DECSAI
Departamento de Ciencias de la Computación e I.A.
Universidad de Granada

Softmax

Fernando Berzal, berzal@acm.org

Softmax

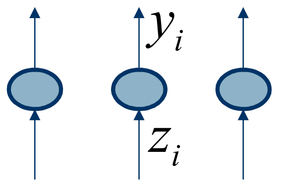
El uso del error cuadrático como medida de error tiene algunos inconvenientes:

- Si la salida deseada es 1 y la salida actual es 0.000001 el gradiente es prácticamente 0, por lo que una unidad logística difícilmente conseguirá corregir el error.

- Si estamos ante un problema de clasificación y queremos estimar la probabilidad de cada clase, sabemos que la suma de las salidas debería ser 1, pero no estamos usando esa información para entrenar la red neuronal.

Softmax

¿Existe alguna función de coste alternativa que funcione mejor? Sí, una que fuerza que las salidas de la red representen una distribución de probabilidad.



grupo softmax

$$y_i = \frac{e^{z_i}}{\sum_{j \in group} e^{z_j}}$$

$$\frac{\partial y_i}{\partial z_i} = y_i(1 - y_i)$$

Softmax

Entropía cruzada [cross-entropy]

La función de coste asociada a softmax: $C = -\sum_j t_j \log y_j$

- El gradiente de C es muy grande si el valor deseado (t) es 1 pero la salida obtenida (y) está cercana a 0.

$$\frac{\partial C}{\partial z_i} = \sum_j \frac{\partial C}{\partial y_j} \frac{\partial y_j}{\partial z_i} = y_i - t_i$$

La pendiente de δC/δy
compensa el valor bajo de δy/δz

Softmax

Una interpretación alternativa
UFLDL Tutorial, http://ufldl.stanford.edu/tutorial/

La regresión lineal consiste en encontrar una función $h_\theta(x)=\theta^T x$ en la que los parámetros θ se eligen de forma que se minimiza una función de coste $J(\theta)$:

$$J(\theta) = \frac{1}{2}\sum_i (h_\theta(x^{(i)}) - y^{(i)})^2 = \frac{1}{2}\sum_i (\theta^T x^{(i)} - y^{(i)})^2$$

Para minimizar dicha función usando el gradiente descendente, calculamos $\nabla_\theta J(\theta)$:

$$\frac{\partial J(\theta)}{\partial \theta_j} = \sum_i x_j^{(i)}(h_\theta(x^{(i)}) - y^{(i)}) = \sum_i x_j^{(i)}(\theta^T x^{(i)} - y^{(i)})$$

Softmax

Una interpretación alternativa
UFLDL Tutorial, http://ufldl.stanford.edu/tutorial/

De la misma forma, podemos predecir una variable discreta utilizando regresión logística:

$$P(y=1\,|\,x) = h_\theta(x) = \sigma(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$$

$$P(y=0\,|\,x) = 1 - P(y=1\,|\,x) = 1 - h_\theta(x)$$

donde $\sigma(z)$ es la función logística.

Softmax

Una interpretación alternativa
UFLDL Tutorial, http://ufldl.stanford.edu/tutorial/

Nuestro objetivo en regresión logística es buscar valores para θ de forma que $\mathbf{h_\theta(x)}$ sea grande cuando $\mathbf{x}$ pertenece a la clase 1 y pequeño si pertenece a la clase 0.

La siguiente función de coste nos sirve:

$$J(\theta) = -\sum_i \left(y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)})(1 - \log h_\theta(x^{(i)})) \right)$$

NOTA: Sólo uno de los dos términos es distinto de cero para cada ejemplo (según sea de una clase u otra).

Softmax

Una interpretación alternativa
UFLDL Tutorial, http://ufldl.stanford.edu/tutorial/

¿De dónde sale esa función de coste?

