Behavior Recognition



Example task: bAbI 1+2

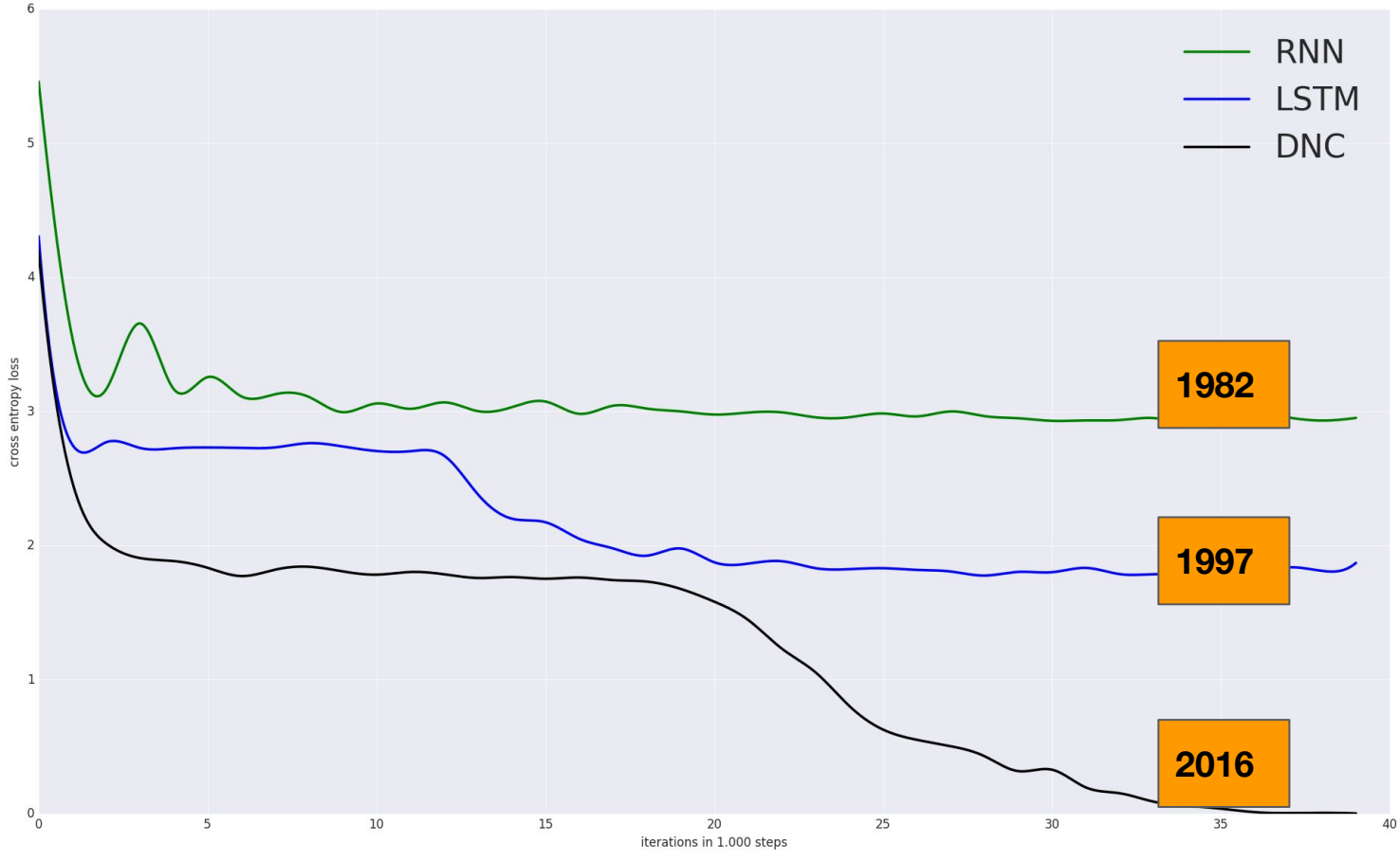
John is in the playground. John picked up the football. Where is the football? playground Daniel went to the hallway. Sandra moved to the garden. Where is Daniel? hallway John moved to the office. Sandra journeyed to the bathroom. Where is Daniel? hallway Mary moved to the hallway. Daniel travelled to the office. Where is Daniel? office

- Input: One-hot vector [112, 23, 10, _, 21, … 43, 66, 23, _]
- Output: One-hot vector [_, _, _, 22, _, … _, _, _, 78]

- Vocabulary size: 40
- 4000 Samples
- Sequence length between 29 and 486

Source: https://github.com/facebook/bAbI-tasks , 24.07.2017

Performance on bAbI task 1+2



Keys - Basic

Matrix Transfer

Vector Transfer

Scalar Transfer

Copy

Keys (all variables are vectors)

1- $y = 1 - x$ $(elementwise)$

× $y = a * b$ $(elementwise)$

*-1 $y = x * (-1)$ $(elementwise)$

+ $y = a + b$ $(elementwise)$

mat mul $y_{ik} = \sum_j a_{ij} * b_{jk}$

C $C(u, v) = \frac{u \cdot v}{|u| * |v|}$

sum $y = \sum_i x_i$

soft max $SM(x) = \frac{\exp^x}{\sum_i \exp^{x_i}}$

Keys - Neural Network Layers

/

$$linear(x|\theta) = \sum \mathbf{x} * \theta$$

σ

$$sigmiod(x|\theta) = \frac{1}{1 + \exp^{-\sum \mathbf{x} * \theta}}$$

soft max

$$softmax(x|\theta) = \frac{\exp^{\sum \mathbf{x} * \theta}}{\sum \exp^{\sum \mathbf{x} * \theta}}$$

RNN

Recurrent **N**eural **N**etwork

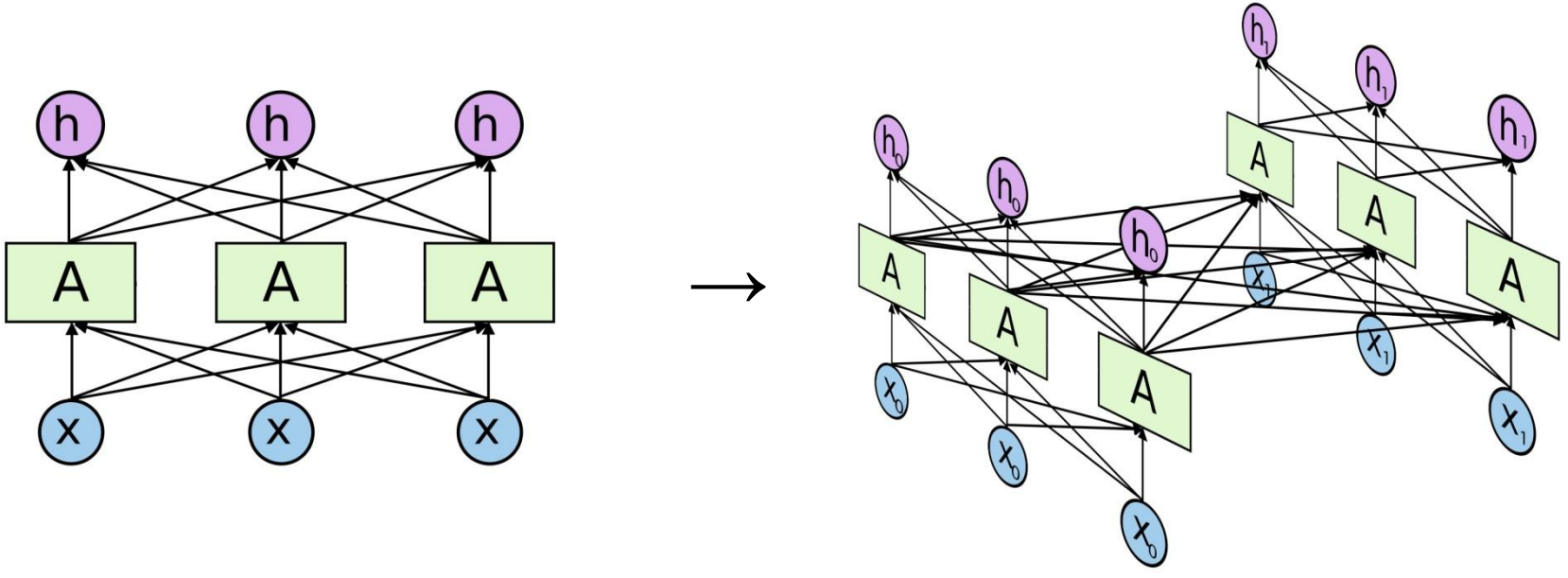
RNN - Basic Idea

- Current output is a part of the next input.
- Main purpose: Process arbitrary sequences of inputs.

- Introduced by John Hopfield in 1982 [1], refined by Jeffrey Elman [2] and Michael Jordan (not the basketballer Michael Jordan).

Modelling sequences with RNN

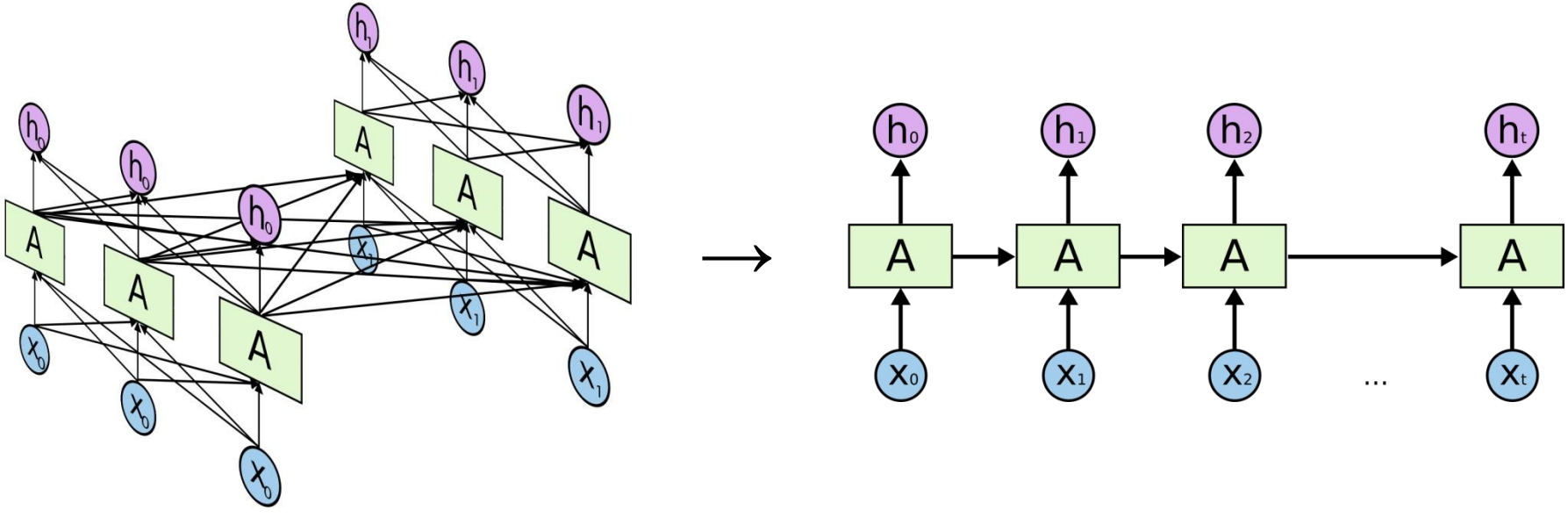
- Neural Network with time links



h	*Output*
x	*Input*

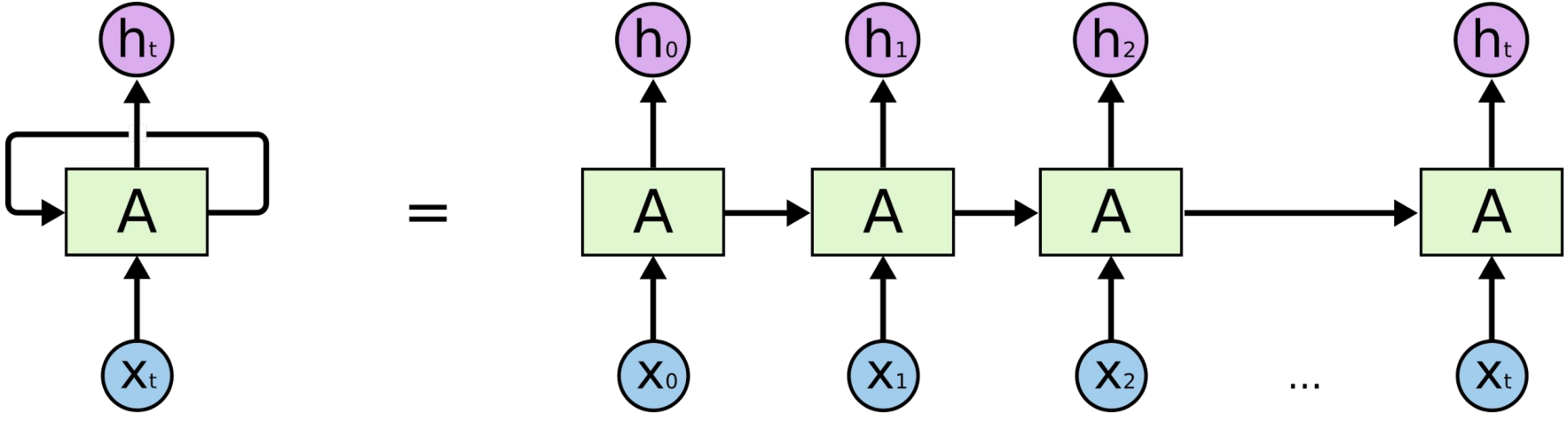
Modelling sequences with RNN

- Simpler view without layers



Modelling sequences with RNN

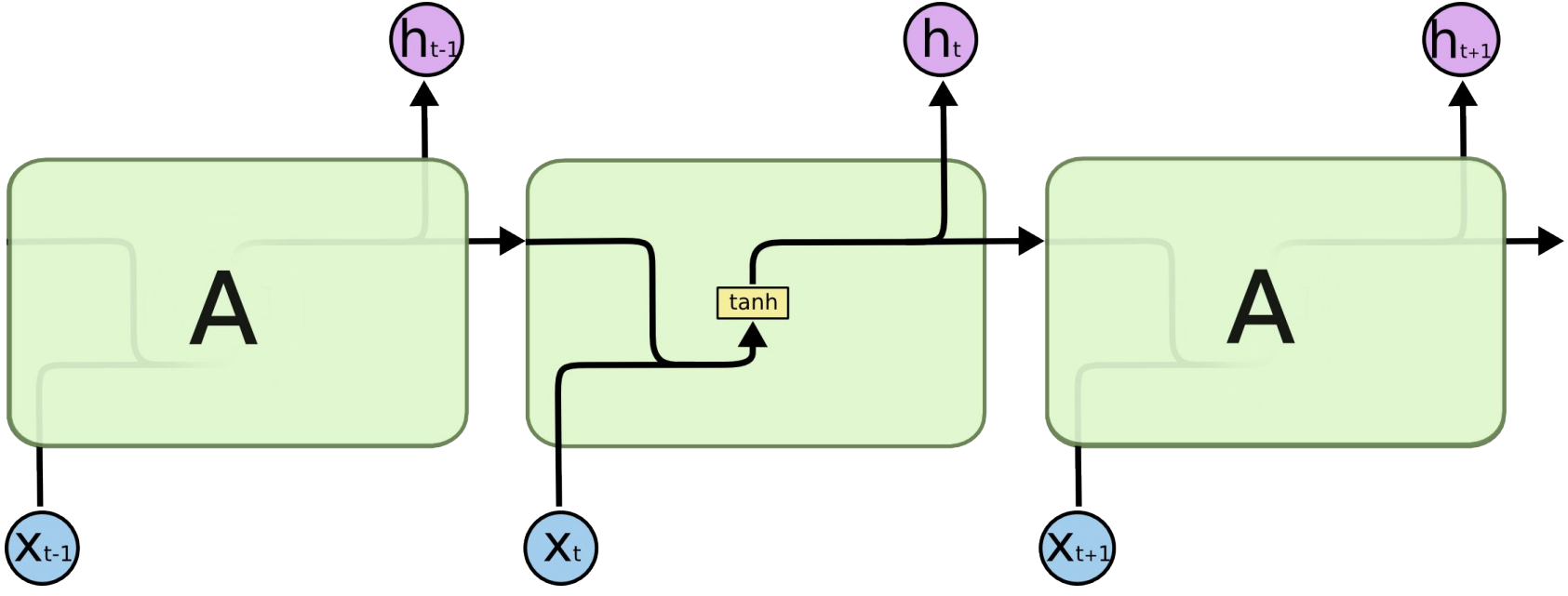
- Unfolded RNN in time, **weights in all time steps the same**



