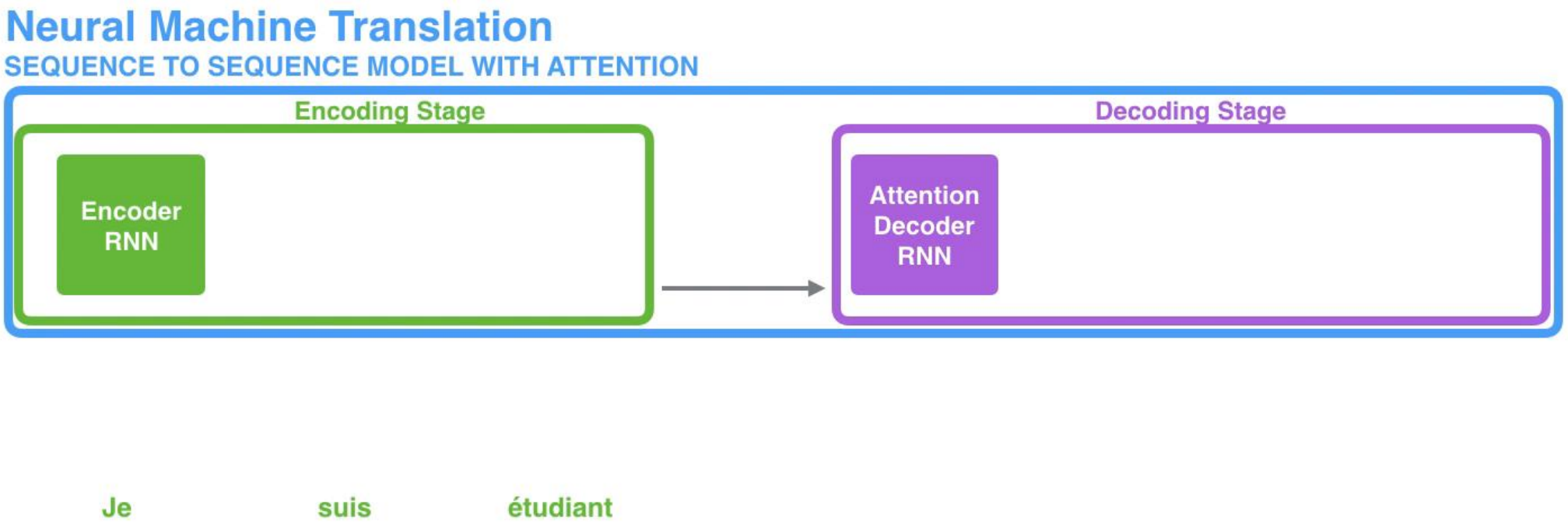
Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

Sistemas Urbanos Inteligentes

Autoatención

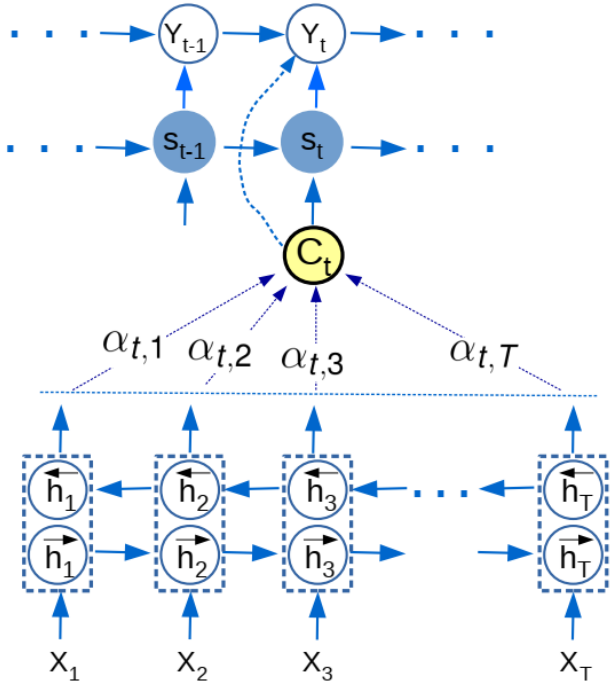
Hans Löbel
Dpto. Ingeniería de Transporte y Logística
Dpto. Ciencia de la Computación

Modelos seq2seq con atención conceptualmente funcionan muy bien

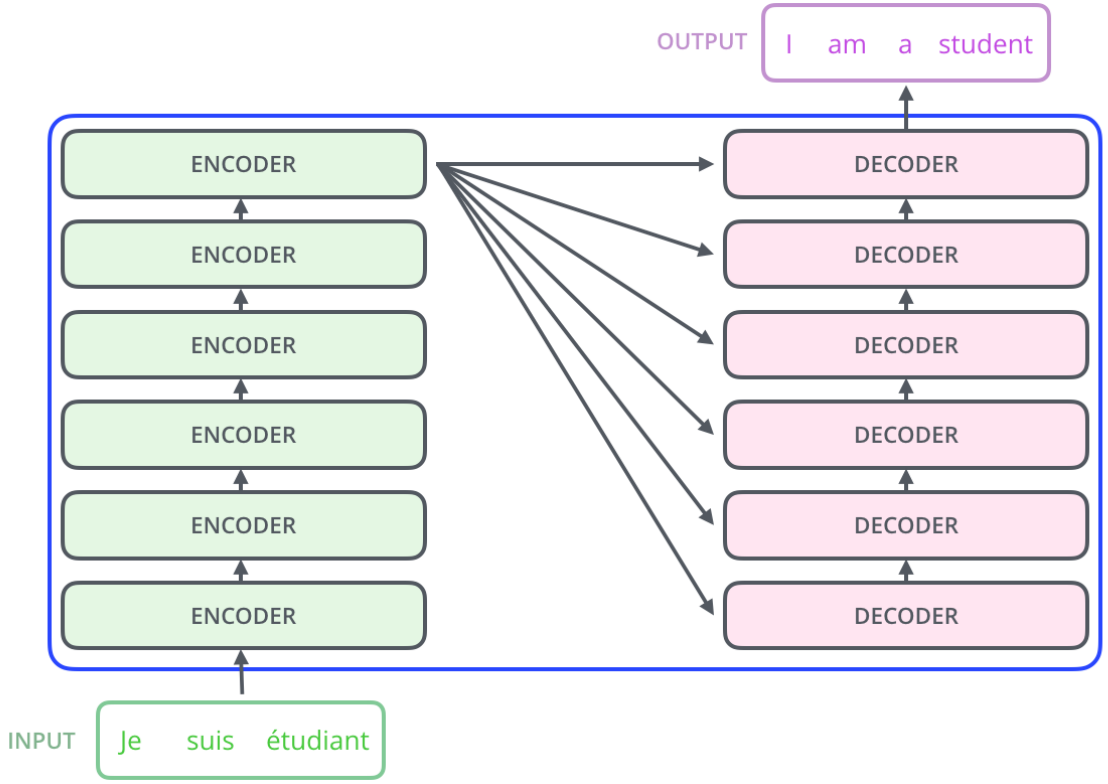


Lamentablemente, acarrean muchos de los problemas de las RNN

- Poco eficientes computacionalmente.
- Problemas con secuencias muy largas.
- Estos problemas complican su aplicación a sets de datos gigantescos, que potencialmente entregan mayor conocimiento.

