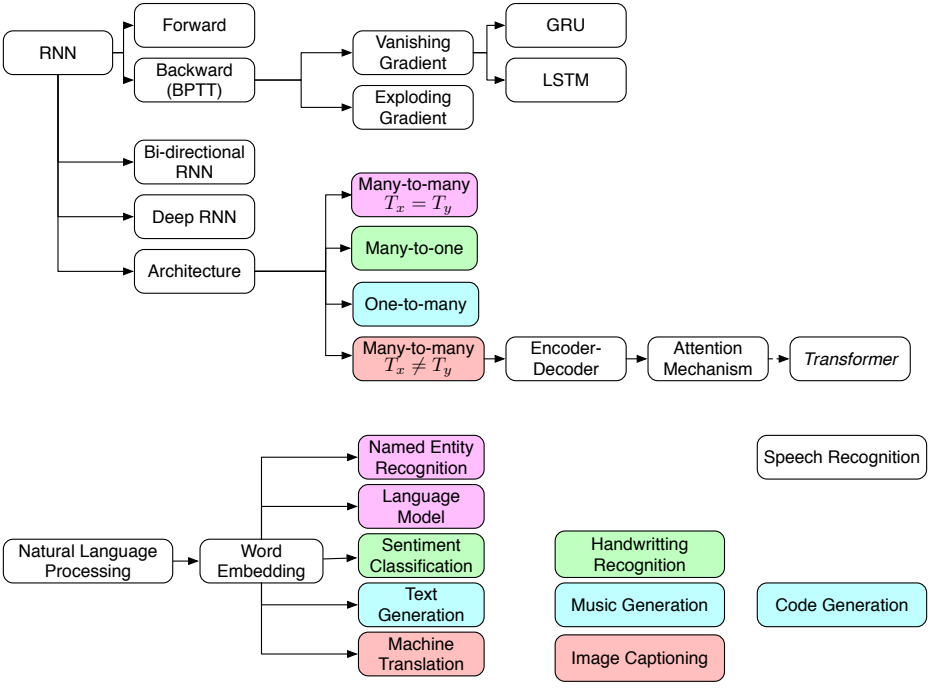
Deep Learning
Topics: Recurrent Neural Networks for Sequential Data

Geoffroy Peeters

LTCI, Télécom Paris, IP Paris

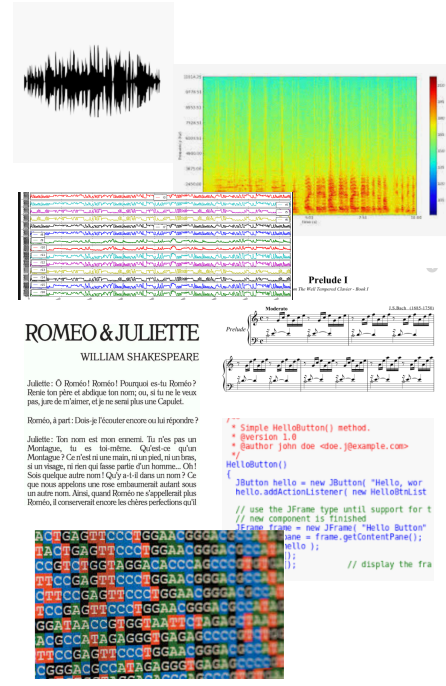
-


RNN
Forward
Backward (BPTT)
Vanishing Gradient
Exploding Gradient
GRU
LSTM
Bi-directional RNN
Deep RNN
Architecture
Many-to-many $T_x = T_y$
Many-to-one
One-to-many
Many-to-many $T_x \neq T_y$
Encoder-Decoder
Attention Mechanism
Transformer
Natural Language Processing
Word Embedding
Named Entity Recognition
Language Model
Sentiment Classification
Text Generation
Machine Translation
Speech Recognition
Handwritting Recognition
Music Generation
Code Generation
Image Captioning

Recurrent Neural Networks for Sequencial Data

What are sequential data ?

- **Sequential data are any type of data for which the order matters**
- Examples :
  - **Time series :**
    - Audio (sequence of sound pressures or sequence of Fourier transforms)
    - Sequence of musical notes ©
    - Video (sequence of images)
    - Sensors (heart-beat, plane altitude)
    - Stock market
  - **Ordered :** (but not with time)
    - Text/Words ©
    - Genes/ADN ©
- Common models :
  - Vector linear autoregression (VAR)
  - Markov model (Discrete-State HMMs)
  - Kalman filters (Continuous-State HMMs)
  - ...



Notations

- Inputs :
  - $\underline{x}^{\langle t \rangle}$: input vectors $\underline{x}$ at time t (of dimension $n^{[0]}$)
  - $\{\underline{x}^{\langle 1 \rangle}, \underline{x}^{\langle 2 \rangle}, \dots \underline{x}^{\langle T_x \rangle}\}$ the sequence of inputs of length T_x
- Outputs :
  - $\underline{y}^{\langle t \rangle}$: output vectors $\underline{y}$ at time t
  - $\{\underline{y}^{\langle 1 \rangle}, \underline{y}^{\langle 2 \rangle}, \dots \underline{y}^{\langle T_y \rangle}\}$ the sequence of outputs of length T_y
- T_x and T_y can be of different lengths
- $x^{(i)<t>}$ is the i^{th} example in the training-set
  - ${T_x}^{(i)}$ the length of the i^{th} input sequence
  - ${T_y}^{(i)}$ the length of the i^{th} output sequence

Notations

- Different types of output $\underline{y}$:
  - Continuous values $\underline{y} \in \mathbb{R}^{n_y}$ (Activation : linear, Loss : MSE)
  - Binary-classes $\underline{y} \in \{0,1\}$ (Activation : sigmoid, Loss : Binary-Cross-Entropy)
  - Multi-classes $y \in \{1 \dots K\} \Rightarrow \underline{y} \in \{0,1\}^K$ (Activation : softmax, Loss : Cross-Entropy)
- Different types of input $\underline{x}$
  - Continuous values $\underline{x} \in \mathbb{R}^{n_x}$
  - Categorical values $x \in \{1 \dots K\} \Rightarrow \underline{x} \in \{0,1\}^K$
- **Example** : categorical inputs/ binary outputs
  - "Named Entity Recognition"

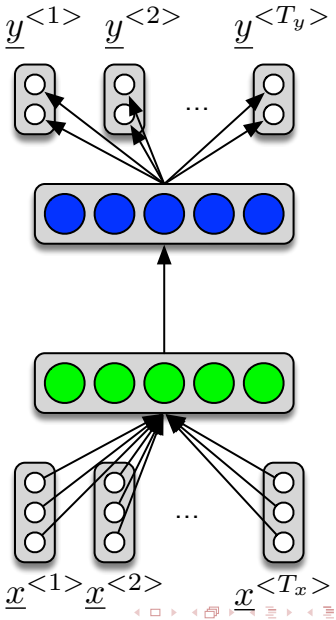
{x}	$x^{<1>}$	$x^{<2>}$	…				$x^{<7>}$
	Emmanuel	Macron	is	the	president	of	France
{y}	$y^{<1>}$	$y^{<2>}$	…				$y^{<7>}$
	1	1	0	0	0	0	1

- **How do we process categorical ©values as input ?**
  - One-hot-encoding
  - Embedding

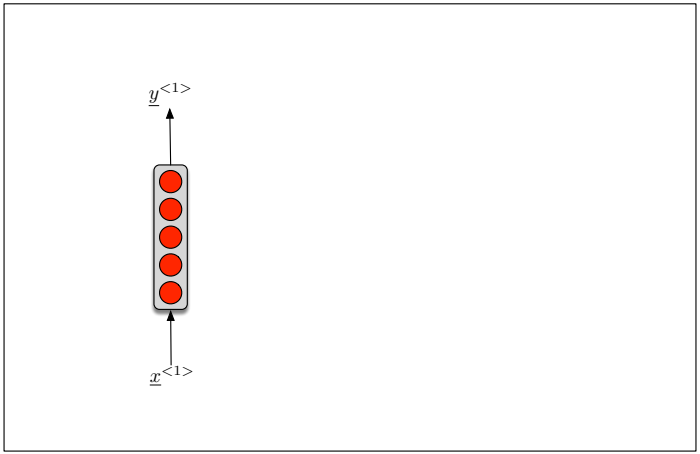
How can we process Sequential data with a Neural Network

- **Why Multi-Layer-Perceptron/Fully-connected networks are not appropriate ?**

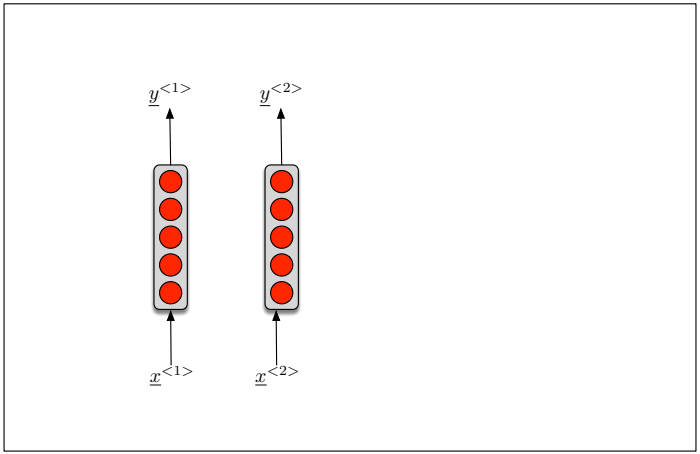
- We could represent an input sequence $\{\underline{x}^{\langle 1 \rangle} \cdots \underline{x}^{\langle T_x \rangle}\}$ as a large-vector of dimension $(n^{[0]} T_x)$
- but (because $T_x{}^{(i)}$ can be different for each input sequence) the dimension of this large-vector would change for each
  - zero-padding ?
- still require a huge input dimension $(n^{[0]} T_x)$
- the network will have to learn different weights for the same dimension $n^{[0]}$ at various time $\langle t \rangle$ in the sequence
  - no weight sharing !
- **Solutions :**
  - Recurrent Neural Network
  - 1D Convolutional Neural Network (convolution only over time)



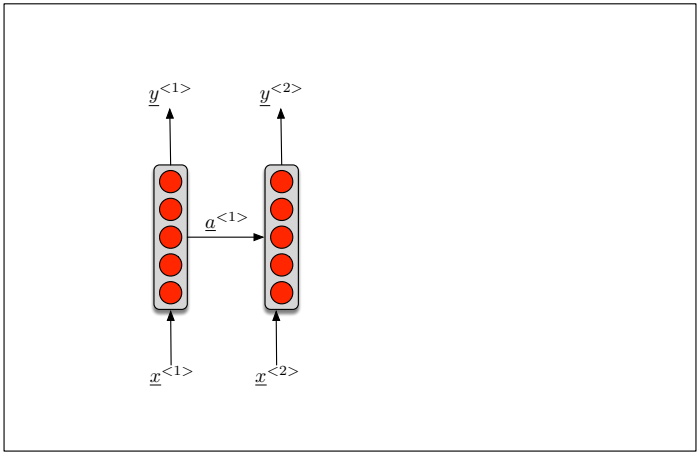
What is an RNN ?



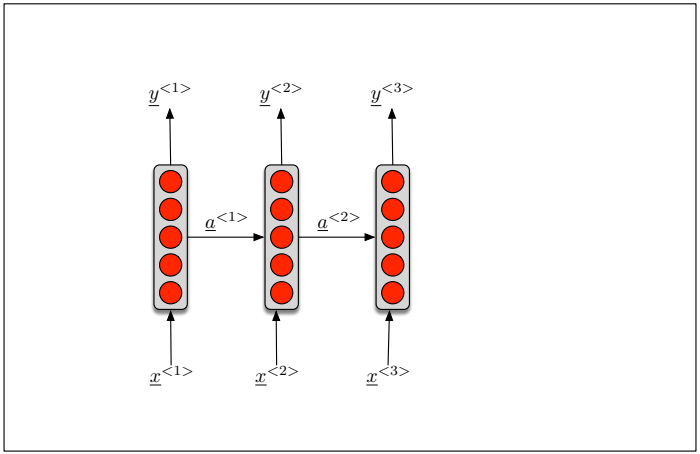
What is an RNN ?



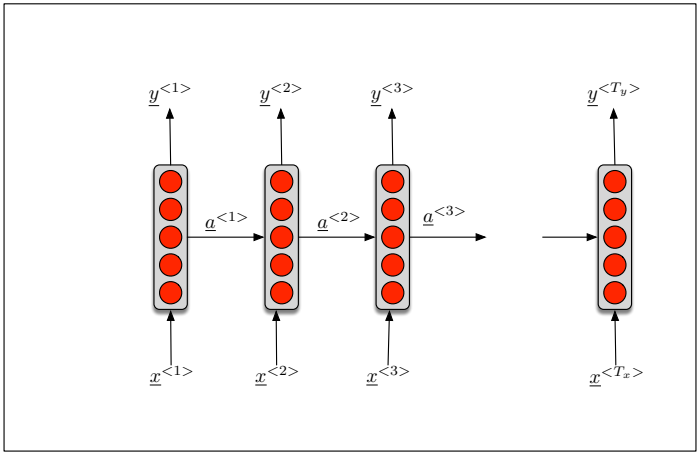
What is an RNN ?



What is an RNN ?

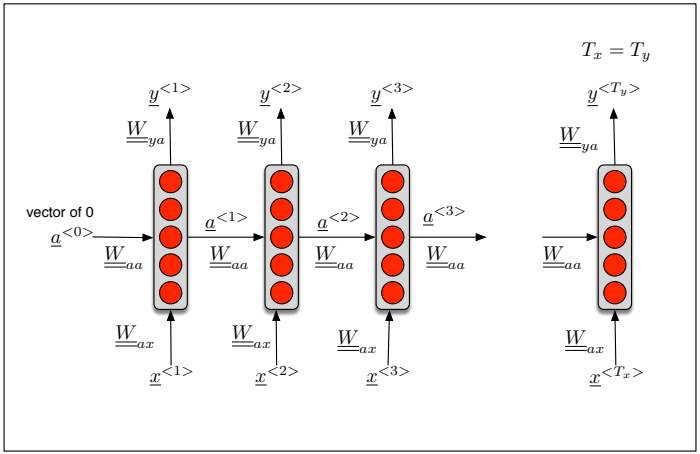


What is an RNN ?



