Aprendizaje profundo

Mecanismos de atención y memoria externa

Gibran Fuentes-Pineda

Diciembre 2020/Enero 2021

Modelos secuencia a secuencia (seq2seq)

- Necesitan codificar todo el contexto de la entrada en un sólo vector

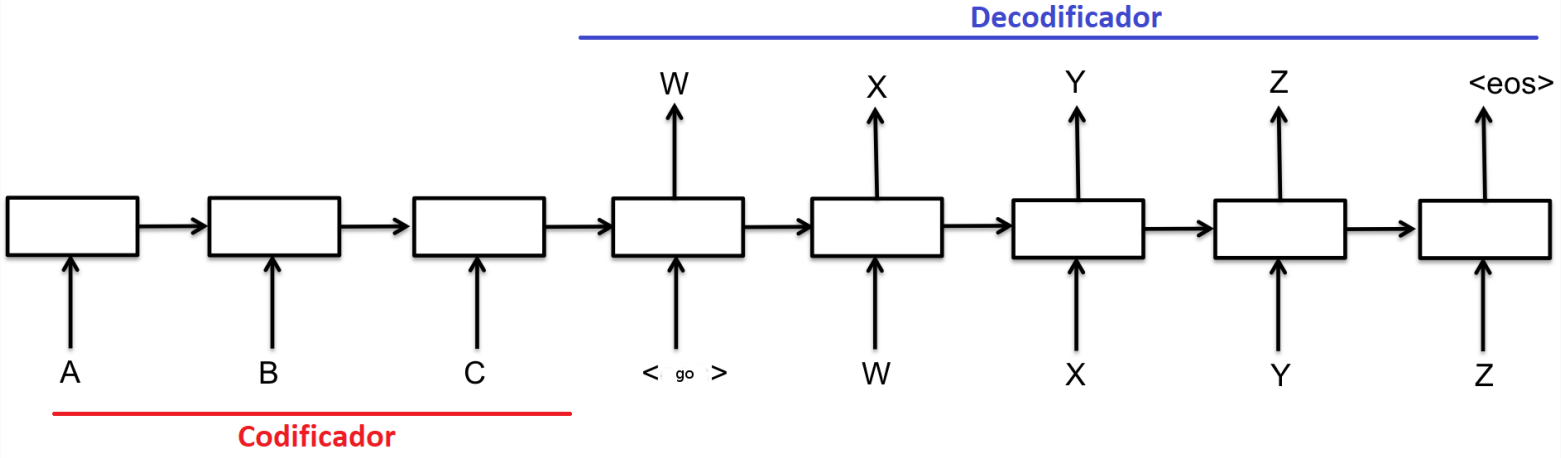

Imagen derivada de `https://www.tensorflow.org/tutorials/seq2seq`

Ejemplo de seq2seq para traducción

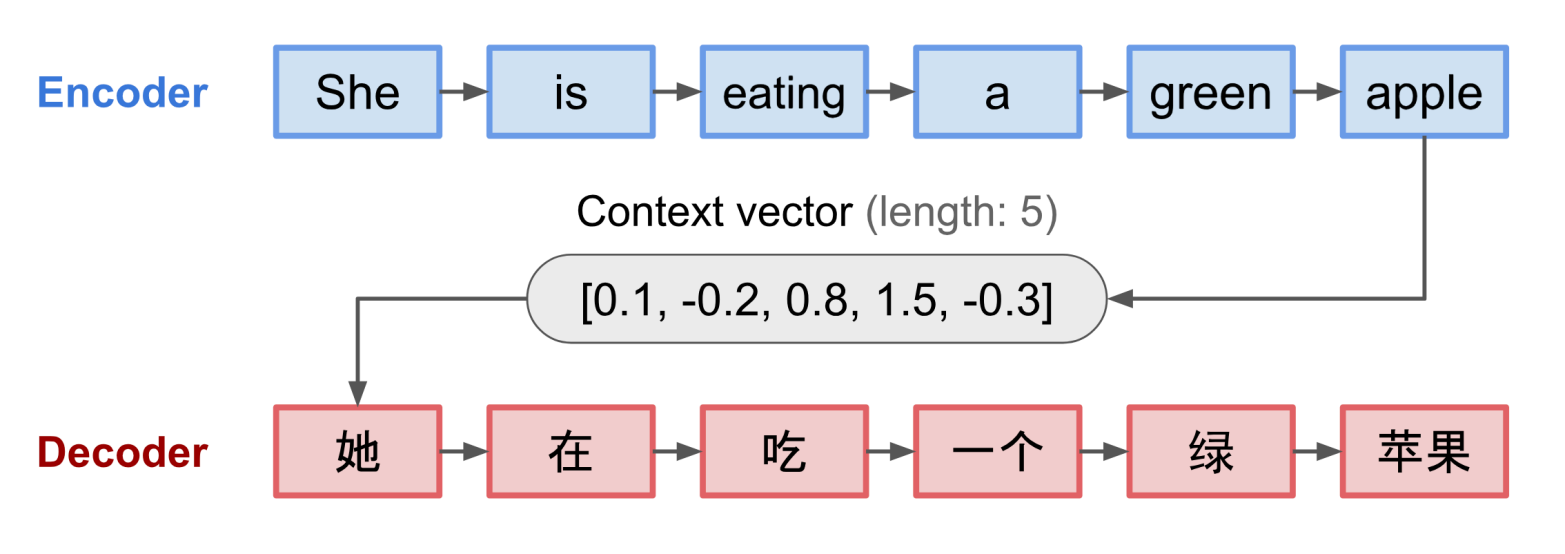

Imagen tomada de `https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html`

Mecanismos de atención

- Información relevante de la entrada se codifica en un solo vector de contexto (último estado de una red recurrente)
  - Difícil en secuencias largas

- Mecanismos de atención: se calcula un vector de contexto distinto por cada paso del decodificador a partir de
  1. Estados del decodificador
  2. Estados del codificador
  3. Función de alineación

Esquema general de los mecanismos de atención

- En cada paso t del decodificador calcular un vector de contexto a partir de todos los estados del codificador

$$\begin{aligned}
\mathbf{c}^{[t]} &= \sum_{i=1}^{T} \alpha_{t,i} \cdot \hat{\mathbf{h}}^{[i]} \\
\alpha_{t,i} &= \text{alineación}(\mathbf{h}^{[t]}, \hat{\mathbf{h}}^{[i]}) \\
&= \mathit{softmax}(\mathit{puntaje}(\mathbf{h}^{[t]}, \hat{\mathbf{h}^{[i]}})) \\
&= \frac{\mathit{puntaje}(\mathbf{h}^{[t]}, \hat{\mathbf{h}^{[i]}})}{\sum_k \mathit{puntaje}(\mathbf{h}^{[t]}, \hat{\mathbf{h}}^{[k]}}))
\end{aligned}$$

- donde *puntaje* es una función definida para 2 estados, $\mathbf{h}^{[t]}$ es el estado actual del decodificador y $\hat{\mathbf{h}}^{[i]}$ es el estado del codificador en el paso i

Funciones puntaje

- Basadas en contenido

$$\text{puntaje}(\mathbf{h}^{[t]}, \hat{\mathbf{h}}^{[i]}) = \mathbf{h}^{[t]\top}\hat{\mathbf{h}}^{[i]} \text{ (producto punto)}$$

$$\text{puntaje}(\mathbf{h}^{[t]}, \hat{\mathbf{h}}^{[i]}) = \frac{\mathbf{h}^{[t]\top}\hat{\mathbf{h}}^{[i]}}{\sqrt{T}} \text{ (producto punto escalado)}$$

$$\text{puntaje}(\mathbf{h}^{[t]}, \hat{\mathbf{h}}^{[i]}) = \mathbf{h}^{[t]\top}\mathbf{W}_a\hat{\mathbf{h}}^{[i]} \text{ (general)}$$

$$\text{puntaje}(\mathbf{h}^{[t]}, \hat{\mathbf{h}}^{[i]}) = \mathbf{W}_a \left[\mathbf{h}^{[t]}; \hat{\mathbf{h}}^{[i]}\right] \text{ (concatenación)}$$

- Basada en ubicación

$$\alpha_{t,i} = \mathbf{W}_a\mathbf{h}^{[t]}$$

Atención de Bahdanau

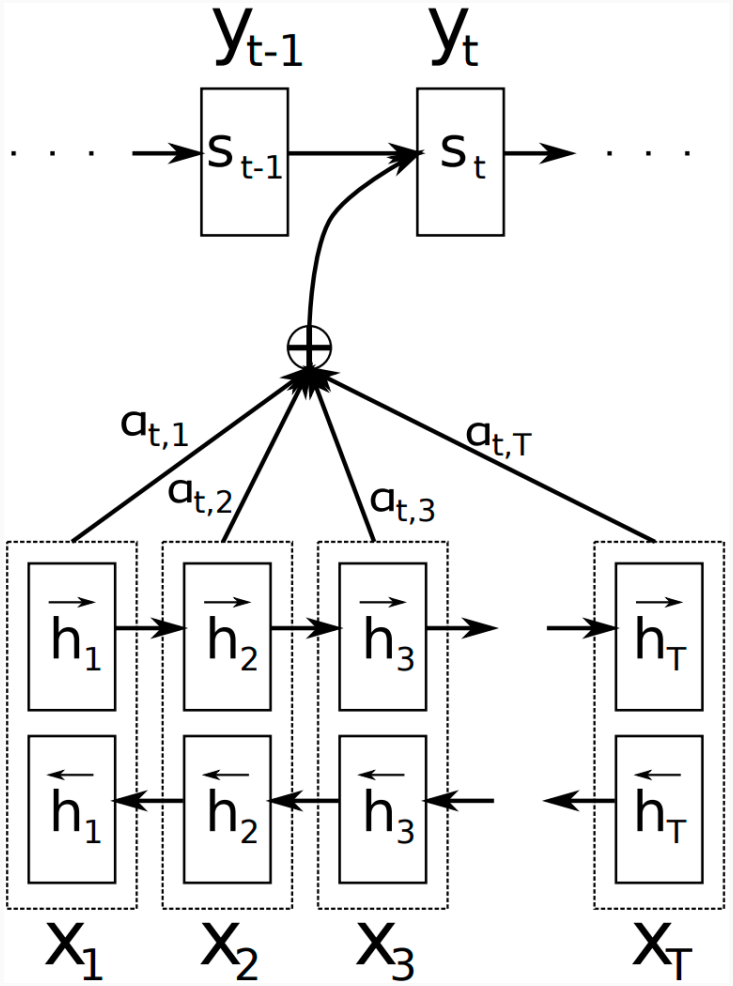

Imagen tomada de Bahdanau et al. *Neural Machine Translation by Jointly Learning to align and Translate, arXiv:1409.0473, 2014*

Atención de Luong (global)