Source: http://colah.github.io/posts/2015-08-Understanding-LSTMs/, 24.07.2017

Modelling sequences with RNN

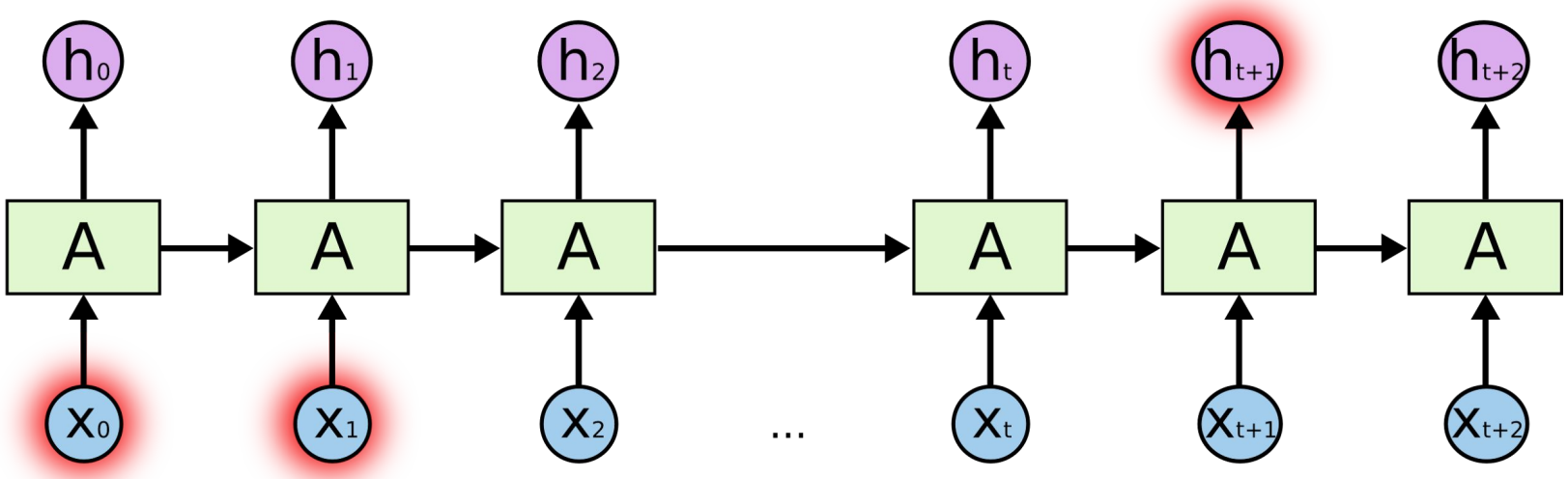
- Use tanh as activation function



Source: http://colah.github.io/posts/2015-08-Understanding-LSTMs/, 24.07.2017

Modelling sequences with RNN

- Weak for long dependencies (vanishing gradient)



Source: http://colah.github.io/posts/2015-08-Understanding-LSTMs/, 24.07.2017

LSTM

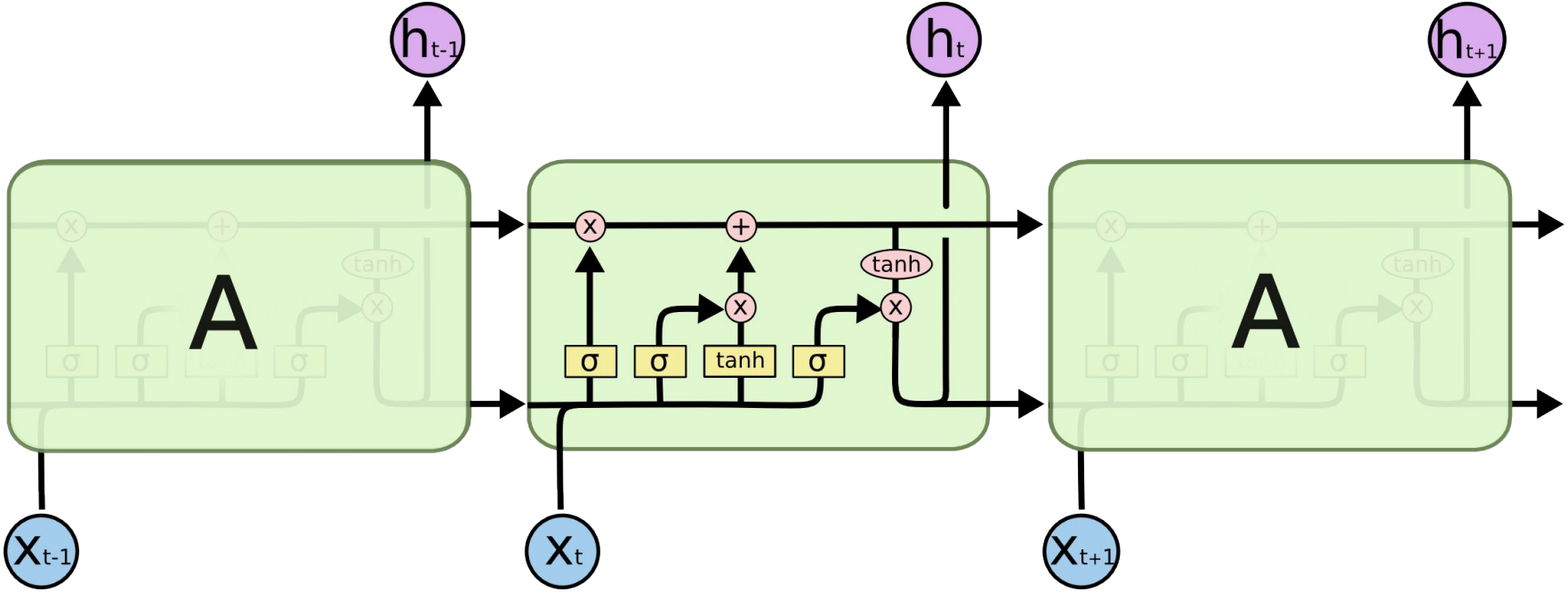
Long **S**hort-**T**erm **M**emory

LSTM – Basic Idea

- Each network node stores an internal cell state (CS).
- It can forget information and add input information from one time step to the next.
- Uses gates to determine influence to and impact out of cell state.
- Introduced to get rid of the vanishing gradient problem

- State of the Art technique in recurrent neural network setting.
- Introduced by Hochreiter & Schmidhuber in 1997 [3].

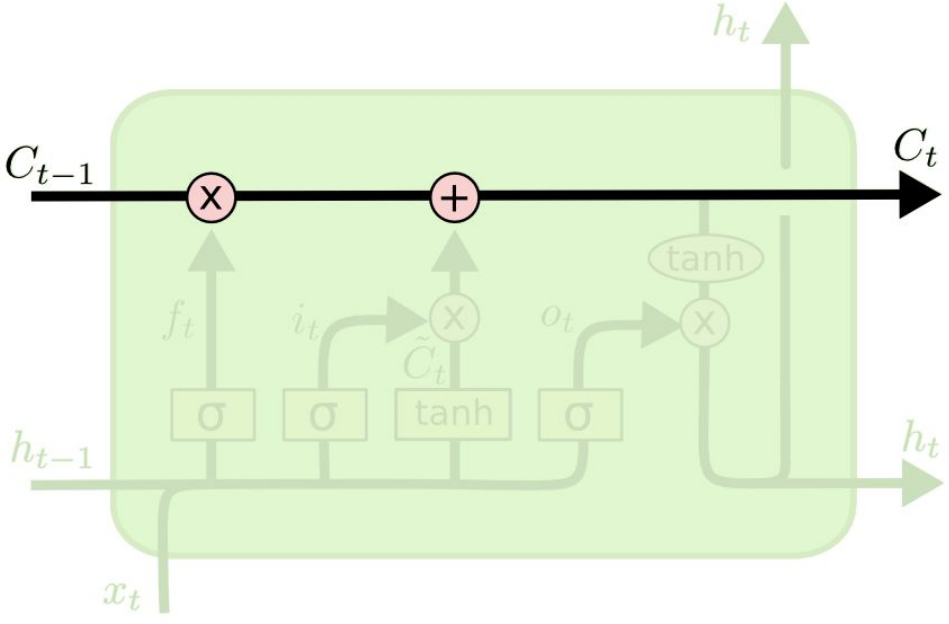
LSTM – Sophisticated activation function



h	*Output*
x	*Input*

Source: http://colah.github.io/posts/2015-08-Understanding-LSTMs/, 24.07.2017

LSTM – The internal cell state

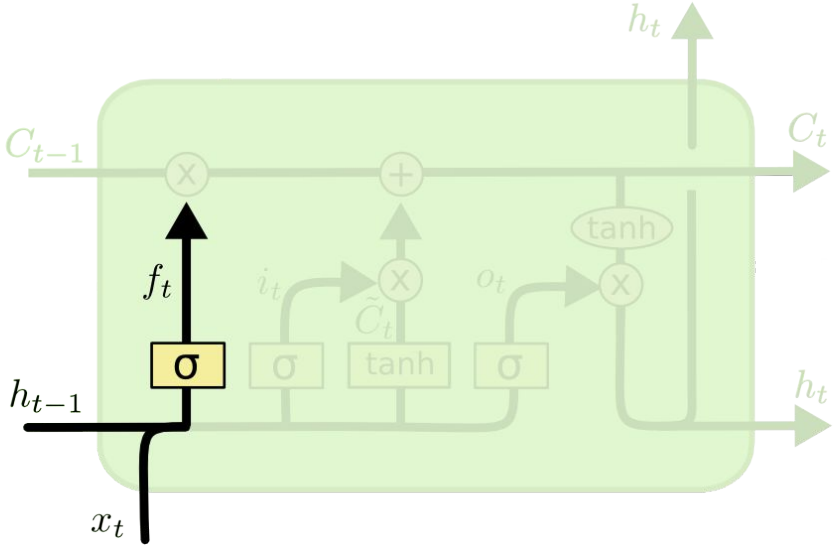


C | *Cell state*

Source: http://colah.github.io/posts/2015-08-Understanding-LSTMs/, 24.07.2017

LSTM – Forget Gate

$$f_t = \sigma\left(W_f * [h_{t-1}, x_t] + b_f\right)$$



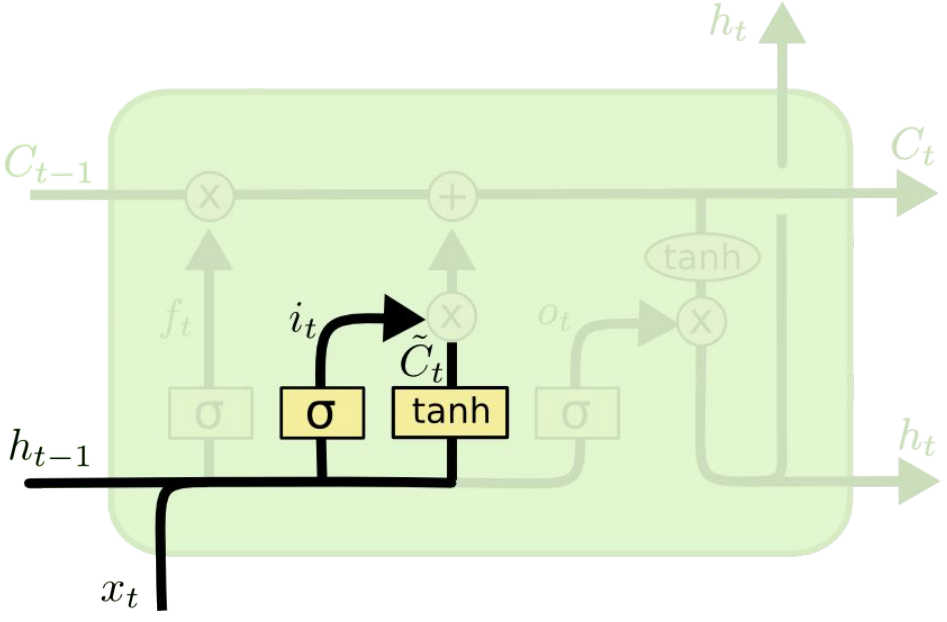
h	*Output*
x	*Input*
f	*Forget Gate*
W	*Weight*
b	*Bias*

Source: http://colah.github.io/posts/2015-08-Understanding-LSTMs/, 24.07.2017

LSTM – Input activation and input gate

$$i_t = \sigma(W_i * [h_{t-1}, x_t] + b_i)$$
$$\tilde{C}_t = tanh(W_C * [h_{t-1}, x_t] + b_C)$$