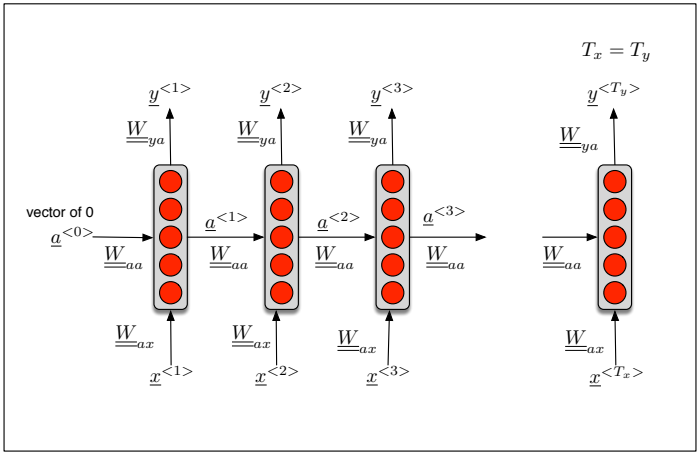
- At each time step $< t >$, the RNN passes its activation $a^{\langle t \rangle}$ to the next time step $< t + 1 >$

What is an RNN ?



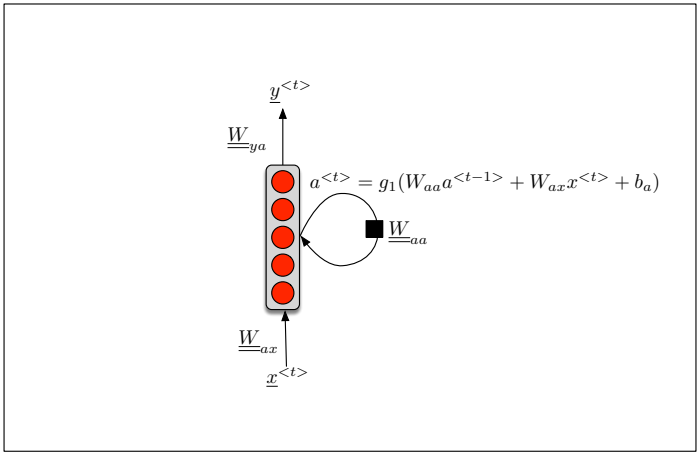
- Initialize the first activation $a^{\langle 0 \rangle}$ as zero

What is an RNN ?



- $\underline{\underline{W}}_{ax}$ ($x^{\langle t \rangle} \rightarrow a^{\langle t \rangle}$) is the same for every time step
- $\underline{\underline{W}}_{aa}$ ($a^{\langle t-1 \rangle} \rightarrow a^{\langle t \rangle}$) is the same for every time step
- $\underline{\underline{W}}_{ya}$ ($a^{\langle t \rangle} \rightarrow y^{\langle t \rangle}$) is the same for every time step

Other possible representation



- $\underline{\underline{W}}_{ax}$ ($x^{\langle t \rangle} \rightarrow a^{\langle t \rangle}$) is the same for every time step
- $\underline{\underline{W}}_{aa}$ ($a^{\langle t-1 \rangle} \rightarrow a^{\langle t \rangle}$) is the same for every time step
- $\underline{\underline{W}}_{ya}$ ($a^{\langle t \rangle} \rightarrow y^{\langle t \rangle}$) is the same for every time step

- **Forward propagation** :

$$a^{\langle 0 \rangle} = 0$$

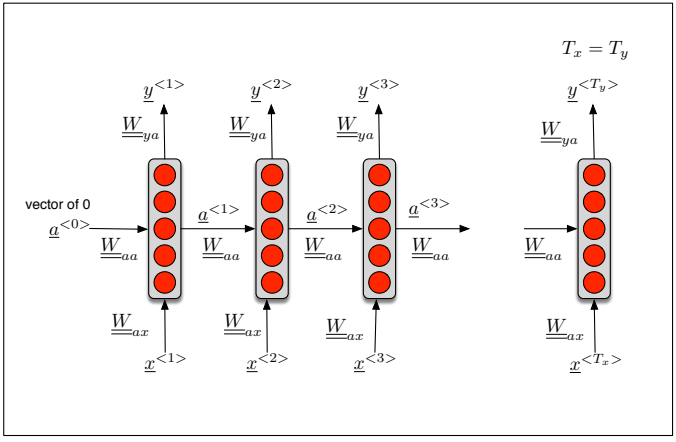
$$a^{\langle 1 \rangle} = g_1\left(W_{aa}a^{\langle 0 \rangle} + W_{ax}x^{\langle 1 \rangle} + b_a\right)$$

$$\hat{y}^{\langle 1 \rangle} = g_2\left(W_{ya}a^{\langle 1 \rangle} + b_y\right)$$

$$\dots$$

$$a^{\langle t \rangle} = g_1\left(W_{aa}a^{\langle t-1 \rangle} + W_{ax}x^{\langle t \rangle} + b_a\right)$$

$$\hat{y}^{\langle t \rangle} = g_2\left(W_{ya}a^{\langle t \rangle} + b_y\right)$$



- **Compact notations** :

$$\underset{(n_a,n_a)}{[\,W_{aa}} \mid \underset{(n_a,n_x))}{W_{ax}\,}] \begin{bmatrix} a^{\langle t-1 \rangle} \\ {\scriptstyle (n_a,m)} \\ x^{\langle t \rangle} \\ {\scriptstyle (n_x,m)} \end{bmatrix}$$

$$a^{\langle t \rangle} = g_1\left(W_a[a^{\langle t-1 \rangle}; x^{\langle t \rangle}] + b_a\right)$$

$$\hat{y}^{\langle t \rangle} = g_2\left(W_y a^{\langle t \rangle} + b_y\right)$$

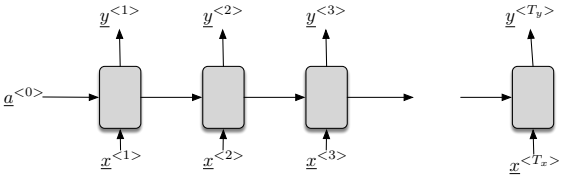
Various types of sequential data

Various types of sequential data/ Many-To-Many $T_y = T_x$

Input $\underline{x}$	Output $\underline{y}$	Type	Examples
sequence $T_x > 1$	sequence $T_y = T_x$	Many-To-Many	Named entity recognition

In 1917, Einstein applied the general theory of relativity to model the large-scale structure of the universe. He was visiting the United States when Adolf Hitler came to power in 1933 and did not go back to Germany, where he had been a professor at the Berlin Academy of Sciences. He settled in the U.S., becoming an American citizen in 1940. On the eve of World War II, he endorsed a letter to President Franklin D. Roosevelt alerting him to the potential development of "extremely powerful bombs of a new type" and recommending that the U.S. begin similar research. This eventually led to what would become the Manhattan Project. Einstein supported defending the Allied forces, but largely denounced using the new discovery of nuclear fission as a weapon. Later, with the British philosopher Bertrand Russell, Einstein signed the Russell–Einstein Manifesto, which highlighted the danger of nuclear weapons. Einstein was affiliated with the Institute for Advanced Study in Princeton, New Jersey, until his death in 1955.

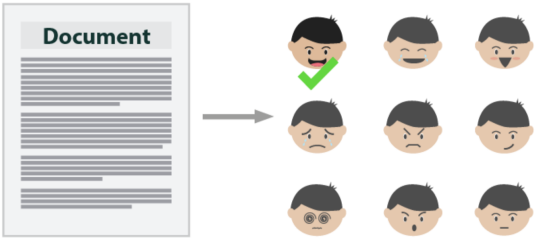
Tag colours:
LOCATION TIME PERSON ORGANIZATION MONEY PERCENT DATE

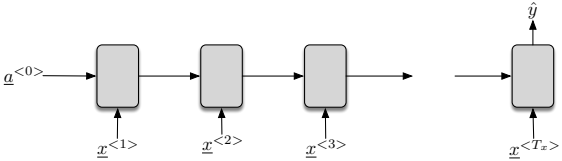


- **Many-to-many** : $T_x = T_y$

Various types of sequential data/ Many-To-One $T_x > 1, T_y = 1$

Input $\underline{x}$	Output $\underline{y}$	Type	Examples
sequence $T_x > 1$	single $T_y = 1$	Many-To-One	Sentiment analysis, Video activity detection



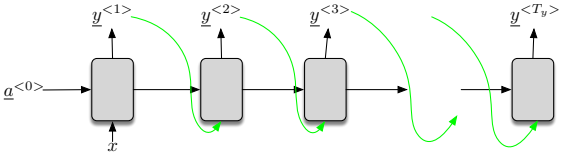


- **Many-to-one** : sentiment classification
  - x = text (movie review)
  - y = 0/1 or $1 \cdots 5$ rating

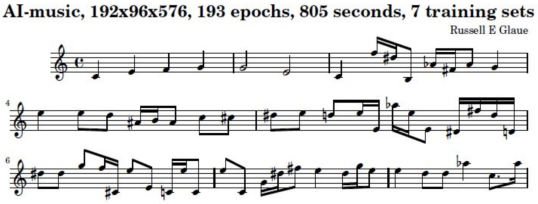
Various types of sequential data/ One-To-Many $T_x = 1, T_y > 1$

Input $\underline{x}$	Output $\underline{y}$	Type	Examples
single $T_x = 1$	sequence, $T_y > 1$	One-To-Many	Text generation, music generation

```
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.]]
```

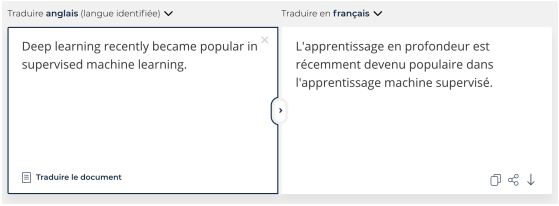


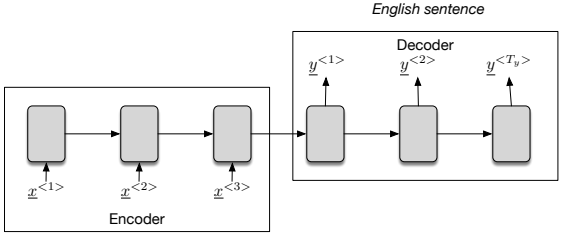
- **One-to-many :** text, code, music generation



Various types of sequential data/ Many-To-Many $T_y \neq T_x$

Input $\underline{x}$	Output $\underline{y}$	Type	Examples
sequence $T_x > 1$	sequence $T_y > 1, T_y \neq T_x$	Many-To-Many	Automatic Speech Recognition, Machine translation

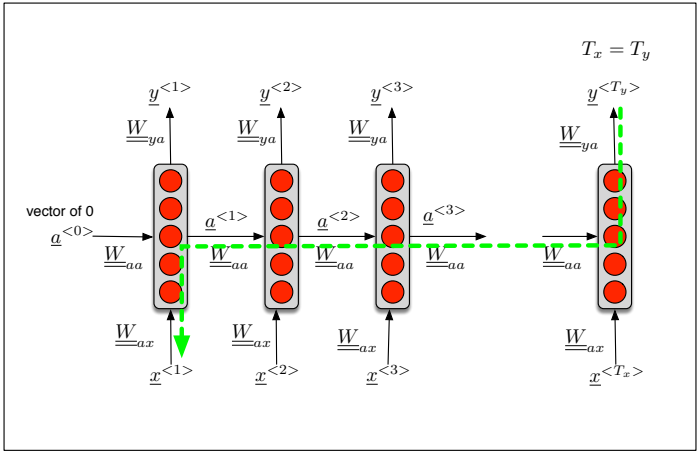




- **Many-to-many** : $T_x \neq T_y$: Machine translation
  - Encoder/Decoder
  - **Attention** mechansim
  - **Transformer** architecture

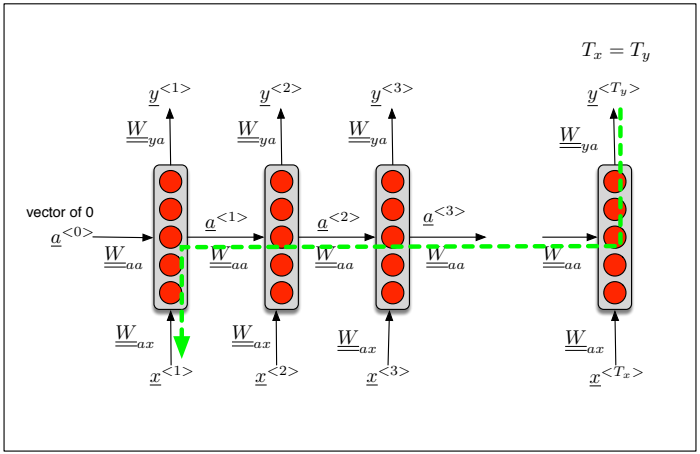
Different architectures

RNN



- RNN scans through the data from left to right ⇒ can do skip connections in the past
- Question : is "Apple" a named entity ?
  - "The new **iPhone** is sold in Apple stores."
  - "A refreshing and healthy **beverages** is an Apple juice."
- ⇒ we can predict (since RNN depends on the past)

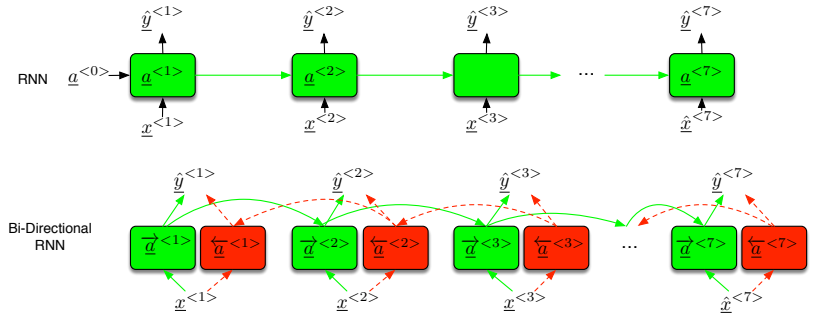
Bi-directional RNN



- RNN scans through the data from left to right ⇒ can do skip connections only in the past
- Question : is "Apple" a named entity ?
  - "Apple stores sale the new **iPhone**."
  - "Apple juice is a refreshing and healthy **beverages**."
- ⇒ we cannot predict (since RNN does not depend on the future)

