Attention mechanism

杜岳華

2019.3.30

About me

- Julia Taiwan 社群發起人
- AI Tech 社群常規成員與講師
- 工研院 機器學習理論與實作 講師
- 著作：《Julia程式設計》

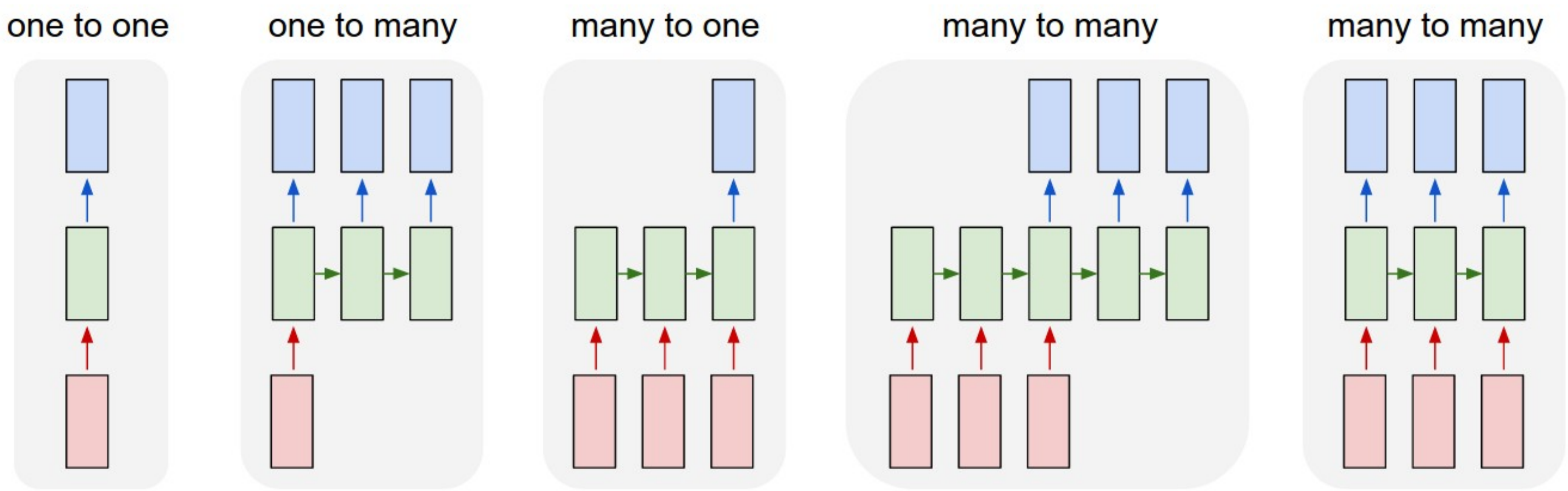
- 專長：系統生物學、計算生物學、機器學習
- 碩論：Identification of cell state using super-enhancer RNA

- 陽明 生物醫學資訊所 碩士
- 成大 雙主修 醫學檢驗生物技術學系 學士，資訊工程學系 學士

Outline

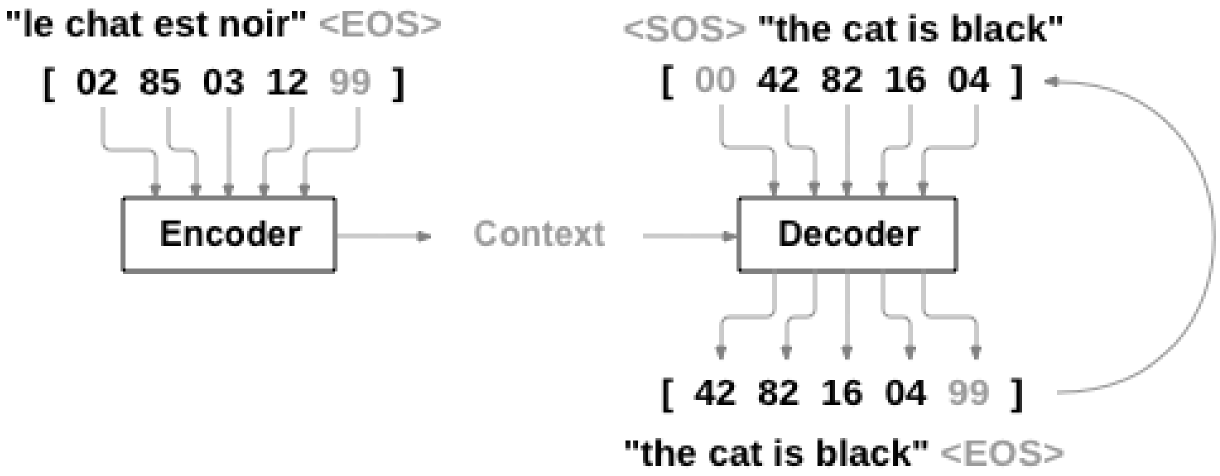
- RNN 的問題
- Seq2Seq encoder-decoder 架構
- Attention model 解決的問題
- Attention types
- Applications of attention
    - Translation
    - Summarization
    - Image caption
- Transformer