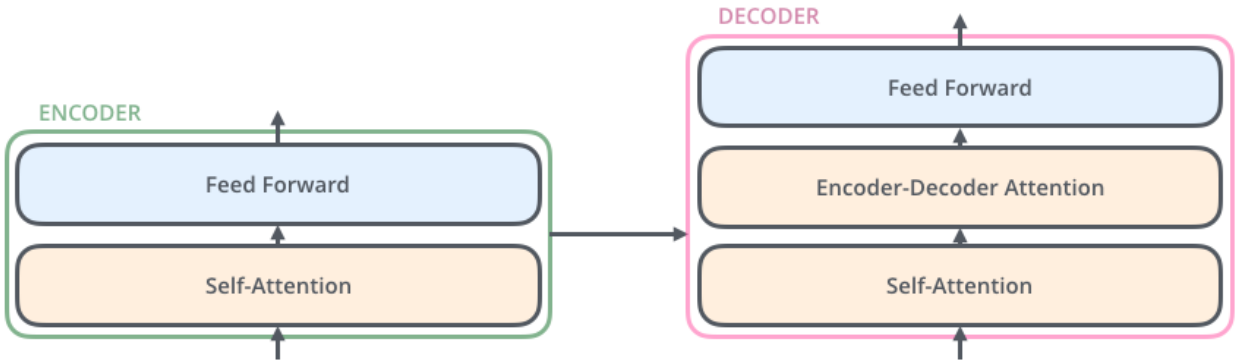


Una posible solución a estos problemas la entrega la arquitectura Transformer



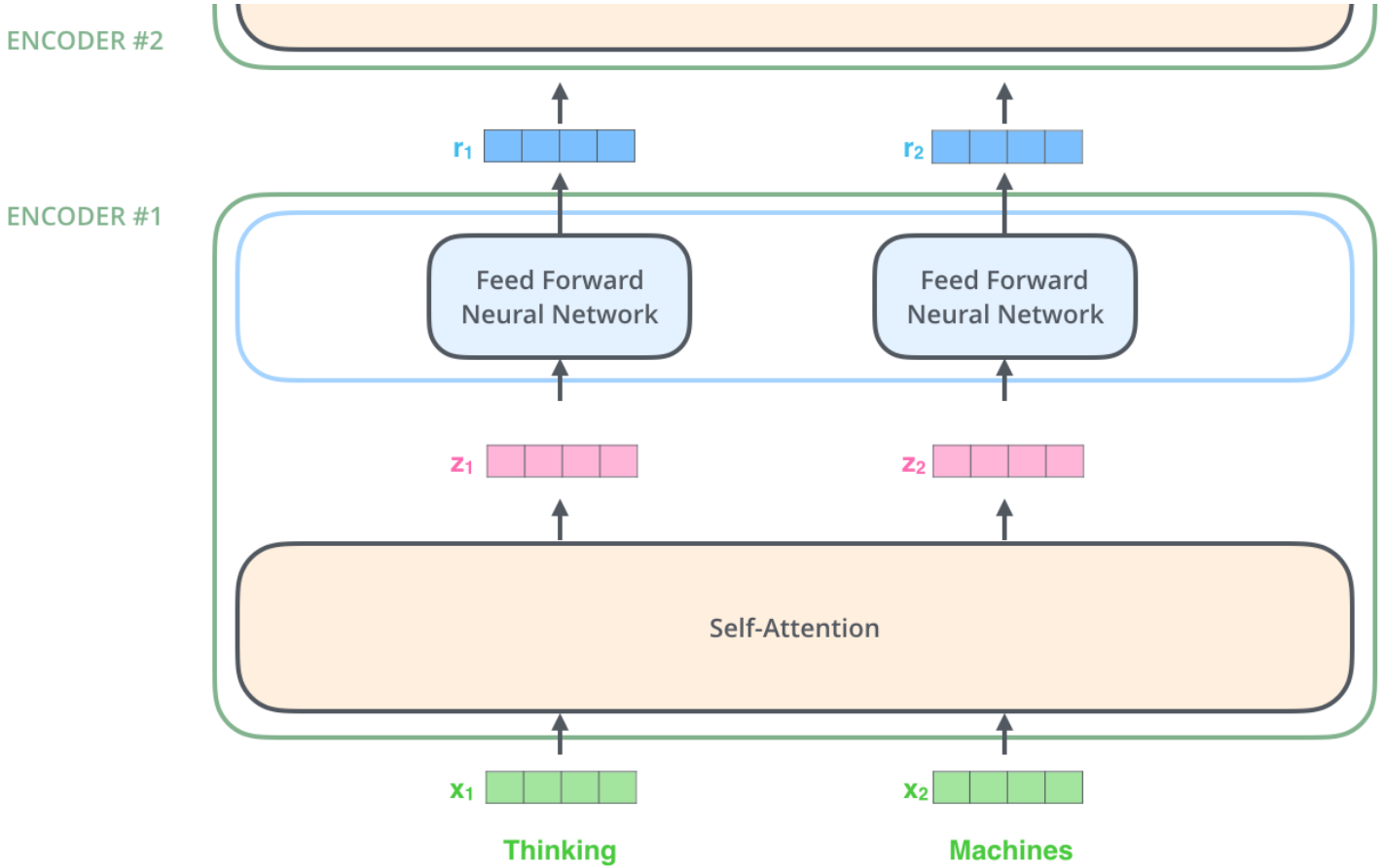
- Esta arquitectura profunda está completamente basada en mecanismos de atención.
- Su gran aporte es ser más eficiente y permitir dependencias de mayor largo que los modelos seq2seq.

Si bien también están formados por *encoder* y *decoders*, estos no son recurrentes, sino combinaciones de atención y capas densas

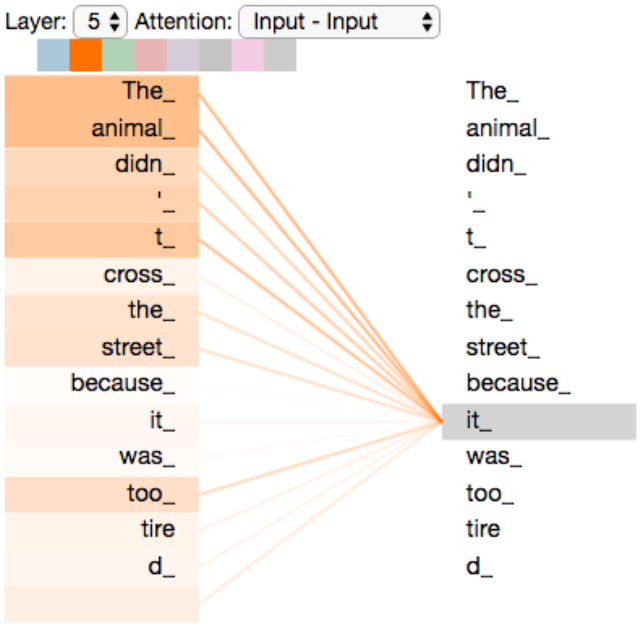


- Atención en Transformers no es igual a la de un modelo seq2seq.
- En este caso se utiliza la auto-atención, que indica para cada elementos de una secuencia, su dependencia con otros elementos de la misma.