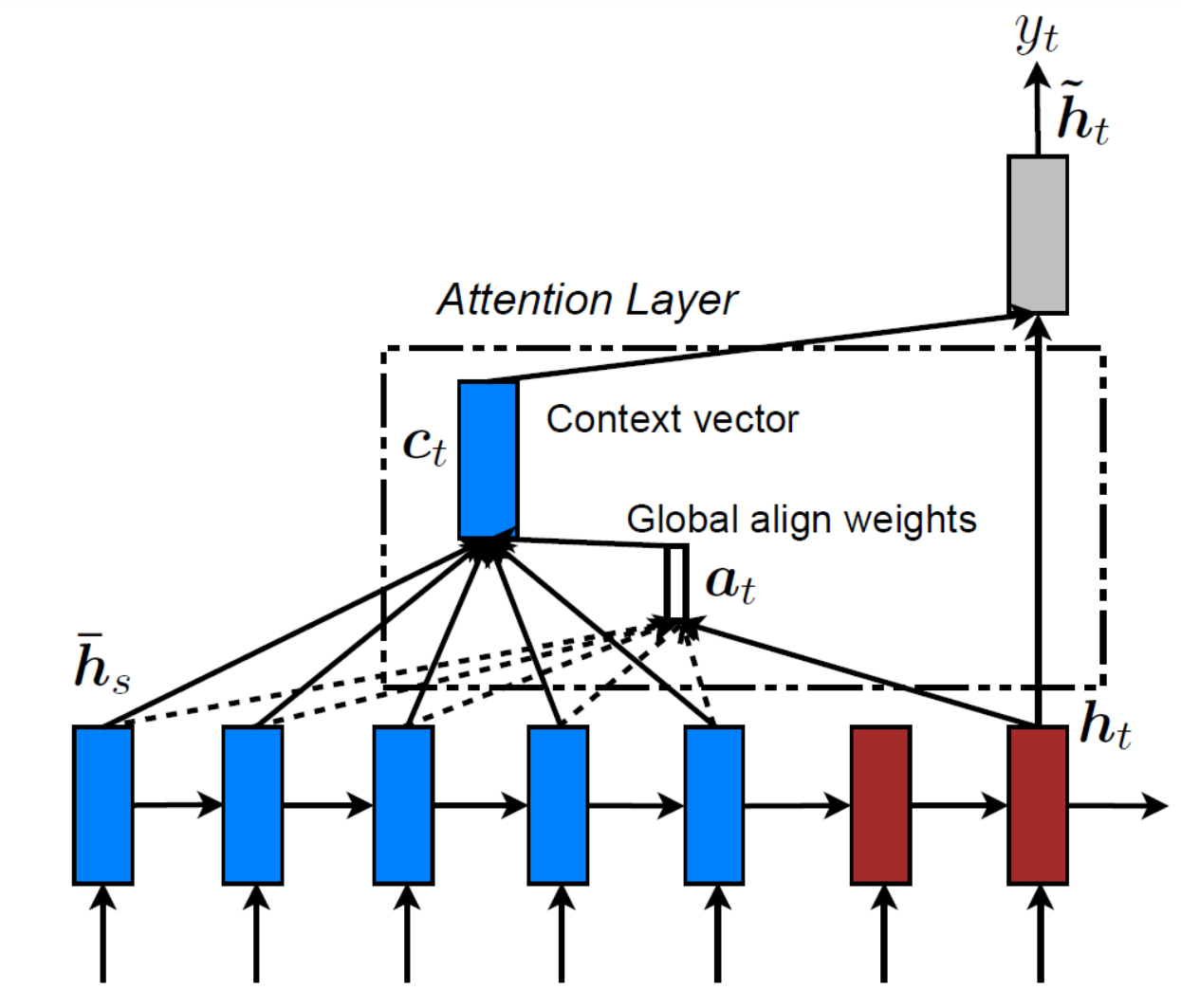

Imagen tomada de Luong et al. *Effective Approaches to Attention-based Neural Machine Translation*, EMNLP, 2015

Atención de Luong (local)

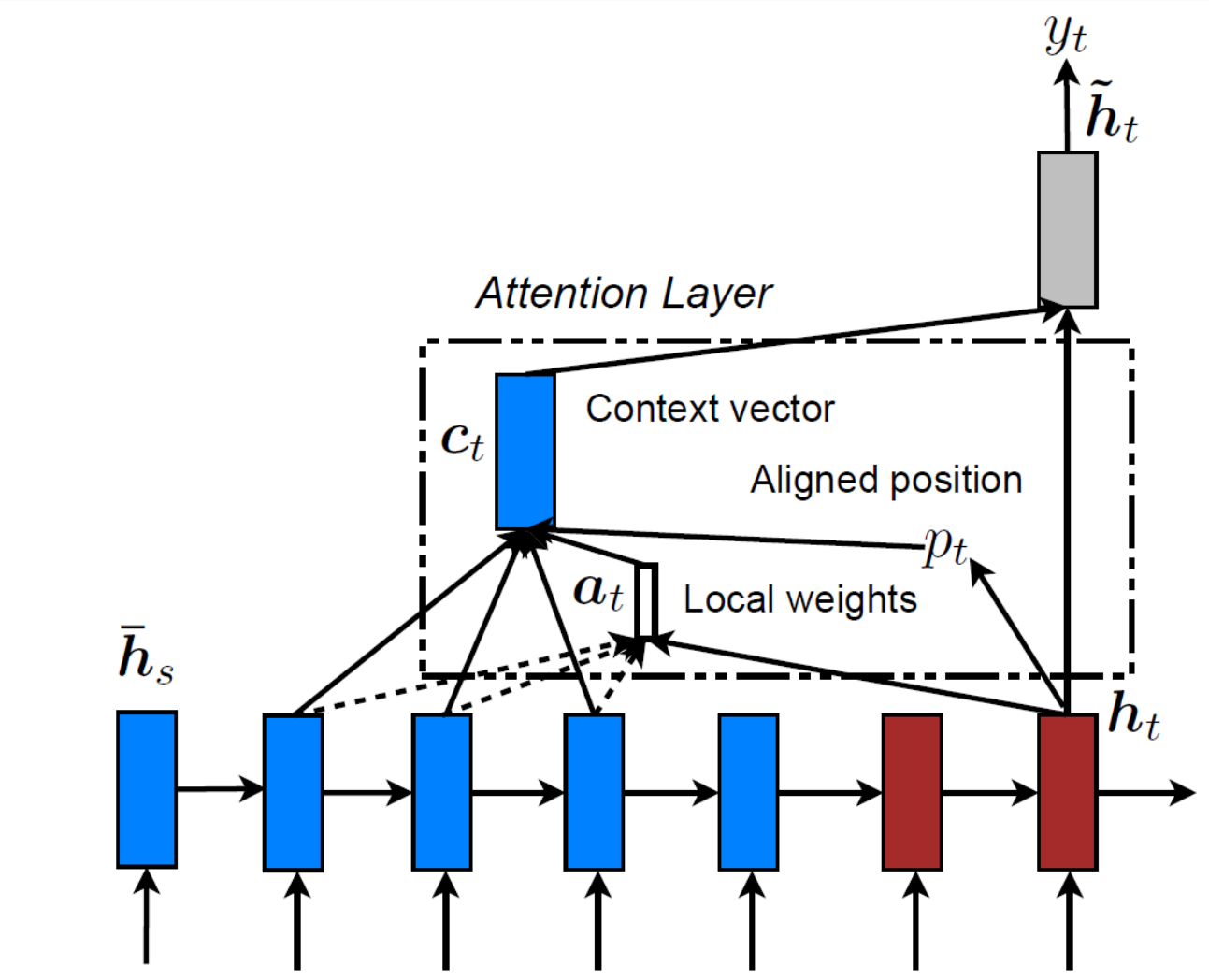

Imagen tomada de Luong et al. *Effective Approaches to Attention-based Neural Machine Translation*, EMNLP, 2015

Atención con Luong (alimentación de entradas)

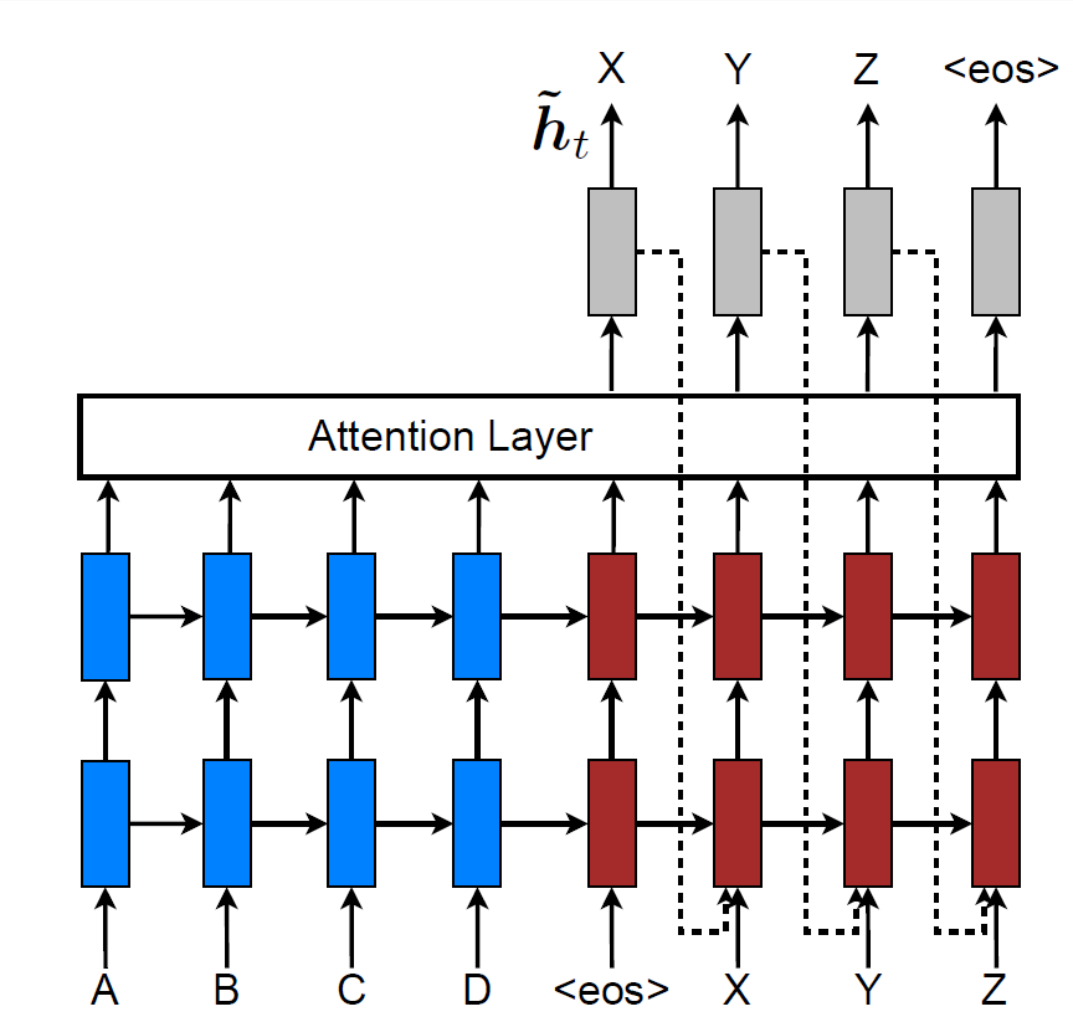

Imagen tomada de Luong et al. *Effective Approaches to Attention-based Neural Machine Translation*, EMNLP, 2015