C	*Cell state*
h	*Output*
x	*Input*
i	*Input Gate*
W	*Weight*
b	*Bias*

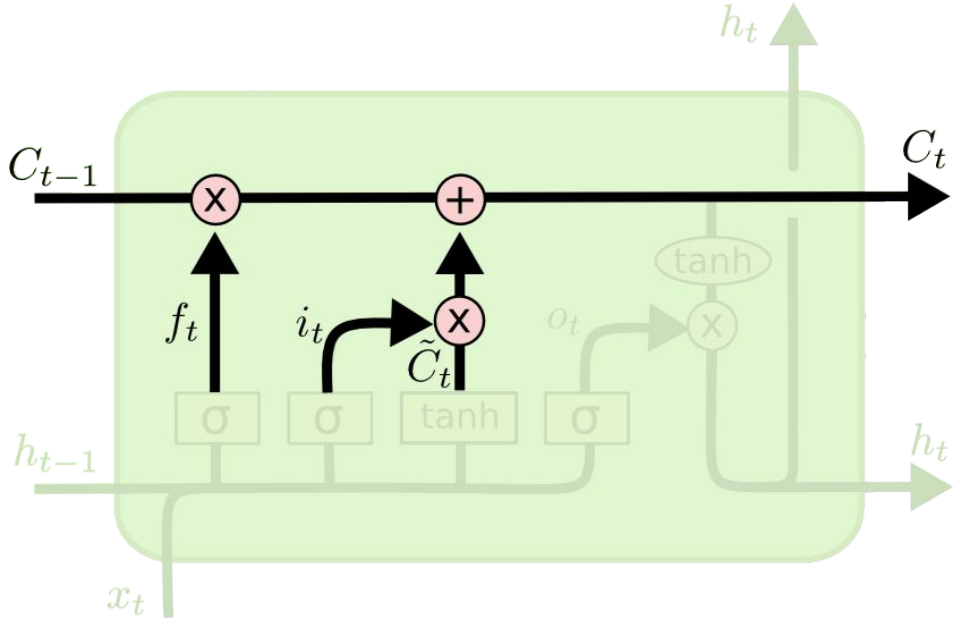


Source: http://colah.github.io/posts/2015-08-Understanding-LSTMs/, 24.07.2017

LSTM – Updating the cell state

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

C	*Cell state*
f	*Forget Gate*
i	*Input Gate*

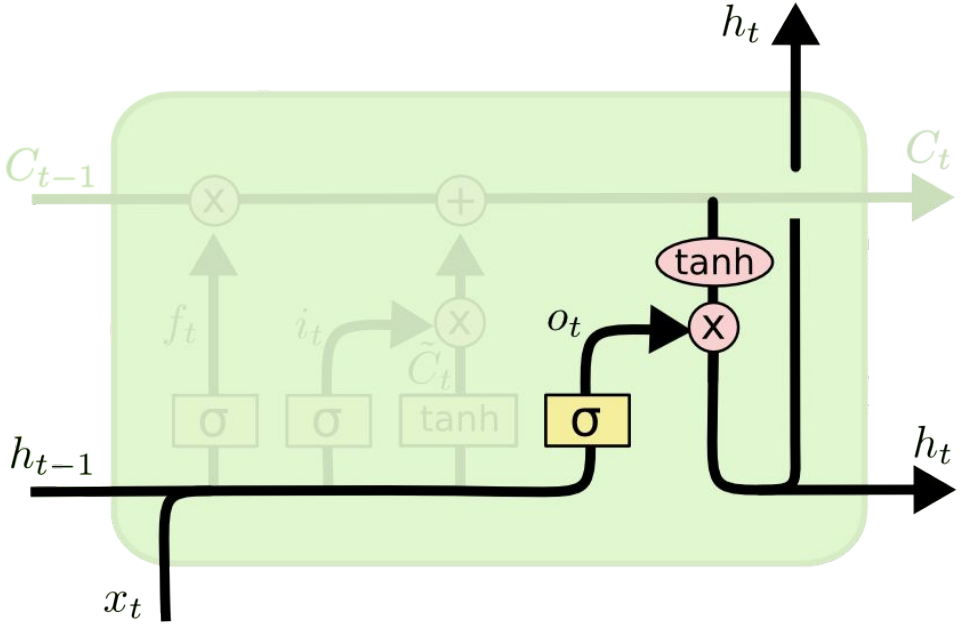


Source: http://colah.github.io/posts/2015-08-Understanding-LSTMs/, 24.07.2017

LSTM – Output activation and output gate

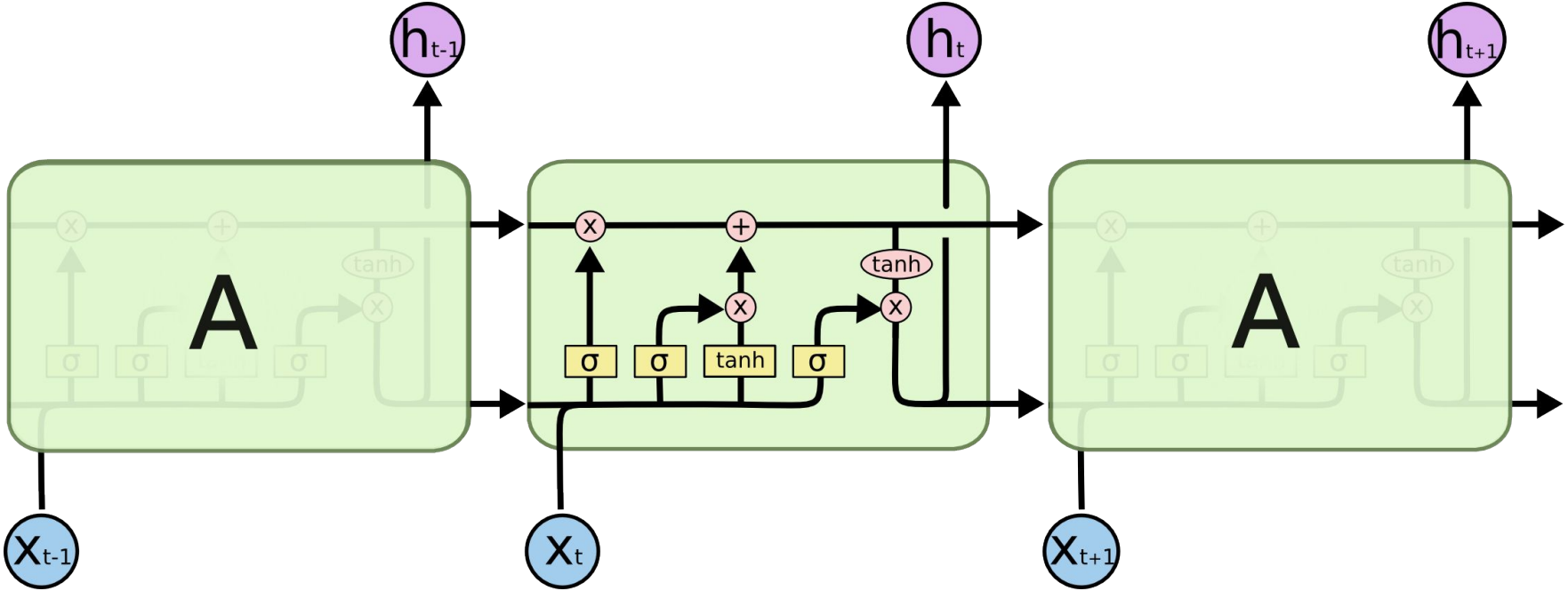
$$f_o = \sigma(W_o * [h_{t-1}, x_t] + b_o)$$
$$h_t = f_o * \tanh(C_t)$$

h	*Output*
x	*Input*
o	*Output Gate*
W	*Weight*
b	*Bias*



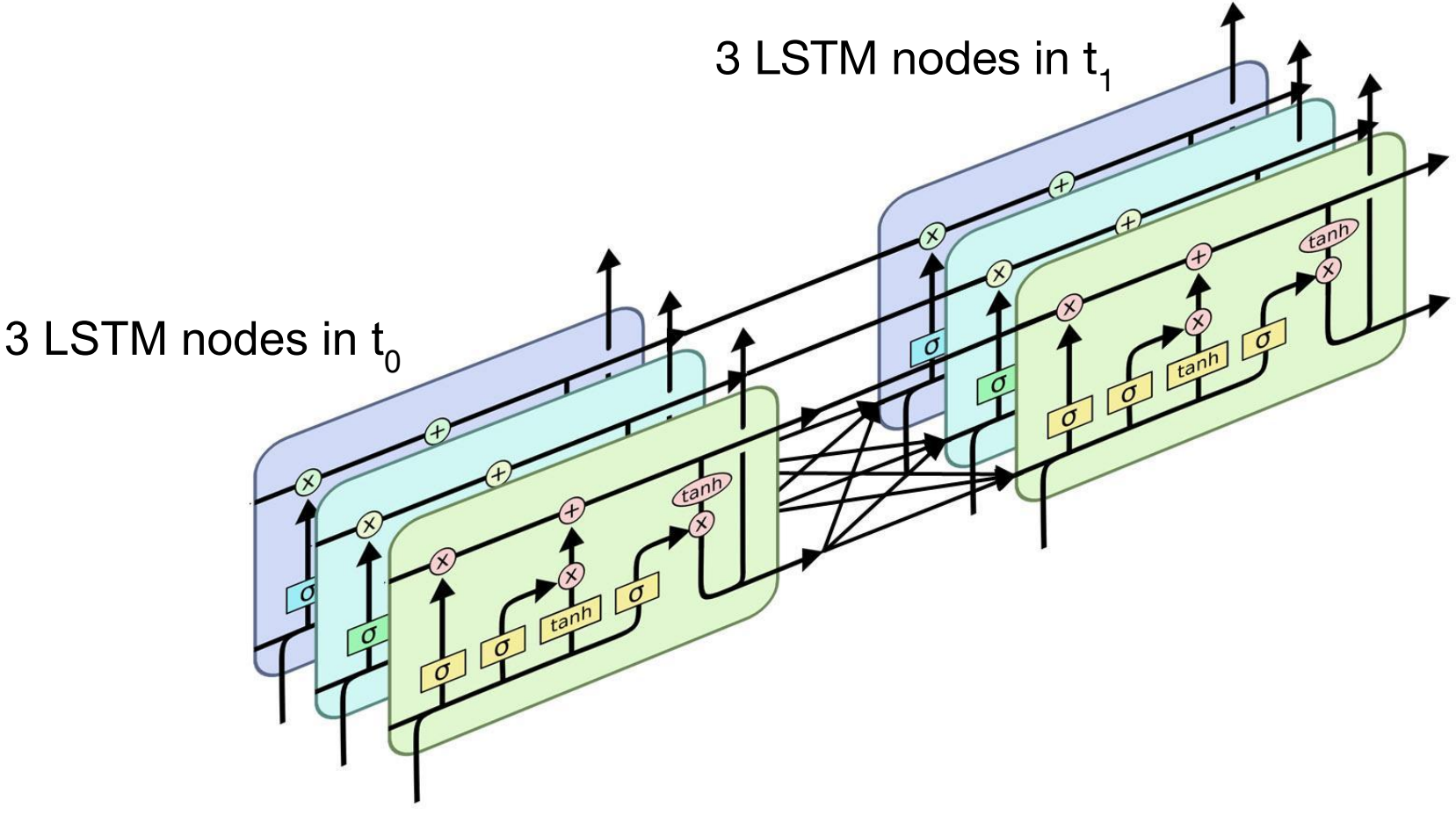
Source: http://colah.github.io/posts/2015-08-Understanding-LSTMs/, 24.07.2017

LSTM - Overview



Source: http://colah.github.io/posts/2015-08-Understanding-LSTMs/, 24.07.2017

LSTM – Multiple hidden nodes



DNC

Differential **N**eural **C**omputer

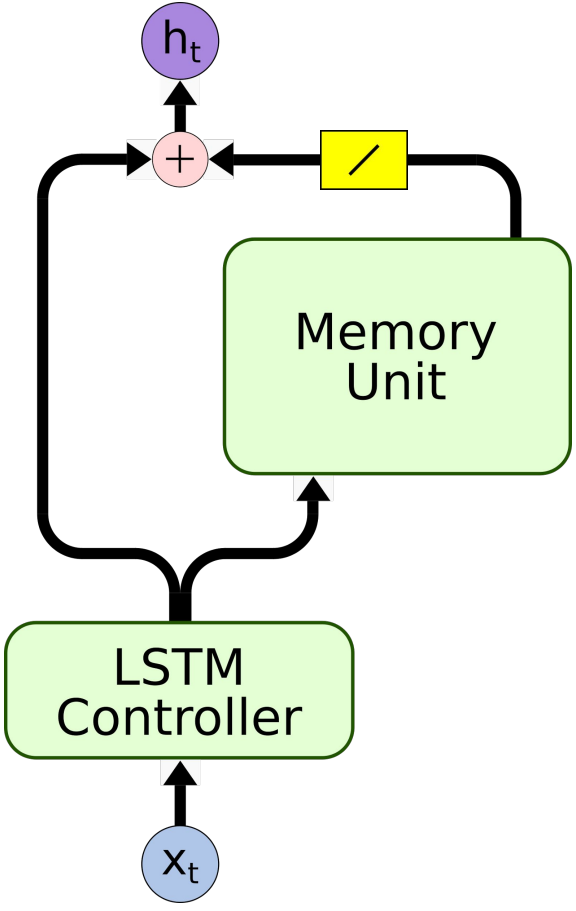
Differential Neural Computer – Basic Idea

- Memory-augmented Neural Network (MANN).

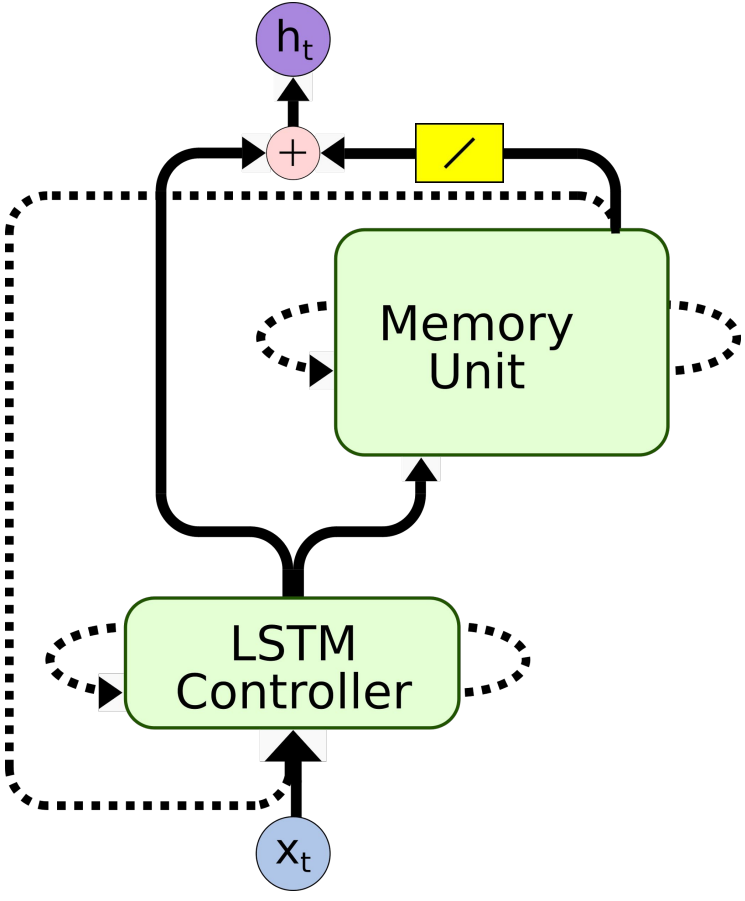
- A controller learns to use an internal memory.
- Content can be written to the memory as well as erased.
- A linkage matrix stores information about the order of writes.
- Reading from the memory either by similar content or via order from linkage matrix.
- Differentiable end-to-end.