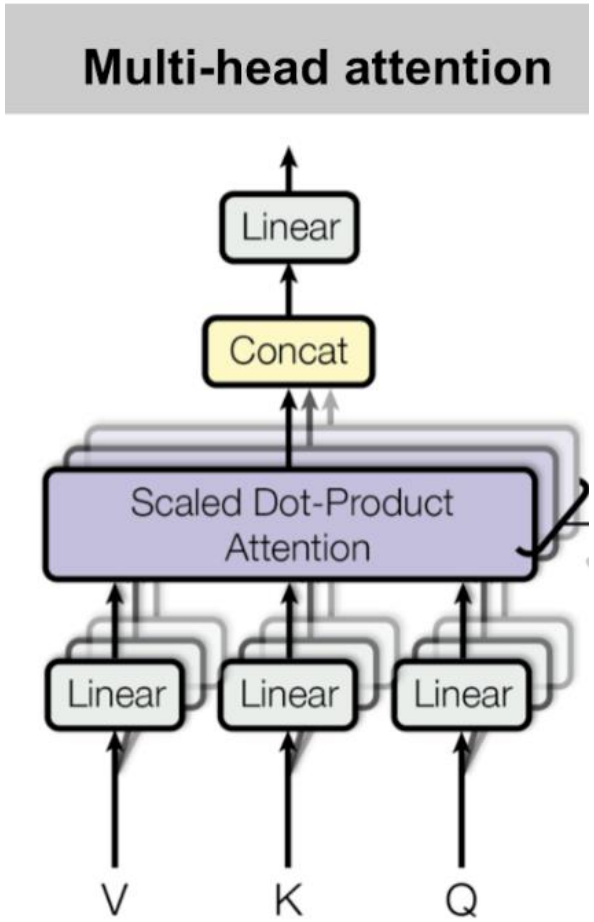
Si bien también están formados por *encoder* y *decoders*, estos no son recurrentes, sino combinaciones de atención y capas densas



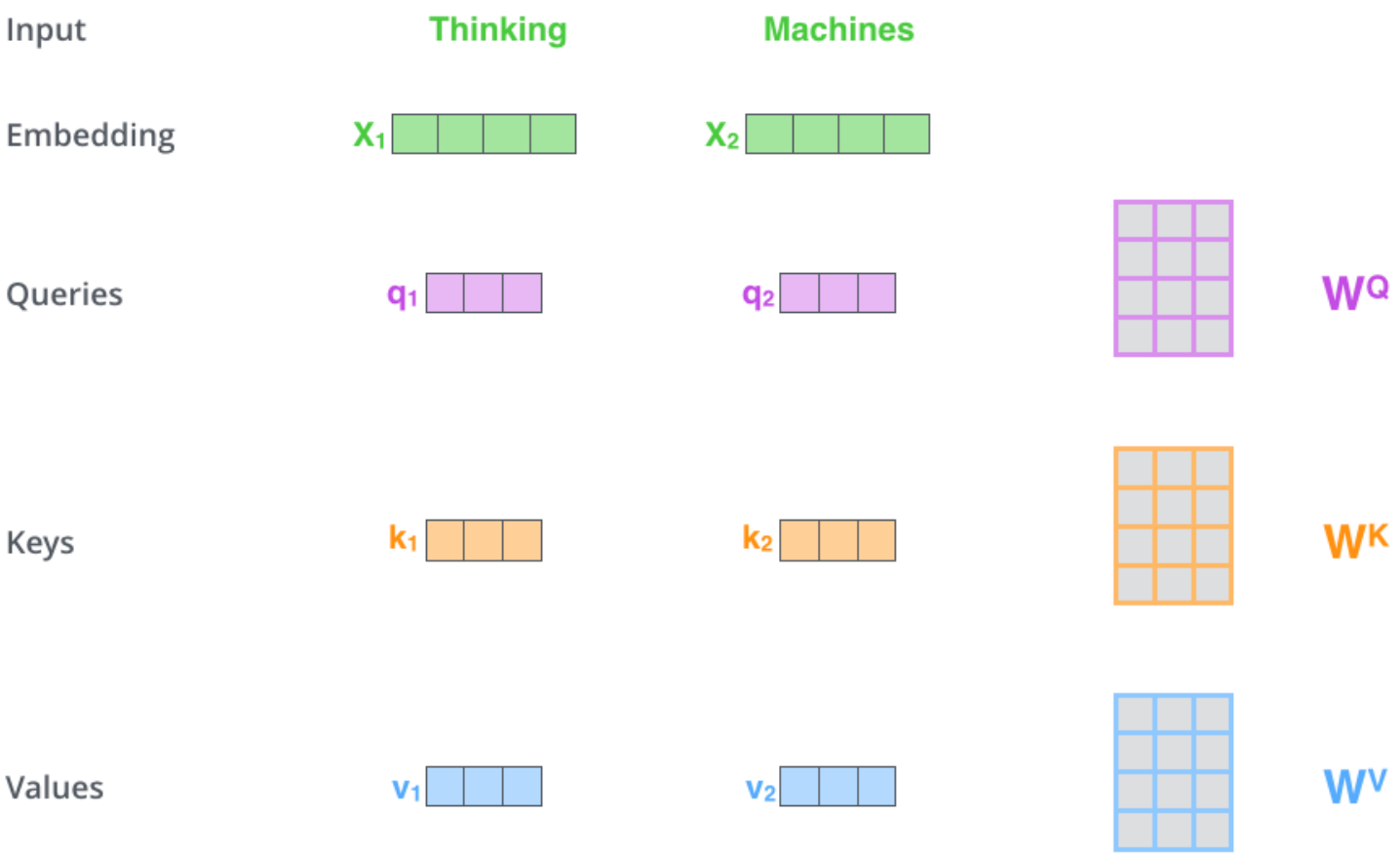
Auto-atención resulta ser el elemento clave