Aprendizaje con mecanismos de atención en traducción

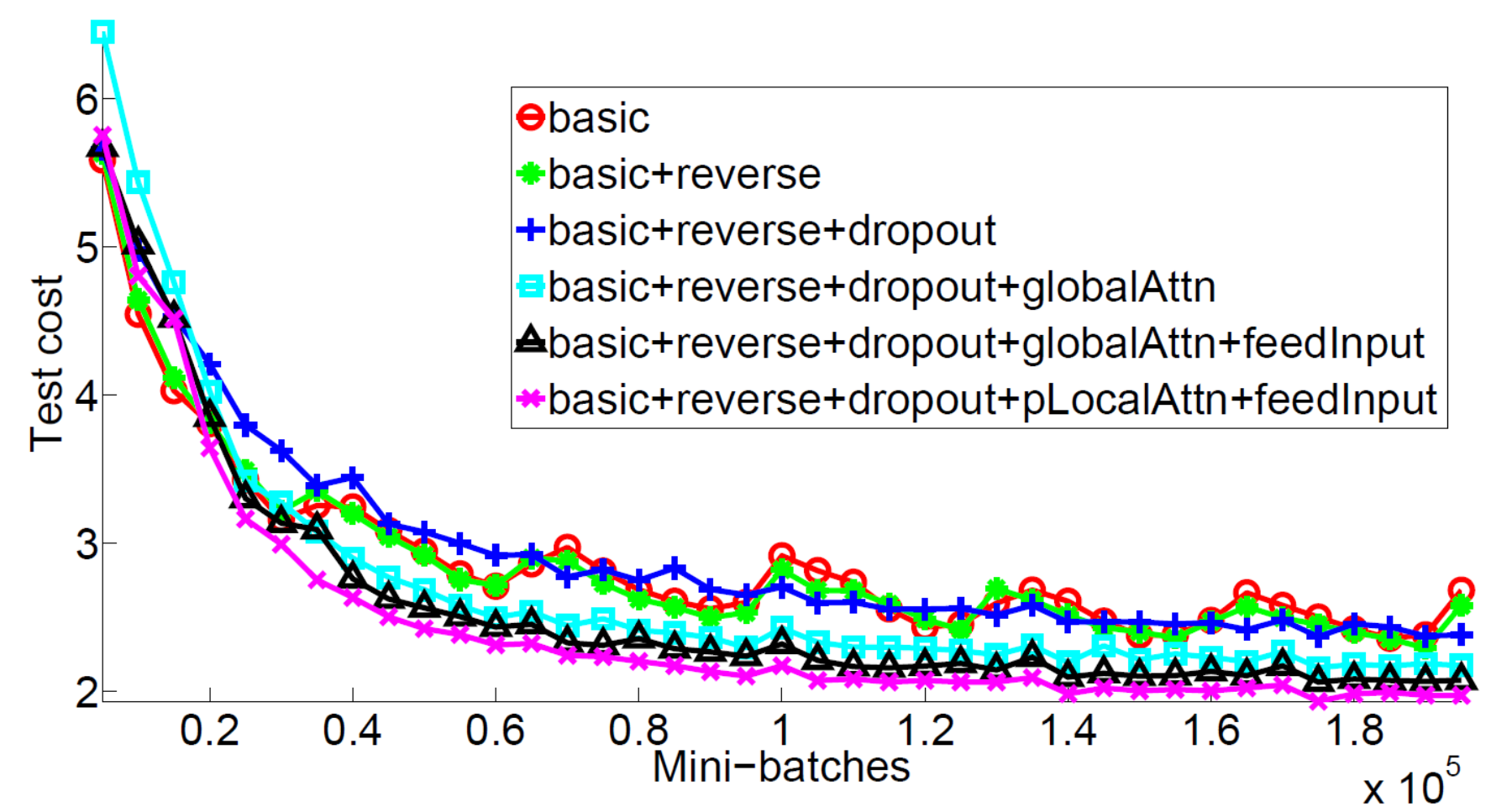

Imagen tomada de Luong et al. *Effective Approaches to Attention-based Neural Machine Translation*, EMNLP, 2015

Efecto del tamaño de secuencia en la atención

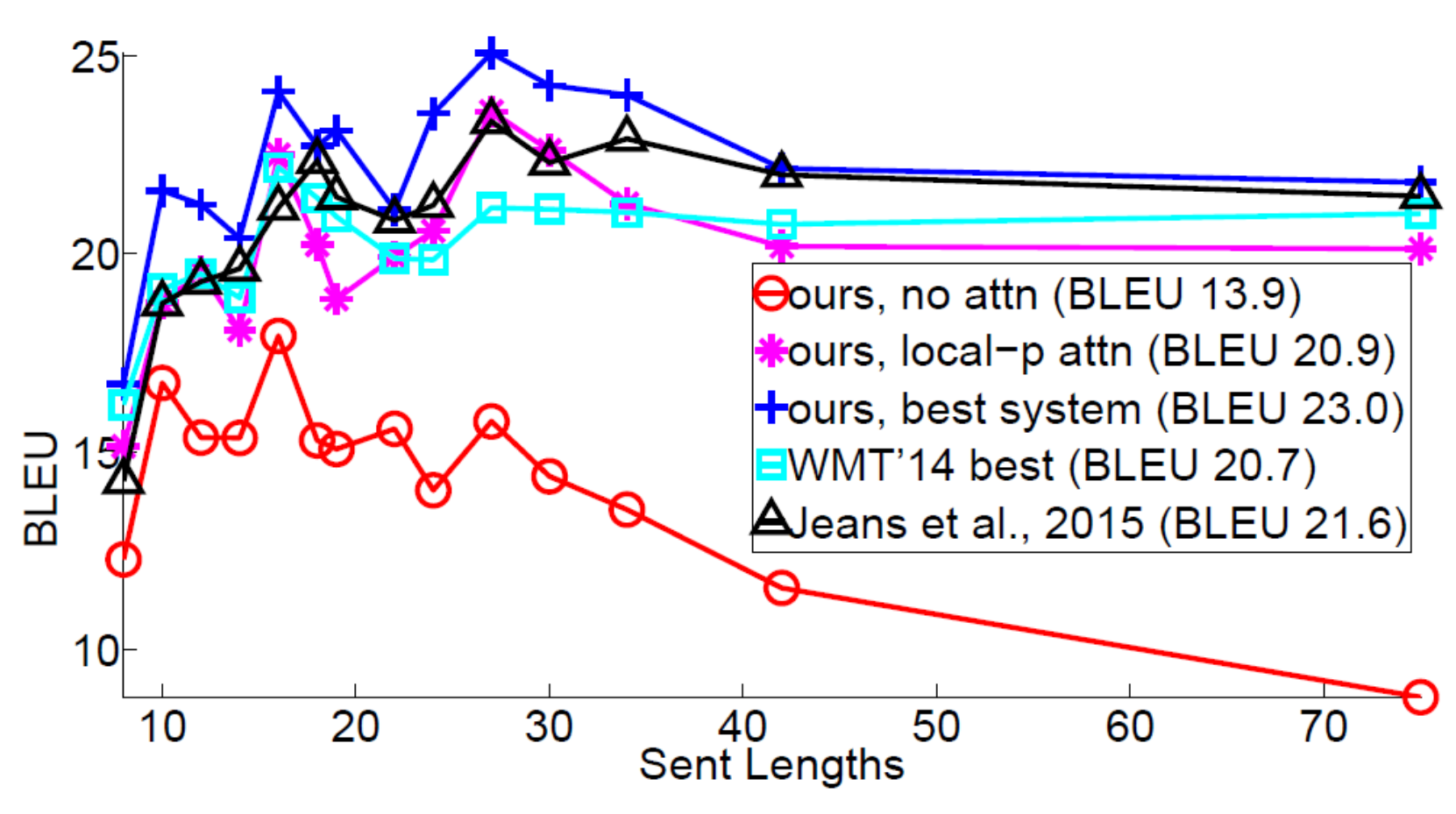

Imagen tomada de Luong et al. *Effective Approaches to Attention-based Neural Machine Translation*, EMNLP, 2015

Atención en traducción inglés-alemán

English-German translations	
src	Orlando Bloom and Miranda Kerr still love each other
ref	Orlando Bloom und *Miranda Kerr* lieben sich noch immer
best	Orlando Bloom und *Miranda Kerr* lieben einander noch immer .
base	Orlando Bloom und **Lucas Miranda** lieben einander noch immer .
src	″ We ′ re pleased the FAA recognizes that an enjoyable passenger experience is not incompatible with safety and security , ″ said Roger Dow , CEO of the U.S. Travel Association .
ref	“ Wir freuen uns , dass die FAA erkennt , dass ein angenehmes Passagiererlebnis nicht im Widerspruch zur Sicherheit steht ” , sagte *Roger Dow* , CEO der U.S. Travel Association .
best	″ Wir freuen uns , dass die FAA anerkennt , dass ein angenehmes ist nicht mit Sicherheit und Sicherheit *unvereinbar* ist ″ , sagte *Roger Dow* , CEO der US - die .
base	″ Wir freuen uns über die <unk> , dass ein <unk> <unk> mit Sicherheit nicht **vereinbar** ist mit Sicherheit und Sicherheit ″ , sagte *Roger* **Cameron** , CEO der US - <unk> .

Imagen tomada de Luong et al. *Effective Approaches to Attention-based Neural Machine Translation*, EMNLP, 2015

Atención en traducción alemán-inglés

German-English translations	
src	In einem Interview sagte Bloom jedoch , dass er und Kerr sich noch immer lieben .
ref	However , in an interview , Bloom has said that he and *Kerr* still love each other .
best	In an interview , however , Bloom said that he and *Kerr* still love .
base	However , in an interview , Bloom said that he and **Tina** were still <unk> .
src	Wegen der von Berlin und der Europäischen Zentralbank verhängten strengen Sparpolitik in Verbindung mit der Zwangsjacke , in die die jeweilige nationale Wirtschaft durch das Festhalten an der gemeinsamen Währung genötigt wird , sind viele Menschen der Ansicht , das Projekt Europa sei zu weit gegangen
ref	The *austerity imposed by Berlin and the European Central Bank , coupled with the straitjacket* imposed on national economies through adherence to the common currency , has led many people to think Project Europe has gone too far .
best	Because of the strict *austerity measures imposed by Berlin and the European Central Bank in connection with the straitjacket* in which the respective national economy is forced to adhere to the common currency , many people believe that the European project has gone too far .
base	Because of the pressure **imposed by the European Central Bank and the Federal Central Bank with the strict austerity** imposed on the national economy in the face of the single currency , many people believe that the European project has gone too far .

