RNN&LSTM

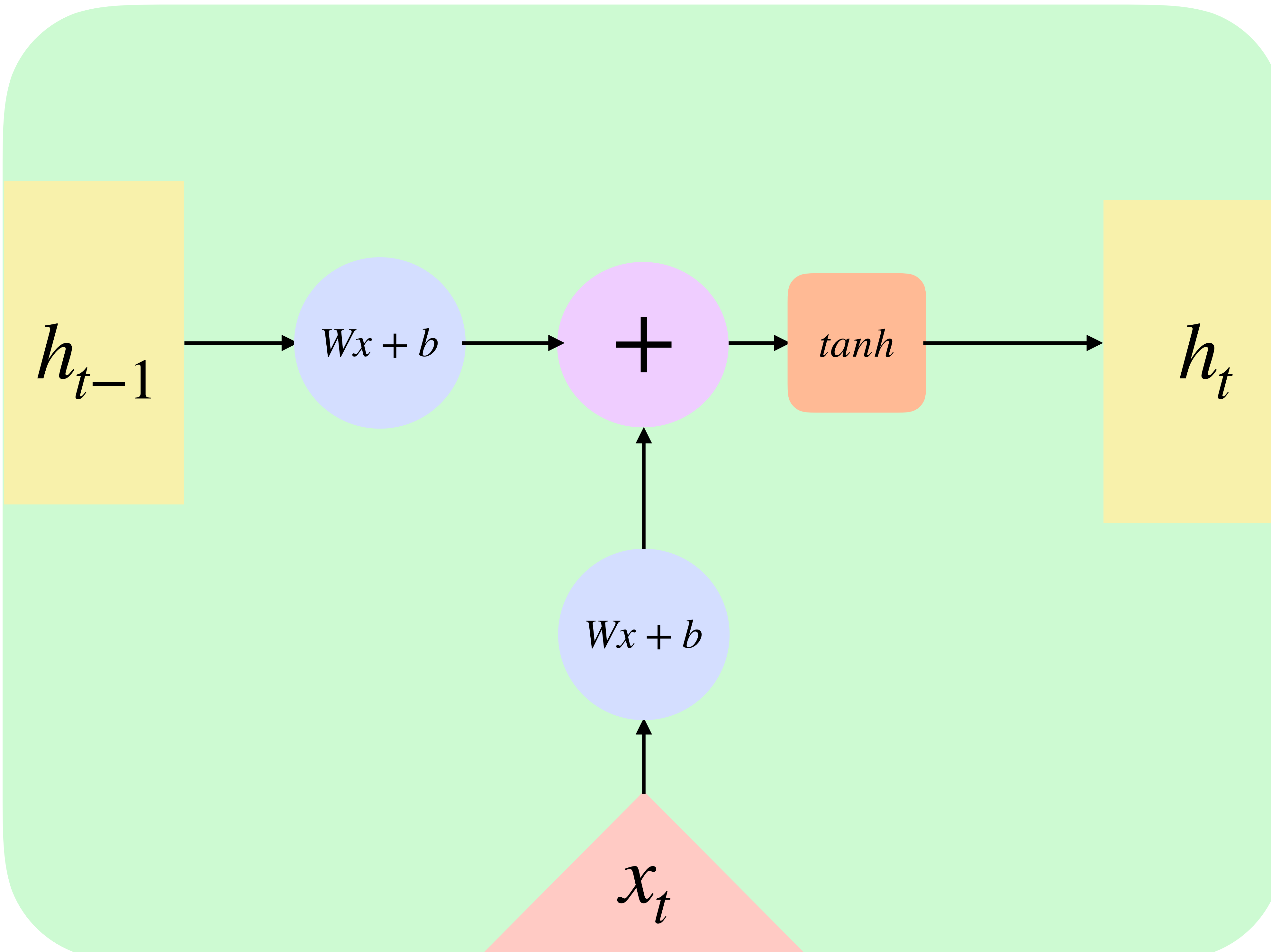
Troshin Sergei
Student at HSE

2018

Plot

- Limitations of Vanilla RNN
  - RNN - hard to train
  - Problems with long-term dependences
  - 1-bit problem
- From RNN to LSTM
  - residual connections
  - LSTM in detail
  - GRU
- Tips&Tricks
  - Gradients Clipping
  - Layer Norm
  - Batch Norm
  - Dropout
- Modern Architectures
  - Image Captoning
  - Seq2seq
  - Attention

RNN step

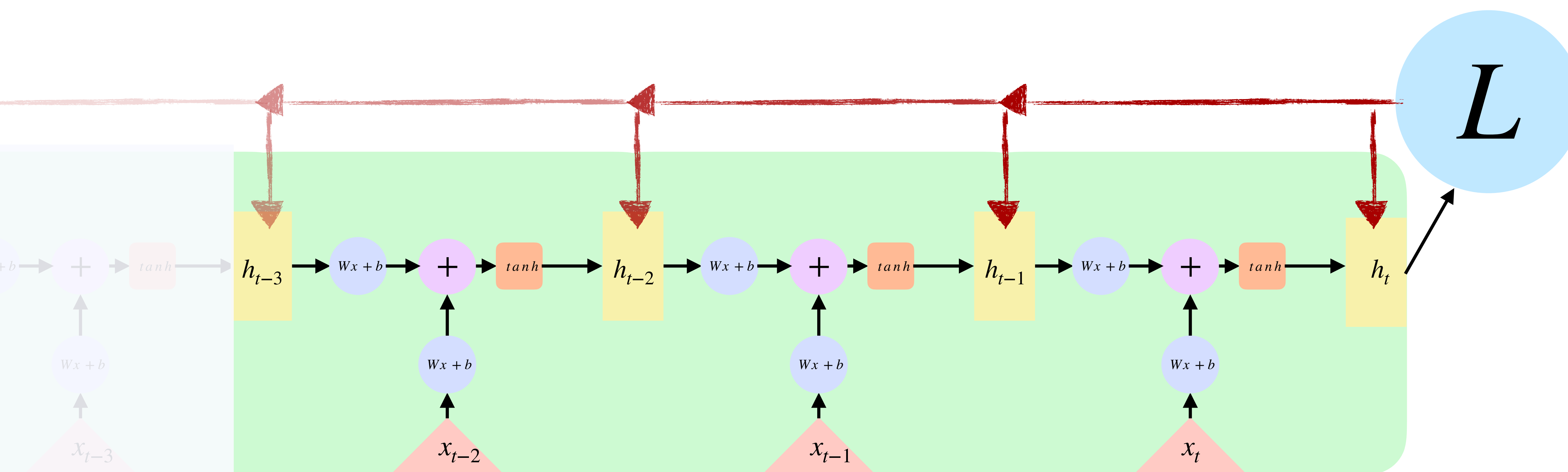


Reliable storing - Vanishing gradients.

$$L = L(h_t(h_{t-1}(\ldots h_{\tau+1}(h_\tau))))$$

$$\frac{\partial L}{\partial h_\tau} = \frac{\partial L}{\partial h_t} \cdot \frac{\partial h_t}{\partial h_{t-1}} \cdot \frac{\partial h_{t-1}}{\partial h_{t-2}} \cdot \ldots \cdot \frac{\partial h_{\tau+1}}{\partial h_\tau}$$

if spectral radius (maximal eigen value) > 1 than propagated gradient vanish

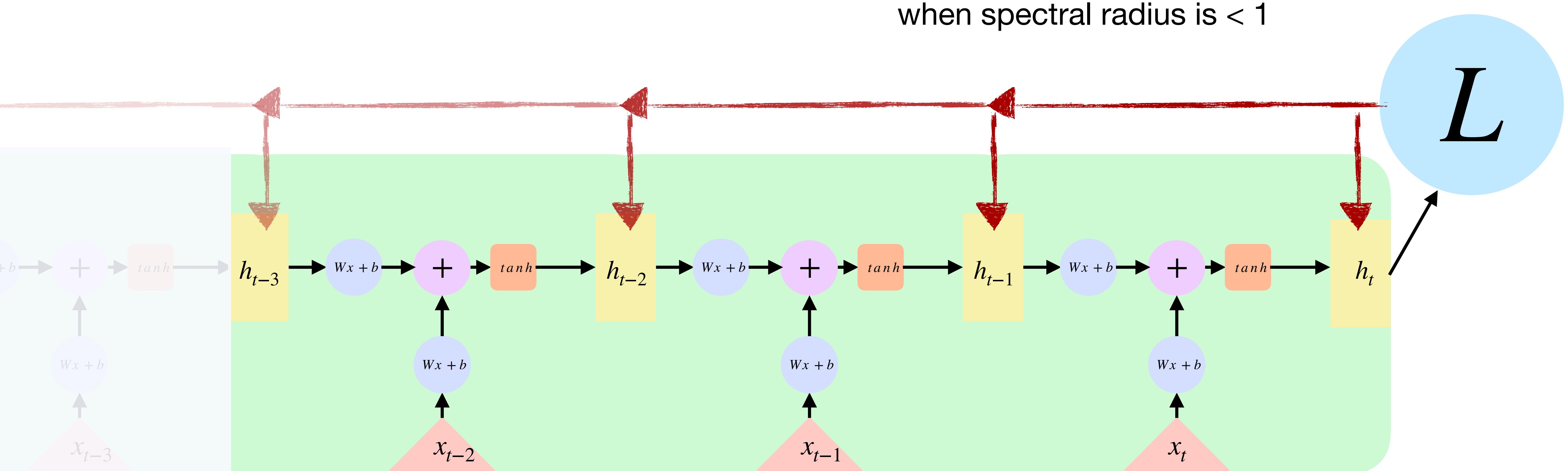


Why this is a problem

- long-term dependences correspond to updating the state with an exponentially smaller weight than short-term dependences.

$$\frac{\partial L}{\partial W} = \sum_{\tau \leq t} \frac{\partial L}{\partial h_\tau} \cdot \frac{\partial h_\tau}{\partial W} = \sum_{\tau \leq t} \frac{\partial L}{\partial h_t} \cdot \boxed{\frac{\partial h_t}{\partial h_\tau}} \cdot \frac{\partial h_\tau}{\partial W}$$

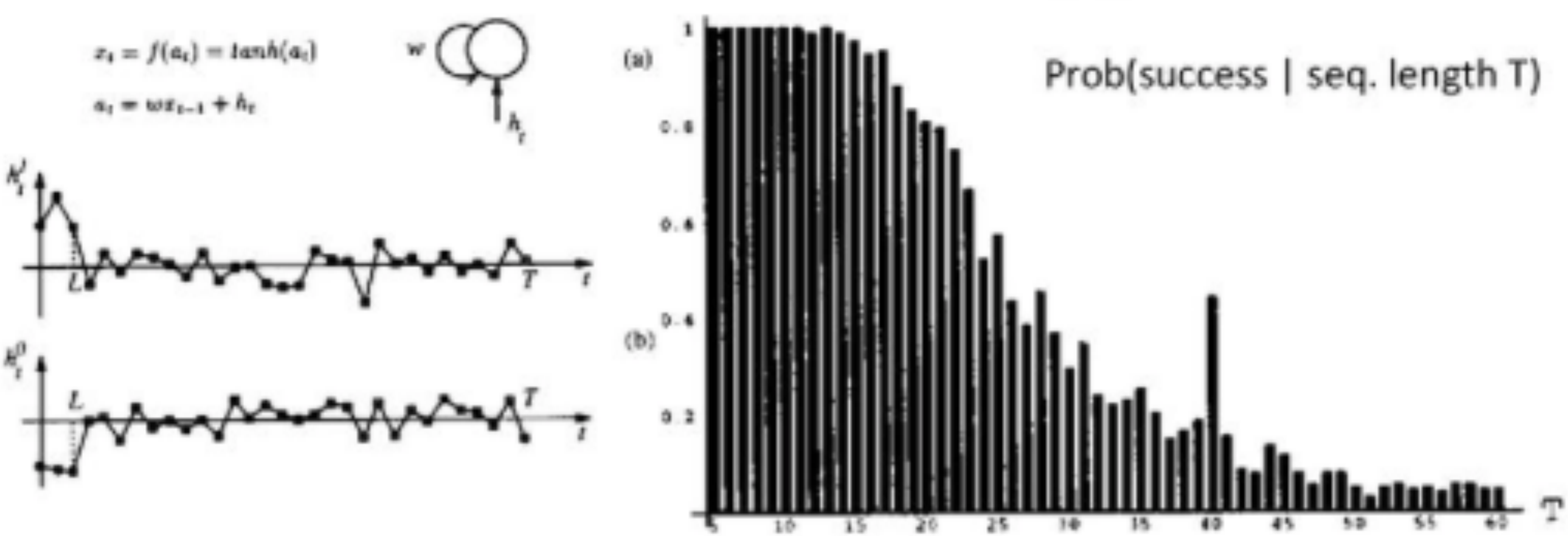
tends to vanish exponentially
for long time dependences
when spectral radius is < 1