RNN 的問題



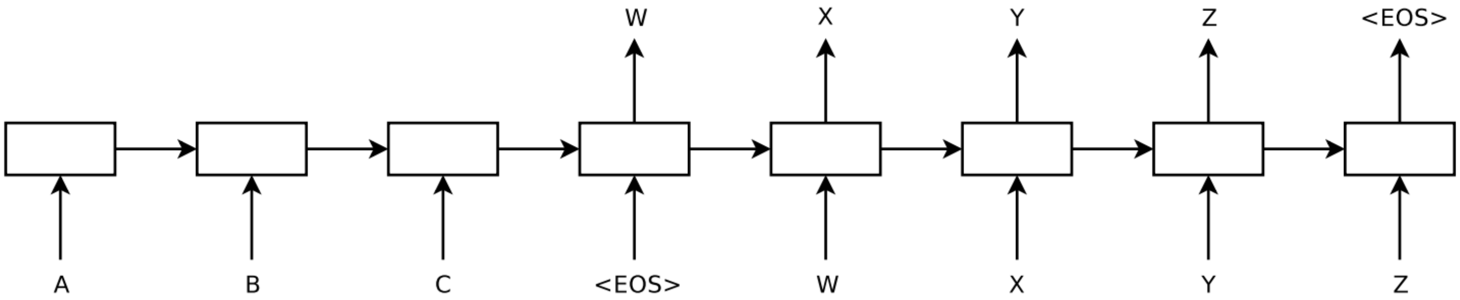
[picture source (http://karpathy.github.io/2015/05/21/rnn-effectiveness/)](http://karpathy.github.io/2015/05/21/rnn-effectiveness/)

Seq2Seq encoder-decoder 架構



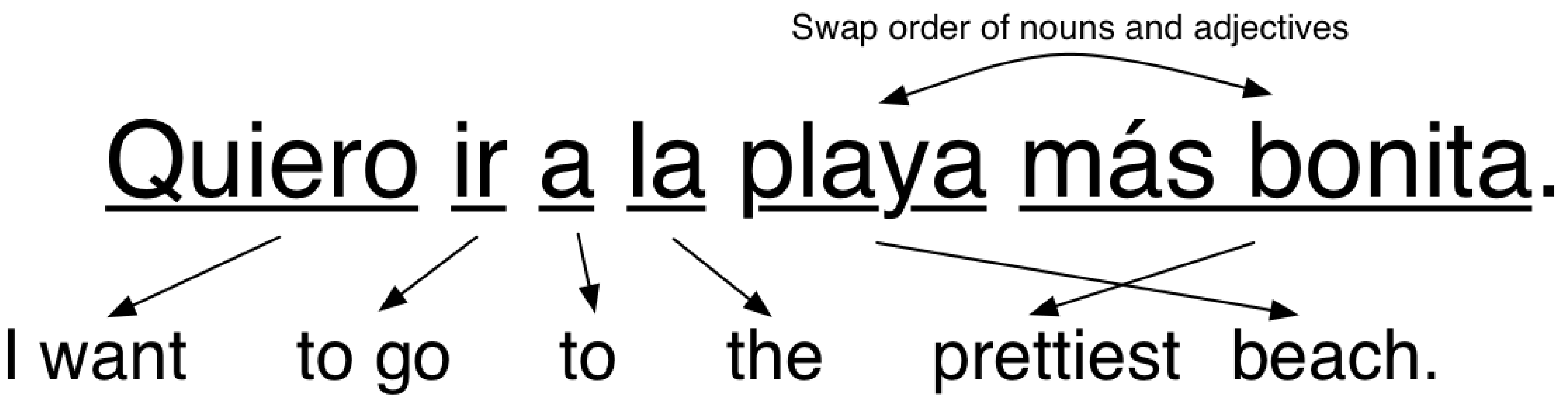
[picture source (https://pytorch.org/tutorials/intermediate/seq2seq_translation_tutorial.html)](https://pytorch.org/tutorials/intermediate/seq2seq_translation_tutorial.html)

