

# I SPDL

## Deep Learning for Natural Language Processing

**César  
Bragagnini**

[cesarbrma91@gmail.com](mailto:cesarbrma91@gmail.com)  
[@MarchBragagnini](https://twitter.com/MarchBragagnini)



# Outline

1. Natural Language Processing
2. Word Embeddings
3. Neural networks
  - a. RNN & Vanish/Exploding Gradient
  - b. LSTM, GRU.
  - c. Bidirectional-RNN
  - d. Recursive Neural Networks
  - e. Stack RNN
  - f. Leakey units/residual connections
4. Neural Language Modeling
5. Sentiment Analysis
6. Name Entity Recognition(NER)
7. Opinion Mining
8. Attention/self Attention
9. Machine Translation
  - a. Encoder-Decoder Sutskever
  - b. GNMT
  - c. Conv2Seq
  - d. Transformer
10. Image Caption
11. Question Answering/Machine Comprehension / Chat bot
12. Visual Question Answering
13. Transfer Learning in NLP
14. References

# 1. Natural Language Processing

NLP es la intersección entre ciencia de la computación, inteligencia artificial y lingüística.

## Objetivo

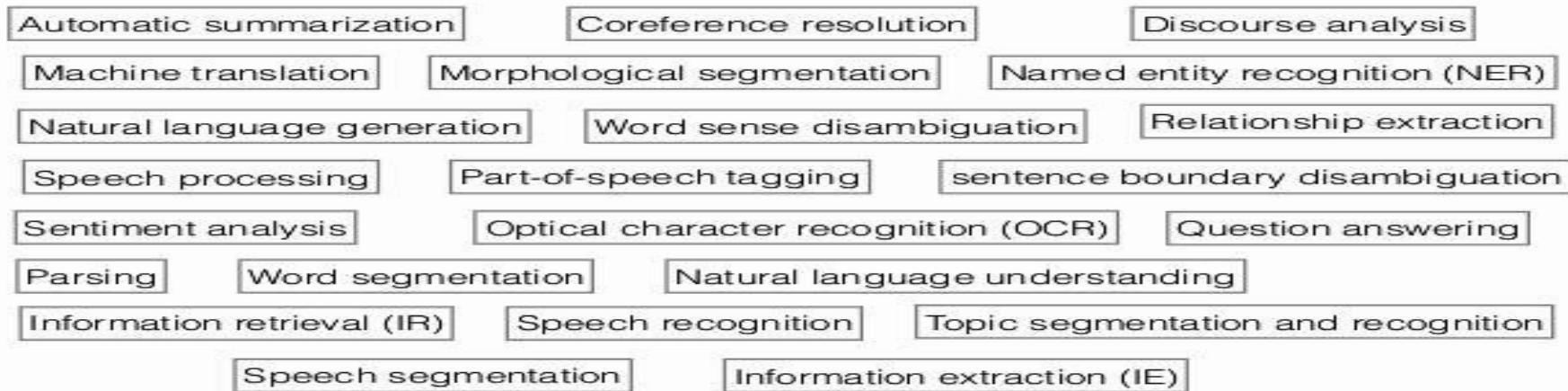
Lograr que las computadoras procesan o “entiendan” el lenguaje Natural para realizar tareas que son útiles, por ejemplo:

- Atender citas, realizar compras, atender clientes(bots)
- Responder preguntas (Question Answering)

**El perfecto entendimiento del lenguaje, significa una IA completa.**

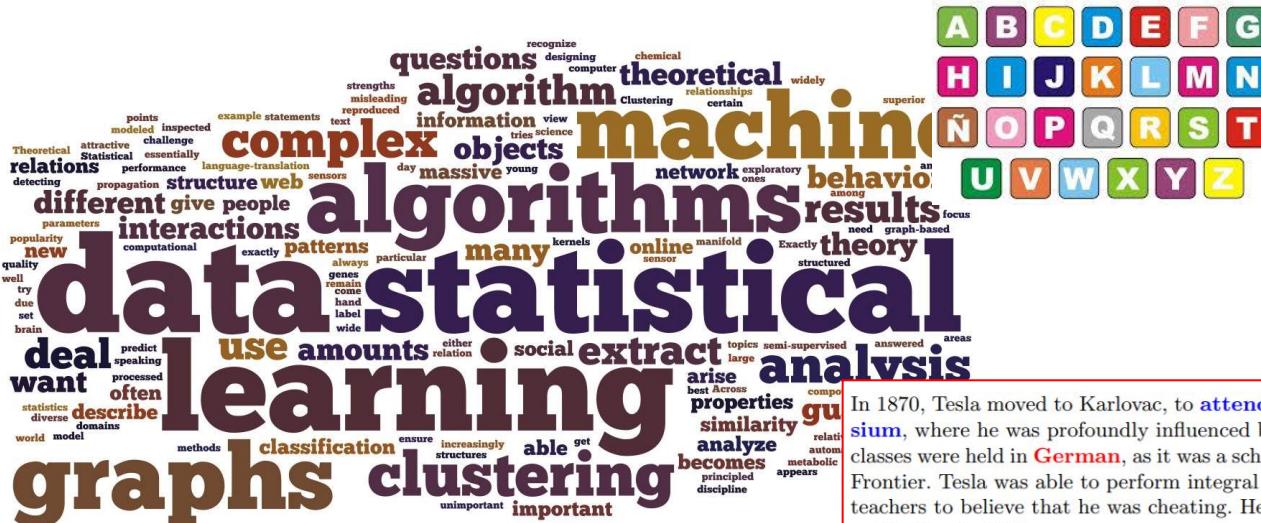
# 1. Natural Language Processing

En general



# 1. Natural Language Processing

# Datos de *Natural Language Processing*



In 1870, Tesla moved to Karlovac, to attend school at the Higher Real Gymnasium, where he was profoundly influenced by a math teacher Martin Sekulic. The classes were held in German, as it was a school within the Austro-Hungarian Military Frontier. Tesla was able to perform integral calculus in his head, which prompted his teachers to believe that he was cheating. He finished a four-year term in three years, graduating in 1873.

## 2. Word Embeddings

Representación de los datos - *One Hot Encoding*

$$w^{aardvark} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, w^a = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, w^{at} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \dots w^{zebra} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$$

Problema? No representa bien la relación entre 2 palabras

$$(w^{hotel})^T w^{motel} = (w^{hotel})^T w^{cat} = 0$$

## 2. Word Embeddings

Embeddings - Base de todo de Natural Language Processing

Es una función que mapea un token, palabra o carácter de cierto lenguaje a un vector de una dimensión alta.

$$W : \text{words} \rightarrow \mathbb{R}^n$$

$$W(\text{"cat"}) = (0.2, -0.4, 0.7, \dots)$$

$$W(\text{"mat"}) = (0.0, 0.6, -0.1, \dots)$$

## 2. Word Embeddings

Base de todo de Natural Language Processing



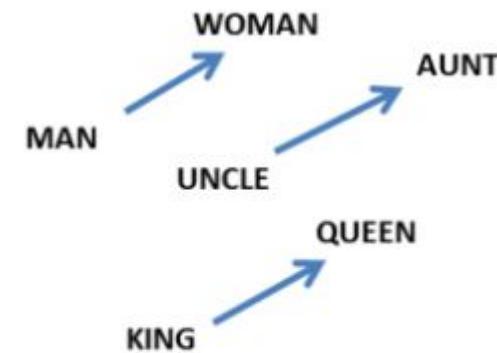
## 2. Word Embeddings

Base de todo de Natural Language Processing

Podemos realizar operaciones aritméticas como:

$$W(\text{"woman"}) - W(\text{"man"}) \simeq W(\text{"aunt"}) - W(\text{"uncle"})$$

$$W(\text{"woman"}) - W(\text{"man"}) \simeq W(\text{"queen"}) - W(\text{"king"})$$



From Mikolov *et al.*  
(2013a)

## 2. Word Embeddings

Base de todo de Natural Language Processing

¿Qué define el significado de una palabra ?

**El significado de una palabra está determinado por su contexto**

1. El **perro** está ladrando y moviendo la colita
2. El **can** está ladrando y moviendo la colita.
3. El **boby** está ladrando y moviendo la colita.
4. El **Chiquito** está ladrando y moviendo la colita.

**¿Qué representa la palabra boby y Chiquito ?**



## 2. Word Embeddings

Base de todo de Natural Language Processing

**El significado de una palabra está determinado por su contexto**

1. El **perro** está ladrando y moviendo la colita
2. El **can** está ladrando y moviendo la colita.
3. El **boby** está ladrando y moviendo la colita.
4. El **Chiquito** está ladrando y moviendo la colita.



**Si un modelador de lenguaje, genera la siguiente frase:**

1. El **boby** ha maullado muy fuerte, y ahora está hablando con Lari.



## 2. Word Embeddings

Base de todo de Natural Language Processing

**El significado de una palabra está determinado por su contexto**

1. El **perro** está ladrando y moviendo la colita
2. El **can** está ladrando y moviendo la colita.
3. El **boby** está ladrando y moviendo la colita.
4. El **Chiquito** está ladrando y moviendo la colita.



**Si un modelador de lenguaje, genera la siguiente frase:**

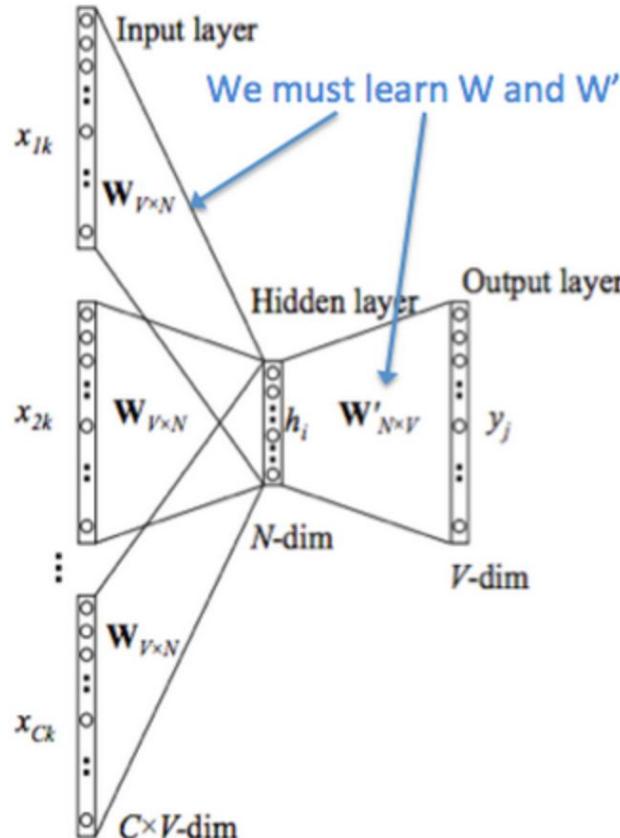
1. El **boby** ha maullado muy fuerte, y ahora está hablando con Lari.



**Tal vez, sea un mal generador de lenguaje**

## 2. Word Embeddings

Word2Vec - CBOW Model

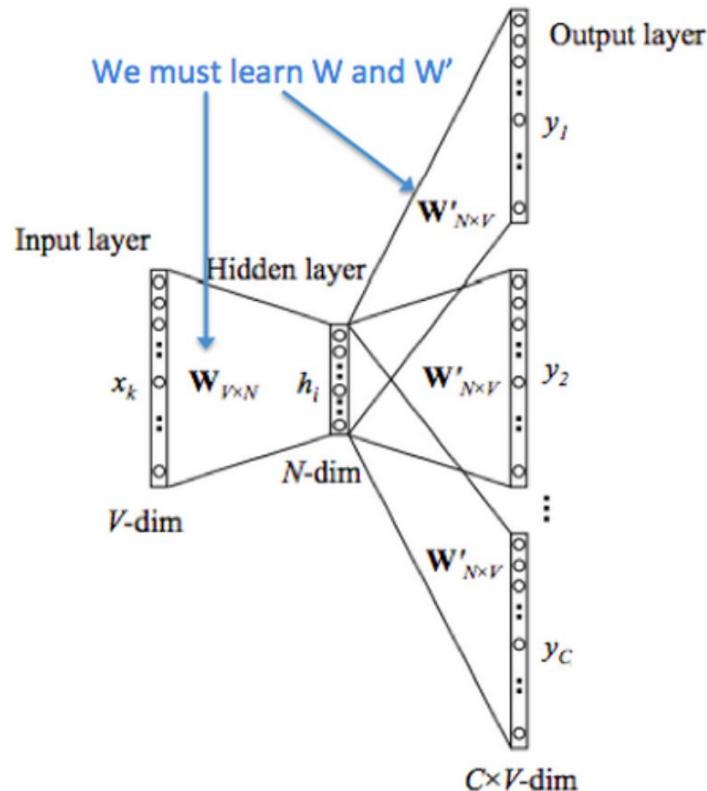


**Notation for CBOW Model:**

- $w_i$ : Word  $i$  from vocabulary  $V$
- $\mathcal{V} \in \mathbb{R}^{n \times |V|}$ : Input word matrix
- $v_i$ :  $i$ -th column of  $\mathcal{V}$ , the input vector representation of word  $w_i$
- $\mathcal{U} \in \mathbb{R}^{n \times |V|}$ : Output word matrix
- $u_i$ :  $i$ -th row of  $\mathcal{U}$ , the output vector representation of word  $w_i$

## 2. Word Embeddings

Word2Vec - Skip Gram Model



### Notation for Skip-Gram Model:

- $w_i$ : Word  $i$  from vocabulary  $V$
- $\mathcal{V} \in \mathbb{R}^{n \times |V|}$ : Input word matrix
- $v_i$ :  $i$ -th column of  $\mathcal{V}$ , the input vector representation of word  $w_i$
- $\mathcal{U} \in \mathbb{R}^{n \times |V|}$ : Output word matrix
- $u_i$ :  $i$ -th row of  $\mathcal{U}$ , the output vector representation of word  $w_i$

## 2. Word Embeddings

Problema con los word embeddings

1. La visita al zoológico no se **cobra**, y se puede observar la **cobra** recién adquirida
2. Visitamos las plantaciones de **COCA**, mientras tomábamos una **COCA** cola
3. Poniendo tu **cara** bonita, difícilmente te salga **cara** la entrada

### 3. Neural Networks

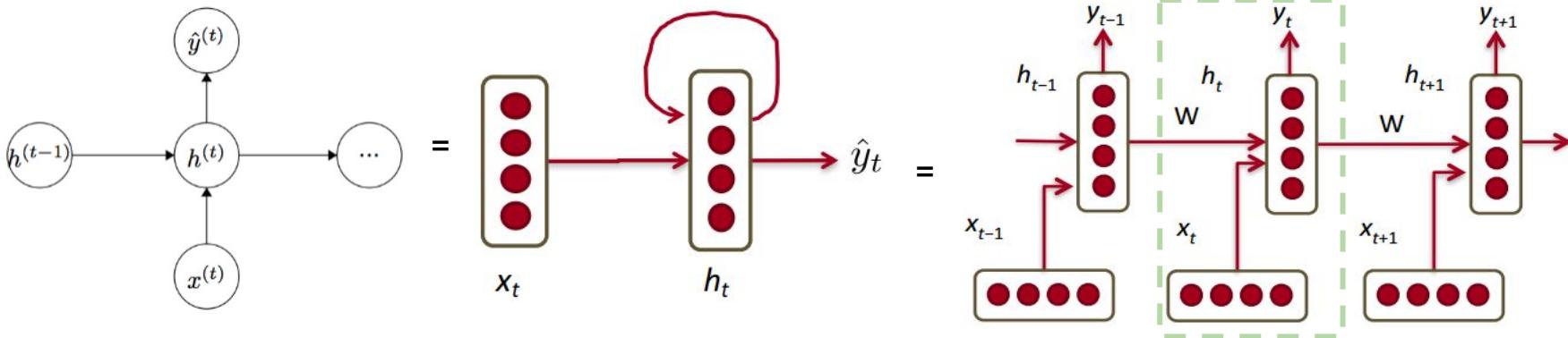
#### 3.1 Introducción

el                    **delicioso**            me  
adobo                parece                 .

1. adobo delicioso me el .
2. el adobo me parece delicioso .
3. . adobo me el delicioso

# 3. Neural Networks

## 3.1 Recurrent Neural Networks & Vanish/Exploding Gradient



Given a list of words:

$x_1, \dots, x_{t-1}, x_t, x_{t+1}, \dots, x_T$

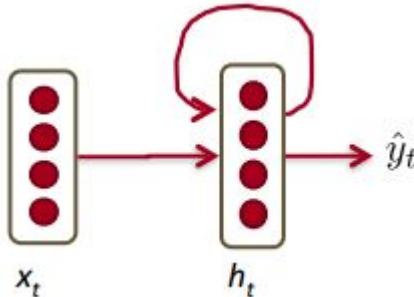
At a single time step:

$$\boxed{h_t = \sigma(W^{hh}h_{t-1} + W^{hx}x_{[t]})}$$
$$\boxed{\hat{y}_t = \text{softmax}(W^{(S)}h_t)}$$

# 3. Neural Networks

## 3.1 Recurrent Neural Networks & Vanish/Exploding Gradient

$$\begin{aligned} h_t &= Wf(h_{t-1}) + W^{(hx)}x_{[t]} \\ \hat{y}_t &= W^{(S)}f(h_t) \end{aligned}$$



$$\frac{\partial E}{\partial W} = \sum_{t=1}^T \frac{\partial E_t}{\partial W}$$

$$\frac{\partial E_t}{\partial W} = \sum_{k=1}^t \frac{\partial E_t}{\partial y_t} \frac{\partial y_t}{\partial h_t} \frac{\partial h_t}{\partial h_k} \frac{\partial h_k}{\partial W}$$

$$\frac{\partial E_t}{\partial W} = \sum_{k=1}^t \frac{\partial E_t}{\partial y_t} \frac{\partial y_t}{\partial h_t} \boxed{\frac{\partial h_t}{\partial h_k}} \frac{\partial h_k}{\partial W}$$

$$\frac{\partial h_t}{\partial h_k} = \boxed{\prod_{j=k+1}^t \frac{\partial h_j}{\partial h_{j-1}}}$$

# 3. Neural Networks

## 3.1 Recurrent Neural Networks & Vanish/Exploding Gradient

Each partial is a Jacobian

$$\frac{d\mathbf{f}}{d\mathbf{x}} = \begin{bmatrix} \frac{\partial \mathbf{f}}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix}$$
$$= \begin{bmatrix} \frac{\partial x_1}{\partial x_1} & \cdots & \frac{\partial x_n}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

Analyzing the norms of the Jacobians, yields:

$$\left\| \frac{\partial h_j}{\partial h_{j-1}} \right\| \leq \|W^T\| \|\text{diag}[f'(h_{j-1})]\| \leq \beta_W \beta_h$$

Where we defined  $\beta$ 's as upper bounds of the norms

The gradient is a product of Jacobian matrices, each associated with a step in the forward computation.

$$\left\| \frac{\partial h_t}{\partial h_k} \right\| = \left\| \prod_{j=k+1}^t \frac{\partial h_j}{\partial h_{j-1}} \right\| \leq (\beta_W \beta_h)^{t-k}$$

This can become very small or very large quickly

# 3. Neural Networks

## 3.1 Recurrent Neural Networks & Vanish/Exploding Gradient

The basic problem is that gradients propagated over many stages tend to either vanish (most of the time) or explode (rarely, but with much damage to the optimization)

$$\mathbf{h}^{(t)} = \mathbf{W}^\top \mathbf{h}^{(t-1)}$$

$$\mathbf{h}^{(t)} = (\mathbf{W}^t)^\top \mathbf{h}^{(0)}, \quad \mathbf{W} = \mathbf{Q} \boldsymbol{\Lambda} \mathbf{Q}^\top$$

$$\mathbf{h}^{(t)} = \mathbf{Q}^\top \boldsymbol{\Lambda}^t \mathbf{Q} \mathbf{h}^{(0)}$$

The eigenvalues are raised to the power of  $t$ , causing eigenvalues with magnitude less than one to decay to zero and eigenvalues with magnitude greater than one to explode.