Bi-directional RNN

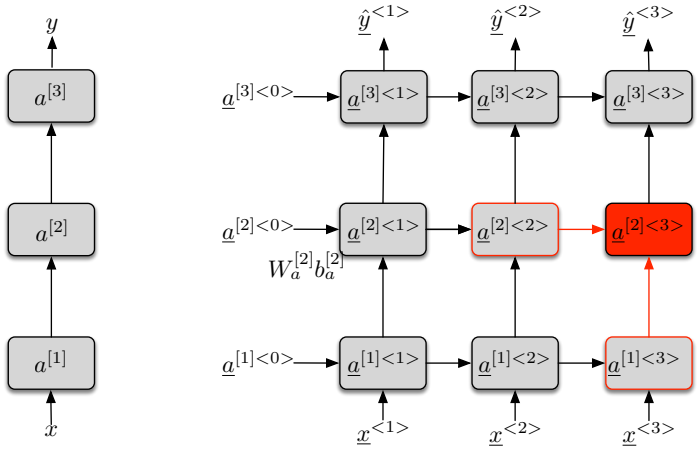
- Solution ?
  - Bi-directional RNN (BRNN)
- Question : is "Apple" a named entity ?
  - "Apple stores sale the new **iPhone**."
  - "Apple juice is a refreshing and healthy **beverages**."
- $\Rightarrow$ we can predict since we now depend on the past and the future



- The prediction $\hat{y}^{\langle t \rangle}$ is done from the concatenation of
  - the **left-to-right** $\overrightarrow{a}^{\langle t \rangle}$ and
  - the **right-to-left** $\overleftarrow{a}^{\langle t \rangle}$ hidden states of RNN

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

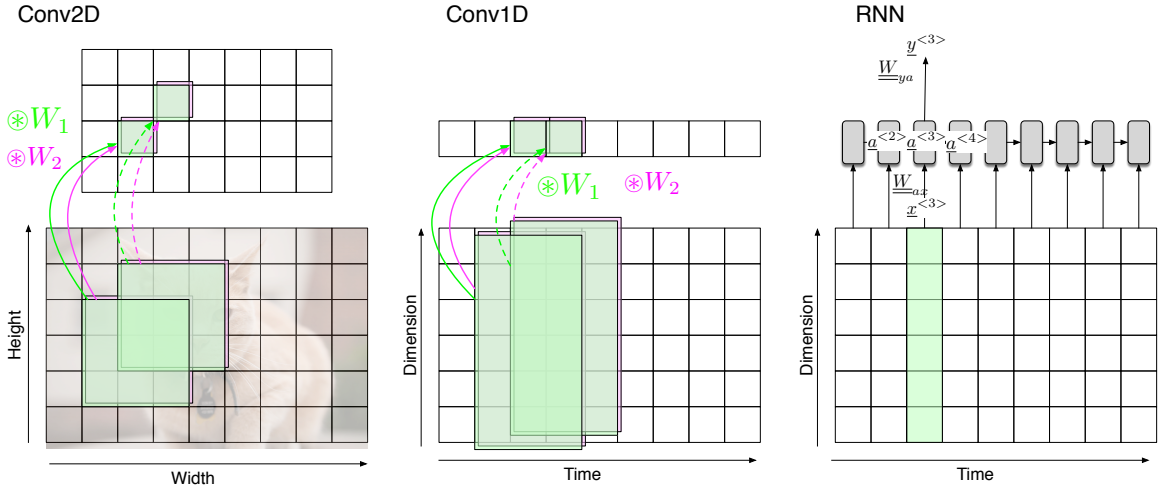
Deep RNN



- For the first layer, the inputs of the cell at time $< t >$ are
  - the input x at the current time $\langle t \rangle$: $x^{\langle t \rangle}$
  - the value of the cell at the previous time $\langle t-1 \rangle$: $a^{\langle t-1 \rangle}$
- For the layer $[l]$, the inputs of the cell at time $\langle t \rangle$ are
  - the value of the cell of the previous layer $[l-1]$ at the current time $\langle t \rangle$: $a^{[l-1]\langle t \rangle}$
  - the value of the cell of the current layer $[l]$ at the previous time $\langle t-1 \rangle$: $a^{[l]\langle t-1 \rangle}$

$$a^{[l]<t>} = g\left(W_a^{[l]}[a^{[l]\langle t-1 \rangle}, a^{[l-1]\langle t \rangle}] + b_a^{[l]}\right)$$

Using 1D Convolution instead of RNN



Back Propagation Through Time (BPTT)

Forward pass

- Compute $a^{<t>}$, $\hat{y}^{<t>}$, $\mathcal{L}^{<t>}$, $\mathcal{L}$

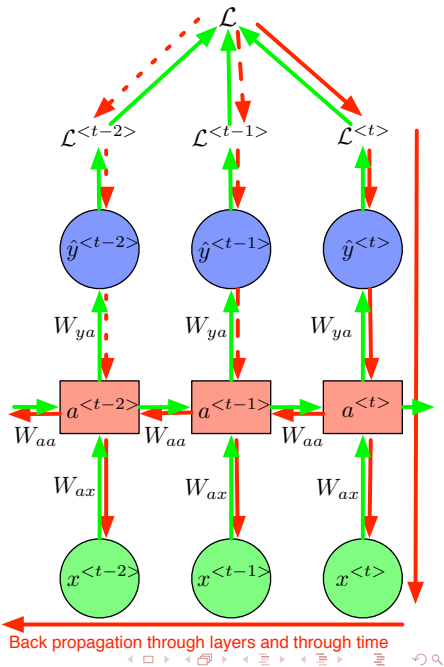
$$a^{<t>} = g_a(W_{ax}x^{<t>} + W_{aa}a^{<t-1>} + b_a)$$

$$\hat{y}^{<t>} = g_y(W_{ya}a^{<t>} + b_y)$$

- Compute the Loss
  - Loss for time $< t >$: $\mathcal{L}^{<t>}(y^{<t>}, \hat{y}^{<t>})$
    - $y^{<t>}$: true label
    - $\hat{y}^{<t>}$: predicted label
  - Total loss :

$$\mathcal{L} = \sum_t \mathcal{L}^{<t>}(y^{<t>}, \hat{y}^{<t>})$$

- **Backward pass**
  - compute $\frac{\partial \mathcal{L}}{\partial W_{ya}}$, $\frac{\partial \mathcal{L}}{\partial W_{ax}}$, $\frac{\partial \mathcal{L}}{\partial W_{aa}}$, $\frac{\partial \mathcal{L}}{\partial b_x}$, $\frac{\partial \mathcal{L}}{\partial b_a}$
  - All weights are shared across time steps ! ! !



Backward pass : 1) compute $\frac{\partial L}{\partial W_{ya}}$

- We measure how much varying W_{ya} affect $\mathcal{L}^{\langle t \rangle}$.

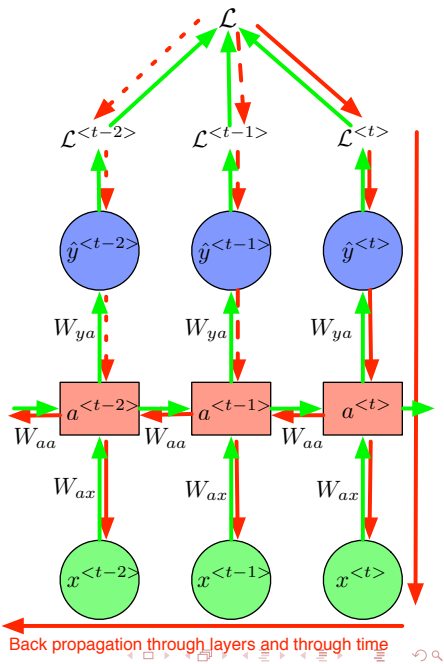
$$\frac{\partial \mathcal{L}}{\partial W_{ya}} = \sum_{t=1}^{T} \frac{\partial \mathcal{L}^{\langle t \rangle}}{\partial W_{ya}}$$

- We can use the chain rule.

$$\frac{\partial \mathcal{L}^{\langle t \rangle}}{\partial W_{ya}} = \frac{\partial \mathcal{L}^{\langle t \rangle}}{\partial \hat{y}^{\langle t \rangle}} \frac{\partial \hat{y}^{\langle t \rangle}}{\partial W_{ya}}$$

- This is very simple since $\hat{y}_t$ depends on W_{ya} on one place (indeed $a^{\langle t \rangle}$ does not depend on W_{ya}).

$$\hat{y}^{\langle t \rangle} = g_y(W_{ya} a^{\langle t \rangle} + b_y)$$



Backward pass : 2) compute $\frac{\partial \mathcal{L}}{\partial W_{aa}}$

- We measure how much varying W_{aa} affect $\mathcal{L}^{\langle t \rangle}$.

$$\frac{\partial \mathcal{L}}{\partial W_{aa}} = \sum_{t=1}^{T} \frac{\partial \mathcal{L}^{\langle t \rangle}}{\partial W_{aa}}$$

- We cannot use the chain rule ($\frac{\partial \mathcal{L}^{\langle t \rangle}}{\partial W_{aa}} = \frac{\partial \mathcal{L}^{\langle t \rangle}}{\partial \hat{y}^{\langle t \rangle}} \frac{\partial \hat{y}^{\langle t \rangle}}{\partial a^{\langle t \rangle}} \frac{\partial a^{\langle t \rangle}}{\partial W_{aa}}$) because a_t depends on W_{aa} also through $a^{\langle t-1 \rangle}$ which itself ...

$$a^{\langle t \rangle} = g_a(W_{ax} x^{\langle t \rangle} + W_{aa} a^{\langle t-1 \rangle} + b_a)$$

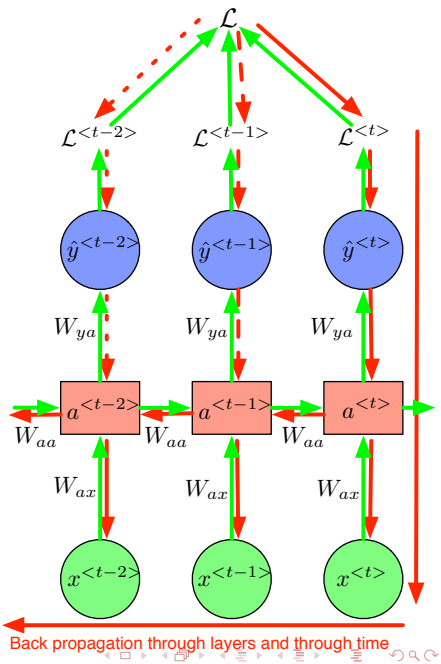
- We use a formula of total derivative

$$\frac{d\mathcal{L}^{\langle t \rangle}}{dW_{aa}}$$

$$= \frac{\partial \mathcal{L}^{\langle t \rangle}}{\partial \hat{y}^{\langle t \rangle}} \frac{\partial \hat{y}^{\langle t \rangle}}{\partial a^{\langle t \rangle}} \left(\frac{\partial a^{\langle t \rangle}}{\partial W_{aa}} + \frac{\partial a^{\langle t \rangle}}{\partial a^{\langle t-1 \rangle}} \frac{\partial a^{\langle t-1 \rangle}}{\partial W_{aa}} + \ldots + \frac{\partial a^{\langle t \rangle}}{\partial a^{\langle t-1 \rangle}} \ldots \frac{\partial a^{\langle 0 \rangle}}{\partial W_{aa}} \right)$$

$$= \frac{\partial \mathcal{L}^{\langle t \rangle}}{\partial \hat{y}^{\langle t \rangle}} \frac{\hat{y}^{\langle t \rangle}}{a^{\langle t \rangle}} \sum_{k=0}^{t} \frac{\partial a^{\langle t \rangle}}{\partial a^{\langle t-1 \rangle}} \cdots \frac{\partial a^{\langle k+1 \rangle}}{\partial a^{\langle k \rangle}} \frac{\partial a^{\langle k \rangle}}{\partial W_{aa}}$$

$$= \frac{\partial \mathcal{L}^{\langle t \rangle}}{\partial \hat{y}^{\langle t \rangle}} \frac{\partial \hat{y}^{\langle t \rangle}}{\partial a^{\langle t \rangle}} \sum_{k=0}^{t} \left(\prod_{i=k+1}^{t} \frac{\partial a^{\langle i \rangle}}{\partial a^{\langle i-1 \rangle}} \right) \frac{\partial a^{\langle k \rangle}}{\partial W_{aa}}$$



Backward pass : 3) compute $\frac{\partial \mathcal{L}}{\partial W_{ax}}$

- We measure how much varying W_{ax} affect $\mathcal{L}^{\langle t \rangle}$.

$$\frac{\partial \mathcal{L}}{\partial W_{ax}} = \sum_{t=1}^{T} \frac{\partial \mathcal{L}^{\langle t \rangle}}{\partial W_{ax}}$$

- It is also necessary to go backwards in time to calculate $\frac{\partial \mathcal{L}}{\partial W_{ax}}$
  - same situation as with W_{aa} :
  - $a^{\langle t \rangle}$ depends on W_{ax} also through $a^{\langle t-1 \rangle}$ which itself ...

$$a^{\langle t \rangle} = g_a(W_{ax}x^{\langle t \rangle} + W_{aa}a^{\langle t-1 \rangle} + b_a)$$