Imagen tomada de Luong et al. *Effective Approaches to Attention-based Neural Machine Translation*, EMNLP, 2015

Visualización de puntuaciones de la alineación

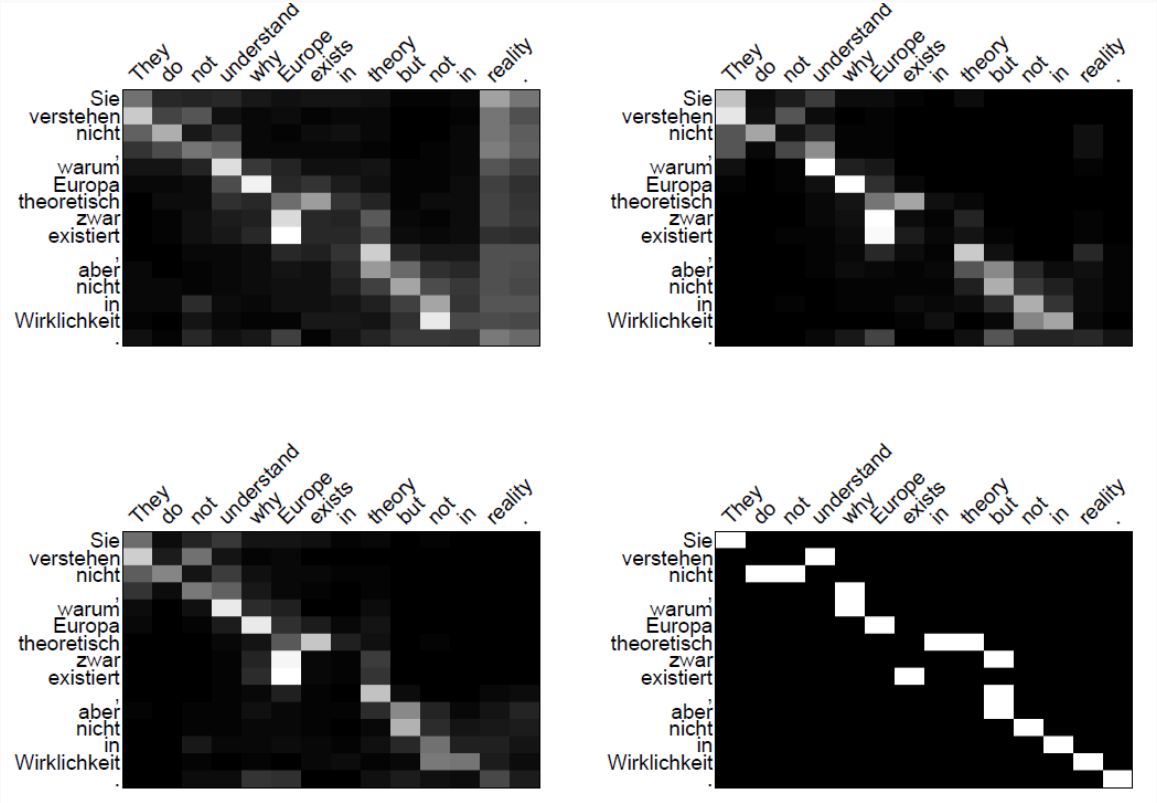

Imagen tomada de Bahdanau et al. *Neural machine translation by jointly learning to align and translate*, ICLR, 2015

Atención en imágenes: descripción

- Permiten enforcarse a sólo ciertas partes de la entrada al producir cada salida
- Modelo aprende a qué partes ponerle atención en cada paso

Imagen tomada de Xu et al. *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention*, ICML, 2015

Atención en imágenes: arquitectura

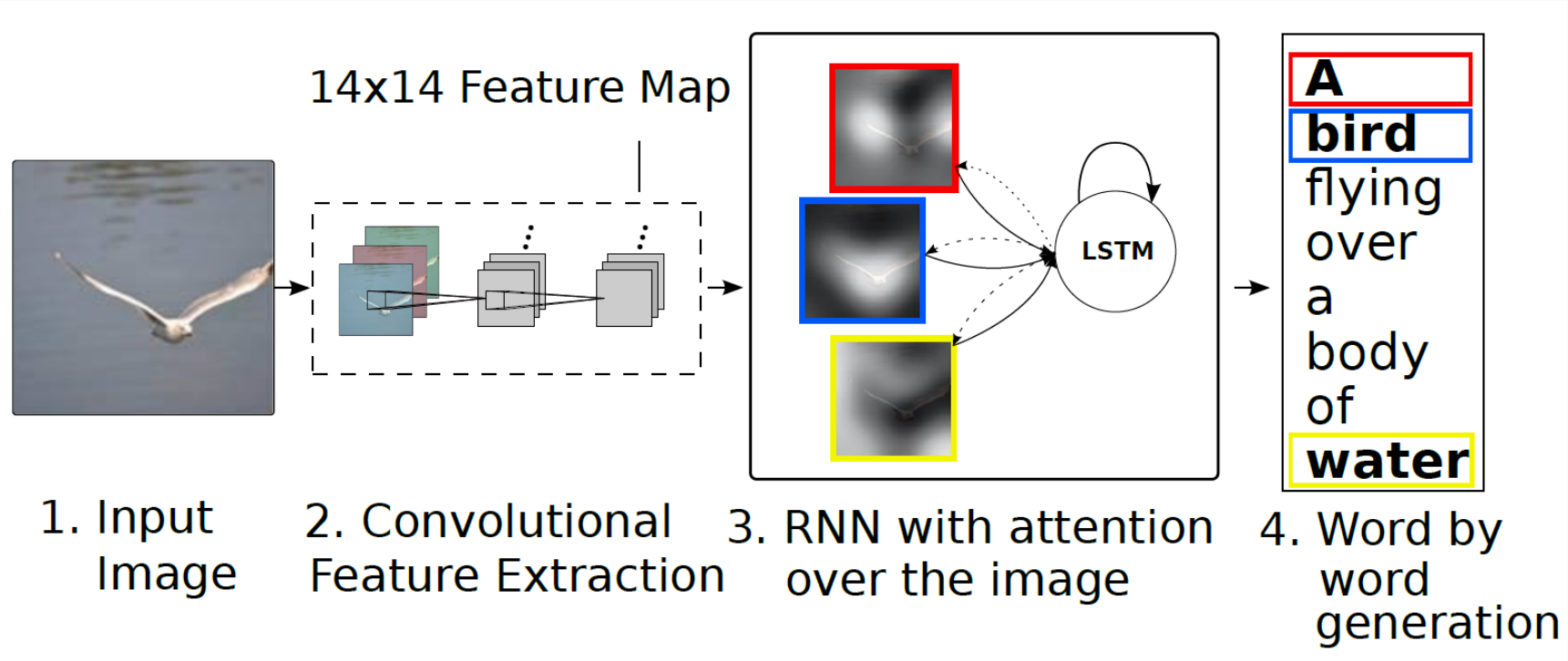

Imagen tomada de Xu et al. *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention*, ICML, 2015

Atención en imágenes: suave vs dura

- **Dura:** se toma en cuenta una sola region de la imagen

- **Suave:** se toma en cuenta cada región de la imagen en distinta proporción, de forma similar a la atención de Bahdanau

Figure 2. Attention over time. As the model generates each word, its attention changes to reflect the relevant parts of the image. "soft" (top row) vs "hard" (bottom row) attention. (Note that both models generated the same captions in this example.)

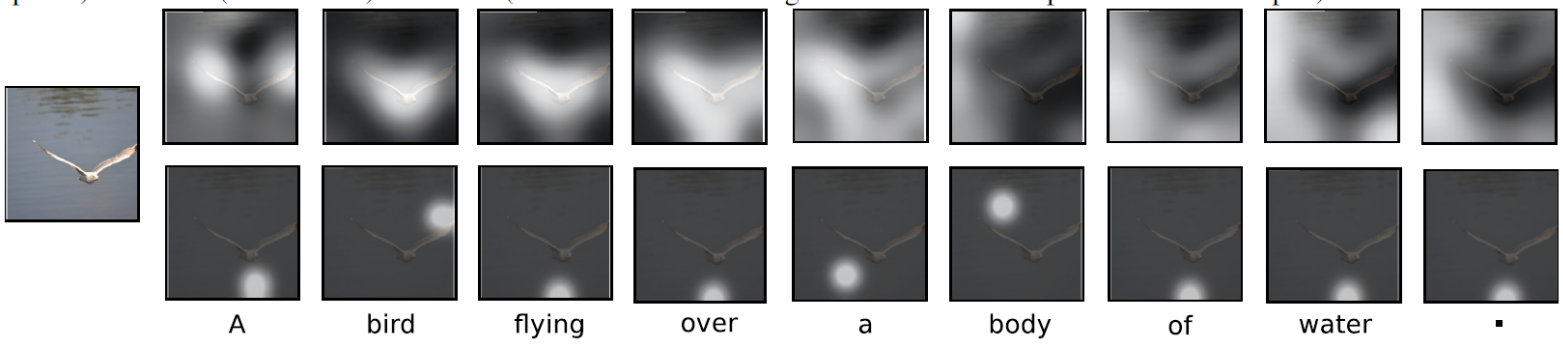

Imagen tomada de Xu et al. *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention*, ICML, 2015

Atención en imágenes: errores

Figure 5. Examples of mistakes where we can use attention to gain intuition into what the model saw.

