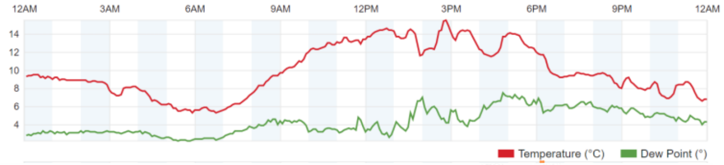
Deep Learning

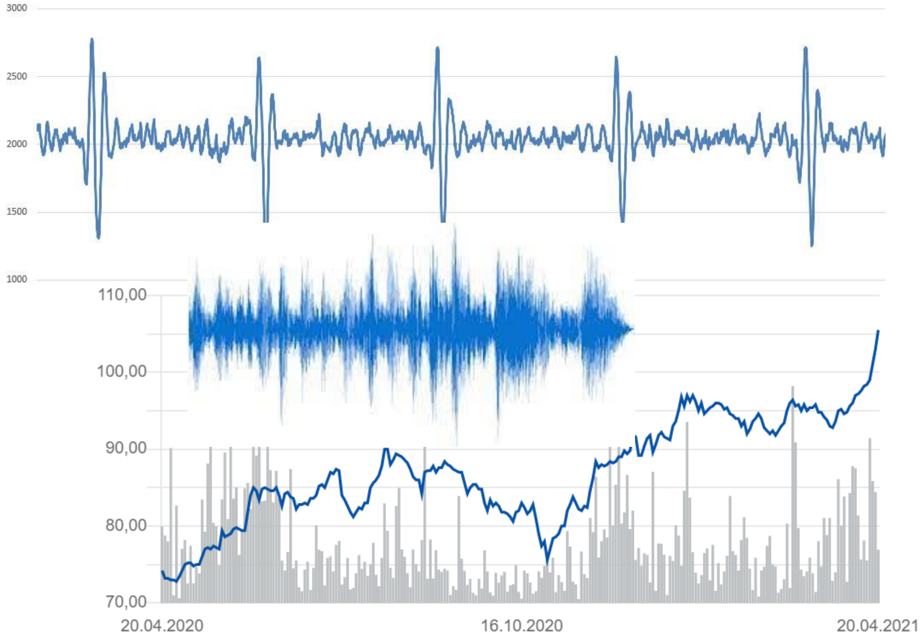
Recurrent Neural Networks

Danijel Skočaj
University of Ljubljana
Faculty of Computer and Information Science

Academic year: 2022/23

Sequential data

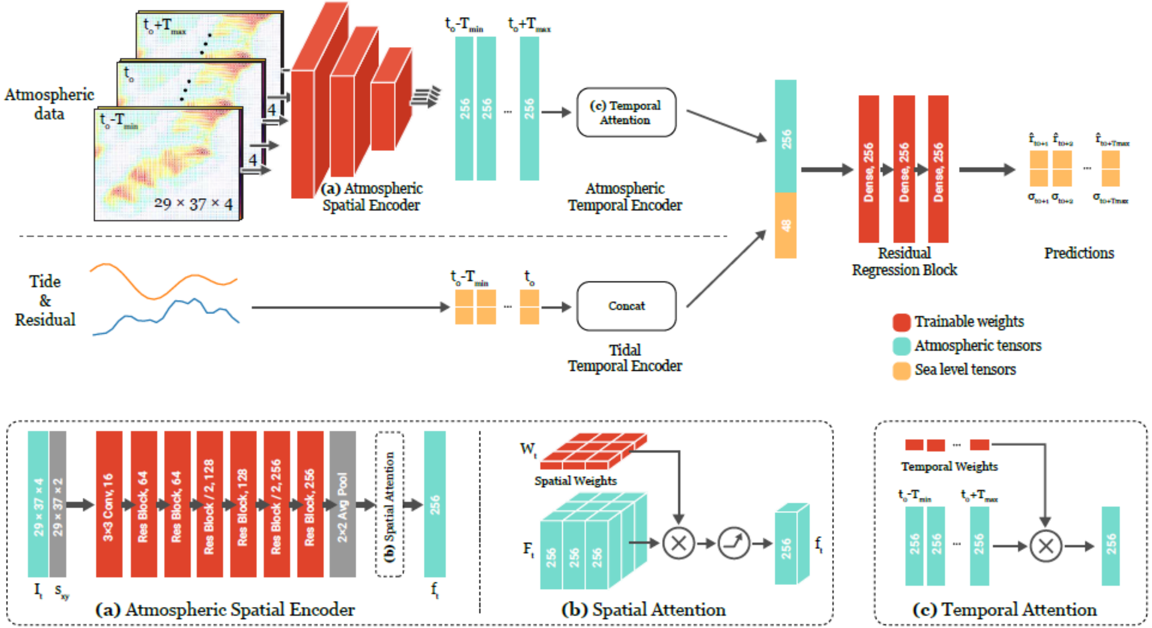


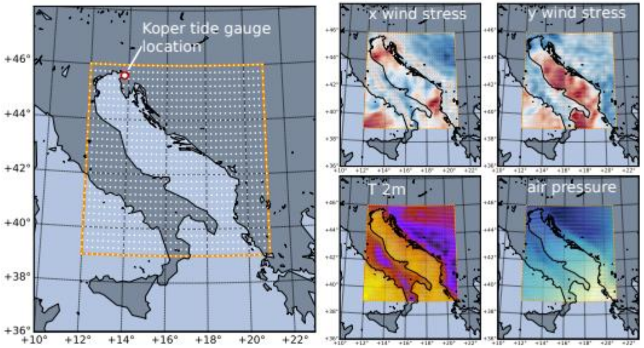


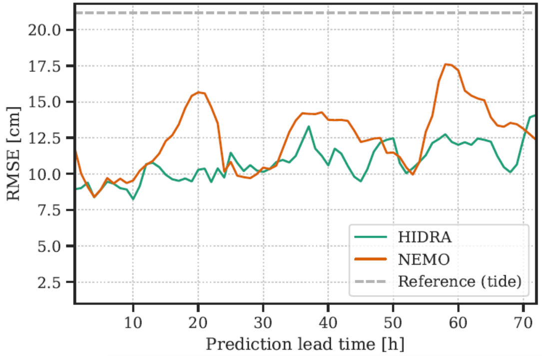
Deep learning is a type of machine learning that uses deep artificial neural networks for modelling acquired knowledge. **Artificial intelligence** is a research field dealing with the development of algorithms and systems for solving tasks that require intelligence to be solved.

CNN-based approach

- Sea level forecasting
- Stack a window of sequential data into a fixed-length tensor and use ANN/CNN
- Predict a fixed number of parameters