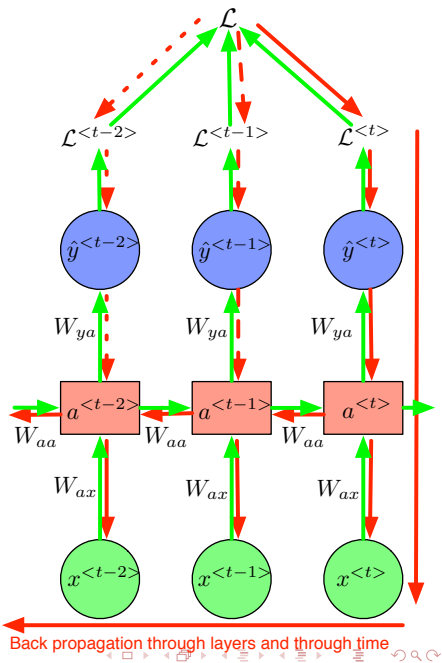
- We use a formula of total derivative

$$\frac{d\mathcal{L}^{\langle t \rangle}}{dW_{ax}}$$

$$= \frac{\partial \mathcal{L}^{\langle t \rangle}}{\partial \hat{y}^{\langle t \rangle}} \frac{\partial \hat{y}^{\langle t \rangle}}{\partial a^{\langle t \rangle}} \left(\frac{\partial a^{\langle t \rangle}}{\partial W_{ax}} + \frac{\partial a^{\langle t \rangle}}{\partial a^{\langle t-1 \rangle}} \frac{\partial a^{\langle t-1 \rangle}}{\partial W_{ax}} + \ldots + \frac{\partial a^{\langle t \rangle}}{\partial a^{\langle t-1 \rangle}} \cdots \frac{\partial a^{\langle 0 \rangle}}{\partial W_{ax}} \right)$$

$$= \frac{\partial \mathcal{L}^{\langle t \rangle}}{\partial \hat{y}^{\langle t \rangle}} \frac{\partial \hat{y}^{\langle t \rangle}}{\partial a^{\langle t \rangle}} \sum_{k=0}^{t} \frac{\partial a^{\langle t \rangle}}{\partial a^{\langle t-1 \rangle}} \cdots \frac{\partial a^{\langle k+1 \rangle}}{\partial a^{\langle k \rangle}} \frac{\partial a^{\langle k \rangle}}{\partial W_{ax}}$$

$$= \frac{\partial \mathcal{L}^{\langle t \rangle}}{\partial \hat{y}^{\langle t \rangle}} \frac{\partial \hat{y}^{\langle t \rangle}}{\partial a^{\langle t \rangle}} \sum_{k=0}^{t} \left(\prod_{i=k+1}^{t} \frac{\partial a^{\langle i \rangle}}{\partial a^{\langle i-1 \rangle}} \right) \frac{\partial a^{\langle k \rangle}}{\partial W_{ax}}$$



Vanishing and exploding gradients

Vanishing gradient

- Long-term dependency
  - The student, which already followed 6 hours of lessons, was tired.
  - The students, which already followed 6 hours of lessons, were tired.
- Very deep NN
  - very difficult to propagate back the gradient to affect the weights of the early Layers

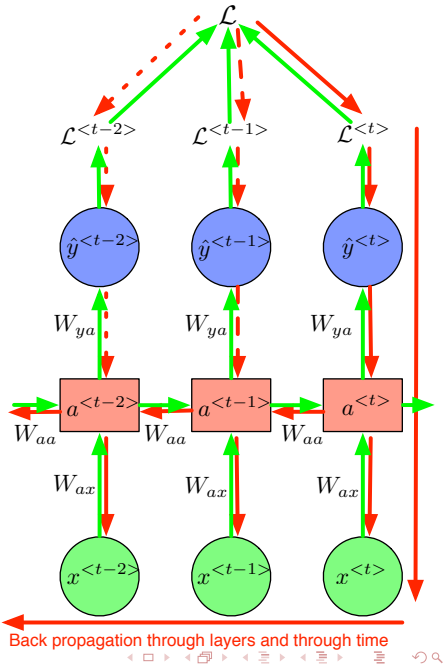
Vanishing and exploding gradients

- To train an RNN we need to backpropagate through layers and through time

$$\frac{\partial \mathcal{L}}{\partial W_{aa}} = \sum_{t=0}^{T} \frac{\partial \mathcal{L}^{\langle t \rangle}}{\partial W_{aa}}$$

$$\frac{\partial \mathcal{L}^{\langle t \rangle}}{\partial W_{aa}} \propto \sum_{k=0}^{t} \left(\prod_{i=k+1}^{t} \frac{\partial a^{\langle i \rangle}}{\partial a^{\langle i-1 \rangle}} \right) \frac{\partial a^{\langle k \rangle}}{\partial W_{aa}}$$

  - each term in the sum is the contribution of
    - a state at time k
    - to the gradient of the loss at time step t
  - the more steps between k and t, the more elements in this product
  - the values of these Jacobian matrices have particularly severe impact on the contributions of fareways steps



Vanishing and exploding gradients

$$\frac{\partial \mathcal{L}}{\partial W_{aa}} = \sum_{t=0}^{T} \frac{\partial \mathcal{L}^{\langle t \rangle}}{\partial W_{aa}}$$

$$\frac{\partial \mathcal{L}^{\langle t \rangle}}{\partial W_{aa}} \propto \sum_{k=0}^{t} \left(\prod_{i=k+1}^{t} \frac{\partial a^{\langle i \rangle}}{\partial a^{\langle i-1 \rangle}} \right) \frac{\partial a^{\langle k \rangle}}{\partial W_{aa}}$$

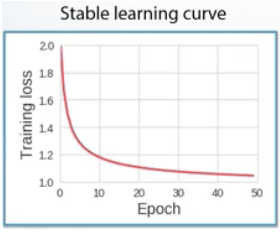
- Suppose only one hidden units, then a_i is a scalar and consequently $\frac{\partial a^{\langle i \rangle}}{\partial a^{\langle i-1 \rangle}}$ is also a scalar
  - If $\left| \frac{\partial a^{\langle i \rangle}}{\partial a^{\langle i-1 \rangle}} \right| < 1$ then the product goes to 0 exponentially fast
  - If $\left| \frac{\partial a^{\langle i \rangle}}{\partial a^{\langle i-1 \rangle}} \right| > 1$ then the product goes to infinity exponentially fast
- Vanishing gradients
  - $\left| \frac{\partial a^{\langle i \rangle}}{\partial a^{\langle i-1 \rangle}} \right| < 1$
    - contributions from faraway steps vanish and don't affect the training
    - difficult to learn long-range dependencies
- Exploding gradients
  - $\left| \frac{\partial a^{\langle i \rangle}}{\partial a^{\langle i-1 \rangle}} \right| > 1$
    - make the learning process unstable
    - gradient could even become NaN

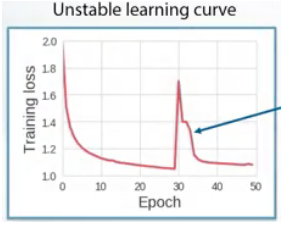
Dealing with the exploding gradient problem

- **Solution 1 : gradient clipping**
  - gradient $d\theta = \frac{\partial \mathcal{L}}{\partial \theta}$, where θ are all the parameters of the network
  - if $\|d\theta\| > threshold$

$$d\theta \leftarrow \frac{threshold}{\|d\theta\|} d\theta$$

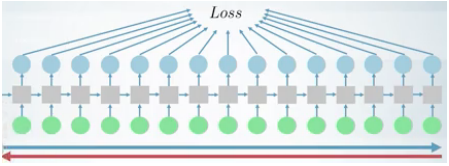
  - clipping doesn't change the direction of the gradient but change its length
  - we can clip only the norm of the part which causes the problem
  - we choose the threshold manually : start with a large threshold and then reduce it

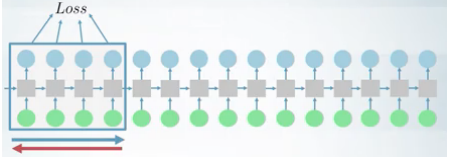


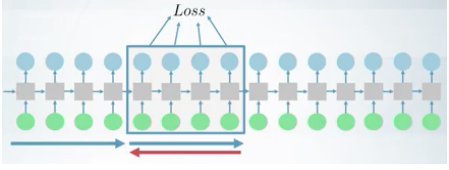


Dealing with the exploding gradient problem

- **Solution 2 : truncated BPTT**
  - Training very long sequences
    - time consuming !
    - exploding gradients !
  - Truncated BPTT :
    - run forward and backward passes through the chunks of the sequence (instead of the whole sequence)
  - Forward
    - We carry hidden states forward in time forever
  - Backward
    - only backpropagate in the chunks (small number of steps)
  - Much faster but dependencies longer than the chunk size don't affect the training (at least they still work at Forward pass)

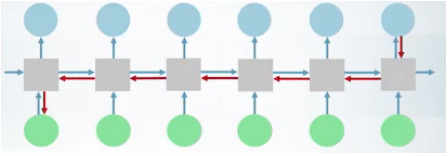


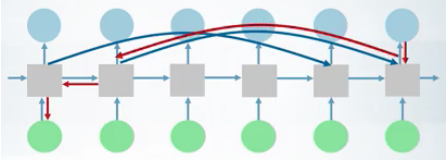




Dealing with the vanishing gradient problem

- **Solution 2 : use skip-connections**
  - in RNN : because at each step we multiply the Jacobian matrices → vanishing gradient
    - we cannot learn long-term dependencies
  - Solution : add short-cuts between hidden states that are separated by more than one time step
    - are usual connections (with their own parameter matrices)
    - create much shorter paths between far away time-steps
  - Backpropagation through the shortcuts : the gradients vanish slower
    - we can learn long-term dependencies
  - not exclusive to RNN (see ResNet)



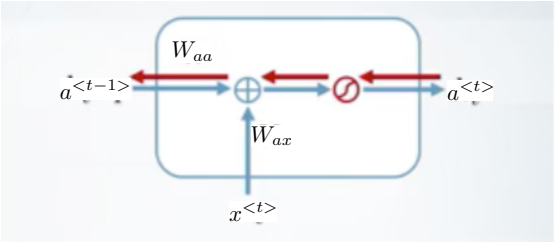


Gated units (LSTM and GRU)

Main ideas : Recurrent Neural Network (RNN)

$$a^{\langle t \rangle} = g\left(W_{ax}x^{\langle t \rangle} + W_{aa}a^{\langle t-1 \rangle} + \underline{b}_a\right)$$

- non-linearities and/or multiplication creates the vanishing gradient problem
- Solution ?
  - create a short-way for backpropagation without any non-linearities or multiplication



Main ideas : Long Short Term Memory Units (LSTM)

[Hochreiter and Schmidhuber 1997. "Long short-term memory", Neural Computation Volume 9 | Issue 8 | November 15, 1997 p.1735-1780]

- Add a **memory cell**
  - the same dimension as hidden layer
- **Simplified LSTM** (no possibility to erase anything) :

Candidate cell value

$$\tilde{c}^{\langle t \rangle} = g\left(W_{ax}^{c} x^{\langle t \rangle} + W_{aa}^{c} a^{\langle t-1 \rangle} + b_g\right)$$

Update gate

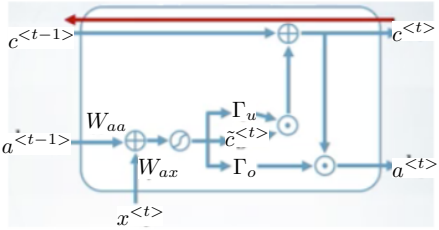
$$\Gamma_u = \sigma\left(W_{ax}^{u} x^{\langle t \rangle} + W_{aa}^{u} a^{\langle t-1 \rangle} + b_u\right)$$

Output gate

$$\Gamma_o = \sigma\left(W_{ax}^{o} x^{\langle t \rangle} + W_{aa}^{o} a^{\langle t-1 \rangle} + b_o\right)$$

$$c^{\langle t \rangle} = c^{\langle t-1 \rangle} + \Gamma_u \odot \tilde{c}^{\langle t \rangle}$$

$$a^{\langle t \rangle} = \Gamma_o \odot f(c^{\langle t \rangle})$$



- no non-linearity or multiplication in the memory cell-path
- $\frac{\partial c^{\langle t \rangle}}{\partial c^{\langle t-1 \rangle}} = diag(1)$
- no vanishing gradients ! ! !

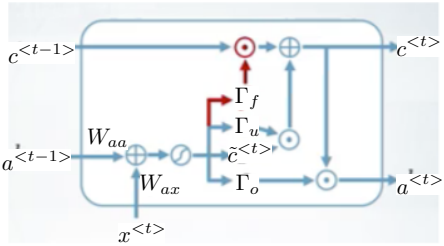
Main ideas : Long Short Term Memory Units (LSTM)

[Hochreiter and Schmidhuber 1997. "Long short-term memory", Neural Computation Volume 9 | Issue 8 | November 15, 1997 p.1735-1780]

- **Forget gate LSTM** :

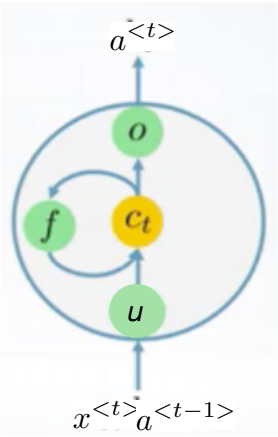
  Forget gate

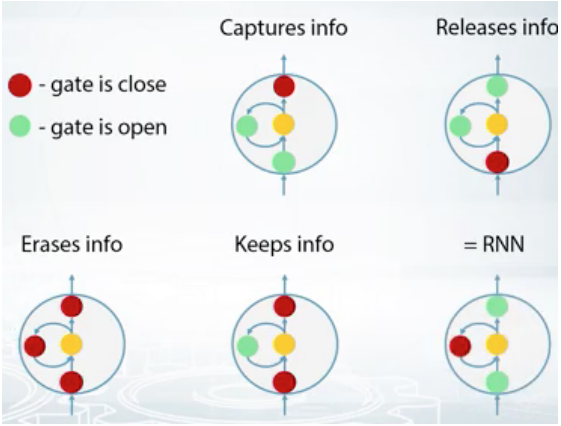
$$\Gamma_f = \sigma\left(W_{ax}^f x^{\langle t\rangle} + W_{aa}^f a^{\langle t-1\rangle} + b_f\right)$$
$$c^{\langle t\rangle} = \Gamma_f \odot c^{\langle t-1\rangle} + \Gamma_u \odot \tilde{c}^{\langle t\rangle}$$
$$a^{\langle t\rangle} = \Gamma_o \odot f(c^{\langle t\rangle})$$



- We now have a multiplication on the short-way
  - $\frac{\partial c^{\langle t\rangle}}{\partial c^{\langle t-1\rangle}} = \text{diag}(\Gamma_f)$
  - with Γ_f the forget gate : $\Gamma_f = \sigma(W_{ax}^f x^{\langle t\rangle} + W_{aa}^f a^{\langle t-1\rangle} + b_f)$
  - Γ_f takes value between 0 and 1
    - $\Rightarrow$ set a high initial value for b_f
    - at the beginning of the training, the LSTM doesn't forget and can find long-term dependencies in the data
- LSTM are very nice but has four times more parameters than RNN

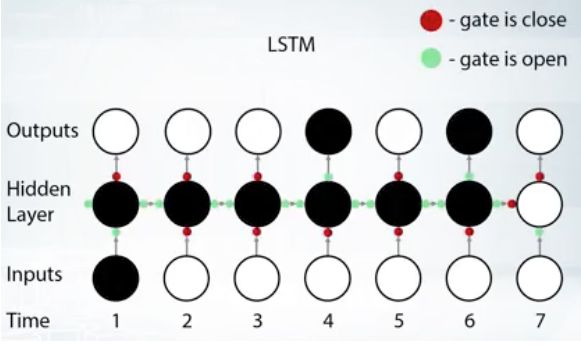
LSTM regimes





Main ideas : Long Short Term Memory Units (LSTM)

- LSTM regimes

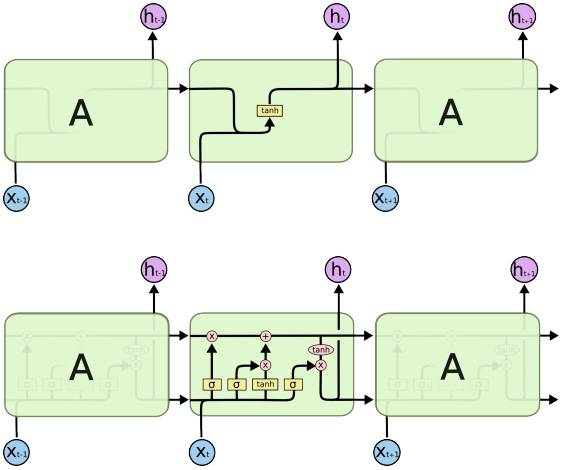


LSTM Example usage

Based on http ://colah.github.io/posts/2015-08-Understanding-LSTMs/

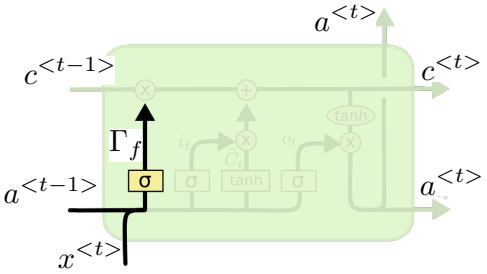
Instead of having a single neural network layer, there are four, interacting in a very special way.