Any component of  $\mathbf{h}(0)$  that is not aligned with the largest eigenvector will eventually be discarded

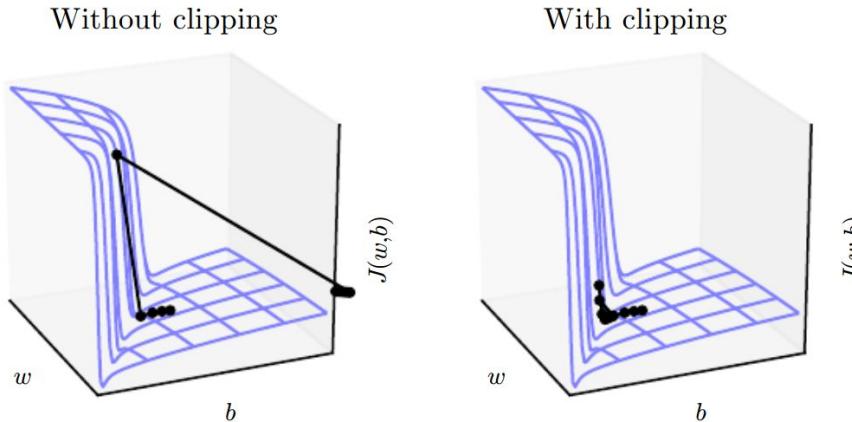
**Problem: Long Term Dependencies, information whose outputs depend on very early entries can not be remembered.**

**Input:** Teo es la mascota de Lari, es un perrito pequeño y bien cariñoso, siempre sale a pasear todos los domingos, le gusta ladrar a la gente y a las perritas. A su mascota no le gustan las chalinas

Can RNN predict the missing word? : La mascota de Lari se llama \_\_\_\_

# 3. Neural Networks

## 3.1 Recurrent Neural Networks & Vanish/Exploding Gradient



Example of the effect of gradient clipping, in a recurrent network with two parameters  $w$  and  $b$ . For solving exploding gradient.

\* **(Left)**Gradient descent without gradient clipping overshoots the bottom of this small ravine, then receives a very large gradient from the cliff face. The large gradient catastrophically propels the parameters outside the axes of the plot.

\* **(Right)**Gradient descent with gradient clipping has a more moderate reaction to the cliff.

# 3. Neural Networks

## 3.1 Recurrent Neural Networks & Vanish/Exploding Gradient

---

**Algorithm 1** Pseudo-code for norm clipping the gradients whenever they explode

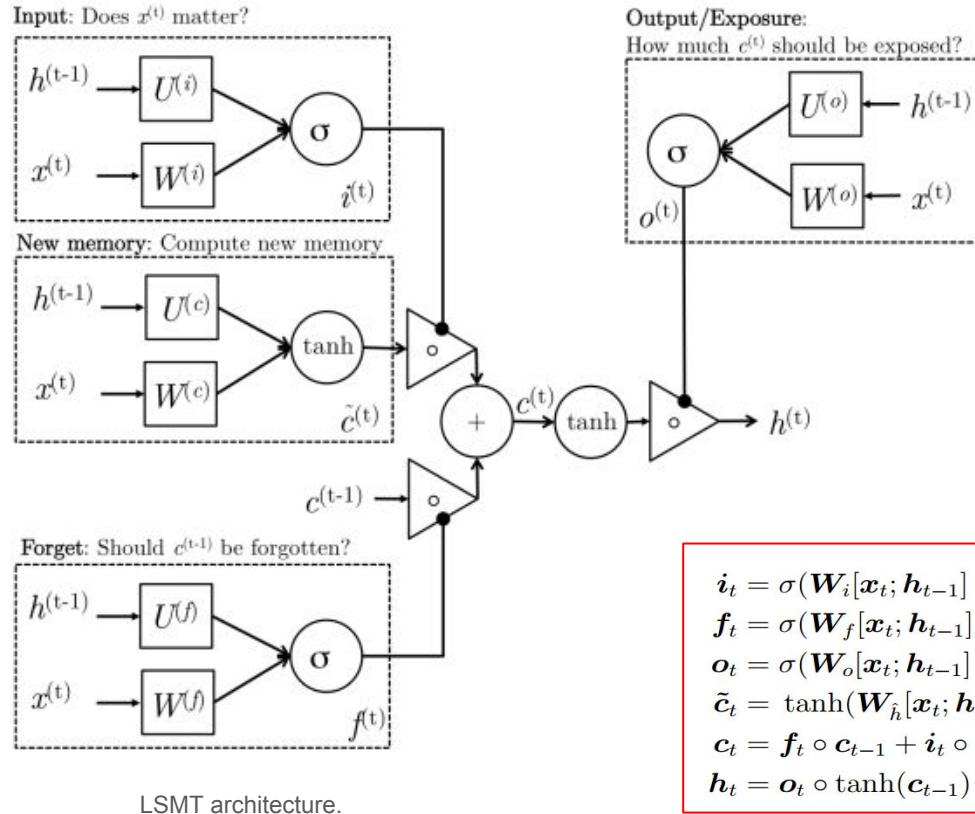
---

```
     $\hat{g} \leftarrow \frac{\partial \mathcal{E}}{\partial \theta}$ 
    if  $\|\hat{g}\| \geq threshold$  then
         $\hat{g} \leftarrow \frac{threshold}{\|\hat{g}\|} \hat{g}$ 
    end if
```

---

# 3. Neural Networks

## 3.2 Long Short Term Memory(LSTM) & ...

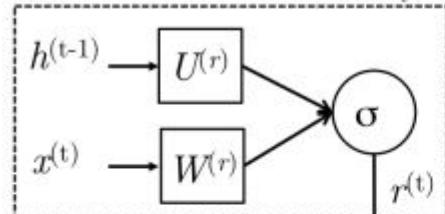


$i_t = \sigma(\mathbf{W}_i[\mathbf{x}_t; \mathbf{h}_{t-1}] + \mathbf{b}_i)$	Input Gate
$f_t = \sigma(\mathbf{W}_f[\mathbf{x}_t; \mathbf{h}_{t-1}] + \mathbf{b}_f)$	Forget Gate
$\mathbf{o}_t = \sigma(\mathbf{W}_o[\mathbf{x}_t; \mathbf{h}_{t-1}] + \mathbf{b}_o)$	Output Gate
$\tilde{\mathbf{c}}_t = \tanh(\mathbf{W}_c[\mathbf{x}_t; \mathbf{h}_{t-1}] + \mathbf{b}_c)$	New Cell Memory
$\mathbf{c}_t = f_t \circ \mathbf{c}_{t-1} + i_t \circ \tilde{\mathbf{c}}_t$	Final Memory
$\mathbf{h}_t = \mathbf{o}_t \circ \tanh(\mathbf{c}_t)$	Hidden State

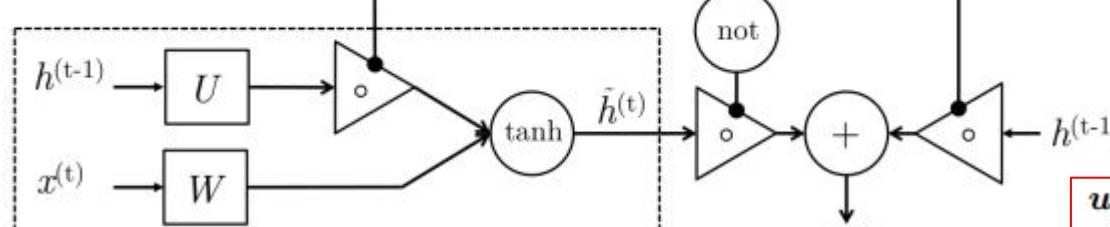
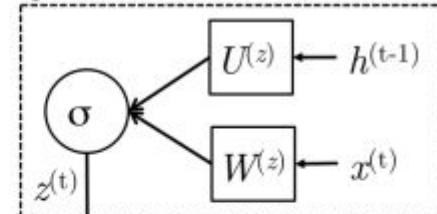
# 3. Neural Networks

## 3.2 Gated Recurrent Unit(GRU)

Reset: Include  $h^{(t-1)}$  in new memory?



Update: How much  $h^{(t-1)}$  in next state?



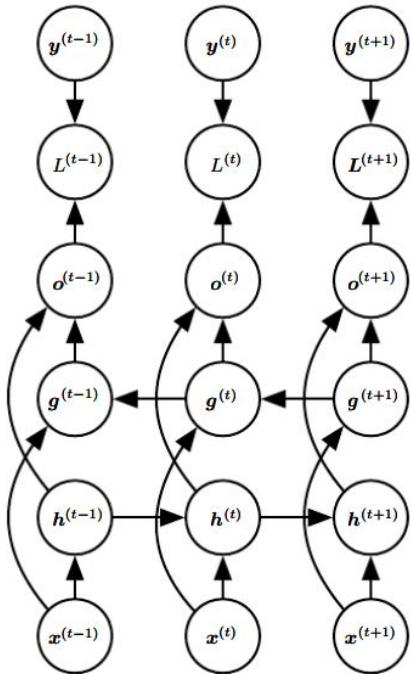
New memory: Compute new memory based on current word input  $x^{(t)}$  and potentially  $h^{(t-1)}$

GRU architecture.

$u_t = \sigma(\mathbf{W}_u[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_u)$	Update Gate
$r_t = \sigma(\mathbf{W}_r[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_r)$	Restart Gate
$\tilde{\mathbf{h}}_t = \tanh(\mathbf{W}_h[r_t \circ \mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_h)$	New Memory
$\mathbf{h}_t = u_t \circ \tilde{\mathbf{h}}_t + (\vec{1} - u_t) \circ \mathbf{h}_{t-1}$	Hidden State

# 3. Neural Networks

## 3.4 Bidirectional - RNN



$$\mathbf{h}^{(t-1)} = f(\mathbf{W}_h[\mathbf{h}^{(t-2)}, \mathbf{x}^{(t-1)}] + \mathbf{b}_h)$$

$$\mathbf{h}^{(t)} = f(\mathbf{W}_h[\mathbf{h}^{(t-1)}, \mathbf{x}^{(t)}] + \mathbf{b}_h)$$

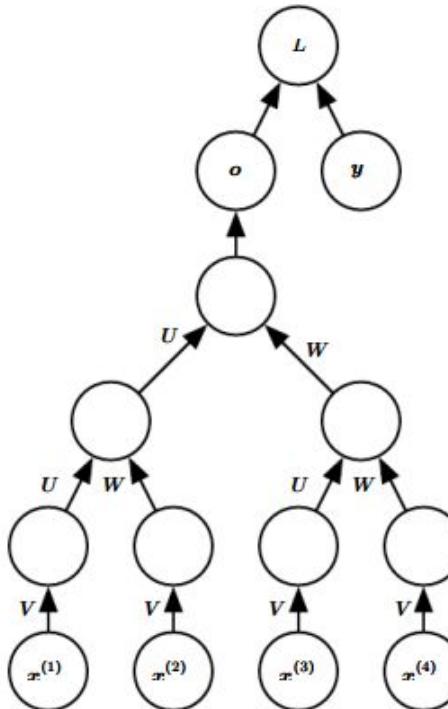
$$\mathbf{g}^{(t+1)} = f(\mathbf{W}_g[\mathbf{g}^{(t+2)}, \mathbf{x}^{(t+1)}] + \mathbf{b}_g)$$

$$\mathbf{g}^{(t)} = f(\mathbf{W}_g[\mathbf{g}^{(t+1)}, \mathbf{x}^{(t)}] + \mathbf{b}_g)$$

$$\mathbf{o}^{(t)} = \hat{\mathbf{y}}_{(t)} = \text{Softmax}(\mathbf{W}_o[\mathbf{h}^{(t)}, \mathbf{g}^{(t)}] + \mathbf{b}_o)$$

# 3. Neural Networks

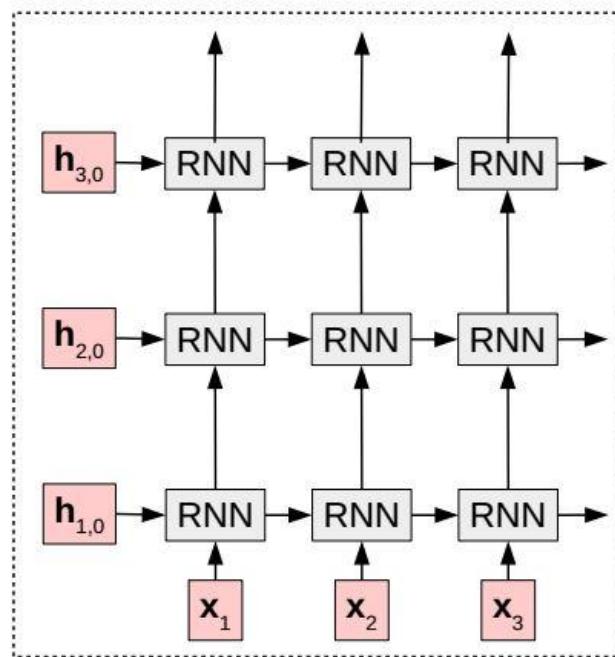
## 3.5 Recursive Neural Networks



A recursive network has a computational graph that generalizes that of the recurrent network from a chain to a tree. A variable-size sequence  $x(1), x(2), \dots, x(t)$  can be mapped to a fixed-size representation (the output  $o$ ), with a fixed set of parameters (the weight matrices  $U, V, W$ ).

# 3. Neural Networks

## 3.6 Stack RNN



$$\mathbf{h}_{3,t} = \text{RNN}_3(\mathbf{h}_{2,t}, \mathbf{h}_{3,t-1}, \mathbf{b}_3)$$

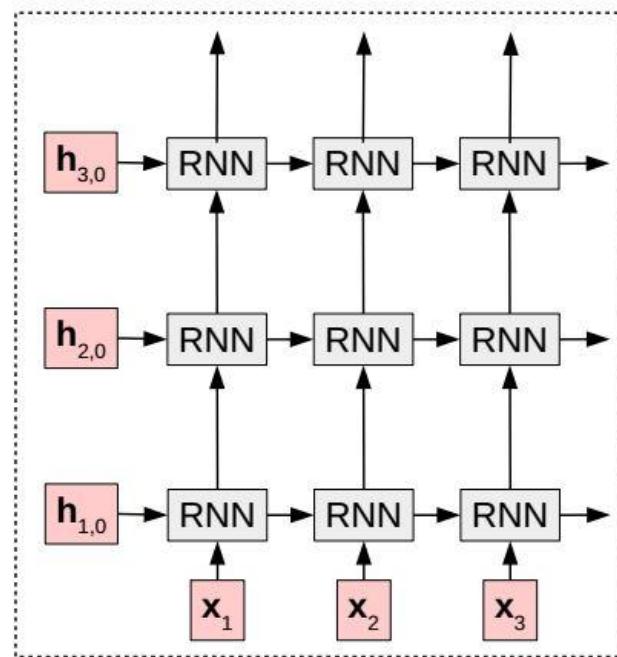
$$\mathbf{h}_{2,t} = \text{RNN}_2(\mathbf{h}_{1,t}, \mathbf{h}_{2,t-1}, \mathbf{b}_2)$$

$$\mathbf{h}_{1,t} = \text{RNN}_1(\mathbf{x}_t, \mathbf{h}_{1,t-1}, \mathbf{b}_1)$$

# 3. Neural Networks

## 3.6 Stack RNN

Trouble?



$$\mathbf{h}_{3,t} = \text{RNN}_3(\mathbf{h}_{2,t}, \mathbf{h}_{3,t-1}, \mathbf{b}_3)$$

$$\mathbf{h}_{2,t} = \text{RNN}_2(\mathbf{h}_{1,t}, \mathbf{h}_{2,t-1}, \mathbf{b}_2)$$

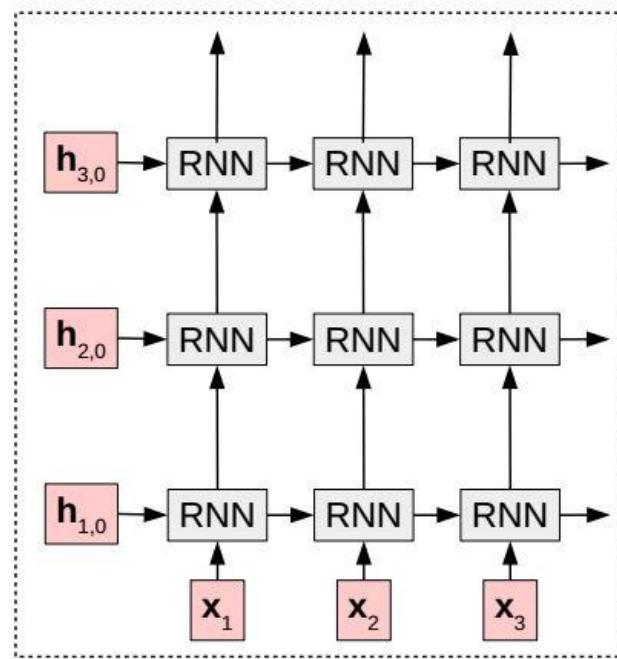
$$\mathbf{h}_{1,t} = \text{RNN}_1(\mathbf{x}_t, \mathbf{h}_{1,t-1}, \mathbf{b}_1)$$

# 3. Neural Networks

## 3.6 Stack RNN

Trouble?

Maybe appear  
gradient  
vanish  
problem in  
vertical



$$\mathbf{h}_{3,t} = \text{RNN}_3(\mathbf{h}_{2,t}, \mathbf{h}_{3,t-1}, \mathbf{b}_3)$$

$$\mathbf{h}_{2,t} = \text{RNN}_2(\mathbf{h}_{1,t}, \mathbf{h}_{2,t-1}, \mathbf{b}_2)$$

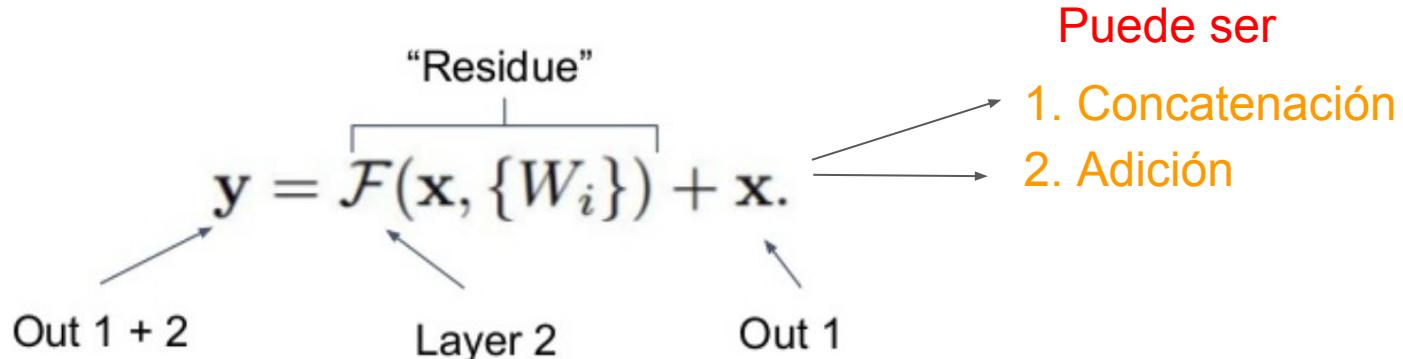
$$\mathbf{h}_{1,t} = \text{RNN}_1(\mathbf{x}_t, \mathbf{h}_{1,t-1}, \mathbf{b}_1)$$

# 3. Neural Networks

## 3.6 Residual Connections/Skip Connections & ...

Gradients may vanish or explode exponentially with respect to the number of time steps.

