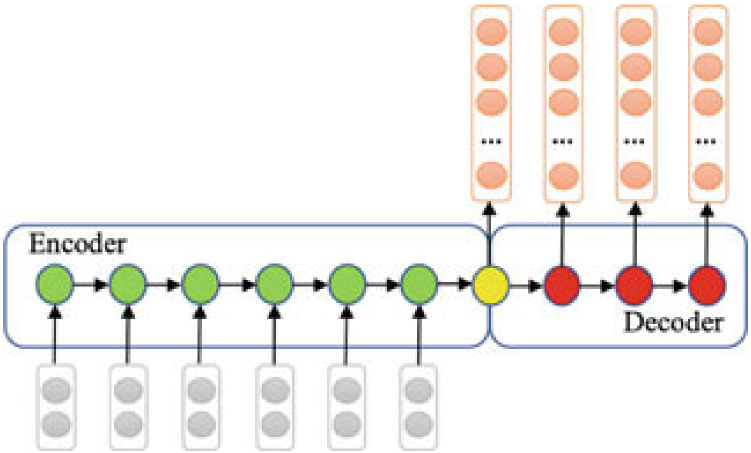
DECSAI
Departamento de Ciencias de la Computación e I.A.
Universidad de Granada

Mecanismos de atención

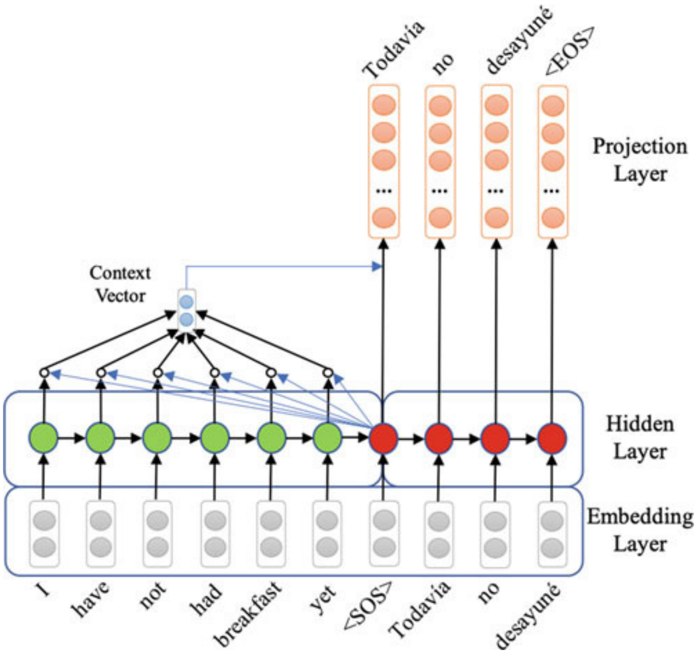
Fernando Berzal, berzal@acm.org

Motivación

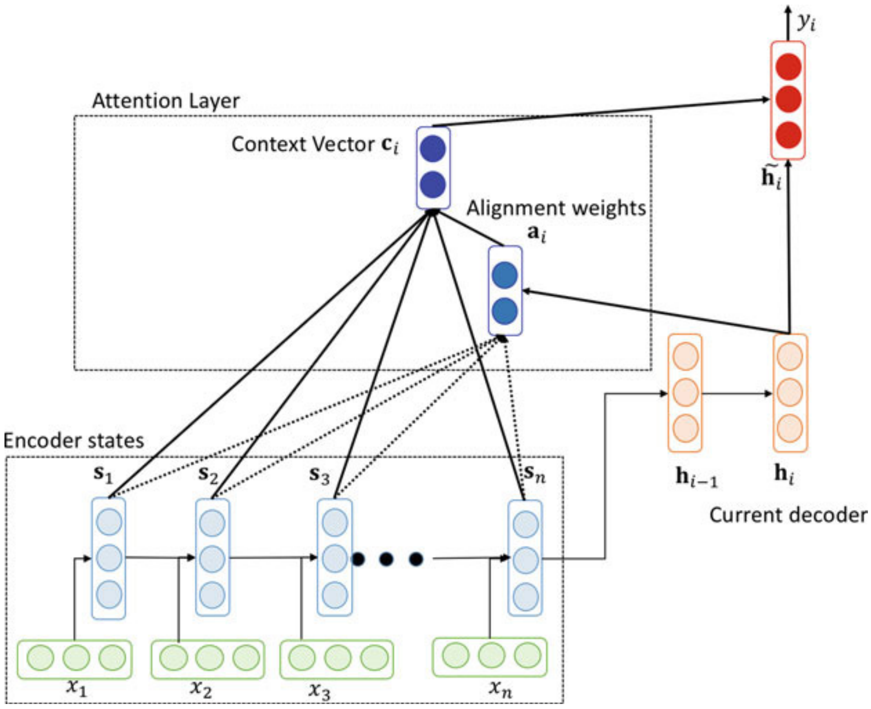
seq2seq



Motivación



Soft attention



Soft attention

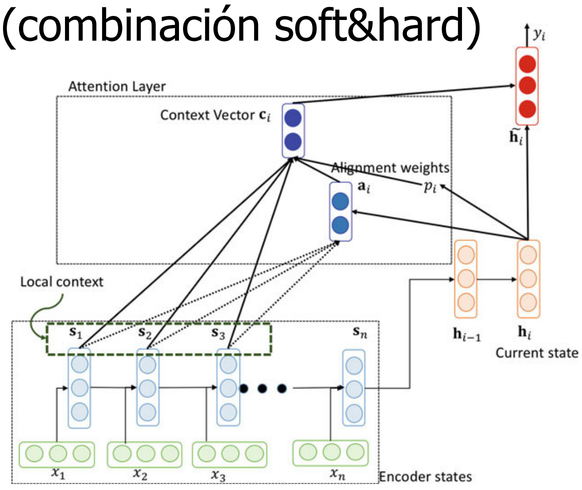
Score-based attention

Score name	Score description	Parameters
Concat (additive)	$\text{score}(\mathbf{s}_j, \mathbf{h}_i) = \mathbf{v}_a^\intercal \tanh(\mathbf{W}_a[\mathbf{s}_j; \mathbf{h}_i])$	$\mathbf{v}_a$ and $\mathbf{W}_a$ trainable
Linear (additive)	$\text{score}(\mathbf{s}_j, \mathbf{h}_i) = \mathbf{v}_a^\intercal \tanh(\mathbf{W}_a \mathbf{s}_j + \mathbf{U}_a \mathbf{h}_i)$	$\mathbf{v}_a$, $\mathbf{U}_a$, and $\mathbf{W}_a$ trainable