Si asumimos que los ejemplos del conjunto de entrenamiento se generaron de forma independiente, la función de verosimilitud [likelihood] de los parámetros es

$$\begin{aligned} L(\theta) &= p(\vec{y} \mid X; \theta) \\ &= \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}; \theta) \\ &= \prod_{i=1}^{m} \left(h_\theta(x^{(i)})\right)^{y^{(i)}} \left(1 - h_\theta(x^{(i)})\right)^{1-y^{(i)}} \end{aligned}$$

Softmax

Una interpretación alternativa
UFLDL Tutorial, http://ufldl.stanford.edu/tutorial/

¿De dónde sale esa función de coste?

Dicha función de verosimilitud [likelihood] resulta más fácil de maximizar tomando logaritmos [log-likelihood]:

$$\begin{aligned} \ell(\theta) &= \log L(\theta) \\ &= \sum_{i=1}^{m} y^{(i)} \log h(x^{(i)}) + (1 - y^{(i)}) \log(1 - h(x^{(i)})) \end{aligned}$$

Softmax

Una interpretación alternativa
UFLDL Tutorial, http://ufldl.stanford.edu/tutorial/

Para maximizar el log-likelihood, minimizamos la función de coste **$J(\theta)=-\log L(\theta)$**, para lo que calculamos **$\nabla_\theta J(\theta)$**:

$$\frac{\partial J(\theta)}{\partial \theta_j} = \sum_i x_j^{(i)} (h_\theta(x^{(i)}) - y^{(i)})$$

O, en forma vectorial:

$$\nabla_\theta J(\theta) = \sum_i x^{(i)} (h_\theta(x^{(i)}) - y^{(i)})$$

Softmax

La regresión softmax (o regresión logística multinomial) no es más que una generalización de la regresión logística cuando tenemos más de dos clases distintas.

Si tenemos K clases,
tendremos K vectores de parámetros θ_k:

$$h_\theta(x) = \begin{bmatrix} P(y=1 \mid x) \\ P(y=2 \mid x) \\ \vdots \\ P(y=K \mid x) \end{bmatrix} = \frac{1}{\sum_{j=1}^{K} e^{\theta_j^T x}} \begin{bmatrix} e^{\theta_1^T x} \\ e^{\theta_2^T x} \\ \vdots \\ e^{\theta_K^T x} \end{bmatrix}$$

Softmax

La función de coste asociada a la regresión logística la podríamos reescribir como:

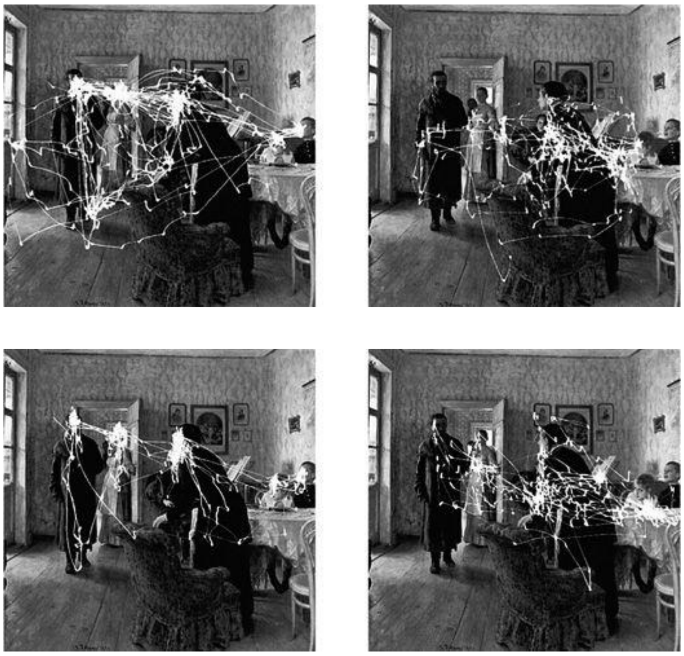
$$J(\theta) = -\sum_i \left(y_i \log h_\theta(x_i) + (1-y_i)(1-\log h_\theta(x_i)) \right)$$
$$= -\sum_i t_i \log P(t_i \mid x_i)$$