To introduce recurrent connections with a time-delay of  $d$  to mitigate this problem. Gradients now diminish exponentially as a function of  $T/d$  rather than  $T$ .

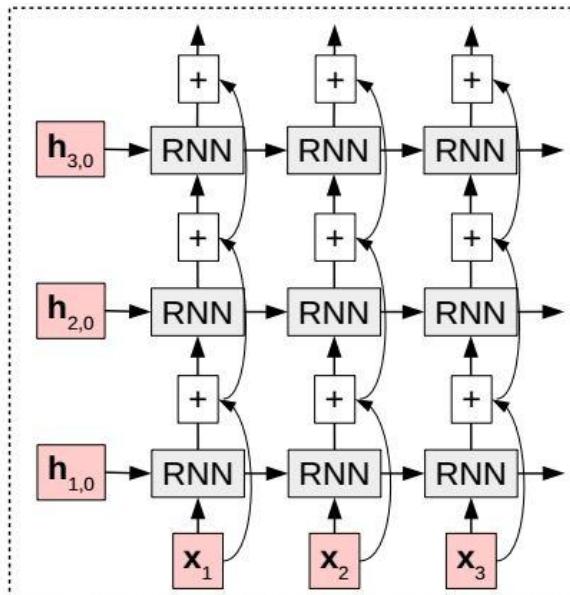


# 3. Neural Networks

## 3.6 Residual Connections/Skip Connections & ...

Gradients may vanish or explode exponentially with respect to the number of time steps.

To introduce recurrent connections with a time-delay of  $d$  to mitigate this problem. Gradients now diminish exponentially as a function of  $T/d$  rather than  $T$ .



$$\mathbf{h}_{3,t} = \text{RNN}(\mathbf{h}_{2,t}, \mathbf{h}_{3,t-1}, \mathbf{b}_3) + \mathbf{h}_{2,t}$$

$$\mathbf{h}_{2,t} = \text{RNN}(\mathbf{h}_{1,t}, \mathbf{h}_{2,t-1}, \mathbf{b}_2) + \mathbf{h}_{1,t}$$

$$\mathbf{h}_{1,t} = \text{RNN}(\mathbf{x}_t, \mathbf{h}_{1,t-1}, \mathbf{b}_1) + \mathbf{x}_t$$

# 3. Neural Networks

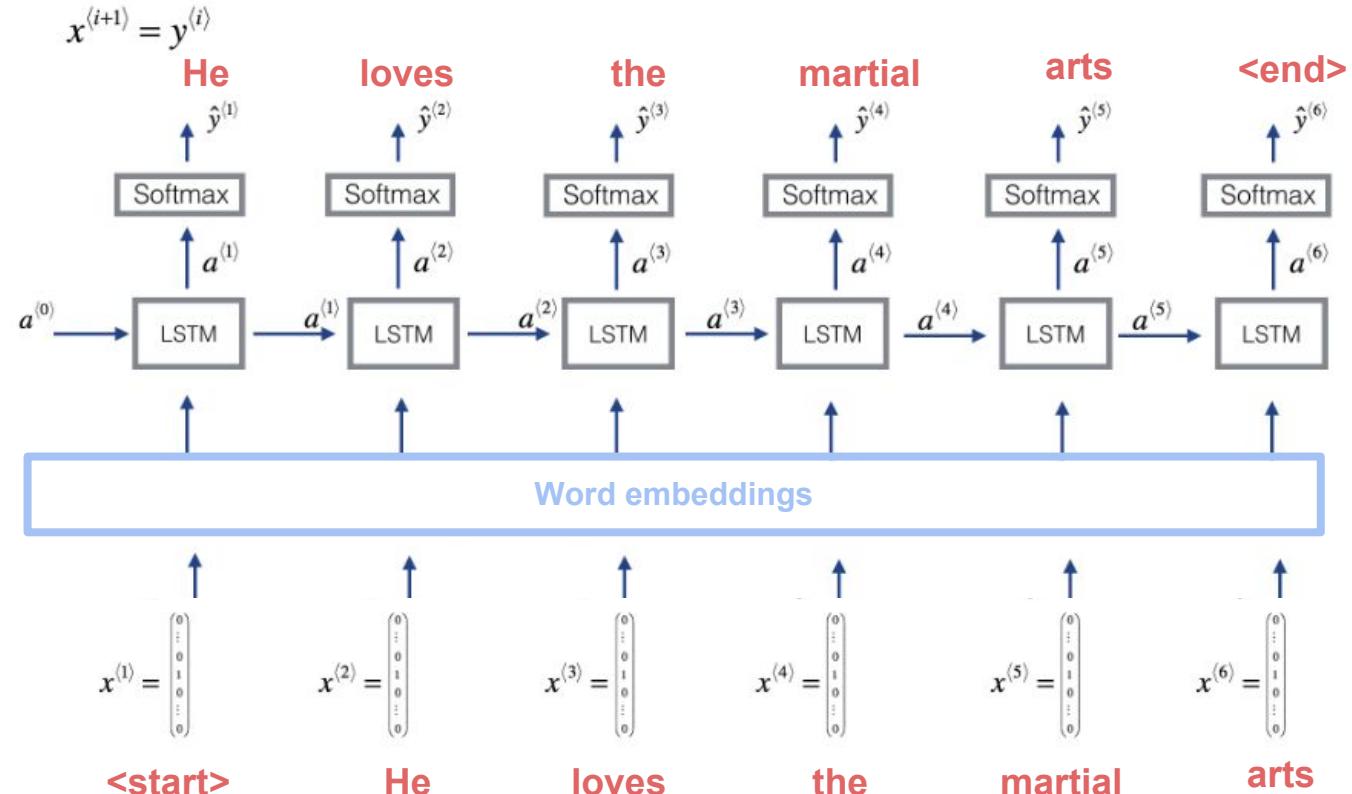
## 3.6 Leaky Units

$$\mathbf{y} = \alpha * \mathcal{F}(\mathbf{x}, \{\mathbf{W}_i\}) + (1 - \alpha) * \mathbf{x}$$

$$0 \leq \alpha \leq 1$$

# 4. Neural Language Modeling

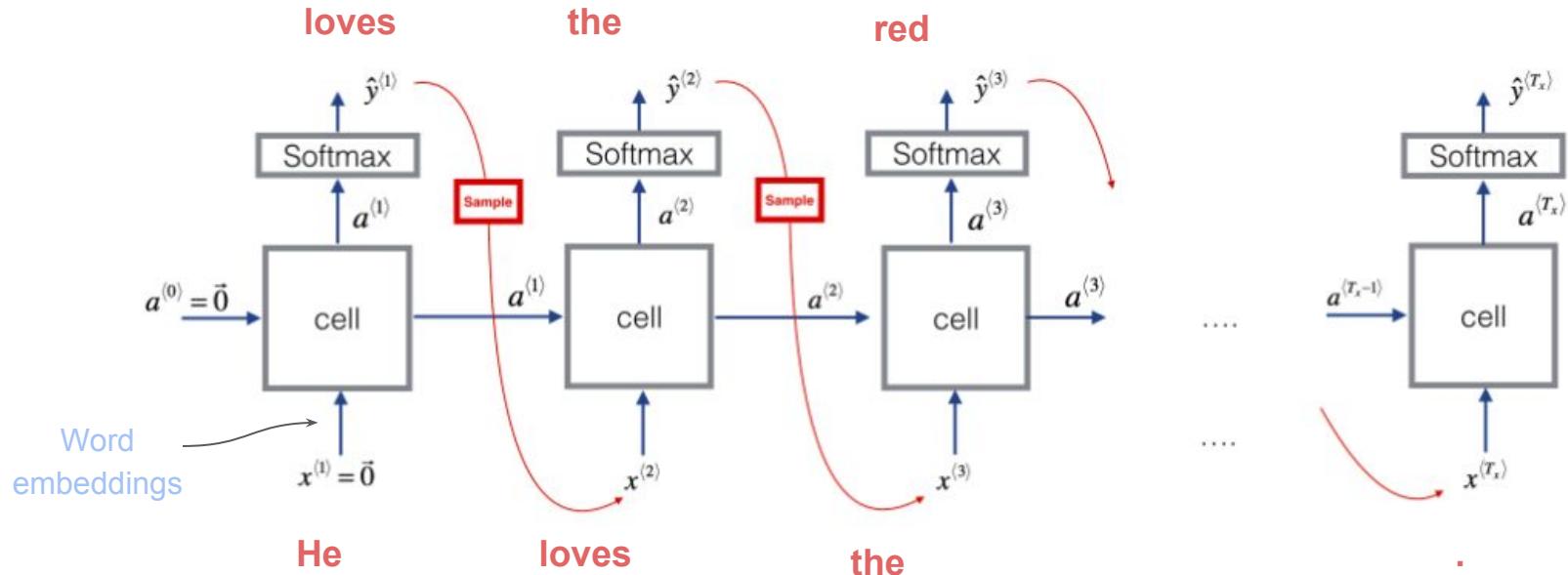
## Language Modeling Train



# 4. Neural Language Modeling

## Application

### Language Modeling Testing



# 5. Sentiment Analysis

Representa un sentimiento positivo, negativo o neutro ??

Aplicación: ¿Yo como empresa quiero saber cuantas personas les gusta el juego “Call of Duty World War II” ?

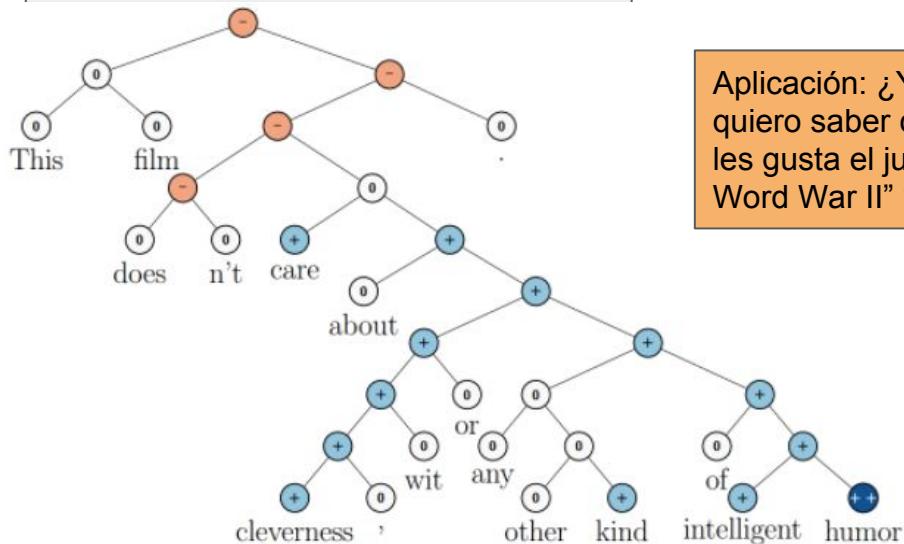
The image shows a vertical stack of five tweets from the Twitter interface:

- Xavier Soto @SheGra\_Soul** · 1 día  
Hey guys, @TheTeamGO is looking for a lead designer, lead editor and a team captain for CoD: WWII. If you have any interests or questions please DM me or @ImCalledWonder RT's are appreciated #videoeditors #gfxdesigners #CallofDutyWWII  
Traducir del inglés
- Gu01312 @Gu01312** · 1 día  
Going live for #CallofDutyWWII at 9CT... tune in... New year, New Gu.  
@HypeSquadTV  
Traducir del inglés
- Patrick 🐻 @HavoKFraged** · 1 día  
Scoping Nubs  
#CallofDutyWWII  
[twitch.tv/havokfraged](http://twitch.tv/havokfraged)  
@HavoKShark @HvKeSports  
#SupportSmallStreamers #EmbraceTheHavok  
Traducir del inglés
- MemeGamesYT @MemeGames0721** · 2 días  
Hoy me e lavantado con ganas de sacarme mi primer arma en oro en #CallofDutyWWII [pic.twitter.com/yRixyPCgzk](http://pic.twitter.com/yRixyPCgzk)

Below the tweets is a video player showing a scene from the game Call of Duty: WWII. The video has a play button in the center and the text "0.00 | 122 reproducciones".

# 5. Sentiment Analysis

Representa un sentimiento positivo, negativo o neutro ??



Aplicación: ¿Yo como empresa quiero saber cuantas personas les gusta el juego “Call of Duty Word War II” ?

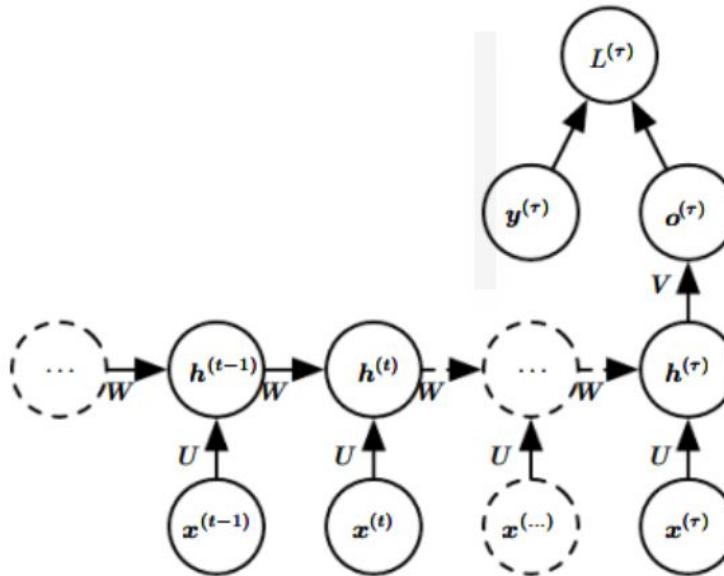
A screenshot of a Twitter feed showing four tweets related to Call of Duty WWII:

- Xavier Soto @SheGra\_Soul · 1 día Hey guys, @TheTeamGō is looking for a lead designer, lead editor and a team captain for CoD: WWII. If you have any interests or questions please DM me or @ImCalledWonder RT's are appreciated #videoeditors #gfxdesigners #CallofDutyWWII  
Traducir del inglés
- Gu01312 @Gu01312 · 1 día Going live for #CallofDutyWWII at 9CT... tune in... New year, New Gu. @HypeSquadTV  
Traducir del inglés
- Patrick ⚡ @HavokFragged · 1 día Scoping Nubs  
#CallofDutyWWII  
twitch.tv/havokfragged  
@HavokShark @HvKeSports  
#SupportSmallStreamers #EmbraceTheHavok  
Traducir del inglés
- MemeGamesYT @MemeGames0721 · 2 días Hoy me e lavantado con ganas de sacarme mi primer arma en oro en #CallofDutyWWII pic.twitter.com/yRixyPCgzk

Below the tweets is a video player showing a scene from Call of Duty: World War II.

# 5. Sentiment Analysis

## RNN for many-to-one



(a) RNN model to mapping an input sequence  $(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(\tau)})$  to a fixed vector  $\mathbf{o}^{\tau}$ .

# 6. Text Summarization

The screenshot shows a Bing search results page. The search bar contains the query "automatic summarization". Below the search bar, there are tabs for "Todos" (selected), "Imágenes", "Vídeos", and "Noticias", followed by a link to "Mis elementos guardados". The search results indicate 209.000 resultados. There are filters for "Fecha", "Idioma", and "Región". The first result is a Wikipedia page titled "Automatic summarization - Wikipedia" with the URL [https://en.wikipedia.org/wiki/Automatic\\_summarization](https://en.wikipedia.org/wiki/Automatic_summarization). A red dashed box highlights the first paragraph of the summary: "Automatic summarization is the process of shortening a text document with software, ... Conceptual artwork using automatic summarization software in Microsoft Word 2008." Below this, there are links for "Types · Applications and ... · Evaluation techniques". The second result is a PDF titled "A Survey on Automatic Text Summarization - Carnegie ..." with the URL [www.cs.cmu.edu/~nsmith/L2/das-martins.07.pdf](http://www.cs.cmu.edu/~nsmith/L2/das-martins.07.pdf). A red dashed box highlights the author information: "A Survey on Automatic Text Summarization Dipanjan Das Andr e F.T. Martins Language Technologies Institute Carnegie Mellon University fdipanjan, afmg@cs.cmu.edu".

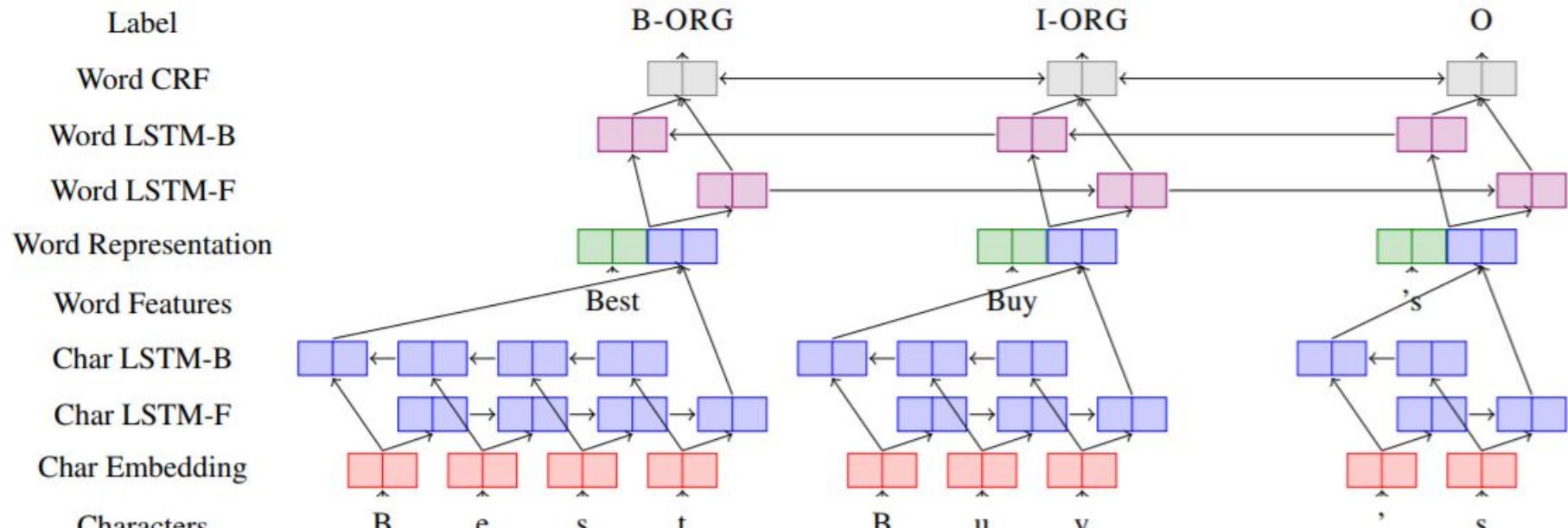
Pueden ser resumenes

A Survey on Automatic Text Summarization - Carnegie ...