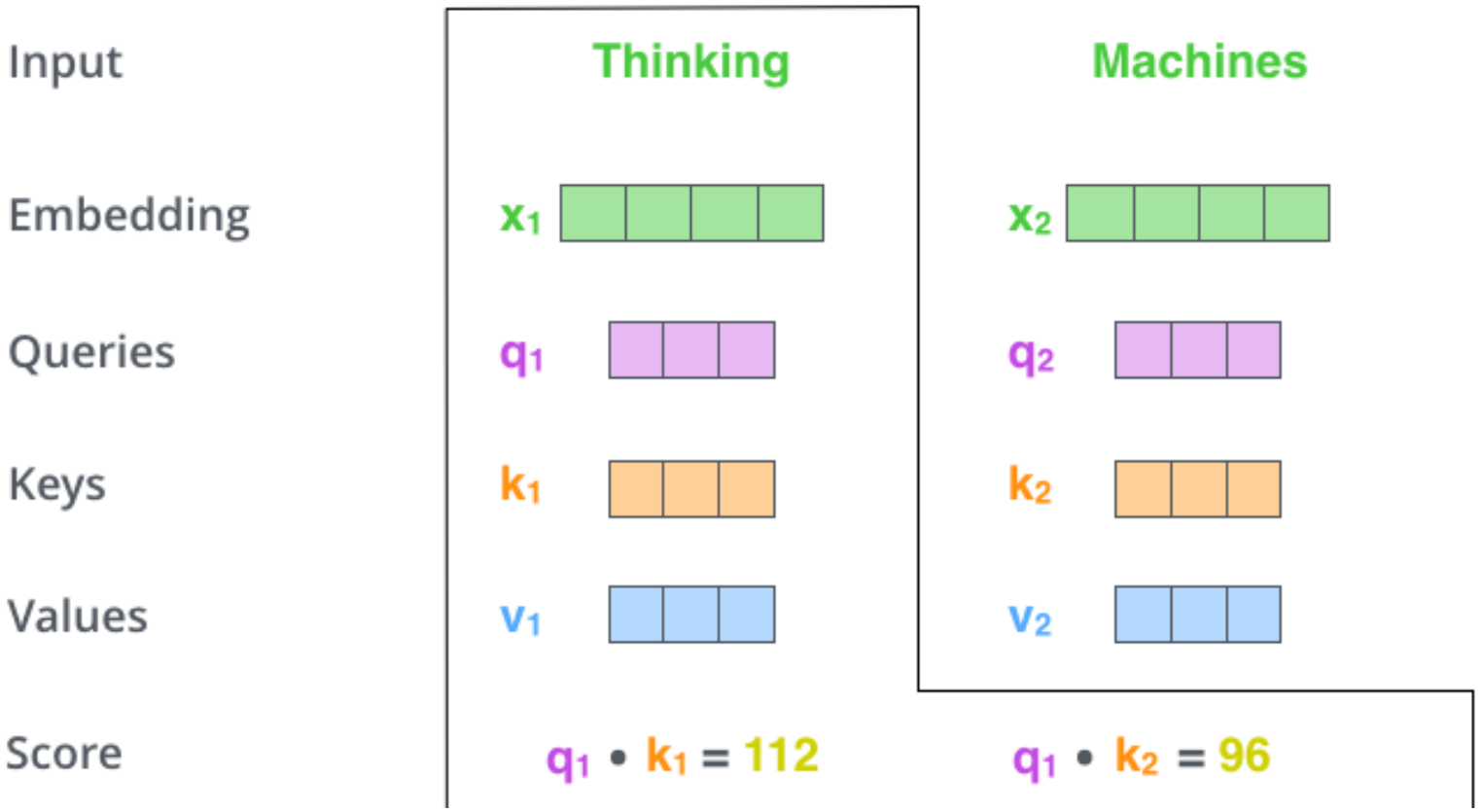
Auto-atención resulta ser el elemento clave



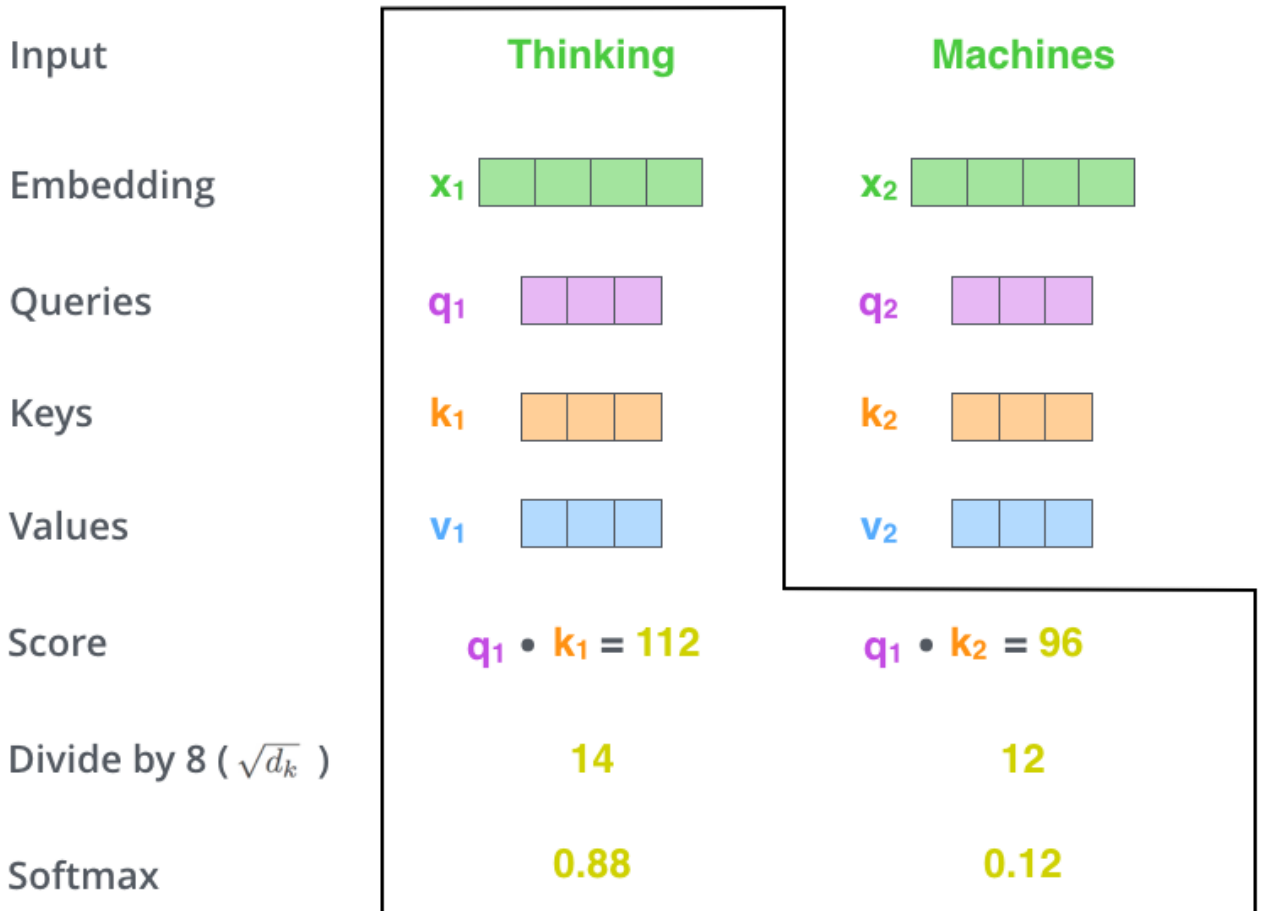
Auto-atención resulta ser el elemento clave



