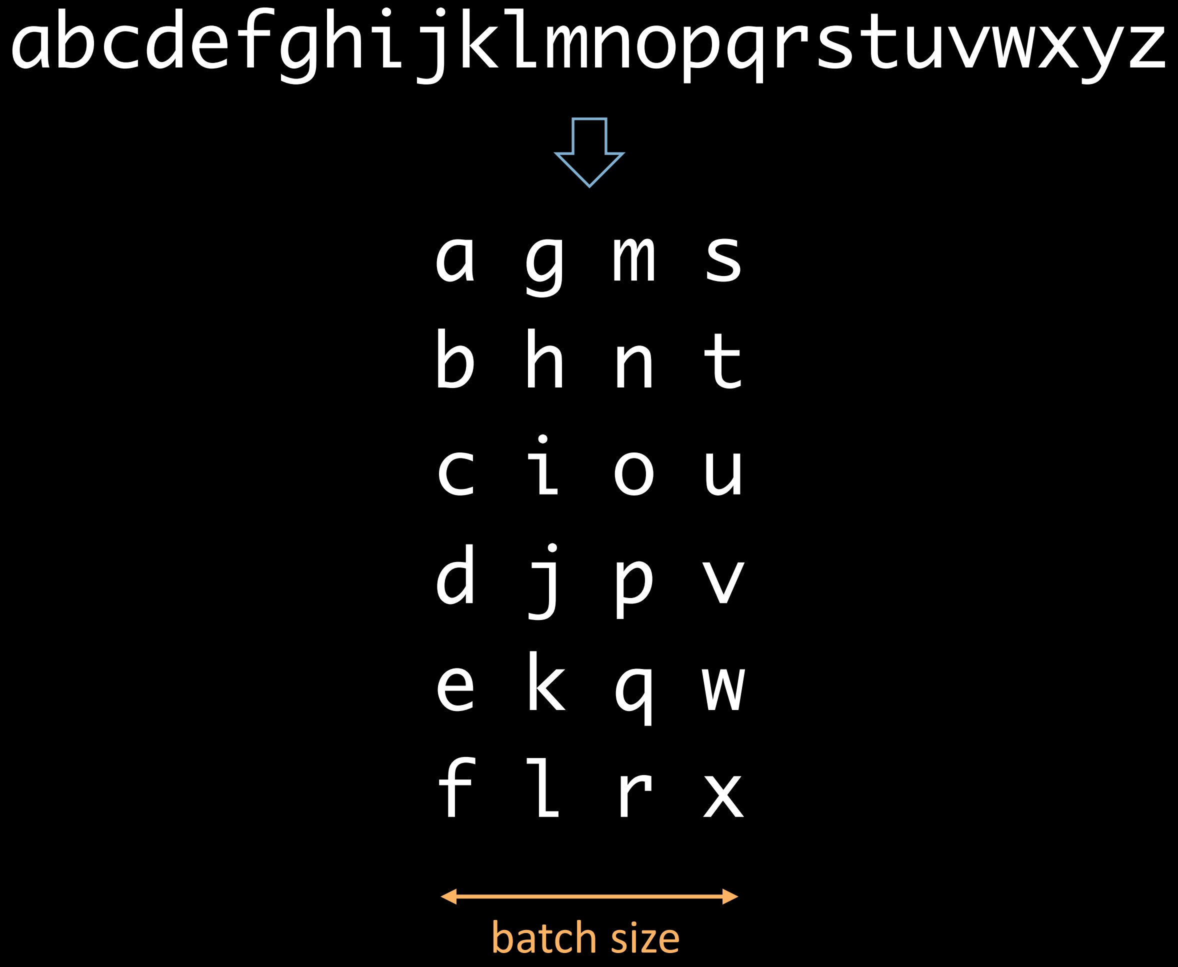
Training example

Language modelling

Batch-ification

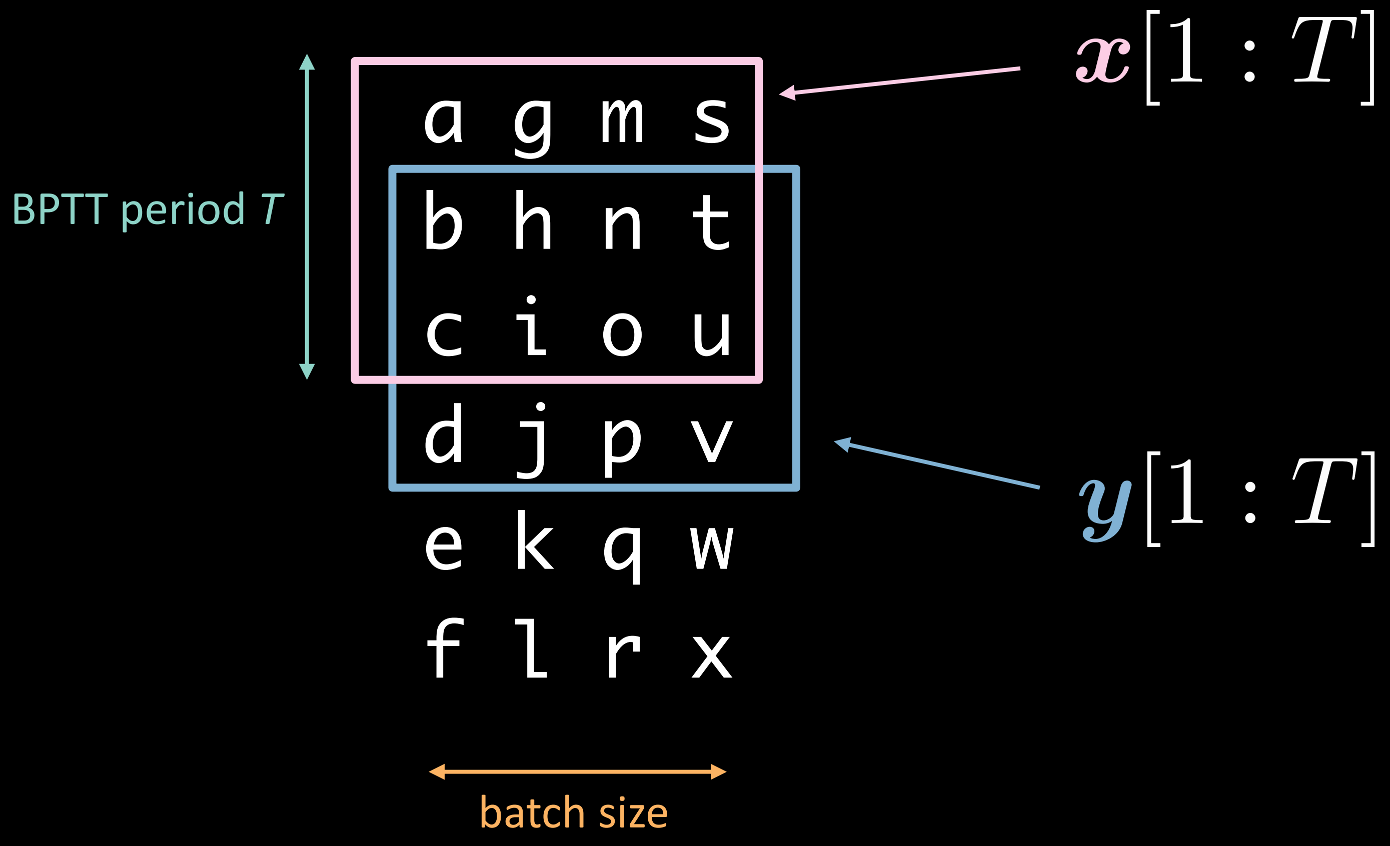


abcdefghijklmnopqrstuvwxyz

a	g	m	s
b	h	n	t
c	i	o	u
d	j	p	v
e	k	q	w