- Main purpose: Dealing with huge long term relationships.
- Successor of Neural Turing Machine [4]
- Introduced by Graves in 2016 [5]

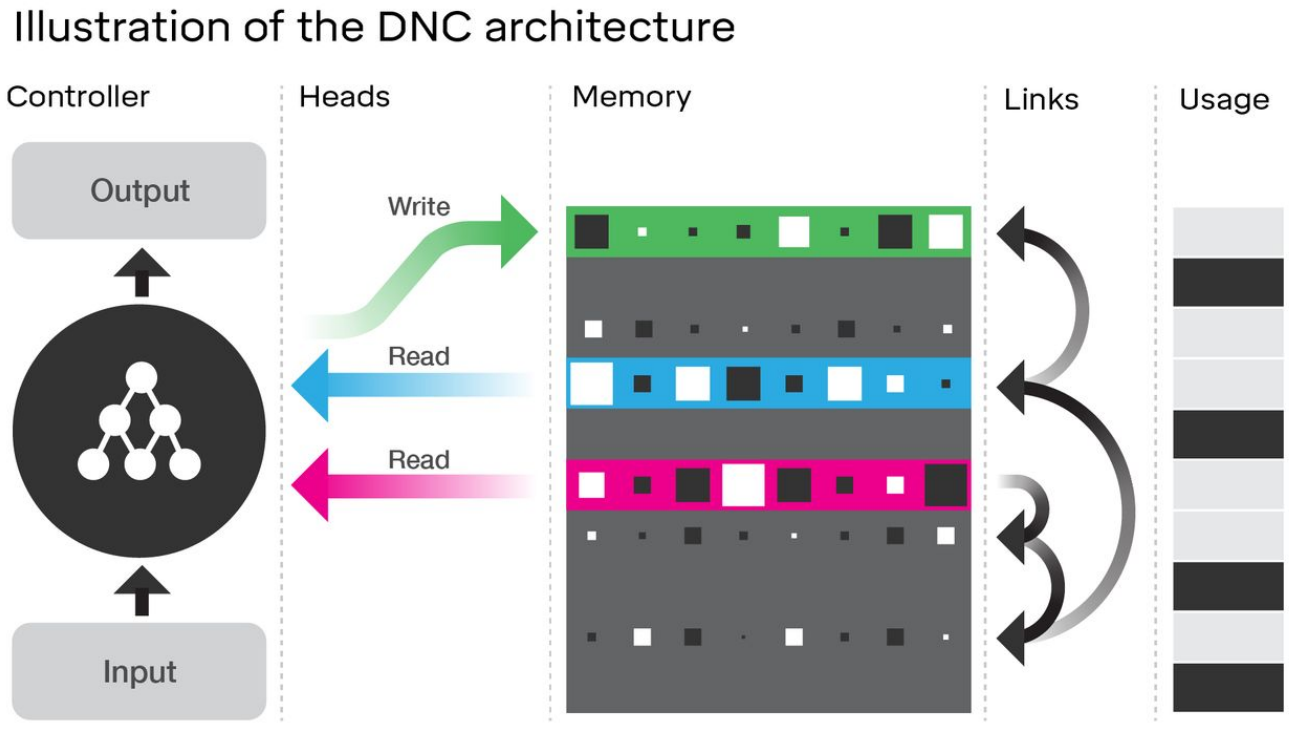
DNC - Architecture



DNC - Architecture (recurrent connections)

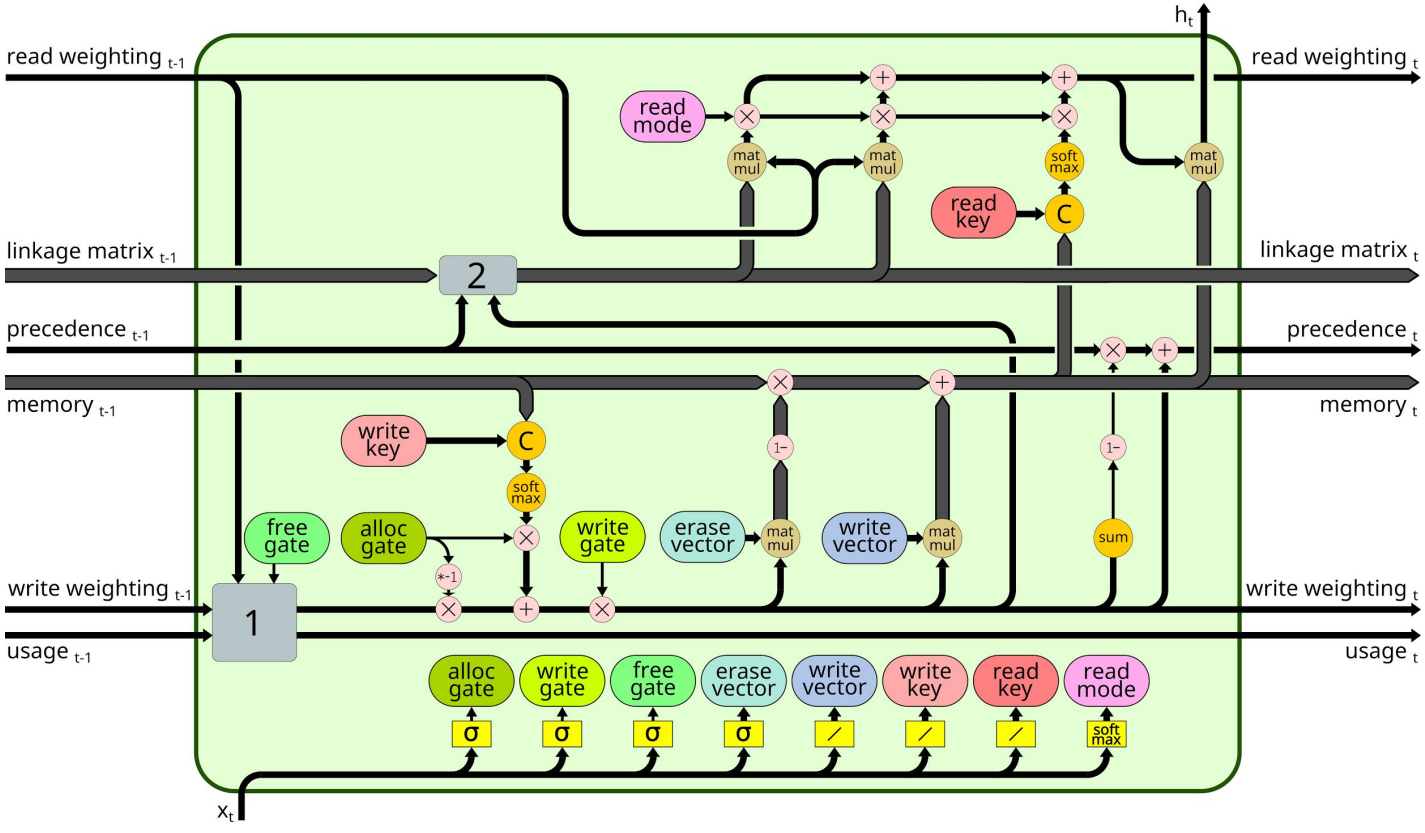


DNC - Figure from Nature Paper



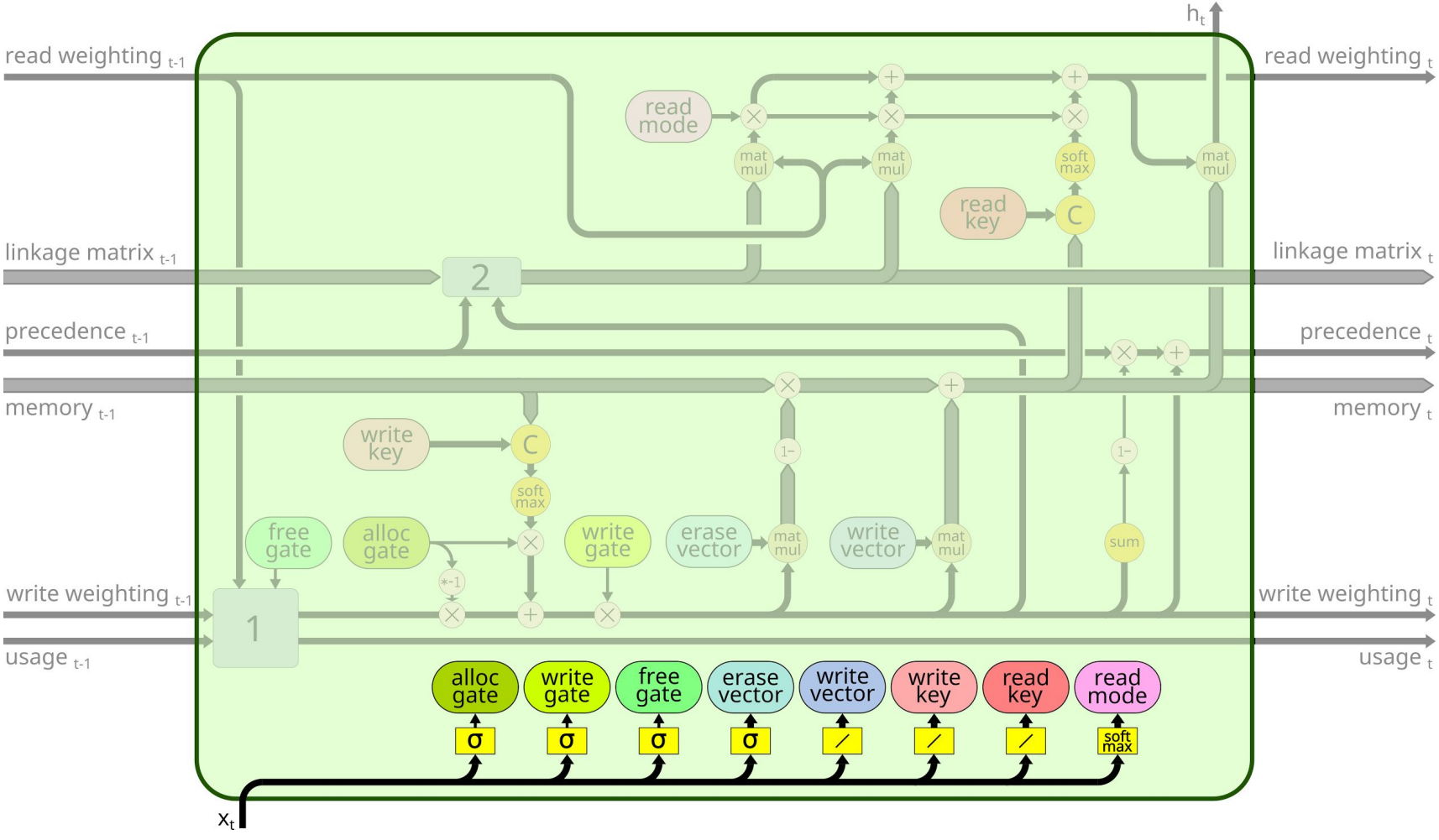
Source: Graves et al. 2016 [5]

DNC - Memory Unit

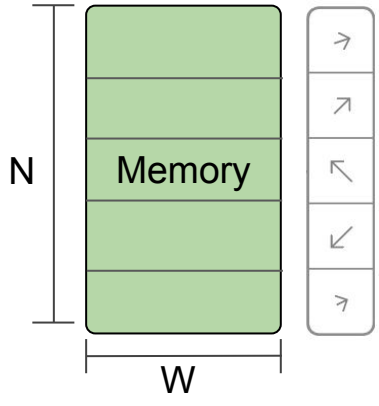


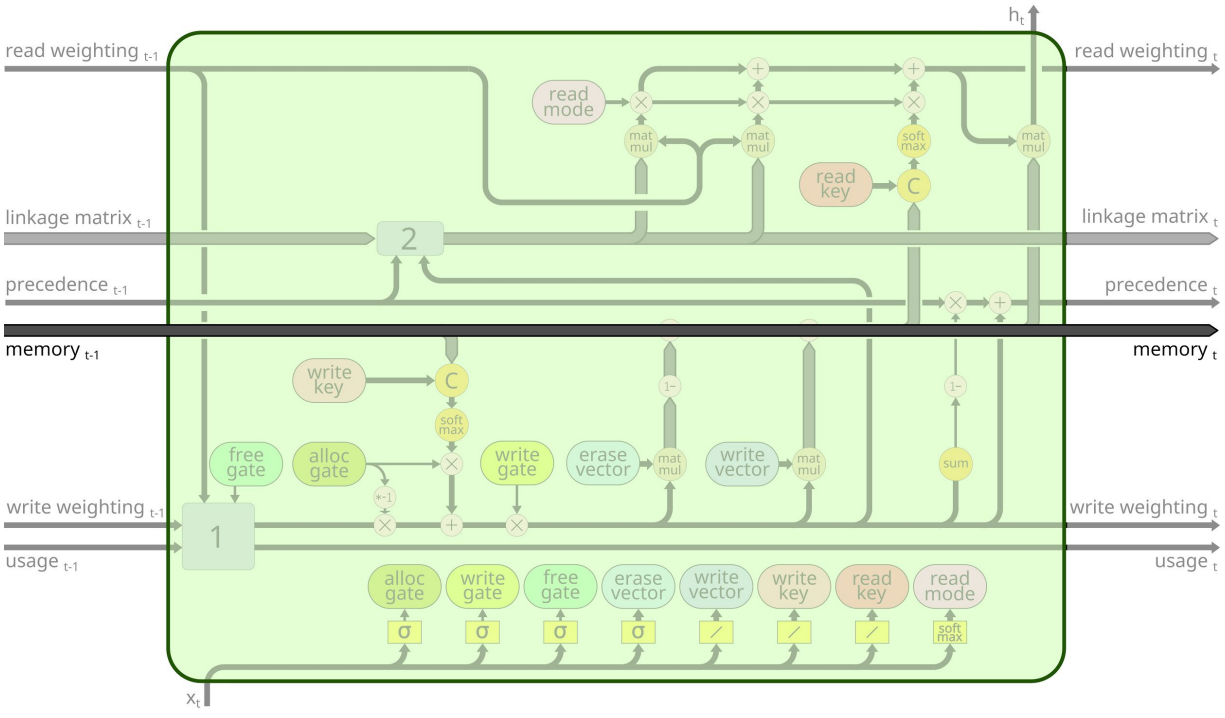
This figure has some simplifications, e.g.boxed calculations or simple softmax functions.