Bilinear (multiplicative)	$\text{score}(\mathbf{s}_j, \mathbf{h}_i) = \mathbf{h}_i^\intercal \mathbf{W}_a \mathbf{s}_j$	$\mathbf{W}_a$ trainable
Dot (multiplicative)	$\text{score}(\mathbf{s}_j, \mathbf{h}_i) = \mathbf{h}_i^\intercal \mathbf{s}_j$	No parameters
Scaled dot (multiplicative)	$\text{score}(\mathbf{s}_j, \mathbf{h}_i) = \frac{\mathbf{h}_i^\intercal \mathbf{s}_j}{\sqrt{n}}$	No parameters
Location-based	$\text{score}(\mathbf{s}_j, \mathbf{h}_i) = \text{softmax}(\mathbf{W}_a \mathbf{h}_i^\intercal)$	$\mathbf{W}_a$ trainable

Variantes

Selecciona puntos concretos en lugar de ir acumulando la secuencia de entrada en un vector de contexto…

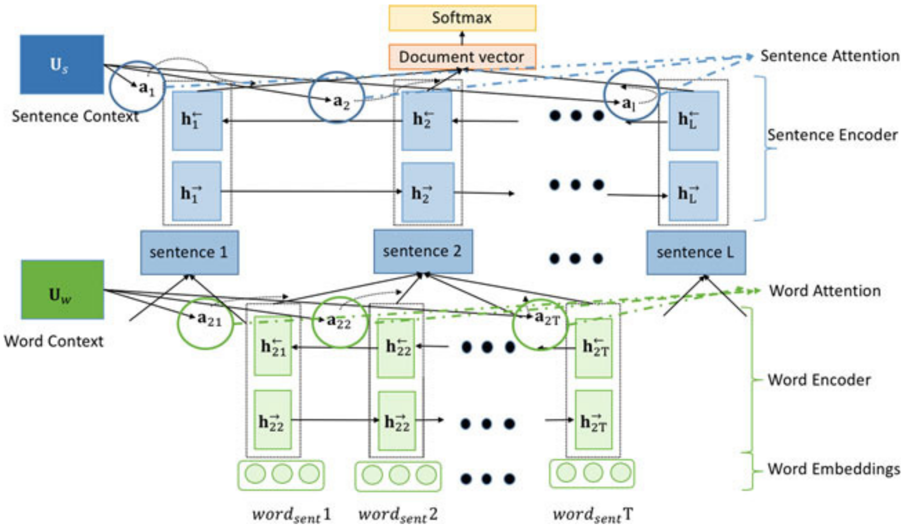
- Pointer networks
- Hard attention
- Local attention (combinación soft&hard)