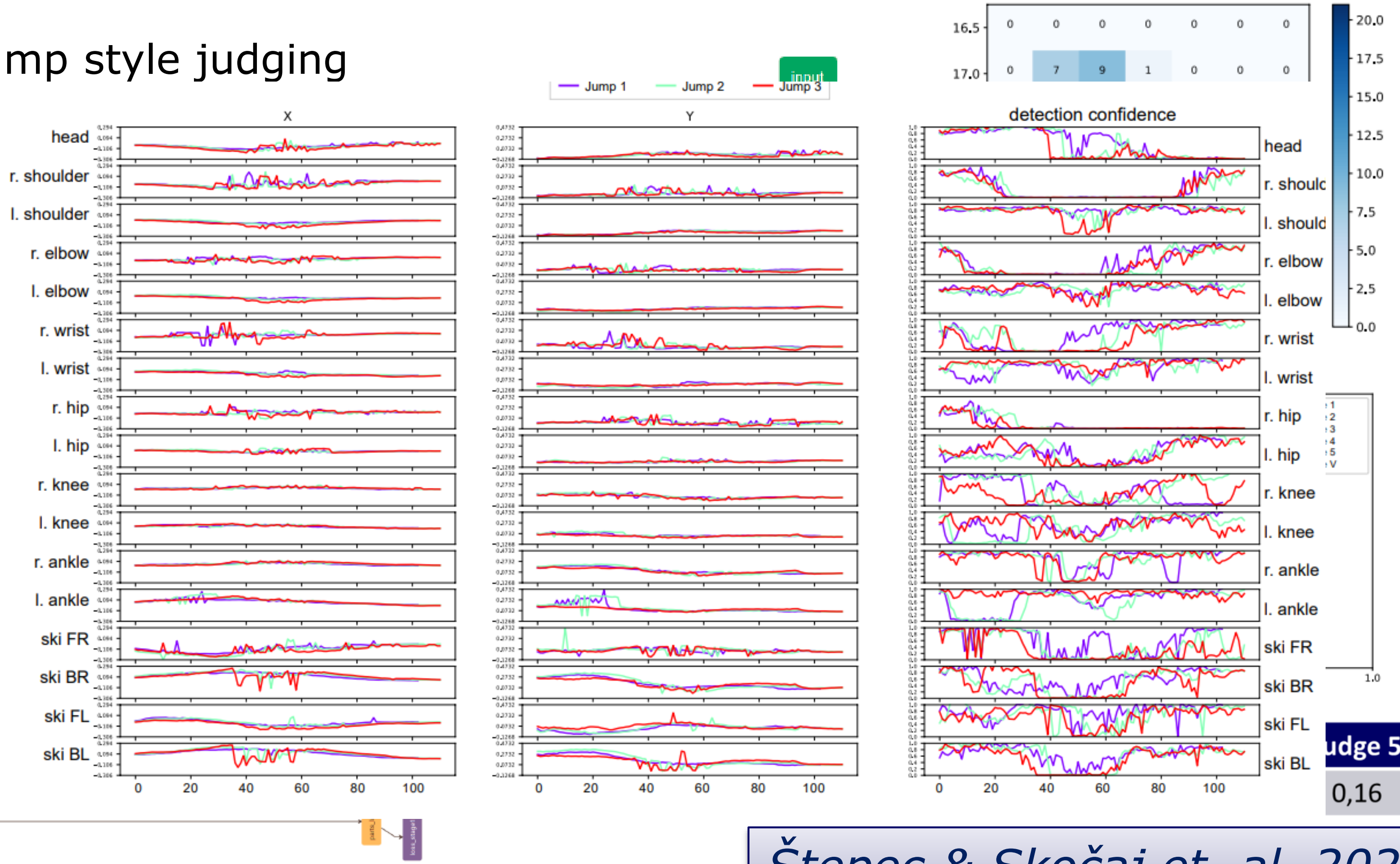




Žust et. al, 2021

CNN-based approach

- Ski jump style judging



- Faste
- CPM f
- Shallo
  - (1

udge 5	Virt.judge
0,16	0,21

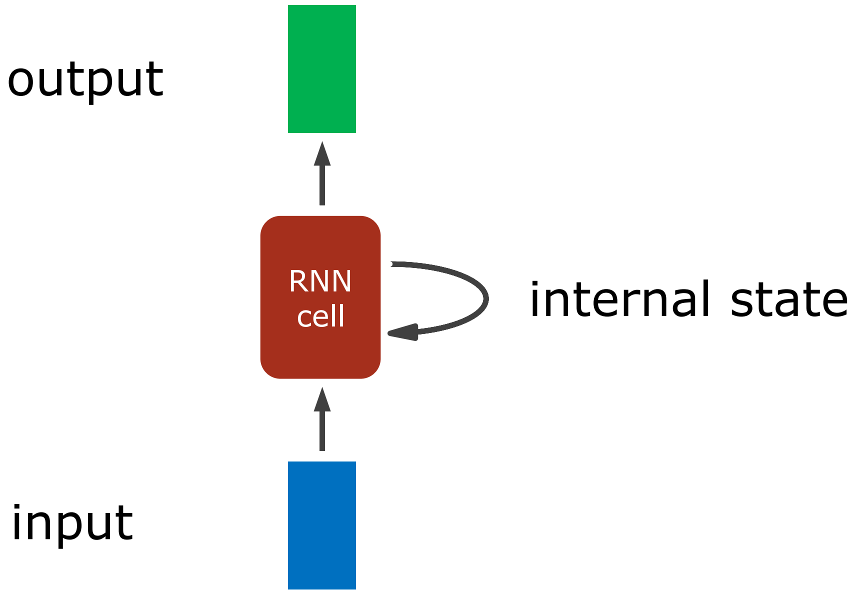
Štepec & Skočaj et. al, 2022

Naive approach

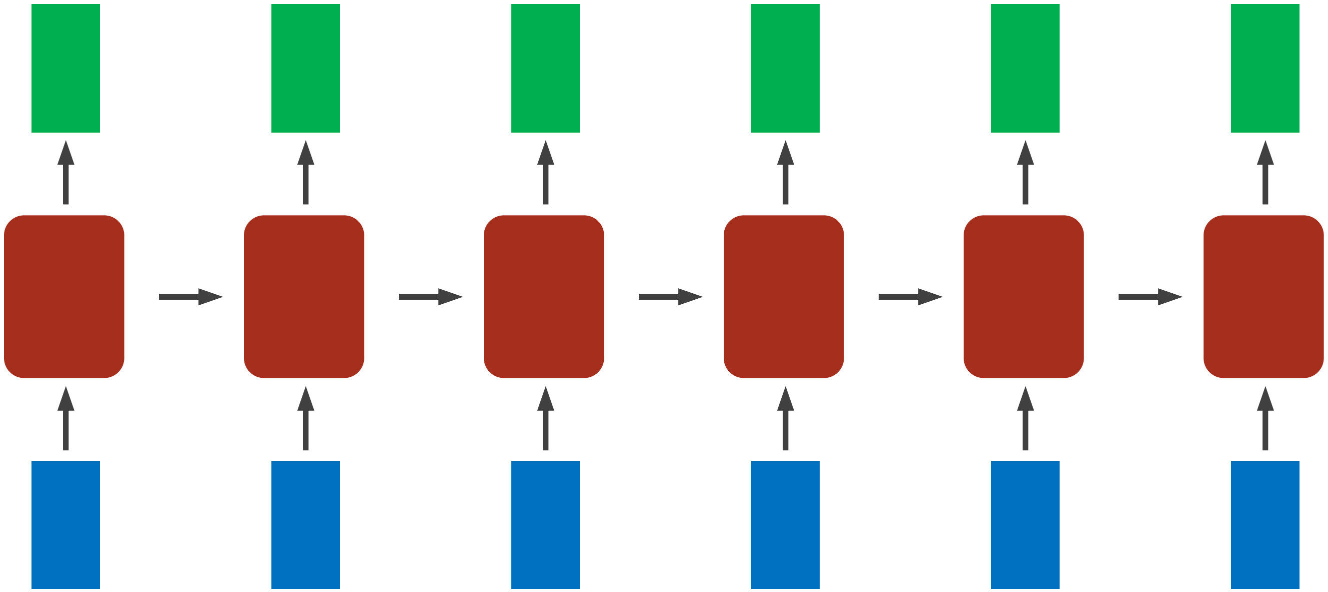
- Task: predict the next word.
  - `Deep learning is a type of machine learning.`
- Naive approach 1: Use the fixed window
  - `Deep learning is a type of machine learning.`
  - Too small, rigid, the important information might be at the beggining of the sequence: **`Deep learning is a not so new technique, which has been frequently applied lately. It is a type of machine learning.`**
- Naive approach 2: Bag of words
  - Count the number of the individual words
  - Counts don`t preserve the order:
    - **`Luka Dončić played extremely good tonight, not as bad as LeBron.`**
    - **`Luka Dončić played extremely bad tonight, not as good as LeBron.`**
- Requirements:
  - Sequence, variable length of sequences
  - Time (order) dependency, long term dependencies

Recurrent Neural Network

At every
time step:

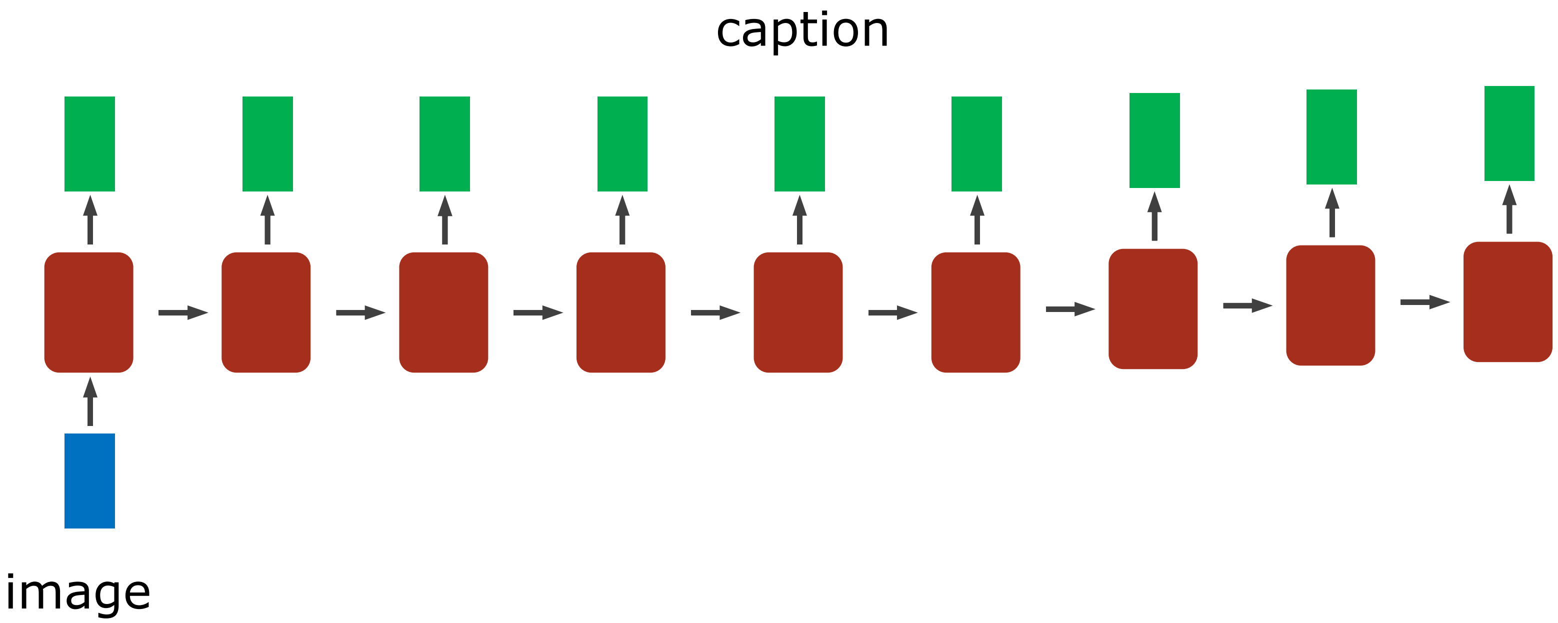


Recurrent Neural Network



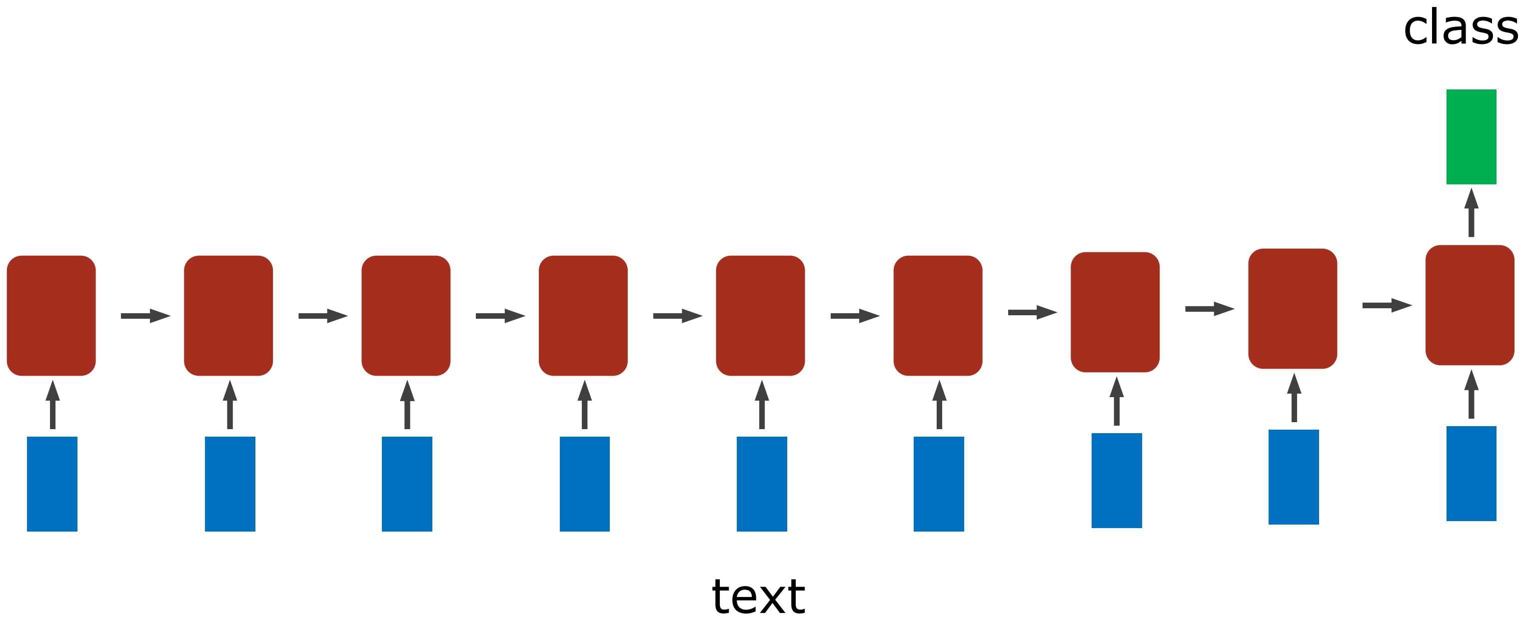
One-to-many RNN

- E.g., image captioning, text generation, music generation, etc.