Extendiendo esta función de coste a la regresión softmax, la función de coste resultante es, como esperamos, la entropía cruzada:

$$C = J(\theta) = -\sum_i t_i \log P(t_i \mid x_i) = -\sum_j t_j \log y_j$$

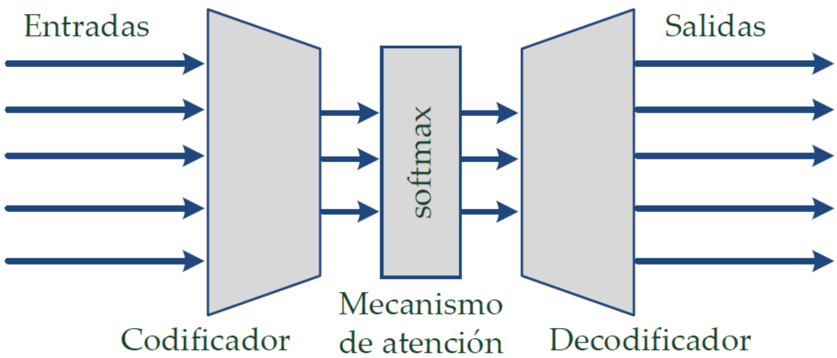
Mecanismos de atención

Atención visual



12

Mecanismos de atención



Arquitectura de red codificador-decodificador

13

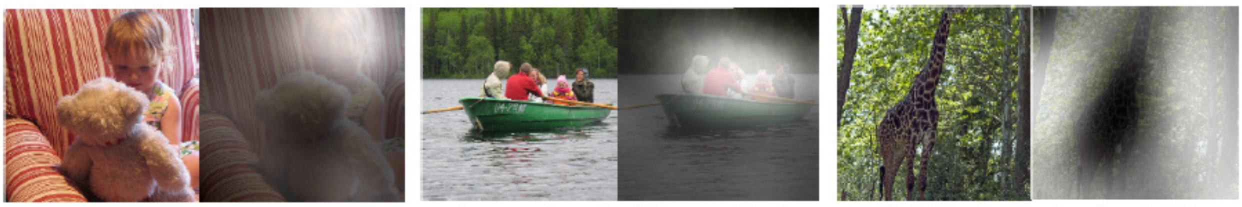
Mecanismos de atención

Descripción textual de imágenes [image captioning]

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.



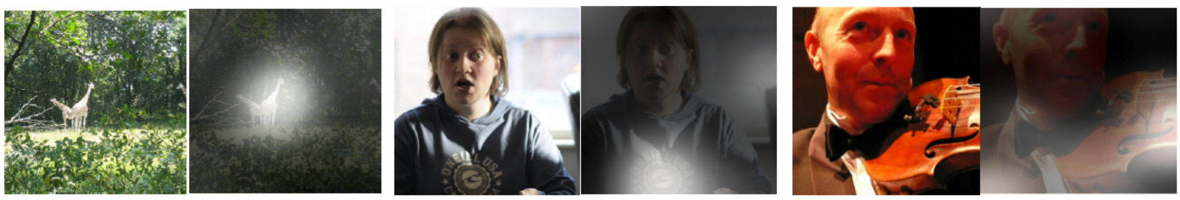
A little girl sitting on a bed with a teddy bear.

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

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

Mecanismos de atención

Descripción textual de imágenes [image captioning]



A large white bird standing in a forest.

A woman holding a clock in her hand.

A man wearing a hat and a hat on a skateboard.

A person is standing on a beach with a surfboard.