f	l	r	x

Check word_language_model @ github.com/pytorch/examples/

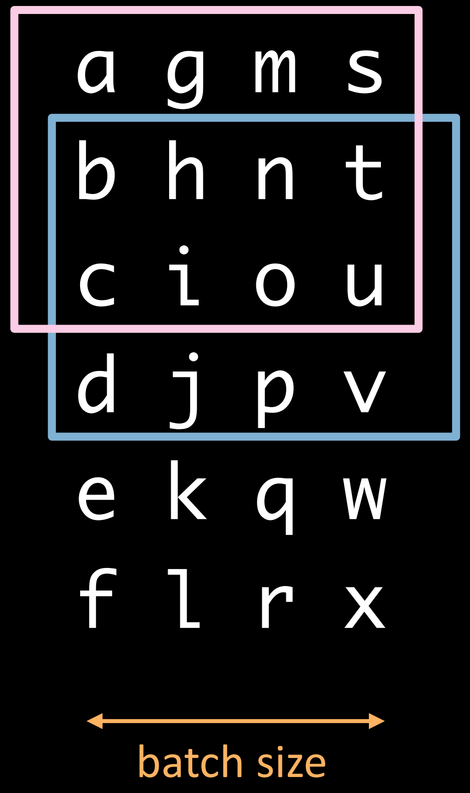
Get batch (I)



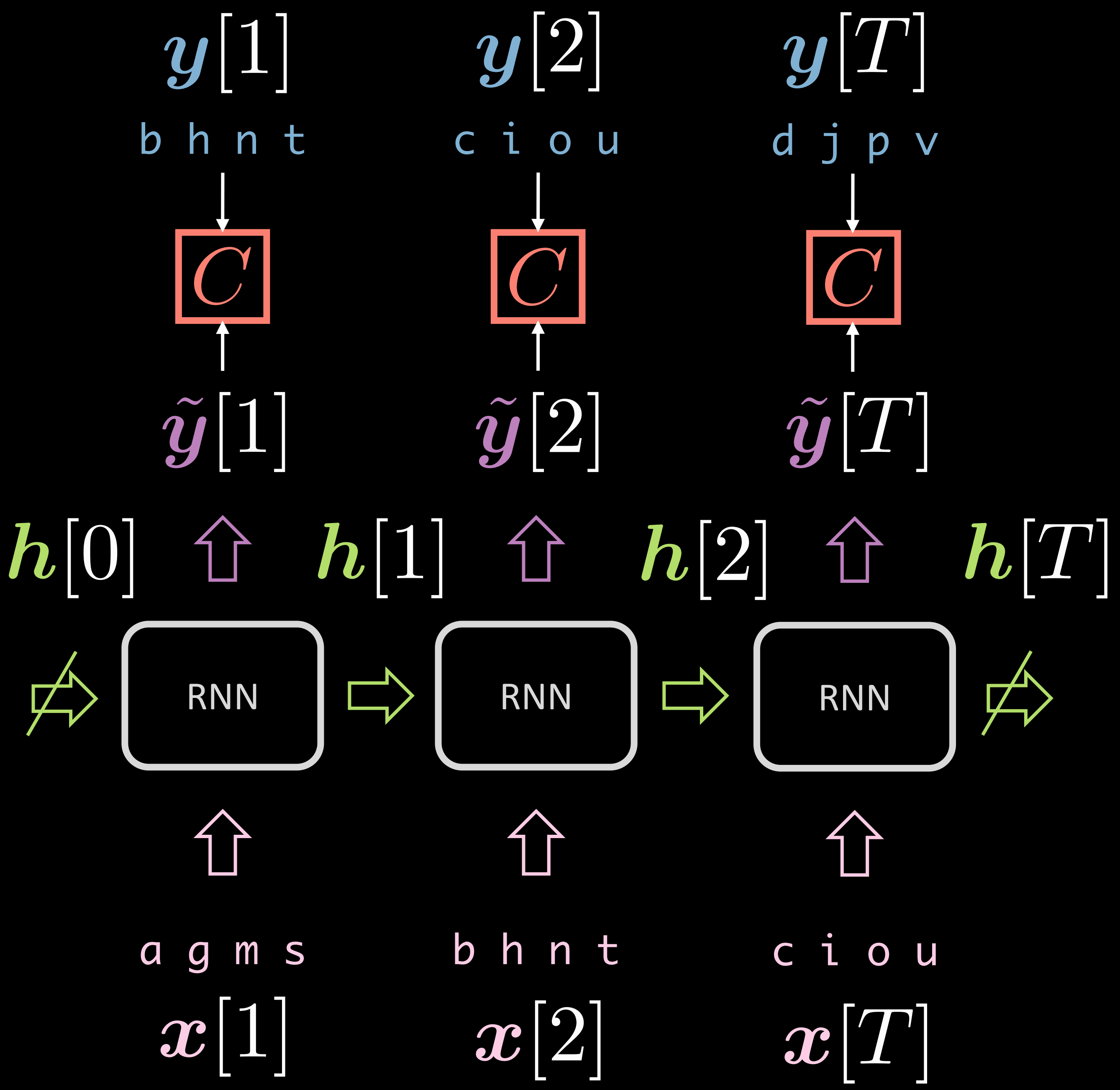
a	g	m	s
b	h	n	t
c	i	o	u
d	j	p	v
e	k	q	w
f	l	r	x