**Tipo Extractivo:** Si todo el resumen está conformado por palabras presente en el texto.

**Tipo abstracto:** Si el resumen no necesariamente usa palabras dentro del texto. Extrapolación.

# 6. Name Entity Recognition(NER)



Word+Character level NN architecture for NER.

# 7. Opinion Mining

¿ Realmente tiene  
depresion ?

The screenshot shows a Twitter search results page for the hashtag #depression. The interface includes a header with navigation links: Destacados, Más reciente, Personas, Fotos, Videos, Noticias, and Transmisiones. Below the header, there's a sidebar with 'Tendencias para ti' (Trends for you) featuring hashtags like #Centenario, #Dakar2018, #PinoXfiles, #Pasamayo, #Kenji, #Aveengers, #MaritzaGarcía, #Latina, #FridayFeeling, and #FranciscoenPerú. The main content area displays several tweets from users like myselfandhealth, Finding Relief, mimi, Depression Roots, and Dr Keith Grimes, each discussing depression. Red dashed boxes highlight specific tweets from myselfandhealth, Finding Relief, mimi, and Dr Keith Grimes.

#depression

Destacados Más reciente Personas Fotos Videos Noticias Transmisiones

Encuentra a personas que conoces Importa tus contactos de Gmail

Conecta otras libretas de direcciones

Tendencias para ti · Cambiar

Centenario 6.691 Tweets

#Dakar2018 @pcperu está twitteando sobre esto

PinoXfiles 13.8 K Tweets

Pasamayo 7.227 Tweets

Kenji 13.9 K Tweets

Aveengers 14.1 K Tweets

Maritza García

Latina 55.3 K Tweets

#FridayFeeling 31.3 K Tweets

#FranciscoenPerú

© 2018 Twitter · Sobre nosotros · Centro de Ayuda · Condiciones · Política de privacidad · Cookies · Información sobre anuncios

myselfandhealth @myselfandhealth · 31 dic. 2017

Don't let anyone make you feel bad for not wanting to be around people or declining an invitation. Instead, applaud yourself for knowing when to say no to someone else's social agenda! #depression #anxiety

Finding Relief @FindingRelief · 29 dic. 2017

There's nothing to feel weak or ashamed about if we take a day or three to rest and rejuvenate because of symptoms of mental illness. Part of how we treat all illness, physical or mental, is through appropriate care when it is needed. #recovery #depression #anxiety #SickNotWeak

mimi @BPDmimi · 30 dic. 2017

I'm struggling so much. Feel so lonely it hurts. Horrible time of year. I should've planned to do more/see friends between xmas & new year. But unfortunately I've lost a lot of friends due to my depression. Ironic, but now I have no one when I need them the most. #BPD #Depression

Depression Roots @DepressionRoots · 1 día

Childhood with neglect, unequal parental treatment of siblings, physical or sexual abuse increases the risk of developing #depression.

Dr Keith Grimes @keithgrimes · 1 día

Day 14: Talking to myself

Globally, #Depression affects more than 300million people and is the leading cause of disability

<http://www.twitter.com>

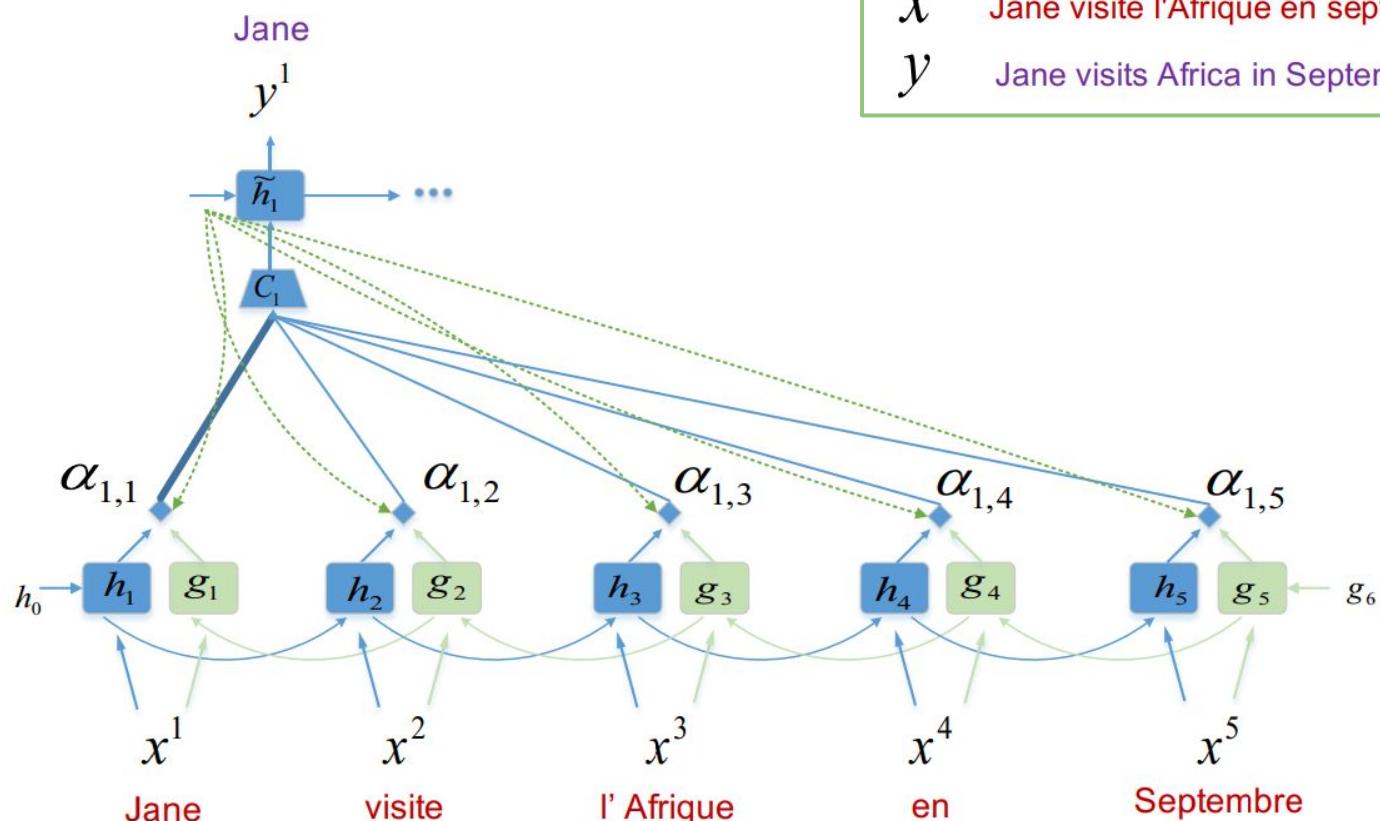
## 8. Attention / Self Attention

$$y = f(x)$$

$x$  Jane visite l'Afrique en septembre

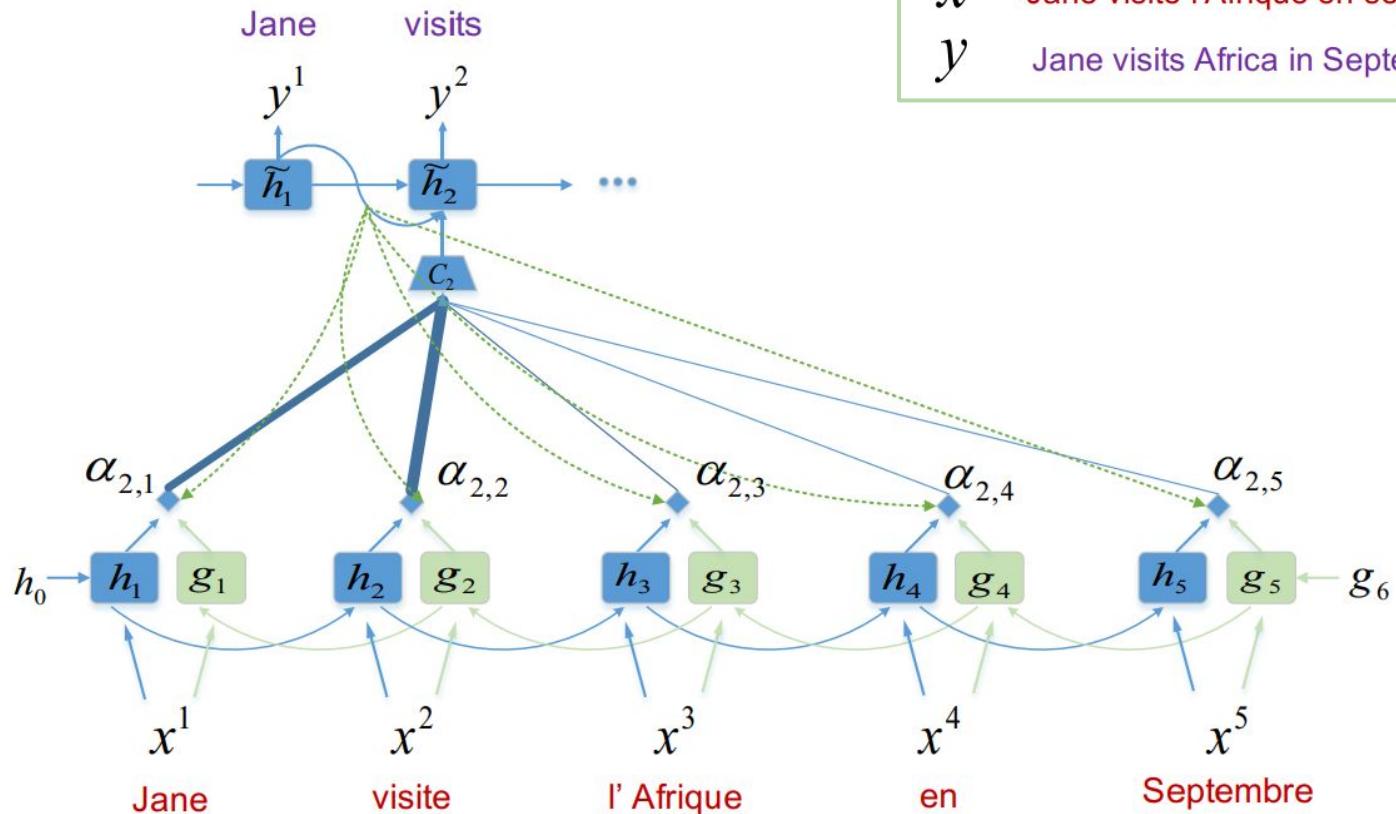
$y$  Jane visits Africa in September

# 8. Attention / Self Attention



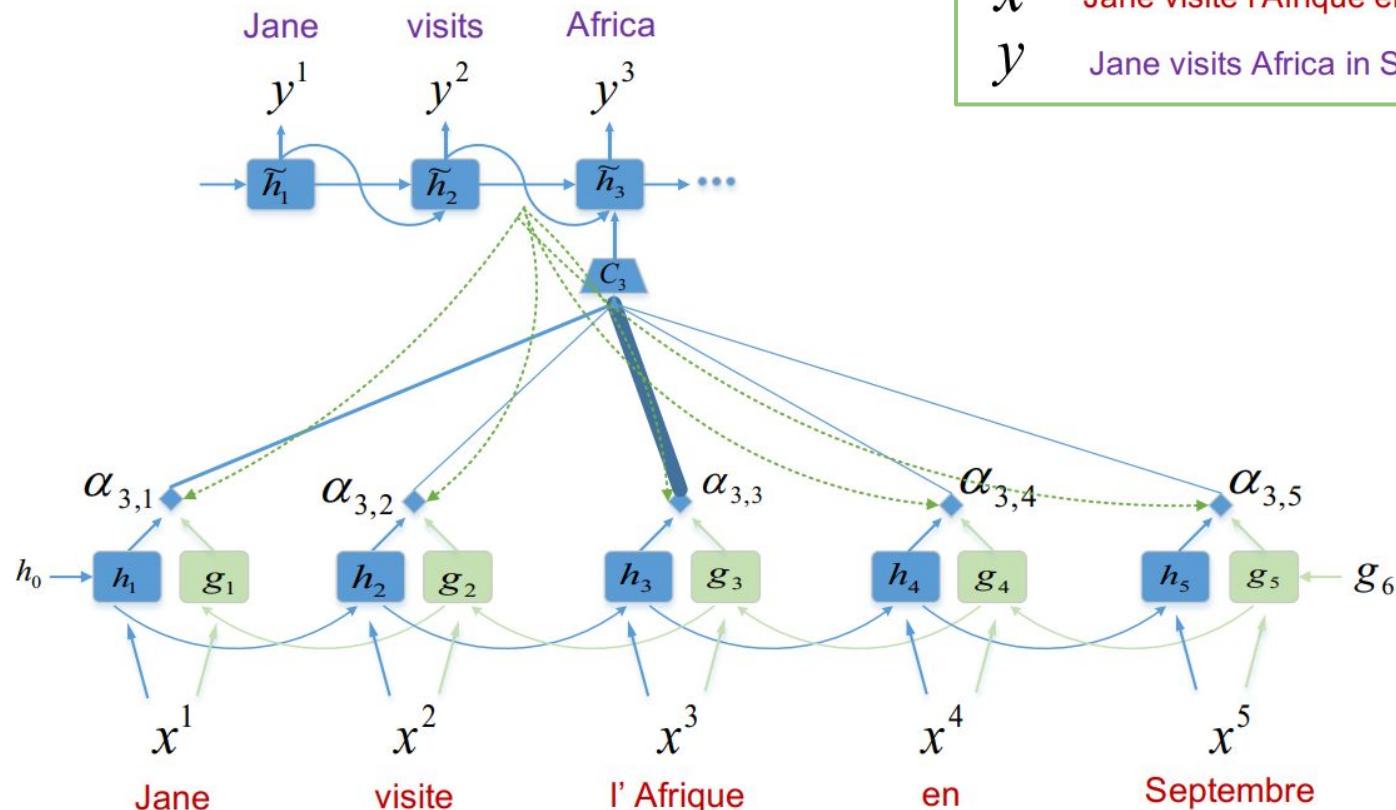
Atención en *Machine Translation* en el paso  $t=1$

# 8. Attention / Self Attention



Atención en *Machine Translation* en el paso  $t=1$

# 8. Attention / Self Attention



Atención en *Machine Translation* en el paso  $t=2$

# 8. Attention / Self Attention

Formulación matemática

Score value

$$h_1^c = [h_1, g_1]$$

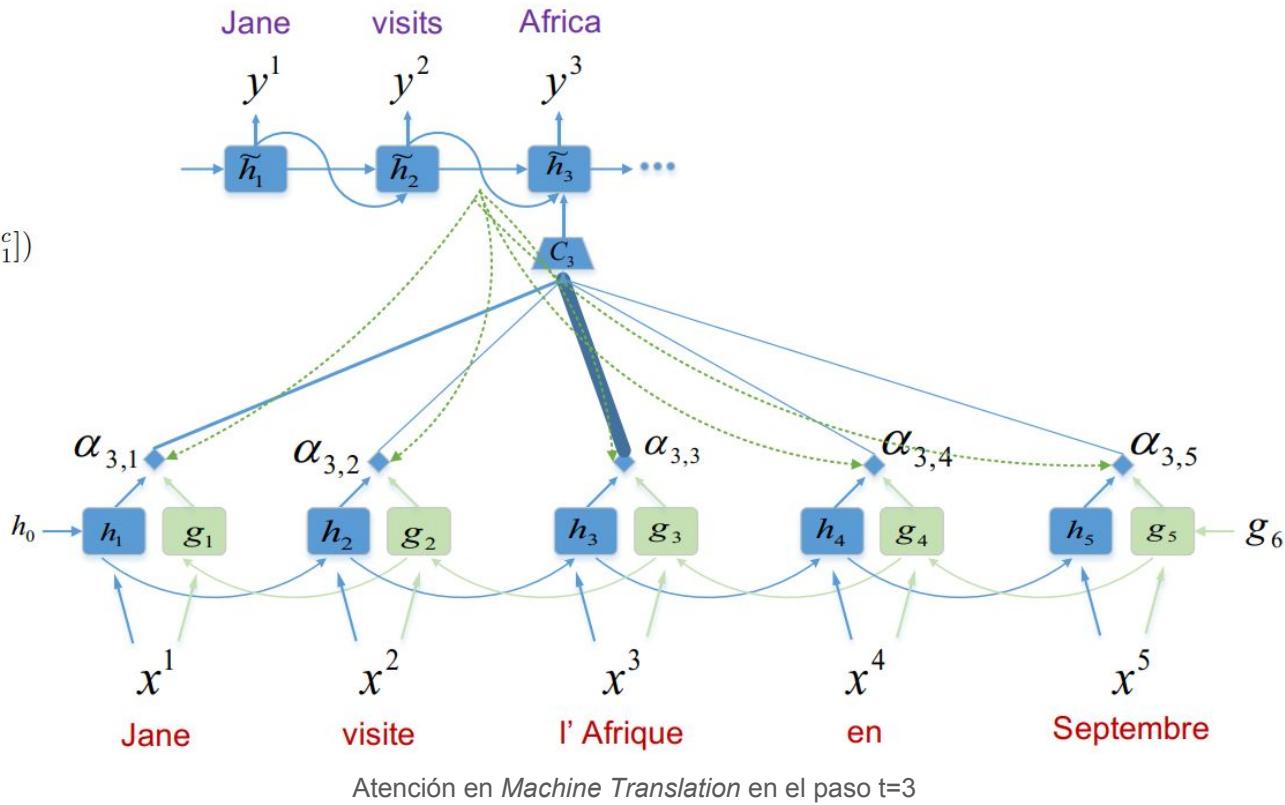
$$e_{3,1} = v_a^T \tanh(W_a[\tilde{h}_2, h_1^c])$$