- Lines carry a vector
- A recurrent net transmits its hidden states
- LSTM introduces a second channel : the cell state
- It acts as a memory
- Gates control the memory



LSTM Example usage

- **What should be forgotten from the previous cell state ?**



$$\Gamma_f = \sigma(W_f[a^{<t-1>}, x^{<t>}] + b_f)$$

Action

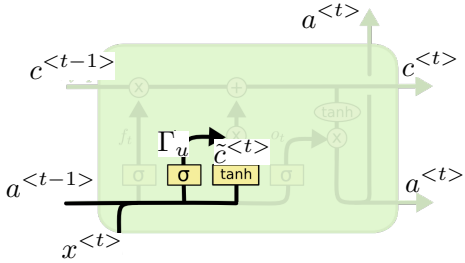
- The sigmoid (forget gate) answers for each component :
  - 1 : to keep it,
  - 0 to forget it, or
  - a value in-between to mitigate its influence

Intuition for language modeling

- The cell state might embed the gender of the present subject :
  - keep it to predict the correct pronouns,
  - or forget it, when a new subject appears

LSTM Example usage

- **What should be taken into account ?**



$$\tilde{c}^{<t>} = \tanh(W_c[a^{<t-1>}, x^{<t>}] + b_c)$$
$$\Gamma_u = \sigma(W_u[a^{<t-1>}, x^{<t>}] + b_u)$$

Action

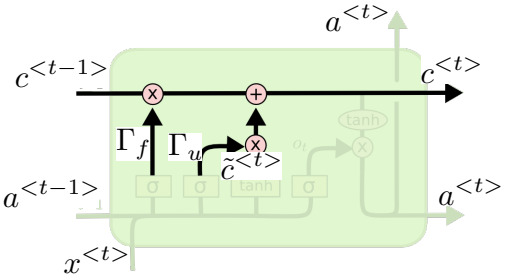
- Create the update $\tilde{c}^{\langle t \rangle}$ of the cell state
- and its contribution Γ_u (the update gate with a sigmoid activation)

Intuition for language modeling

- Add the gender of the new subject to the cell state,
- to replace the old one we're forgetting. appears

LSTM Example usage

- **Write the new state**



$$c^{<t>} = \Gamma_u \odot \tilde{c}^{<t>} + \Gamma_f \odot c^{<t-1>}$$

Action

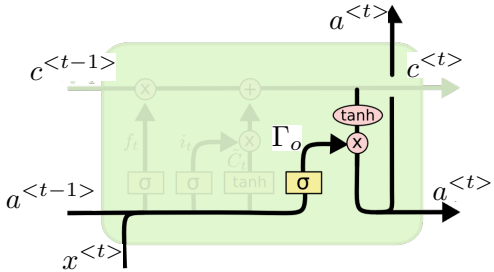
- Merge the old cell state modified by the forget gate
- with the new input

Intuition for language modeling

- Decide to drop the information about the old subject
- Refresh the memory appears

LSTM Example usage

- **Write the new hidden state**



$$\Gamma_o = \sigma(W_o[a^{<t-1>}, x^{<t>}] + b_o)$$
$$a^{<t>} = \Gamma_o \odot \tanh(c^{<t>})$$

Action

- Decide what parts of the (filtered) cell state to output $a^{\langle t \rangle}$
- Compute the hidden state

Intuition for language modeling

- Since we just saw a subject,
- output the relevant information for the future (gender, number) appears

Gated Recurrent Unit (GRU)

[Cho et al. 2014. "On the Properties of Neural Machine Translation: Encoder–Decoder Approaches"]

[Chung et al. 2014 "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling"]

- **Simple GRU**

  Candidate cell value

$$\tilde{c}^{<t>} = \tanh(W_c[c^{<t-1>}, x^{<t>}] + b_c)$$

  Update gate

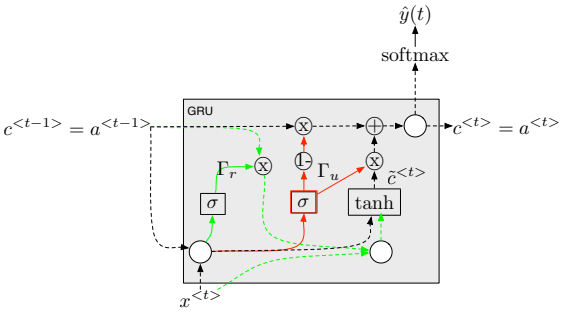
$$\Gamma_u = \sigma(W_c[c^{<t-1>}, x^{<t>}] + b_c)$$

  Cell update

$$c^{<t>} = \Gamma_u \odot \tilde{c}^{<t>} + (1 - \Gamma_u) \odot c^{<t-1>}$$

  Output

$$c^{<t>} = a^{<t>}$$



Gated Recurrent Unit (GRU)

[Cho et al. 2014. "On the Properties of Neural Machine Translation: Encoder–Decoder Approaches"]

[Chung et al. 2014 "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling"]

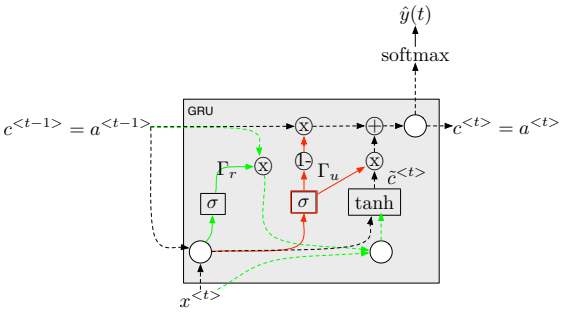
- **Full GRU**
  - adding a relevance gate Γ_r
  - instead of using all $c^{<t-1>}$ to compute $\tilde{c}^{<t>}$, we will first filter it using a "Relevant gate" Γ_r.
  - The "Relevant gate" Γ_r is computed using $c^{<t-1>}$ and $x^{<t>}$.

Relevant gate

$$\Gamma_r = \sigma\left(W_r[c^{<t-1>}, x^{<t>}] + b_c\right)$$

Cell update

$$\tilde{c}^{<t>} = \tanh(W_c[\Gamma_r \odot c^{<t-1>}, x^{<t>}] + b_c)$$



Natural Language Processing

Language model

- **Applications**
  - Automatic Speech Recognition, Machine Translation, OCR, ...
- **Automatic Speech Recognition** :
  - "This is an example of an homophone." or "This is an egg-sample of an homophone."
    - $p(\text{"}example\text{"}) = 5.7 \cdot 10^{-10}$
    - $p(\text{"}egg - sample\text{"}) = 3.2 \cdot 10^{-13}$
- **Language model :** $p(sentence)$
  - $p(y^{\langle 1 \rangle}, y^{\langle 2 \rangle}, \ldots, y^{\langle T_y \rangle})$

Language model

- **Notation :**

  $\mathbf{w} = w_1^L : w_1, w_2, \ldots, w_L$

- **Goal**
  - **Estimate the (non-zero) probability of a word sequence for a given vocabulary**

$$P(w_1^L) = P(w_1, w_2, \ldots, w_L) = \prod_{i=1}^{L} P(w_i | w_1^{i-1}) \qquad \forall i, w_i \in V$$
$$= P(w_1)P(w_2|w1)P(w_3|w_1, w_2)P(w_4|w_1, w_2, w_3)\ldots P(w_L|w_1 \ldots w_{L-1})$$

  - with the **N-gram assumption** :

$$P(w_1^L) = \prod_{i=1}^{L} P(w_i | w_{i-n+1}^{i-1}), \forall i, w_i \in V$$
$$= P(w_1)P(w_2|w1)P(w_3|w_2)P(w_4|w_2)\ldots P(w_L|w_{L-1})$$

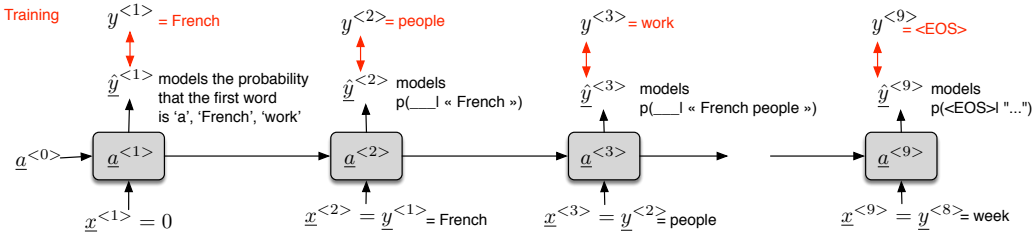
  - in the **recurrent way** :

$$P(w_i | w_1^{i-1})$$

- **How to build a language model ?**
  - need a large corpus of English text
  - define a dictionary
  - tokenize the text
    - convert to one-hot-vector
    - \+ <EOS> (end of sentence)
    - \+ <UNK> (unknown word, very uncommon vocalular) : Gif-sur-Yvette

Language model/ ① Training using a RNN

- **Algorithm :**
  - start with empty context, empty input word
  - given the previous word as input $x^{\langle t \rangle}$ and the context (in the hidden cell $a^{\langle t \rangle}$) predict (maximize the probability to observe) the next word $y^{\langle t \rangle}$ as output $\hat{y}^{\langle t \rangle}$
- "French people work an average of 35 hours a week."



- **Training :**
  - output of the network (softmax) is $\hat{y}^{\langle t \rangle}$ (probability of each word)
  - ground-truth is $y^{\langle t \rangle}$ (one-hot-encoding)
  - minimize the loss

$$\mathcal{L}_t(\hat{y}^{\langle t \rangle}, y^{\langle t \rangle}) = -\sum_{c=1}^{K} y_c^{\langle t \rangle} \log(\hat{y}_c^{\langle t \rangle})$$

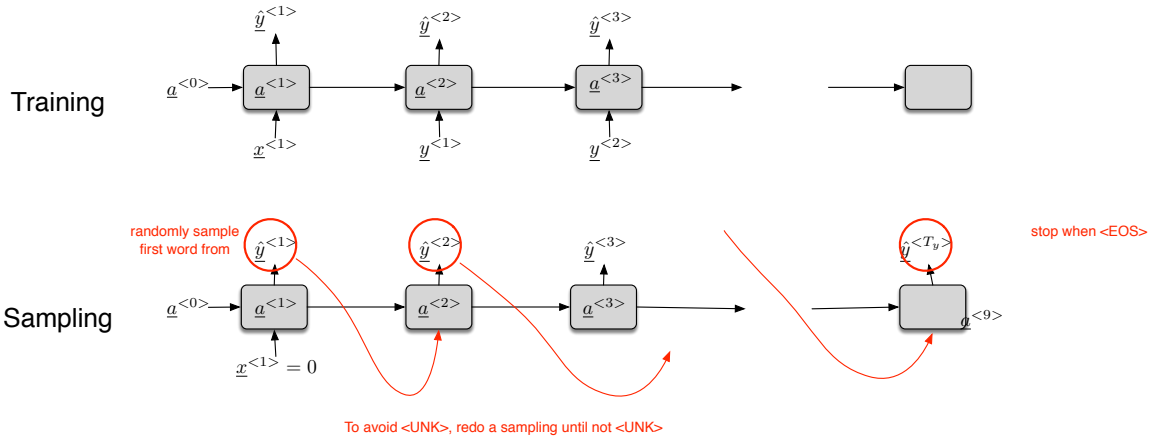
$$\mathcal{L} = \sum_t \mathcal{L}_t(\hat{y}^{\langle t \rangle}, y^{\langle t \rangle})$$

Language model/ ② Testing using a RNN

- Given a new sentence of three words $(y^{(1)}, y^{(2)}, y^{(3)})$
  - what is the probability of this sentence ?

$$
\begin{aligned}
P(y^{(1)}, y^{(2)}, y^{(3)}) = & p(y^{(1)}) && \text{given by the first softmax} \\
& p(y^{(2)}|y^{(1)}) && \text{given by the second softmax} \\
& p(y^{(3)}|y^{(1)}, y^{(2)}) && \text{given by the third softmax}
\end{aligned}
$$

Language model/ ③ Sampling novel sequences using RNN



- Word-level RNN
- Character-level RNN :
  - Vocabulary= ['a', 'b', 'c', . . . , '.', ',', ';', 'A', 'B', 'C', . . .]
- Examples :
  - Shakespeare, wikipedia, source-code generation
  - http://karpathy.github.io/2015/05/21/rnn-effectiveness/

One-hot-encoding

One-hot-encoding

- How to represent the individual words in a sentence ?
- What will be $x^{\langle t \rangle}$?