Check `word_language_model` @ `github.com/pytorch/examples/`

Get batch (II)



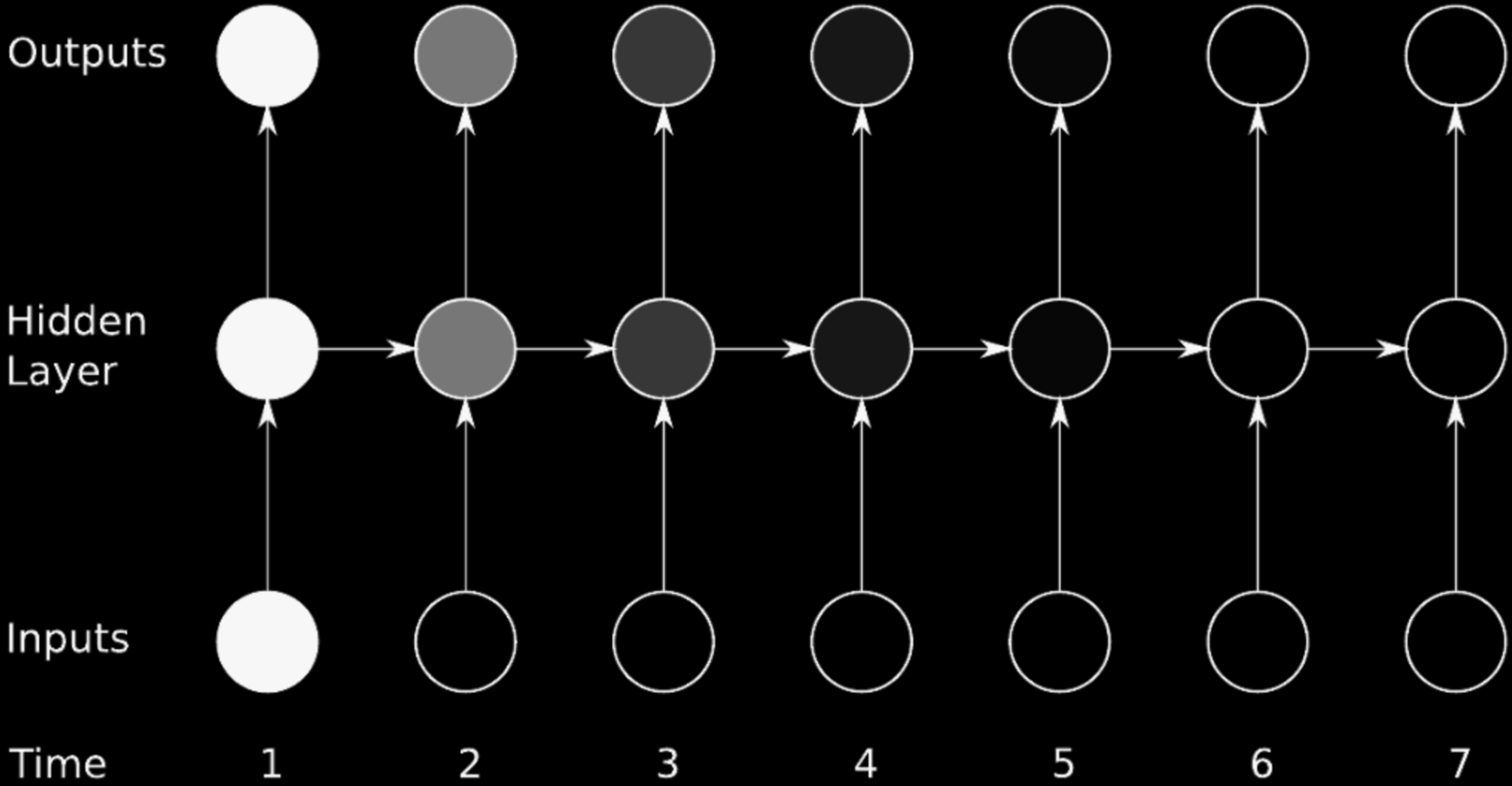
a g m s
b h n t
c i o u
d j p v
e k q w
f l r x



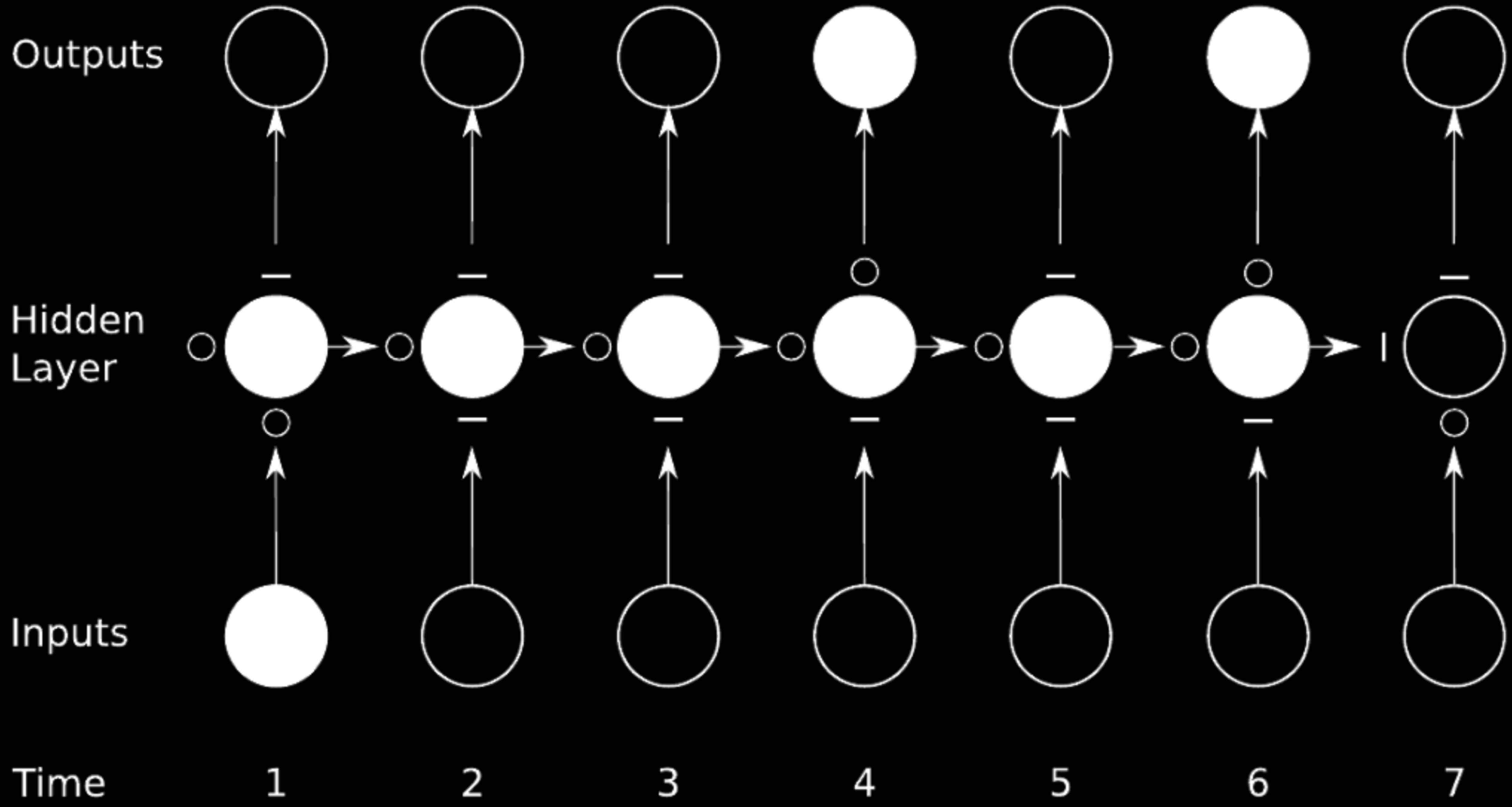
Check word_language_model @ github.com/pytorch/examples/