Memory Unit - Input signals

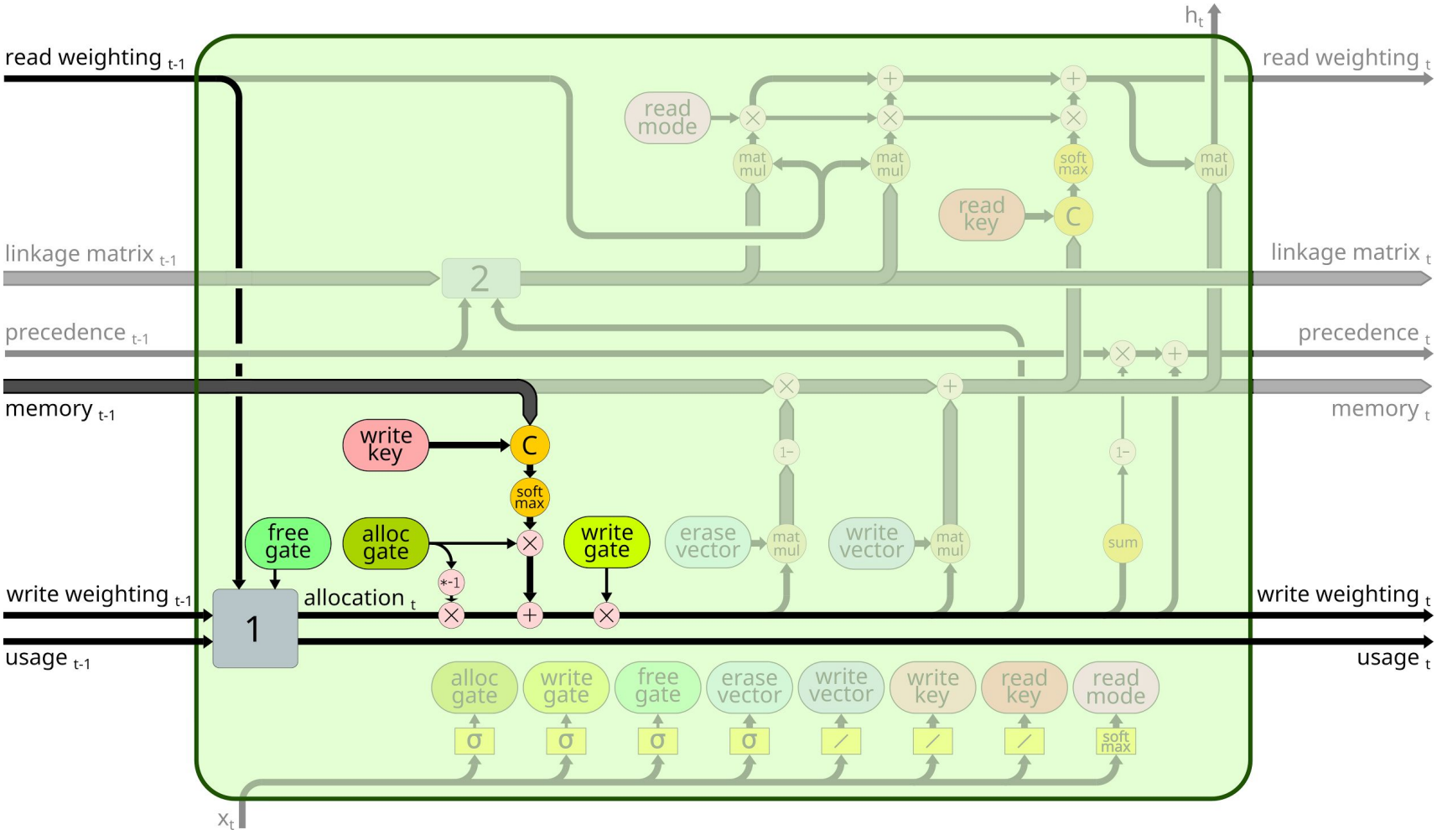


Memory Unit - Memory

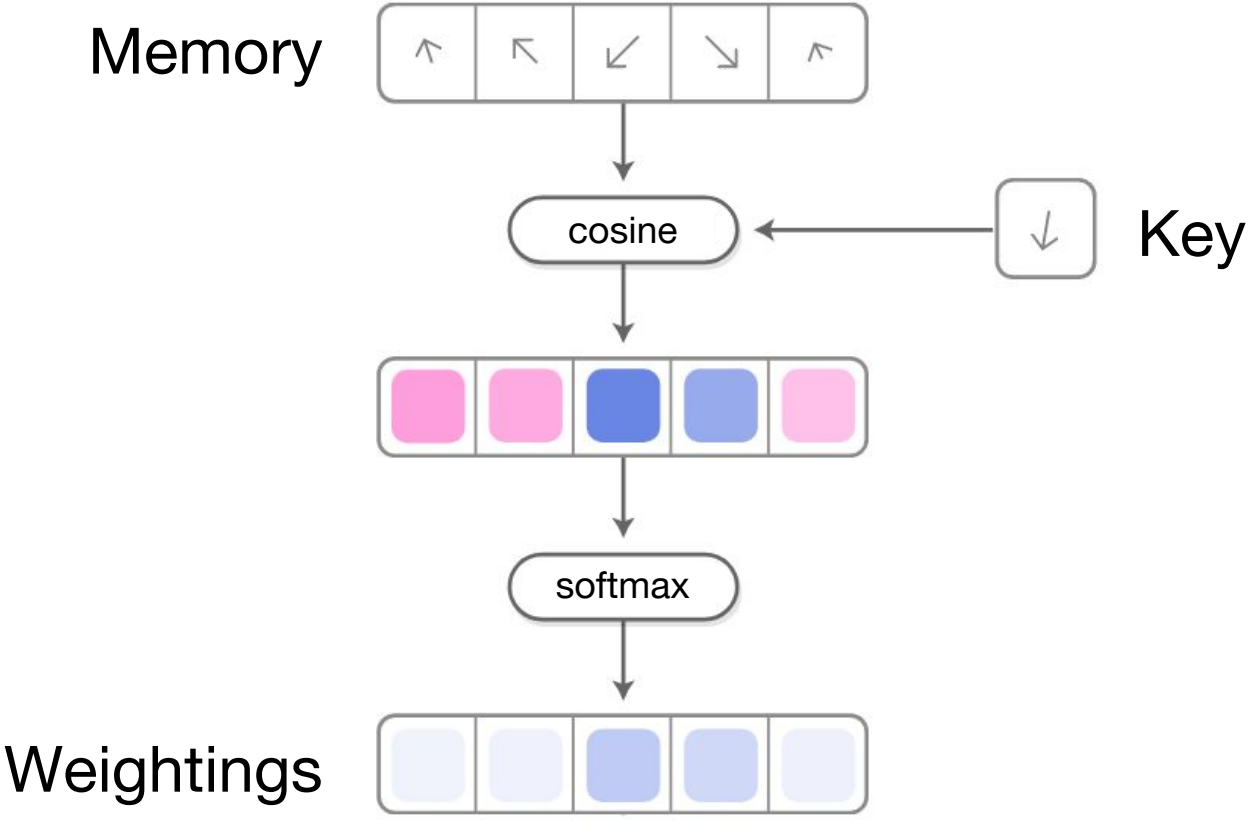




Memory Unit - Create new write weightings

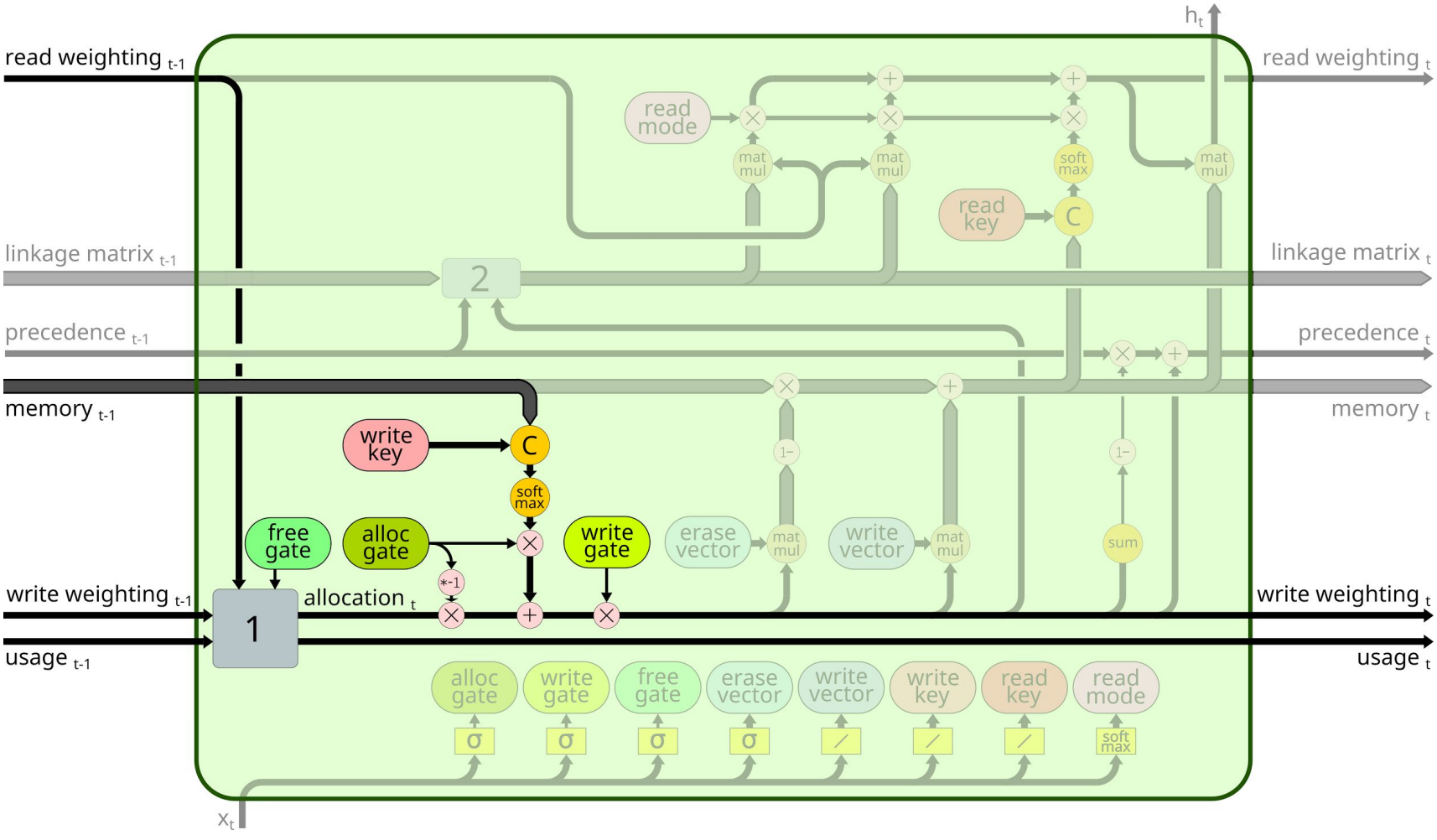


Memory addressing

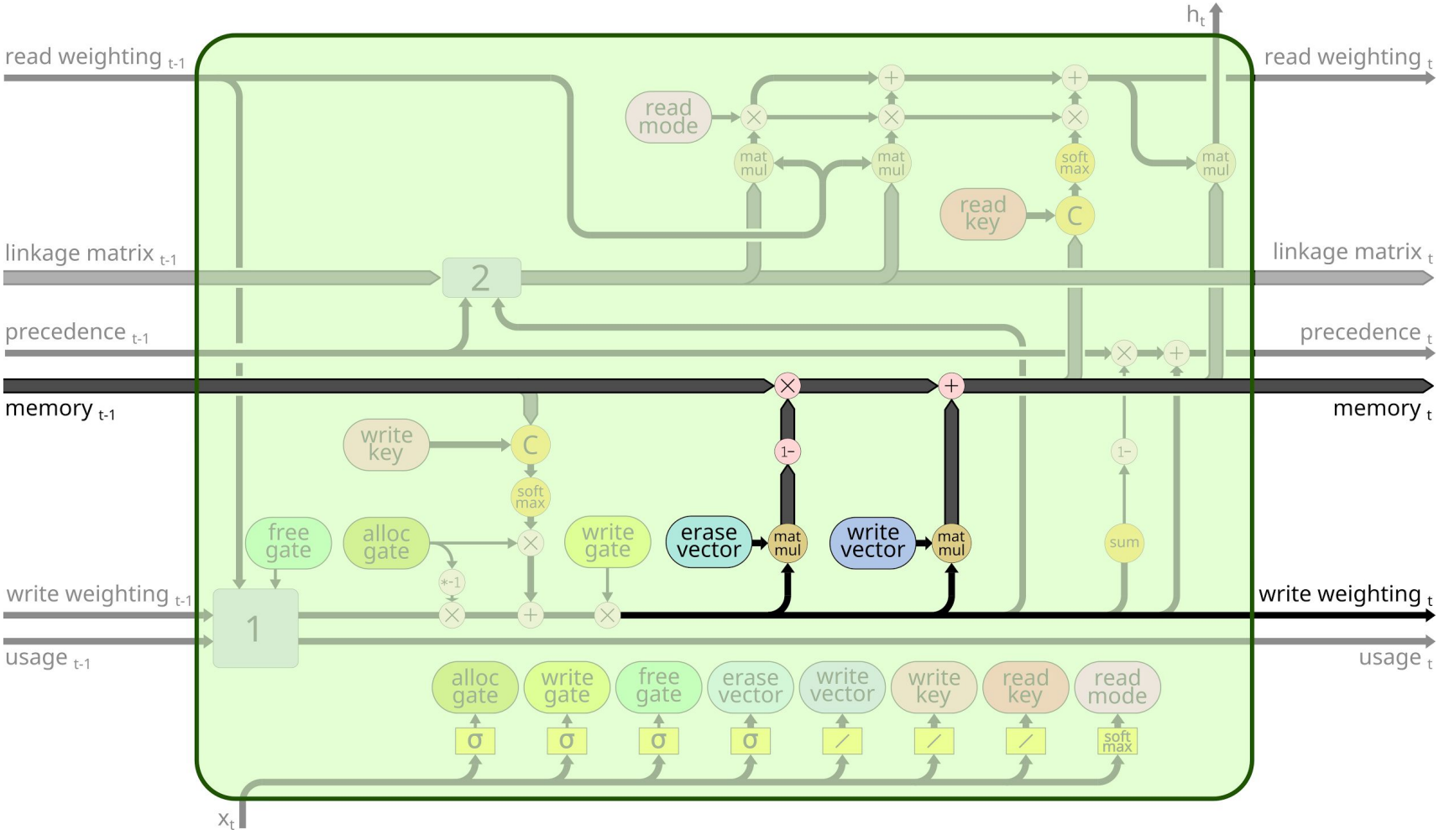


Source: https://distill.pub/2016/augmented-rnns/#neural-turing-machines 24.07.2017

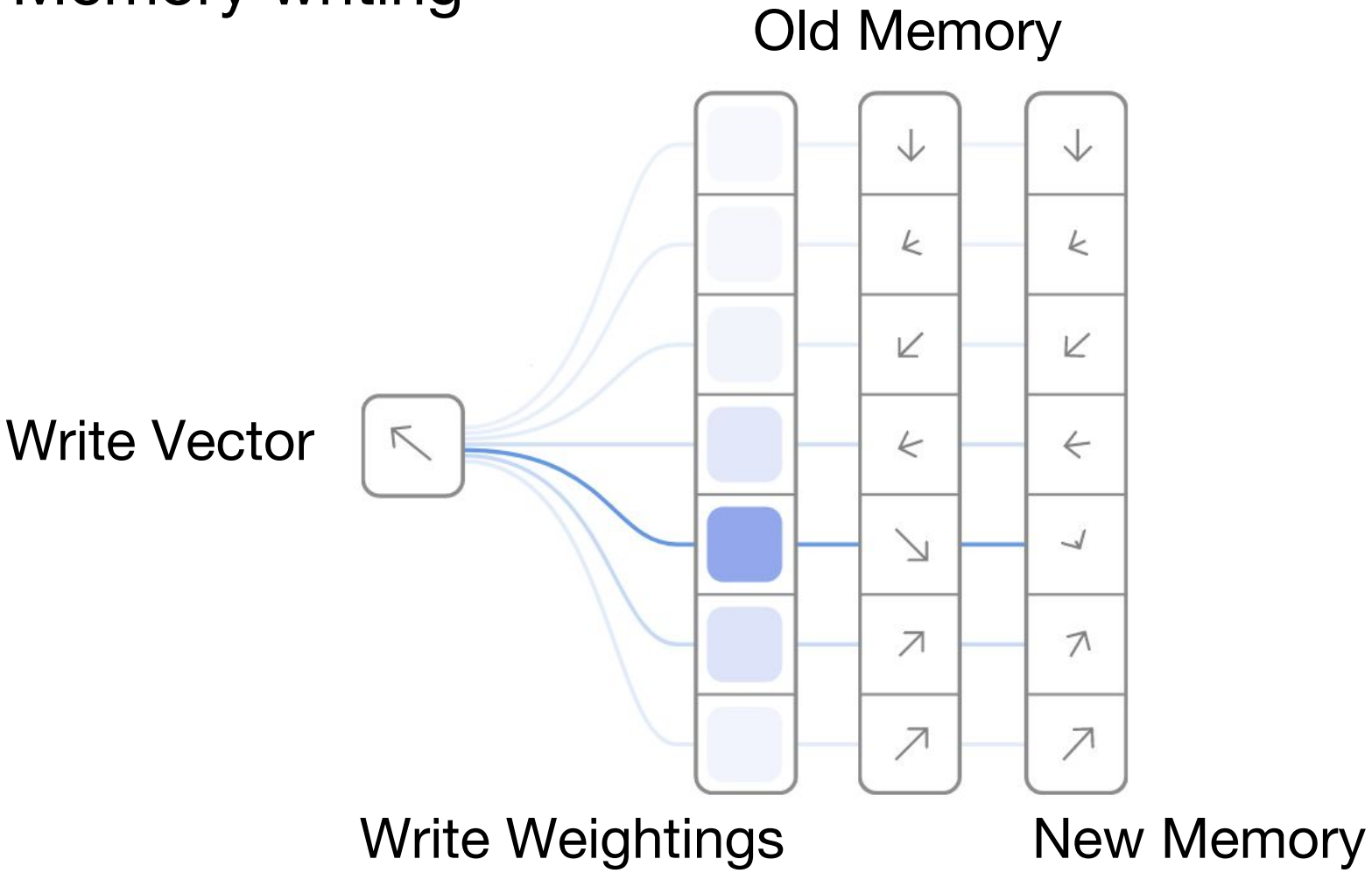
Memory Unit - Create new write weightings