Variantes

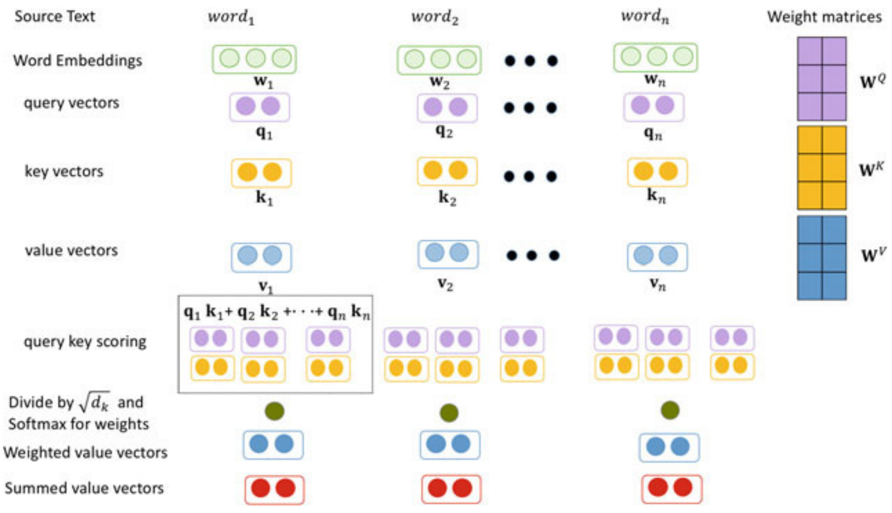
Atención jerárquica

Sentencias/palabras para clasificación de documentos



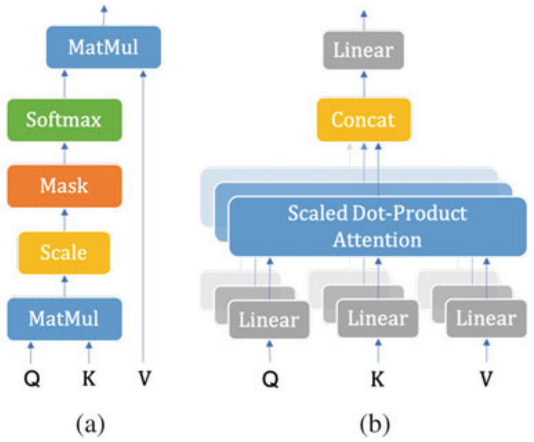
Variantes

Self-attention / intra-attention



Transformer networks

Aplica el mecanismo de atención directamente sobre la entrada, reduciendo o incluso eliminando la necesidad de conexiones recurrentes en la red neuronal



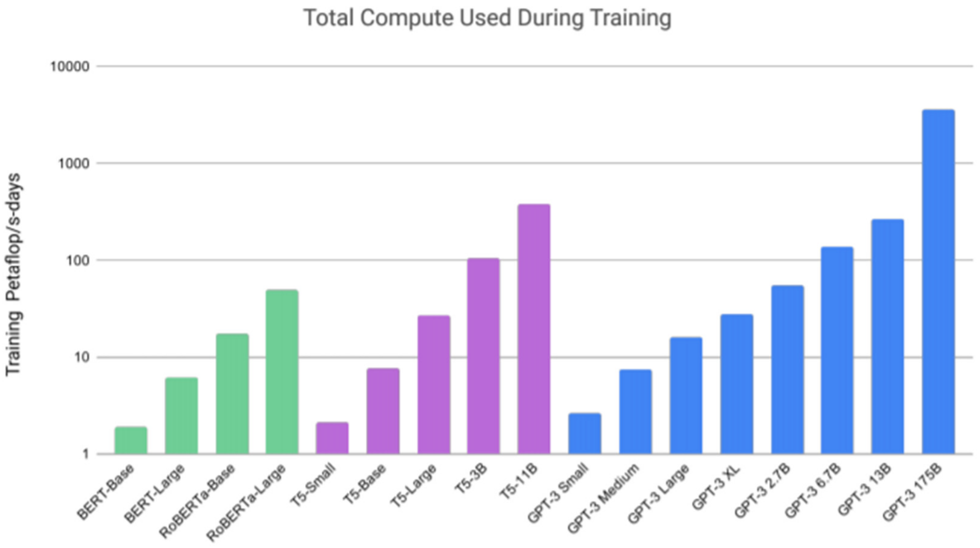
Scaled dot-product & multihead attention

Transformer networks

- BERT, Google, 2018 340M parameters
  [Bidirectional Encoder Representations from Transformers]
- GPT, OpenAI, 2018 110M parameters
  [Generative Pre-trained Transformer]
- MT-DNN, Microsoft, 2019 330M parameters
  [MultiTask Deep Neural Network]
- Transformer ELMo, AI2, 2019 465M parameters
- GPT-2, OpenAI, 2019 1.5B parameters
- MegatronLM, NVIDIA, 2019 8.3B parameters
- T5, Google, 2020 11B parameters
  [Text-to-Text Transfer Transformer]
- Turing-NLG, Microsoft, 2020 17B parameters
- GPT-3, OpenAI, 2020 175B parameters !!!

Transformer networks

1 petaflop/s·day = 10^{15} ops/s for 1 day ≈ 10^{20} ops


Total Compute Used During Training
Training Petaflop/s-days
BERT-Base
BERT-Large
RoBERTa-Base
RoBERTa-Large
T5-Small
T5-Base
T5-Large
T5-3B
T5-11B
GPT-3 Small
GPT-3 Medium
GPT-3 Large
GPT-3 XL
GPT-3 2.7B
GPT-3 6.7B
GPT-3 13B
GPT-3 175B