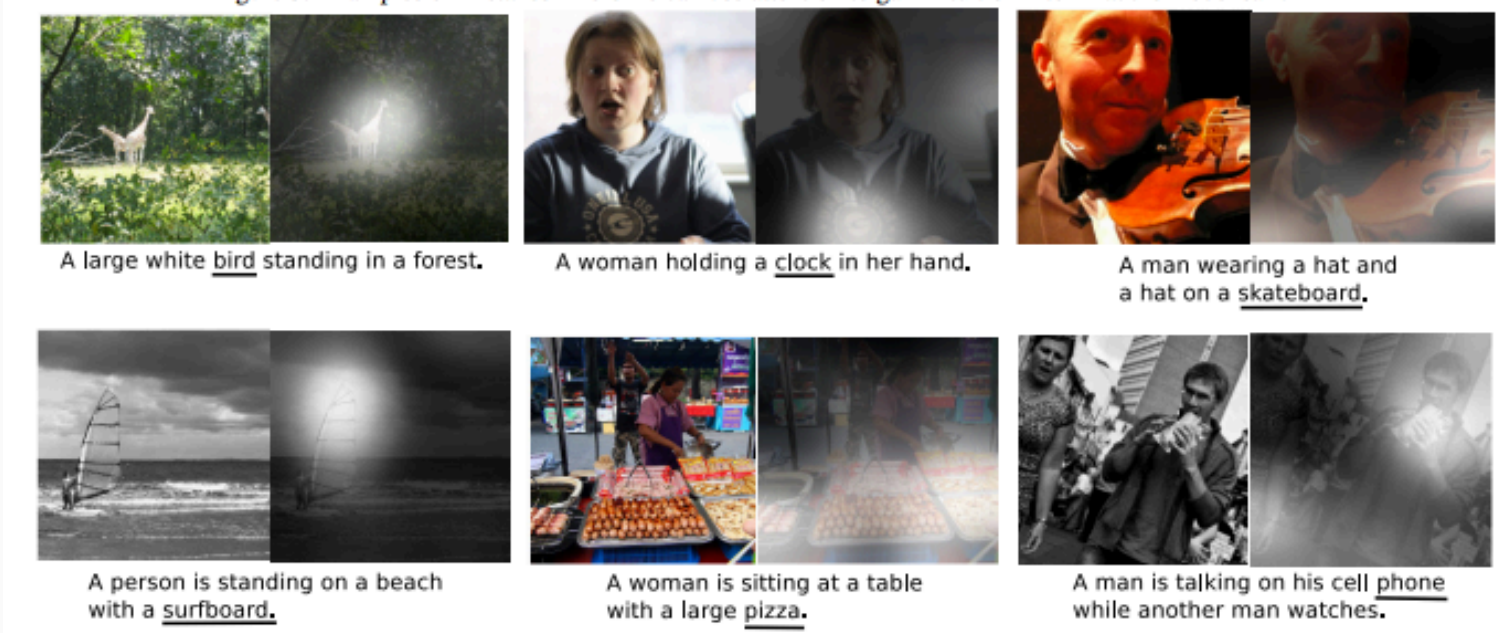

Imagen tomada de Xu et al. *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention*, ICML, 2015

Autoatención

- Cada salida $\mathbf{y}^{[i]}$ es simplemente la suma ponderada de todas las entradas x_j en la secuencia:

$$\mathbf{y}^{[i]} = \sum_j \alpha_{i,j} \cdot \mathbf{x}_j, \text{ donde } \sum_j \alpha_{i,j} = 1$$

- Cada ponderación $\alpha_{i,j}$ se obtiene a partir de una función de la entrada $\mathbf{x}^{[i]}$ correspondiente a la salida $\mathbf{y}^{[i]}$ y cada entrada $\mathbf{x}_j$. Por ej.,

$$\alpha_{i,j} = softmax(\mathbf{x}^{[i]\top}\mathbf{x}_j)$$

- La operación de atención propia no considera el orden

Arquitectura *Transformer*

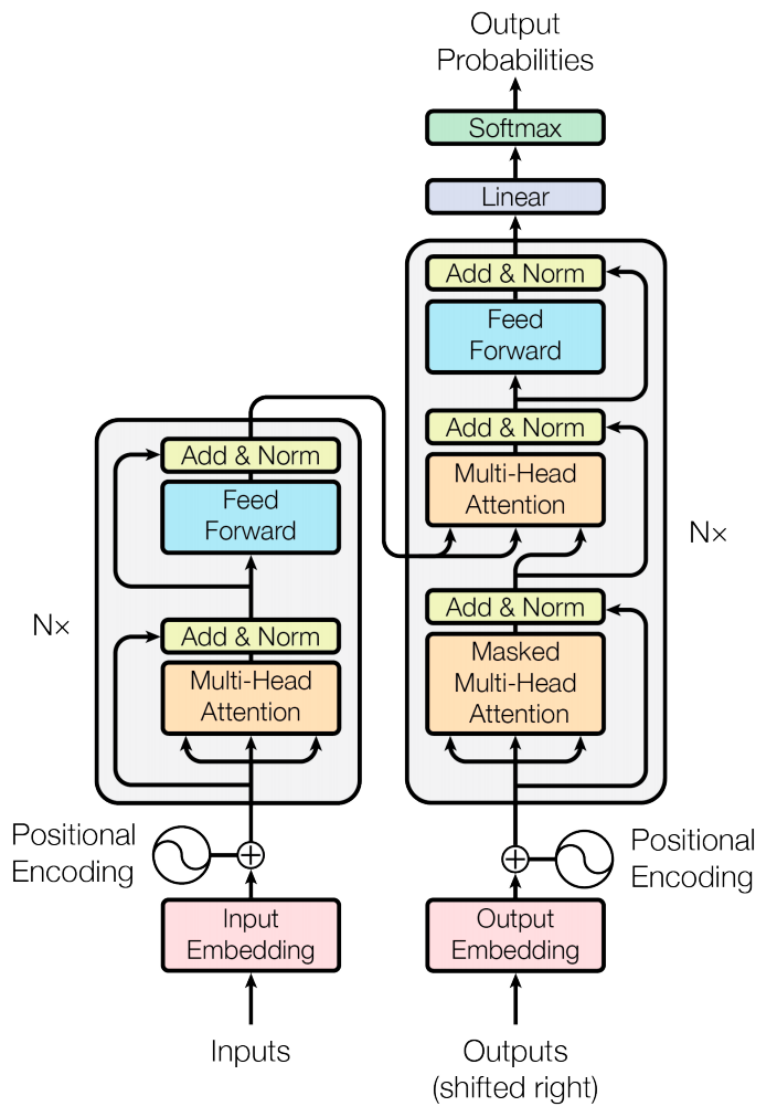

Imagen tomada de Vaswani et al. *Attention Is All You Need*, NIPS, 2017

Transformer: autoatención

- Se transforma linealmente cada entrada $\mathbf{x}^{[i]}$ a los vectores consulta ($\mathbf{q}^{[i]}$), llave ($\mathbf{k}^{[i]}$) y valor ($\mathbf{v}^{[i]}$)

$$\mathbf{q}^{[i]} = \mathbf{W}_q \mathbf{x}^{[i]}$$
$$\mathbf{k}^{[i]} = \mathbf{W}_k \mathbf{x}^{[i]}$$
$$\mathbf{v}^{[i]} = \mathbf{W}_v \mathbf{x}^{[i]}$$

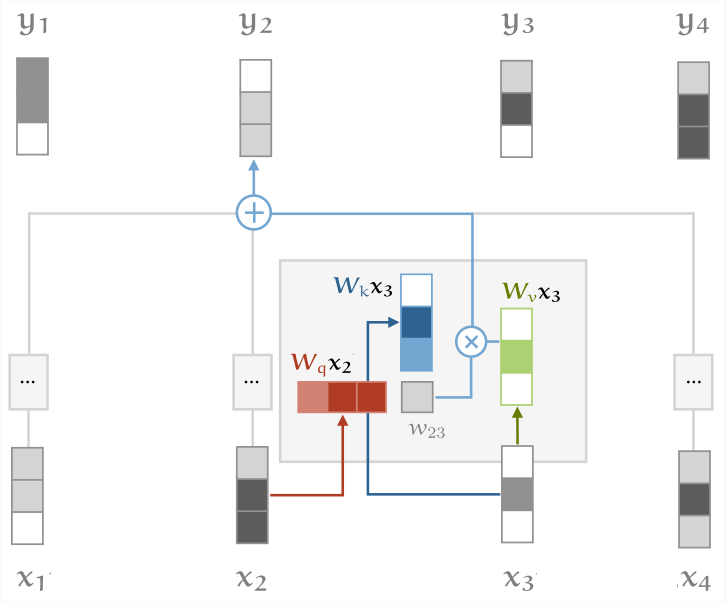

Imagen tomada de `http://www.peterbloem.nl/blog/transformers`

- Cada salida $\mathbf{y}^{[i]}$ es la suma ponderada de todos los valores $\mathbf{v}_j, j = 1, \ldots S$ por la alineación $\alpha_{i,j}$, esto es, $\mathbf{y}^{[i]} = \sum_j \alpha_{i,j} \mathbf{v}_j$

Transformer: alineación por producto punto normalizado

- Considerando que $\mathbf{q}^{[i]}, \mathbf{k}^{[i]} \in \mathrm{R}^{d_k}$ y $\mathbf{v}^{[i]} \in \mathrm{R}^{d_v}$, la función de puntaje está dada por el producto punto normalizado

$$puntaje(\mathbf{x}^{[i]}, \mathbf{x}_j) = \frac{\mathbf{q}^{[i]\top}\mathbf{k}^{[i]}}{\sqrt{d_k}}$$

- Por lo tanto, la alineación quedaría como

$$\alpha_{i,j} = softmax(\frac{\mathbf{q}^{[i]\top}\mathbf{k}^{[i]}}{\sqrt{d_k}})$$

Transformer: autoatención multicabeza

- Se transforma cada entrada con h distintos $\mathbf{W}_q$, $\mathbf{W}_k$ y $\mathbf{W}_v$ y se calcula la autoatención para cada una

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = [\mathbf{Y}^1; \ldots; \mathbf{Y}^h]\mathbf{W}_o$$

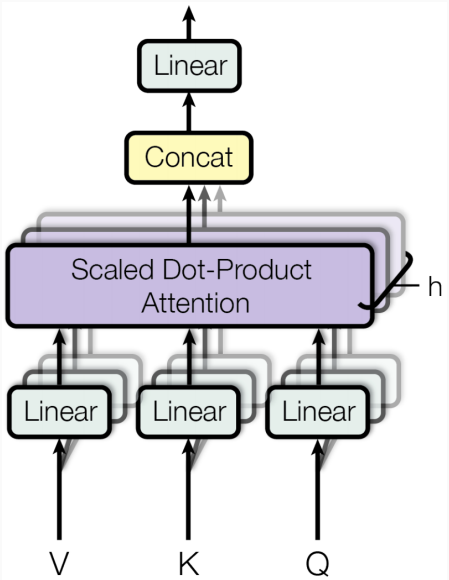

Imagen tomada de Vaswani et al. *Attention Is All You Need*, NIPS, 2017

Transformer: codificación posicional

- Para representar el orden en una secuencia, se codifica la posición de cada entrada, la cual se agrega a su *embedding*
- Vaswani et al. proponen funciones sinusoidales

$$PE(pos, 2i) = sin\left[\frac{pos}{10000^{\left(\frac{2i}{d_{model}}\right)}}\right]$$

$$PE(pos, 2i+1) = cos\left[\frac{pos}{10000^{\left(\frac{2i}{d_{model}}\right)}}\right]$$

  donde pos es la posición en la secuencia, i la dimensión y d_{model} es igual a la dimensionalidad del *embedding*
- Es posible también aprender la codificación[1]

[1] Por ej. Gehring et al. *Convolutional Sequence to Sequence Learning*, arxiv:1705.03122, 2017

Transformer: red hacia adelante por posición

- Las salidas de los bloques de autoatención se conectan a 2 capas densas, la última con función de activación ReLU

$$FFN(X) = \mathbf{máx}(0, x \cdot \mathbf{W}^{\{1\}} + b^{\{1\}}) \cdot \mathbf{W}^{\{2\}} + b^{\{2\}}$$