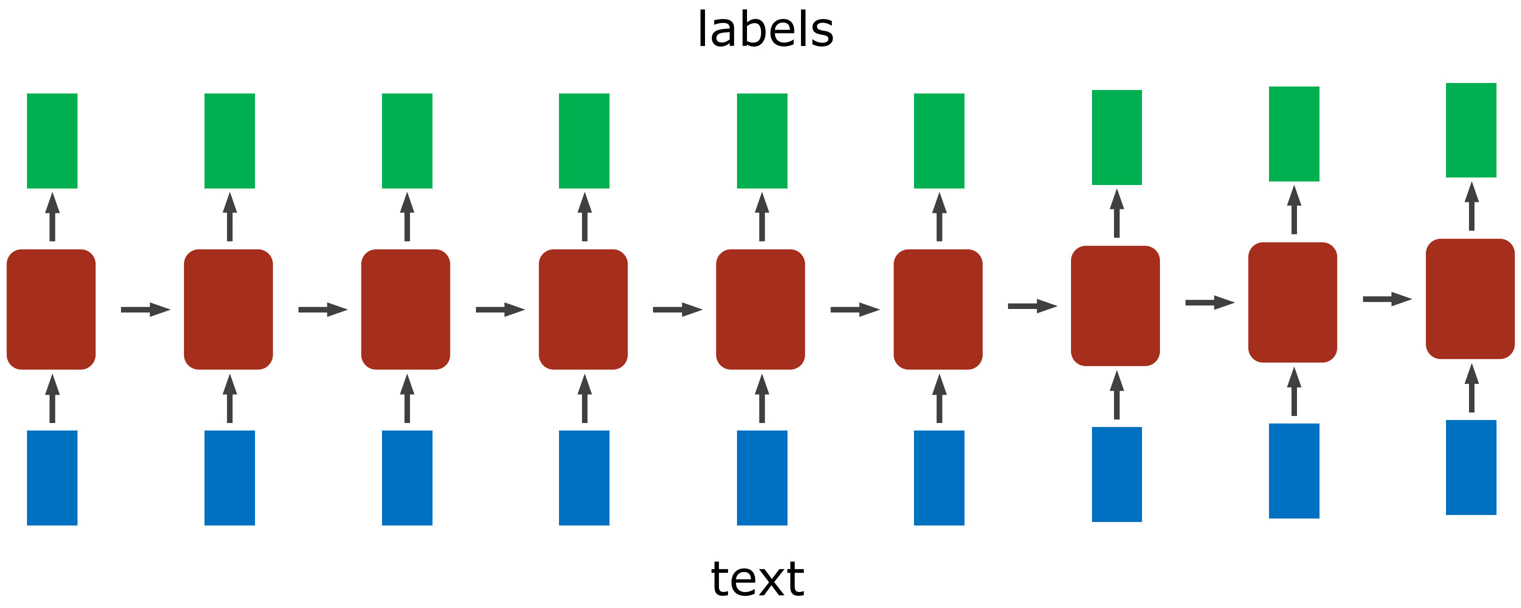
Many-to-one RNN

- E.g., text classification, action recognition



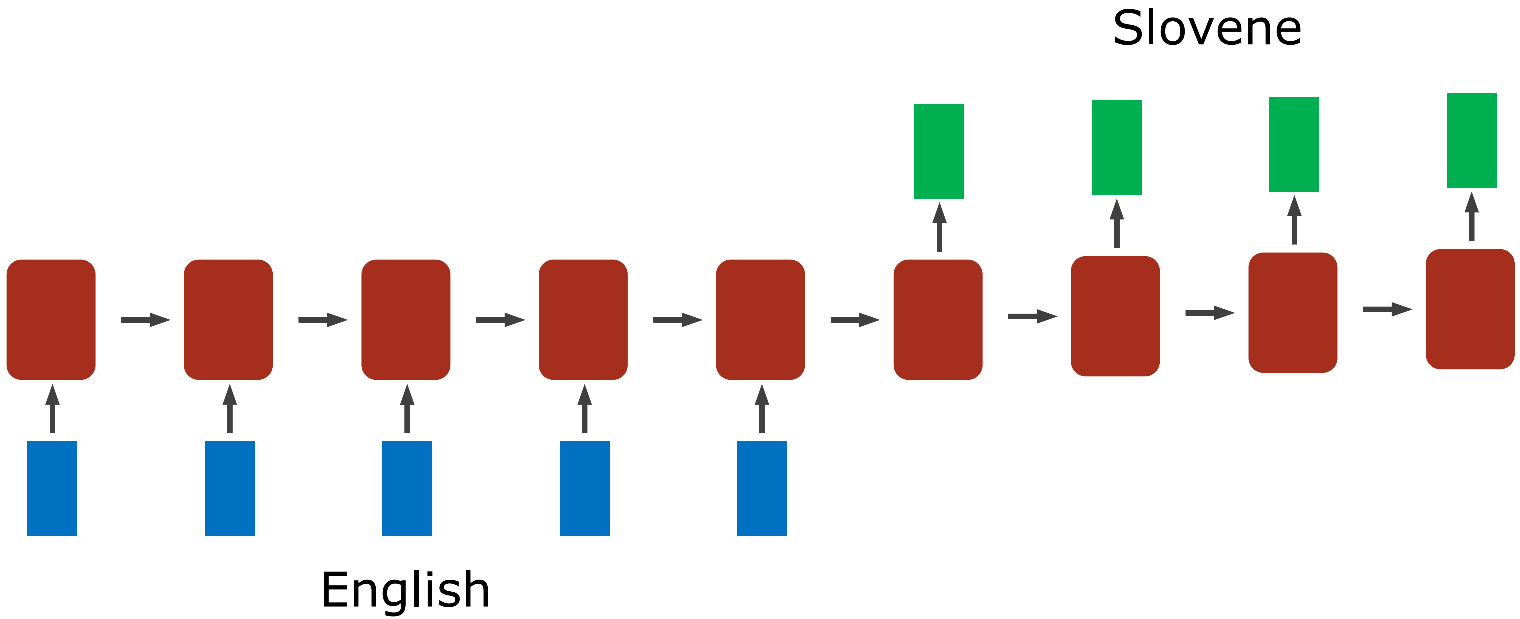
Many-to-many RNN

- E.g., named entity recognition, video segmentation

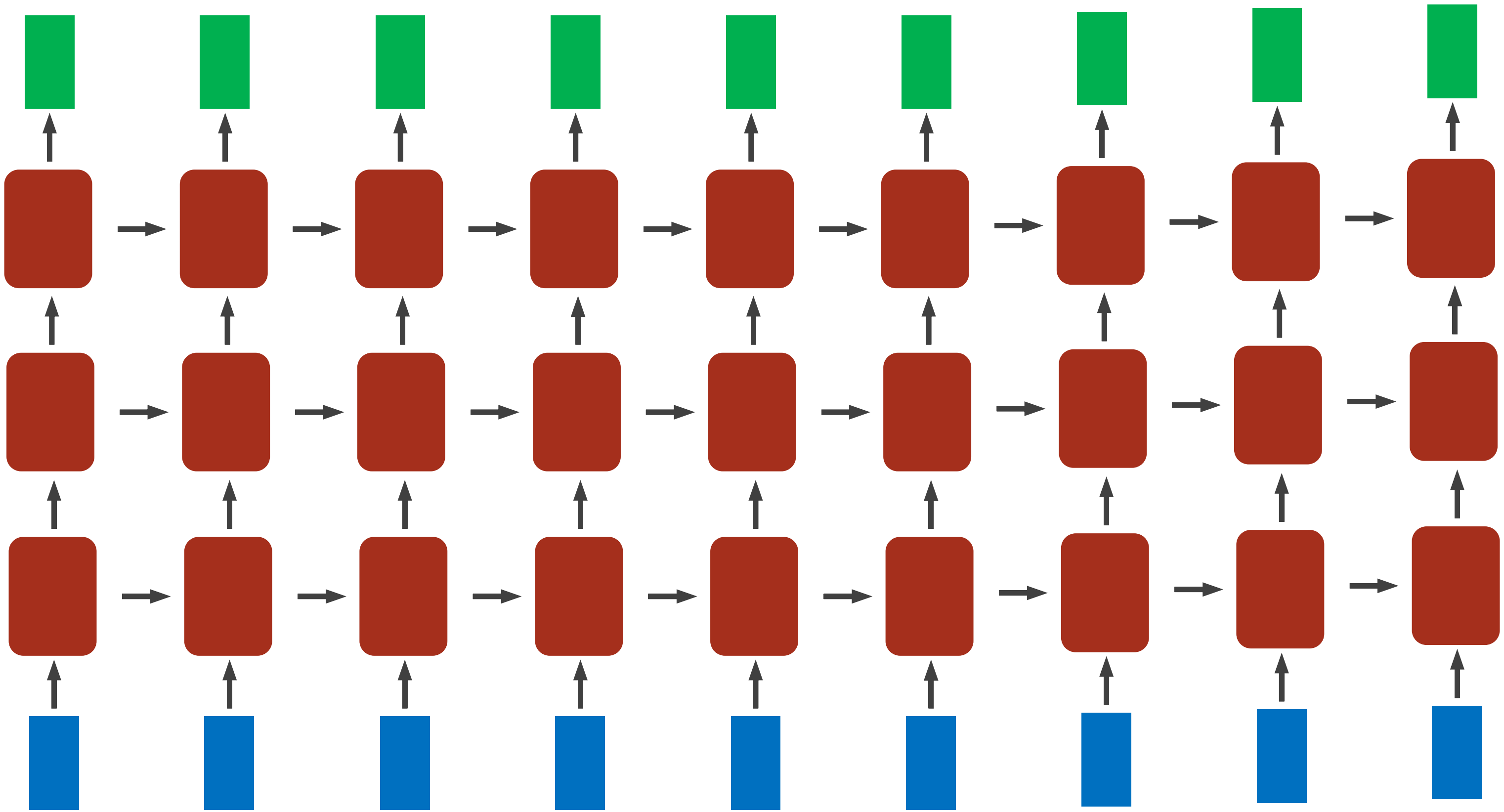


Many-to-many (many-to-one + one-tomany) RNN

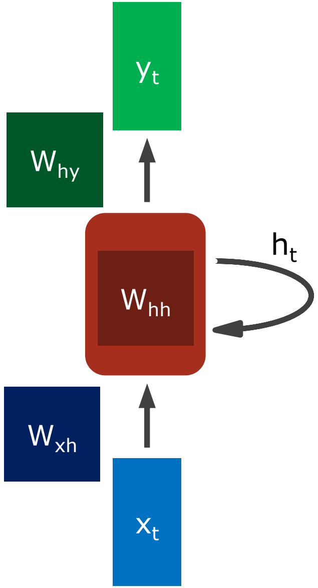
- E.g., machine translation, sequence to sequence



Multilayer RNN



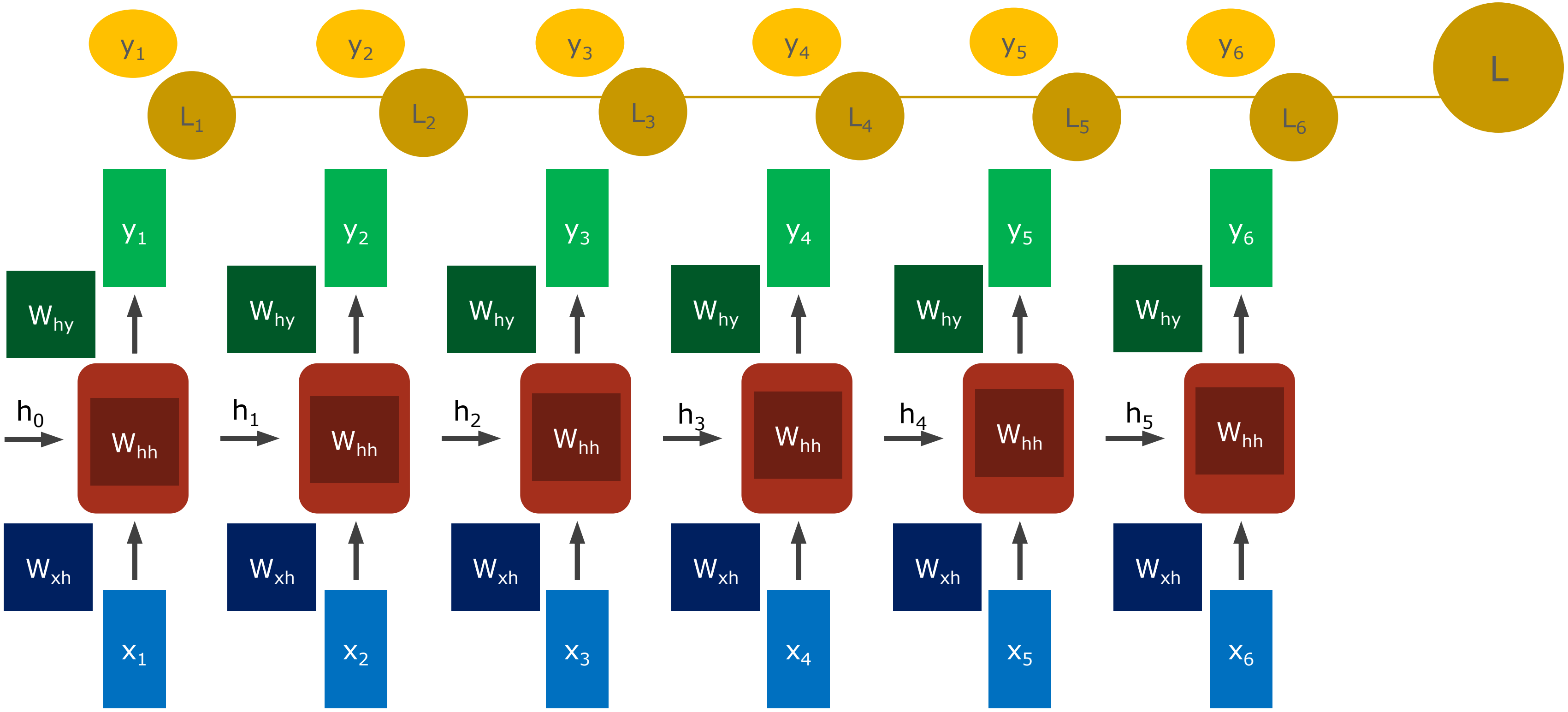
Recurrence formula



$$y_t = W_{hy} h_t$$

$$h_t = \tanh(W_{hh} h_{t-1} + W_{xh} x_t)$$

Computational graph



Backpropagation through time

$$\mathbf{h}_t = tanh(W_{hh}\mathbf{h}_{t-1} + W_{xh}\mathbf{x}_t + \mathbf{b}_\mathbf{h})$$

$$z_t = softmax(W_{hz}\mathbf{h}_t + \mathbf{b}_z)$$

$$\alpha_t = W_{hz}\mathbf{h}_t + \mathbf{b}_z$$

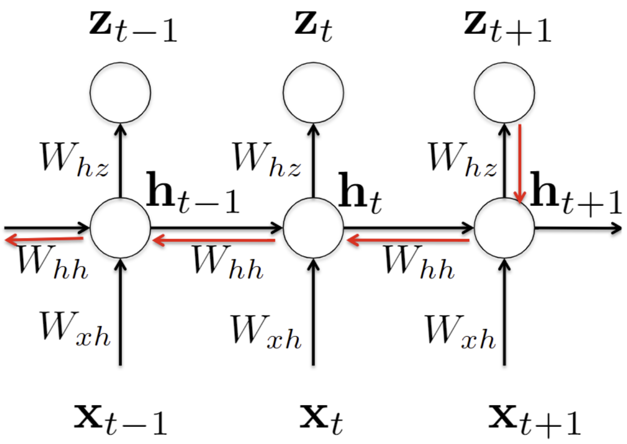
$$\mathcal{L}(\mathbf{x}, \mathbf{y}) = -\sum_t y_t log z_t \qquad \frac{\partial \mathcal{L}}{\partial \alpha_t} = -(y_t - z_t)$$

$$\frac{\partial \mathcal{L}}{\partial W_{hz}} = \sum_t \frac{\partial \mathcal{L}}{\partial z_t}\frac{\partial z_t}{\partial W_{hz}} \qquad \frac{\partial \mathcal{L}}{\partial b_z} = \sum_t \frac{\partial \mathcal{L}}{\partial z_t}\frac{\partial z_t}{\partial b_z}$$

$$\frac{\partial \mathcal{L}(t+1)}{\partial W_{hh}} = \sum_{k=1}^{t} \frac{\partial \mathcal{L}(t+1)}{\partial z_{t+1}}\frac{\partial z_{t+1}}{\partial \mathbf{h}_{t+1}}\frac{\partial \mathbf{h}_{t+1}}{\partial \mathbf{h}_k}\frac{\partial \mathbf{h}_t}{\partial W_{hh}}$$