- Construct the list of the N most used categories (words) in the corpus
  - = **"vocabulary/dictionary"**
  - N is usually in the range 10.000 words - 50.000.
  - add a <UNK> (unknown word) for the words that are not in vocabulary/dictionary
- Each word is now associated with an ID

1	Angela
...	Boris
...	chancelor
367	Emmanuel
...	France
...	Germany
4075	is
...	Johnson
6830	Macron
...	Merkel
...	of
...	president
...	prime-minister
...	the
10.000	UK

One-hot-encoding

- **One-hot-encoding** :
  - the one-hot vector of an ID is a vector filled with 0s, except for a 1 at the position associated with the ID
  - each word w is associated with a one-hot-vector vector o_{ID}
  - If $N = 10.000$, $x^{\langle t \rangle}$ has dimension 10.000
- a one-hot encoding makes no assumption about word similarity
  - all words are equally different from each other

{x}	$x^{<1>}$	$x^{<2>}$	$x^{<3>}$	$x^{<4>}$	$x^{<5>}$	$x^{<6>}$	$x^{<7>}$
	Emmanuel	Macron	is	the	president	of	France
1	0	0	0				
2	0	0	0				
…	…	…	…				
367	1	0	0				
…	…	…	…				
4075	0	0	1				
…	…	…	…				
6830	0	1	0				
…	…	…	…				
10.000	0	0	0				

One-hot-encoding

- **Weaknesses of one-hot-encoding**
  - it is very high-dimensional
    - vulnerability to overfitting, computationally expensive
  - all words are equally different from each other
    - the inner product between any two one-hot vectors is zero
    - as a consequence, we cannot generalize
- Example :
  - we have a trained a language model to complete the sentence
    - The president is on a **boat** __?__" $\Rightarrow$ "trip"
  - since there is no specific relationship between "boat" and "plane", the algorithm cannot complete the sentence
    - The president is on a **plane** __?__" $\Rightarrow$?

Word embedding

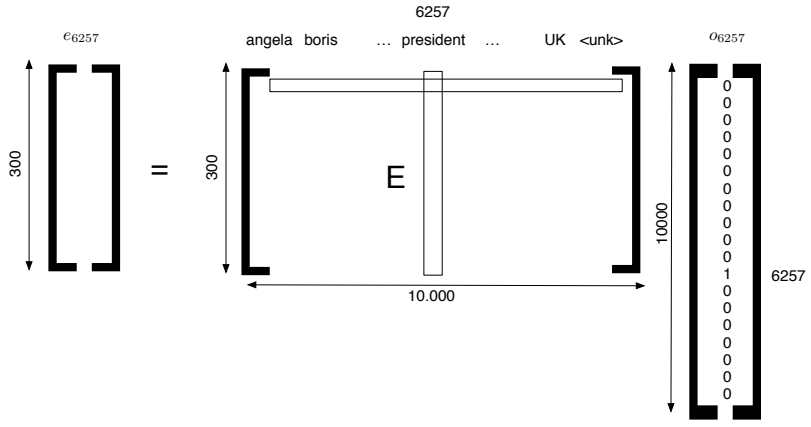
Word embedding

- **Word Embedding** : learn a continuous representation of words
  - each word w is associated with a real-valued vector e_{ID}
    - if embedding size is 300, $e_{\mathbf{5391}}$ is a 300 dimensional vector
  - we would like the distance $||e_{ID1} - e_{ID2}||$ to reflect the meaningfull similarities between words

	Man	Woman	King	Queen	Uncle	Aunt	President	Chancelor	France	Germany	Boat	Plane
	5391	9853	4914	7157	456	6257	7124	3212	6789	1234	5923	8871
Gender	-1	1	-1	1	-1	1	-0,7	-0,3	0	0	0	0
Royal	0	0	1	1	0	0	0,5	0,5	0	0	0	0
Age	0	0	0,7	0,7	0,5	0,5	0,7	0,7	0	0	0	0
Country	0	0	0,5	0,5	0	0	0,5	0,2	1	1	0	0
Transportation	0	0	0	0	0	0	0	0	0	0	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...

- In this representation "boat" and "plane" are quiet similar
  - some of the features may differ but most features are similar
- Therefore the algorithm can find (by generalization) that
  - The president is on a **boat** "__?__" $\Rightarrow$ "trip"
  - The president is on a **plane** "__?__" $\Rightarrow$ "trip"

Word embedding/ Matrix



- $\underset{(10.000,1)}{o_{6257}}$= **one-hot vector** (zero everywhere except 1 at position 6257)
  - dimension 10.000
- $\underset{(300)}{e_{6257}} = \underset{(300,10.000)}{\underline{\underline{E}}} \underset{(10.000,1)}{o_{6257}}$ = **embedding** of o_{6257}
- $e_j = \underline{\underline{E}} \cdot o_j$ = embedding for word j
- In practice, use **look-up table** (take the column vector of $\underline{\underline{E}}$ corresponding to 6257)

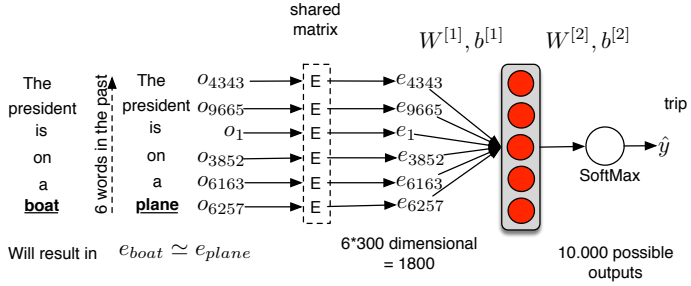
Word embedding/ Learning using

- Various existing methods to learn word embedding :
  - Classic neural language model (Bengio)
  - Collobert and Weston model
  - Word2Vec (Mikolov)
    - Continuous bag-of-words (CBOW)
    - Skip-gram
  - Glove model (Pennington)

Word embedding/ Classic neural language model

[Bengio et al., 2003 "A Neural probabilistic model"]

- **Method** : build a language model (learn to predict the following word) using a MLP
  - Use backpropagation to learn the parameters : $E, W^{[1]}, b^{[1]}, W^{[2]}, b^{[2]}$
    - E is the **embedding matrix**



- **Example :**
  - "The president is on a boat $trip_{target}$ to the United States."
- **Other possibile definition of the context**
  - last 4 words
  - 4 words on the left, 4 words on the right : "is on a boat __?__ to the United States"
  - last 1 words : "boat __?__ "
  - nearby 1 word (skip-gram) : "president __?__ "

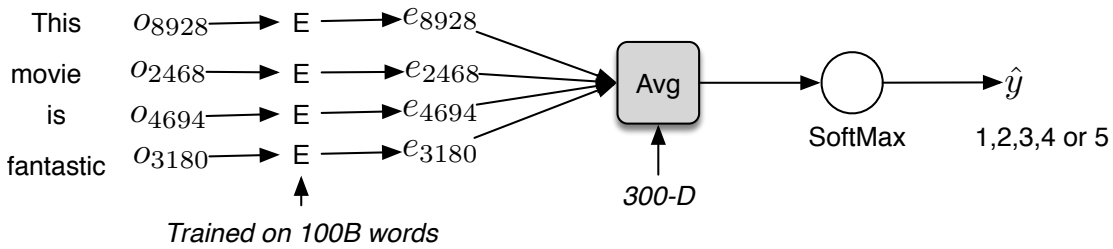
Application : Sentiment Classification

Application : Sentiment Classification

x	y
This movie is fantastic ! I really like it because it is so good !	
Not to my taste, will skip and watch another movie.	
This movie was completely lacking a good script, good actors and a good director.	

Application : Sentiment Classification/ using a simple model

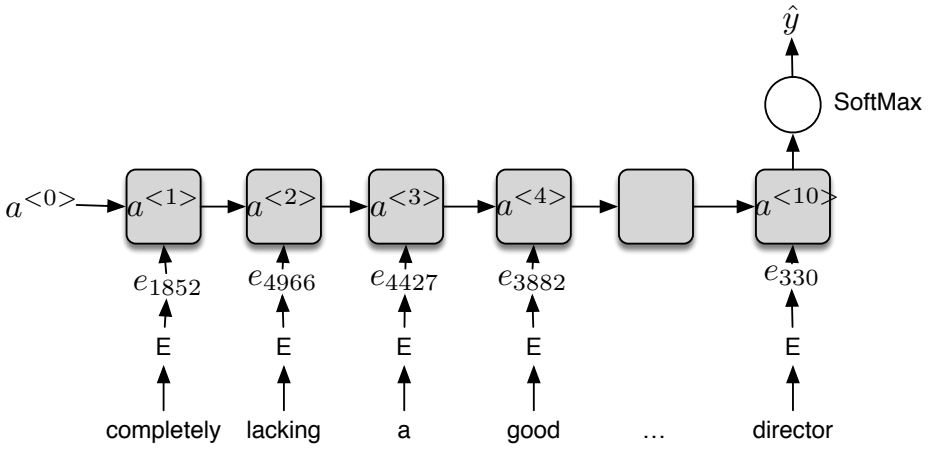
The	movie	is	fantastic
8928	2468	4694	3180



- Does not work for "This movie was completely lacking a **good** script, **good** actors and a **good** director.".
  - Avg= good $\Rightarrow$ does not understand

Application : Sentiment Classification/ using RNN

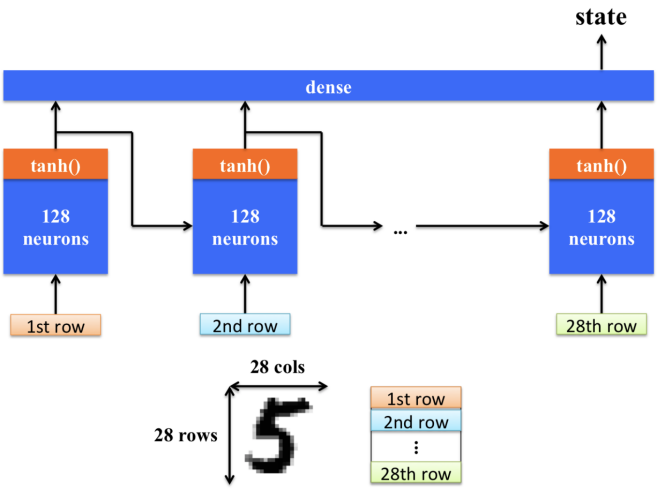
- **Many-to-one** architecture



Application : Handwritten Character Recognition

[Graves et al. 2008 "Unconstrained online handwriting recognition with recurrent neural networks"]
[Graves et al. 2009 "Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks"]

- source
  https://medium.com/machine-learning-algorithms/mnist-using-recurrent-neural-network-2d070a5915a2



RNN data flow

MNIST dataset

Sequence to sequence

Introduction

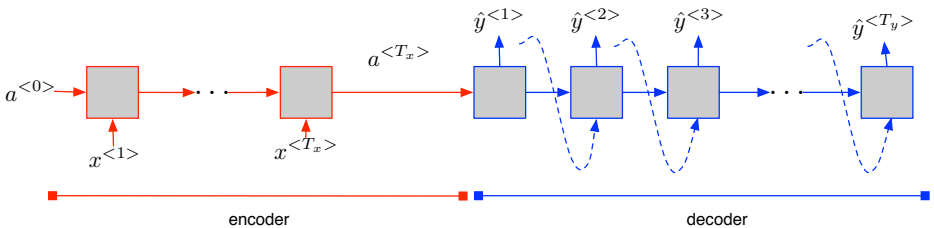
[Sutskever et al. 2014, "Sequence to sequence learning with neural networks"]

[Cho et al. 2014 "Learning phrase representations using rnn encoder-decoder for statistical machine translation"]

- What happens if $T_x \neq T_y$ (many-to-many architecture but with different lengths)
  - Example : Machine Translation

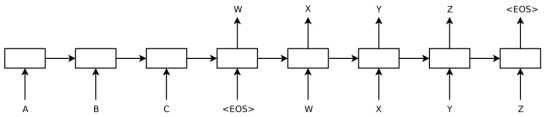
{x}	$x^{<1>}$	$x^{<2>}$	...			$x^{<6>}$		
	Macron	voyage	aux	Etats-Unis	à	Noël		
{y}	$y^{<1>}$	$y^{<2>}$	...					$y^{<8>}$
	Macron	is	travelling	to	the	United-Stat	for	Christmas

- Solution :
  - **Sequence to sequence** [Sutskever] or **Encoder-Decoder** [Cho] architecture

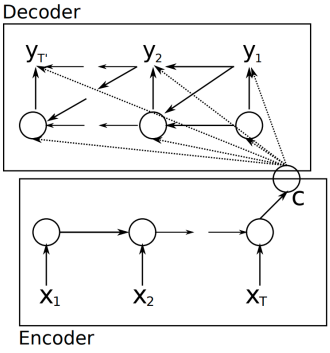


Introduction

[Sutskever et al. 2014, "Sequence to sequence learning with neural networks"]

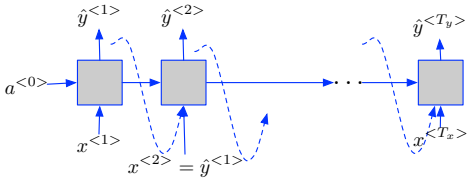


[Cho et al. 2014 "Learning phrase representations using rnn encoder-decoder for statistical machine translation"]

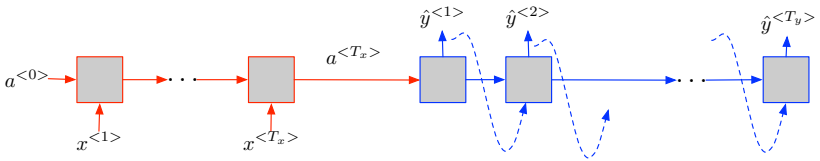


Machine Translation

- **Language model** :
  - $p(y^{\langle 1 \rangle}, \ldots, y^{\langle T_y \rangle})$



- **Machine Translation** :
  - Conditional language model : $p(y^{\langle 1 \rangle}, \ldots, y^{\langle T_y \rangle} | x^{\langle 1 \rangle}, \ldots, x^{\langle T_x \rangle})$
    - French sentence : $x^{\langle 1 \rangle}, \ldots, x^{\langle T_x \rangle}$
    - English sentence : $y^{\langle 1 \rangle}, \ldots, y^{\langle T_y \rangle}$