$$\alpha_{3,1} = \frac{\exp(e_{3,1})}{\sum_{k=1}^5 \exp(e_{3,k})}$$

Score attention

Recordando

$$\sum_{k=1}^{T_x} \alpha_{i,k} = 1$$



Atención en Machine Translation en el paso  $t=3$

# 8. Attention / Self Attention

Formulación matemática

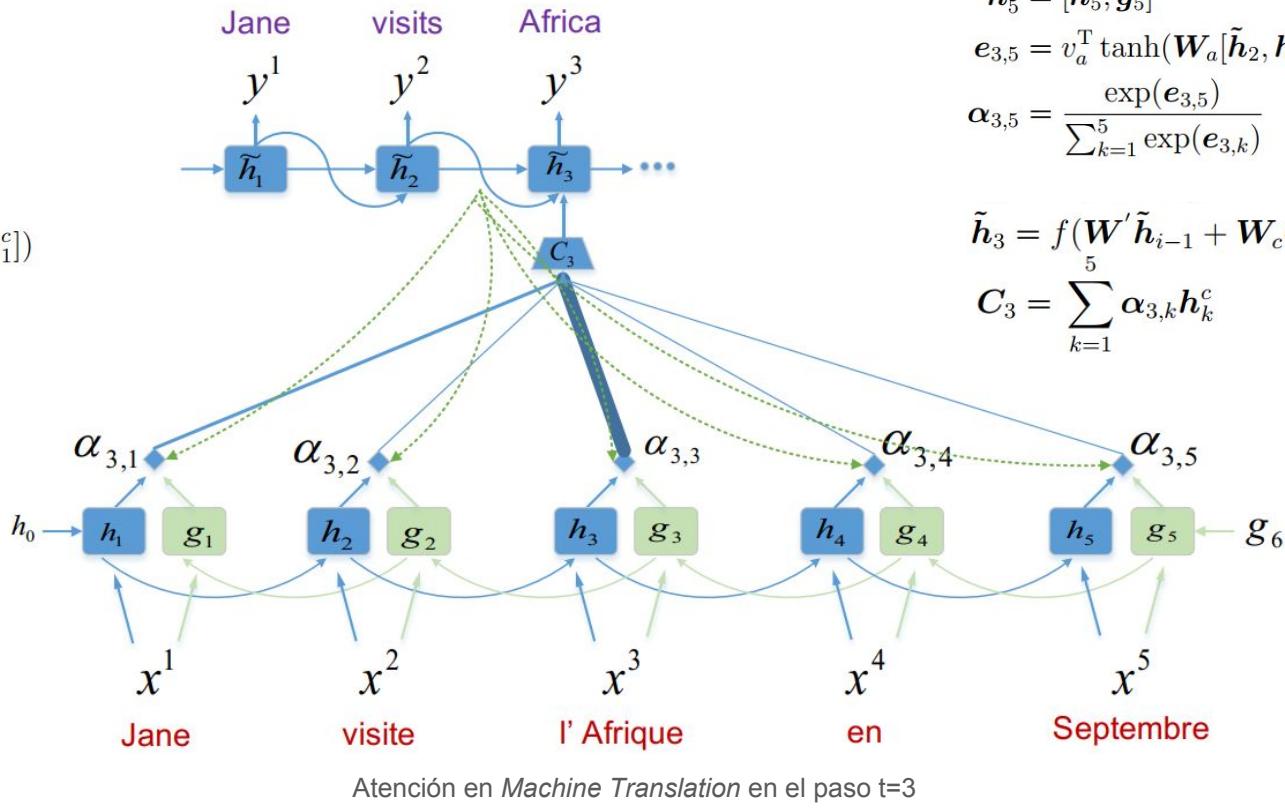
Score value

$$\mathbf{h}_1^c = [\tilde{\mathbf{h}}_1, \mathbf{g}_1]$$

$$e_{3,1} = v_a^T \tanh(\mathbf{W}_a[\tilde{\mathbf{h}}_2, \mathbf{h}_1^c])$$

$$\alpha_{3,1} = \frac{\exp(e_{3,1})}{\sum_{k=1}^5 \exp(e_{3,k})}$$

Score attention



$$\mathbf{h}_5^c = [\tilde{\mathbf{h}}_5, \mathbf{g}_5]$$

$$e_{3,5} = v_a^T \tanh(\mathbf{W}_a[\tilde{\mathbf{h}}_2, \mathbf{h}_5^c])$$

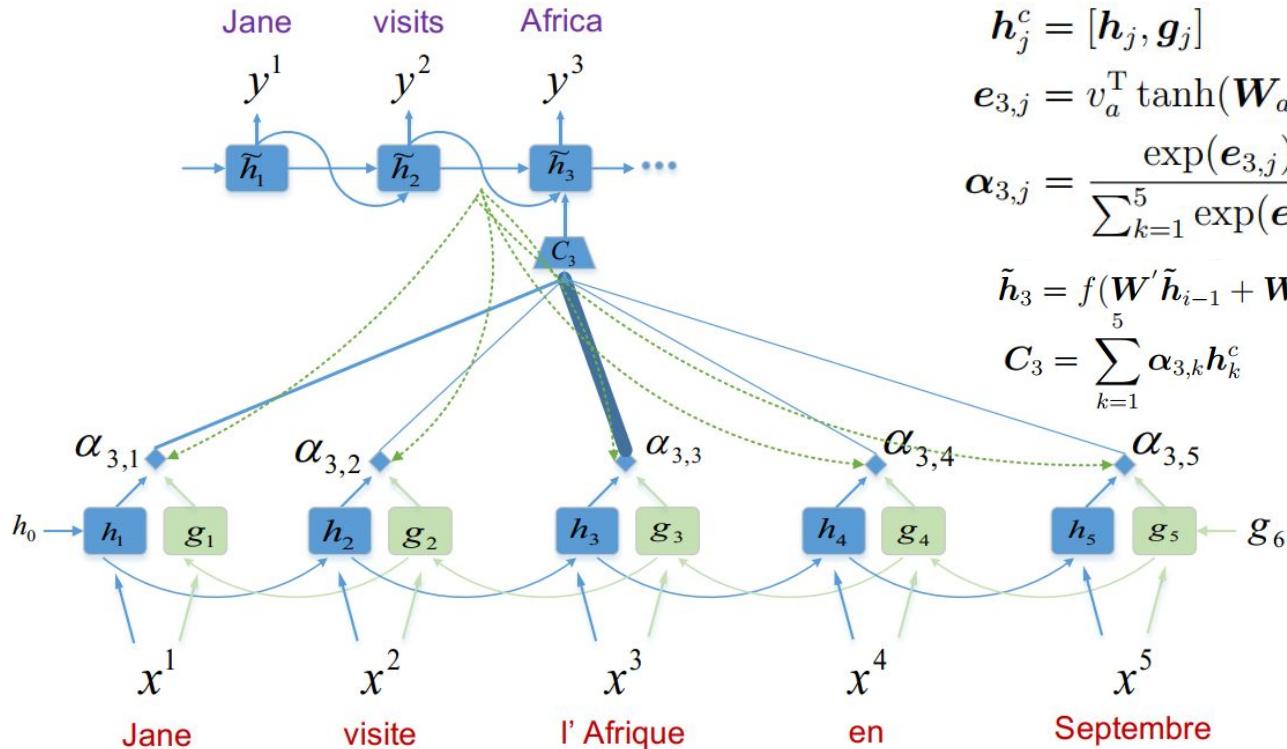
$$\alpha_{3,5} = \frac{\exp(e_{3,5})}{\sum_{k=1}^5 \exp(e_{3,k})}$$

$$\tilde{\mathbf{h}}_3 = f(\mathbf{W}' \tilde{\mathbf{h}}_{i-1} + \mathbf{W}_c \mathbf{C}_3 + \mathbf{b}_{eh})$$

$$\mathbf{C}_3 = \sum_{k=1}^5 \alpha_{3,k} \mathbf{h}_k^c$$

# 8. Attention / Self Attention

Formulación matemática



En general para el paso  $t = 3$

$$\begin{aligned}\mathbf{h}_j^c &= [\mathbf{h}_j, \mathbf{g}_j] \\ e_{3,j} &= v_a^T \tanh(\mathbf{W}_a[\tilde{\mathbf{h}}_2, \mathbf{h}_j^c]) \\ \alpha_{3,j} &= \frac{\exp(e_{3,j})}{\sum_{k=1}^5 \exp(e_{3,k})}\end{aligned}$$

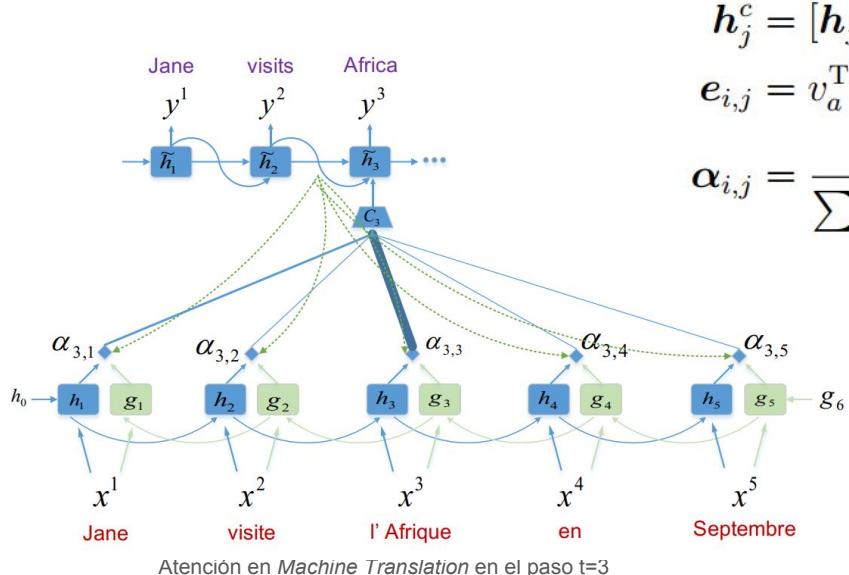
$$\tilde{\mathbf{h}}_3 = f(\mathbf{W}'\tilde{\mathbf{h}}_{i-1} + \mathbf{W}_c C_3 + \mathbf{b}_{eh})$$

$$C_3 = \sum_{k=1}^5 \alpha_{3,k} \mathbf{h}_k^c$$

Atención en *Machine Translation* en el paso  $t=2$

# 8. Attention / Self Attention

Formulación matemática



Para una sentencia  $x$  de tamaño  $T_x$

$$\mathbf{h}_j^c = [\mathbf{h}_j, \mathbf{g}_j]$$

$$e_{i,j} = v_a^T \tanh(\mathbf{W}_a[\tilde{\mathbf{h}}_{i-1}, \mathbf{h}_j^c]) \quad \text{Score value}$$

$$\alpha_{i,j} = \frac{\exp(e_{i,j})}{\sum_{k=1}^{T_x} \exp(e_{i,k})} \quad \text{Score attention}$$

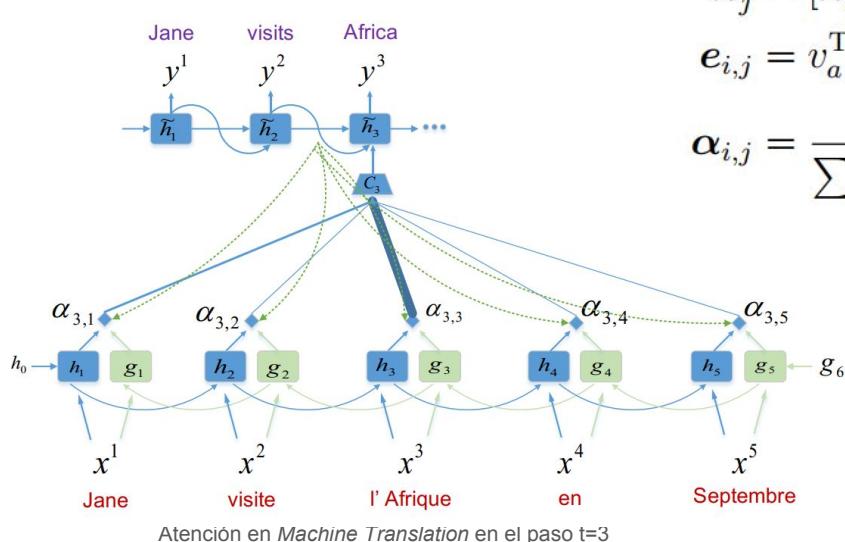
$$\tilde{\mathbf{h}}_3 = f(\mathbf{W}' \tilde{\mathbf{h}}_{i-1} + \mathbf{W}_c C_3 + \mathbf{b}_{eh})$$

$$C_i = \sum_{k=1}^{T_x} \alpha_{i,k} \mathbf{h}_k^c$$

Cual score aprende a quién dar más atención ??

# 8. Attention / Self Attention

Formulación matemática



Para una sentencia  $x$  de tamaño  $T_x$

$$\mathbf{h}_j^c = [\mathbf{h}_j, \mathbf{g}_j]$$

$$e_{i,j} = v_a^T \tanh(\mathbf{W}_a[\tilde{\mathbf{h}}_{i-1}, \mathbf{h}_j^c]) \quad \text{Score value}$$

$$\alpha_{i,j} = \frac{\exp(e_{i,j})}{\sum_{k=1}^{T_x} \exp(e_{i,k})} \quad \text{Score attention}$$

$$\tilde{\mathbf{h}}_3 = f(\mathbf{W}' \tilde{\mathbf{h}}_{i-1} + \mathbf{W}_c C_3 + \mathbf{b}_{eh})$$

$$C_i = \sum_{k=1}^{T_x} \alpha_{i,k} \mathbf{h}_k^c$$

Cual score aprende a quién dar más atención ??

El score value

# 8. Attention / Self Attention

Distintos *attention* ~ distintas formas de calcular *score value*

En resumen un *score value* es calculado a partir de 2 valores, en nuestro ejemplo  $\tilde{h}_{i-1}$  y  $h_j^c$

$$e_{i,j} = \mathcal{F}(\tilde{h}_{i-1}, h_j^c)$$

Mecanismo	Score value
Bahdanau	$e_{i,j} = v_a^T \tanh(\mathbf{W}_a[\tilde{h}_{i-1}, h_j^c])$
Luong	$e_{i,j} = \tilde{h}_{i-1}'^T h_j^c$
Bahdanau Normado	$e_{i,j} = g \frac{v_a}{\ v_a\ } \tanh(\mathbf{W}_a[\tilde{h}_{i-1}, h_j^c])$
Luong Escalado*	$e_{i,j} = g \tilde{h}_{i-1}'^T h_j^c$

Considerar, por fines didácticos usamos  $i-1$

- [] Neural Machine Translation by Jointly Learning to Align and Translate, Bahdanau, 2014.
- [] Online and linear-time attention by enforcing monotonic alignments, Raffel, 2017.
- [] Effective Approaches to Attention-based Neural Machine Translation, Luong, 2015
- [] Attention is all you need, Vaswani, 2017.
- [] Tensorflow Attention [https://www.tensorflow.org/api\\_docs/python/tf/contrib/seq2seq/AttentionWrapper](https://www.tensorflow.org/api_docs/python/tf/contrib/seq2seq/AttentionWrapper)

# 8. Attention / Self Attention

Distintos *attention* ~ distintas formas de calcular *score value*

En resumen un *score value* es calculado a partir de 2 valores, en nuestro ejemplo  $\tilde{h}_{i-1}$  y  $h_j^c$

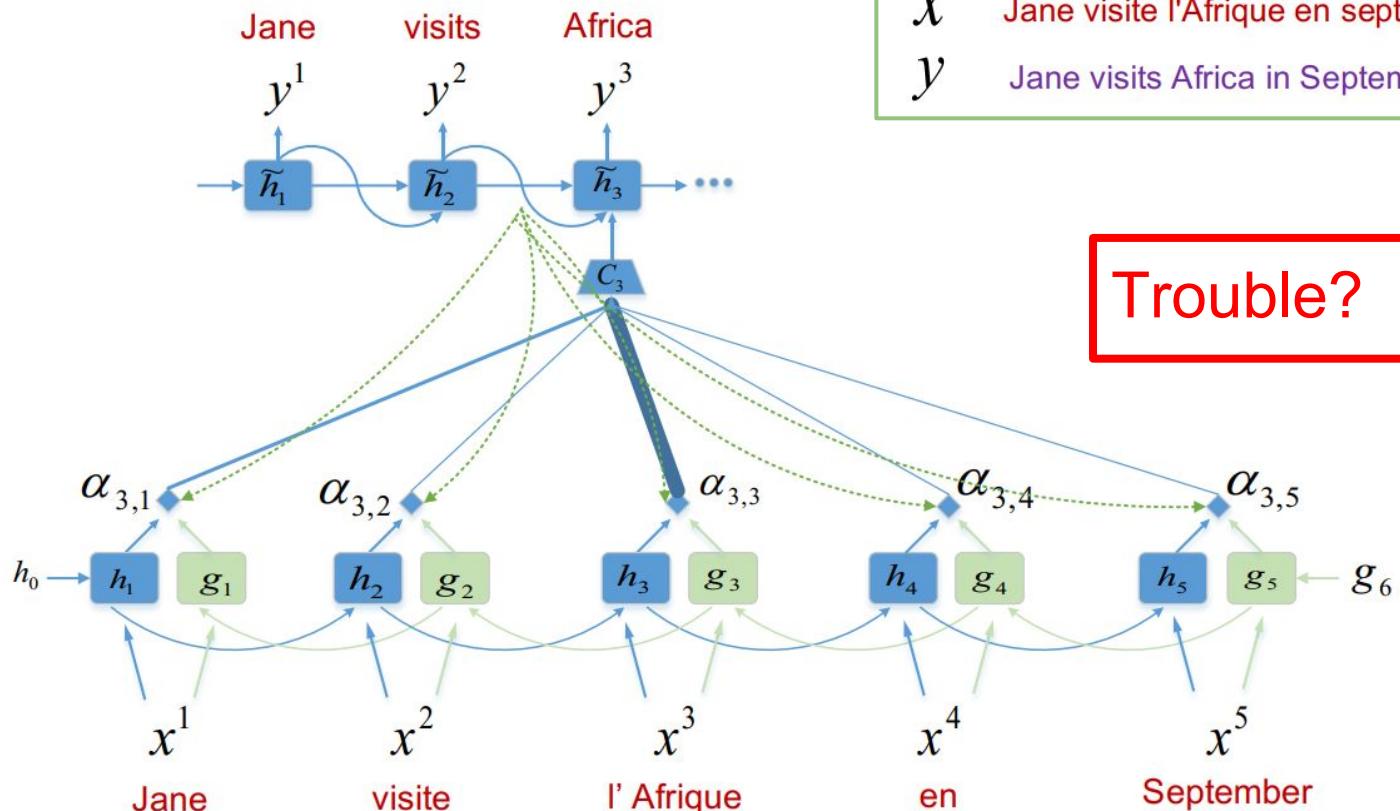
$$e_{i,j} = \mathcal{F}(\tilde{h}_{i-1}, h_j^c)$$

Mecanismo	Score value
Bahdanau	$e_{i,j} = v_a^T \tanh(\mathbf{W}_a[\tilde{h}_{i-1}, h_j^c])$
Luong	$e_{i,j} = \tilde{h}_i'^T h_j^c$
Bahdanau Normado	$e_{i,j} = g \frac{v_a}{\ v_a\ } \tanh(\mathbf{W}_a[\tilde{h}_{i-1}, h_j^c])$
Luong Escalado*	$e_{i,j} = g \tilde{h}_i'^T h_j^c$

Siendo exactos  
con la literatura

- [] Neural Machine Translation by Jointly Learning to Align and Translate, Bahdanau, 2014.
- [] Online and linear-time attention by enforcing monotonic alignments, Raffel, 2017.
- [] Effective Approaches to Attention-based Neural Machine Translation, Luong, 2015
- [] Attention is all you need, Vaswani, 2017.
- [] Tensorflow Attention [https://www.tensorflow.org/api\\_docs/python/tf/contrib/seq2seq/AttentionWrapper](https://www.tensorflow.org/api_docs/python/tf/contrib/seq2seq/AttentionWrapper)