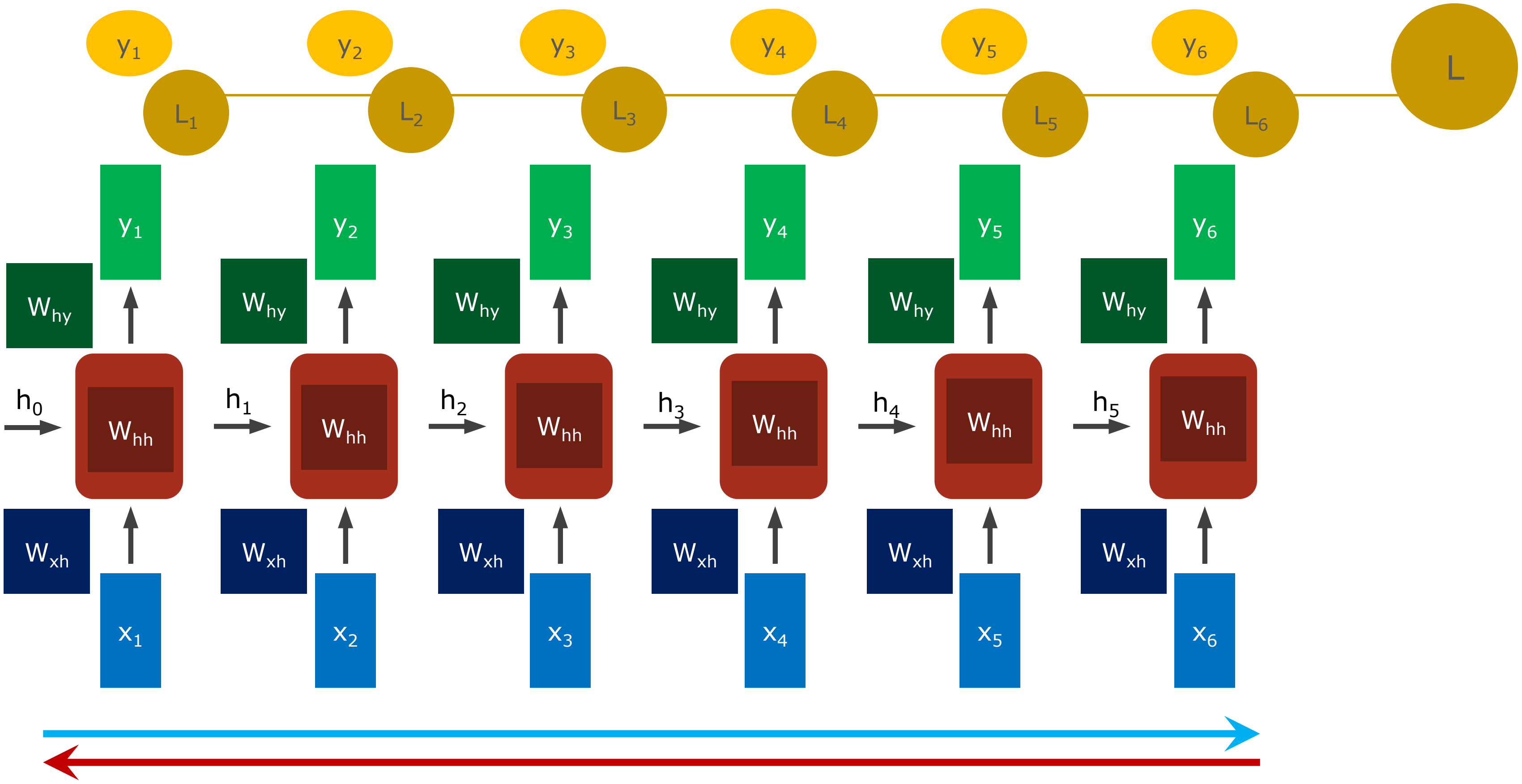
$$\frac{\partial \mathcal{L}}{\partial W_{hh}} = \sum_t \sum_{k=1}^{t+1} \frac{\partial \mathcal{L}(t+1)}{\partial z_{t+1}}\frac{\partial z_{t+1}}{\partial \mathbf{h}_{t+1}}\frac{\partial \mathbf{h}_{t+1}}{\partial \mathbf{h}_k}\frac{\partial \mathbf{h}_k}{\partial W_{hh}} \qquad \frac{\partial \mathcal{L}}{\partial W_{xh}} = \sum_t \sum_{k=1}^{t+1} \frac{\partial \mathcal{L}(t+1)}{\partial z_{t+1}}\frac{\partial z_{t+1}}{\partial \mathbf{h}_{t+1}}\frac{\partial \mathbf{h}_{t+1}}{\partial \mathbf{h}_k}\frac{\partial \mathbf{h}_k}{\partial W_{xh}}$$

Note: y=z

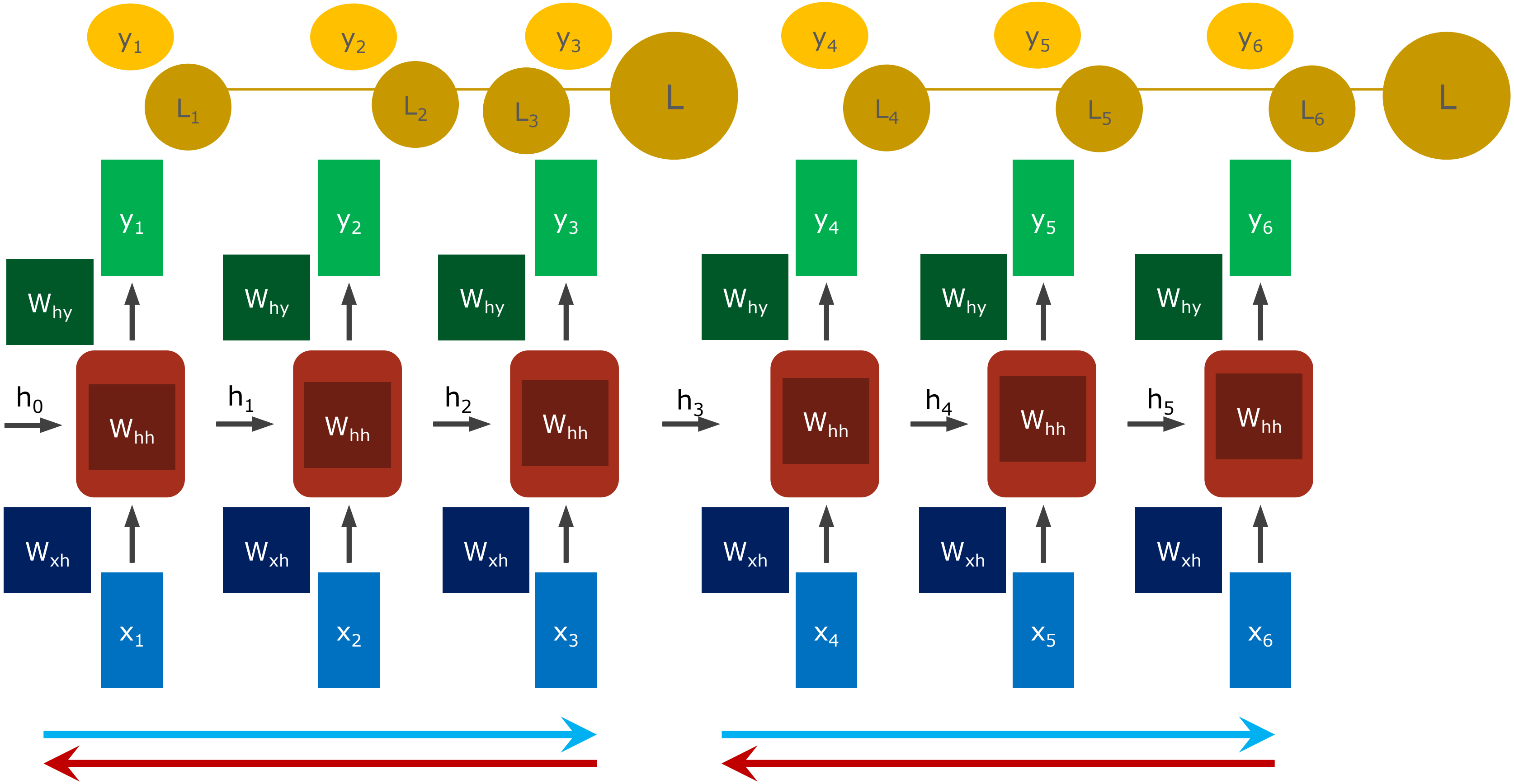


Chen, 2016

Backpropagation through time



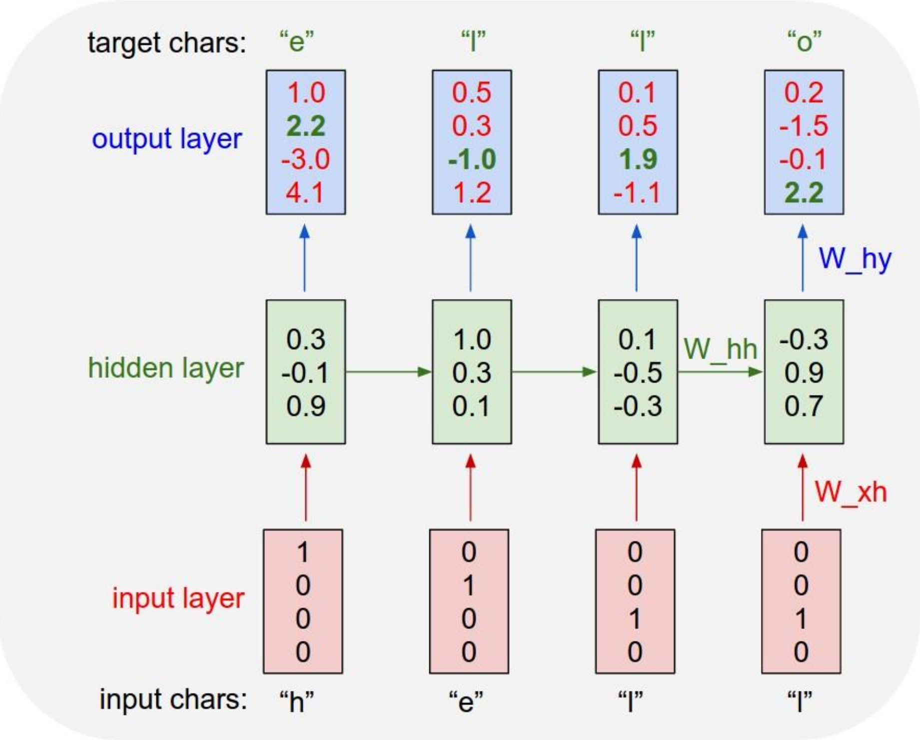
Truncated backpropagation through time



Example – character-level language models

- Task: generate text
- Model the probability distribution of the next character in the sequence given a sequence of previous characters

- Toy example:
  - Vocabulary: {h,e,l,o}
  - Training sample: „hello“



Karpathy, 2015

Example – character-level language models

- Tolstoy, War and peace

[Karpathy, 2015]

```
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
```

```
"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."
```

```
Aftair fall unsuch that the hall for Prince Velzonski's that me of
her hearly, and behs to so arwage fiving were to it beloge, pavu say falling misfort
how, and Gogition is so overelical and ofter.
```

```
"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.
```

Example – character-level language models

- Shakespeare

```
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.
```

Karpathy, 2015

```
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.
```

Example – character-level language models

- LaTeX

For $\bigoplus_{n=1,\ldots,m}$ where $\mathcal{L}_{m_\bullet} = 0$, hence we can find a closed subset $\mathcal{H}$ in $\mathcal{H}$ and any sets $\mathcal{F}$ on X, U is a closed immersion of S, then $U \to T$ is a separated algebraic space.

Proof. Proof of (1). It also start we get

Lemma 0.1. *Assume (3) and (3) by the construction in the description.*

Suppose $X = \lim |X|$ (by the formal open covering X and a single map $\underline{Proj}_X(\mathcal{A}) =$ Spec(B) over U compatible with the complex

$$Set(\mathcal{A}) = \Gamma(X, \mathcal{O}_{X,\mathcal{O}_X}).$$

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

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

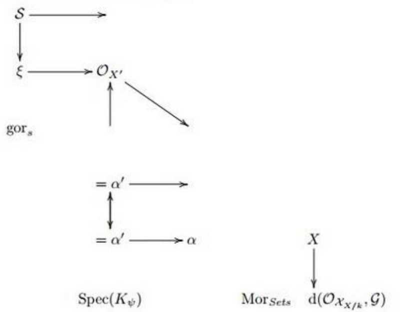
be a morphism of algebraic spaces over S and Y.