Experiment from 1991

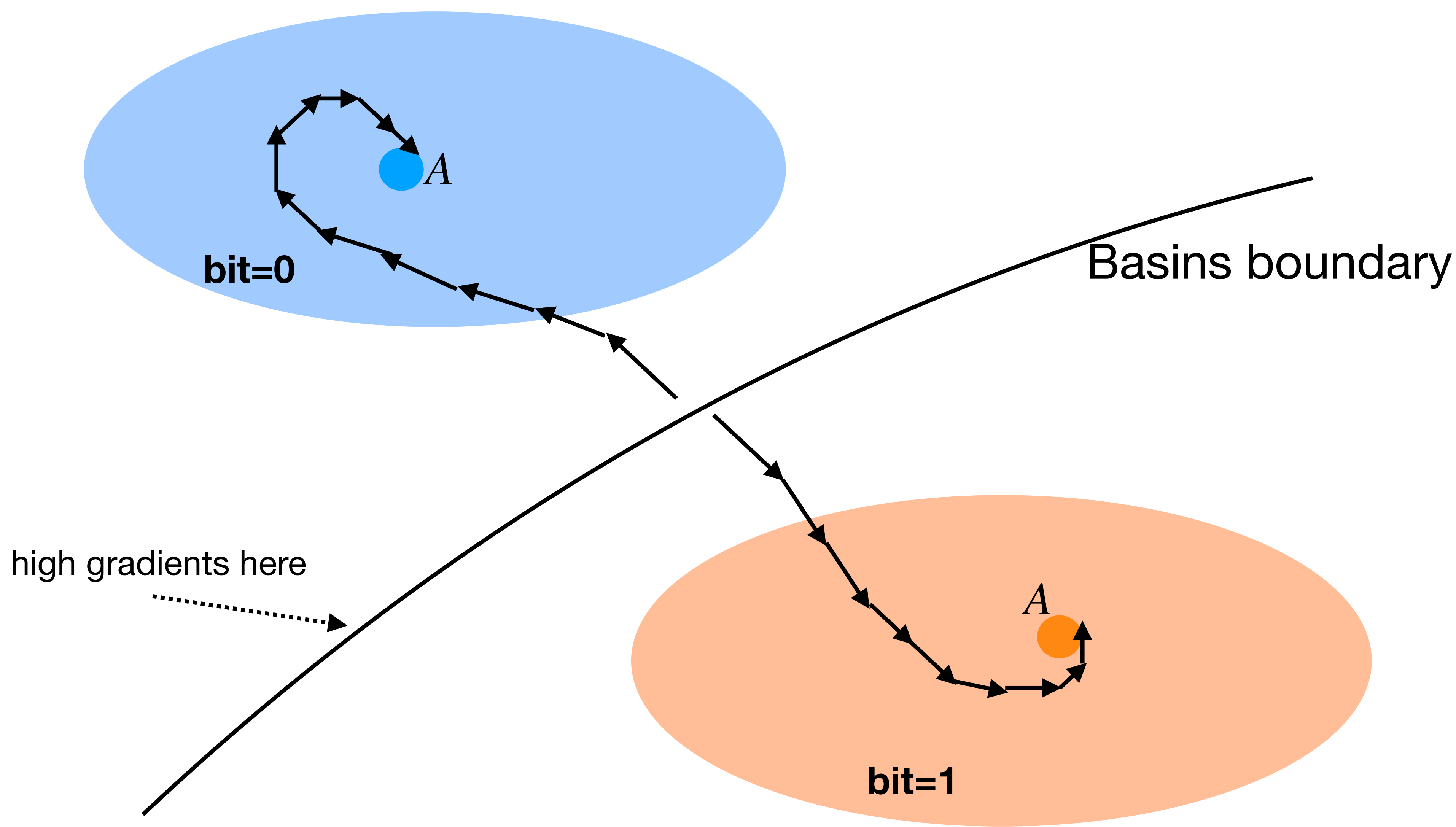
2 categories of sequences

Can the single tanh unit learn to store for T time steps 1 bit of information given by the sign of initial input?



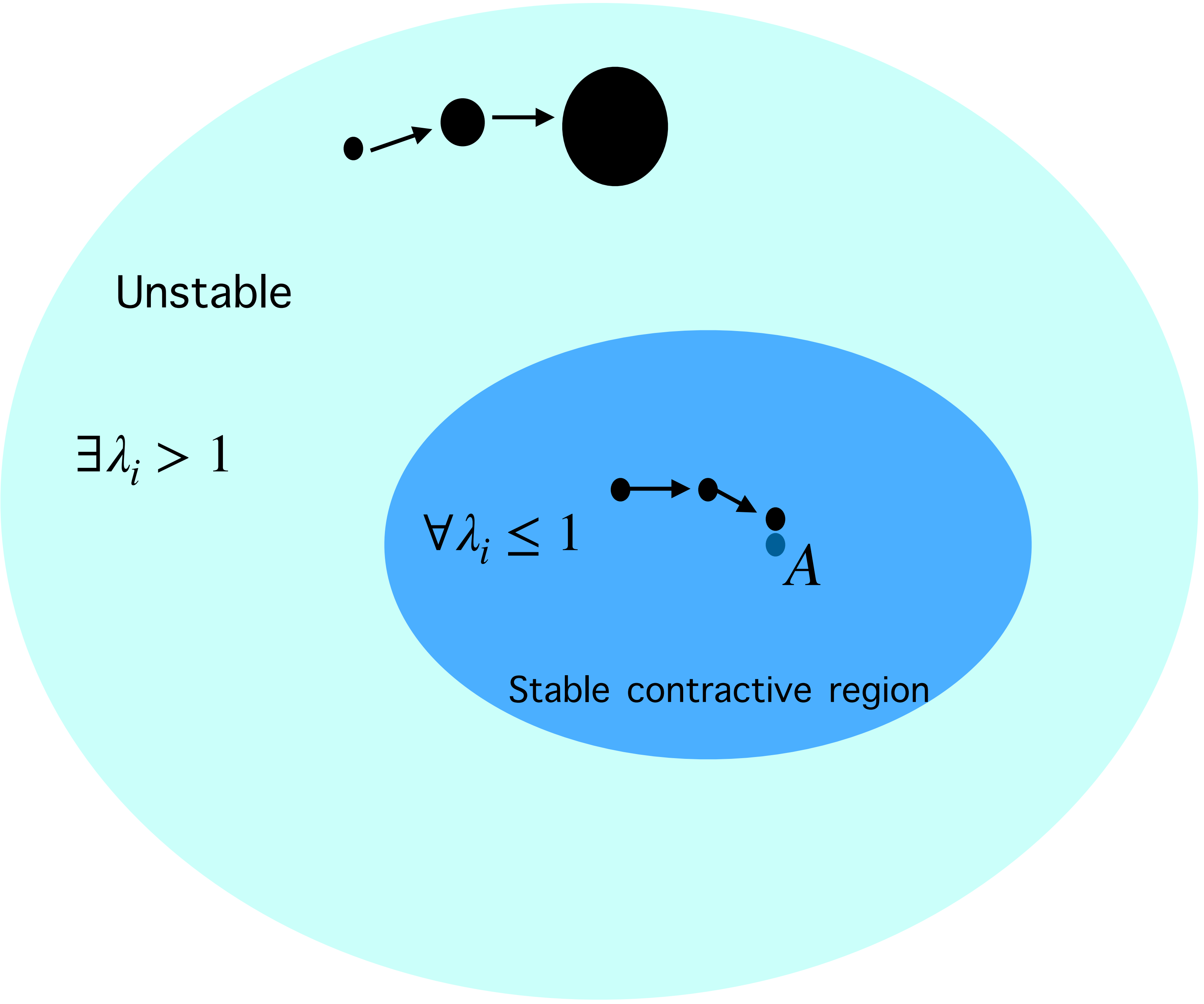
How to store 1 bit?

Some subspace of the state can store a bit (or more) of information if the dynamic system has basins of attraction in some dimensions.



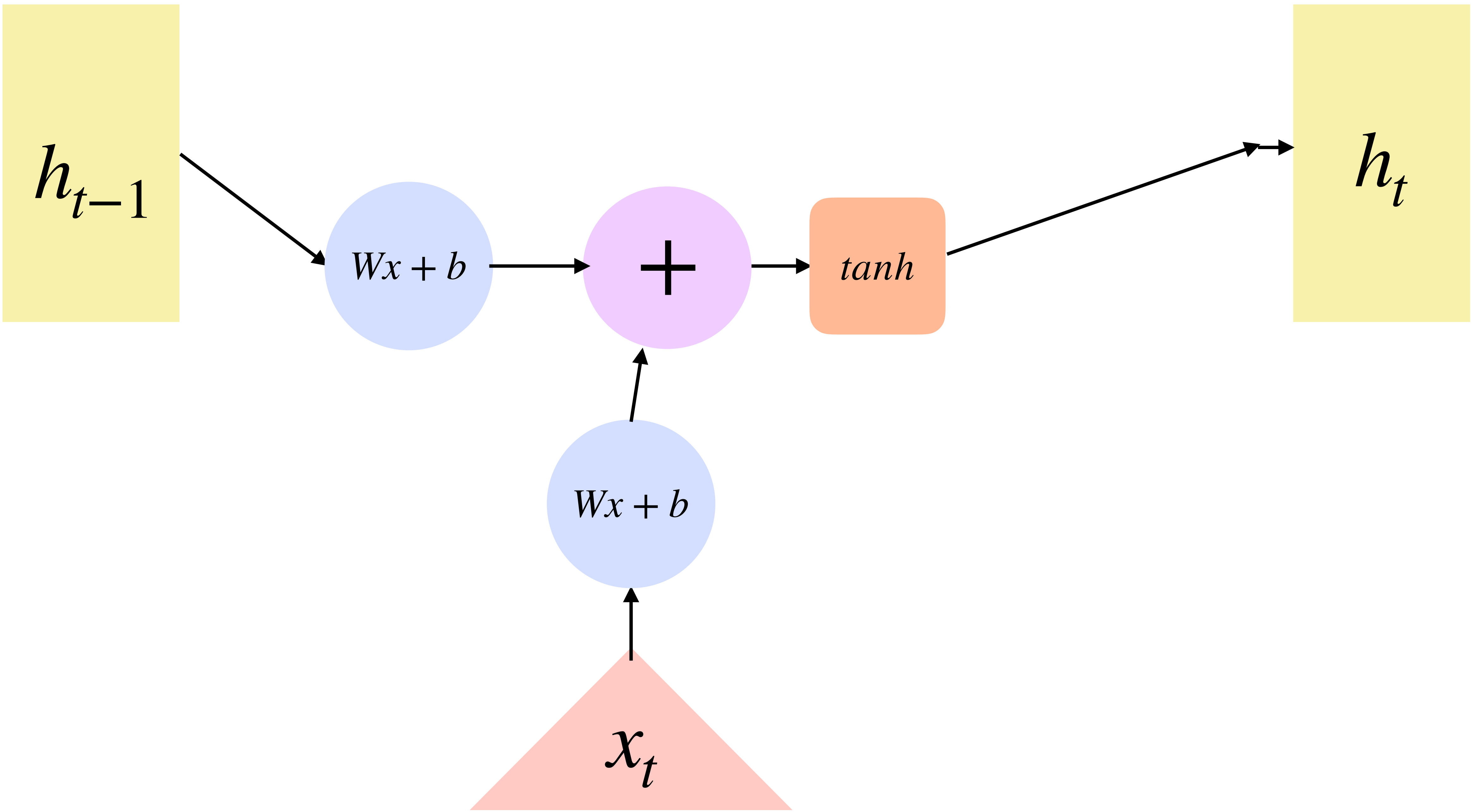
Contractive transformation.

With the spectral radius greater that 1 noise can kick the state out of the attractor. That means we are not likely to store information for long.