Seq2Seq encoder-decoder 架構



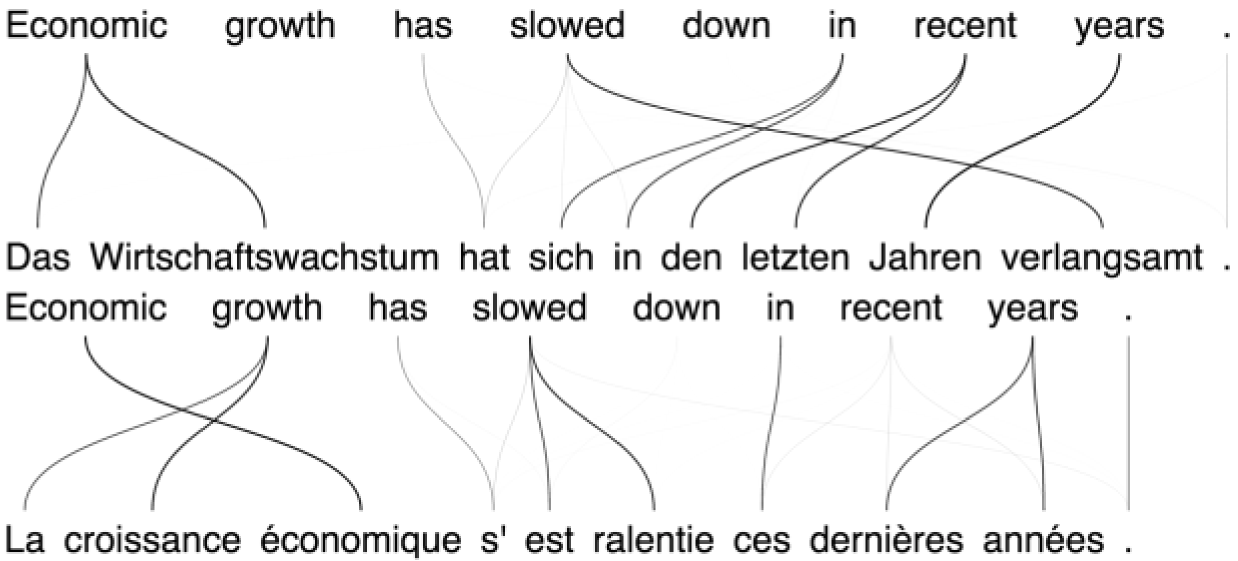
[*picture source (https://machinelearningmastery.com/encoder-decoder-recurrent-neural-network-models-neural-machine-translation/)*](https://machinelearningmastery.com/encoder-decoder-recurrent-neural-network-models-neural-machine-translation/)

Attention model 解決的問題



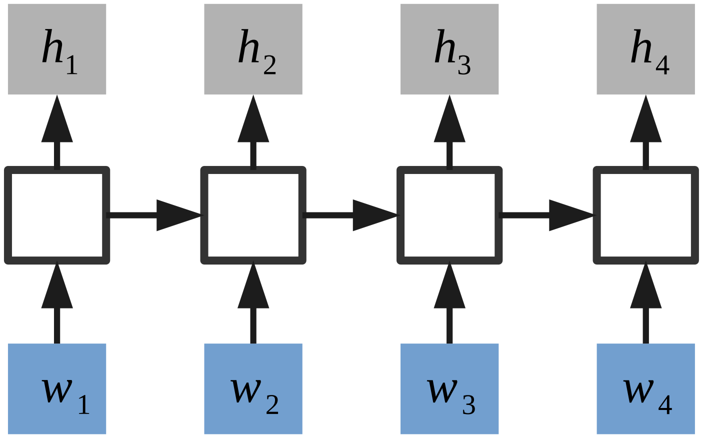
[picture source (https://medium.com/@ageitgey/machine-learning-is-fun-part-5-language-translation-with-deep-learning-and-the-magic-of-sequences-2ace0acca0aa)](https://medium.com/@ageitgey/machine-learning-is-fun-part-5-language-translation-with-deep-learning-and-the-magic-of-sequences-2ace0acca0aa)

How to solve the problem?