Memory Unit - Write/Erase the memory

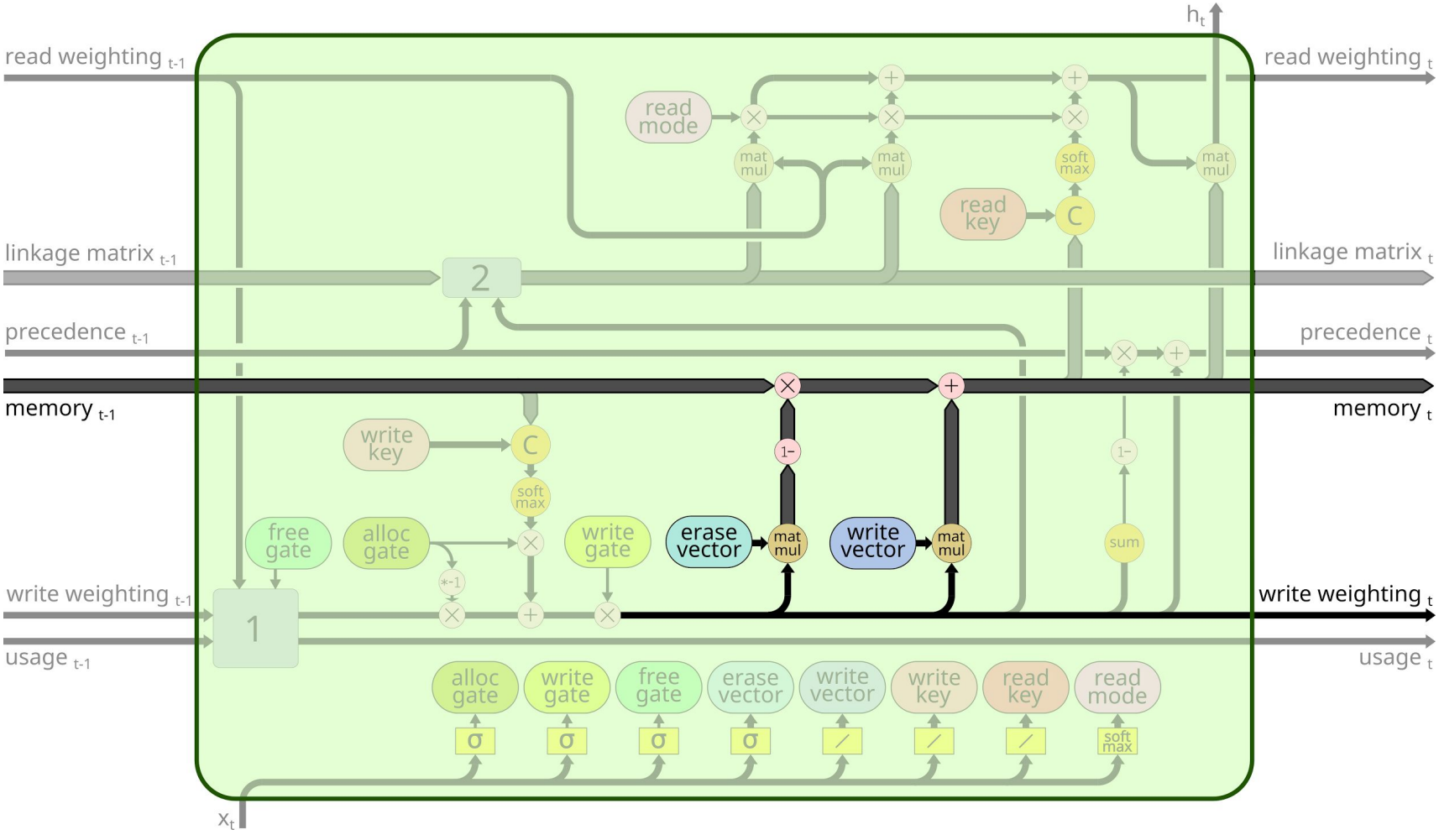


Memory writing

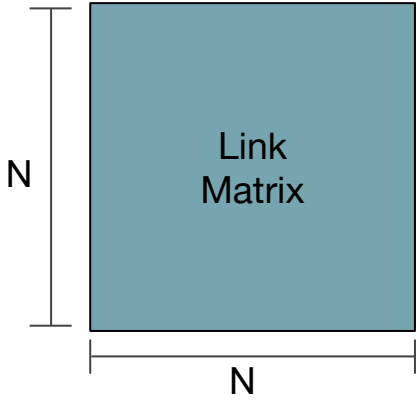


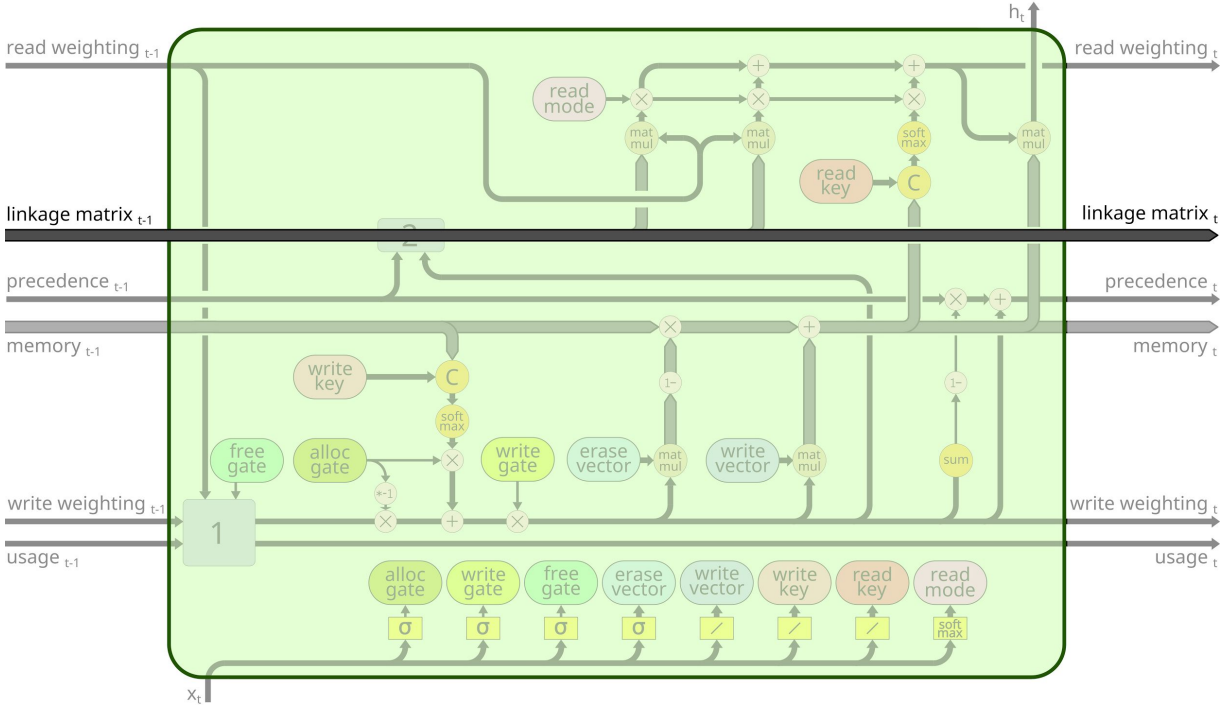
Source: https://distill.pub/2016/augmented-rnns/#neural-turing-machines 24.07.2017