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 $X = \text{Spec}(R)$ 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

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

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.

at $\mathcal{Q} \to \mathcal{C}_{Z/X}$ is stable under the following result
ınd (3). This finishes the proof. By Definition ??
sed subschemes are catenary. If T is surjective we
with residue fields of S. Moreover there exists a
ere U in X' is proper (some defining as a closed
es to check the fact that the following theorem
Since $S = \text{Spec}(R)$ and $Y = \text{Spec}(R)$.

of sheaves on X. But given a scheme U and a
:. Let $U \cap U = \coprod_{i=1,\ldots,n} U_i$ be the scheme X over
$= \lim_i X_i$. □

estrocomposes of this implies that $\mathcal{F}_{x_0} = \mathcal{F}_{x_0} =$

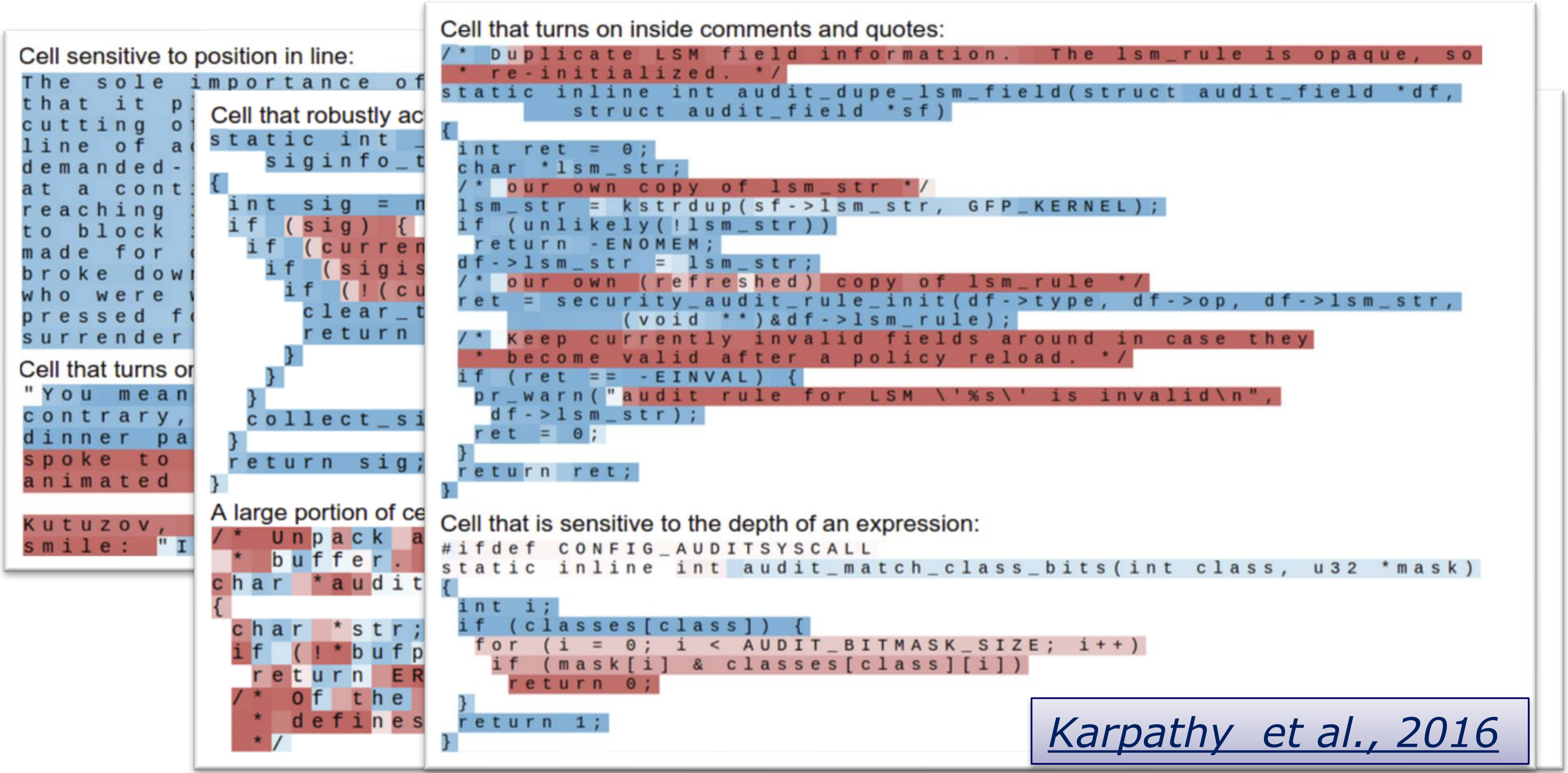
Noetherian scheme over S, $E = \mathcal{F}_{X/S}$. Set $\mathcal{I} =$
zero over $i_0 \leq \mathfrak{p}$ is a subset of $\mathcal{J}_{n,0} \circ \overline{A}_2$ works.

Ience we may assume $\mathfrak{q}' = 0$.

we see that $\mathfrak{p}$ is the mext functor (??). On the
that
$)(\mathcal{O}_{X'}) = \mathcal{O}_X(D)$
$_{n+1}$ is a scheme over S. □

[Karpathy, 2015]

Example – interpreting character-level language models

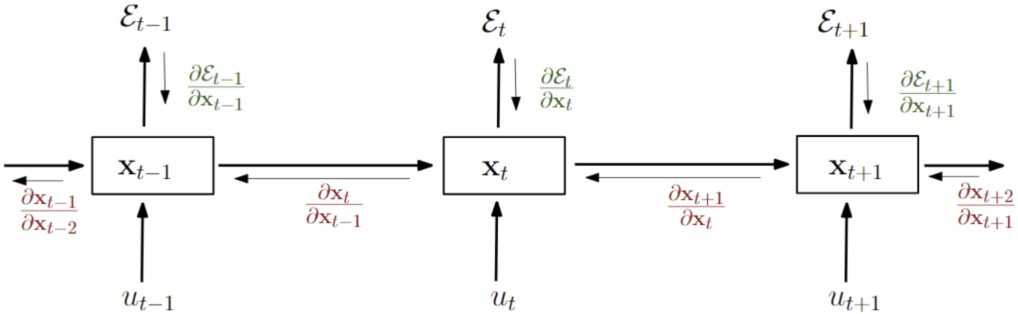


Karpathy et al., 2016

Backpropagation through time problems

$$\frac{\partial \mathcal{L}}{\partial W_{hh}} = \sum_t \sum_{k=1}^{t+1} \frac{\partial \mathcal{L}(t+1)}{\partial z_{t+1}} \frac{\partial z_{t+1}}{\partial \mathbf{h}_{t+1}} \frac{\partial \mathbf{h}_{t+1}}{\partial \mathbf{h}_k} \frac{\partial \mathbf{h}_k}{\partial W_{hh}}$$

$$\frac{\partial \mathcal{L}}{\partial W_{xh}} = \sum_t \sum_{k=1}^{t+1} \frac{\partial \mathcal{L}(t+1)}{\partial z_{t+1}} \frac{\partial z_{t+1}}{\partial \mathbf{h}_{t+1}} \frac{\partial \mathbf{h}_{t+1}}{\partial \mathbf{h}_k} \frac{\partial \mathbf{h}_k}{\partial W_{xh}}$$



$$\forall k, \left\| \frac{\partial \mathbf{x}_{k+1}}{\partial \mathbf{x}_k} \right\| \leq \left\| \mathbf{W}_{rec}^T \right\| \left\| diag(\sigma'(\mathbf{x}_k)) \right\| < \frac{1}{\gamma}\gamma < 1$$

$$\frac{\partial \mathcal{E}_t}{\partial \mathbf{x}_t} \left(\prod_{i=k}^{t-1} \frac{\partial \mathbf{x}_{i+1}}{\partial \mathbf{x}_i} \right) \leq \eta^{t-k} \frac{\partial \mathcal{E}_t}{\partial \mathbf{x}_t}$$

Bengio et al., 1994

Pascanu et.al, 2013

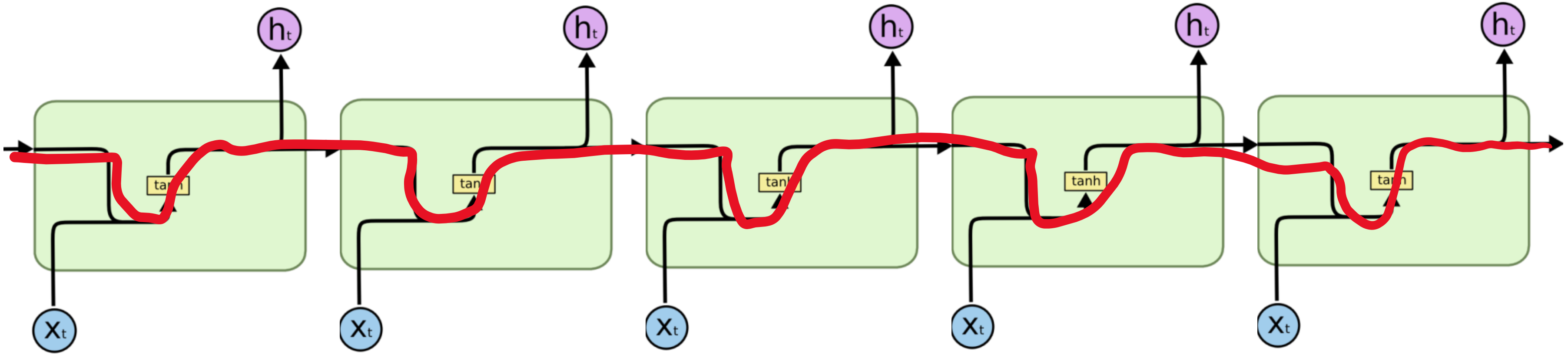
Largest singular value of W:

- >1: Exploding gradients
  -> gradient clipping
- <1: Vanishing gradient

Inherent problem of vanilla RNN!

RNN

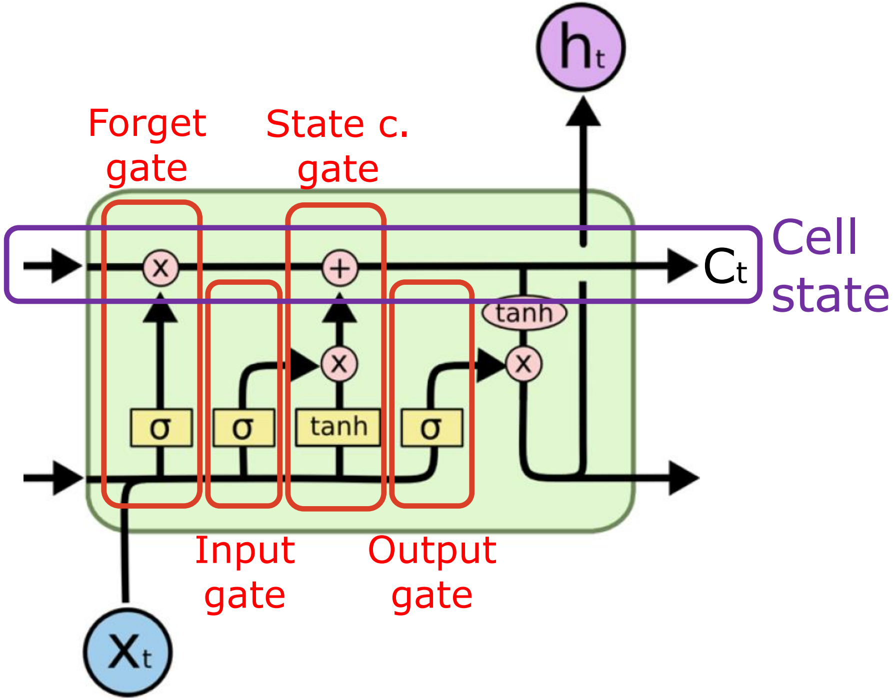
- Backpropagation through time problem



*[Images from:
Christopher Olah,
Understanding LSTM Networks]*

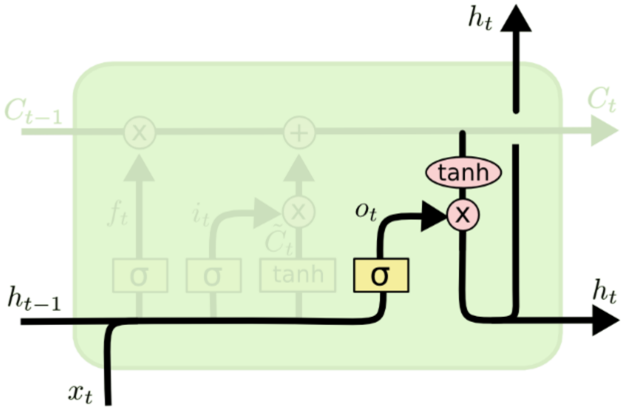
LSTM

- **Long short term memory**
- Additional Cell state
- Forget gate
  - How much to forget the value of the cell state
- Input gate
  - How much to take into account the value of the current input
- State candidate gate
  - Update the old cell state
- Output gate
  - Decide what to output