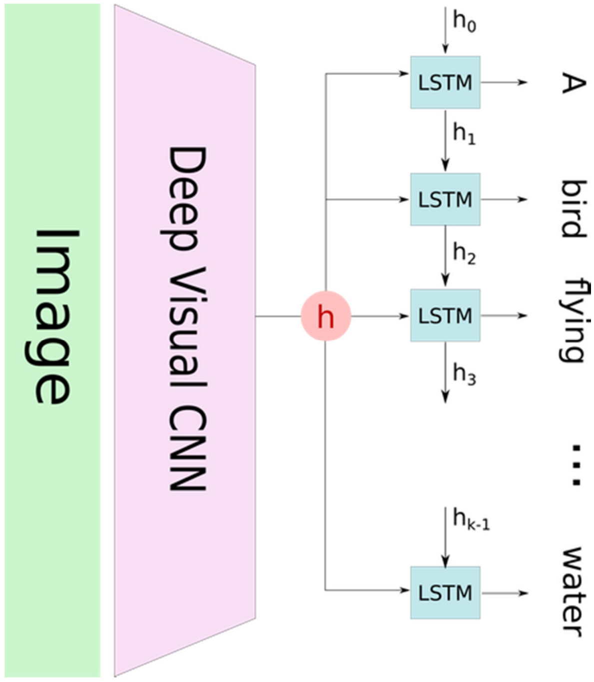
A woman is sitting at a table with a large pizza.

A man is talking on his cell phone while another man watches.

No siempre funciona bien ;-)

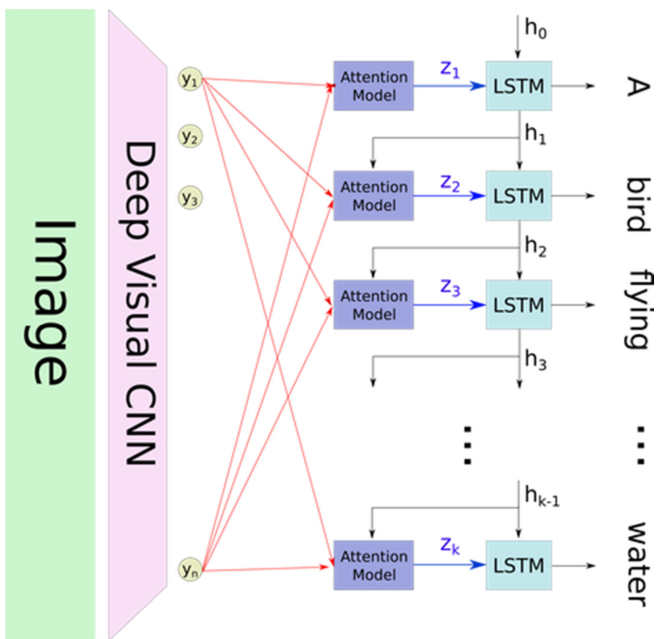
Mecanismos de atención

Descripción textual de imágenes [image captioning]



Mecanismos de atención

Descripción textual de imágenes [image captioning]

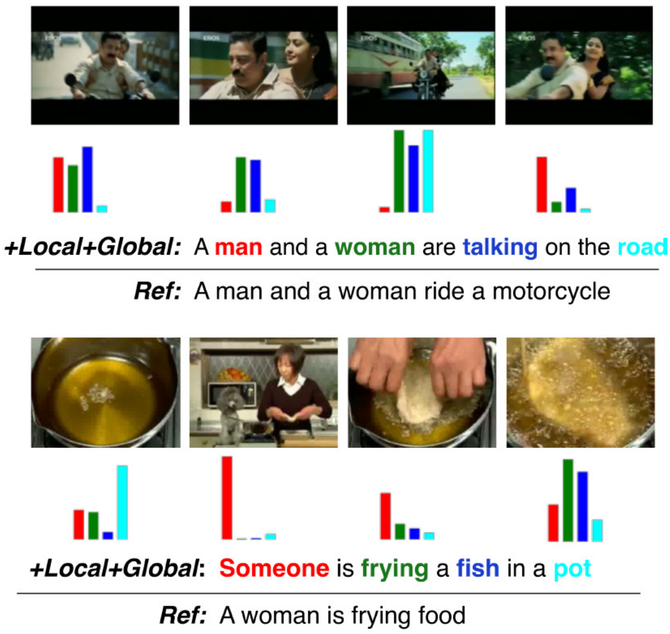


Mecanismos de atención

Descripción textual de imágenes [image captioning]