Vanishing & exploding gradients

Limitations of temporally deep nets



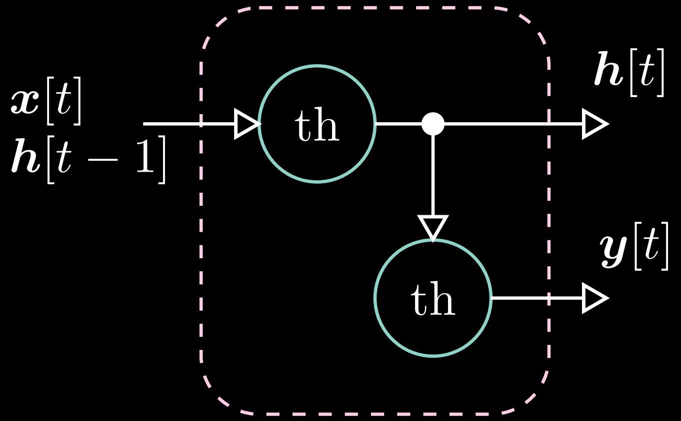
Graves (2012) Supervised sequence labelling



Graves (2012) Supervised sequence labelling

Long Short-Term Memory

Gated RNN



$$\boldsymbol{h}[t] = g\left(\boldsymbol{W}_h \begin{bmatrix} \boldsymbol{x}[t] \\ \boldsymbol{h}[t-1] \end{bmatrix} + \boldsymbol{b}_h\right)$$

$$\hat{\boldsymbol{y}}[t] = g(\boldsymbol{W}_y \boldsymbol{h}[t] + \boldsymbol{b}_y)$$

$$\boldsymbol{i}[t] = \sigma\left(\boldsymbol{W}_i \begin{bmatrix} \boldsymbol{x}[t] \\ \boldsymbol{h}[t-1] \end{bmatrix} + \boldsymbol{b}_i\right)$$

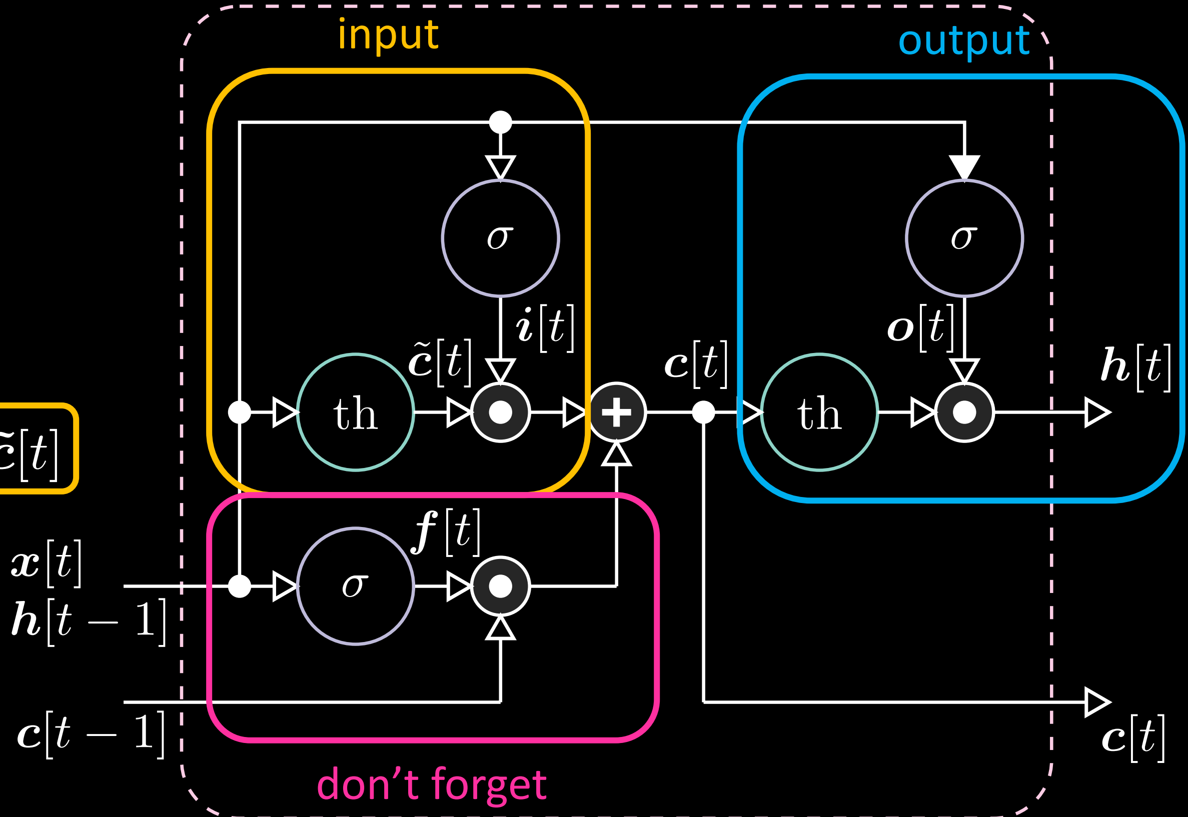
$$\boldsymbol{f}[t] = \sigma\left(\boldsymbol{W}_f \begin{bmatrix} \boldsymbol{x}[t] \\ \boldsymbol{h}[t-1] \end{bmatrix} + \boldsymbol{b}_f\right)$$