Hochreiter & Schmidhuber, 1997

LSTM



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

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

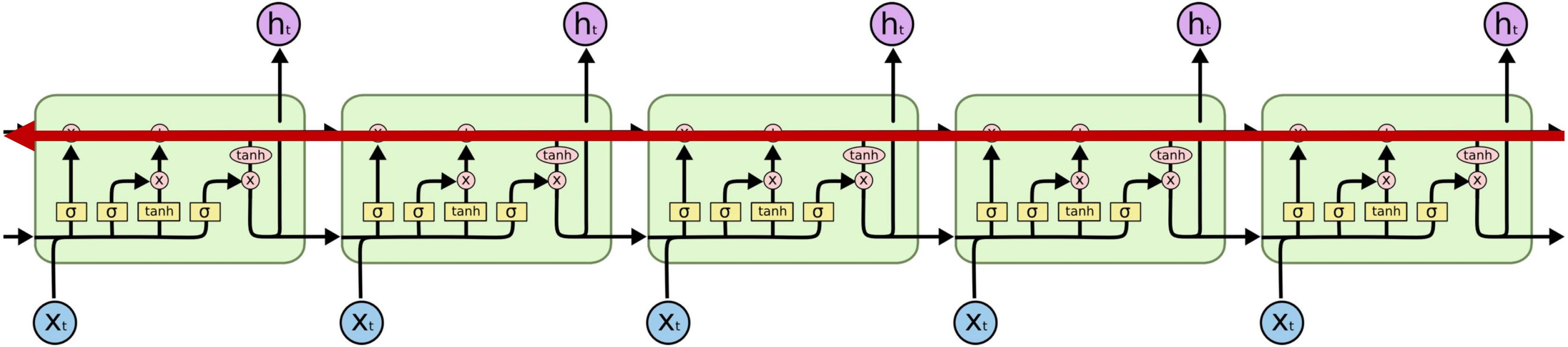
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

$$o_t = \sigma\left(W_o \; [\, h_{t-1}, x_t] \; + \; b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

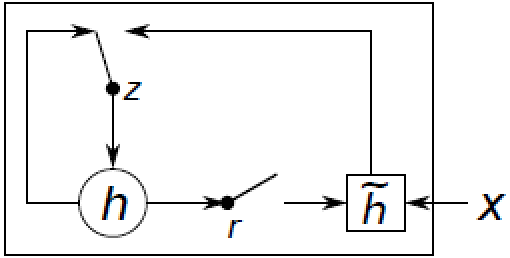
LSTM

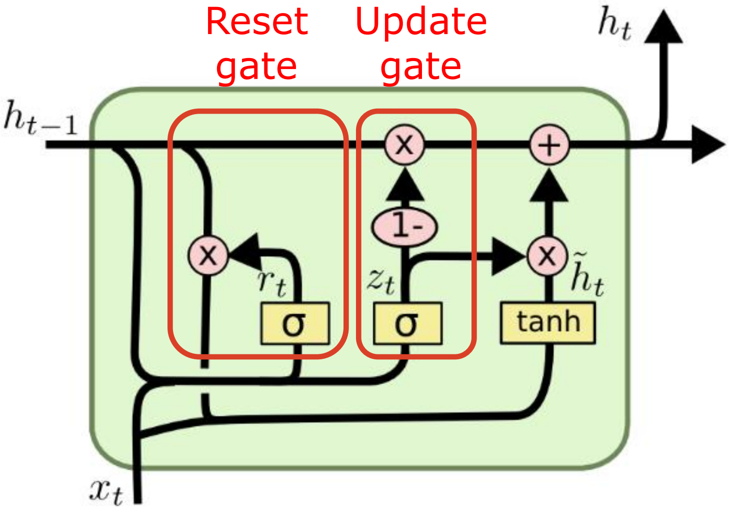
- Backpropagation through time problem solved



GRU

- **Gated Recurrent Units**





$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

Cho et al., 2014

RNN variants

MUT1:

$$
\begin{aligned}
z &= \mathrm{sigm}(W_{\mathrm{xz}} x_t + b_{\mathrm{z}}) \\
r &= \mathrm{sigm}(W_{\mathrm{xr}} x_t + W_{\mathrm{hr}} h_t + b_{\mathrm{r}}) \\
h_{t+1} &= \tanh(W_{\mathrm{hh}}(r \odot h_t) + \tanh(x_t) + b_{\mathrm{h}}) \odot z \\
&+ h_t \odot (1 - z)
\end{aligned}
$$

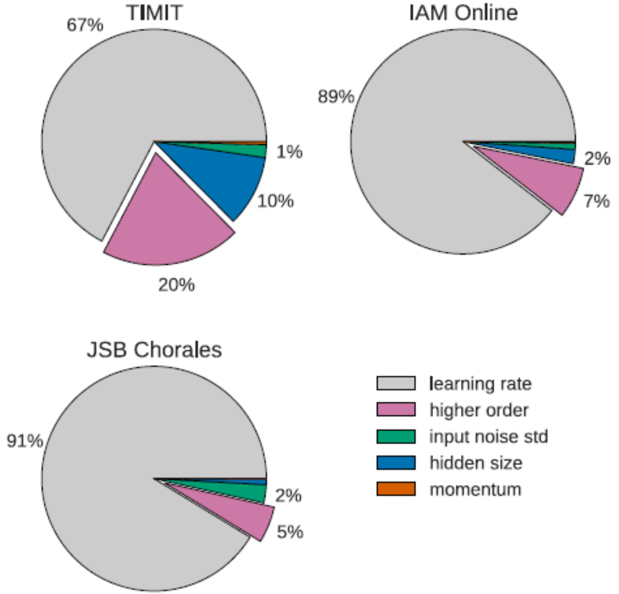
MUT2:

$$
\begin{aligned}
z &= \mathrm{sigm}(W_{\mathrm{xz}} x_t + W_{\mathrm{hz}} h_t + b_{\mathrm{z}}) \\
r &= \mathrm{sigm}(x_t + W_{\mathrm{hr}} h_t + b_{\mathrm{r}}) \\
h_{t+1} &= \tanh(W_{\mathrm{hh}}(r \odot h_t) + W_{xh} x_t + b_{\mathrm{h}}) \odot z \\
&+ h_t \odot (1 - z)
\end{aligned}
$$

MUT3:

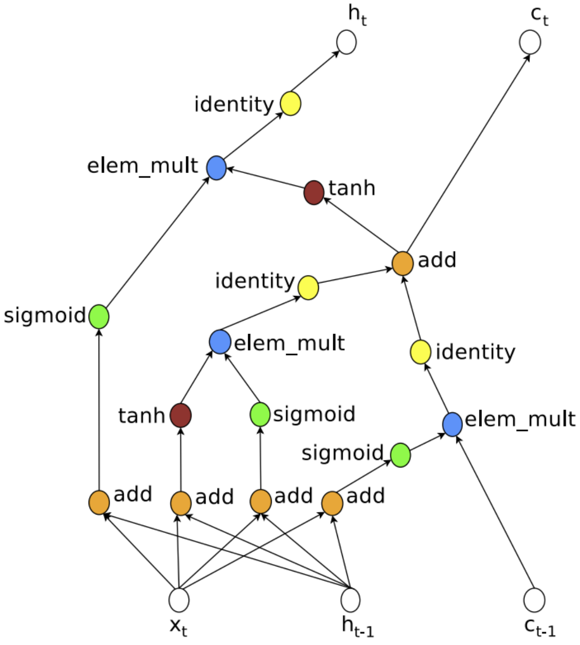
$$
\begin{aligned}
z &= \mathrm{sigm}(W_{\mathrm{xz}} x_t + W_{\mathrm{hz}} \tanh(h_t) + b_{\mathrm{z}}) \\
r &= \mathrm{sigm}(W_{\mathrm{xr}} x_t + W_{\mathrm{hr}} h_t + b_{\mathrm{r}}) \\
h_{t+1} &= \tanh(W_{\mathrm{hh}}(r \odot h_t) + W_{xh} x_t + b_{\mathrm{h}}) \odot z \\
&+ h_t \odot (1 - z)
\end{aligned}
$$

Jozefowicz *et al., 2015*



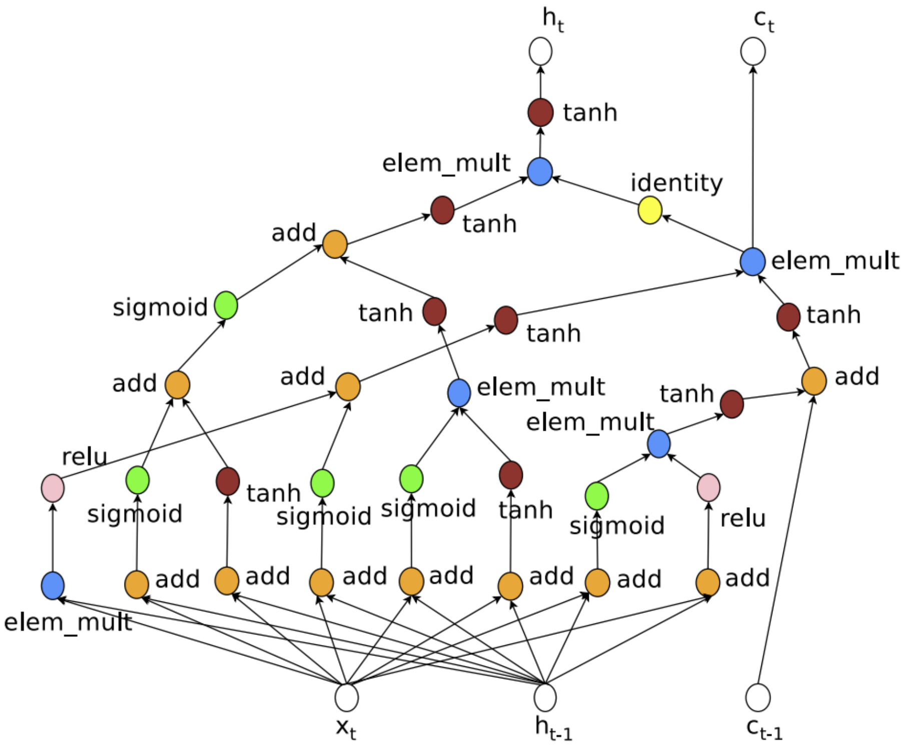
Greff *et al., 2015*

RNN NAS



LSTM cell

[Zoph & Le, *2017*]

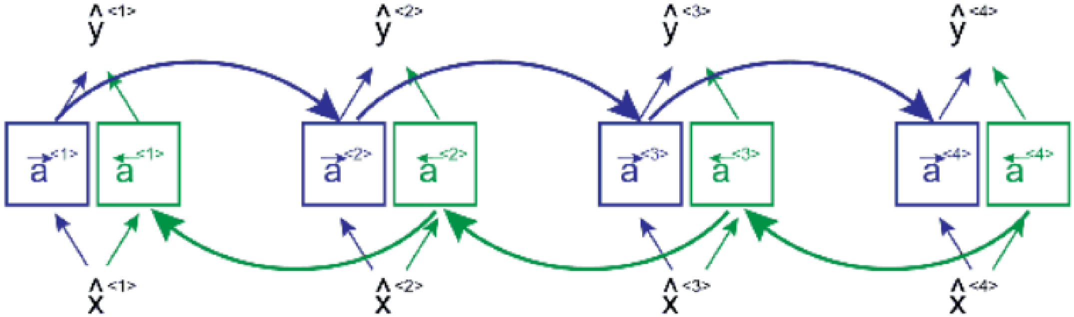


Discovered cell

Bidirectional LSTM

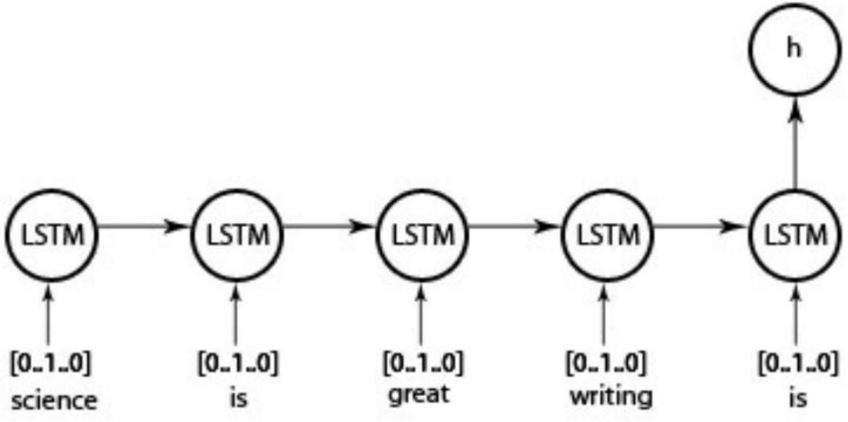
- BRNN
- Two LSTMs
- The output depends on both RNNs
- Considering context from both directions
  - The entire sequence is needed

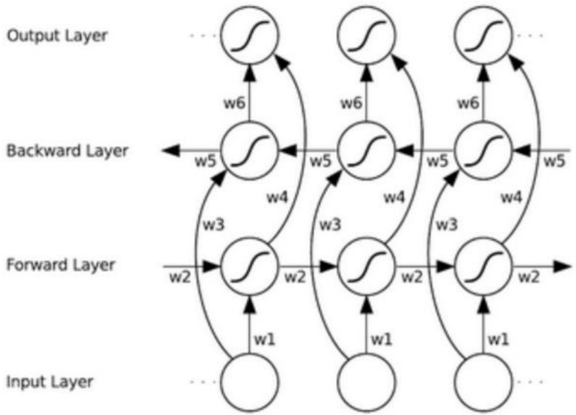
$$\hat{y}^{<t>} = g(W_y [\overrightarrow{a}^{<t>}, \overleftarrow{a}^{<t>}] + b_y)$$