Mecanismos de atención

Descripción textual de vídeos [video clip description]

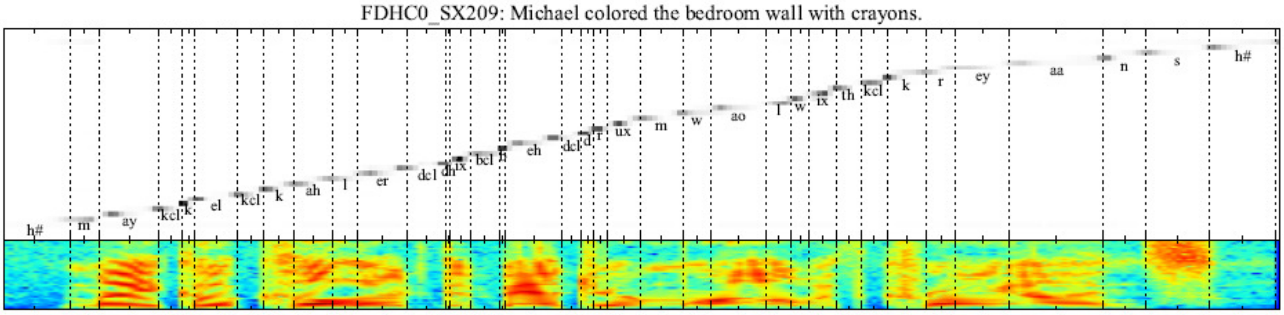


Youtube2Text

Mecanismos de atención

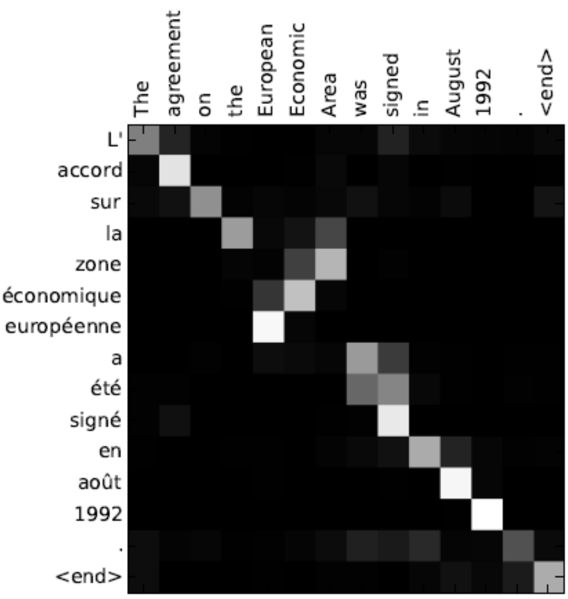
Reconocimiento de voz [speech recognition]

FDHC0_SX209: Michael colored the bedroom wall with crayons.



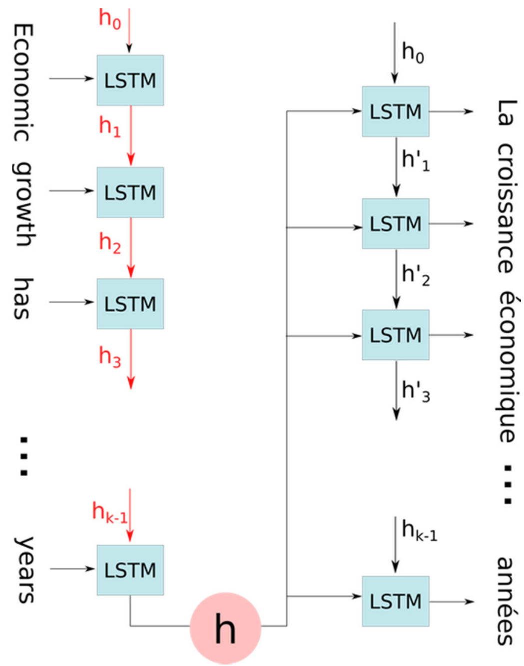
Mecanismos de atención

Traducción automática [machine translation]



