

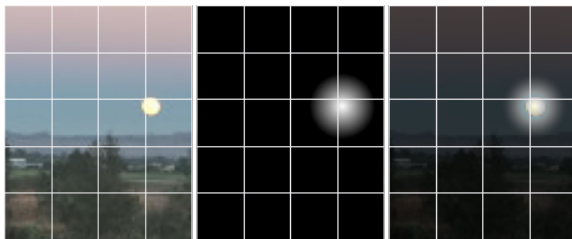


## Definición

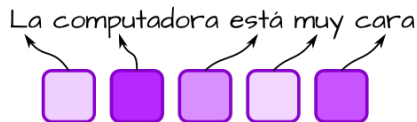
- 1 Definición
- 2 Cuatro aplicaciones

# Atención

- El módulo de atención se representa como un vector de las mismas dimensiones que el objeto sobre el cual se desea seleccionar un área de atención.
- Se utiliza realizando una multiplicación punto a punto  $\odot$ .



(a) Para fijar la atención sobre la luna, la matriz de atención asigna una mayor valor a los pixeles correspondientes y anula los demás.



(b) Para texto el vector es lineal.

## Cuatro aplicaciones

- 1 Definición
- 2 Cuatro aplicaciones

# Neuronal vs neural

## Dato curioso:

*“**Neural** significa que pertenece a un nervio o nervios (los paquetes de fibras semejantes a cuerdas hechos de neuronas), mientras que **neuronal** significa que se refiere a las neuronas (las células conductoras del sistema nervioso).”*

*Cade Hildreth*

[https://bioinformant.com/neural-vs-neuronal/#::~:~:text=The%20short%20answer%20is%20that,cells%20of%20the%20nervous%20system\).](https://bioinformant.com/neural-vs-neuronal/#::~:~:text=The%20short%20answer%20is%20that,cells%20of%20the%20nervous%20system).)

# Temas

## 2 Cuatro aplicaciones

- Máquinas de Turing neurales
- Interfaces con atención
- Tiempo de computación adaptativo
- Programador neural

# Máquinas de Turing neurales

## Lectura:

$$r \leftarrow \sum_i a_i M_i$$

## Escritura:

$$M_i \leftarrow a_i w + (1 - a_i) M_i$$

Olah y Carter 2016

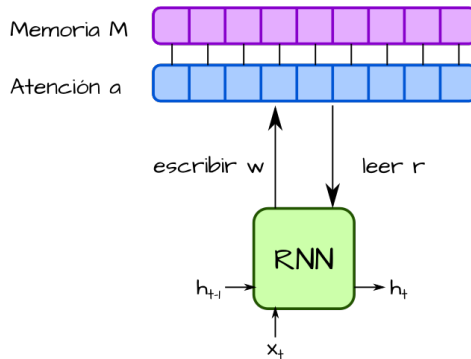
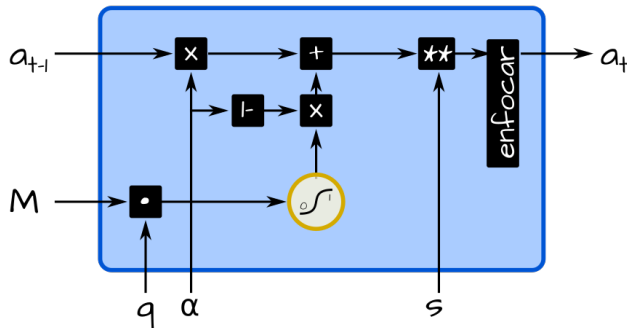


Figura: En cada paso temporal la red escribe y lee de toda la memoria.

# Módulo de atención

Dadas una consulta  $q$ , una constante de interpolación  $\alpha$  y un filtro de corrimiento  $s$  emitidas por la red recurrente:



**Figura:** Cálculo de la atención para una búsqueda al paso  $t$ . Utiliza softmax para determinar el nivel de atención sobre cada componente.



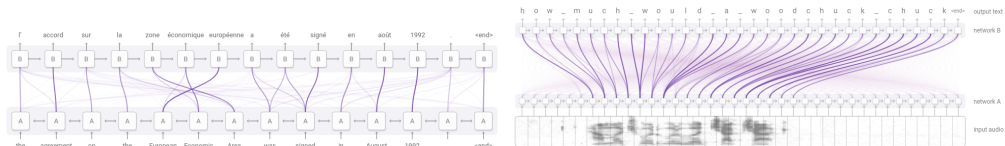
# Temas

## 2 Cuatro aplicaciones

- Máquinas de Turing neurales
- Interfaces con atención
- Tiempo de computación adaptativo
- Programador neural

# Interfaces con atención

- La RNN decodificadora puede analizar las salidas para todo tiempo de la RNN codificadora.
- El vector de atención tiene la longitud de la secuencia más larga.
- Para cada símbolo de entrada calcula la distribución de atención.



**Figura:** Uso de atención para traducción entre idiomas y procesamiento de audio. Ver animaciones en sitio de Olah y Carter 2016. Ejemplo de código en

[https://pytorch.org/tutorials/intermediate/seq2seq\\_translation\\_tutorial.html](https://pytorch.org/tutorials/intermediate/seq2seq_translation_tutorial.html)

# Temas

- 2 Cuatro aplicaciones
  - Máquinas de Turing neurales
  - Interfaces con atención
  - Tiempo de computación adaptativo
  - Programador neural

# Tiempo de computación adaptativo

- Su filosofía es que pasos más complejos en un análisis requieren más tiempo de cómputo.
- Utiliza un módulo de atención para determinar durante cuántos pasos procesar una misma entrada, la correspondiente al tiempo  $t$ .
- Una neurona tipo sigmoide tiene la función de terminar la probabilidad de detenerse en el paso actual.
  - 1 Este valor se almacena como un peso  $w_i$ .
  - 2 Los valores que calcula se restan de un *presupuesto* total de 1.
  - 3 Cuando el presupuesto está por debajo de  $\epsilon$  termina el ciclo y se añade el sobrante al último peso.
- La salida para el tiempo  $t$  es la combinación pesada de las salidas al final de cada paso.

Olah y Carter 2016

# Diagrama resumen

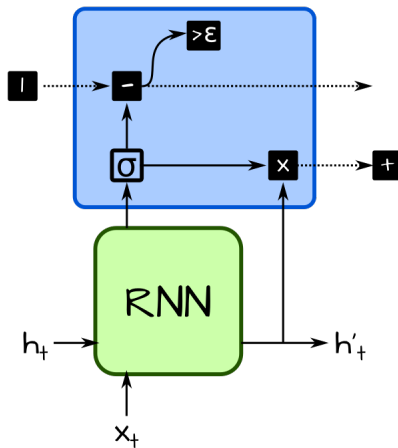


Figura: Un paso en un cómputo detallado con atención adaptativa.

# Temas


## 2 Cuatro aplicaciones

- Máquinas de Turing neurales
- Interfaces con atención
- Tiempo de computación adaptativo
- Programador neural

# Programador neural

[Tiene su propia presentación]

# Referencias I

 Olah, Chris y Shan Carter (sep. de 2016). *Attention and Augmented Recurrent Neural Networks*. Google Brain. DOI: 10.23915/distill.00001. URL: <https://distill.pub/2016/augmented-rnns/>.



# Licencia

Creative Commons  
Atribución-No Comercial-Compartir Igual