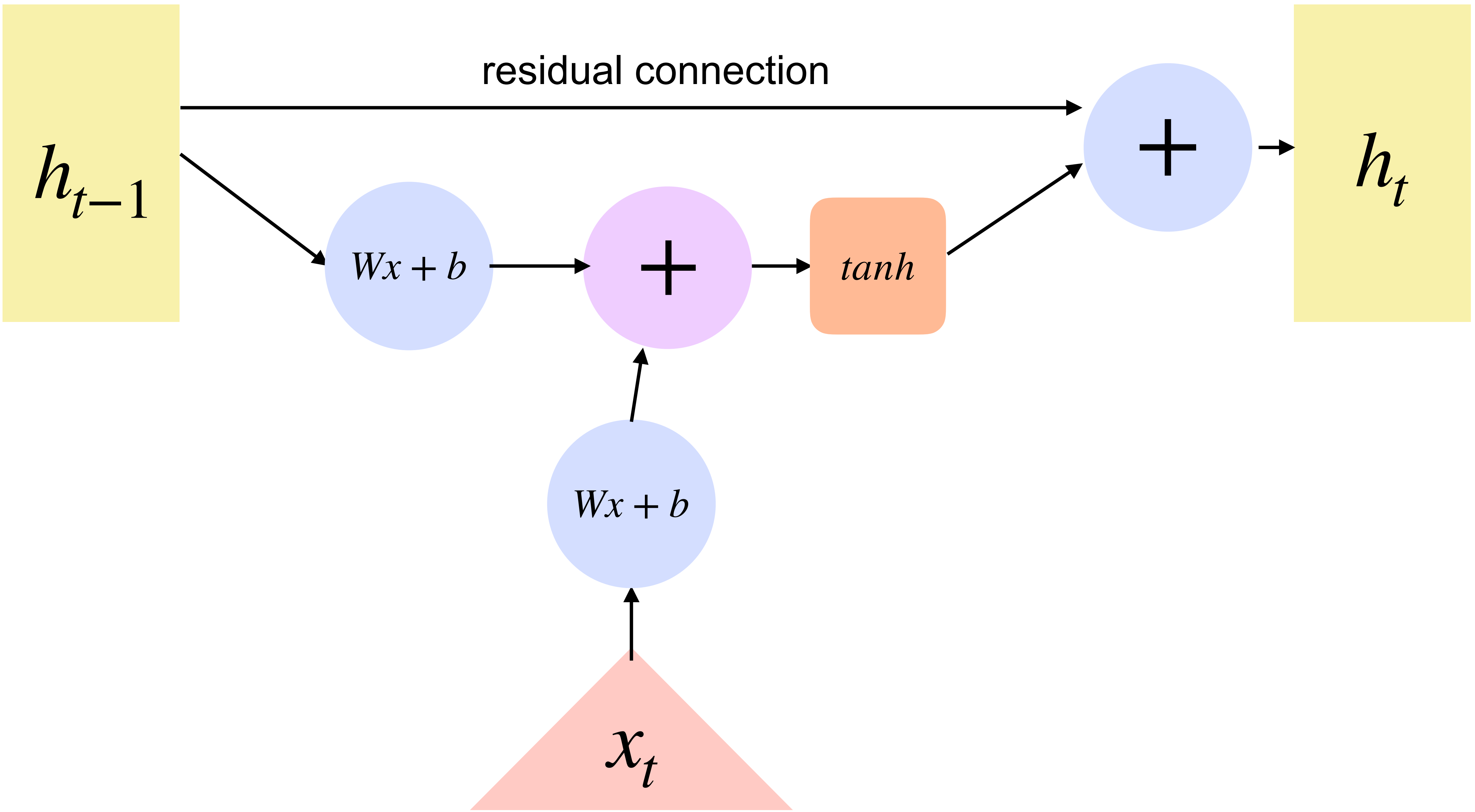


http://ai.dinfo.unifi.it/paolo/ps/tnn-94-gradient.pdf

RNN step



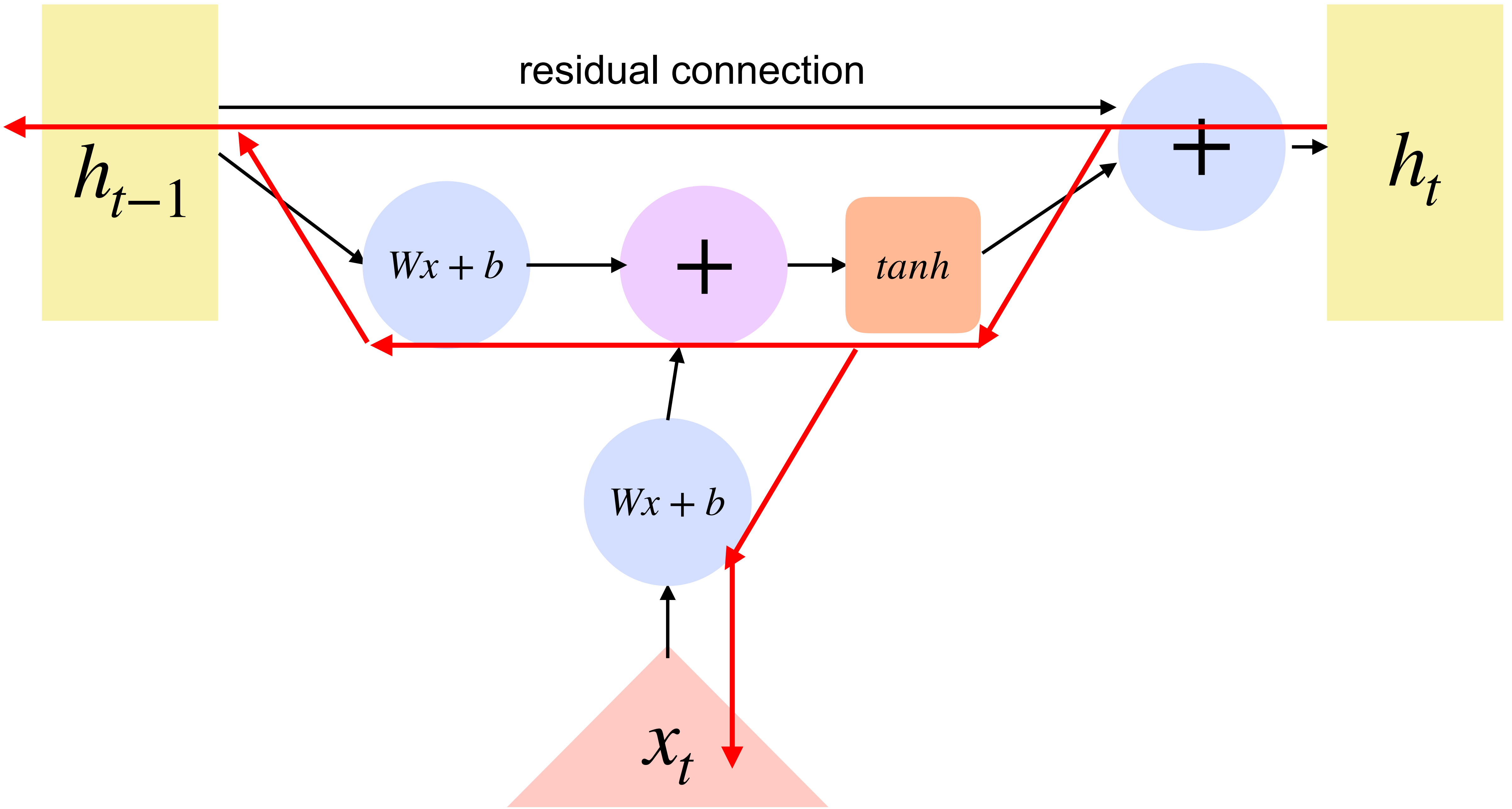
RNN step



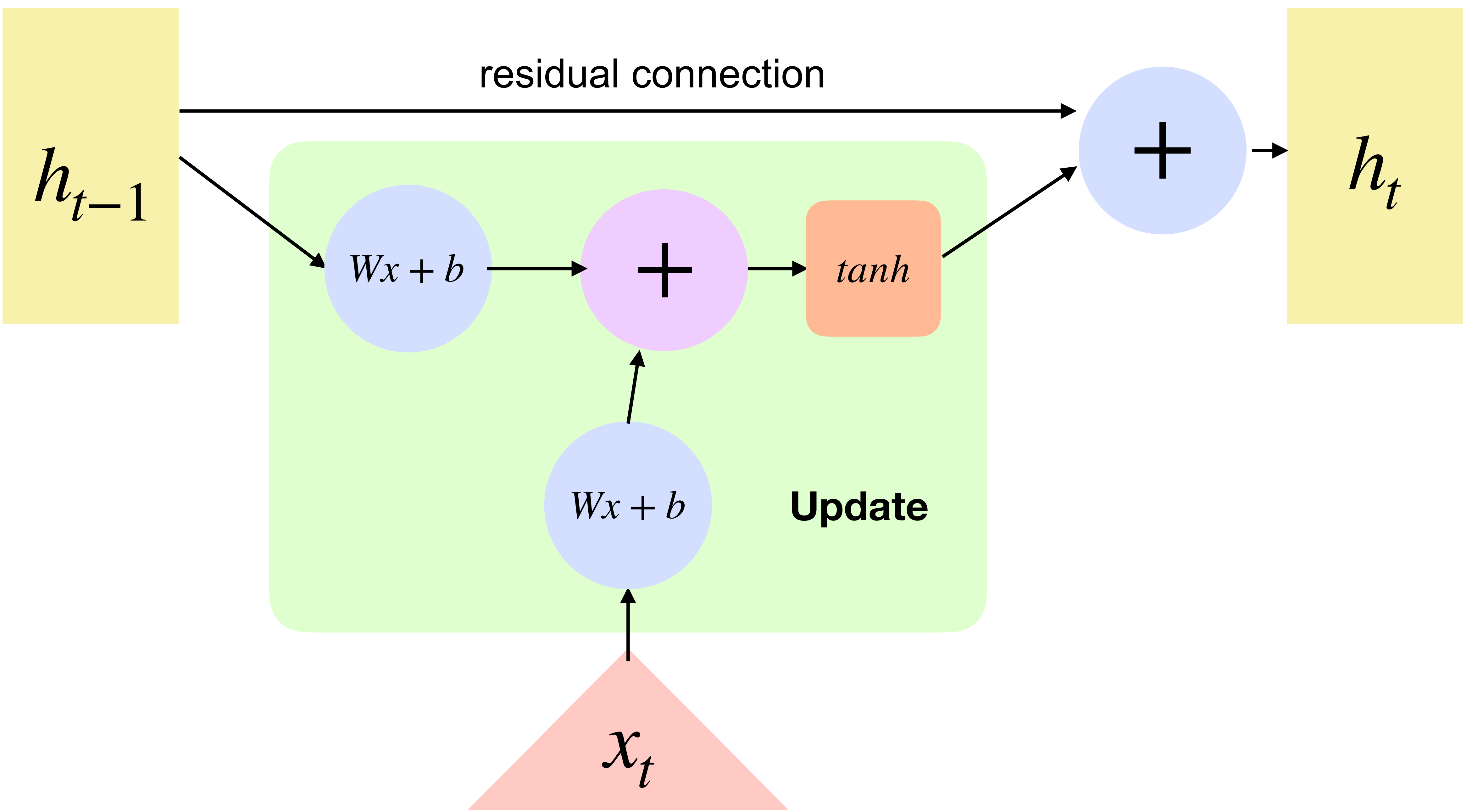
RNN step

$$\frac{\partial h_t}{\partial h_{t-1}} = 1 + \ldots$$ gradients don't vanish