Mecanismos de atención

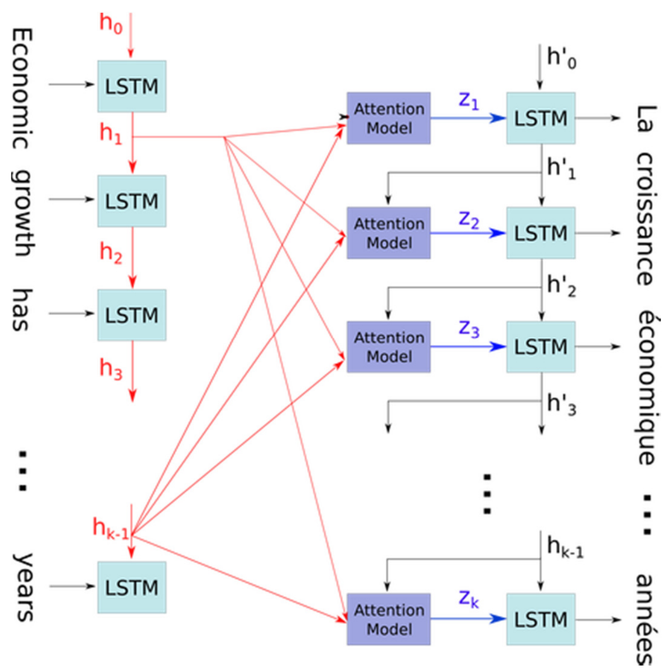
Traducción automática [machine translation]



Sin mecanismo de atención

Mecanismos de atención

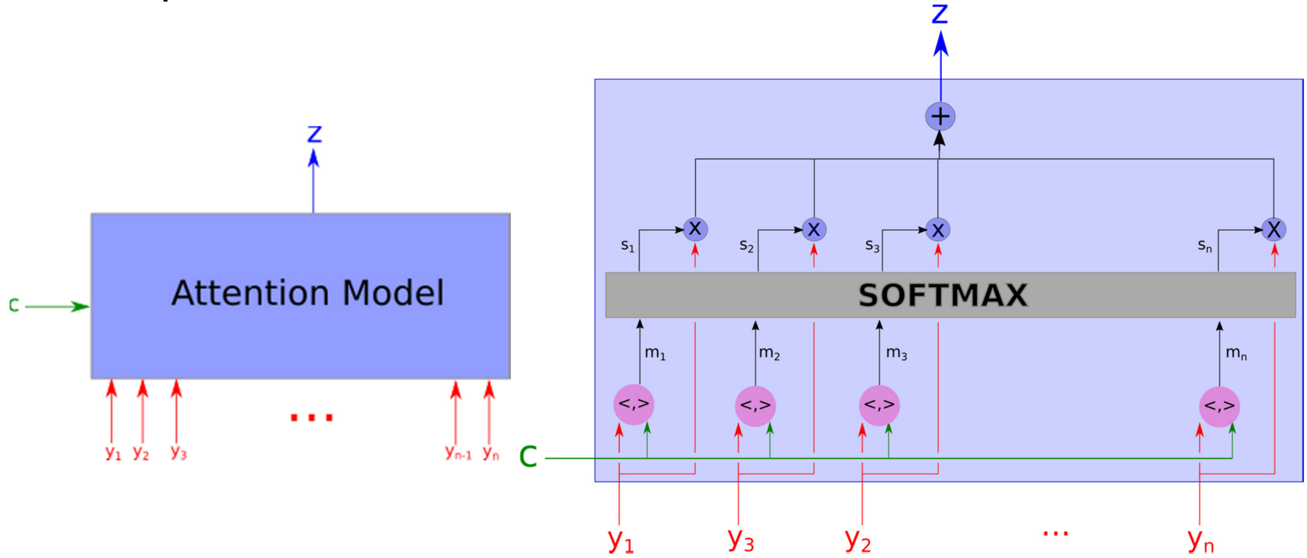
Traducción automática [machine translation]