- Esto se realiza de forma separada por cada posición de la entrada (como una convolución 1D con un filtro de tamaño 1)

- Llamada red hacia adelante por posición o *Position-wise Feed-Forward Networks*

Bloque tipo Transformer

- Generalmente compuestos de:
  1. Autoatención multicabeza
  2. Red hacia adelante por posición
- Con conexiones residuales y normalización por capa (*Layer Normalization*) en ambos

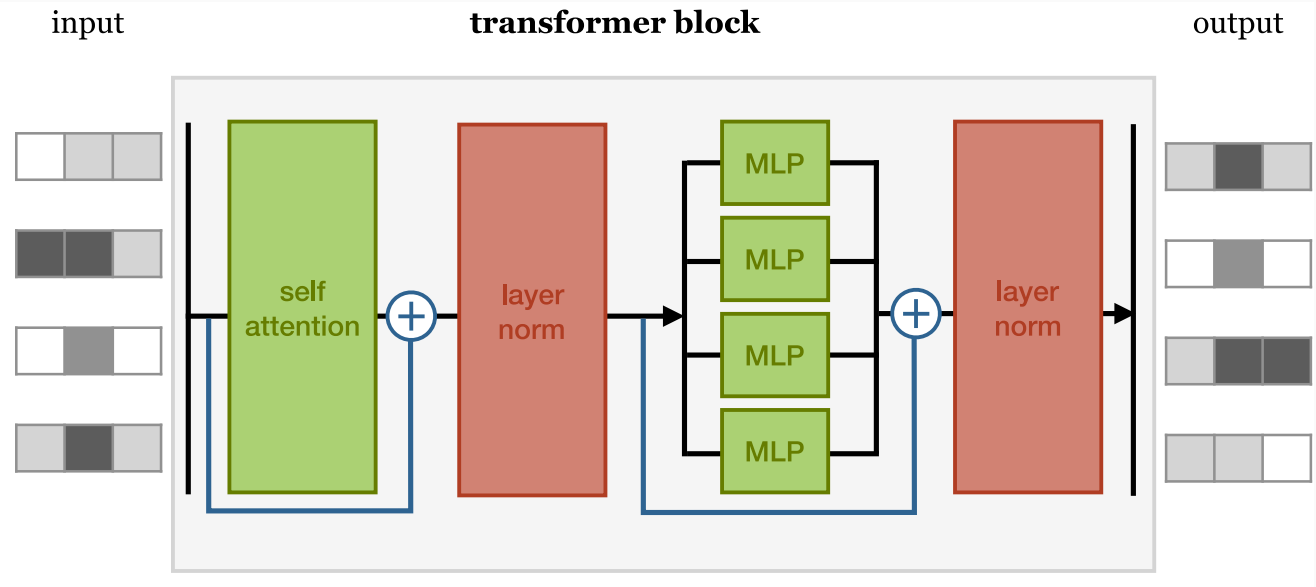

Imagen tomada de `http://www.peterbloem.nl/blog/transformers`

Arquitecturas Transformer: tarea de clasificación

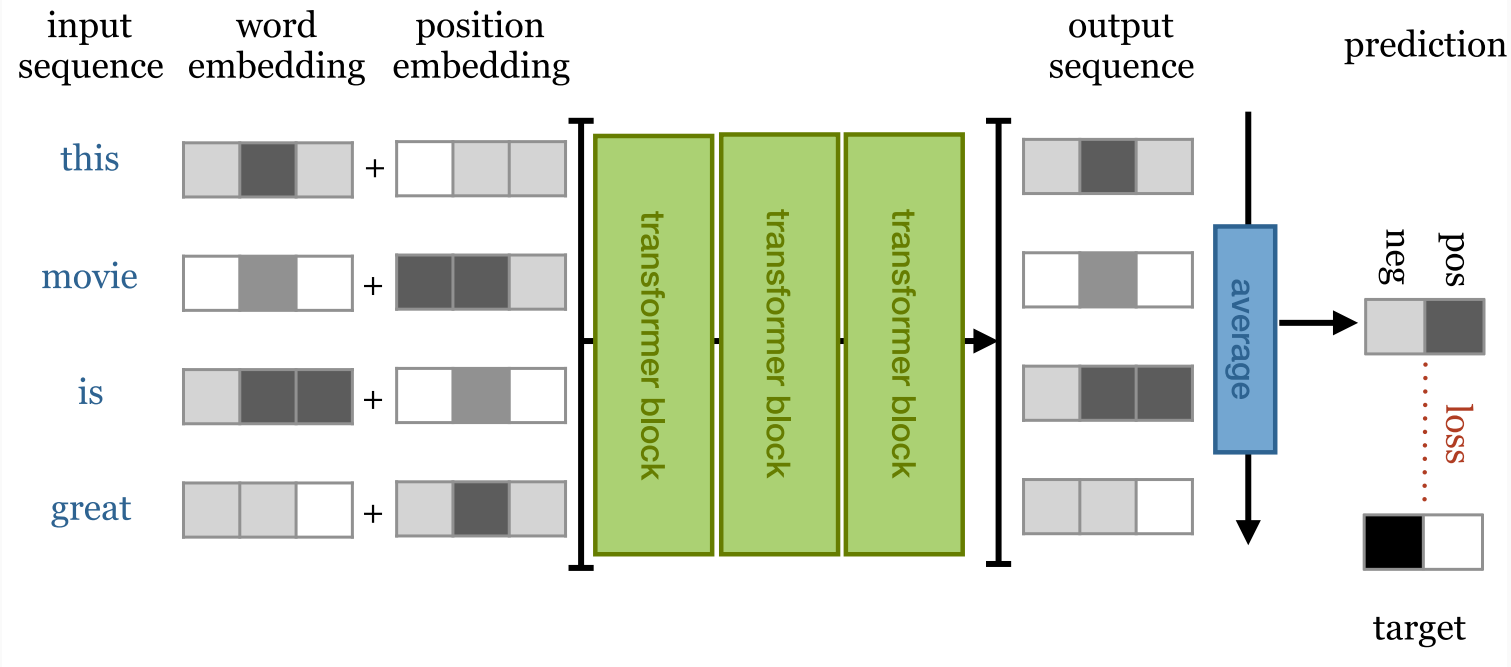

Imagen tomada de `http://www.peterbloem.nl/blog/transformers`

Arquitecturas Transformer: tarea de generación

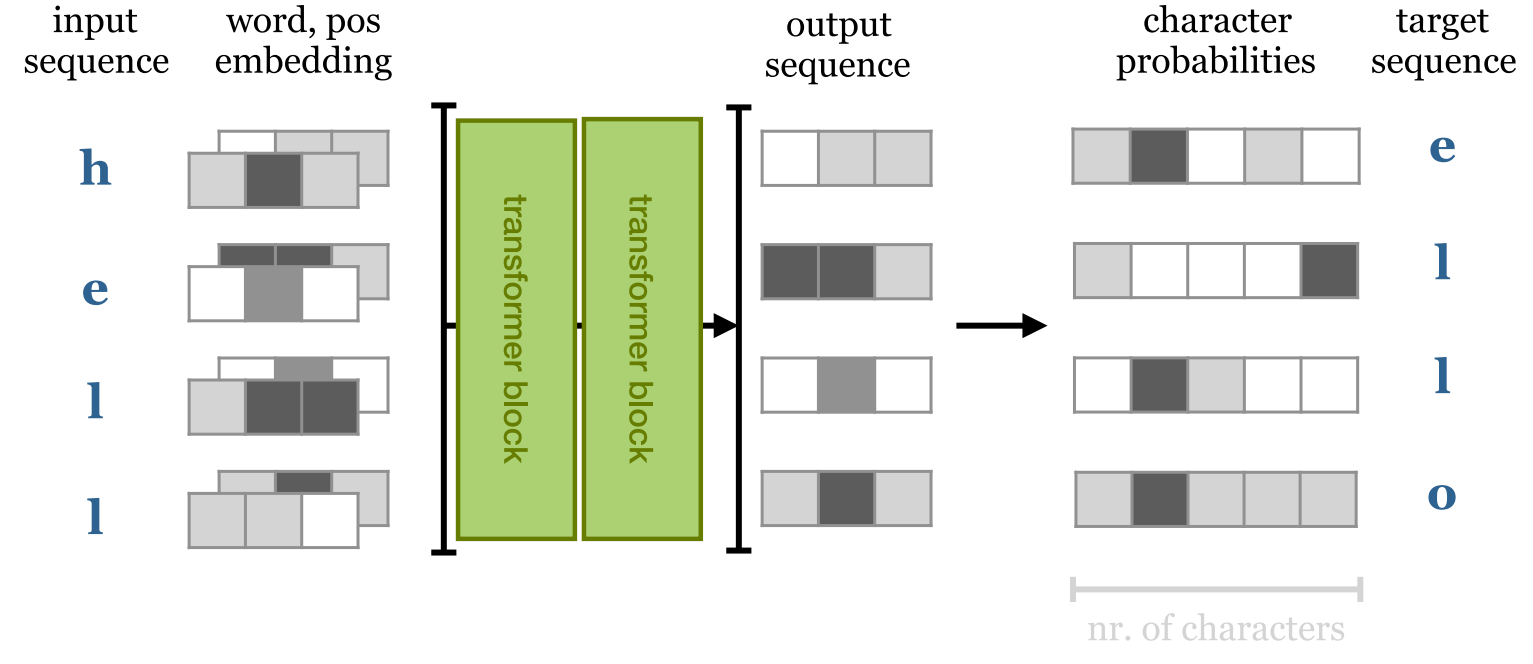

Imagen tomada de `http://www.peterbloem.nl/blog/transformers`

Transformer para generación: enmascaramiento