[Images from medium.com]

Example: sentiment analysis

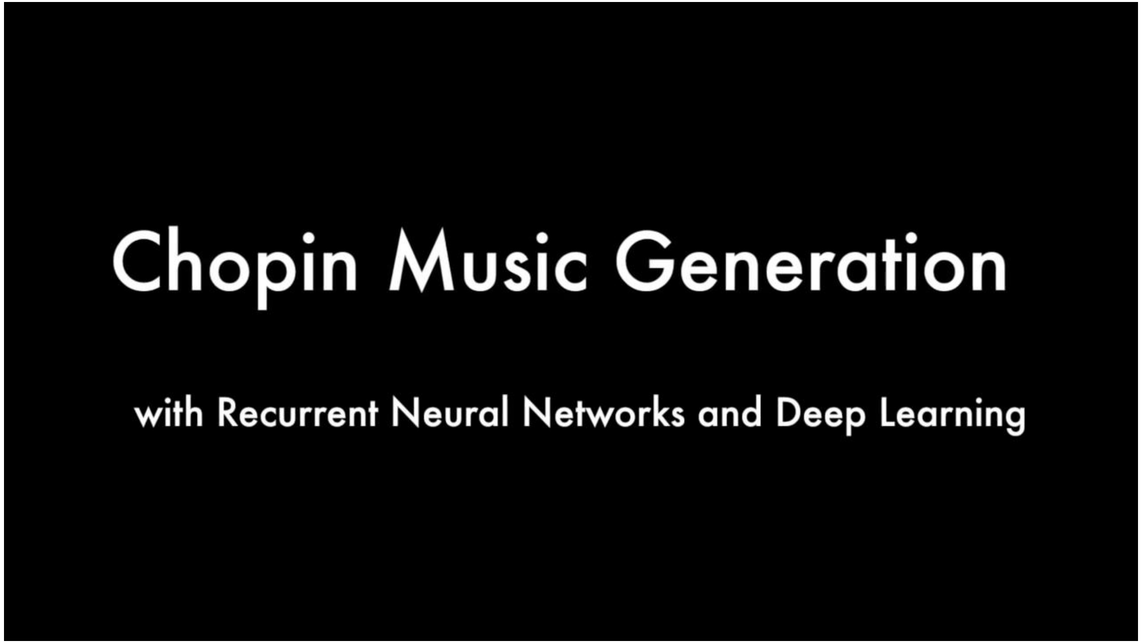




Dictionary size	16201
Number of outputs	3 (good, neutral, bad)
Dimension, hidden layer	140
Accuracy, LSTM	84.415%
Accuracy Bidirectional LSTM	**86.4%**
Accuracy GRU	75.821%

Nowak & Scherer, *2017*

Example: music generation

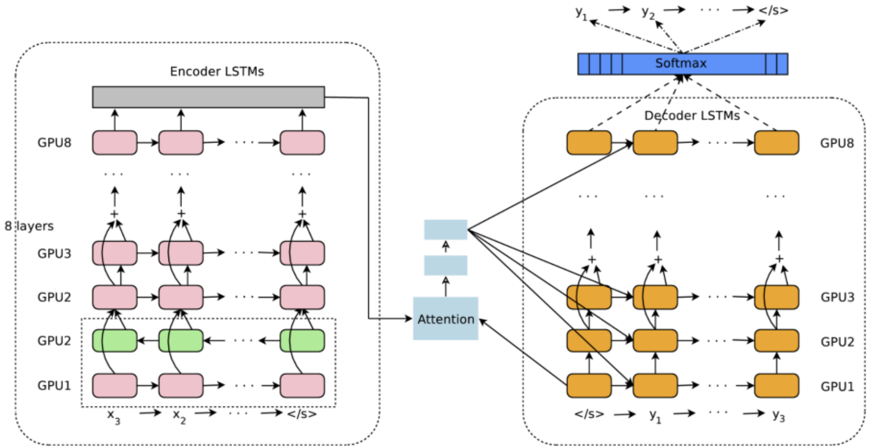


https://www.youtube.com/watch?v=j60J1cGINX4

Example: Machine translation

- Google's Neural Machine Translation system (2016)

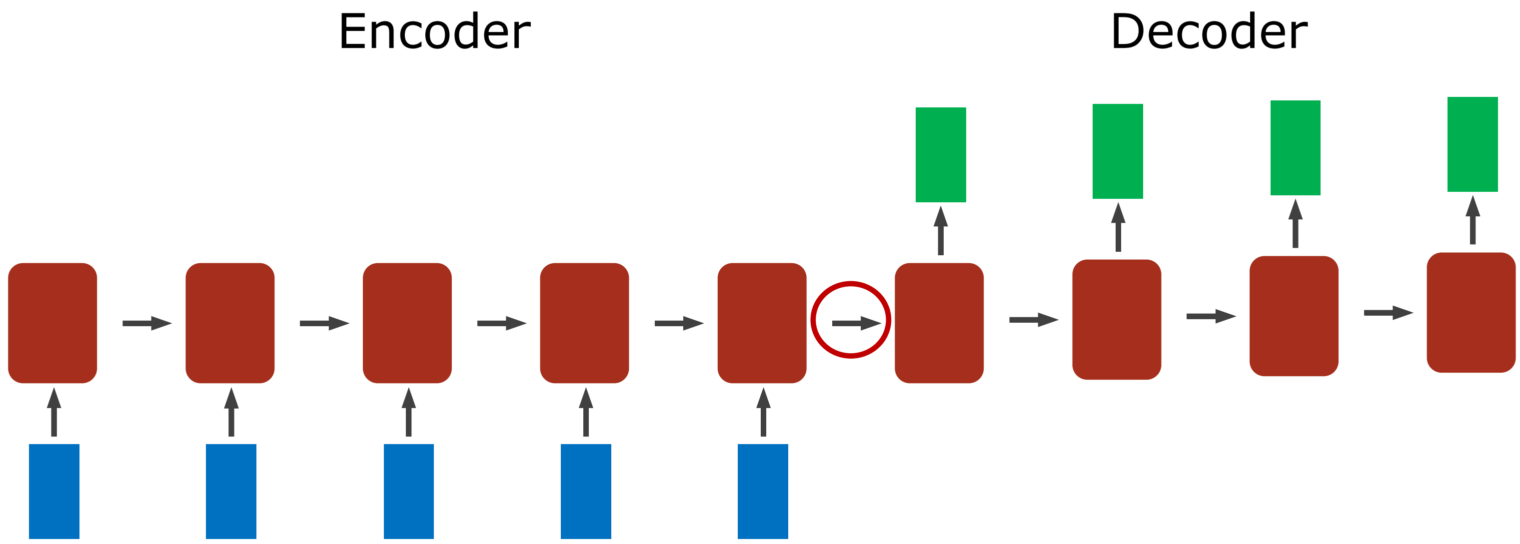
[Wu et al., 2016]



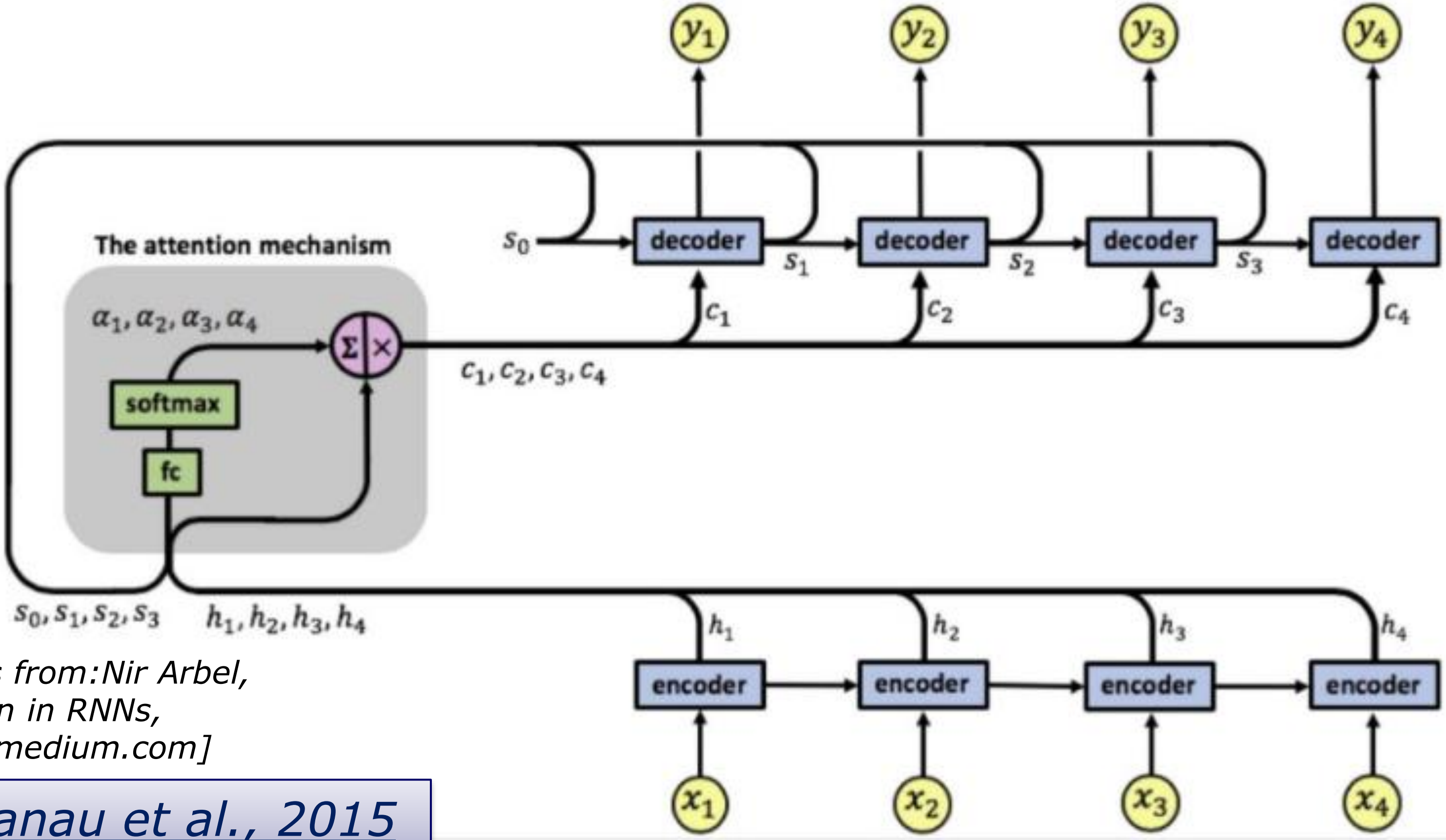
Source	Analysts believe the country is unlikely to slide back into full-blown conflict, but recent events have unnerved foreign investors and locals.	
PBMT	Les analystes estiment que le pays a peu de chances de retomber dans un conflit total, mais les événements récents ont inquiété les investisseurs étrangers et locaux.	5.0
GNMT	Selon les analystes, il est peu probable que le pays retombe dans un conflit généralisé, mais les événements récents ont attiré des investisseurs étrangers et des habitants locaux.	2.0
Human	Les analystes pensent que le pays ne devrait pas retomber dans un conflit ouvert, mais les récents évènements ont ébranlé les investisseurs étrangers et la population locale.	5.0

Encoder – decoder architecture

- E.g., machine translation, sequence to sequence modelling



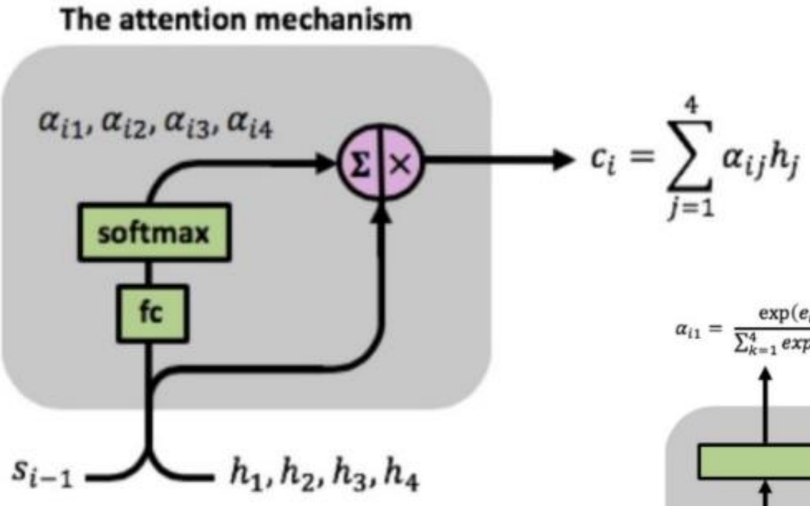
Attention in RNNs

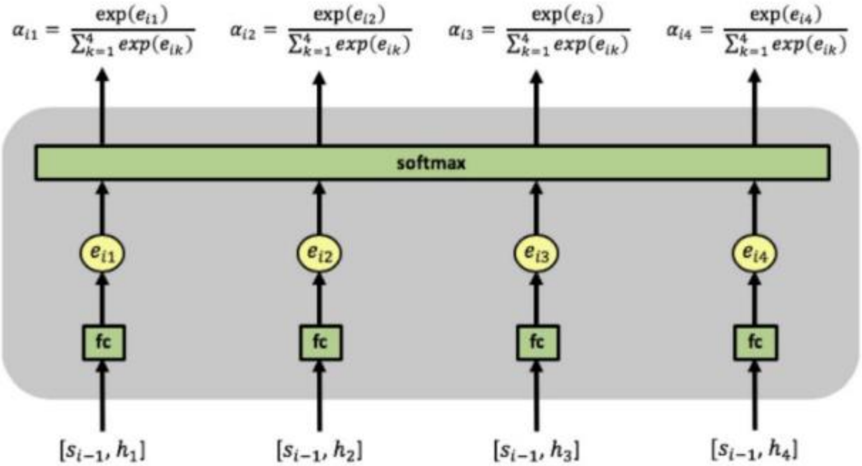


[Images from:Nir Arbel, Attention in RNNs, https://medium.com]