Memory Unit - Write/Erase the memory

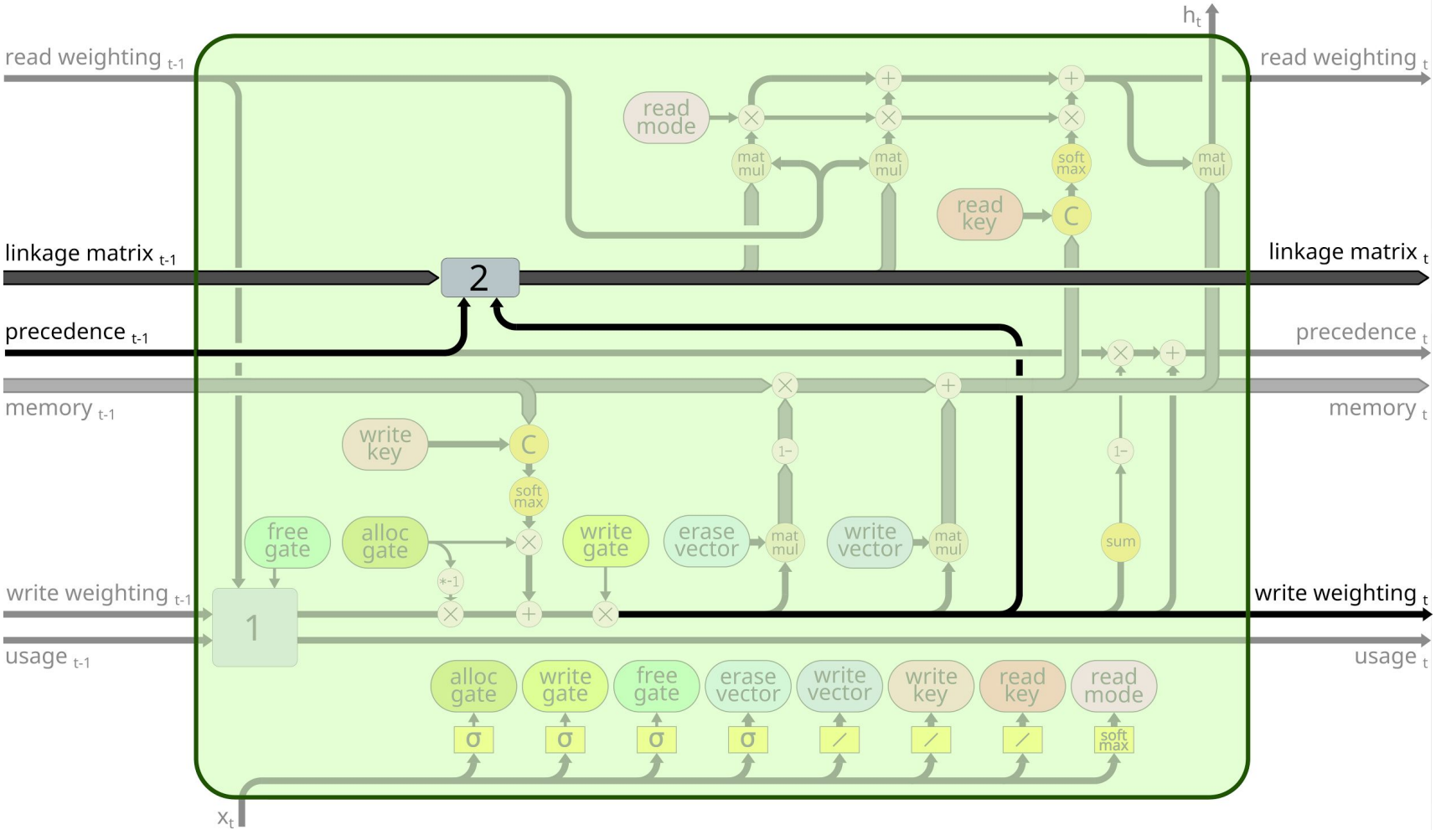


Memory Unit - Linkage Matrix

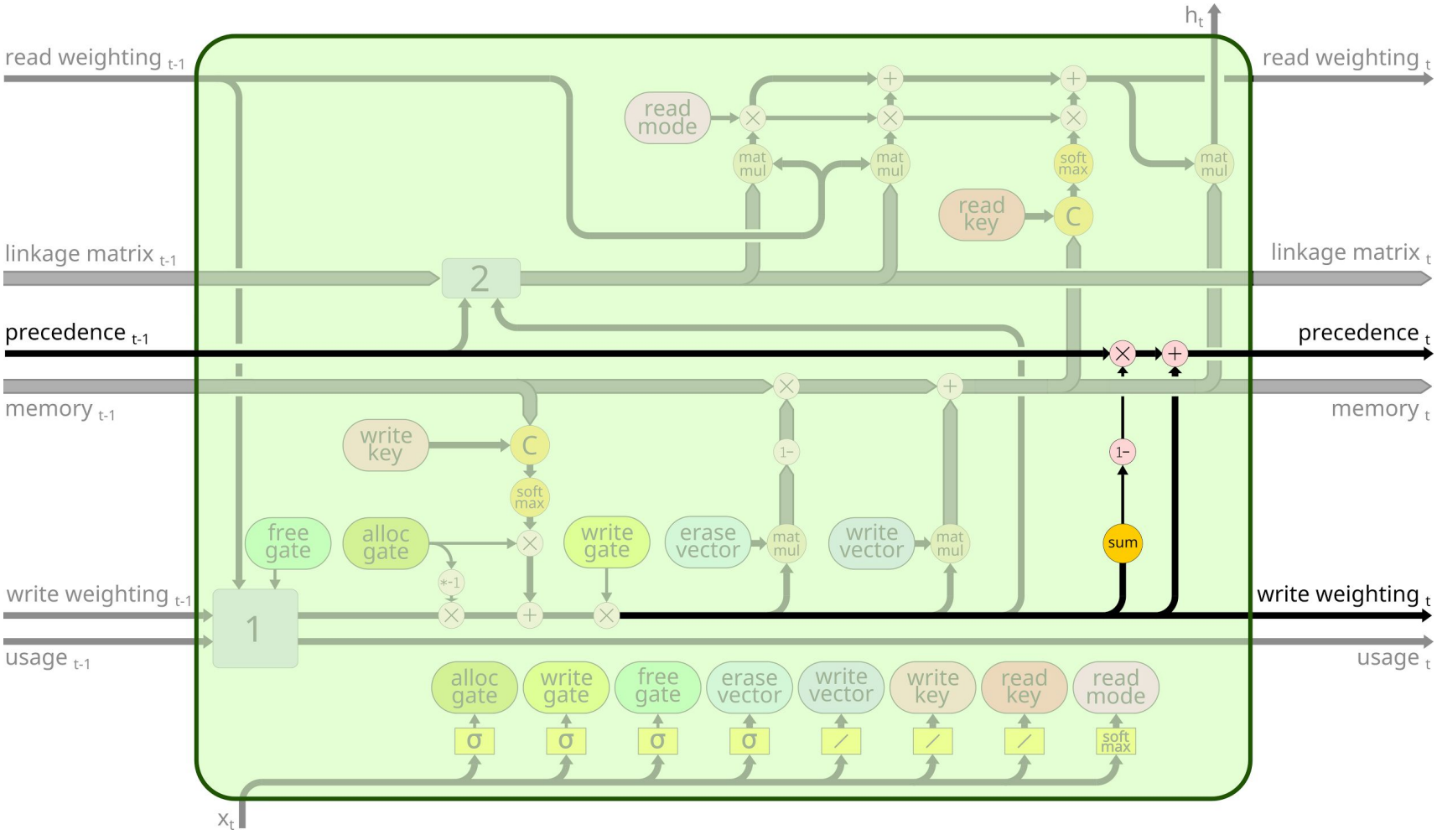




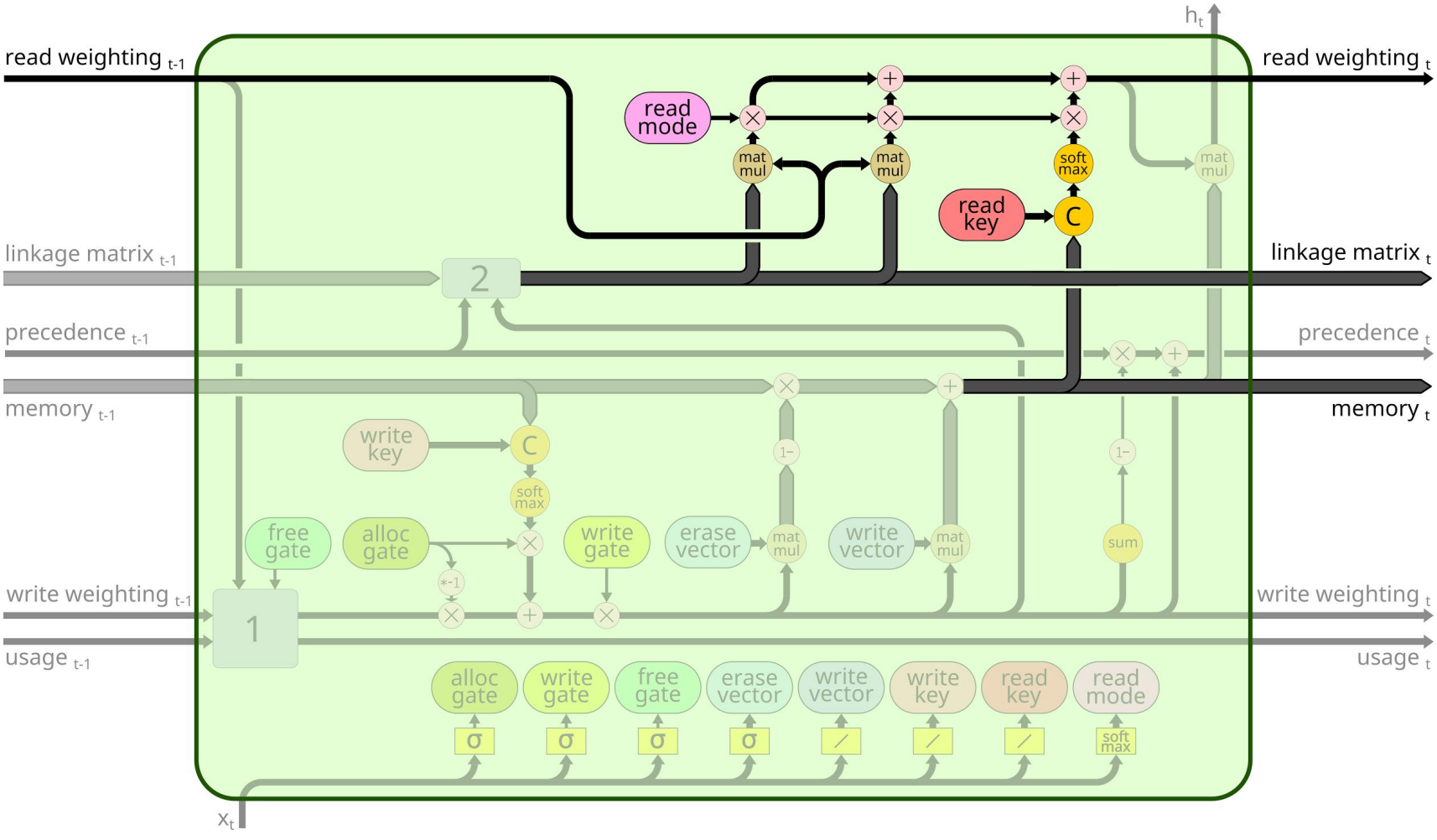
Memory Unit - Update linkage matrix



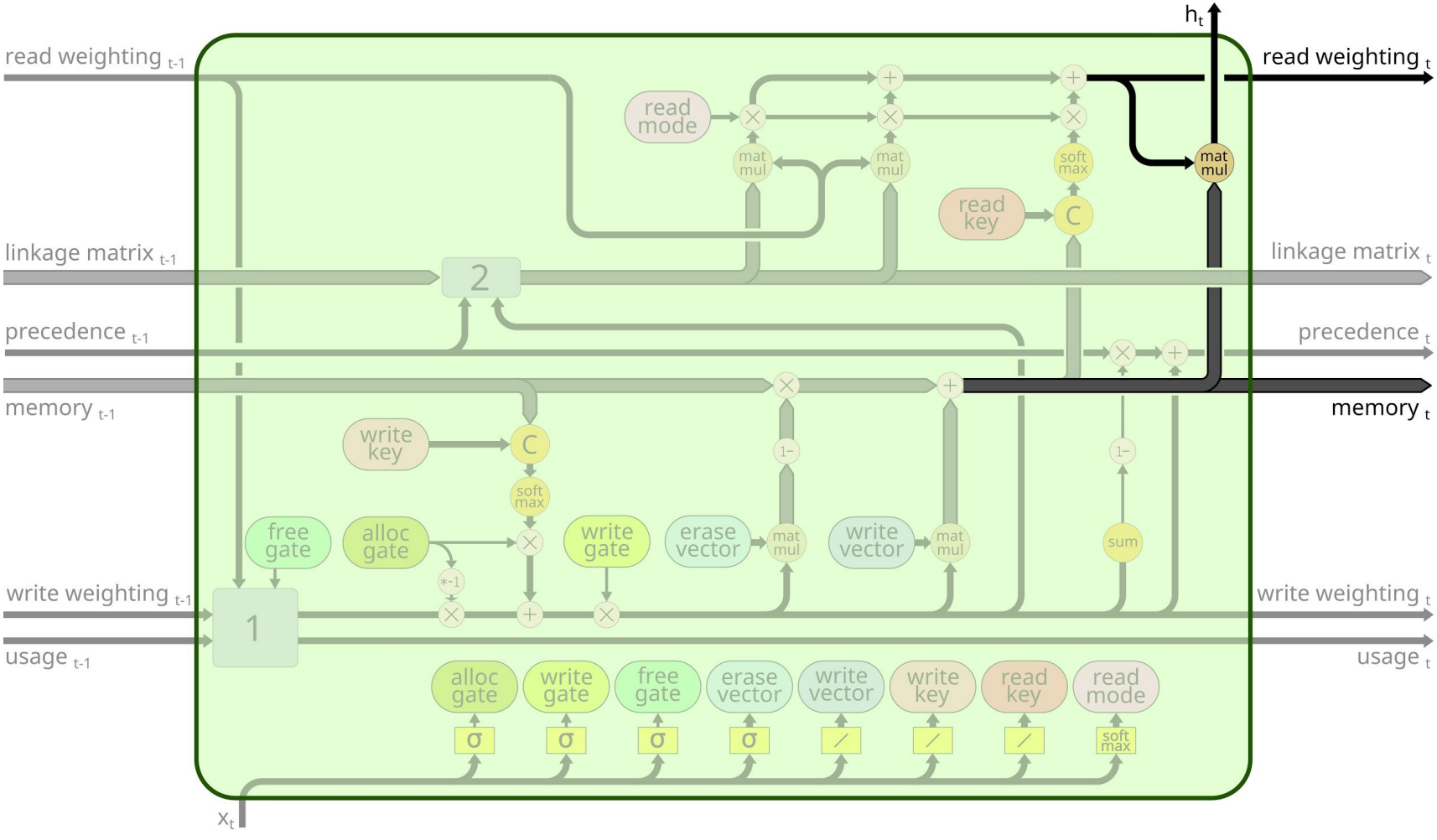
Memory Unit - Update precedence vector



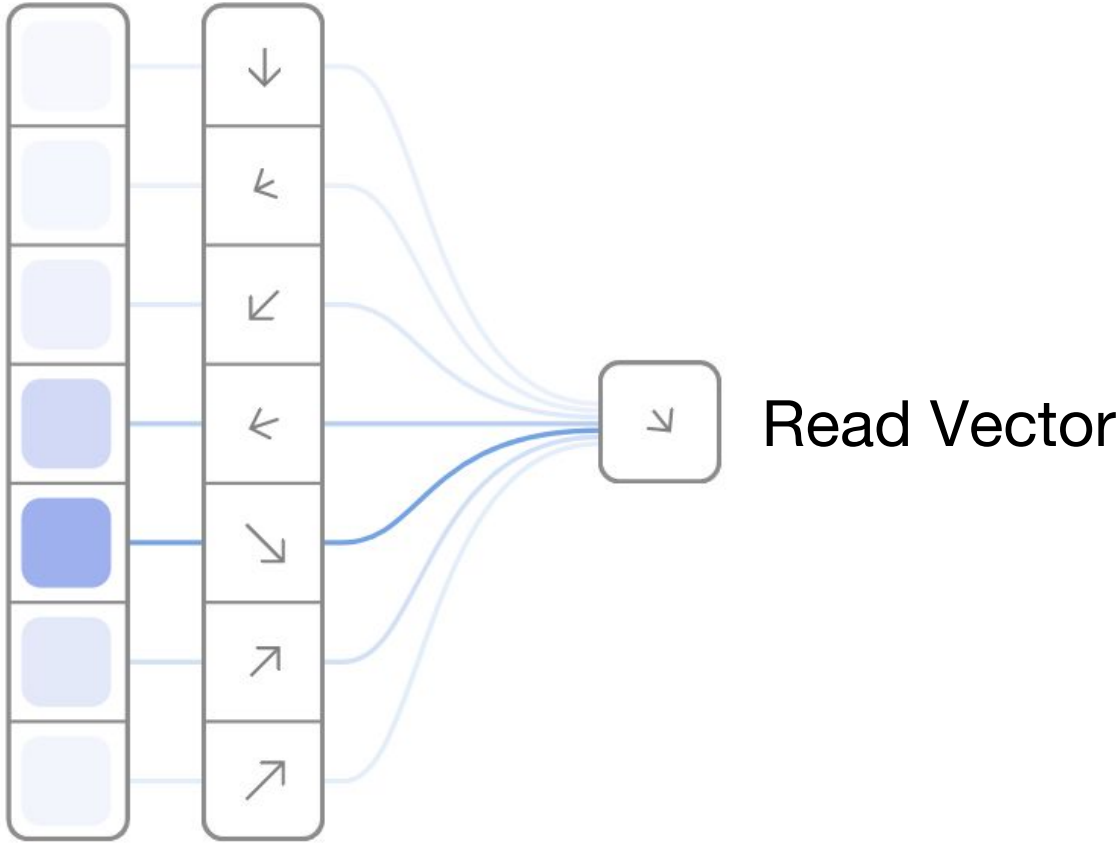
Memory Unit - Create new read weightings



Memory Unit - Read from the memory

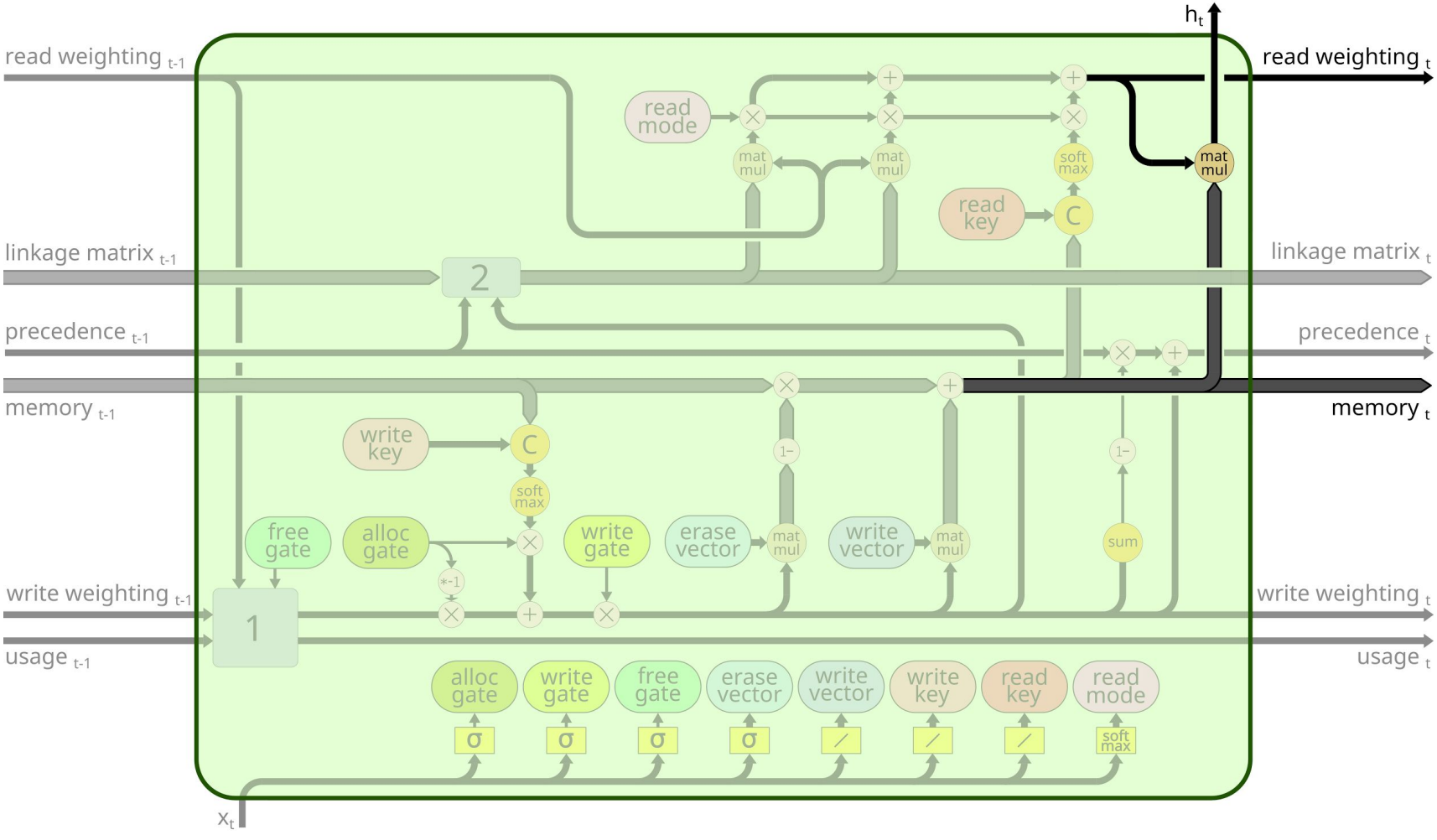


Memory reading

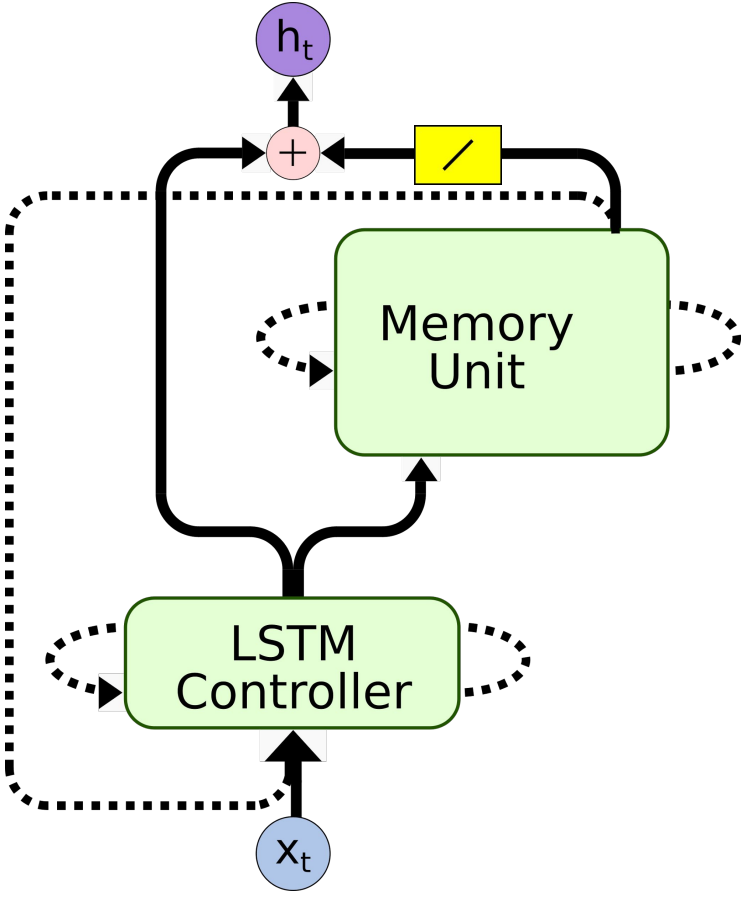


Source: https://distill.pub/2016/augmented-rnns/#neural-turing-machines 24.07.2017