# 8. Attention / Self Attention



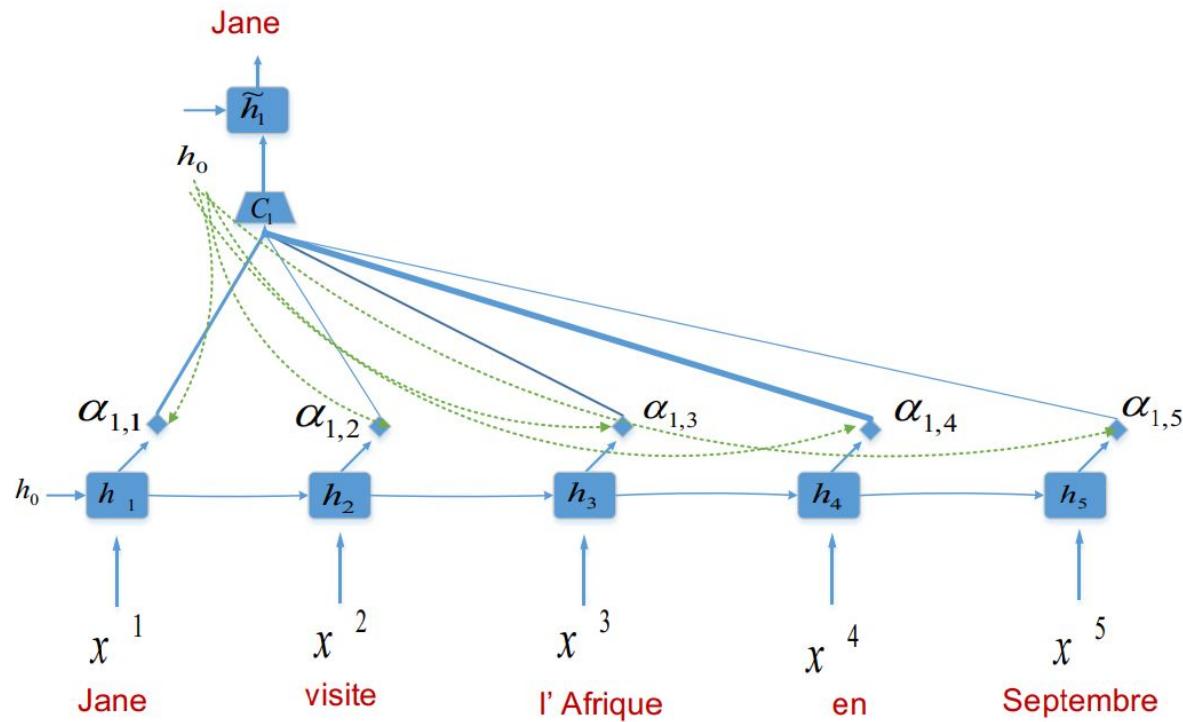
Atención en *Machine Translation* en el paso  $t=3$

# 8. Attention / Self Attention

Expresiones	
Expresión	Add insult to injury.
Significado	To make a bad situation worse.
Traducción equivocada	Agregar un insulto a la herida.
Traducción correcta	De mal en peor.
Expresión	Barking up the wrong tree.
Significado	To look for a solution in the wrong place.
Traducción equivocada	Ladrar al árbol equivocado.
Traducción correcta	Buscas algo en donde no existe
Expresión	Bite the bullet.
Significado	To get something over with because it is inevitable.
Traducción equivocada	Morder la bala.
Traducción correcta	Hacer algo si o si.

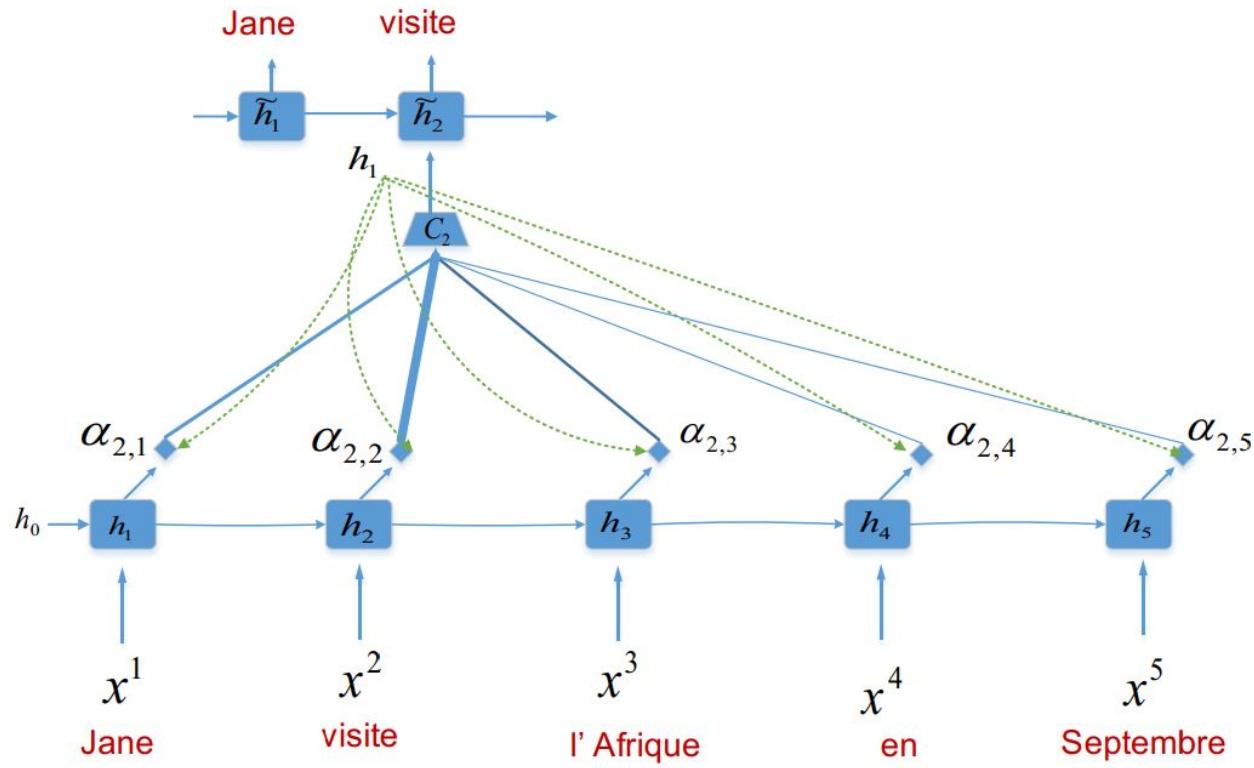
Ejemplo donde fallaría atención para *Machine Translation*

# 8. Attention / Self Attention



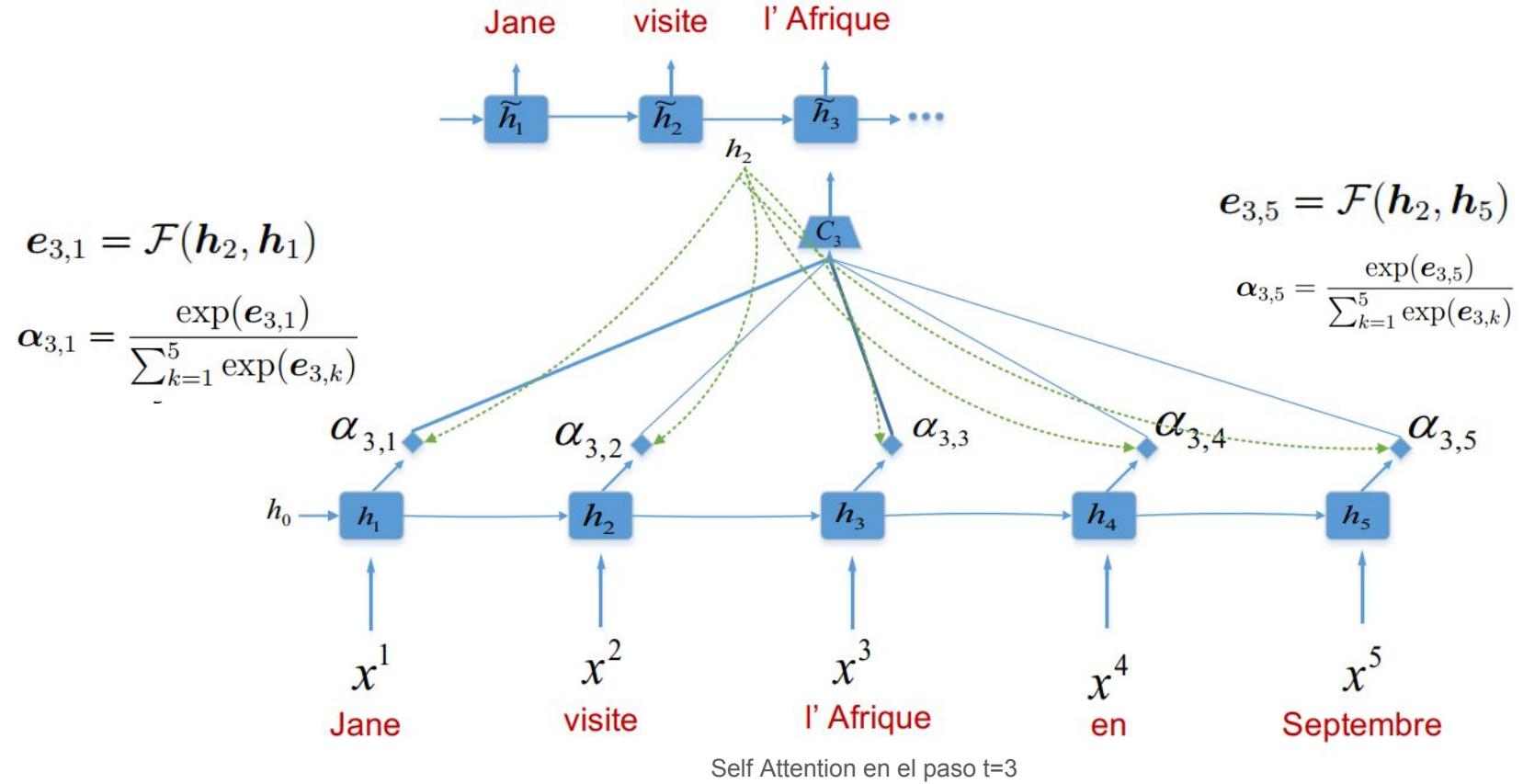
Self Attention en el paso  $t=1$

# 8. Attention / Self Attention

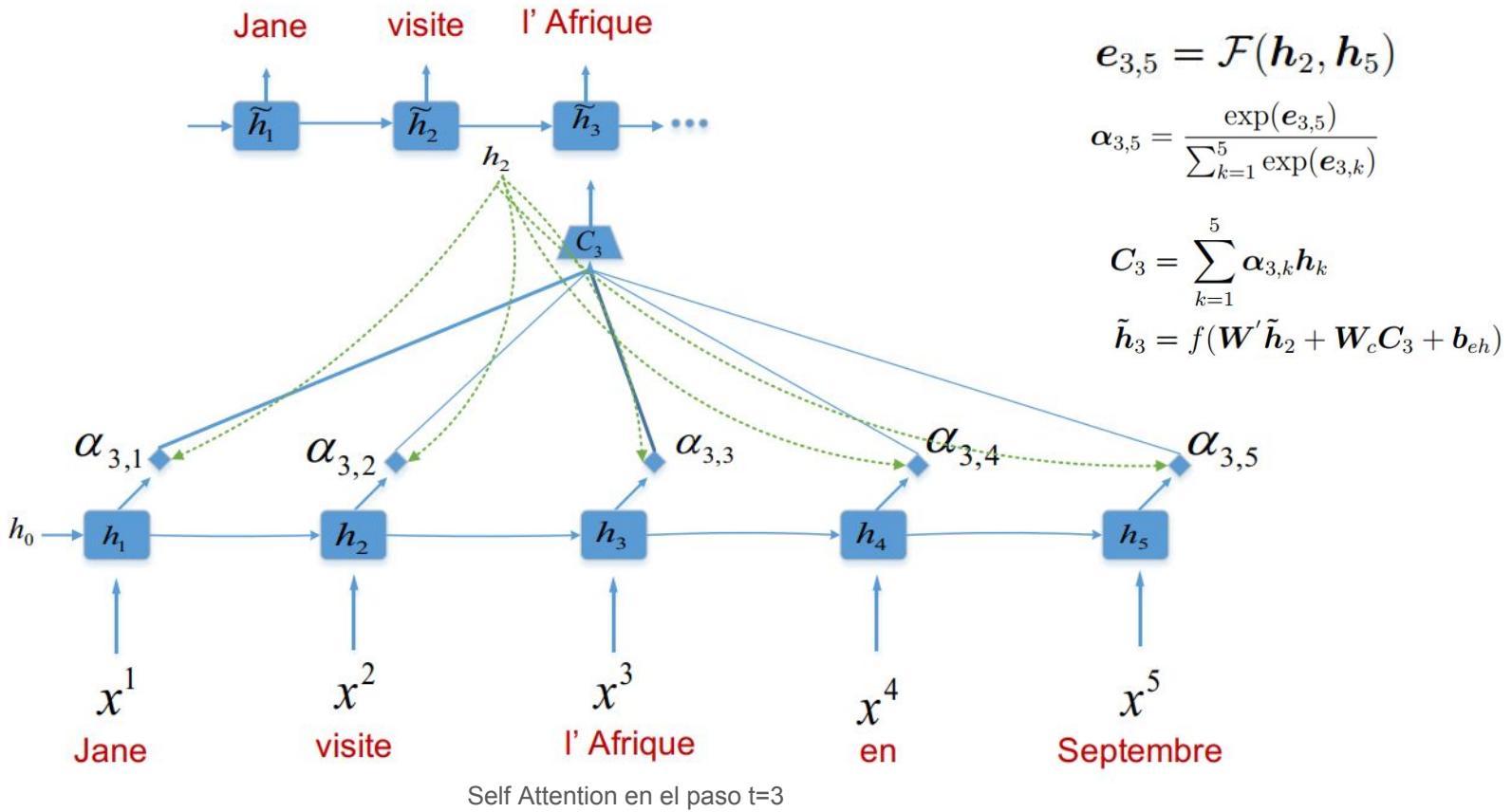


Self Attention en el paso  $t=2$

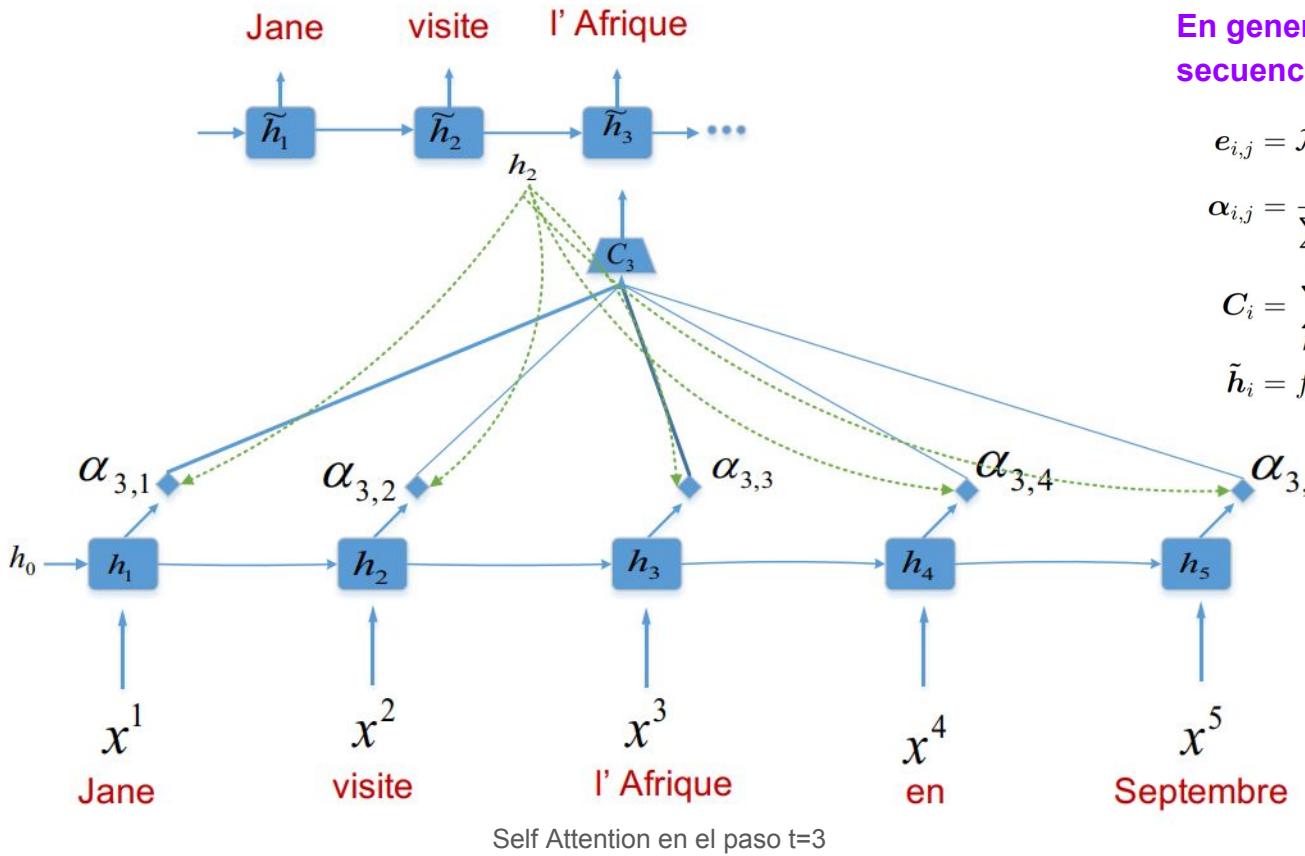
# 8. Attention / Self Attention



# 8. Attention / Self Attention



# 8. Attention / Self Attention



En general para  $t = i$  y una secuencia  $T_x$

$$e_{i,j} = \mathcal{F}(h_{i-1}, h_j)$$

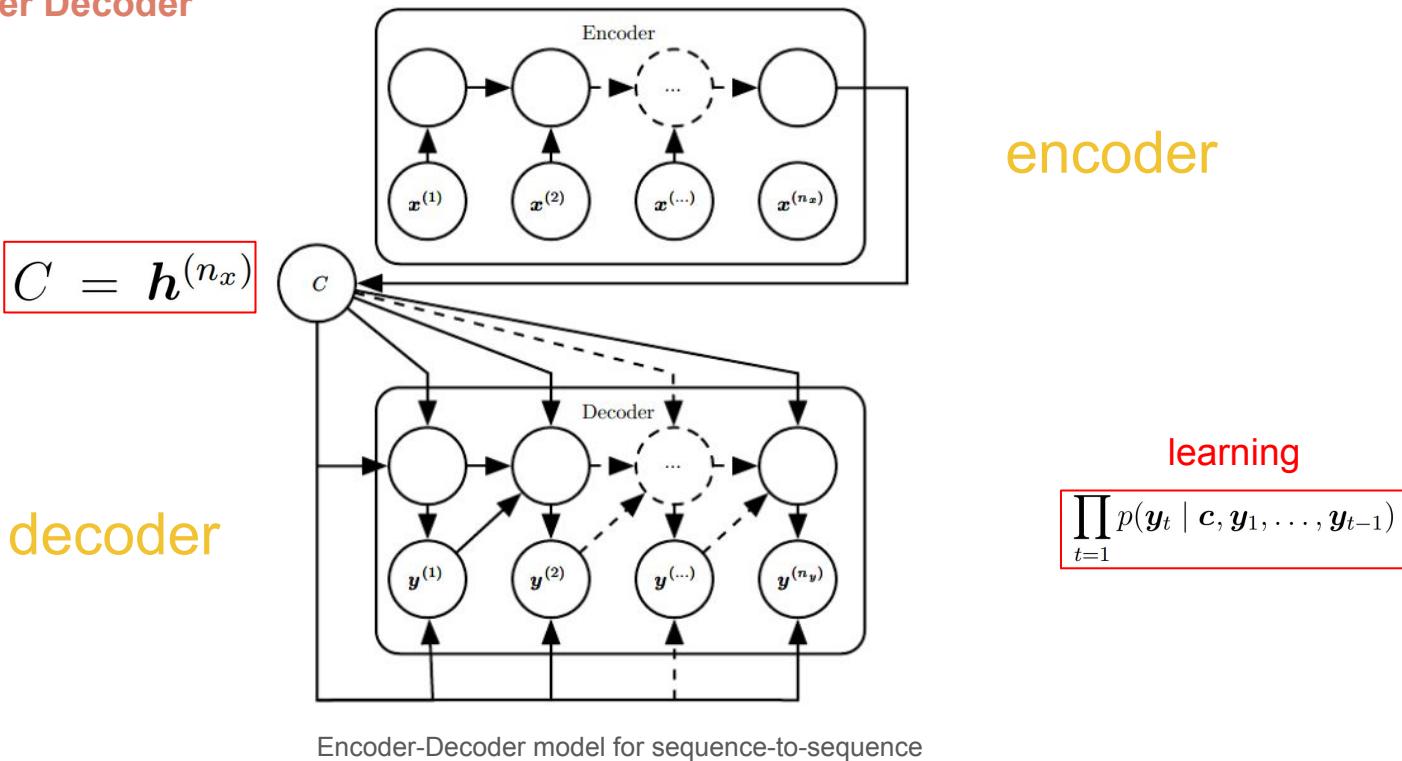
$$\alpha_{i,j} = \frac{\exp(e_{i,j})}{\sum_{k=1}^{T_x} \exp(e_{i,k})}$$