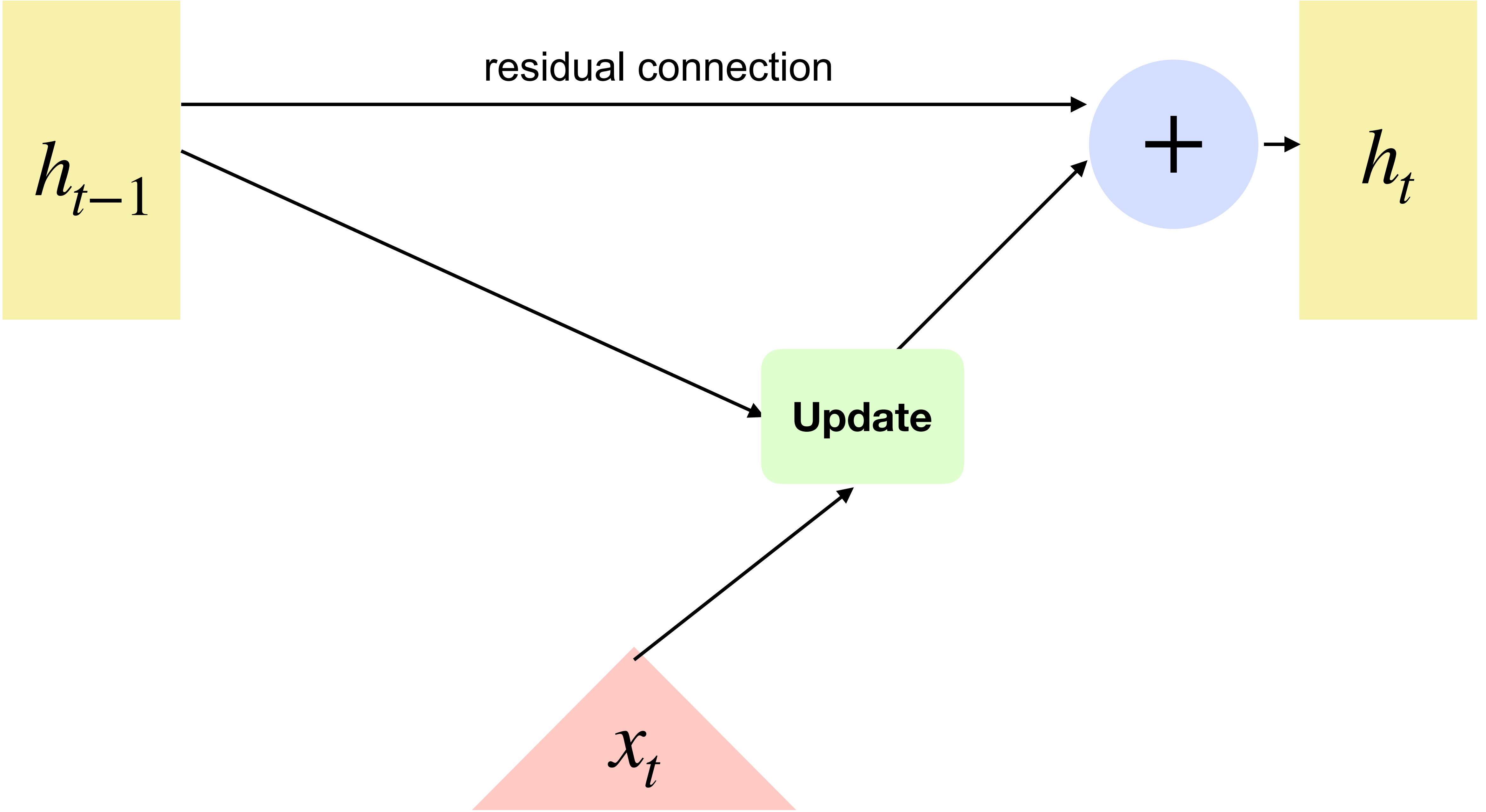
much harder to erase something



RNN step

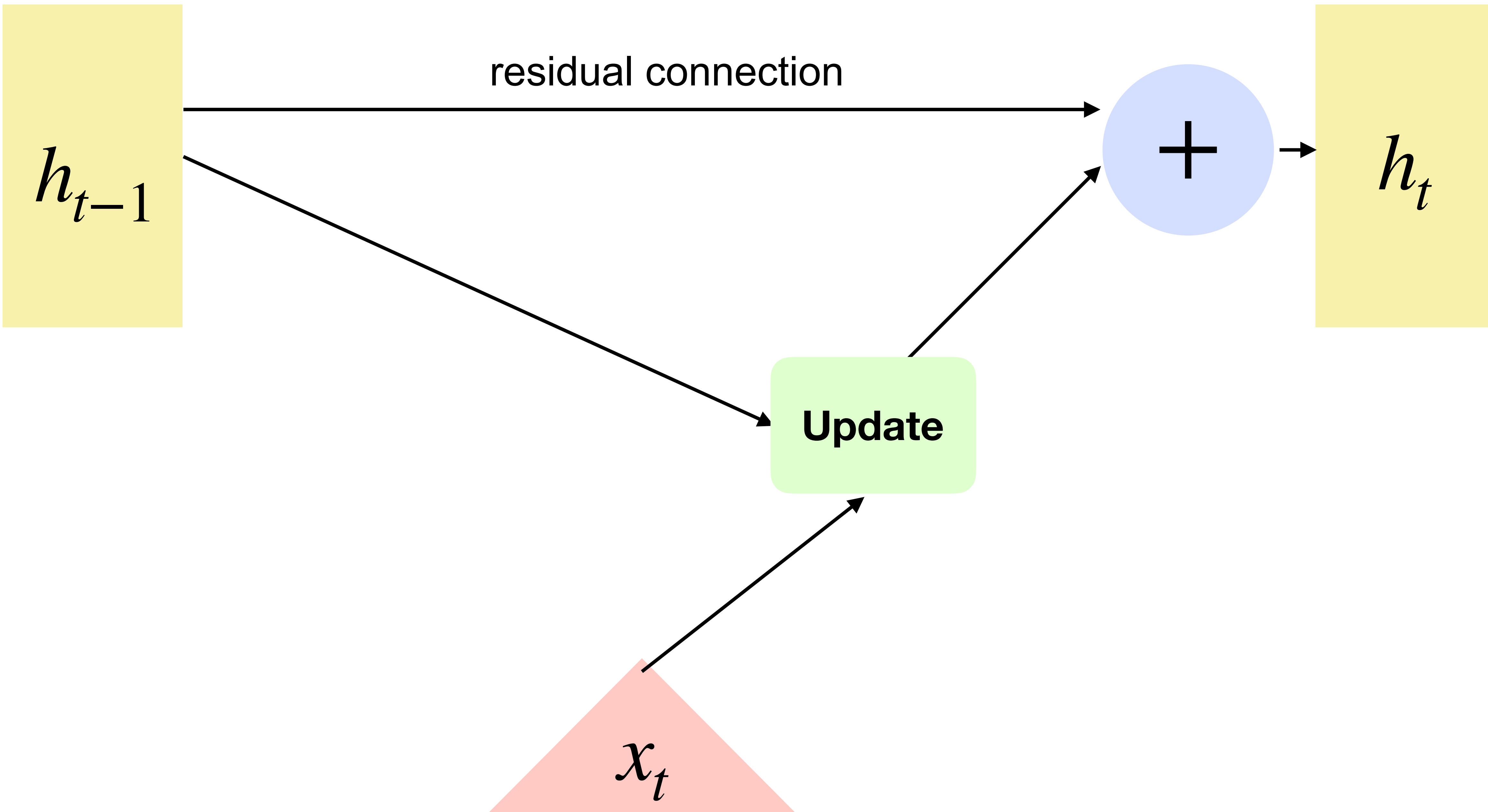


RNN step

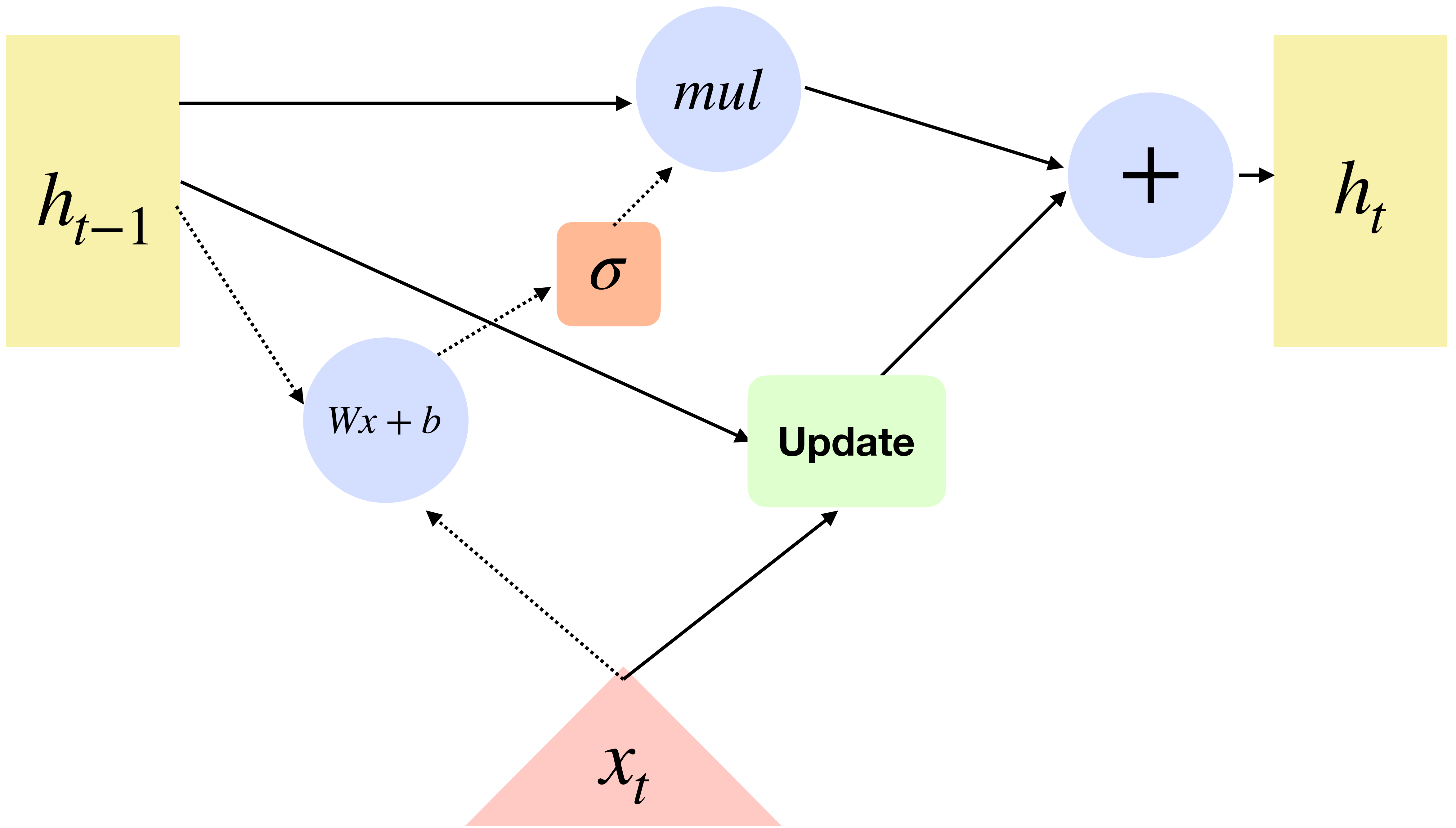


RNN step

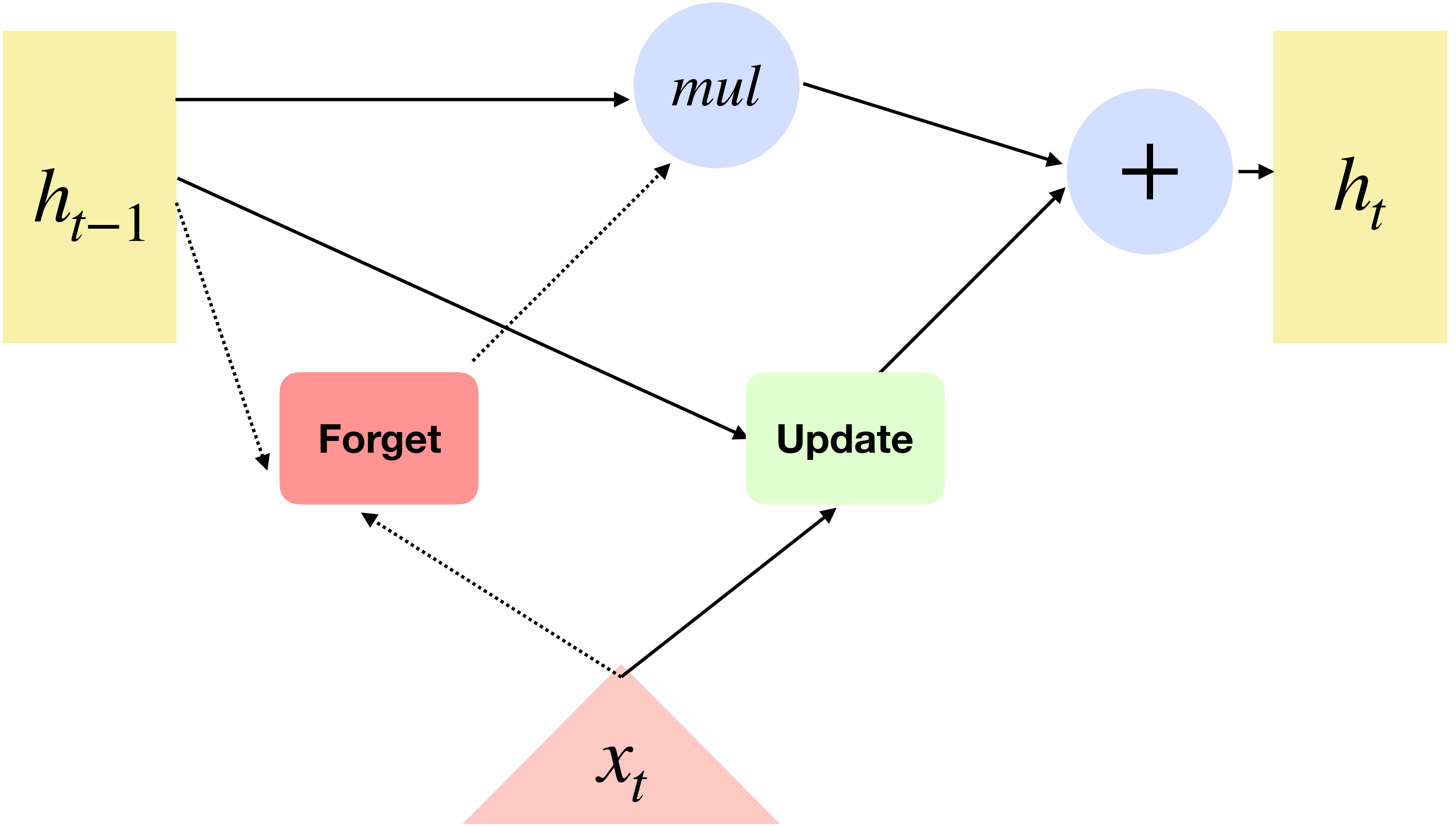
How can we learn to erase?