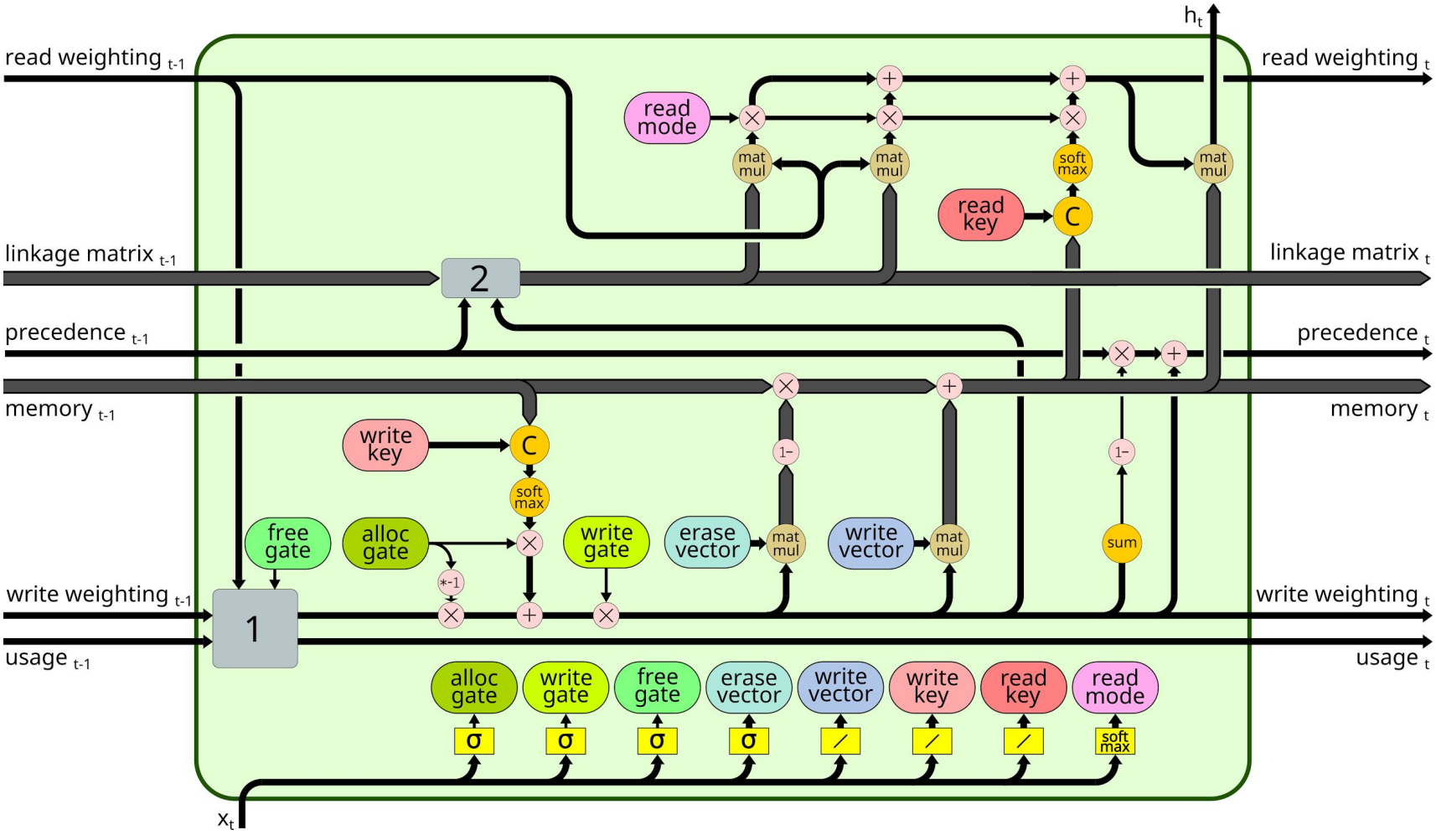
Memory Unit - Read from the memory



DNC - Architecture (recurrent connections)

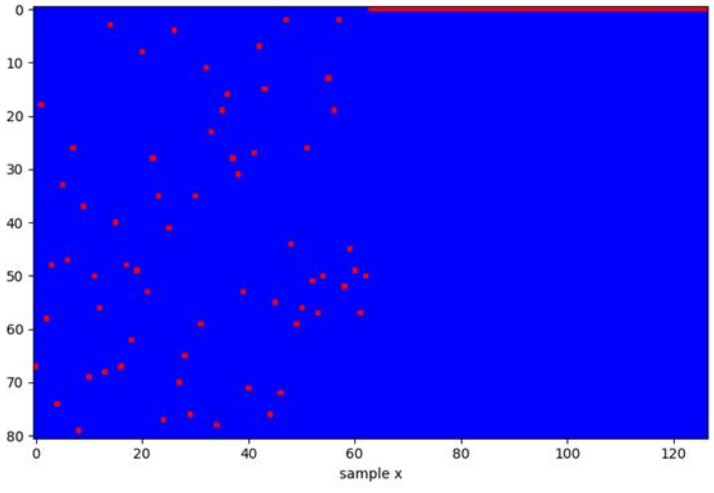


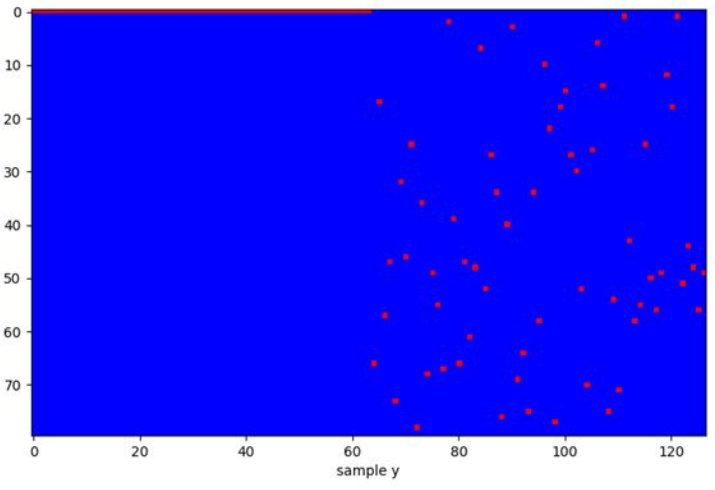
Memory Unit - Overview



Application

Copy Task

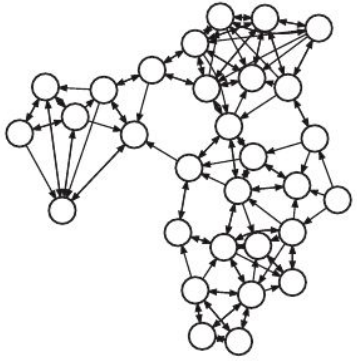




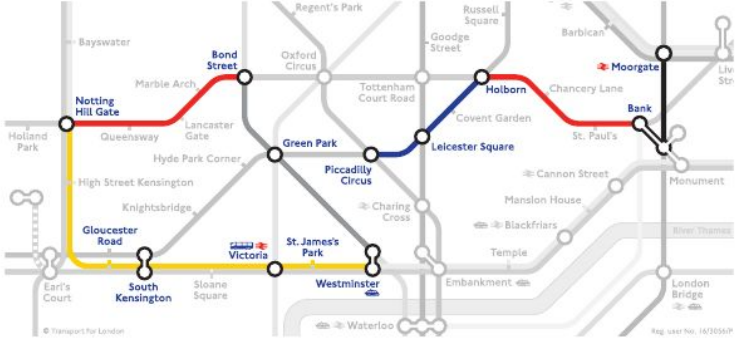
- Input: One-hot vector [12, 23, 01, 23, _, _, _, _]
- Output: One-hot vector [_, _, _, _, 12, 23, 01, 23]
- One-hot vector size up to 12.000
- Sequence length up to 100

Graph Task

a Random graph



b London Underground



Traversal

Underground input:

(OxfordCircus, TottenhamCtRd, Central)
(TottenhamCtRd, OxfordCircus, Central)
(BakerSt, Marylebone, Circle)
(BakerSt, Marylebone, Bakerloo)
(BakerSt, OxfordCircus, Bakerloo)
⋮
(LeicesterSq, CharingCross, Northern)
(TottenhamCtRd, LeicesterSq, Northern)
(OxfordCircus, PiccadillyCircus, Bakerloo)
(OxfordCircus, NottingHillGate, Central)
(OxfordCircus, Euston, Victoria)

84 edges in total

Traversal question:

(BondSt, _, Central),
(_, _, Circle), (_, _, Circle),
(_, _, Circle), (_, _, Circle),
(_, _, Jubilee), (_, _, Jubilee),

Answer:

(BondSt, NottingHillGate, Central)
(NottingHillGate, GloucesterRd, Circle)
⋮
(Westminster, GreenPark, Jubilee)
(GreenPark, BondSt, Jubilee)

Shortest-path

Shortest-path question:

(Moorgate, PiccadillyCircus, _)

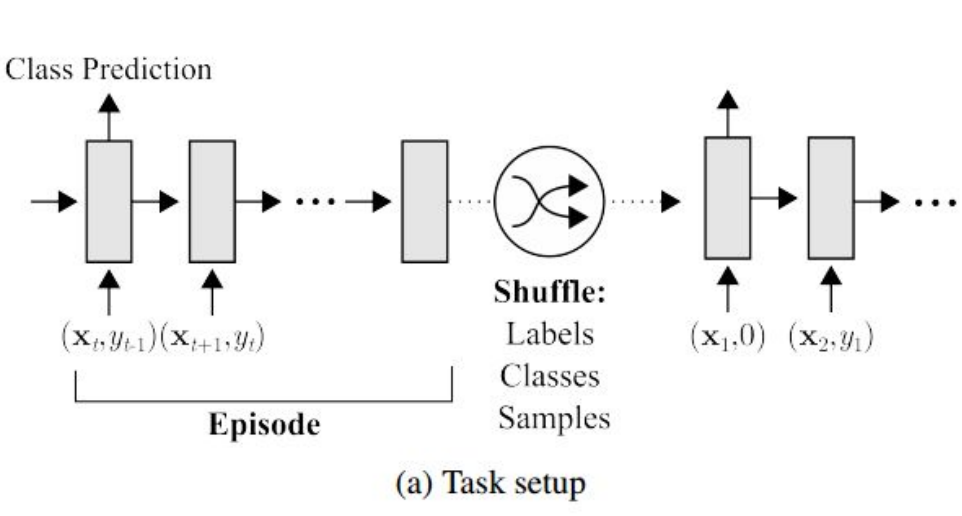
Answer:

(Moorgate, Bank, Northern)
(Bank, Holborn, Central)
(Holborn, LeicesterSq, Piccadilly)
(LeicesterSq, PiccadillyCircus, Piccadilly)

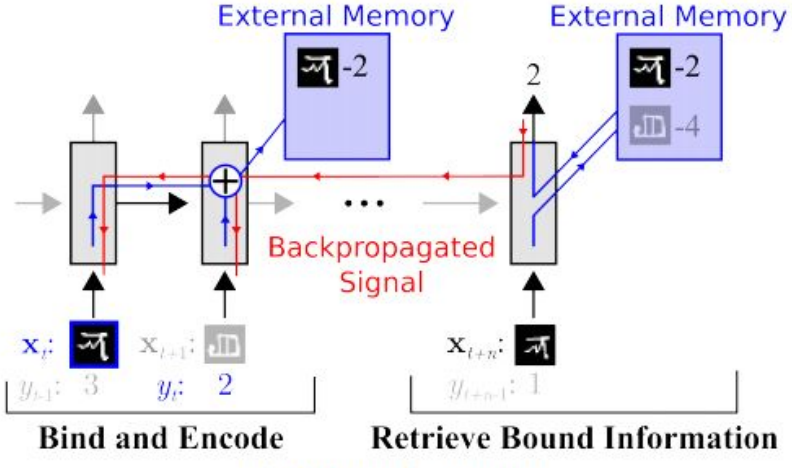
Source: Graves et al. 2016 [5]

One shot learning

One-shot learning with memory-augmented neural networks [6]



(a) Task setup



(b) Network strategy

My personal training insights for DNCs

- RMSprop works better than Adam, AdaDelta or AdaGrad
- Sometimes is takes up to 12 epochs before the loss decreases
- For more complex tasks, curriculum learning works pretty good
- Too many parallel workers in a distributed setting decrease learning vastly
- Adding an LSTM-Layer at output could be a good idea
- Just concatenating memory output with controller is not the best idea

Thanks for your attention!

Any questions?

References

[1] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities", *Proceedings of the National Academy of Sciences of the USA*, vol. 79 no. 8 pp. 2554–2558, April 1982.

[2] Elman, Jeffrey L. "Finding structure in time." *Cognitive science* 14.2 (1990): 179-211.

[3] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural computation* 9.8 (1997): 1735-1780.

[4] Graves, Alex, Greg Wayne, and Ivo Danihelka. "Neural turing machines." *arXiv preprint arXiv:1410.5401* (2014).

[5] Graves, Alex, et al. "Hybrid computing using a neural network with dynamic external memory." *Nature* 538.7626 (2016): 471-476.

[6] Santoro, Adam, et al. "One-shot learning with memory-augmented neural networks. arXiv preprint." *arXiv preprint arXiv:1605.06065* (2016).

Stanford Question Answering Task (SQuAD)

- Vocabulary ~ 120.000
- 100.000+ Question-answer pairs
- 500+ articles

Context	Immigrants arrived from all over the world to search for gold, especially from Ireland and China. Many Chinese miners worked in Victoria, and their legacy is particularly strong in Bendigo and its environs. Although there was some racism directed at them, there was not the level of anti-Chinese violence that was seen at the Lambing Flat riots in New South Wales. However, there was a riot at Buckland Valley.
Question	Where was the 1857 riot?
Answer	Buckland Valley

Make neural networks recurrent