$$\boldsymbol{o}[t] = \sigma\left(\boldsymbol{W}_o \begin{bmatrix} \boldsymbol{x}[t] \\ \boldsymbol{h}[t-1] \end{bmatrix} + \boldsymbol{b}_o\right)$$

$$\tilde{\boldsymbol{c}}[t] = \tanh\left(\boldsymbol{W}_c \begin{bmatrix} \boldsymbol{x}[t] \\ \boldsymbol{h}[t-1] \end{bmatrix} + \boldsymbol{b}_c\right)$$

$$\boldsymbol{c}[t] = \boldsymbol{f}[t] \odot \boldsymbol{c}[t-1] + \boldsymbol{i}[t] \odot \tilde{\boldsymbol{c}}[t]$$

$$\boldsymbol{h}[t] = \boldsymbol{o}[t] \odot \tanh(\boldsymbol{c}[t])$$



Controlling the output - OFF

Saturated sigmoid
- 🟢 = 1
- 🔴 = 0

$$\boldsymbol{i}[t] = \sigma\left(\boldsymbol{W}_i \begin{bmatrix} \boldsymbol{x}[t] \\ \boldsymbol{h}[t-1] \end{bmatrix} + \boldsymbol{b}_i\right)$$

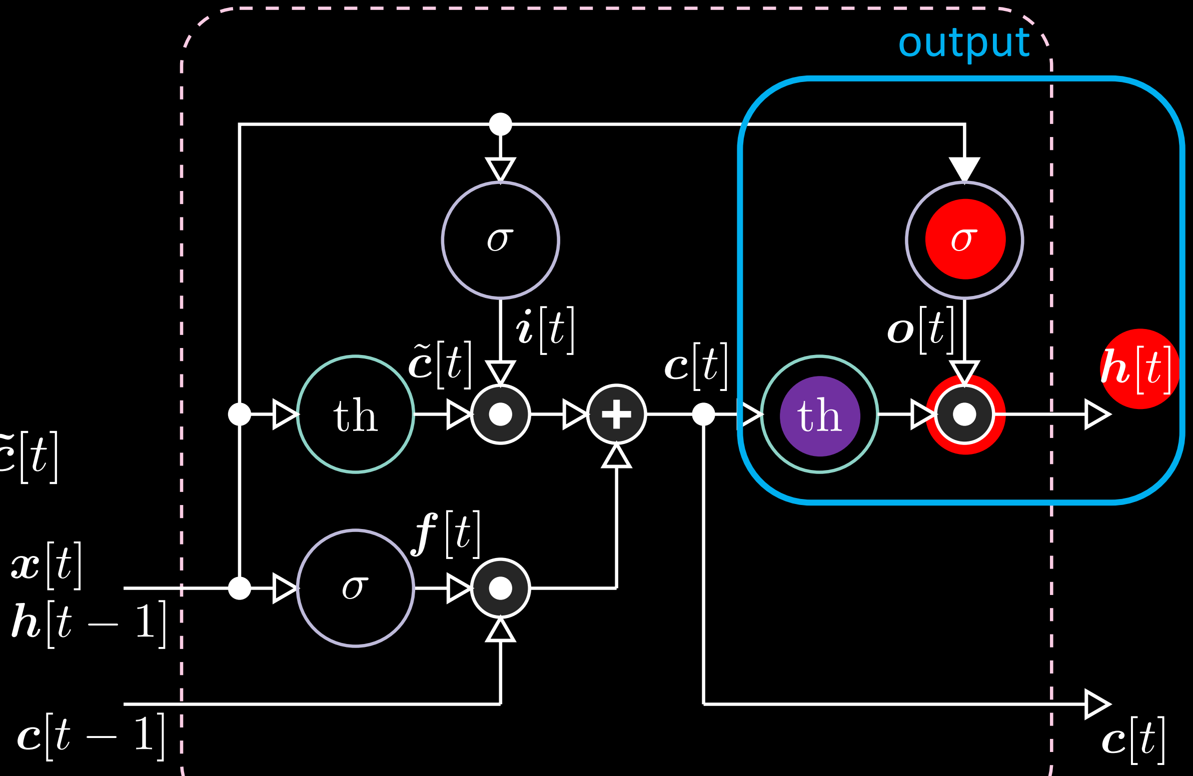
$$\boldsymbol{f}[t] = \sigma\left(\boldsymbol{W}_f \begin{bmatrix} \boldsymbol{x}[t] \\ \boldsymbol{h}[t-1] \end{bmatrix} + \boldsymbol{b}_f\right)$$

$$\boxed{\boldsymbol{o}[t] = \sigma\left(\boldsymbol{W}_o \begin{bmatrix} \boldsymbol{x}[t] \\ \boldsymbol{h}[t-1] \end{bmatrix} + \boldsymbol{b}_o\right)}$$

$$\tilde{\boldsymbol{c}}[t] = \tanh\left(\boldsymbol{W}_c \begin{bmatrix} \boldsymbol{x}[t] \\ \boldsymbol{h}[t-1] \end{bmatrix} + \boldsymbol{b}_c\right)$$

$$\boldsymbol{c}[t] = \boldsymbol{f}[t] \odot \boldsymbol{c}[t-1] + \boldsymbol{i}[t] \odot \tilde{\boldsymbol{c}}[t]$$

$$\boldsymbol{h}[t] = \boxed{\boldsymbol{o}[t] \odot \tanh(\boldsymbol{c}[t])}$$



Controlling the output - ON

Saturated sigmoid
- 🟢 = 1
- 🔴 = 0

$$\boldsymbol{i}[t] = \sigma\left(\boldsymbol{W}_i \begin{bmatrix} \boldsymbol{x}[t] \\ \boldsymbol{h}[t-1] \end{bmatrix} + \boldsymbol{b}_i\right)$$