Con mecanismo de atención

Mecanismos de atención

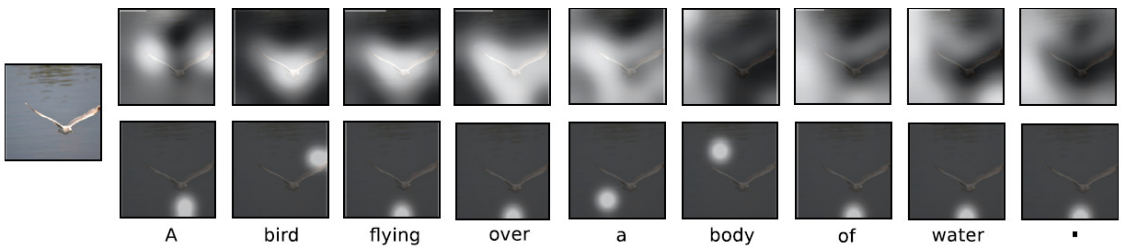
Implementación del mecanismo de atención con softmax

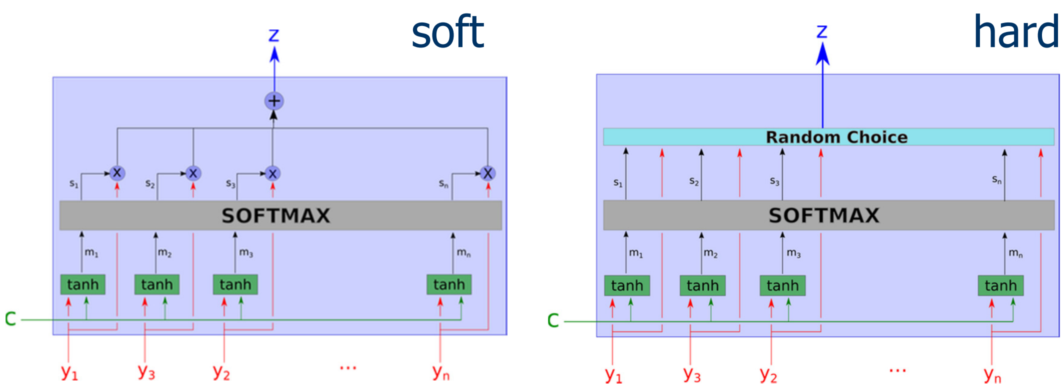


https://blog.heuritech.com/2016/01/20/attention-mechanism/

Mecanismos de atención

Soft vs. Hard attention

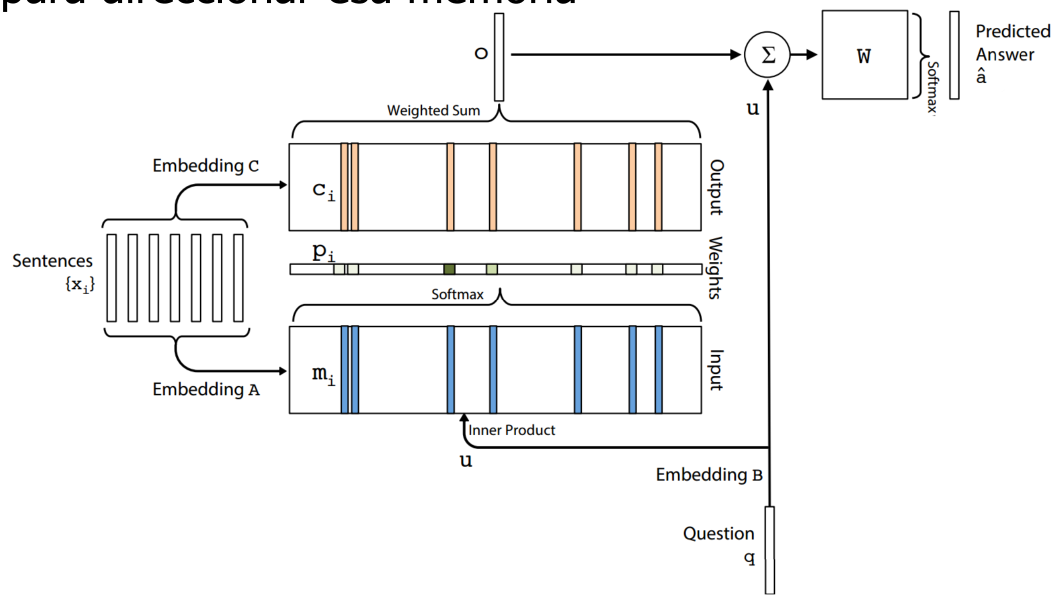




Mecanismos de atención

Redes con memoria (externa):

El mecanismo de atención sirve para direccionar esa memoria



Cursos

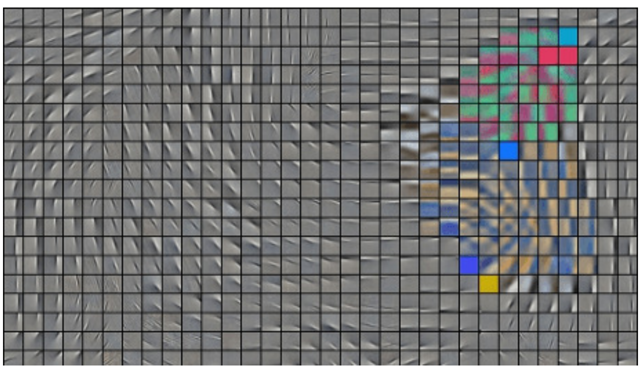
Neural Networks for Machine Learning

by Geoffrey Hinton

(University of Toronto & Google)

https://www.coursera.org/course/neuralnets

coursera



UNIVERSITY OF TORONTO