RNN step



RNN step

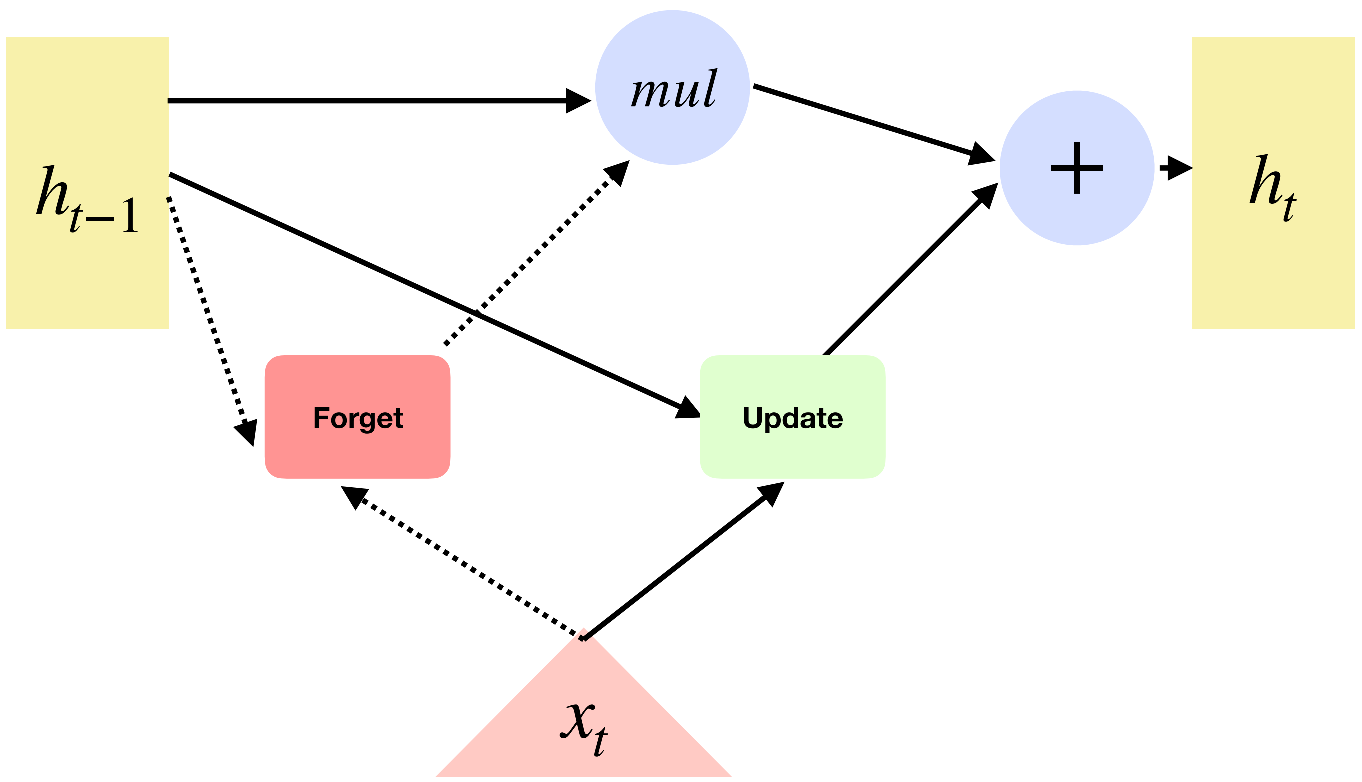


RNN step

$$update(x_i, h_{i-1}) = tanh(W_u^h \cdot h_{i-1} + W_u^i \cdot x_t + b_u)$$

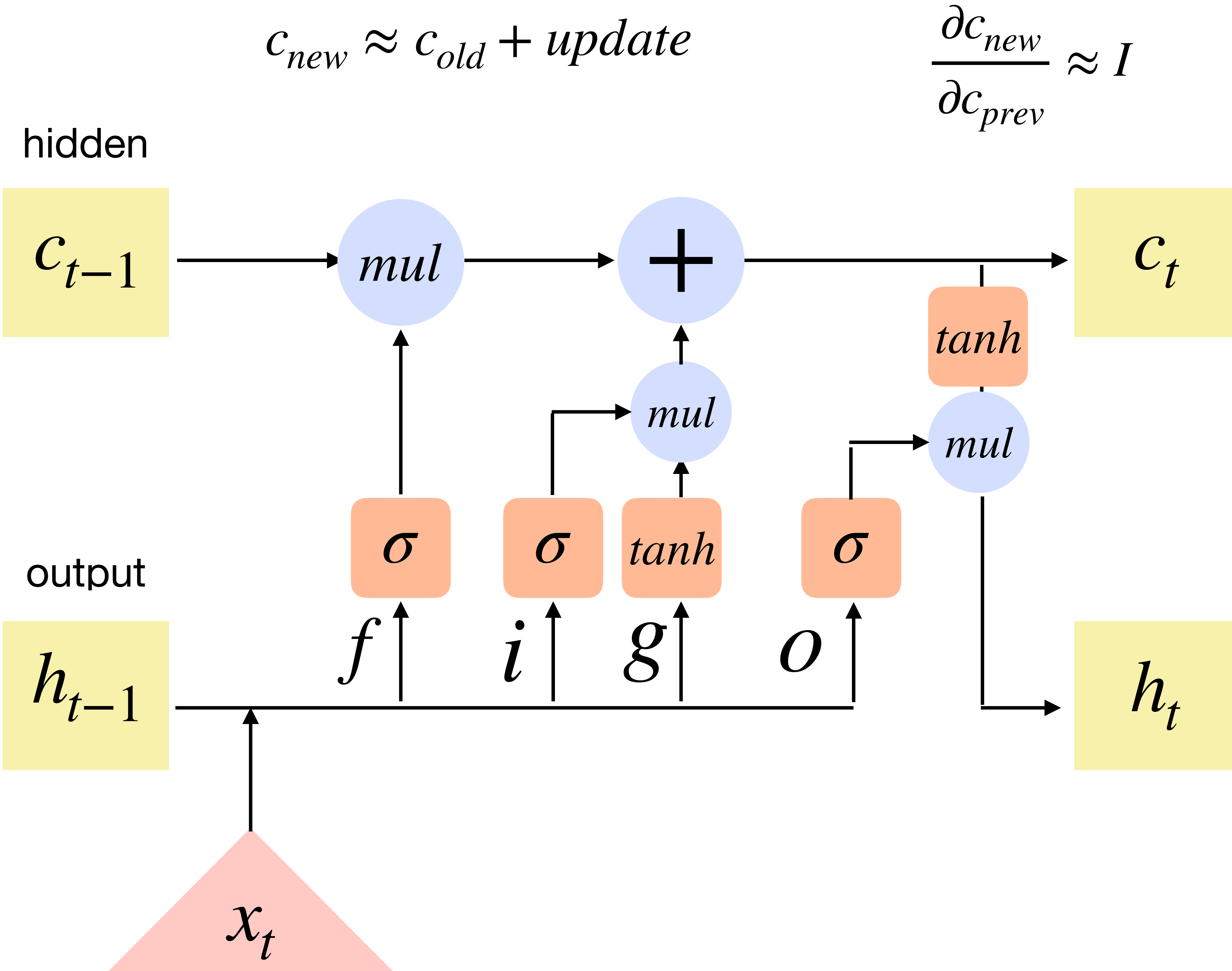
$$forget(x_i, h_{i-1}) = \sigma(W_f^h \cdot h_{i-1} + W_f^i \cdot x_t + b_f)$$

$$h_i(x_i, h_{i-1}) = forget \cdot h_{i-1} + update$$

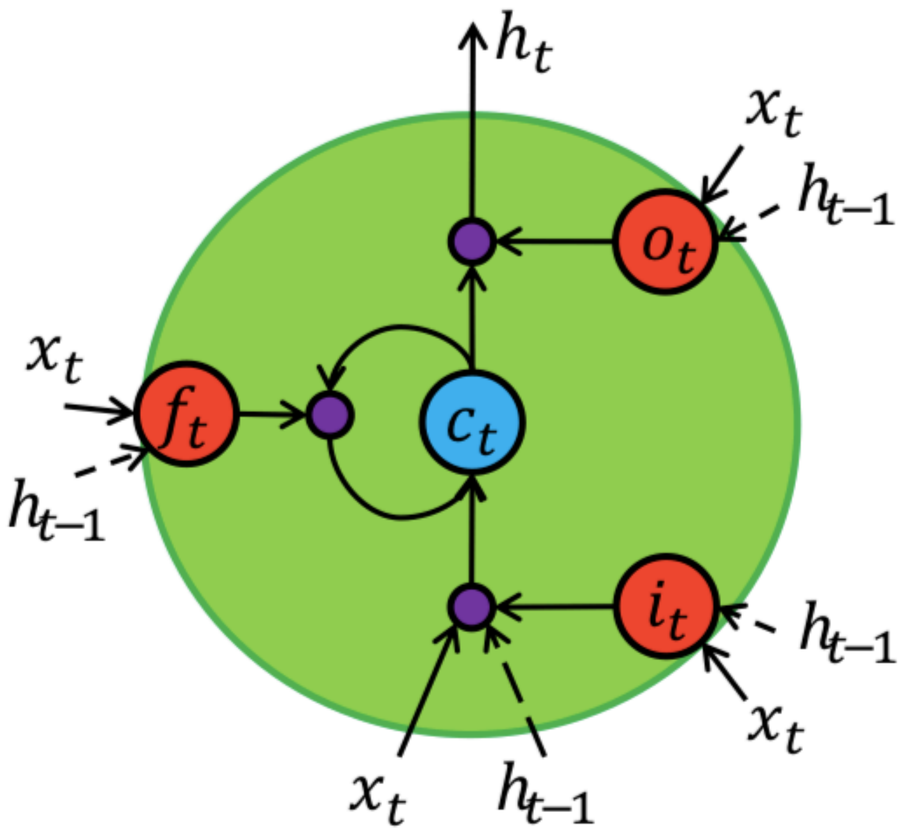


Okey, now we can learn things and forget them

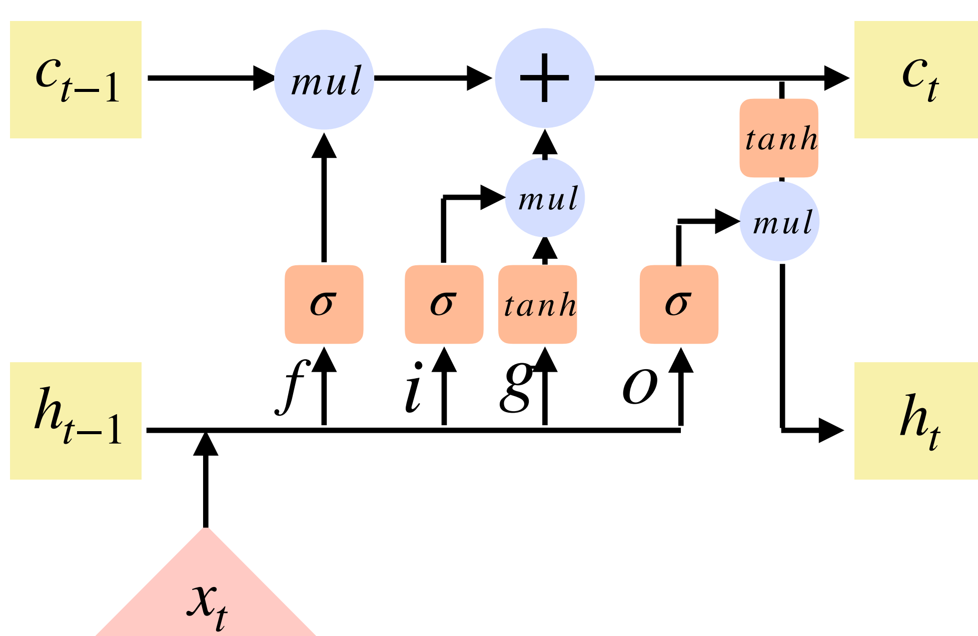
LSTM


$c_{new} \approx c_{old} + update$
$\frac{\partial c_{new}}{\partial c_{prev}} \approx I$
hidden
output
c_{t-1}
h_{t-1}
c_t
h_t
x_t
mul
+
mul
mul
tanh
tanh
σ
σ
σ
f
i
g
o

LSTM



[E. Lobacheva, 2016]



$$i_t = \sigma(V_i x_t + W_i h_{t-1} + b_i)$$

$$f_t = \sigma(V_f x_t + W_f h_{t-1} + b_f)$$

$$o_t = \sigma(V_o x_t + W_o h_{t-1} + b_o)$$

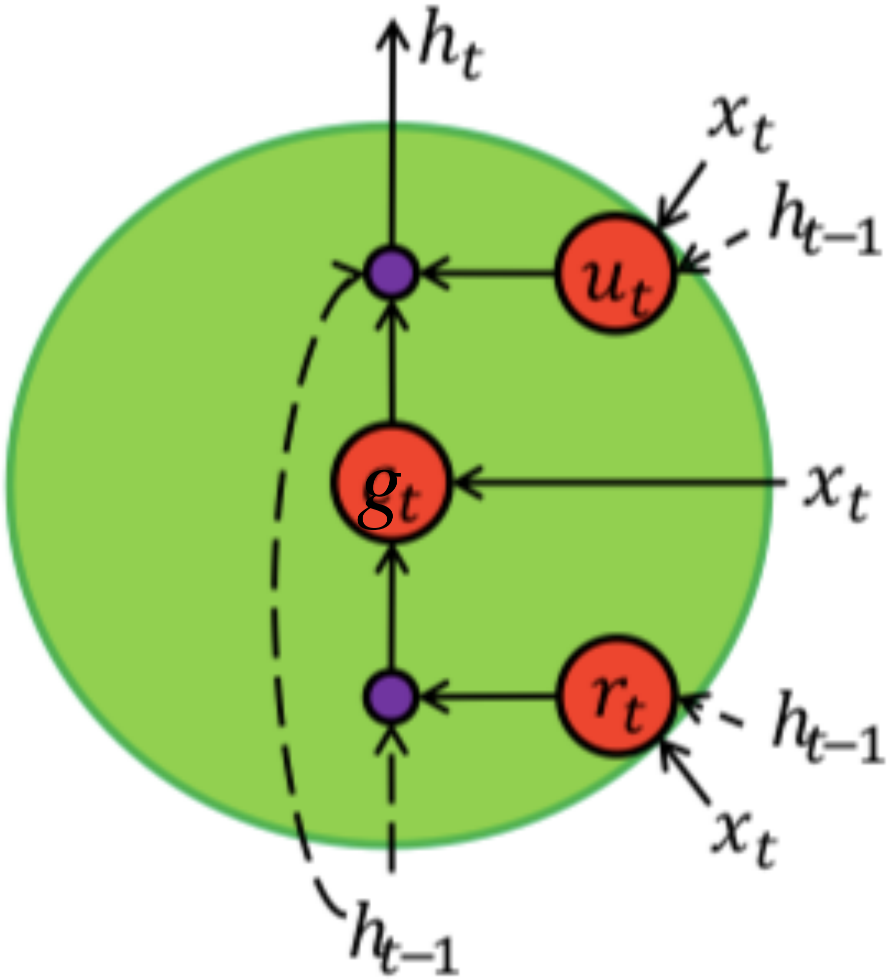
$$g_t = \tanh(V_g x_t + W_g h_{t-1} + b_g)$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot g_t$$ ← self-loop

$$h_t = o_t \cdot tanh(c_t)$$

- create a path where gradients can flow for longer time.
- corresponds to an eigenvectors of the Jacobian matrix slightly less that 1.
- $\frac{\partial c_t}{\partial c_{t-1}} = f_t \rightarrow$ high initial b_f

GRU



$$u_t = \sigma(V_u x_t + W_u h_{t-1} + b_u)$$

$$r_t = \sigma(V_r x_t + W_r h_{t-1} + b_r)$$

$$g_t = tanh(V_g x_t + W_g(h_{t-1} \cdot r_t) + b_g)$$

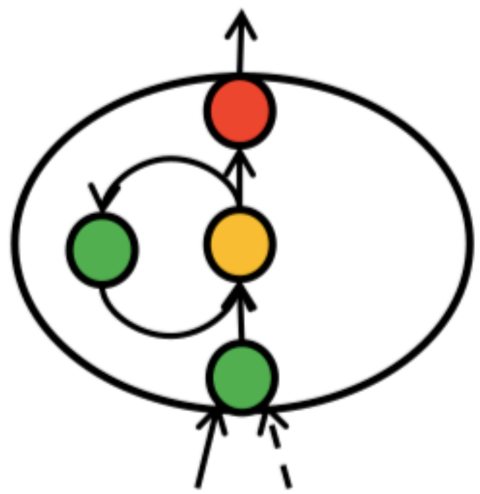
$$h_t = (1 - u_t) \cdot g_t + u_t \cdot h_{t-1}$$

$$\frac{\partial h_t}{\partial h_{t-1}} = u_h + (1 - u_h) \cdot \frac{\partial g_h}{\partial h_{h-1}}$$ ➡ High initial b_u