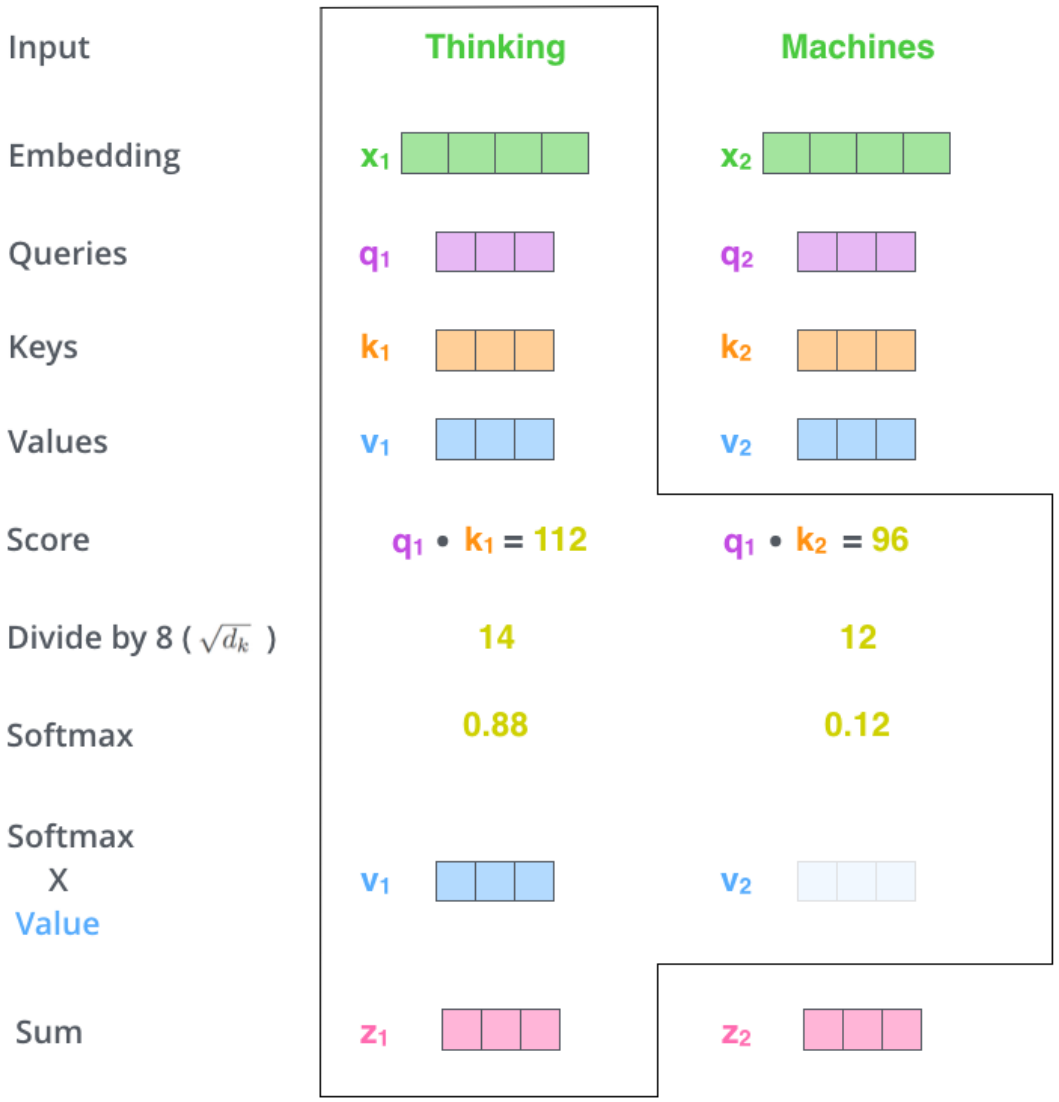
Auto-atención resulta ser el elemento clave



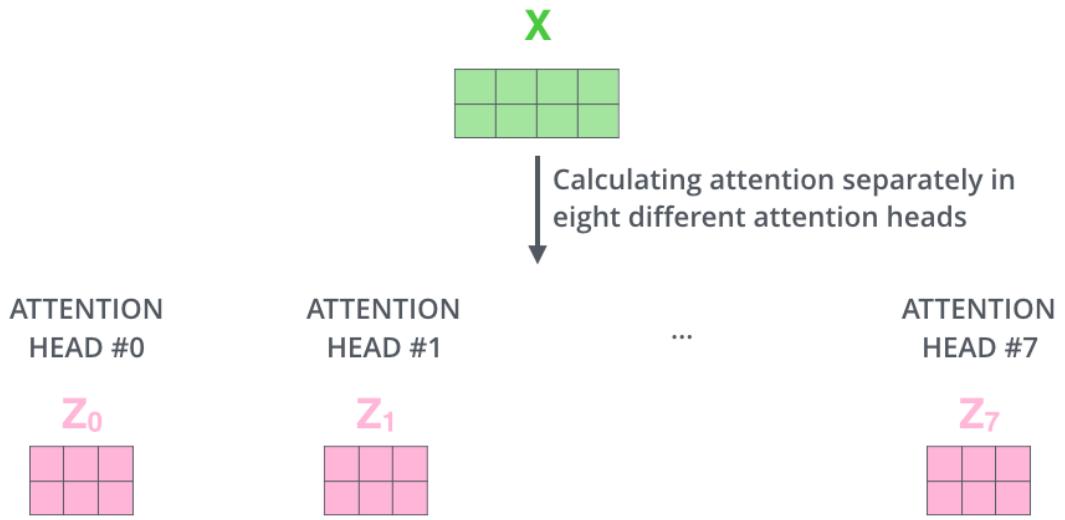
Auto-atención resulta ser el elemento clave



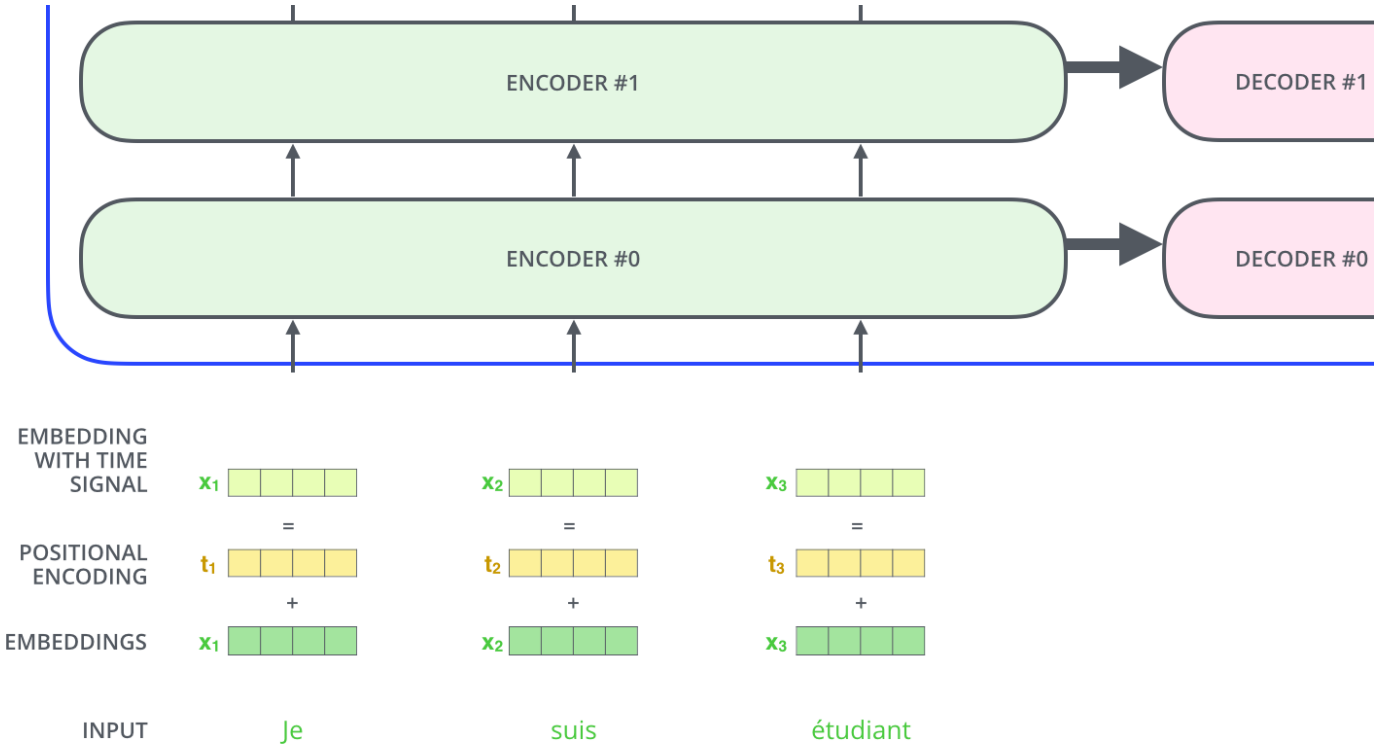
Auto-atención resulta ser el elemento clave