Cursos

Deep Learning Specialization
by Andrew Ng, 2017

- Neural Networks and Deep Learning
- Improving Deep Neural Networks: Hyperparameter tuning, Regularization and Optimization
- Structuring Machine Learning Projects
- Convolutional Neural Networks
- Sequence Models

coursera

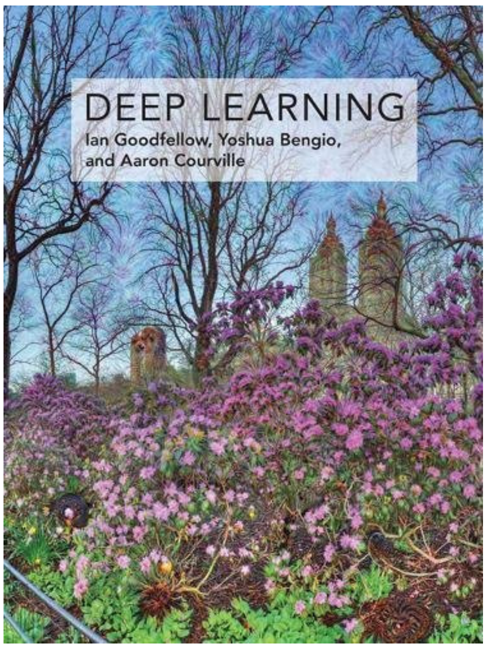
deeplearning.ai

https://www.coursera.org/specializations/deep-learning

Bibliografía

Lecturas recomendadas

Ian Goodfellow,
Yoshua Bengio
& Aaron Courville:
Deep Learning
MIT Press, 2016
ISBN 0262035618



http://www.deeplearningbook.org

Bibliografía

Lecturas recomendadas

Fernando Berzal:
Redes Neuronales & Deep Learning

CAPÍTULO 12
Redes softmax