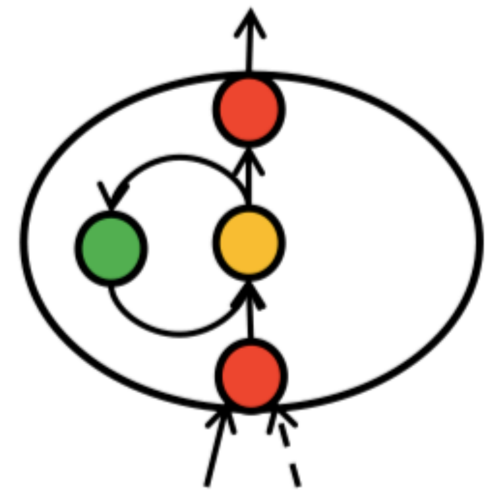
LSTM

Examples

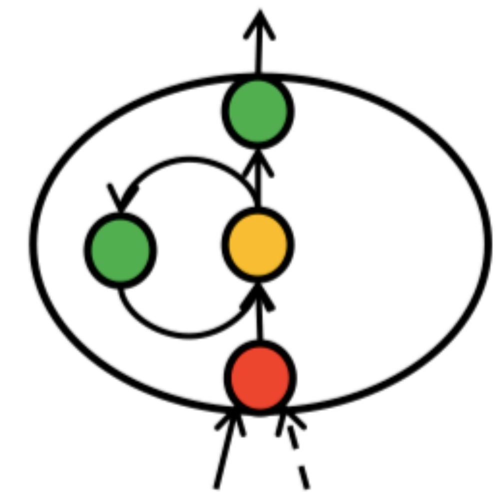
Captures info



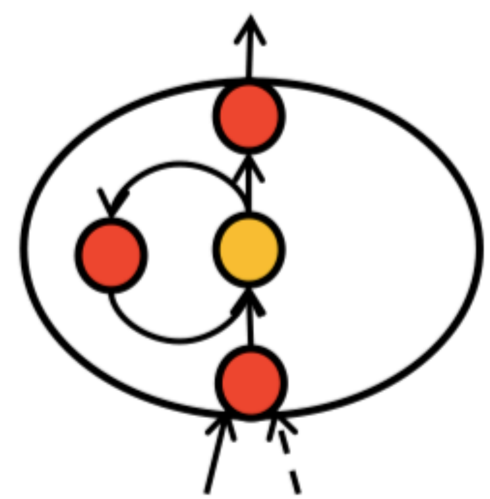
Keeps info



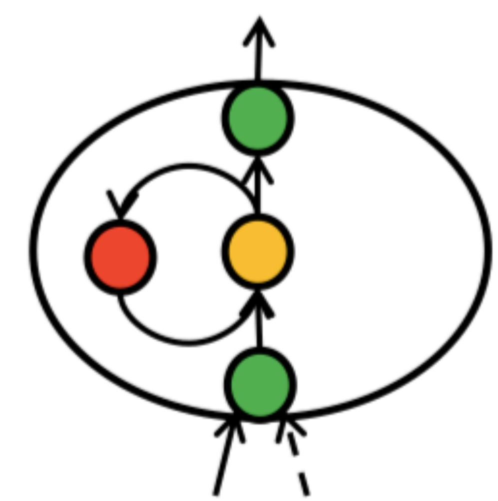
Releases info



Erases info



= RNN



 - gate is close

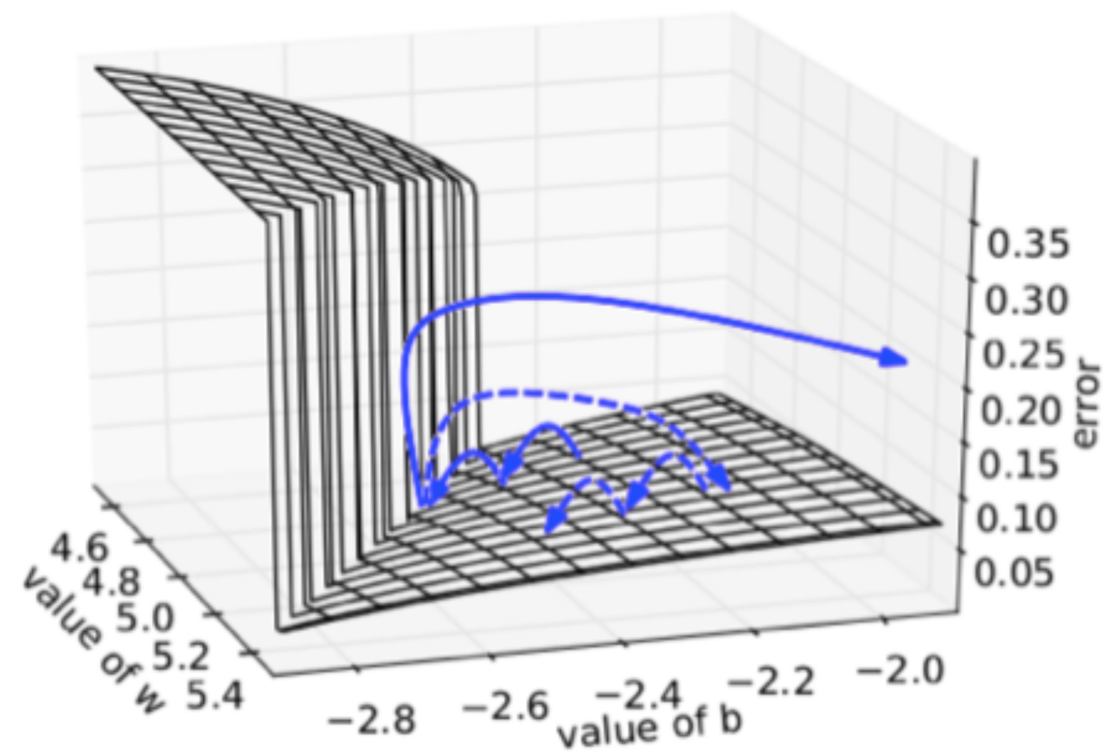
 - gate is open

[pictures: E. Lobacheva, D. Vetrov]

Training Tips&Tricks

Gradient Clipping

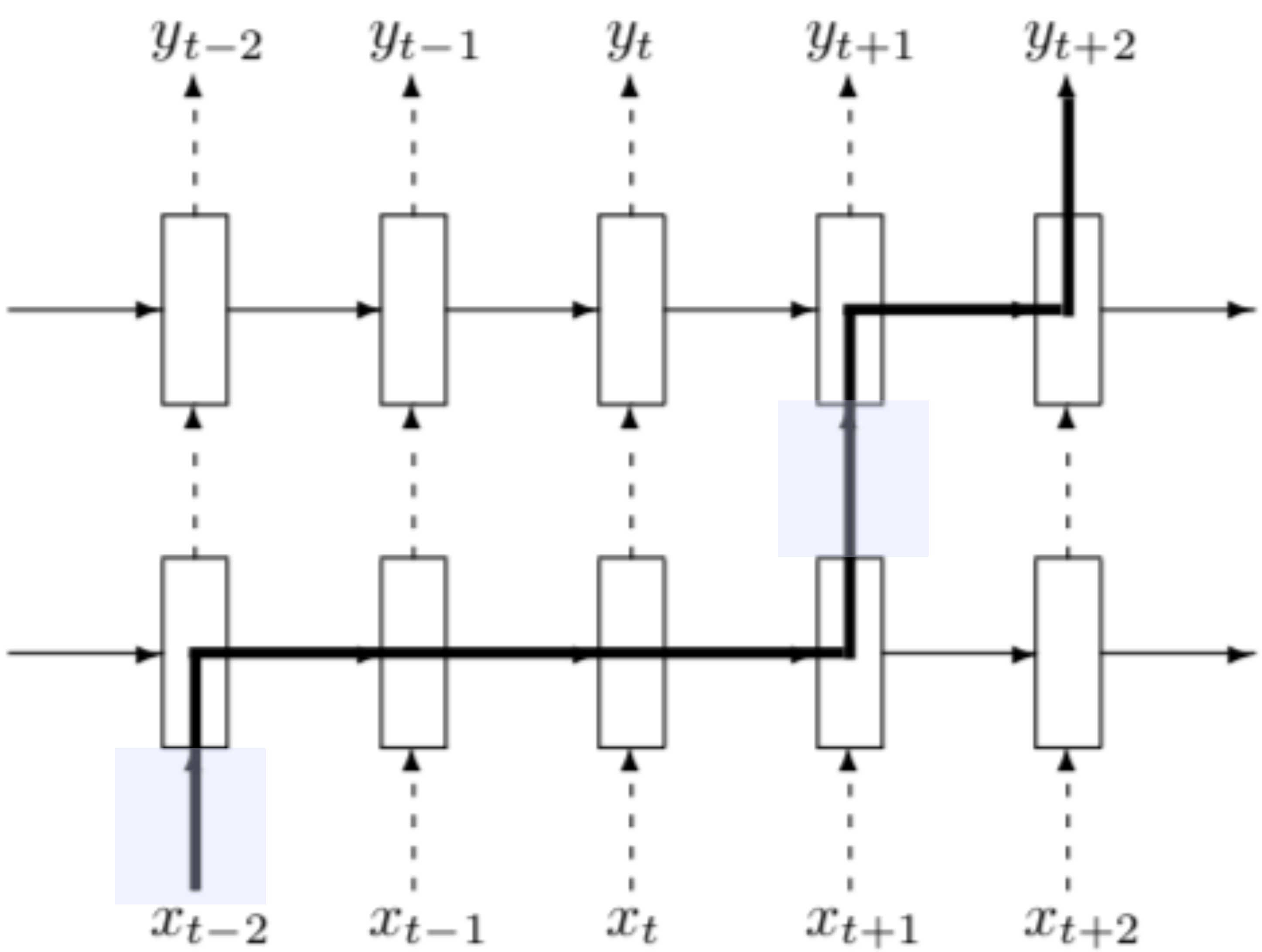
We cannot trust large gradients!



$$
\begin{aligned}
&\hat{\mathbf{g}} \leftarrow \frac{\partial \mathcal{E}}{\partial \theta} \\
&\textbf{if} \quad \|\hat{\mathbf{g}}\| \geq threshold \ \textbf{then} \\
&\quad \hat{\mathbf{g}} \leftarrow \frac{threshold}{\|\hat{\mathbf{g}}\|}\hat{\mathbf{g}} \\
&\textbf{end if}
\end{aligned}
$$

Naive Dropout

For the input connections only



Dropout for LSTM cell

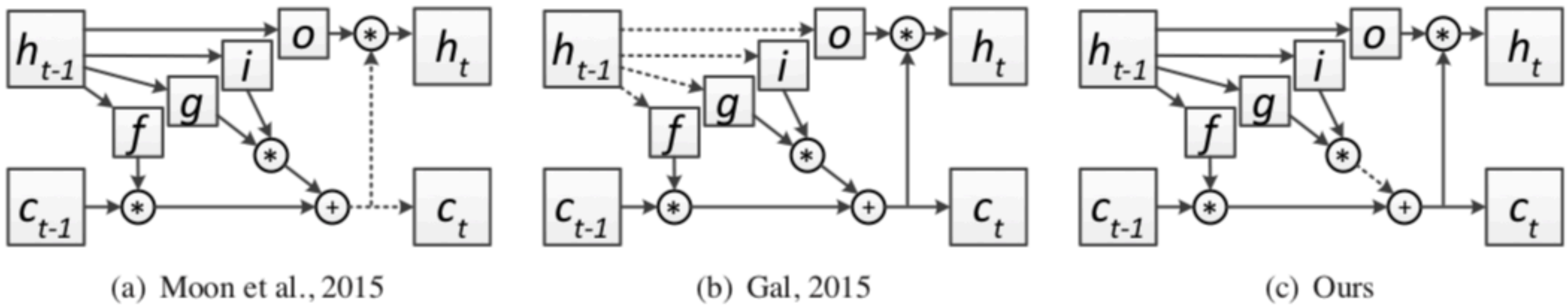


Figure 1: Illustration of the three types of dropout in recurrent connections of LSTM networks. Dashed arrows refer to dropped connections. Input connections are omitted for clarity.

a), b) require sampling a mask per sequence
c) sampling per step

$$c_t = f_t \cdot c_{t-1} + i_t \cdot d(g_t)$$

Recurrent Batch Norm

$$\begin{pmatrix} \tilde{\mathbf{f}}_t \\ \tilde{\mathbf{i}}_t \\ \tilde{\mathbf{o}}_t \\ \tilde{\mathbf{g}}_t \end{pmatrix} = \mathrm{BN}(\mathbf{W}_h \mathbf{h}_{t-1}; \gamma_h, \beta_h) + \mathrm{BN}(\mathbf{W}_x \mathbf{x}_t; \gamma_x, \beta_x) + \mathbf{b}$$

$$\mathbf{c}_t = \sigma(\tilde{\mathbf{f}}_t) \odot \mathbf{c}_{t-1} + \sigma(\tilde{\mathbf{i}}_t) \odot \tanh(\tilde{\mathbf{g}}_t)$$

$$\mathbf{h}_t = \sigma(\tilde{\mathbf{o}}_t) \odot \tanh(\mathrm{BN}(\mathbf{c}_t; \gamma_c, \beta_c))$$

- Statistics are not shared across time!
- need careful initialisation of γ

Layer Norm

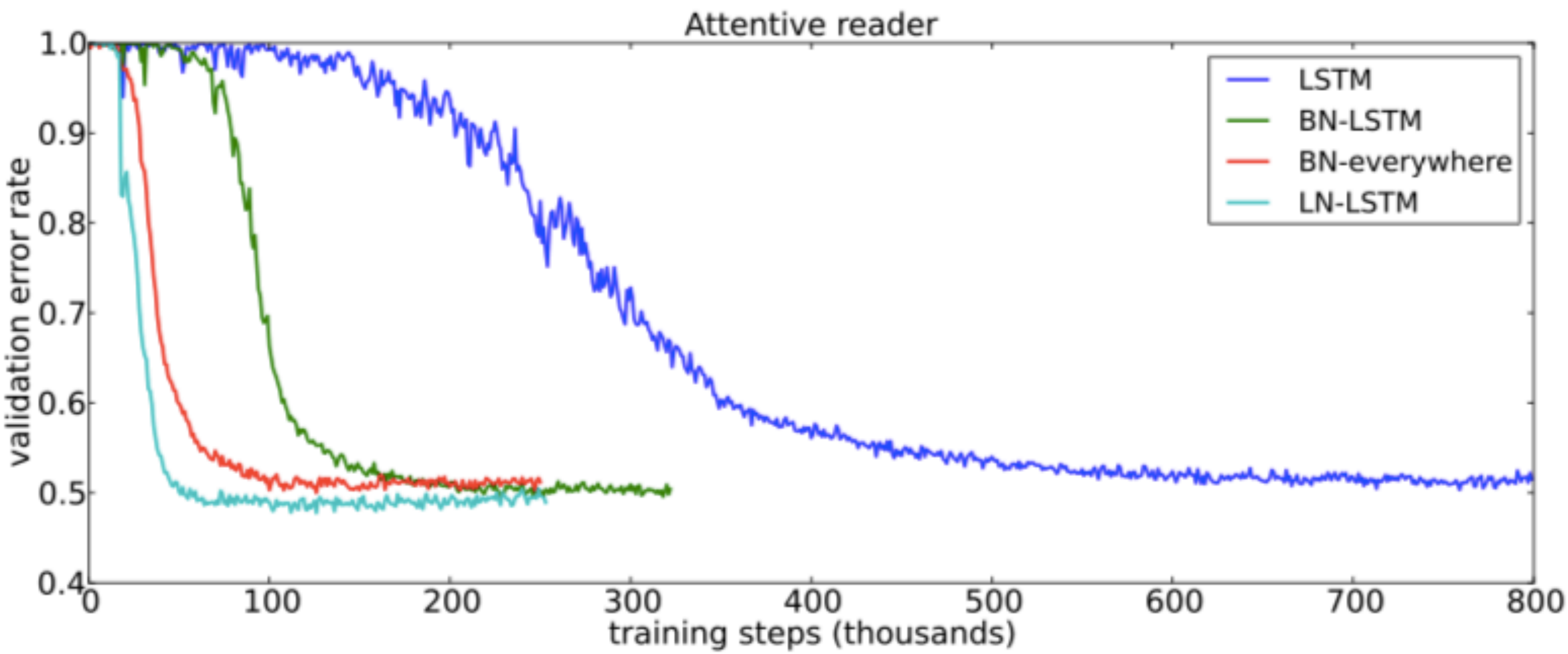
$$\mathbf{a}^t = W_{hh}h^{t-1} + W_{xh}\mathbf{x}^t$$

$$\mathbf{h}^t = f\left[\frac{\mathbf{g}}{\sigma^t} \odot (\mathbf{a}^t - \mu^t) + \mathbf{b}\right] \qquad \mu^t = \frac{1}{H}\sum_{i=1}^{H} a_i^t \qquad \sigma^t = \sqrt{\frac{1}{H}\sum_{i=1}^{H}(a_i^t - \mu^t)^2}$$

- Work same both for training and inference (batchnorm don't)

Layer Norm vs Batch Norm

Question answering task



Experiments by A.Karpathy!