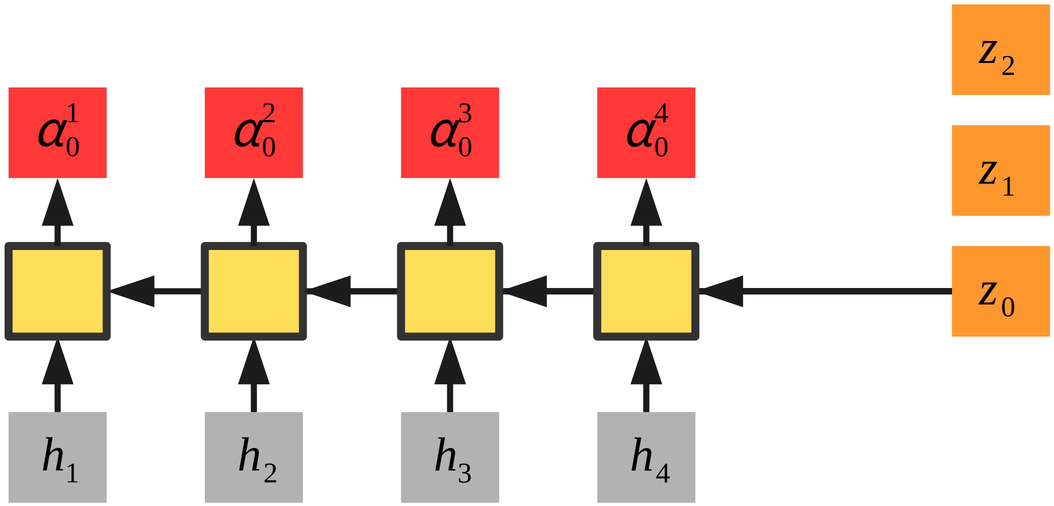


[picture source (https://devblogs.nvidia.com/introduction-neural-machine-translation-gpus-part-3/)](https://devblogs.nvidia.com/introduction-neural-machine-translation-gpus-part-3/)