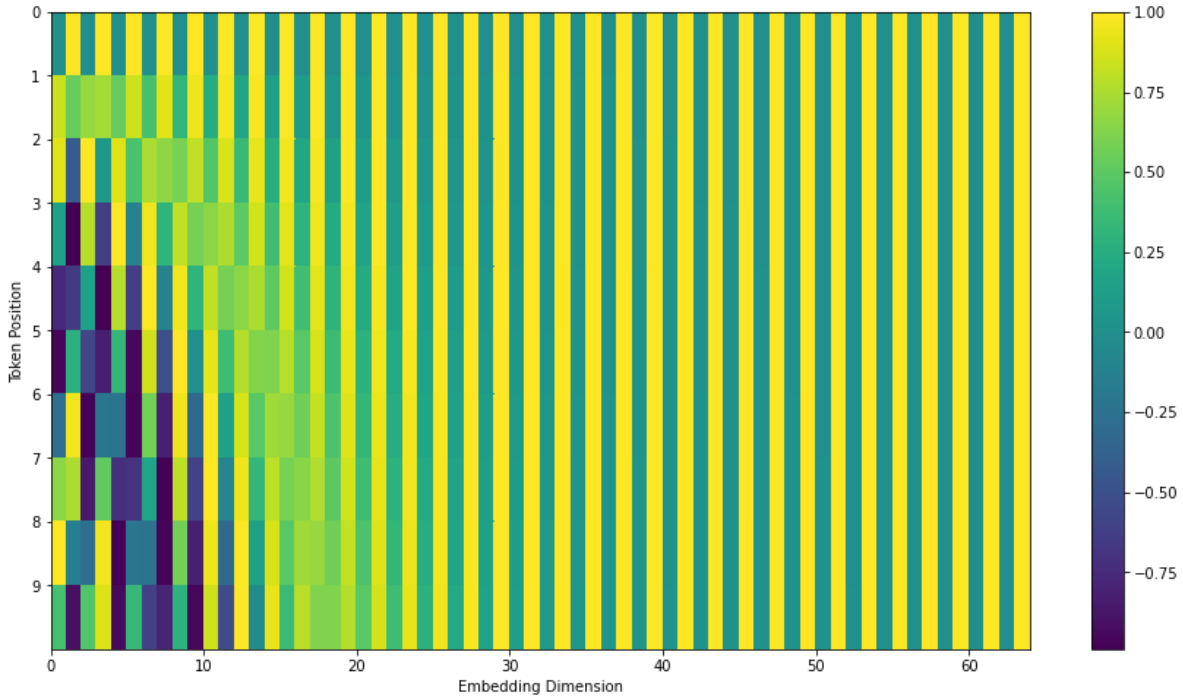
Auto-atención puede ser multimodal (muchas atenciones distintas)



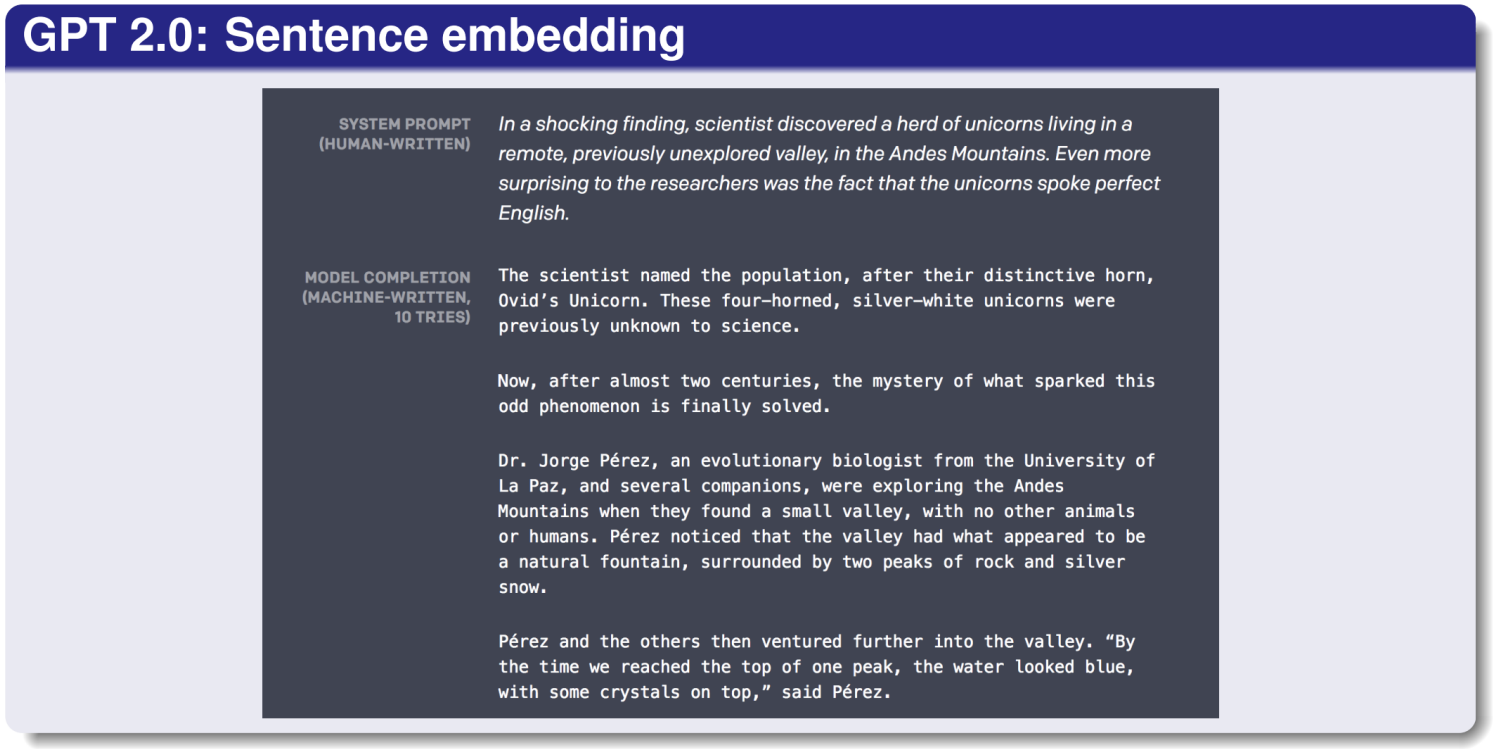
Orden de las secuencias es incluye a través de un *embedding* temporal