Shakespeare

3-layer RNN with 512 hidden nodes on each layer

```
PANDARUS:
Alas, I think he shall be come approached and the day
When little srain would be attain'd into being never
fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.

DUKE VINCENTIO:
Well, your wit is in the care of side and that.

Second Lord:
They would be ruled after this chamber, and
my fair nues begun out of the fact, to be conveyed,
Whose noble souls I'll have the heart of the wars.

Clown:
Come, sir, I will make did behold your worship.

VIOLA:
I'll drink it.
```

```
VIOLA:
Why, Salisbury must find his flesh and thought
That which I am not aps, not a man and in fire,
To show the reining of the raven and the wars
To grace my hand reproach within, and not a fair
are hand,
That Caesar and my goodly father's world;
When I was heaven of presence and our fleets,
We spare with hours, but cut thy council I am
great,
Murdered and by thy master's ready there
My power to give thee but so much as hell:
Some service in the noble bondman here,
Would show him to her wine.

KING LEAR:
O, if you were a feeble sight, the courtesy of your
law,
Your sight and several breath, will wear the gods
With his heads, and my hands are wonder'd at the
deeds,
So drop upon your lordship's head, and your opinion
Shall be against your honour.
```

http://karpathy.github.io/2015/05/21/rnn-effectiveness/

Experiments by A.Karpathy!

Wikipedia


```
<page>
  <title>Antichrist</title>
  <id>865</id>
  <revision>
    <id>15900676</id>
    <timestamp>2002-08-03T18:14:12Z</timestamp>
    <contributor>
      <username>Paris</username>
      <id>23</id>
    </contributor>
    <minor />
    <comment>Automated conversion</comment>
    <text xml:space="preserve">#REDIRECT [[Christianity]]</
text>
  </revision>
</page>
```

Experiments by A.Karpathy!

Wikipedia

```
Naturalism and decision for the majority of Arab countries' capitalide was grounded
by the Irish language by [[John Clair]], [[An Imperial Japanese Revolt]], associated
with Guangzham's sovereignty. His generals were the powerful ruler of the Portugal
in the [[Protestant Immineners]], which could be said to be directly in Cantonese
Communication, which followed a ceremony and set inspired prison, training. The
emperor travelled back to [[Antioch, Perth, October 25|21]] to note, the Kingdom
of Costa Rica, unsuccessful fashioned the [[Thrales]], [[Cynth's Dajoard]], known
in western [[Scotland]], near Italy to the conquest of India with the conflict.
Copyright was the succession of independence in the slop of Syrian influence that
was a famous German movement based on a more popular servicious, non-doctrinal
and sexual power post. Many governments recognize the military housing of the
[[Civil Liberalization and Infantry Resolution 265 National Party in Hungary]],
that is sympathetic to be to the [[Punjab Resolution]]
(PJS)[http://www.humah.yahoo.com/guardian.
cfm/7754800786d17551963s89.htm Official economics Adjoint for the Nazism, Montgomery
was swear to advance to the resources for those Socialism's rule,
was starting to signing a major tripad of aid exile.]]
```

Experiments by A.Karpathy!

Algebraic geometry

Proof. Omitted. ☐

Lemma 0.1. *Let $\mathcal{C}$ be a set of the construction.*

Let $\mathcal{C}$ be a gerber covering. Let $\mathcal{F}$ be a quasi-coherent sheaves of $\mathcal{O}$-modules. We have to show that

$$\mathcal{O}_{\mathcal{O}_X} = \mathcal{O}_X(\mathcal{L})$$

.

Proof. This is an algebraic space with the composition of sheaves $\mathcal{F}$ on $X_{\acute{e}tale}$ we have

$$\mathcal{O}_X(\mathcal{F}) = \{morph_1 \times_{\mathcal{O}_X} (\mathcal{G}, \mathcal{F})\}$$

where $\mathcal{G}$ defines an isomorphism $\mathcal{F} \to \mathcal{F}$ of $\mathcal{O}$-modules. ☐

Lemma 0.2. *This is an integer $\mathcal{Z}$ is injective.*

Proof. See Spaces, Lemma ??. ☐

Lemma 0.3. *Let S be a scheme. Let X be a scheme and X is an affine open covering. Let $\mathcal{U} \subset \mathcal{X}$ be a canonical and locally of finite type. Let X be a scheme. Let X be a scheme which is equal to the formal complex.*

The following to the construction of the lemma follows.

Let X be a scheme. Let X be a scheme covering. Let

$$b : X \to Y' \to Y \to Y \to Y' \times_X Y \to X.$$

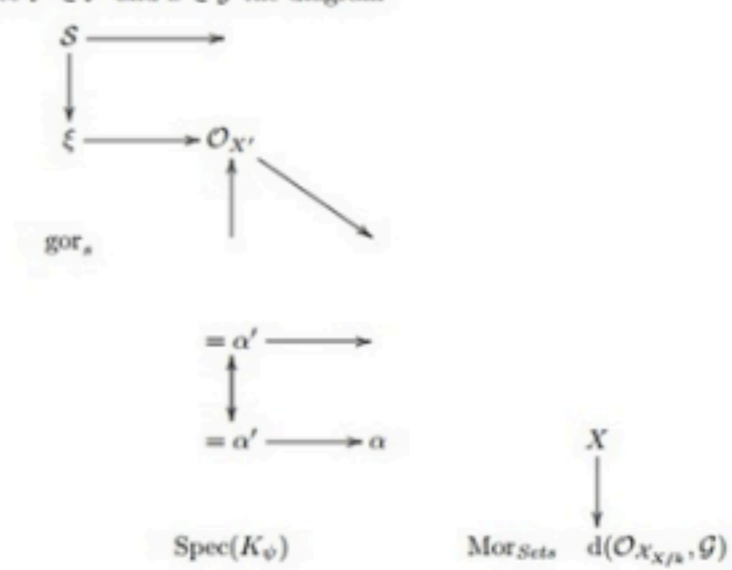
be a morphism of algebraic spaces over S and Y.

Proof. Let X be a nonzero scheme of X. Let X be an algebraic space. Let $\mathcal{F}$ be a quasi-coherent sheaf of $\mathcal{O}_X$-modules. The following are equivalent

(1) $\mathcal{F}$ is an algebraic space over S.
(2) If X is an affine open covering.

Consider a common structure on X and X the functor $\mathcal{O}_X(U)$ which is locally of finite type. ☐

This since $\mathcal{F} \in \mathcal{F}$ and $x \in \mathcal{G}$ the diagram



is a limit. Then $\mathcal{G}$ is a finite type and assume S is a flat and $\mathcal{F}$ and $\mathcal{G}$ is a finite type f_*. This is of finite type diagrams, and

- the composition of $\mathcal{G}$ is a regular sequence,
- $\mathcal{O}_{X'}$ is a sheaf of rings.

☐

Proof. We have see that $X = \mathrm{Spec}(R)$ and $\mathcal{F}$ is a finite type representable by algebraic space. The property $\mathcal{F}$ is a finite morphism of algebraic stacks. Then the cohomology of X is an open neighbourhood of U. ☐

Proof. This is clear that $\mathcal{G}$ is a finite presentation, see Lemmas ??.
A *reduced above* we conclude that U is an open covering of $\mathcal{C}$. The functor $\mathcal{F}$ is a "field

$$\mathcal{O}_{X,x} \longrightarrow \mathcal{F}_{\bar{x}} \quad \text{-1}(\mathcal{O}_{X_{\acute{e}tale}}) \longrightarrow \mathcal{O}_{X_\ell}^{-1}\mathcal{O}_{X_\lambda}(\mathcal{O}_{X_\eta}^{\overline{v}})$$

is an isomorphism of covering of $\mathcal{O}_{X_i}$. If $\mathcal{F}$ is the unique element of $\mathcal{F}$ such that X is an isomorphism.
The property $\mathcal{F}$ is a disjoint union of Proposition ?? and we can filtered set of presentations of a scheme $\mathcal{O}_X$-algebra with $\mathcal{F}$ are opens of finite type over S.
If $\mathcal{F}$ is a scheme theoretic image points. ☐

If $\mathcal{F}$ is a finite direct sum $\mathcal{O}_{X_\lambda}$ is a closed immersion, see Lemma ??. This is a sequence of $\mathcal{F}$ is a similar morphism.

Experiments by A.Karpathy!

Final challenge: Linux Source Code

```
/*
 * Increment the size file of the new incorrect UI_FILTER group information
 * of the size generatively.
 */
static int indicate_policy(void)
{
  int error;
  if (fd == MARN_EPT) {
    /*
     * The kernel blank will coeld it to userspace.
     */
    if (ss->segment < mem_total)
      unblock_graph_and_set_blocked();
    else
      ret = 1;
    goto bail;
  }
  segaddr = in_SB(in.addr);
  selector = seg / 16;
  setup_works = true;
  for (i = 0; i < blocks; i++) {
    seq = buf[i++];
    bpf = bd->bd.next + i * search;
    if (fd) {
      current = blocked;
    }
  }
  rw->name = "Getjbbregs";
  bprm_self_clearl(&iv->version);
  regs->new = blocks[(BPF_STATS << info->historidac)] | PFMR_CLOBATHINC_SECONDS << 12;
  return segtable;
}
```

Experiments by A.Karpathy!

What's going on while training?

Iter 100

```
tyntd-iafhatawiaoihrdemot  lytdws  e ,tfti, astai f ogoh eoase rrranbyne 'nhthnee e
plia tklrgd t o idoe ns,smtt   h ne etie h,hregtrs nigtike,aoaenns lng
```

Iter 300

```
"Tmont thithey" fomesscerliund
Keushey. Thom here
sheulke, anmerenith ol sivh I lalterthend Bleipile shuwy fil on aseterlome
coaniogennc Phe lism thond hon at. MeiDimorotion in ther thize."
```

Iter 500

```
we counter. He stutn co des. His stanted out one ofler that concossions and was
to gearang reay Jotrets and with fre colt otf paitt thin wall. Which das stimn
```

Iter 1200

```
"Kite vouch!" he repeated by her
door. "But I would be done and quarts, feeling, then, son is people...."
```

Iter 2000

```
"Why do what that day," replied Natasha, and wishing to himself the fact the
princess, Princess Mary was easier, fed in had oftened him.
Pierre aking his soul came to the packs and drove up his father-in-law women.
```

Image captioning

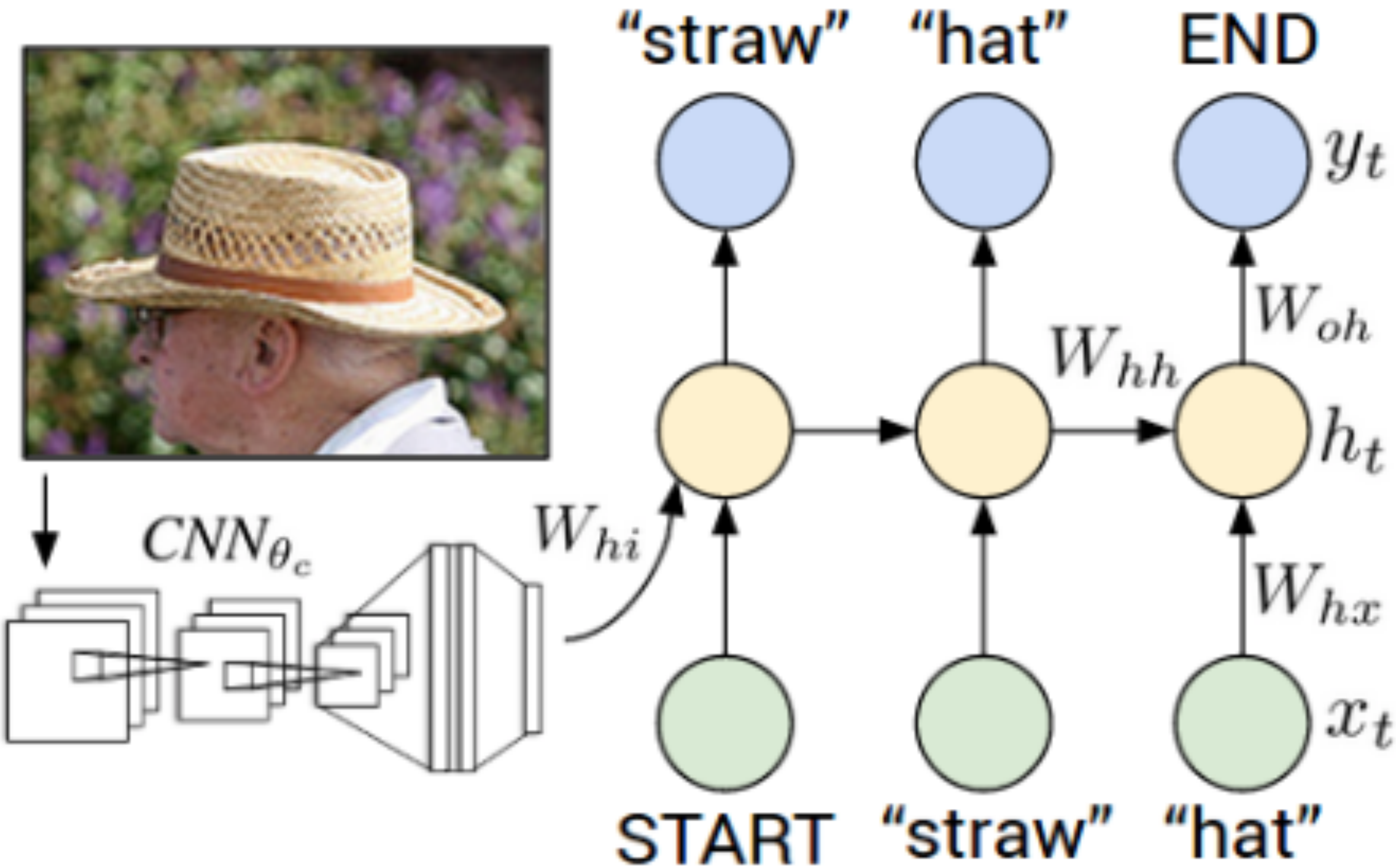
"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