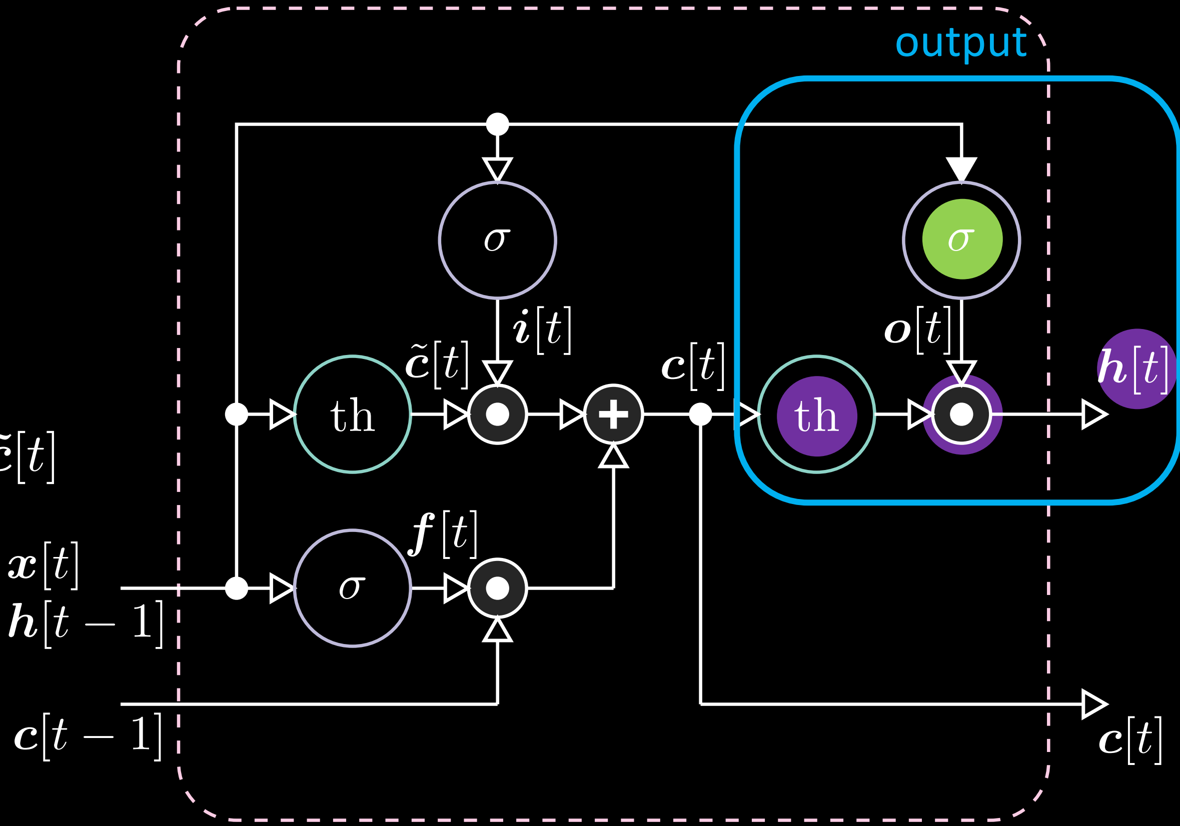
$$\boldsymbol{f}[t] = \sigma\left(\boldsymbol{W}_f \begin{bmatrix} \boldsymbol{x}[t] \\ \boldsymbol{h}[t-1] \end{bmatrix} + \boldsymbol{b}_f\right)$$

$$\boxed{\boldsymbol{o}[t] = \sigma\left(\boldsymbol{W}_o \begin{bmatrix} \boldsymbol{x}[t] \\ \boldsymbol{h}[t-1] \end{bmatrix} + \boldsymbol{b}_o\right)}$$

$$\tilde{\boldsymbol{c}}[t] = \tanh\left(\boldsymbol{W}_c \begin{bmatrix} \boldsymbol{x}[t] \\ \boldsymbol{h}[t-1] \end{bmatrix} + \boldsymbol{b}_c\right)$$

$$\boldsymbol{c}[t] = \boldsymbol{f}[t] \odot \boldsymbol{c}[t-1] + \boldsymbol{i}[t] \odot \tilde{\boldsymbol{c}}[t]$$

$$\boldsymbol{h}[t] = \boxed{\boldsymbol{o}[t] \odot \tanh(\boldsymbol{c}[t])}$$



Controlling the $\boxed{\boxed{\text{memory}}}$ - reset

Saturated sigmoid
- 🟢 = 1
- 🔴 = 0

$$\boxed{\boldsymbol{i}[t] = \sigma\left(\boldsymbol{W}_i \begin{bmatrix} \boldsymbol{x}[t] \\ \boldsymbol{h}[t-1] \end{bmatrix} + \boldsymbol{b}_i\right)}$$