Attention model

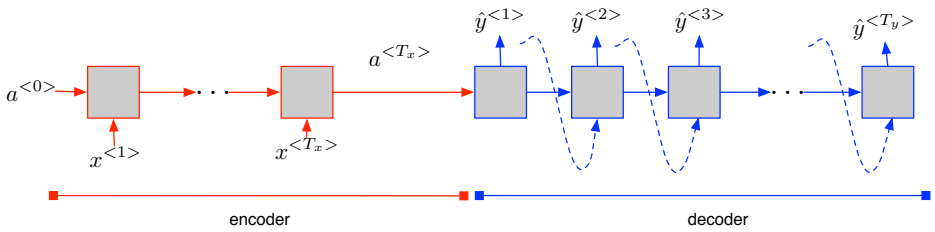
Attention model

[Bahdanau et al. 2015 "Neural machine translation by jointly learning to align and translate"]

- **The problem of long sequences**

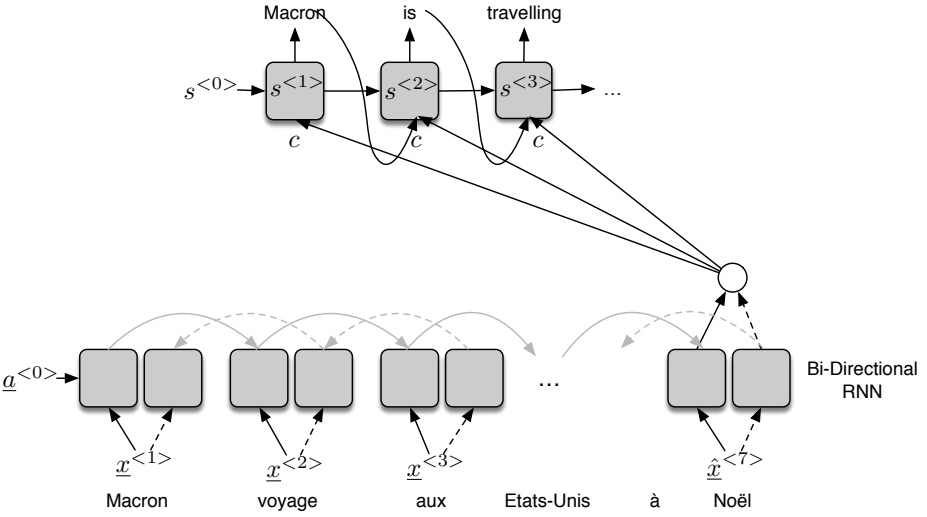
This kind of experience is part of Disney's efforts to "extend the lifetime of its series and build new relationships with audiences via digital platforms that are becoming ever more important," he added.

Ce type d'expérience fait partie des initiatives du Disney pour "prolonger la durée de vie de ses nouvelles et de développer des liens avec les lecteurs numériques qui deviennent plus complexes.



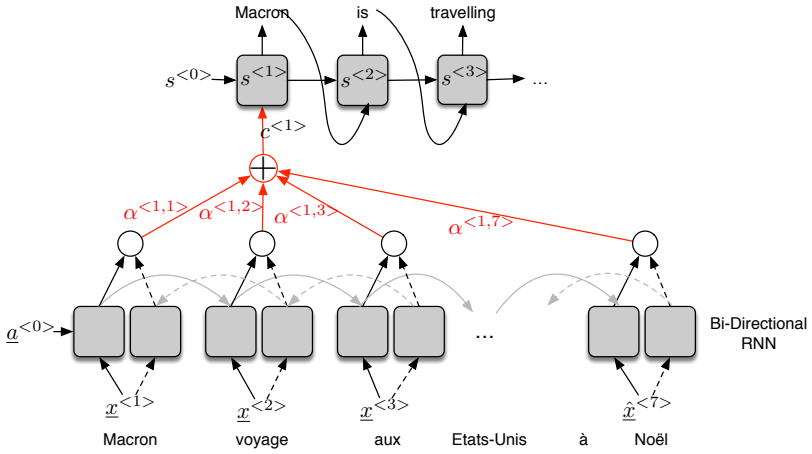
- Encoder/Decoder :
  - $a^{\langle T_x \rangle}$ is supposed to memorize the whole sentence then translate it ;
    - human way : translate each part of a sentence at a time
- **Attention model ?** (modification of the Encoder/Decoder)
  - $a^{\langle T_x \rangle}$ is replaced by a local version in the encoder $\Rightarrow$ we pay attention on specific encoding

Attention model



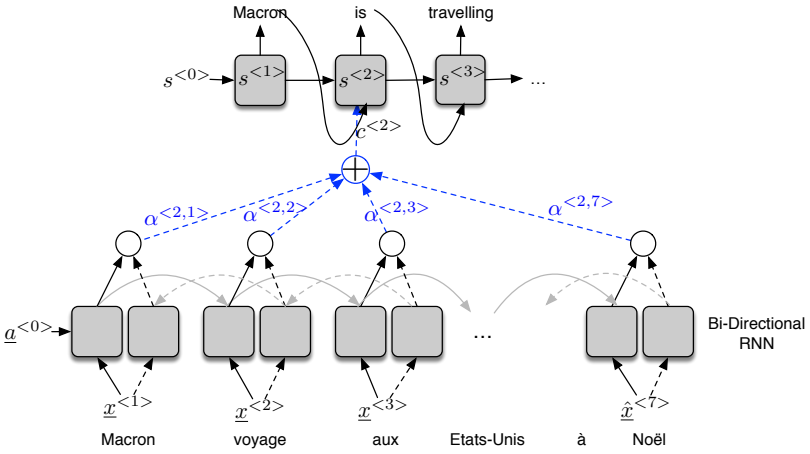
- In usual encoder-decoder,
  - information used for the decoding corresponds to the encoding at T_x : $c = a^{\langle T_x \rangle}$

Attention model



- Attention model
  - replace $c = a^{\langle T_x \rangle}$ by a local version $c^{\langle t \rangle}$
  - $c^{\langle t \rangle}$ is computed as a weighted sum of the encoding hidden states $a^{\langle t \rangle}$
- **Attention weights** $\alpha^{\langle \tau, t \rangle}$:
  - when generating information at time $\tau = 1$, how much, do we have to pay attention on inforrmation at time $t = 1, 2, 3 \ldots 7$
- Notation :
  - $a^{\langle t \rangle}$ / $s^{\langle \tau \rangle}$: hidden states of encoder/ decoder

Attention model



- Attention model
  - replace $c = a^{\langle T_x \rangle}$ by a local version $c^{\langle t \rangle}$
  - $c^{\langle t \rangle}$ is computed as a weighted sum of the encoding hidden states $a^{\langle t \rangle}$
- **Attention weights** $\alpha^{\langle \tau, t \rangle}$:
  - when generating information at time $\tau = 1$, how much, do we have to pay attention on inforrmation at time $t = 1, 2, 3 \ldots 7$
- Notation :
  - $a^{\langle t \rangle}$ / $s^{\langle \tau \rangle}$: hidden states of encoder/ decoder

Attention model

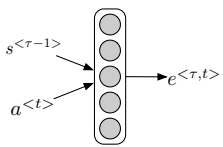
- The **context vector** $c^{\langle \tau \rangle}$ at decoding time τ
  - computed as the sum of the annotations $a^{\langle t \rangle}$ weighted by their **attention weights** $\alpha^{\langle \tau, t \rangle}$

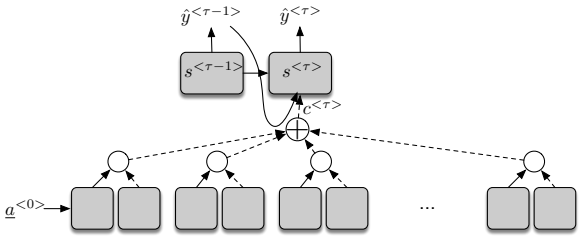
$$c^{\langle \tau \rangle} = \sum_t \alpha^{\langle \tau, t \rangle} a^{\langle t \rangle}$$

- The **attention weight** $\alpha^{\langle \tau, t \rangle}$
  - $\alpha^{\langle \tau, t \rangle}$= among of attention $y^{\langle \tau \rangle}$ should pay to $a^{\langle t \rangle}$
  - computed using a softmax on the **alignment model** $e^{\langle \tau, t \rangle}$

$$\alpha^{\langle \tau, t \rangle} = \frac{\exp(e^{\langle \tau, t \rangle})}{\sum_{t'=1}^{T_x} \exp(e^{\langle \tau, t' \rangle})} \quad \text{with} \quad \sum_t \alpha^{\langle \tau, t \rangle} = 1$$

- The **alignment model** $e^{\langle \tau, t \rangle}$
  - scores how well the inputs around position t match the output at position τ
  - computed using two inputs
    - the RNN hidden state $s^{\langle \tau - 1 \rangle}$ (just before emitting $y^{\langle \tau \rangle}$)
    - the t-th annotation $a^{\langle t \rangle}$ of the input sentence.
  - computed using a feedforward neural network (jointly with all the other components)





Attention model

- Original sentence

This kind of experience is part of Disney's efforts to "extend the lifetime of its series and build new relationships with audiences via digital platforms that are becoming ever more important," he added.

- Encoder/Decoder without attention model

Ce type d'expérience fait partie des initiatives du Disney pour "prolonger la durée de vie de ses nouvelles et de développer des liens avec les lecteurs numériques qui deviennent plus complexes.

- Encoder/Decoder with attention model

Ce genre d'expérience fait partie des efforts de Disney pour "prolonger la durée de vie de ses séries et créer de nouvelles relations avec des publics via des plateformes numériques de plus en plus importantes", a-t-il ajouté.

- Visualization of $\alpha^{\langle \tau, t \rangle}$ (English to French)

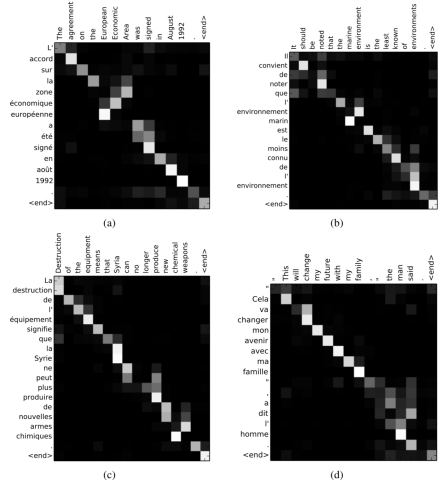
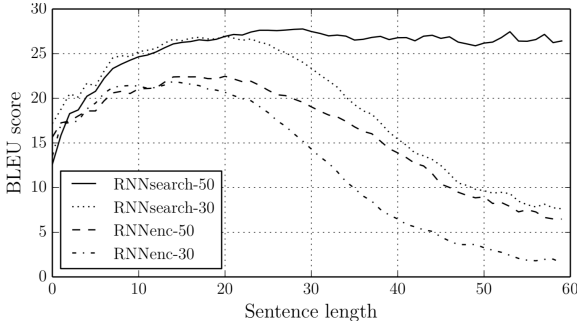


Figure 3: Four sample alignments found by RNNsearch-50. The x-axis and y-axis of each plot correspond to the words in the source sentence (English) and the generated translation (French), respectively. Each pixel shows the weight α_{ij} of the annotation of the j-th source word for the i-th target word (see Eq. (6)), in grayscale (0: black, 1: white). (a) an arbitrary sentence. (b–d) three randomly selected samples among the sentences without any unknown words and of length between 10 and 20 words from the test set.

Attention model

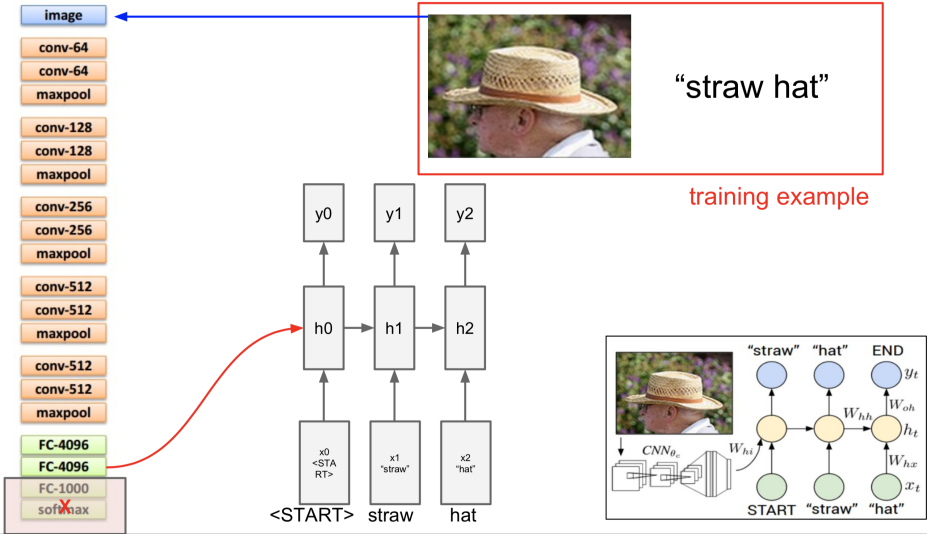
- Bleu score for long sentences



Application : Image captioning

[Mao et al. 2014 "Deep captioning with multimodal recurrent neural networks (m-rnn)"
[Vinyals et al. 2015 "Show and tell: A neural image caption generator"]
[Karpathy et al. 2015 "Deep visual-semantic alignments for generating image descriptions"]

- http://cs231n.stanford.edu



Transformer

Transformer

[Vaswani et al. "Attention Is All You Need", NIPS, 2017]

[J. Alammar "The Illustrated Transformer"]

Transformer

- Global idea : get rid of complex recurrent or convolutional neural networks that include an encoder and a decoder
- a new simple network architecture based solely on attention mechanisms, dispensing with recurrence and convolutions entirely
- still an encoder decoder

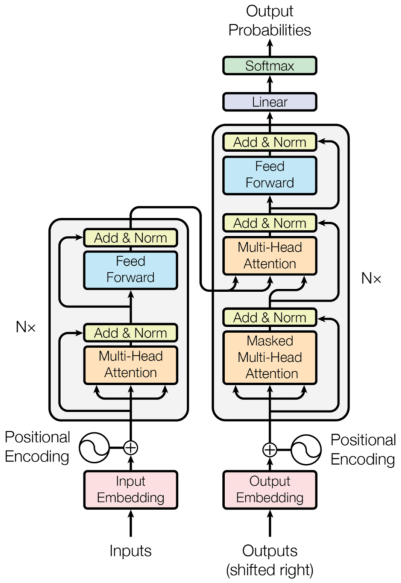


Figure 1: The Transformer - model architecture.