$$C_i = \sum_{k=1}^{T_x} \alpha_{i,k} h_k$$

$$\tilde{h}_i = f(W' \tilde{h}_{i-1} + W_c C_i + b_{eh})$$

# 9. Machine Translation

## 9.1 Encoder Decoder



# 9. Machine Translation

## 9.1 Encoder Decoder

encoder

$$\mathbf{h}_t = f(\mathbf{W}[e(\mathbf{x}_t), \mathbf{h}_{t-1}] + \mathbf{b}_e)$$

$$\vdots = \vdots$$

$$\mathbf{h}_{n_x} = f(\mathbf{W}[e(\mathbf{x}_{n_x}), \mathbf{h}_{n_x-1}] + \mathbf{b}_e)$$

$$\mathbf{c} = g(\mathbf{V} \mathbf{h}_{n_x})$$

decoder

$$\mathbf{h}'_0 = \tanh(\mathbf{V}' \mathbf{c})$$

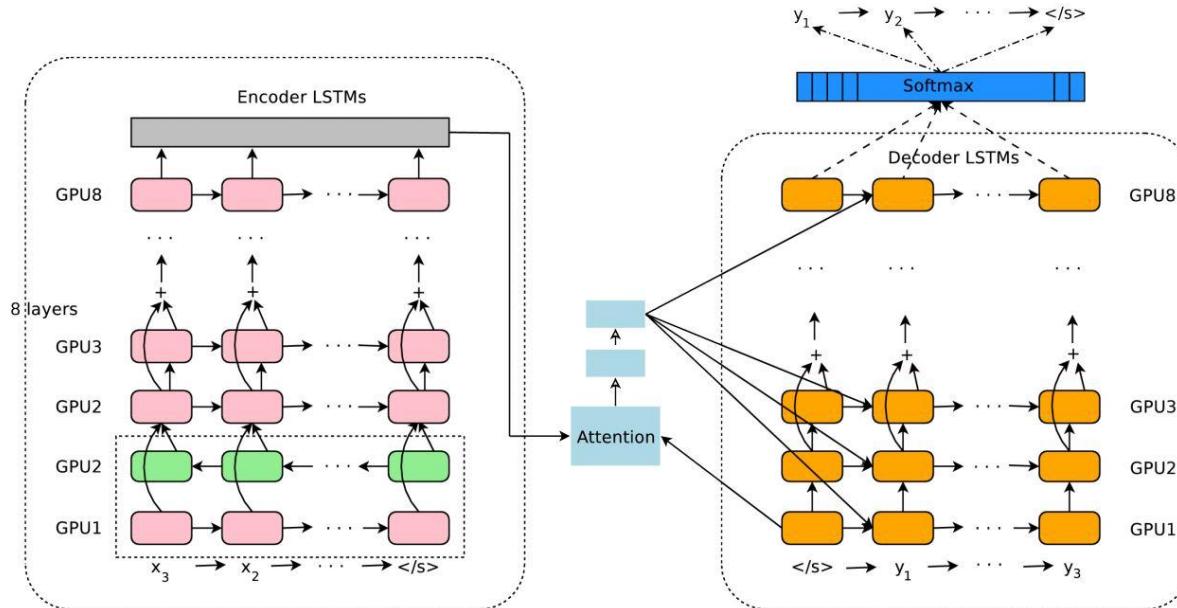
$$\mathbf{h}'_{t'} = f(\mathbf{W}' \mathbf{h}'_{t'-1} + \mathbf{U} e(\mathbf{y}_{t'-1}) + \mathbf{C} \mathbf{c} + \mathbf{b}_{dh})$$

$$\mathbf{s}_{t'} = o(\mathbf{O}_h \mathbf{h}'_{t'} + \mathbf{O}_y e(\mathbf{y}_{t'-1}) + \mathbf{O}_c \mathbf{c} + \mathbf{b}_{do})$$

# 9. Machine Translation

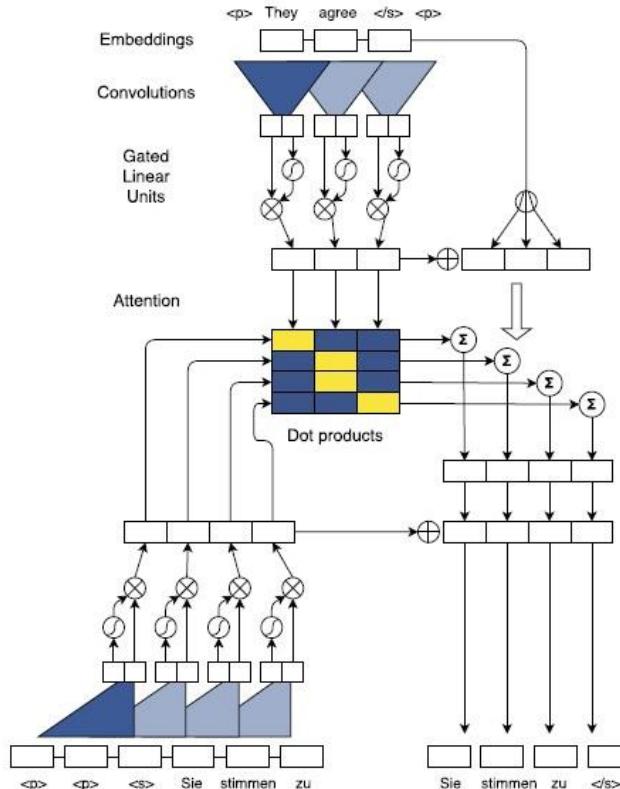
## 9.2 Google Neural Machine Translation

Architecture based Google Neural Machine Translation: Encoder ( stack RNN with first layer BiRNN + residual connections) + Attention Mechanics + Decoder(stack RNN + residual connections ) + **Special Technique of Transfer Learning**.



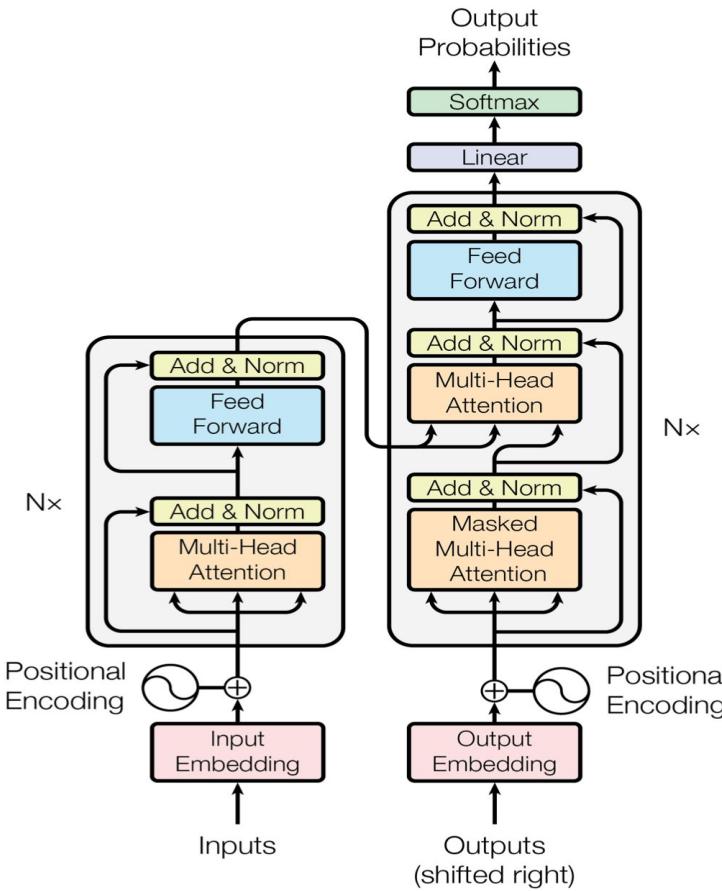
# 9. Machine Translation

## 9.3 Conv2Seq



# 9. Machine Translation

## 9.4 Transformer



# 10. Image Caption

A person riding a motorcycle on a dirt road.



Two dogs play in the grass.



A skateboarder does a trick on a ramp.



A group of young people playing a game of frisbee.



Two hockey players are fighting over the puck.



A little girl in a pink hat is blowing bubbles.



A herd of elephants walking across a dry grass field.



A close up of a cat laying on a couch.



A red motorcycle parked on the side of the road.

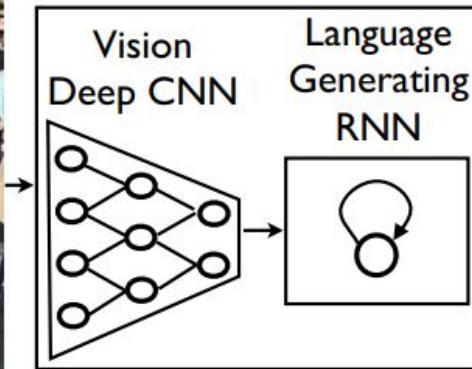


Describes without errors

Describes with minor errors

Somewhat related to the image

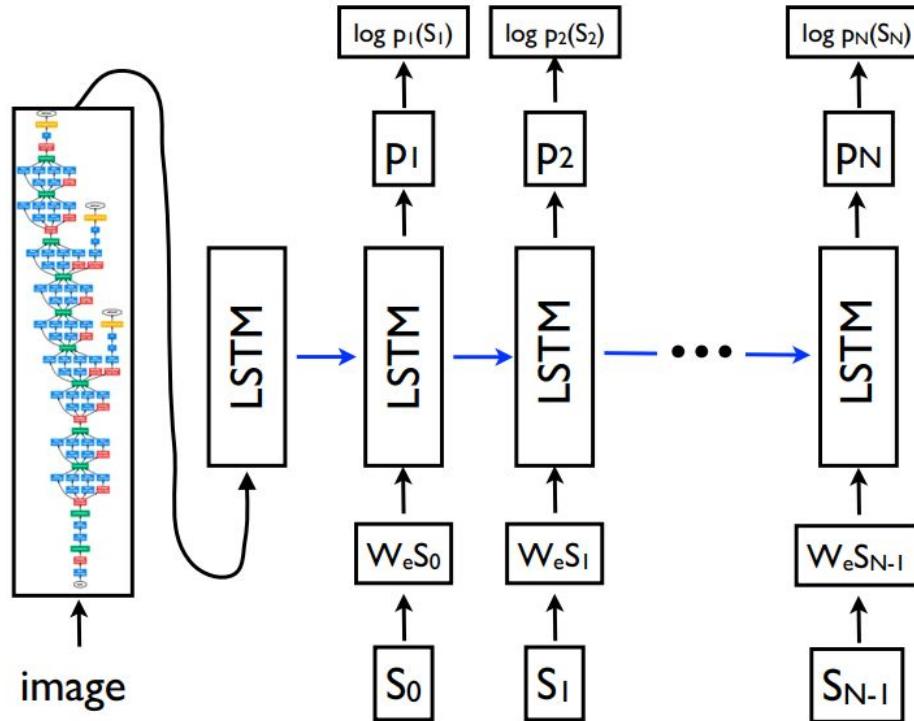
# 10. Image Caption



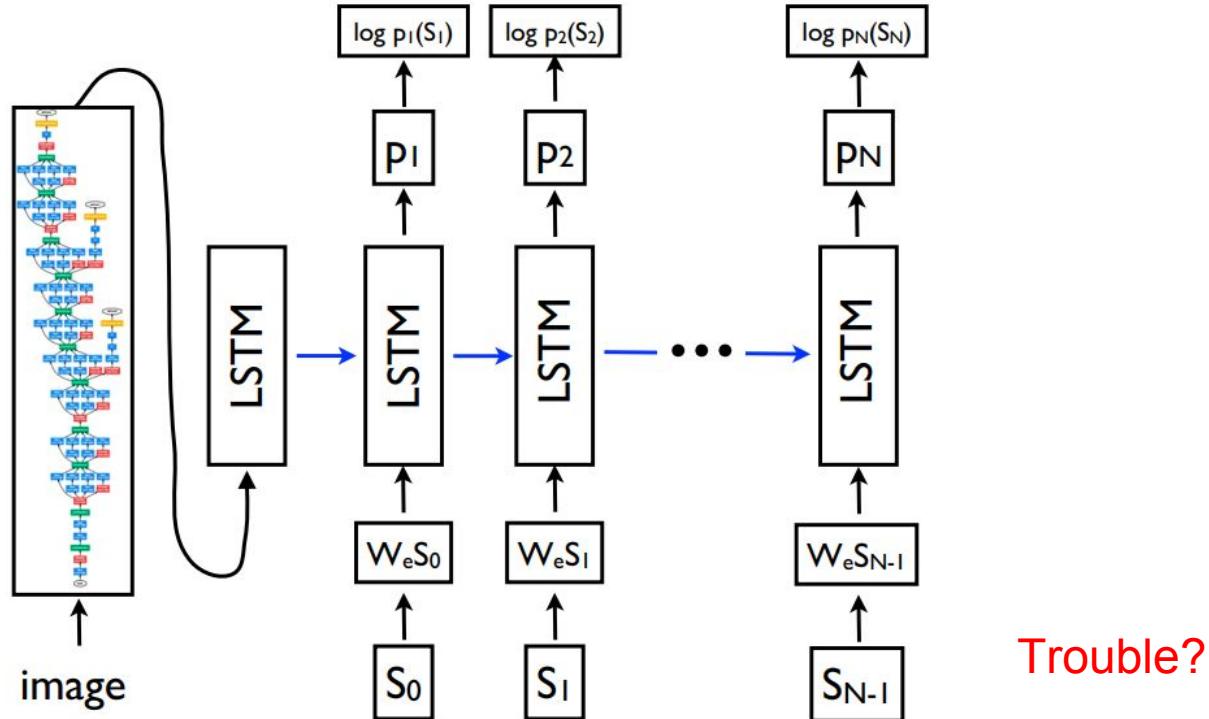
**A group of people  
shopping at an  
outdoor market.**

**There are many  
vegetables at the  
fruit stand.**

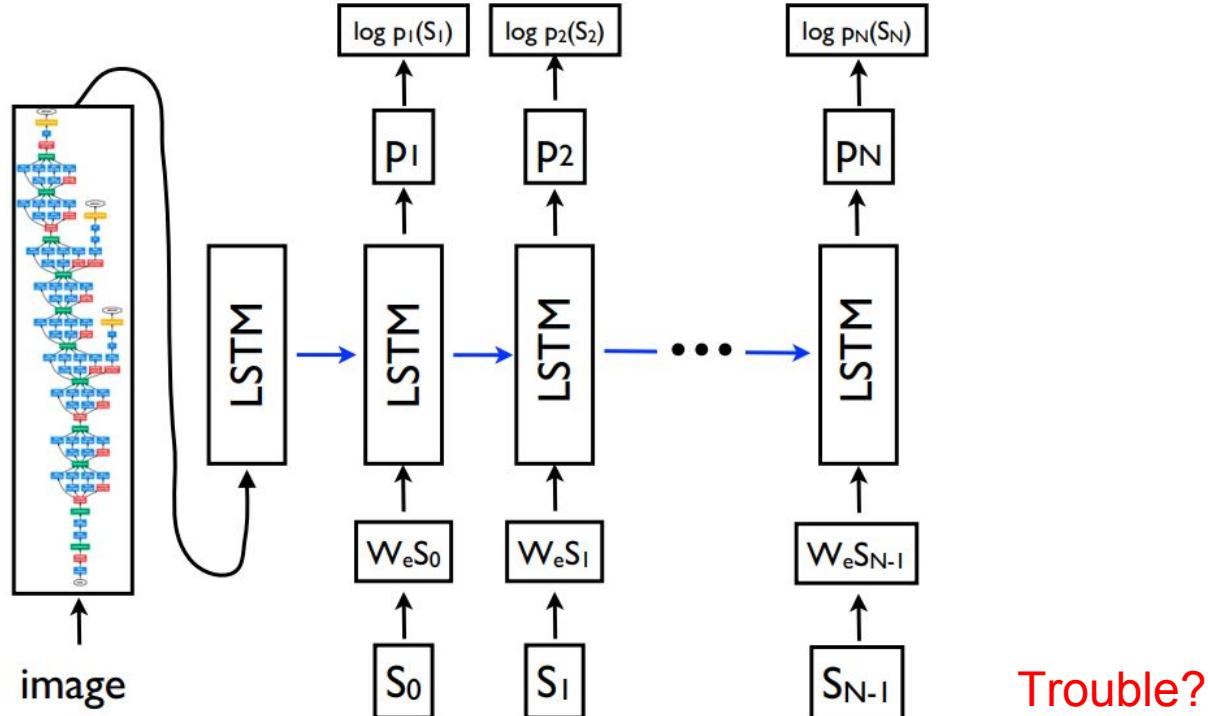
# 10. Image Caption



# 10. Image Caption



# 10. Image Caption



**Problem:** A limitation of the Encoder-Decoder architecture is that a single fixed-length representation is used to hold the extracted features.

# 10. Image Caption

## With Attention



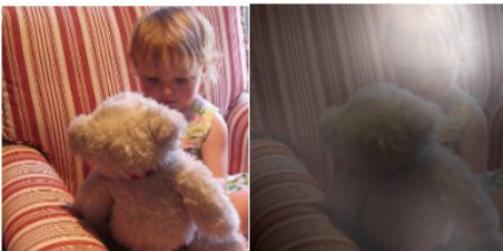
A woman is throwing a frisbee in a park.