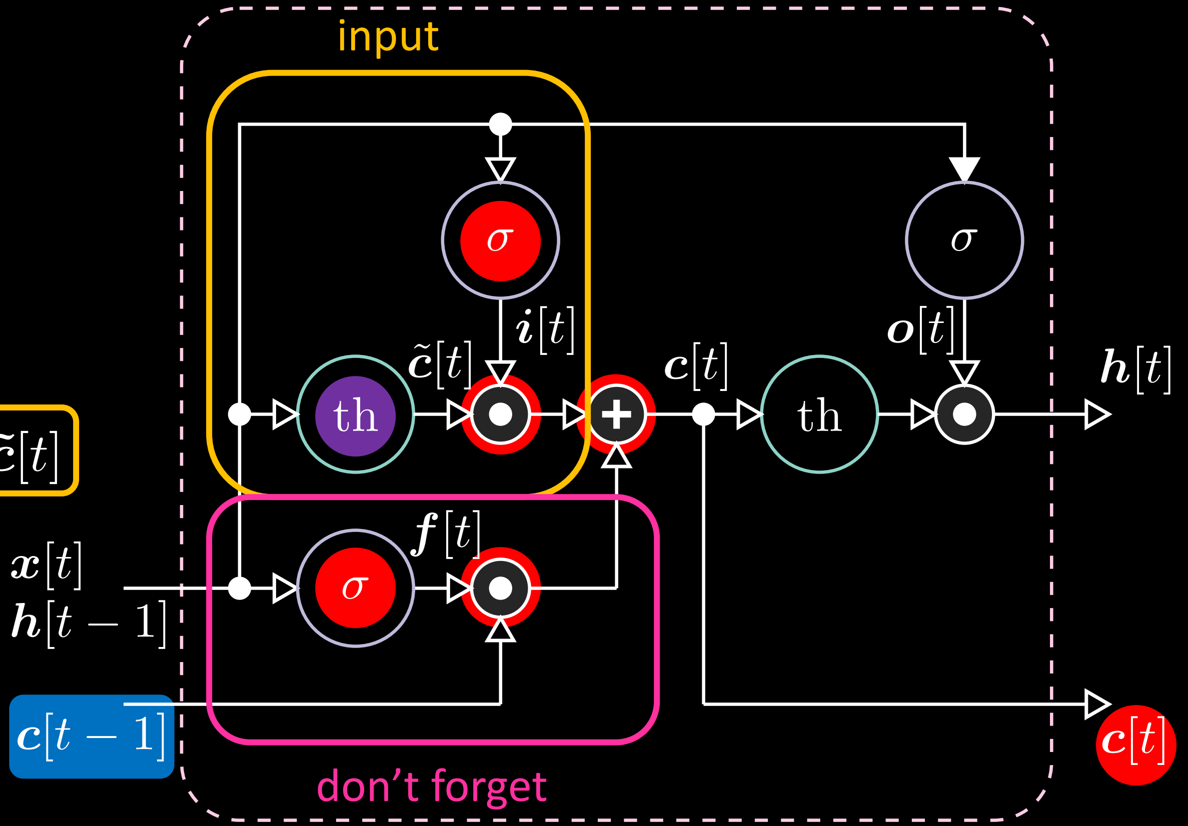
$$\boxed{\boldsymbol{f}[t] = \sigma\left(\boldsymbol{W}_f \begin{bmatrix} \boldsymbol{x}[t] \\ \boldsymbol{h}[t-1] \end{bmatrix} + \boldsymbol{b}_f\right)}$$

$$\boldsymbol{o}[t] = \sigma\left(\boldsymbol{W}_o \begin{bmatrix} \boldsymbol{x}[t] \\ \boldsymbol{h}[t-1] \end{bmatrix} + \boldsymbol{b}_o\right)$$

$$\tilde{\boldsymbol{c}}[t] = \tanh\left(\boldsymbol{W}_c \begin{bmatrix} \boldsymbol{x}[t] \\ \boldsymbol{h}[t-1] \end{bmatrix} + \boldsymbol{b}_c\right)$$

$$\boldsymbol{c}[t] = \boxed{\boldsymbol{f}[t] \odot \boldsymbol{c}[t-1]} + \boxed{\boldsymbol{i}[t] \odot \tilde{\boldsymbol{c}}[t]}$$

$$\boldsymbol{h}[t] = \boldsymbol{o}[t] \odot \tanh(\boldsymbol{c}[t])$$



Controlling the memory - keep

Saturated sigmoid
- 🟢 = 1
- 🔴 = 0

$$\boxed{\boldsymbol{i}[t] = \sigma\left(\boldsymbol{W}_i \begin{bmatrix}\boldsymbol{x}[t]\\ \boldsymbol{h}[t-1]\end{bmatrix} + \boldsymbol{b}_i\right)}$$

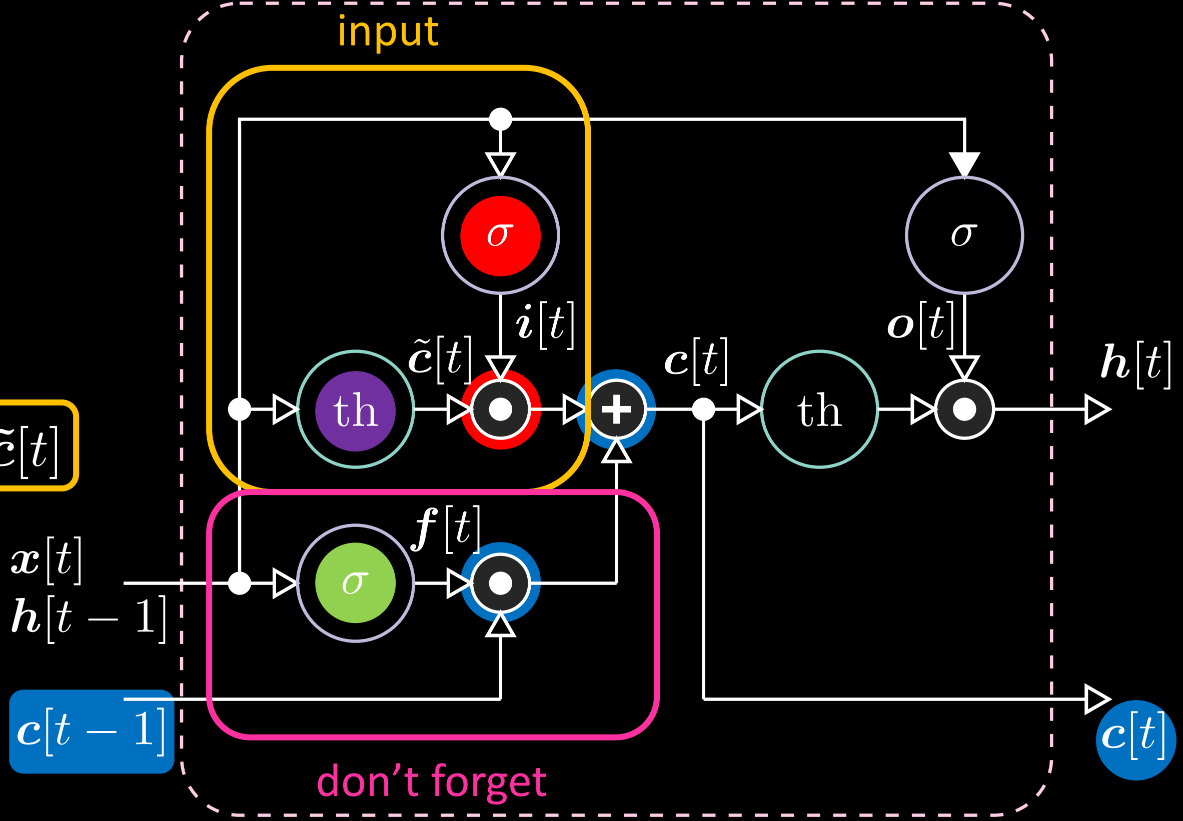
$$\boxed{\boldsymbol{f}[t] = \sigma\left(\boldsymbol{W}_f \begin{bmatrix}\boldsymbol{x}[t]\\ \boldsymbol{h}[t-1]\end{bmatrix} + \boldsymbol{b}_f\right)}$$