Transformer

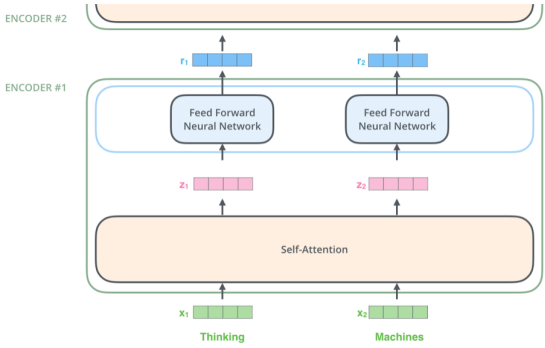
Self-Attention :

- Input sentence we want to translate : "The animal didn't cross the street because it was too tired"
  - What does "it" in this sentence refer to ? Is it referring to the "street" or to the "animal" ?
- When the model is processing the word "it", self-attention allows it to associate "it" with "animal"
- As the model processes each word (each position in the input sequence), self attention allows it to look at other positions in the input sequence for clues that can help lead to a better encoding for this word

Transformer

Now We're Encoding !

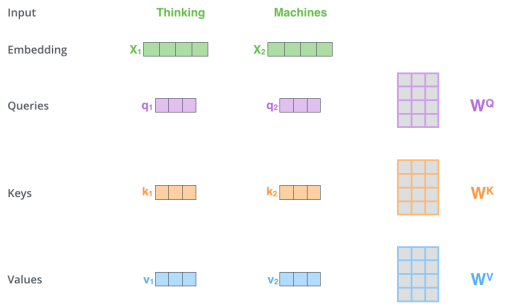
- An encoder receives a list of vectors as input
- It processes this list by passing these vectors into a 'self-attention' layer, then into a feed-forward neural network, then sends out the output upwards to the next encoder



Transformer

First step :

- create three vectors from each of the encoder's input vectors :
  - a Query vector,
  - a Key vector
  - a Value vector
- the vectors are created by multiplying the input by three matrices that we trained during the training process
- the new vectors are smaller in dimension than the embedding vector



Multiplying x1 by the WQ weight matrix produces q1, the "query" vector associated with that word. We end up creating a "query", a "key", and a "value" projection of each word in the input sentence.

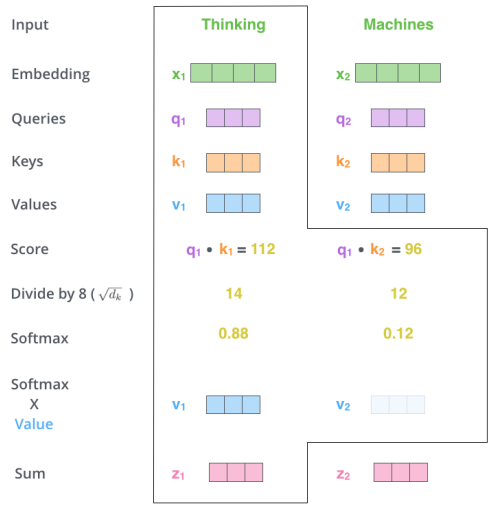
Transformer

Second step : calculate a score

- The score determines how much focus to place on other parts of the input sentence as we encode a word at a certain position.
- Example : calculate self-attention for the first word "Thinking" ⇒ we score each word of the input sentence against this word
- Score ?
  - dot product of the **query** vector with the **key** vector of the respective word we're scoring
  - divide by 8 (the square root of the dimension of the key vectors – 64)
  - pass the result through a softmax operation
    - softmax score determines how much each word will be expressed at this position

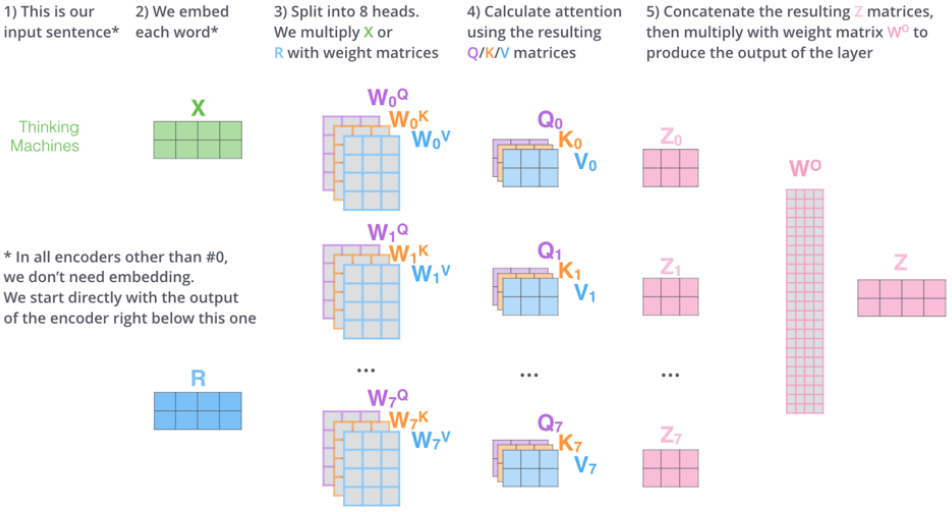
Third step :

- multiply each **value** vector by the **softmax** score
  - keep intact the values of the word(s) we want to focus on, and drown-out irrelevant words
- sum up the weighted value vectors
  - produces the output of the self-attention layer at this position (for the first word).



Transformer

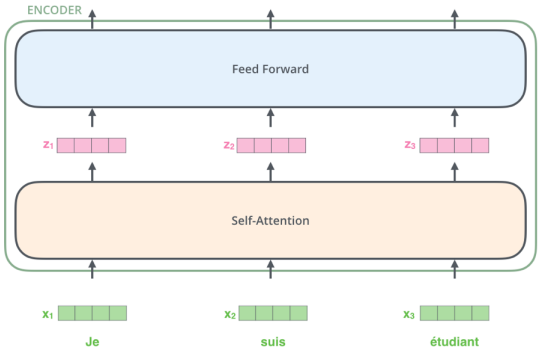
In summary



Transformer

Bringing The Tensors Into The Picture

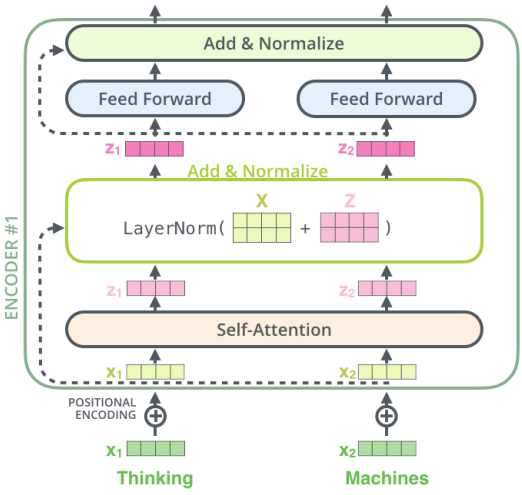
- first turn each input word into a vector using an embedding algorithm (only happens in the bottom-most encoder)
- after embedding the words in our input sequence, each of them flows through each of the two layers of the encoder
- word in each position flows through its own path in the encoder
- There are dependencies between these paths in the self-attention layer. The feed-forward layer does not have those dependencies, however, and thus the various paths can be executed in parallel while flowing through the feed-forward layer



Transformer

Residuals + Normalization

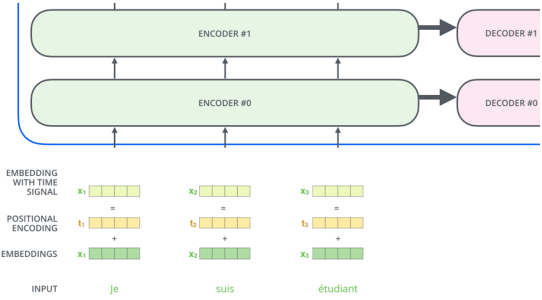
- each sub-layer (self-attention, feed-forward) in each encoder has
  - a residual connection around it
  - is followed by a layer-normalization step



Transformer

Representing The Order of The Sequence Using Positional Encoding

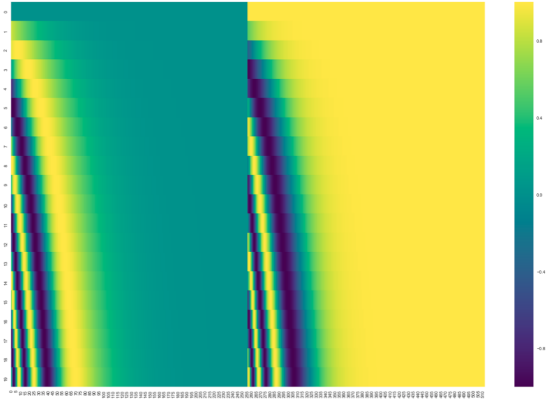
- Find a way to account for the order of the words in the input sequence
  - Transformer adds a vector to each input embedding
  - These vectors follow a specific pattern that the model learns, which helps it determine the position of each word, or the distance between different words in the sequence



To give the model a sense of the order of the words, we add positional encoding vectors -- the values of which follow a specific pattern.

Transformer

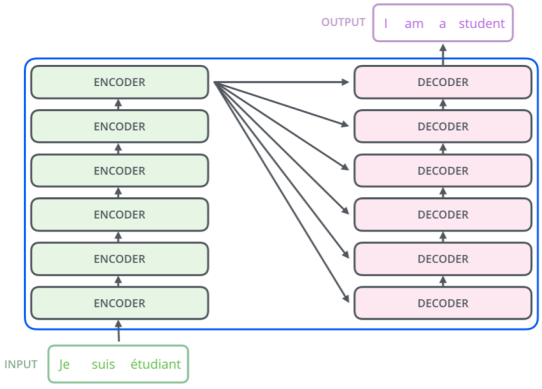
What might this pattern look like ?



A real example of positional encoding for 20 words (rows) with an embedding size of 512 (columns). You can see that it appears split in half down the center. That's because the values of the left half are generated by one function (which uses sine), and the right half is generated by another function (which uses cosine). They're then concatenated to form each of the positional encoding vectors.

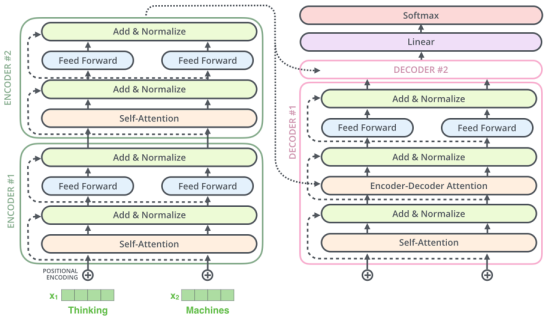
Transformer

- The encoding component is a stack of 6 encoders on top of each other
- The decoding component is a stack of decoders of the same number



Transformer

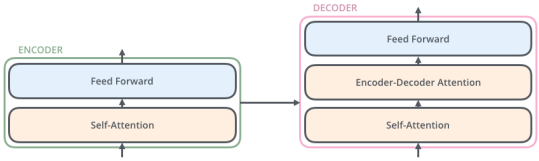
Plugin the encoder to the decoder



Transformer

The decoder has both those layers

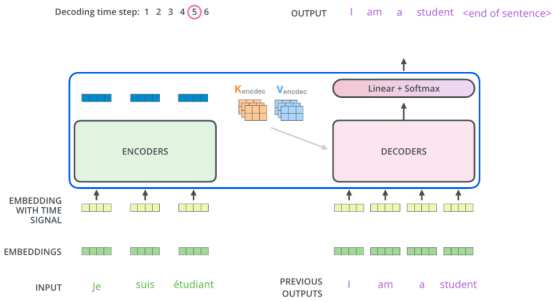
- but between them is an **attention layer** that helps the decoder focus on relevant parts of the input sentence (similar what attention does in seq2seq models)



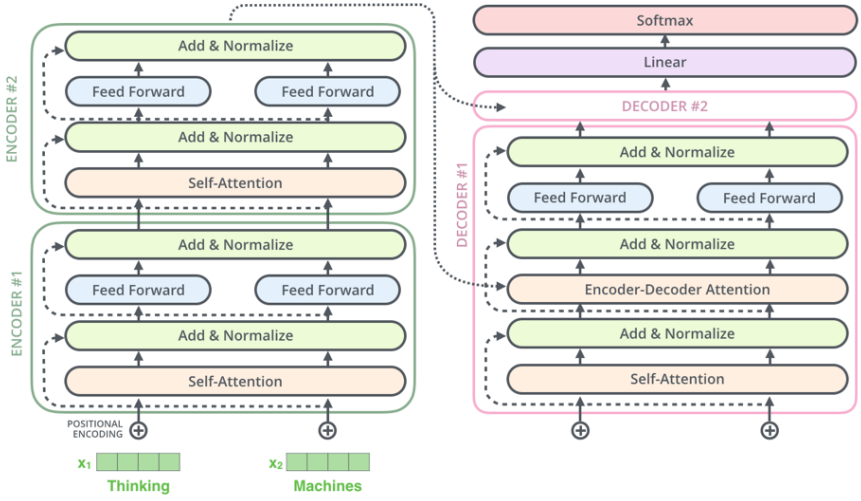
Transformer

Decoder side

- The output of the top encoder is then transformed into a set of attention vectors K and V
  - to be used by each decoder in its "encoder-decoder attention" layer which helps the decoder focus on appropriate places in the input sequence
- The output of the decoder of each step is fed to the bottom in the next step
- As for the encoder, we embed and add positional encoding to those decoder inputs
- In the decoder, the self-attention layer is only allowed to attend to earlier positions in the output sequence
- The "Encoder-Decoder Attention" layer works just like multiheaded self-attention, except it creates its Queries matrix from the layer below it, and takes the Keys and Values matrix from the output of the encoder stack.



Transformer



BERT

BERT

[J. Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", 2019]

[J. Alammar "The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning)"]

- BERT ?
  - a trained Transformer Encoder stack
  - the first input token is supplied with a special [CLS] token. CLS here stands for Classification