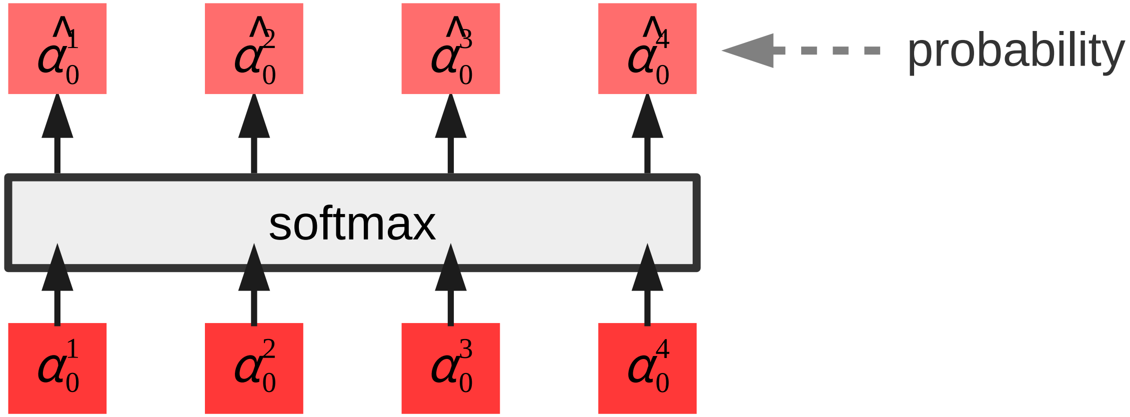
Attention mechanism



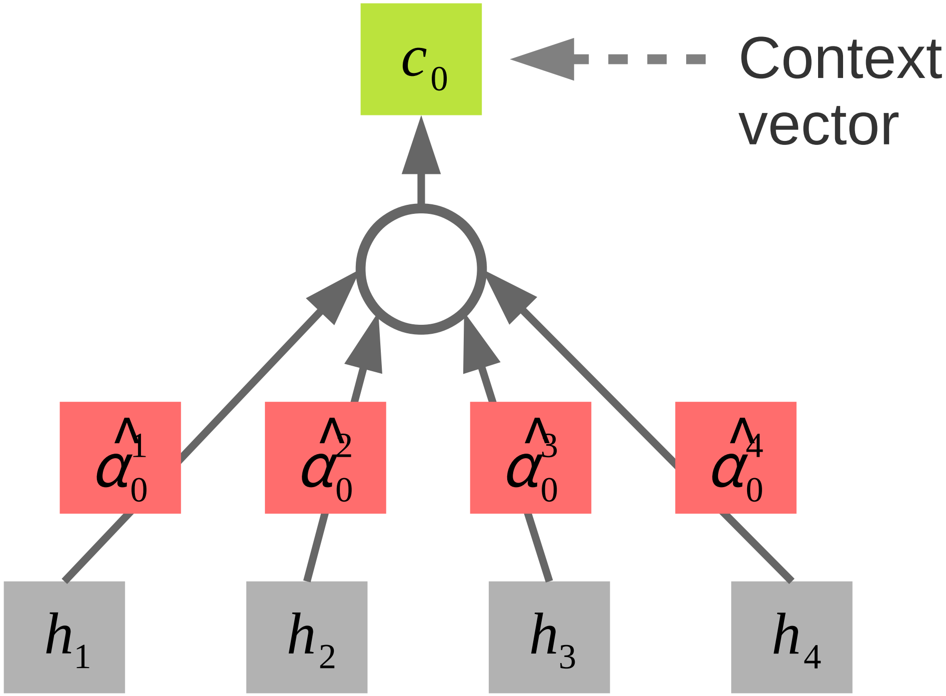
Attention mechanism



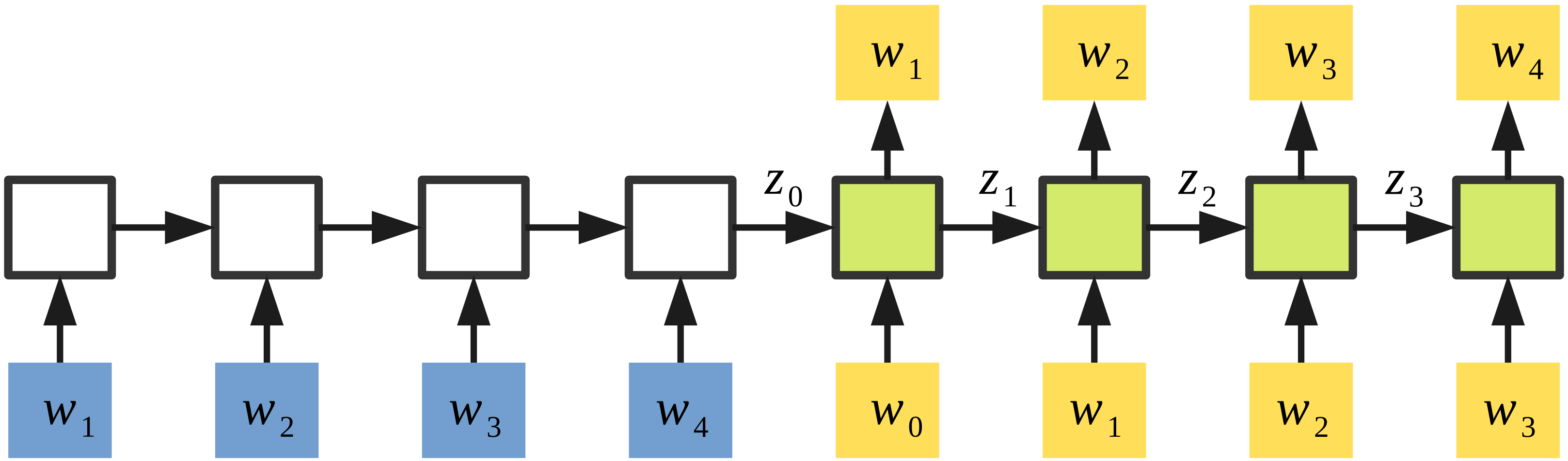
Attention mechanism



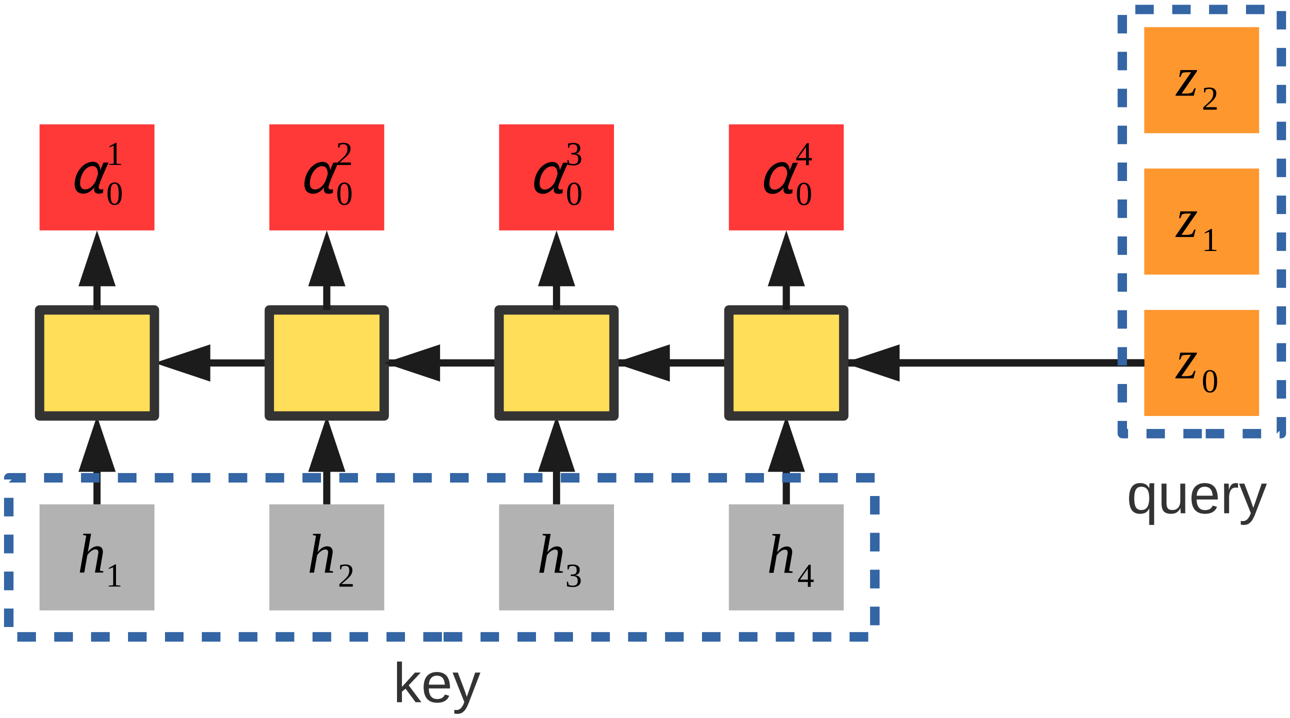
Attention mechanism



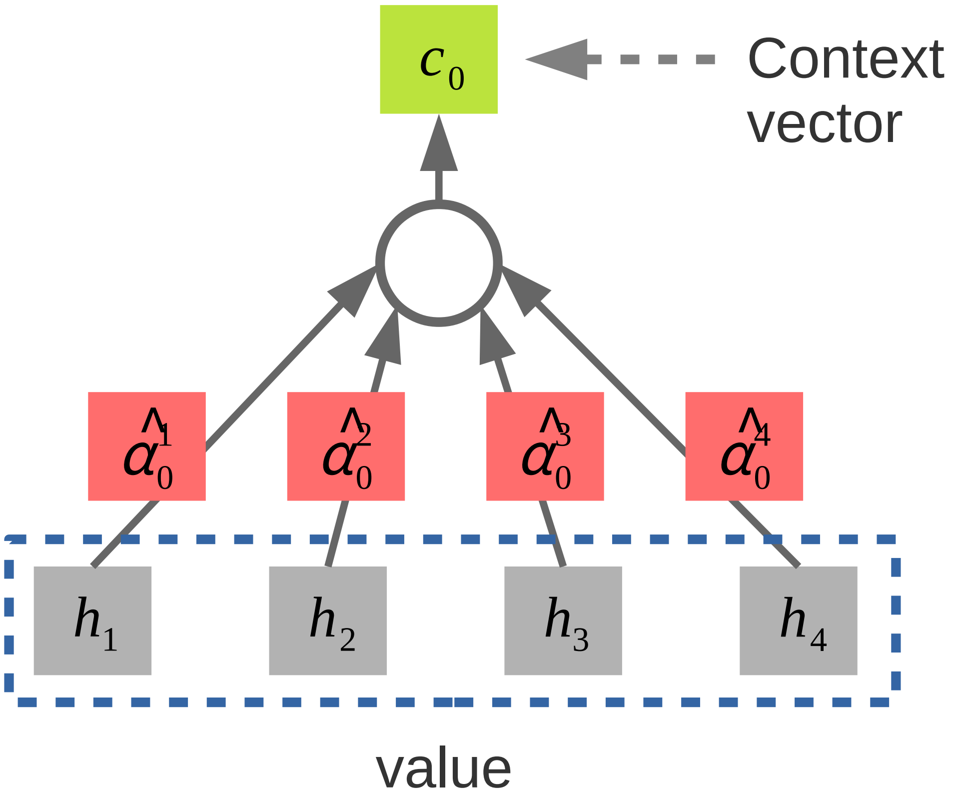
Attention mechanism