A dog is standing on a hardwood floor.



A stop sign is on a road with a mountain in the background.



A little girl sitting on a bed with a teddy bear.



A group of people sitting on a boat in the water.



A giraffe standing in a forest with trees in the background.

Examples of attending to the correct object (white indicates the attended regions, underlines indicate the corresponding word)

# 11. Question Answering / Machine Comprehension / Chatbots

## Nikola Tesla

---

In 1870, Tesla moved to Karlovac, to **attend school at the Higher Real Gymnasium**, where he was profoundly influenced by a math teacher **Martin Sekulic**. The classes were held in **German**, as it was a school within the Austro-Hungarian Military Frontier. Tesla was able to perform integral calculus in his head, which prompted his teachers to believe that he was cheating. He finished a four-year term in three years, graduating in 1873.

---

In what language were the classes given?

**German**

Who was Tesla's main influence in Karlovac?

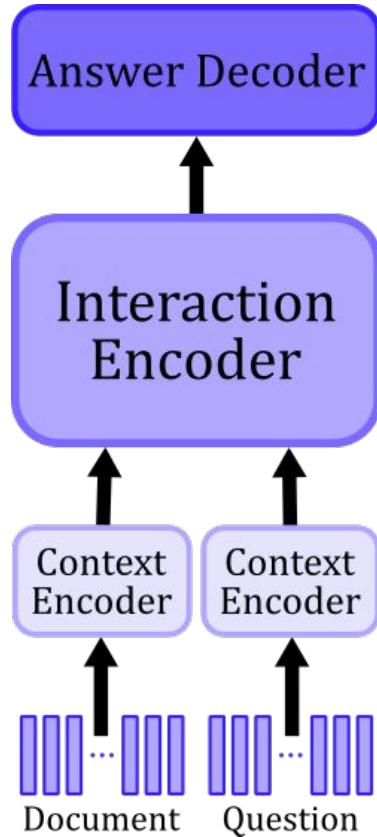
**Martin Sekulic**

Why did Tesla go to Karlovac?

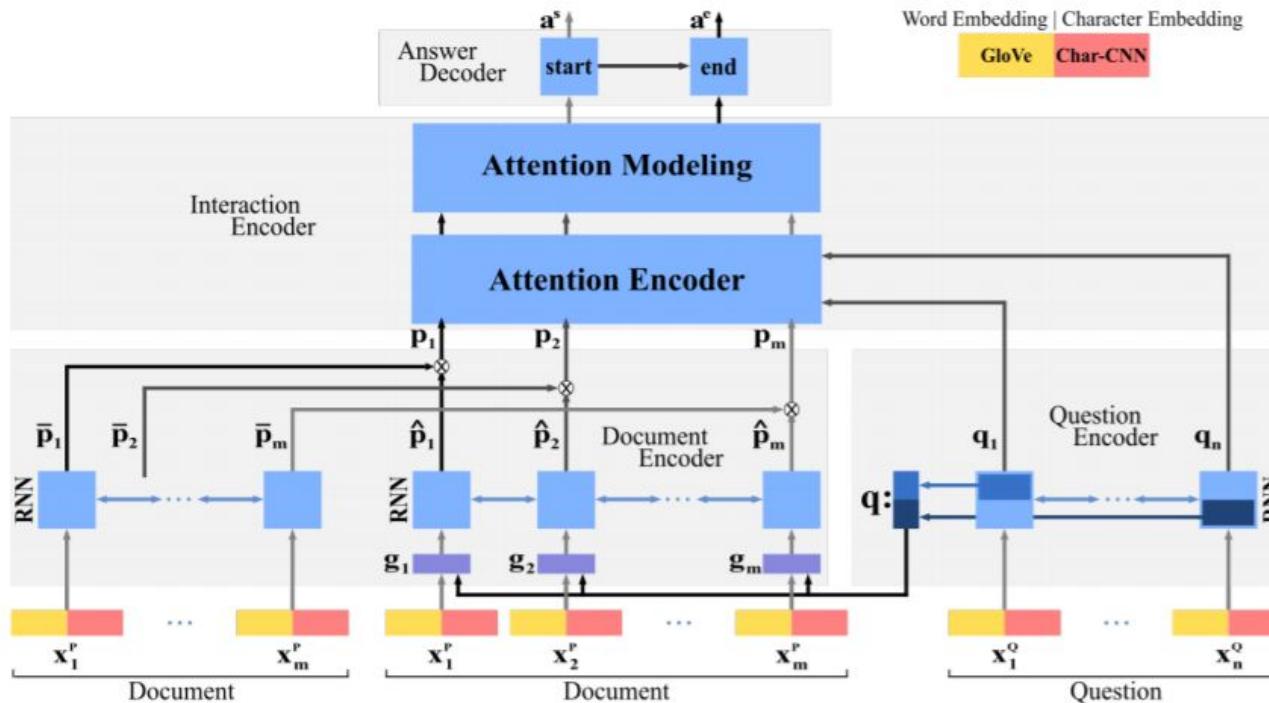
**attend school at the Higher Real Gymnasium**

Pares de Question Answering extraídos de la base Stanford Question Answering Dataset (SQuAD).

# 11. Question Answering / Machine Comprehension / Chatbots



# 11. Question Answering / Machine Comprehension / Chatbots



## 12. Visual Question Answering



What color are her eyes?

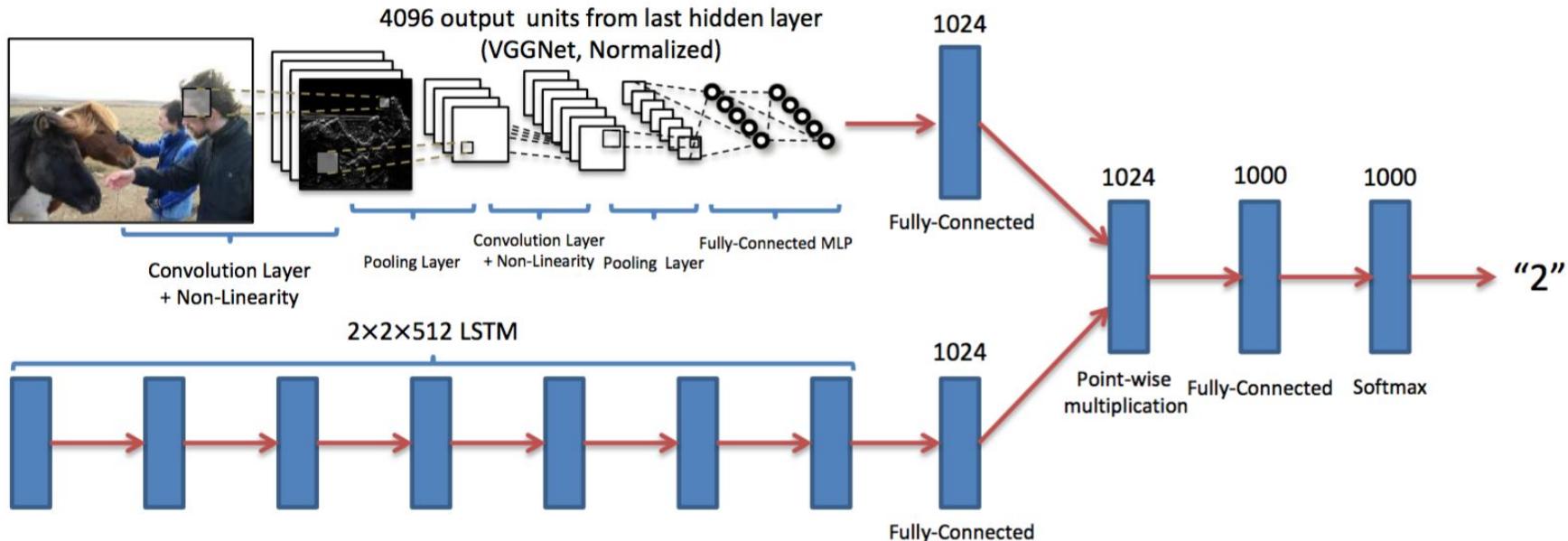
What is the mustache made of?



How many slices of pizza are there?

Is this a vegetarian pizza?

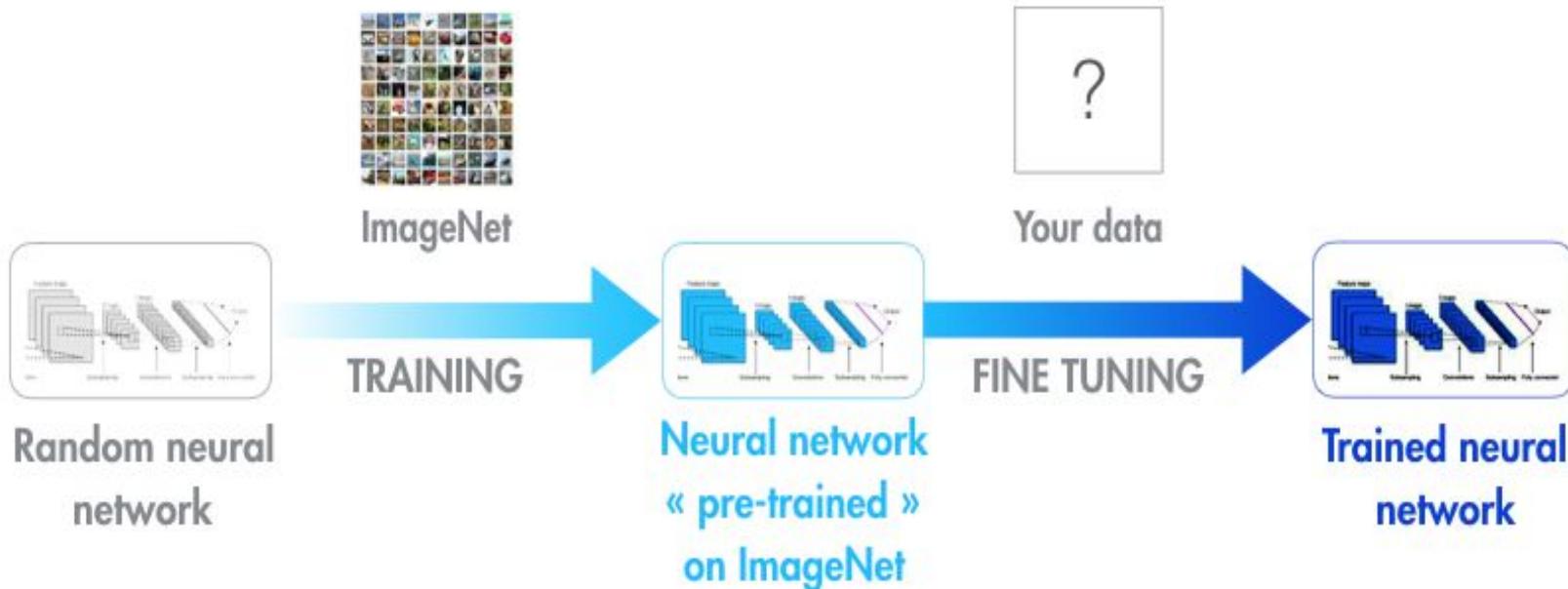
## 12. Visual Question Answering



"How many horses are in this image?"

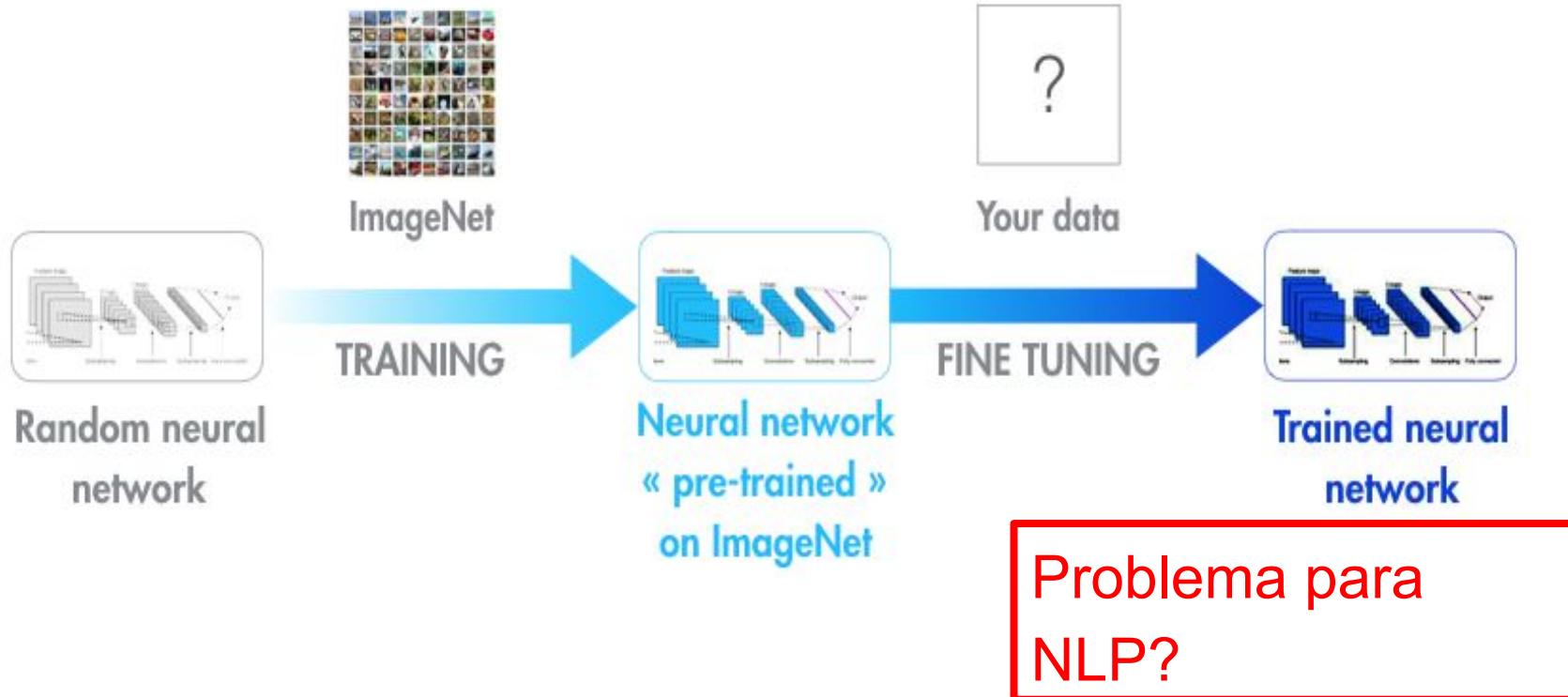
# 13. Transfer learning

In Computer Vision



# 13. Transfer learning

In Computer Vision



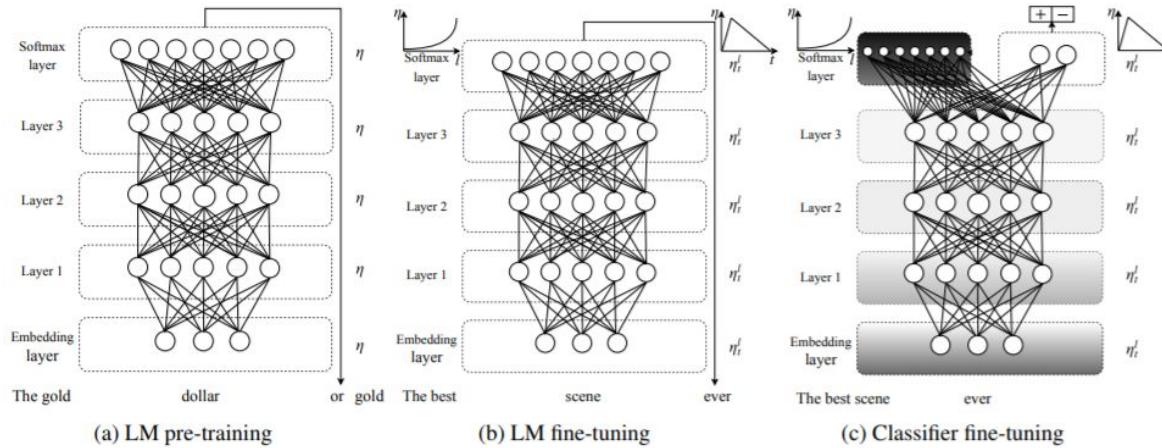
## 13. Transfer learning

### In Natural Language Processing

- No funciona bien usar el enfoque clásico de transfer learning de visión computacional.
- Cual es una arquitectura-concepto ImagetNet para NLP.

# 13. Transfer learning

## In Natural Language Processing



Howard, presenta un enfoque de *transfer learning* exitoso para problemas de clasificación, usando un *Language Modeling* como *pretrained network*. **Su propuesta se centra en una modificación inteligente y dinámica del *learning rate*, sobre ciertas capas durante el entrenamiento.**

# References

- Natural Language Processing with Deep Learning, Christopher Manning and Richard Socher, cs224n,2017.
- Word Embeddings, Christopher Olah.
- Natural Language Processing, Yves Peirsman.
- Word Embeddings, Tomas Mikolov.
- RNN, cs224d, Richard Socher, 2016.
- CS224N/Ling284, Lecture 8, Stanford University, Manning , 2016.
- Learning Long-Term Dependencies with Gradient Descent is Difficult, Bengio, 1994.
- Deep Learning, Ian Goodfellow, Yoshua Bengio, Aaron Courville, 2016.
- Advanced in Optimizing Recurrent Networks, Ian Goodfellow, Yoshua Bengio, Aaron Courville, 2016.
- Lecture note 04 in Deep Learning for Natural Language Processing, Mohammadi, 2015.
- Recurrent Neural Network Regularization, Wojciech Zarembam, Ilya Sutskever and Oriol Vinyals, 2014.
- Learning phrase representations using RNN encoder-decoder for statistical machine translation, Cho, 2014.
- Lecture note 04 in Deep Learning for Natural Language Processing, Mohammadi, 2015.

# References

- Neural Machine Translation and Sequence-to-sequence Models: A Tutorial, Graham Neubig, 2017.
- A Neural Probabilistic Language Model, Yoshua Bengio, Réjean Ducharme, Pascal Vincent, Christian Jauvin, 2003.
- Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank, Socher, 2013.
- Neural Machine Translation for Low Resource Languages, Bragagnini, 2018.
- Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation, Y. Wu, M. huster, Z. Chen, Q.V. Lec, M. Norouzi, 2016
- Convolutional Sequence to Sequence Learning, Jonas Gehring, 2017.
- Attention is All you Need, Vaswani, 2017.
- Show and Tell: A Neural Image Caption Generator, Vinyals, 2015
- Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, Xu, 2015
- Fine-Grained Gating based on Question-Summary for Machine Comprehension, Mayhua, 2018.
- VQA: Visual Question Answering, Agrawal, Lu, Antol, 2017.
- Medium, An Introduction to Transfer Learning in Machine Learning, Curry, 2018.
- Universal Language Model Fine-tuning for Text Classification, Howard, 2018.
- <http://www.iwml.iitbhu.ac.in/presentations/iWML-talk-samarth.pdf>

# I SPDL

## Deep Learning for Natural Language Processing

**César  
Bragagnini**

[cesarbrma91@gmail.com](mailto:cesarbrma91@gmail.com)  
[@MarchBragagnini](https://twitter.com/MarchBragagnini)