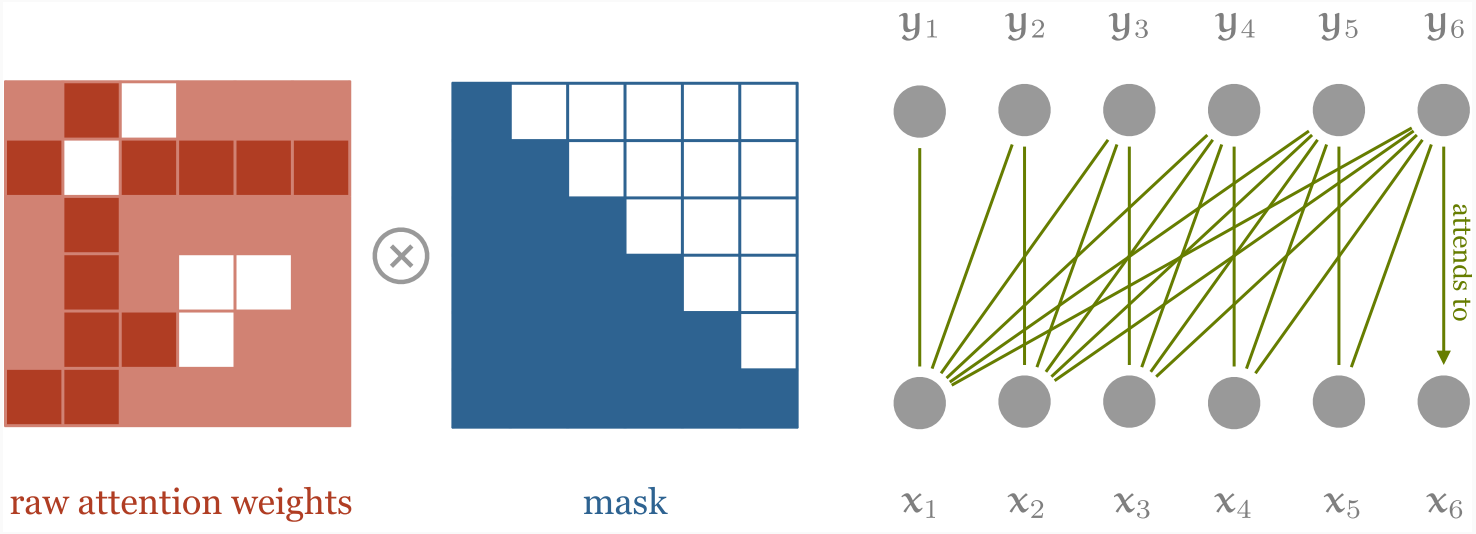

Imagen tomada de `http://www.peterbloem.nl/blog/transformers`

Generative Pre-training Transformer: GPT

- Decodificador con bloques tipo Transformer (descarta codificador)

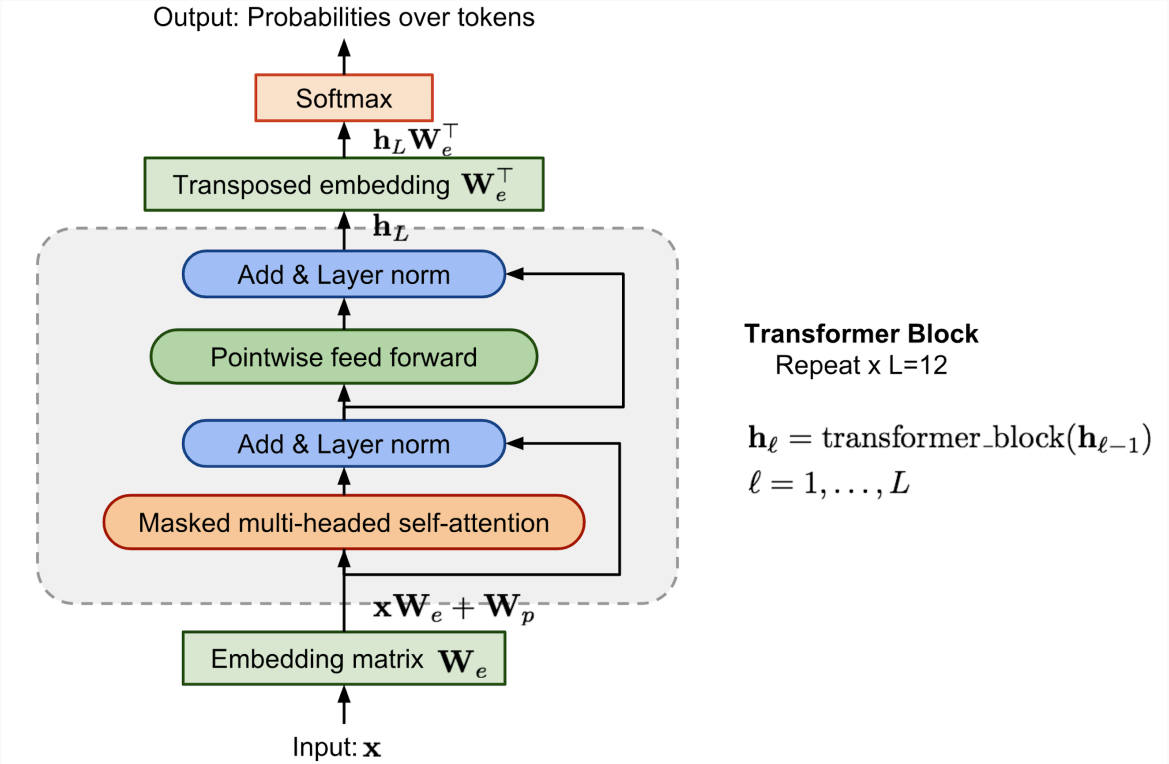

Imagen tomada de *https://lilianweng.github.io/lil-log/2019/01/31/generalized-language-models.html*

GPT pre-entrenamiento

- Pre-entrenamiento autosupervisado
- Aprendizaje por transferencia para distintas tareas de procesamiento del lenguaje natural

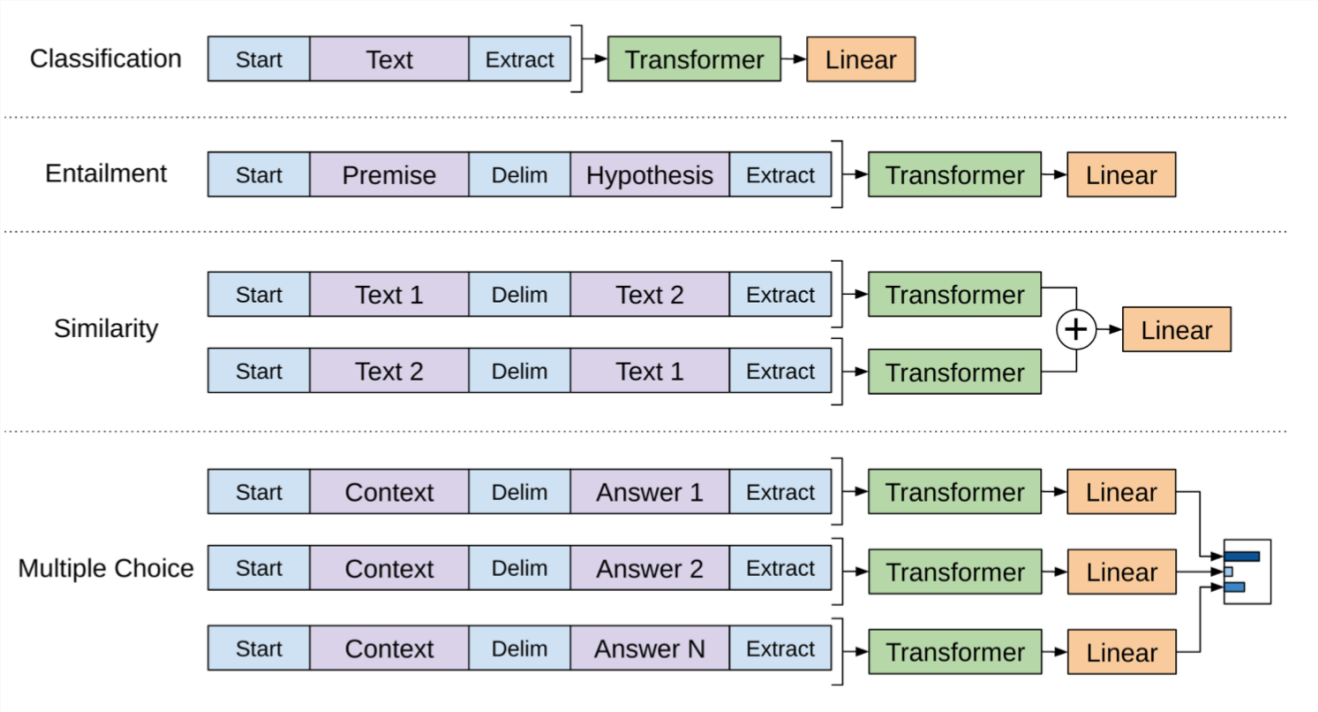

Imagen tomada de *https://lilianweng.github.io/lil-log/2019/01/31/generalized-language-models.html*

Bidirectional Encoder Representations from Transformers: BERT

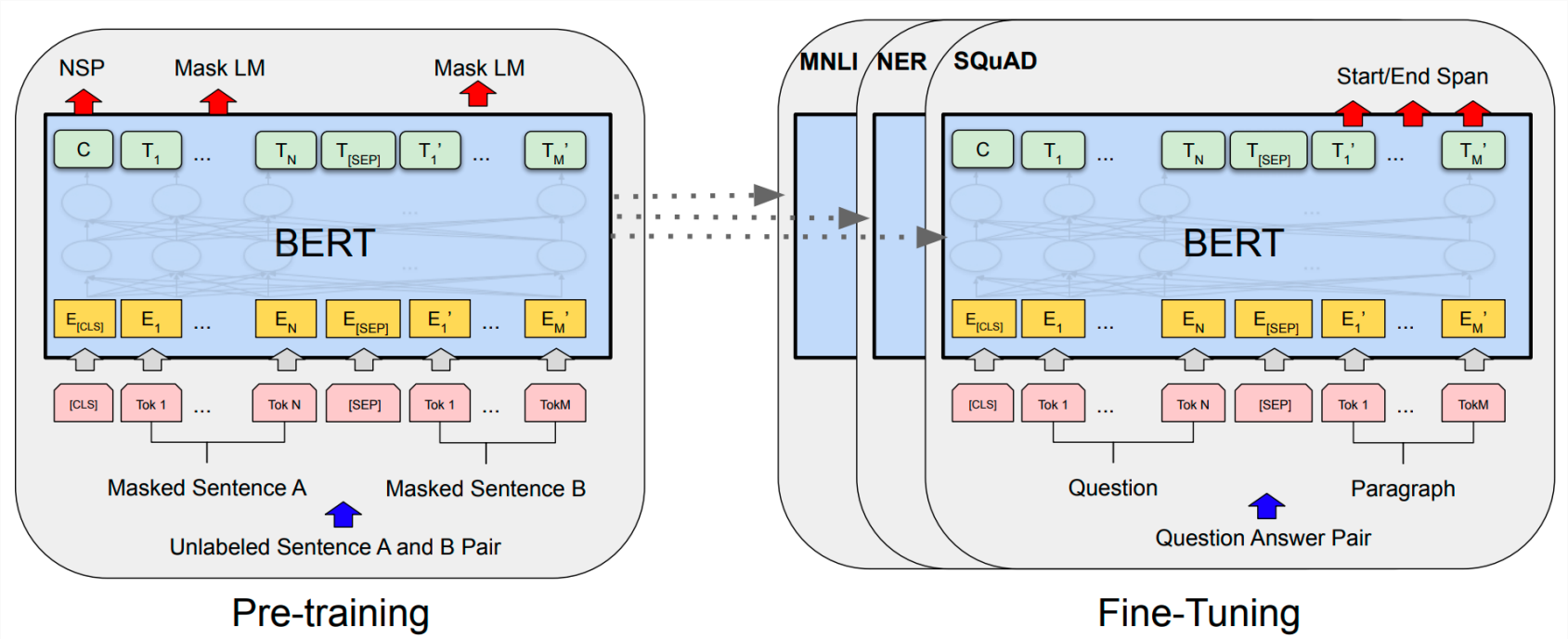

Imagen tomada de Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, arXiv:1810.04805, 2019