$$\boldsymbol{o}[t] = \sigma\left(\boldsymbol{W}_o \begin{bmatrix}\boldsymbol{x}[t]\\ \boldsymbol{h}[t-1]\end{bmatrix} + \boldsymbol{b}_o\right)$$

$$\tilde{\boldsymbol{c}}[t] = \tanh\left(\boldsymbol{W}_c \begin{bmatrix}\boldsymbol{x}[t]\\ \boldsymbol{h}[t-1]\end{bmatrix} + \boldsymbol{b}_c\right)$$

$$\boldsymbol{c}[t] = \boxed{\boldsymbol{f}[t] \odot \boldsymbol{c}[t-1]} + \boxed{\boldsymbol{i}[t] \odot \tilde{\boldsymbol{c}}[t]}$$

$$\boldsymbol{h}[t] = \boldsymbol{o}[t] \odot \tanh(\boldsymbol{c}[t])$$



Controlling the $\boxed{\text{memory}}$ - write

Saturated sigmoid
- 🟢 = 1
- 🔴 = 0

$$\boxed{\boldsymbol{i}[t] = \sigma\left(\boldsymbol{W}_i \begin{bmatrix}\boldsymbol{x}[t]\\ \boldsymbol{h}[t-1]\end{bmatrix} + \boldsymbol{b}_i\right)}$$

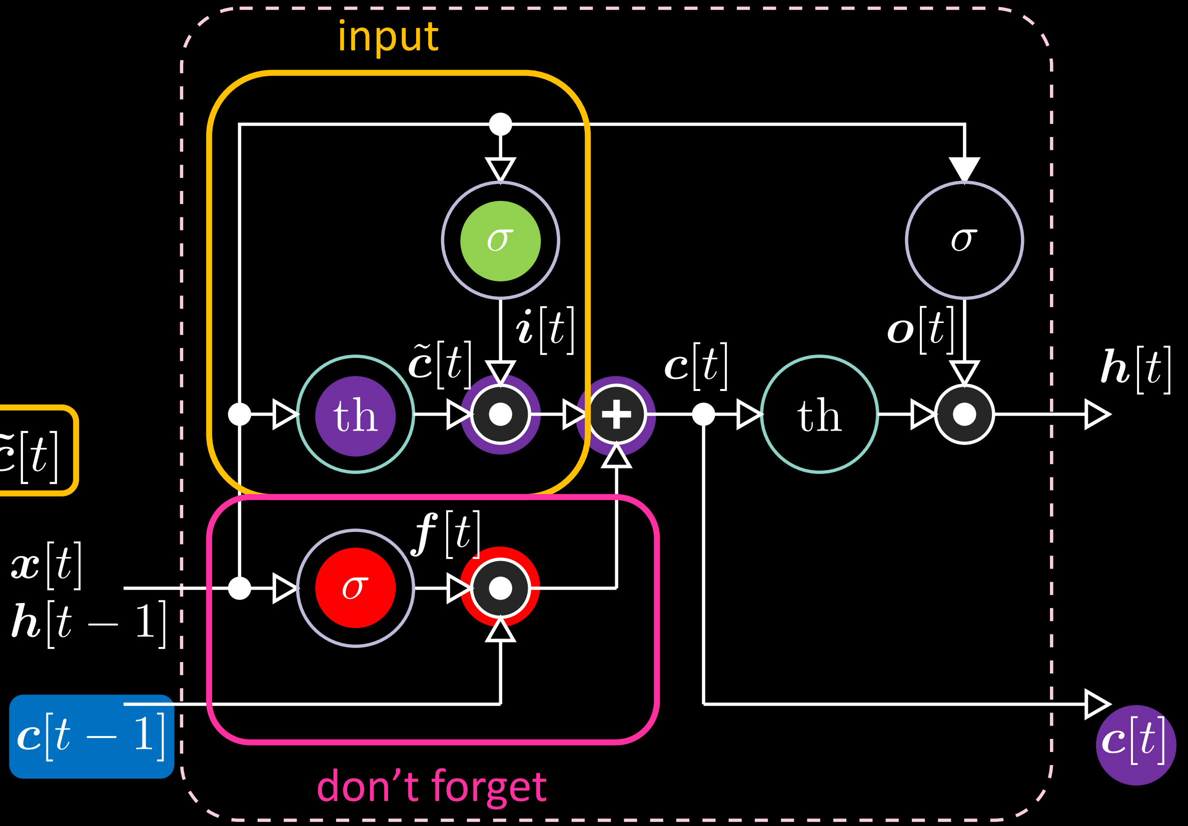
$$\boxed{\boldsymbol{f}[t] = \sigma\left(\boldsymbol{W}_f \begin{bmatrix}\boldsymbol{x}[t]\\ \boldsymbol{h}[t-1]\end{bmatrix} + \boldsymbol{b}_f\right)}$$

$$\boldsymbol{o}[t] = \sigma\left(\boldsymbol{W}_o \begin{bmatrix}\boldsymbol{x}[t]\\ \boldsymbol{h}[t-1]\end{bmatrix} + \boldsymbol{b}_o\right)$$

$$\tilde{\boldsymbol{c}}[t] = \tanh\left(\boldsymbol{W}_c \begin{bmatrix}\boldsymbol{x}[t]\\ \boldsymbol{h}[t-1]\end{bmatrix} + \boldsymbol{b}_c\right)$$

$$\boldsymbol{c}[t] = \boxed{\boldsymbol{f}[t] \odot \boldsymbol{c}[t-1]} + \boxed{\boldsymbol{i}[t] \odot \tilde{\boldsymbol{c}}[t]}$$

$$\boldsymbol{h}[t] = \boldsymbol{o}[t] \odot \tanh(\boldsymbol{c}[t])$$



input
σ
σ
$\boldsymbol{i}[t]$
$\tilde{\boldsymbol{c}}[t]$
$\boldsymbol{o}[t]$
th
$\boldsymbol{c}[t]$
th
$\boldsymbol{h}[t]$
$\boldsymbol{x}[t]$
$\boldsymbol{h}[t-1]$
σ
$\boldsymbol{f}[t]$
$\boldsymbol{c}[t-1]$
$\boldsymbol{c}[t]$
don't forget