Orden de las secuencias es incluye a través de un *embedding* temporal



Algunos ejemplos para cerrar: GPT (1, 2, 3,…)

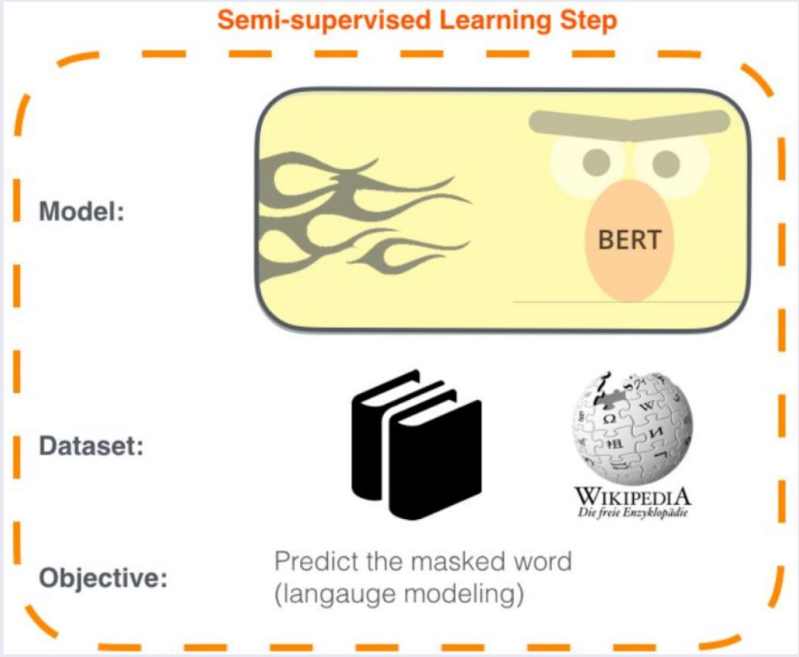


- GPT-2 tiene 1.5 billones de parámetros
- Entrenado usando 8 millones de sitios web

https://github.com/openai/gpt-2

Algunos ejemplos para cerrar: Bert

Bert: Sentence embedding



https://github.com/google-research/bert

Algunos ejemplos para cerrar: Bert

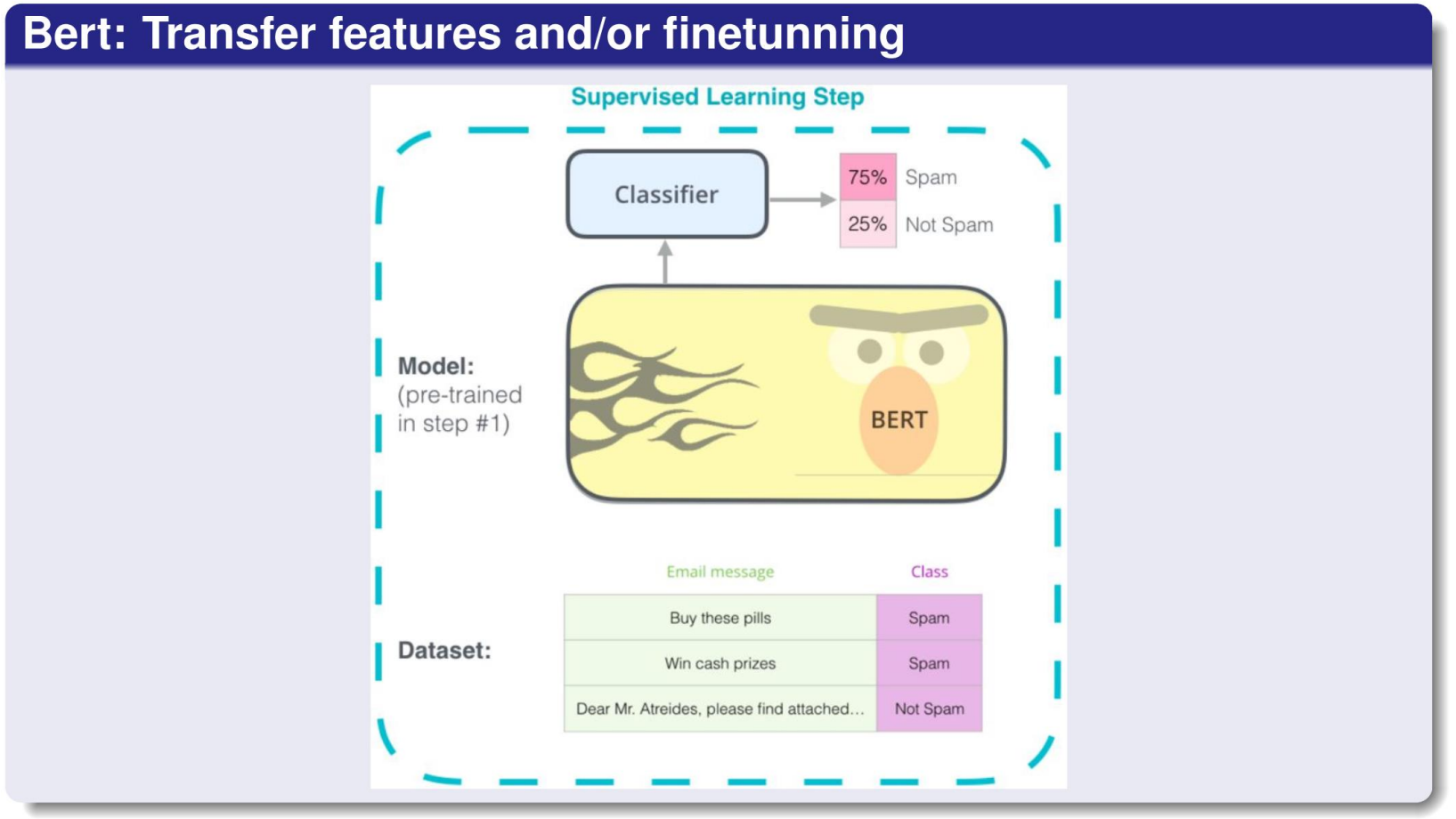
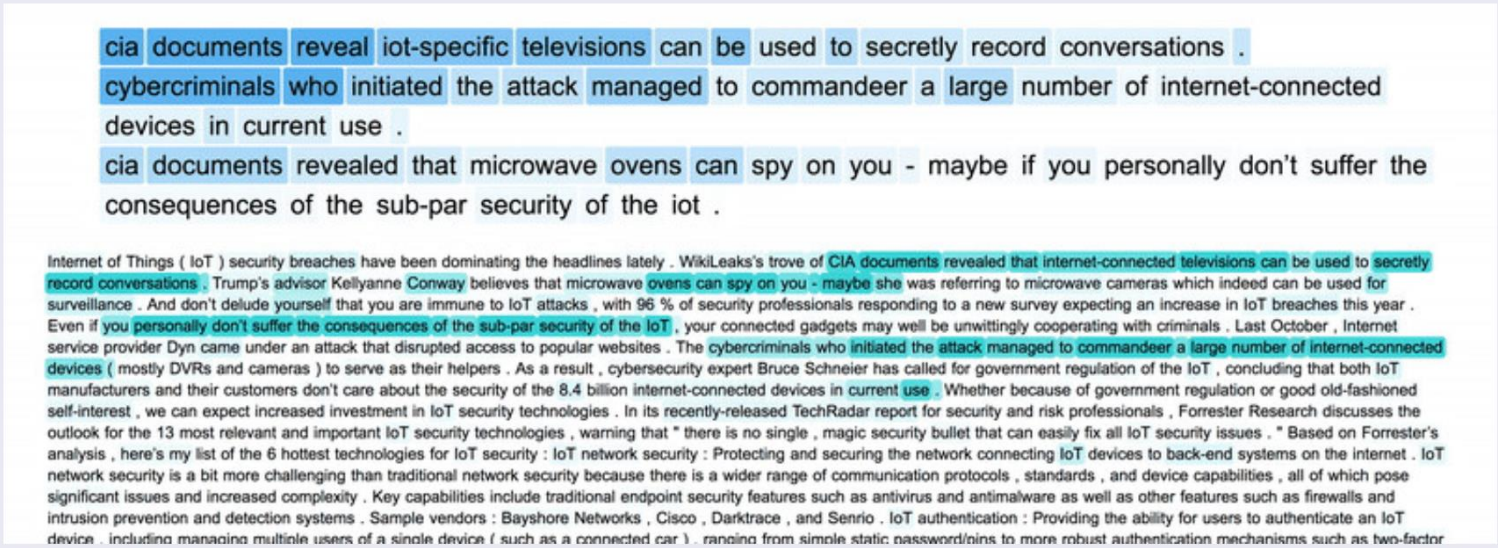