Bahdanau et al., 2015

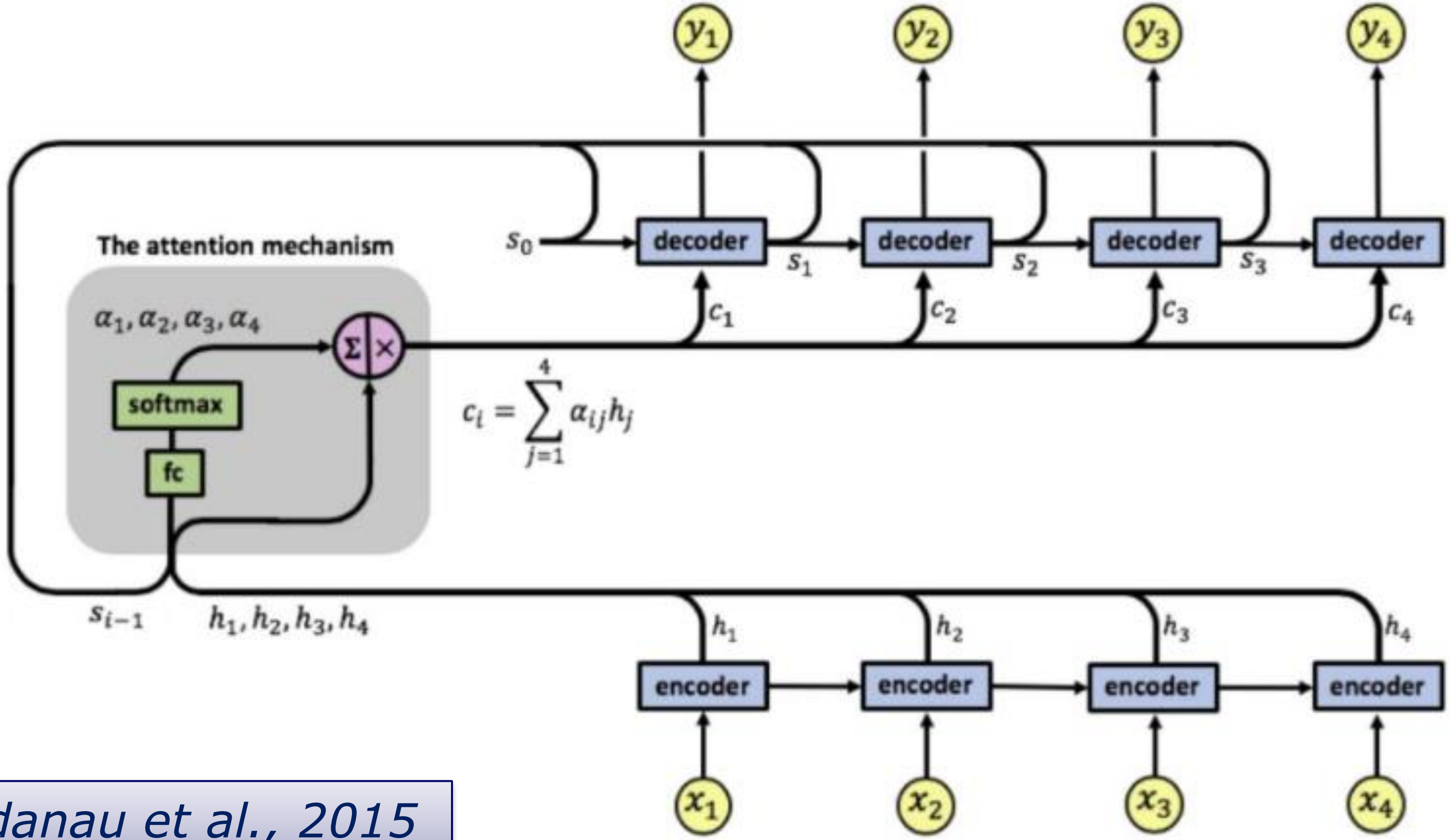
Context vectors





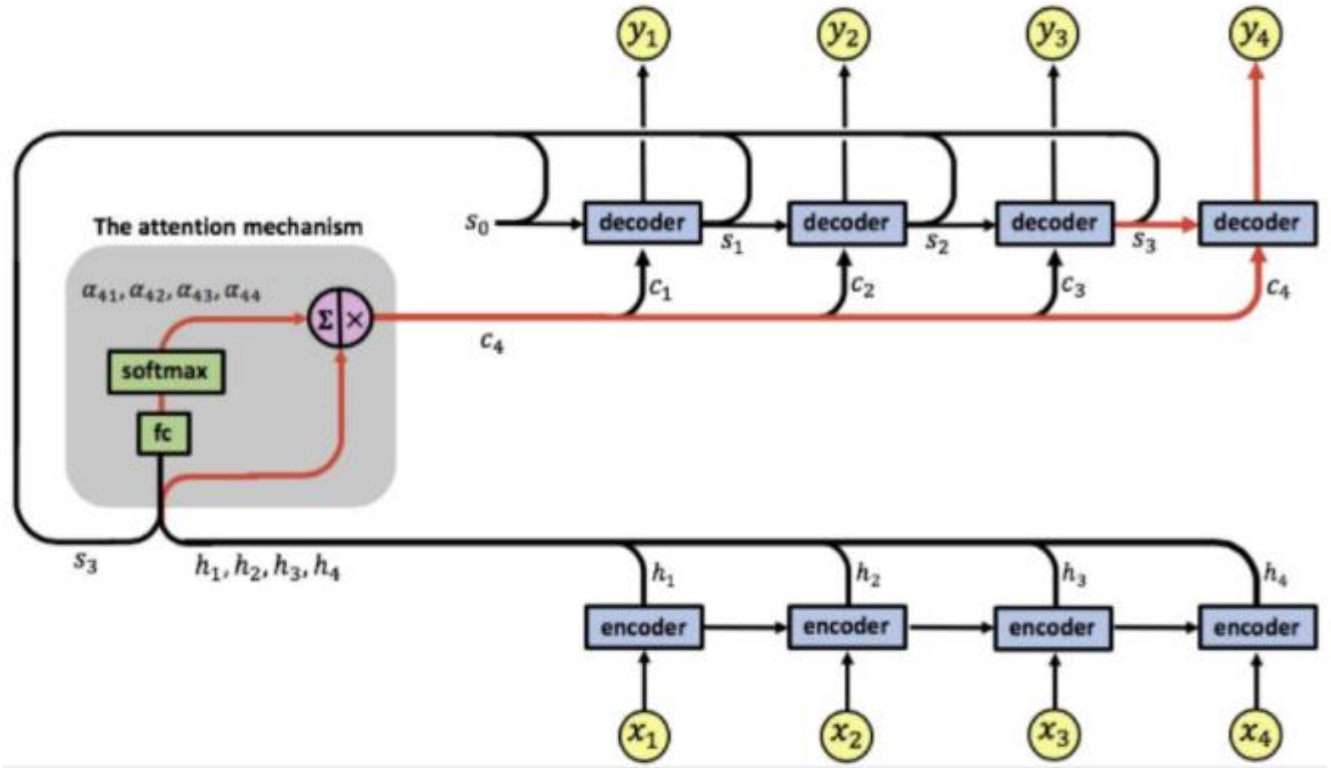
Bahdanau et al., 2015

Attention in RNNs



Bahdanau et al., 2015

Computing the context vectors

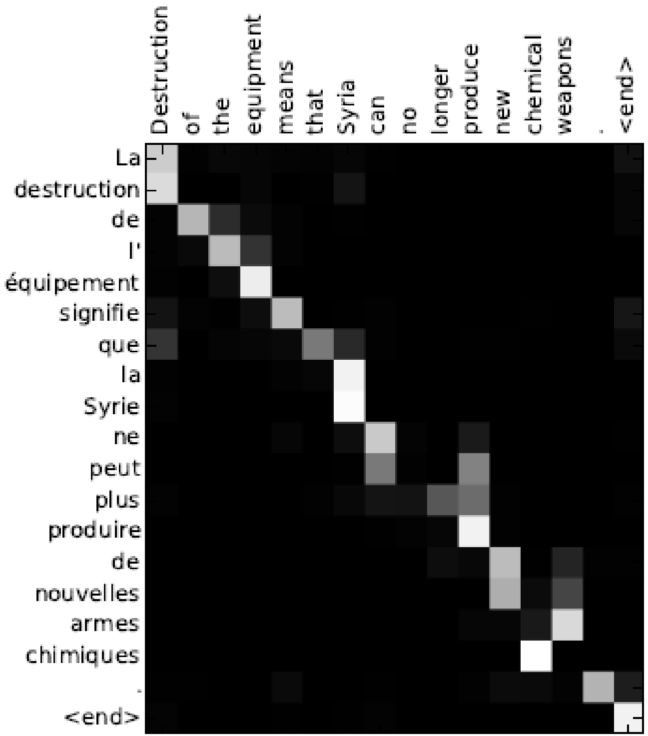


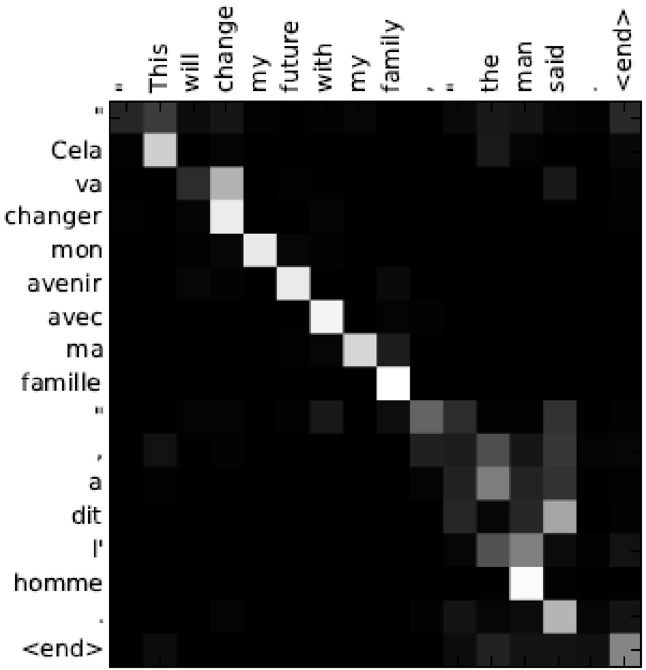
Bahdanau et al., 2015

Example of attention weights

- Translation between English and French

[Bahdanau et al., 2015]

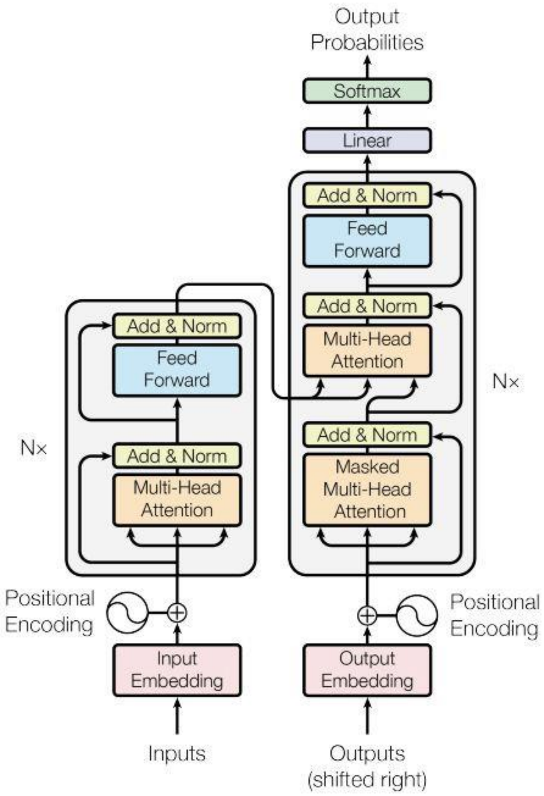




Attention++

- Attention is all you need
- Vaswani et.al, NIPS 2017

- Transformers!



[Vaswani et al., 2017]