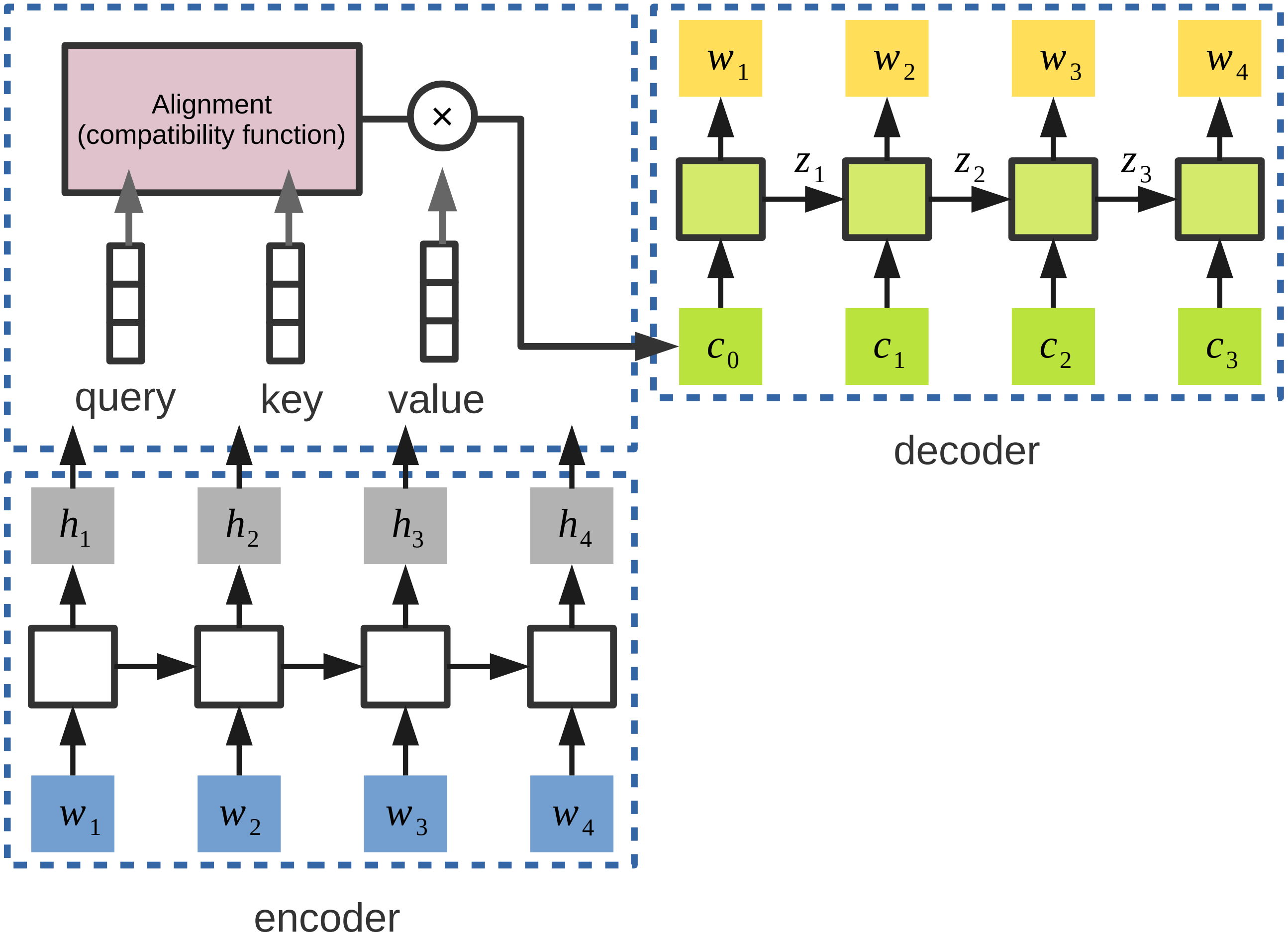
Attention mechanism



Attention mechanism



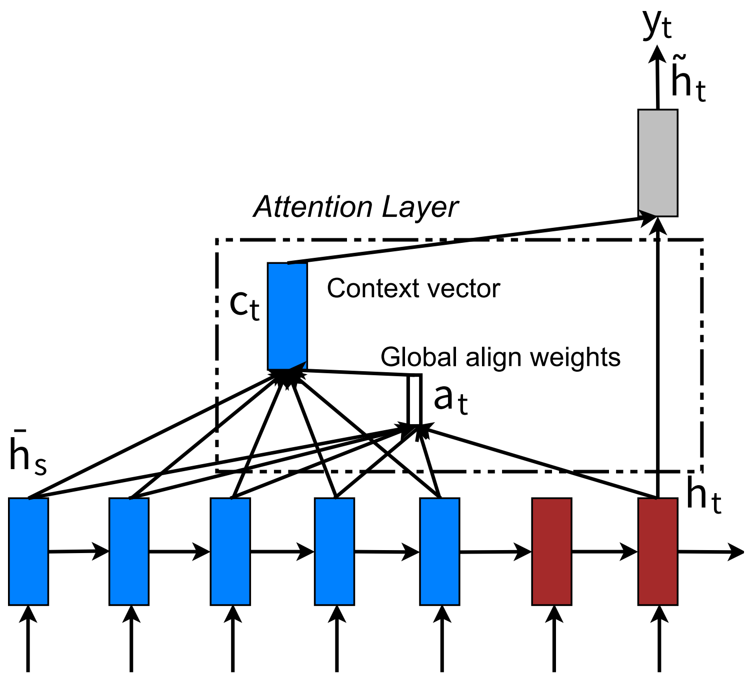
Attention mechanism



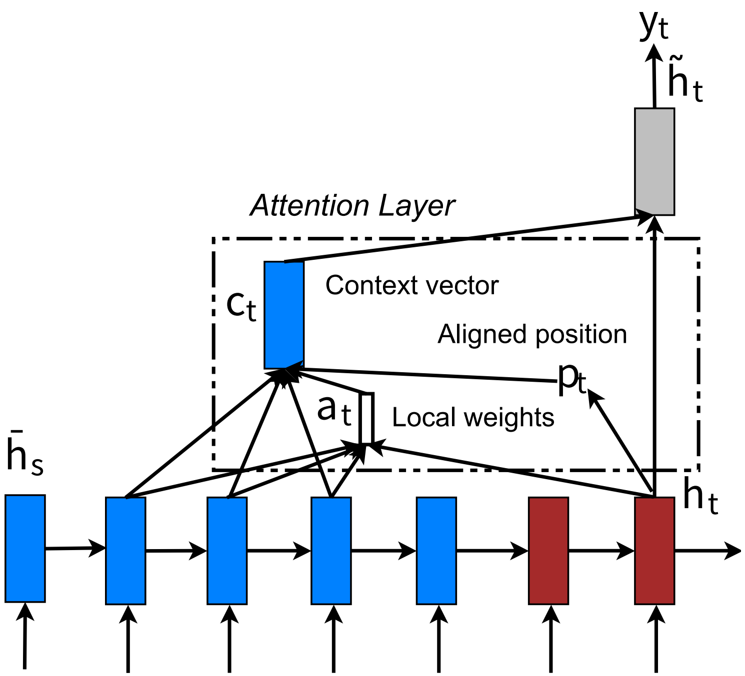
Attention types

- Global/local attention
- Hard/soft attention
- Self-attention

Global/local attention