Modelos de lenguaje modernos

- Basados en arquitecturas profundas con bloques tipo Transformer

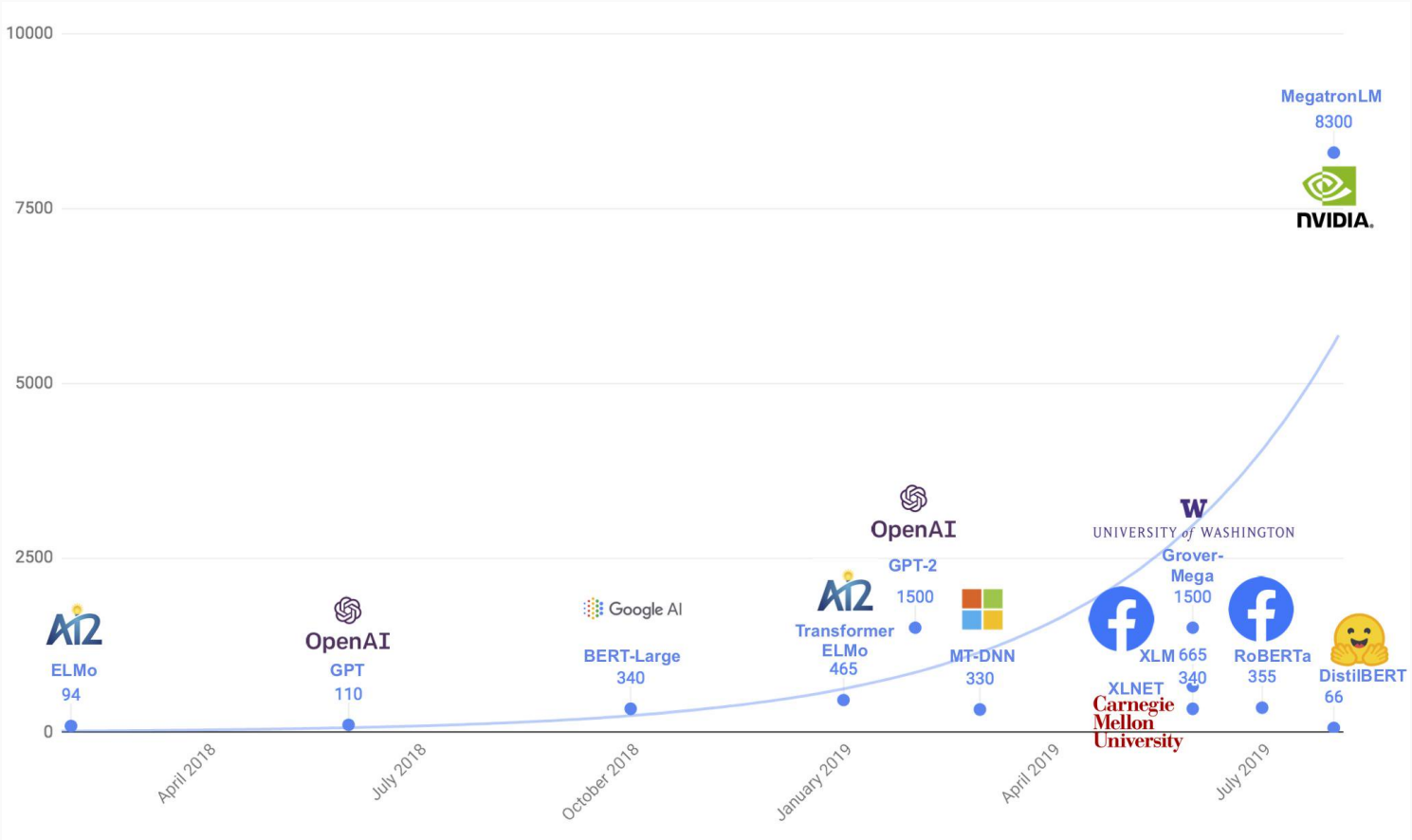

Imagen tomada de `https://medium.com/huggingface/distilbert-8cf3380435b5`.

Máquina neuronal de Turing: arquitectura

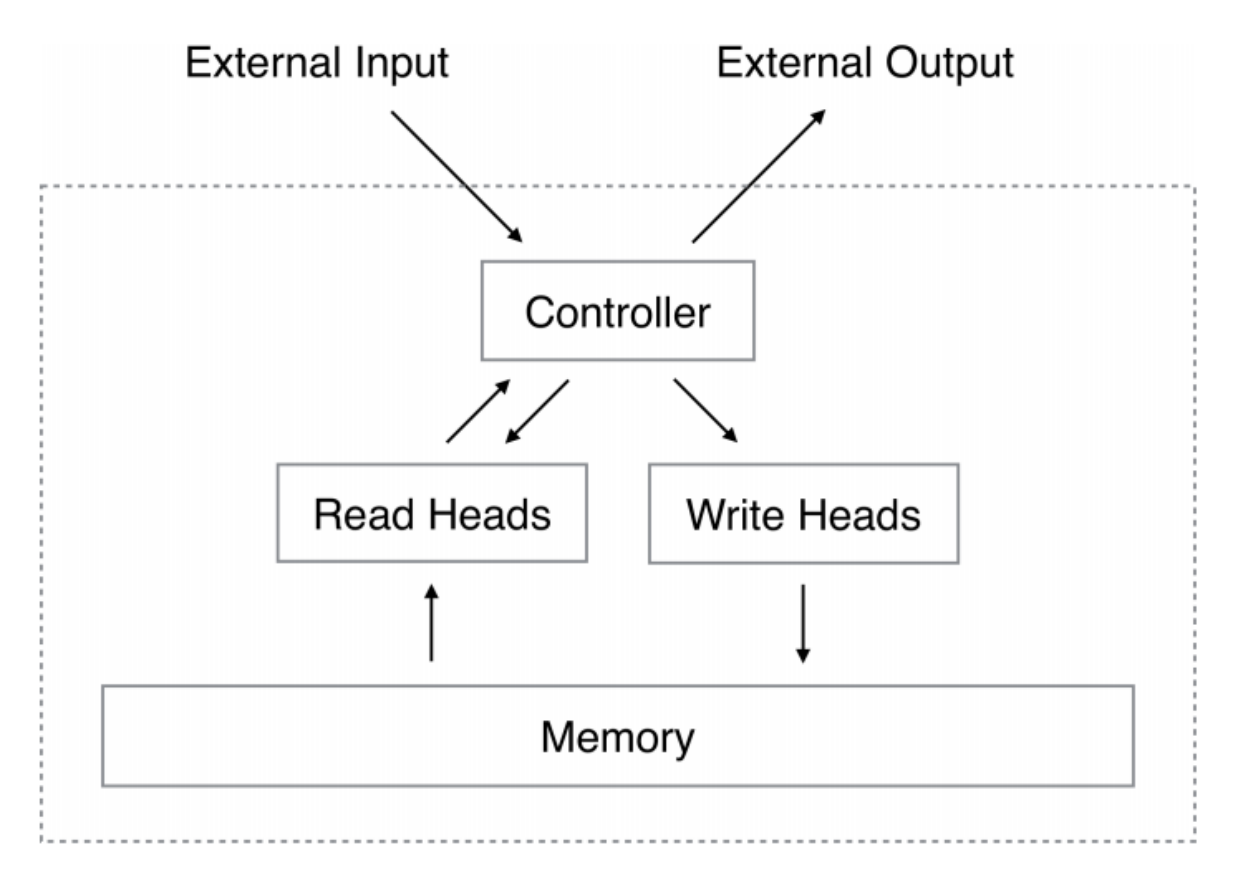

Imagen tomada de Graves et al. *Neural Turing Machines* , arXiv:1410.5401 , 2014

Máquina neuronal de Turing: direccionamiento

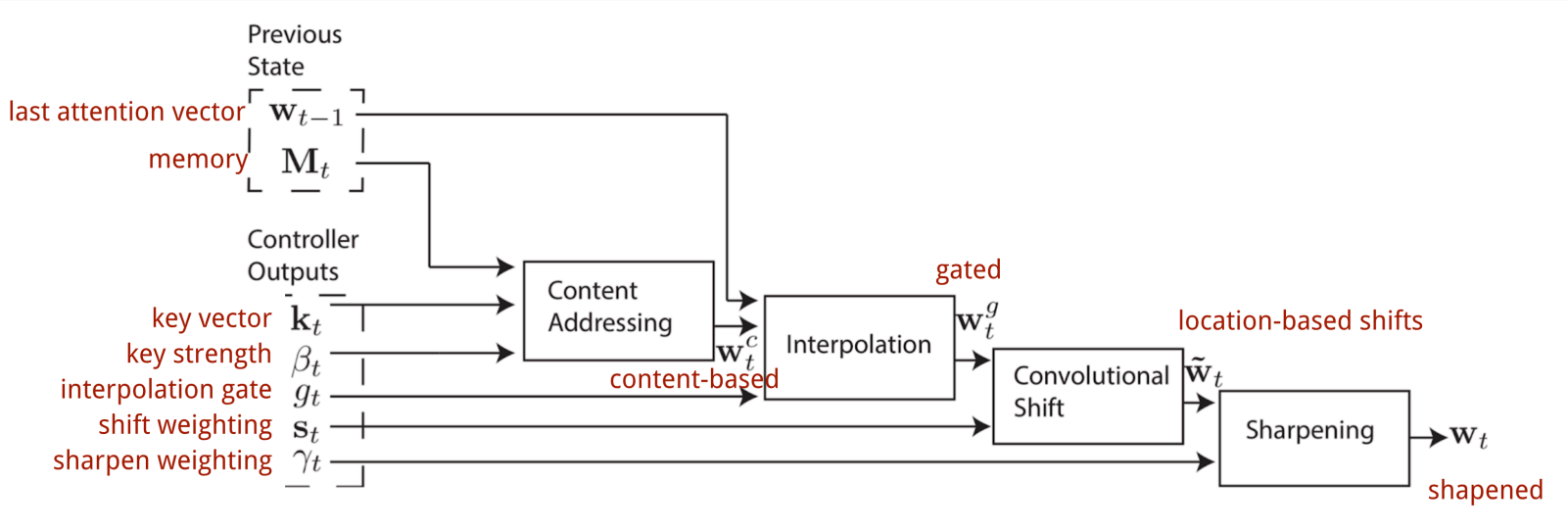

Imagen tomada de Graves et al. *Neural Turing Machines* , arXiv:1410.5401 , 2014