Image Captioning

Image captioning



Attention

Soft attention

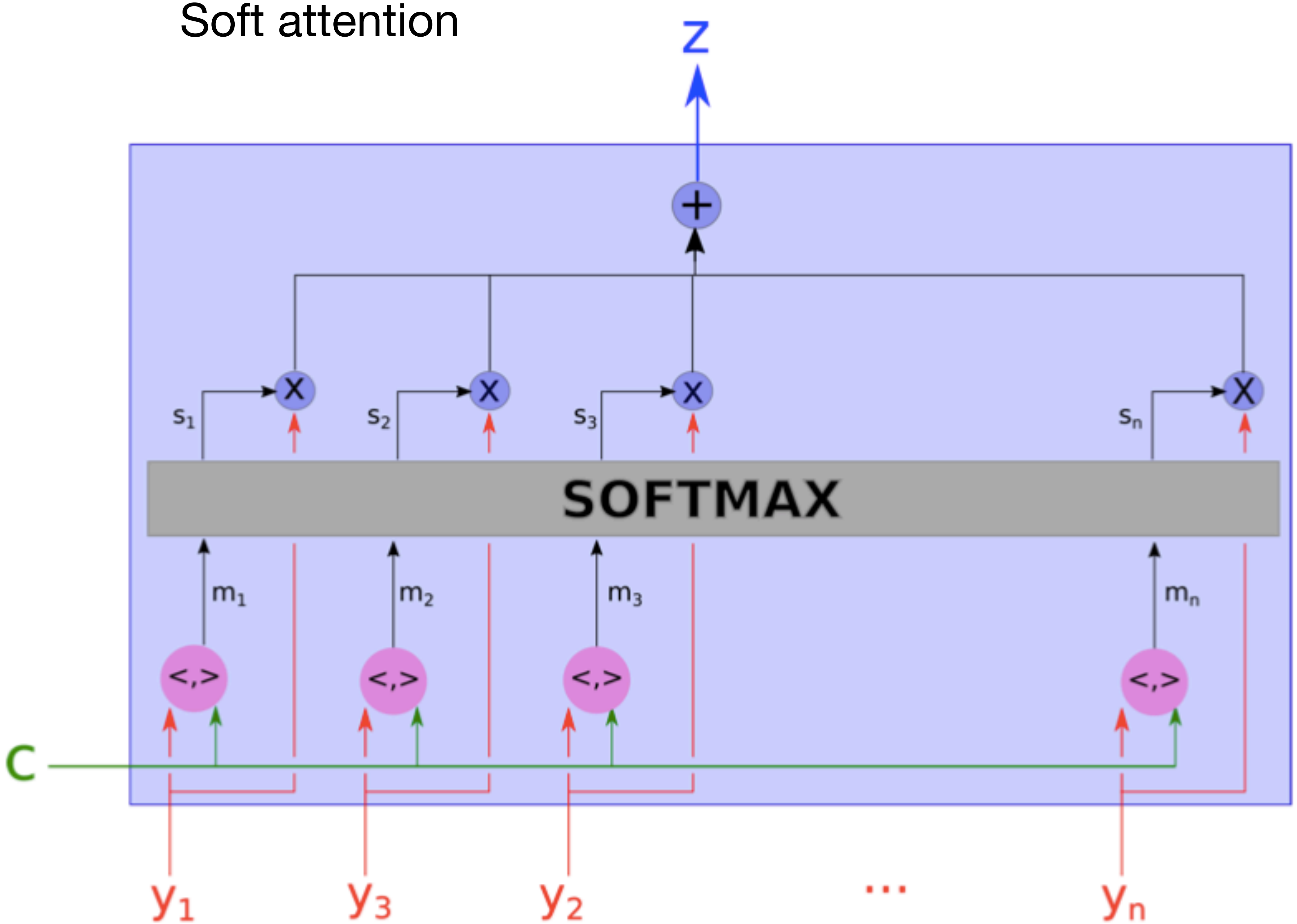
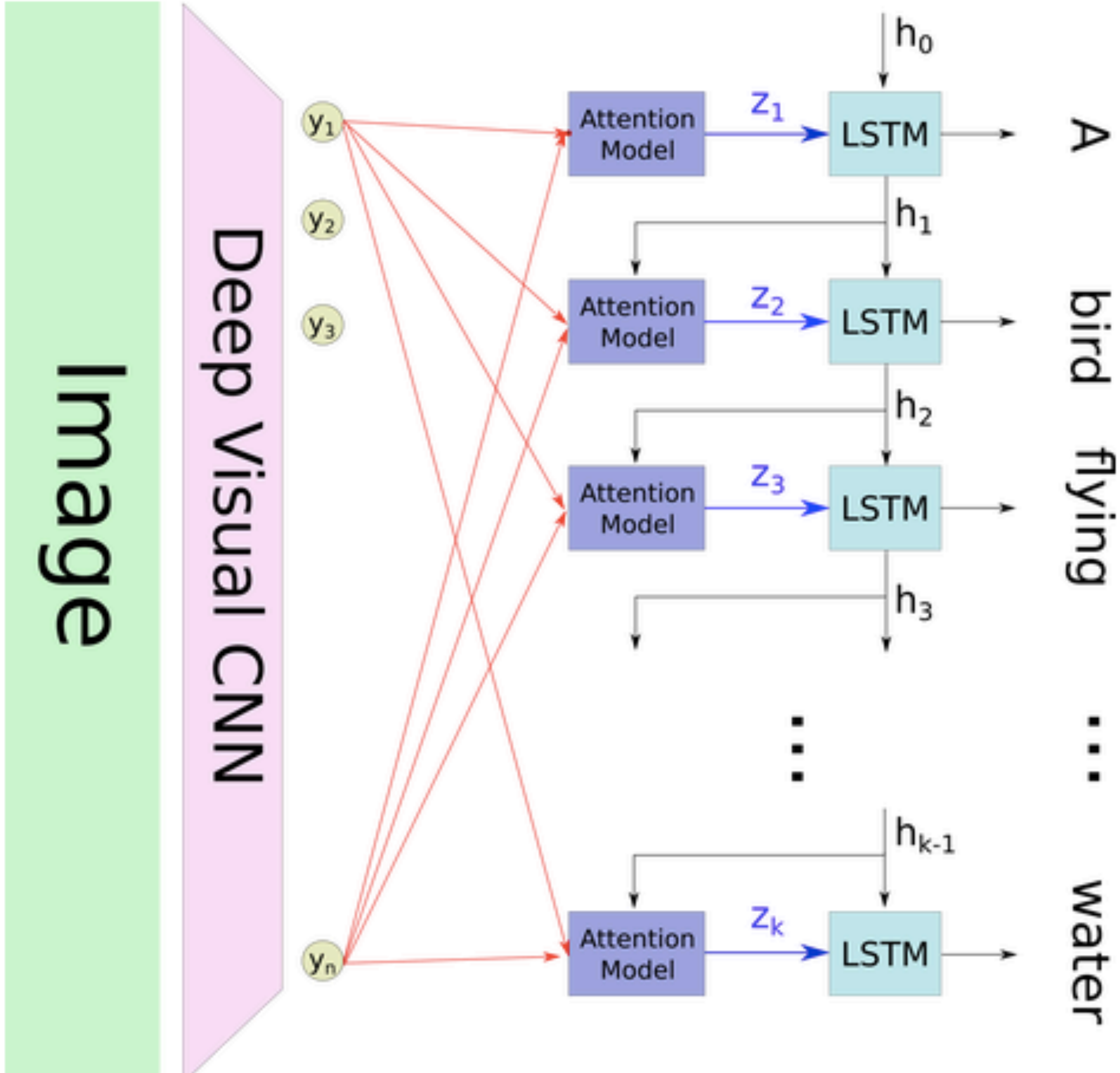
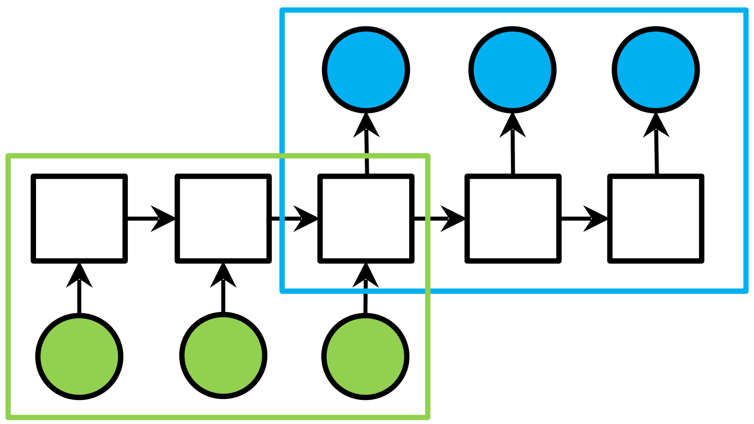


Image captioning + attention

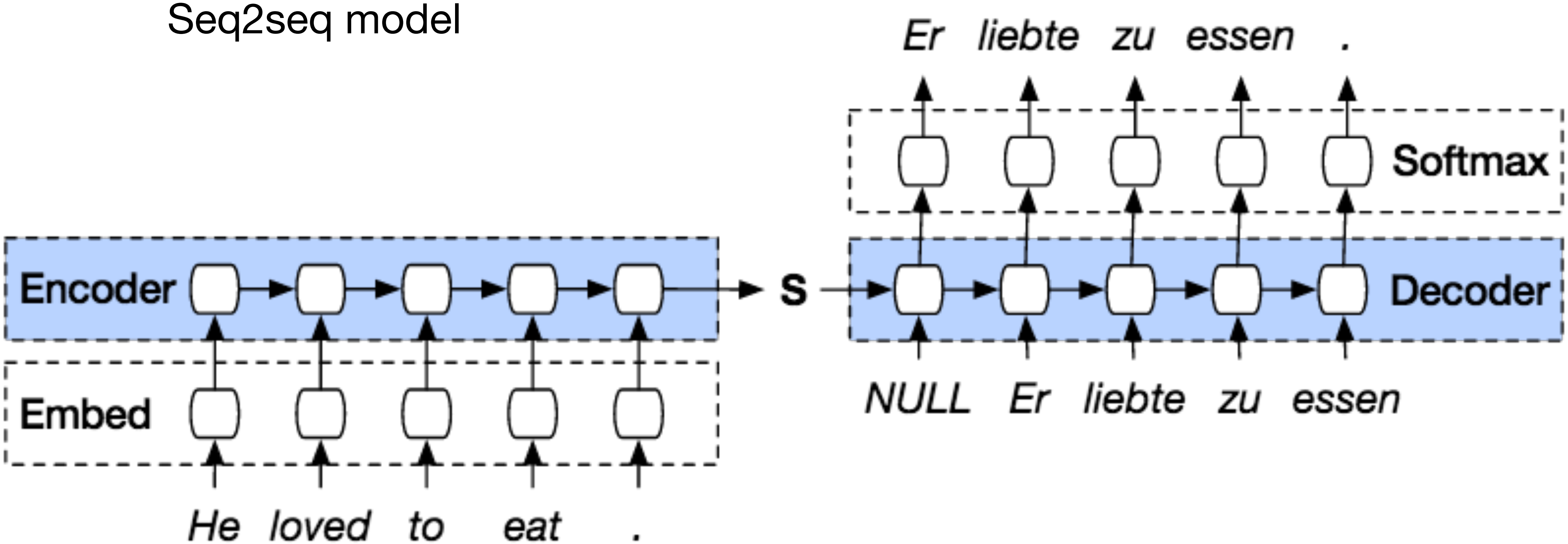


Attention model for image captioning

Machine translation



Seq2seq model



Reference

- https://github.com/justheuristic/Practical_RL/tree/master/week6.5
- http://karpathy.github.io/2015/05/21/rnn-effectiveness/
- Recurrent Batch Normalisation https://arxiv.org/pdf/1603.09025.pdf
- Layer Normalization https://arxiv.org/pdf/1607.06450.pdf
- Bengio et al. 1994 http://ai.dinfo.unifi.it/paolo/ps/tnn-94-gradient.pdf
- Image captioning https://towardsdatascience.com/image-captioning-in-deep-learning-9cd23fb4d8d2
- Attention mechanism https://blog.heuritech.com/2016/01/20/attention-mechanism/
- Lobacheva E. https://compsciclub.ru/media/slides/deep_learning_2016_summer/2016_07_23_deep_learning_2016_summer.pdf

Conclusions

- Recurrent Neural Networks - powerful tool for sequence analysis
- Hard to train. Gradients vanish/explode
- LSTM/GRU can capture long term dependences
- Use Gradient Clipping, BN, LN
- Sometimes need careful initialisation