Figure from: http://jalammar.github.io/illustrated-bert/

Algunos ejemplos para cerrar: Bert

Text summarization using Bert



https://github.com/nlpyang/PreSumm

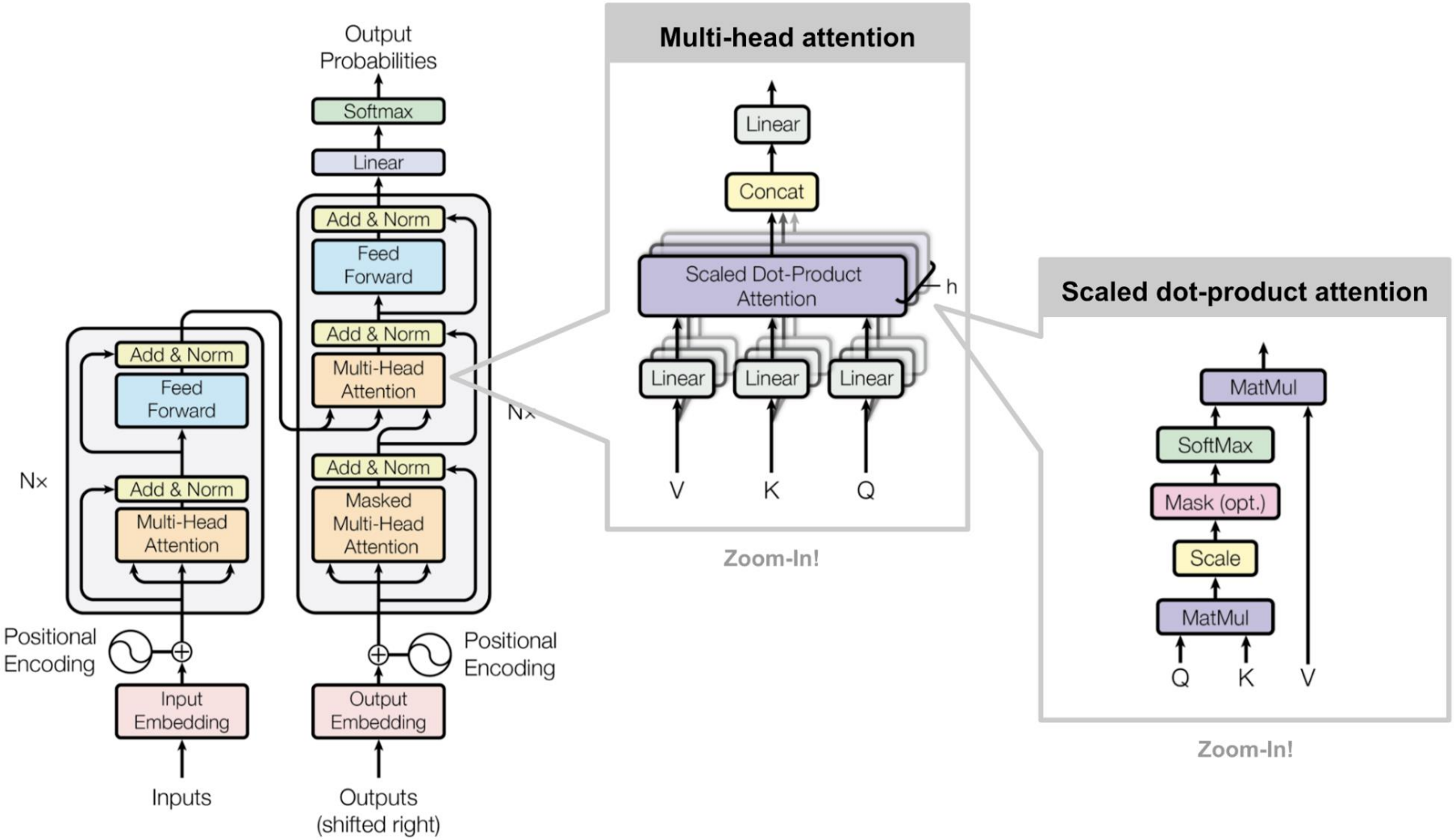
Algunos ejemplos para cerrar: Beto

Beto: Bert's model trained with a Spanich corpus.

Task	BETO-cased	BETO-uncased	Best Multilingual BERT
POS	**98.97**	98.44	97.10 [2]
NER-C	**88.43**	82.67	87.38 [2]
MLDoc	95.60	**96.12**	95.70 [2]
PAWS-X	89.05	89.55	90.70 [8]
XNLI	**82.01**	80.15	78.50 [2]

https://github.com/dccuchile/beto

Antes de terminar, así se ve realmente un Transformer



Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

Sistemas Urbanos Inteligentes

Autoatención

Hans Löbel
Dpto. Ingeniería de Transporte y Logística
Dpto. Ciencia de la Computación