Global attention



Local attention

Hard/soft attention

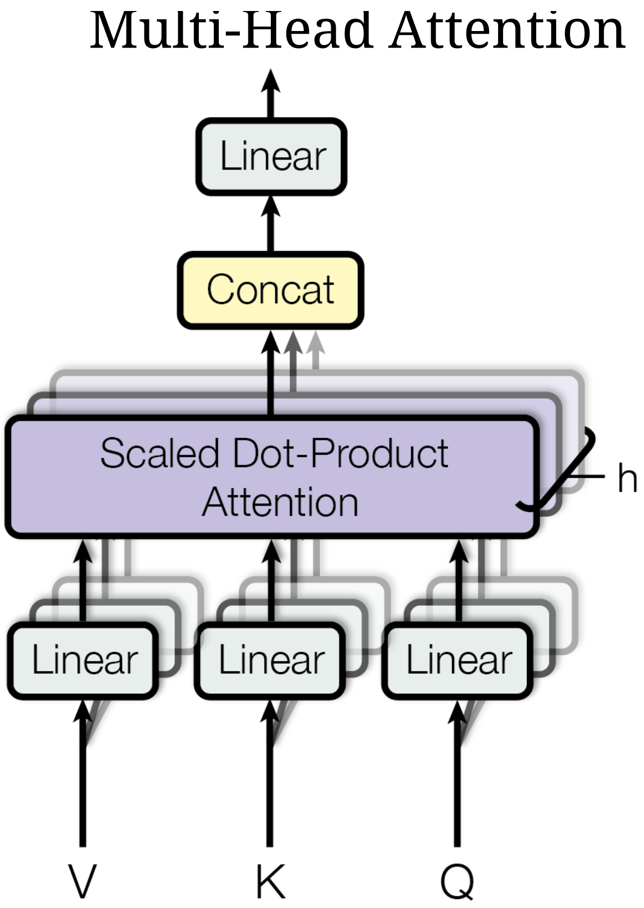
Soft attention

- Alignment weights are learned to attend over all data
- $0 \leq w \leq 1$
- Pro: model is smooth and differentiable
- Con: large computation if input is large

Hard attention

- Select part of data to attend or not at a time
- 0 or 1
- Pro: less inference time
- Con: model is non-differentiable

Self-attention



Alignment (compatibility function)

query: q_j, key: k_i

Location-based

$$\alpha_j^i = softmax(W_\alpha q_j)$$

Content-based

$$score(q_j, k_i) = cos([q_j; k_i])$$

Additive

$$score(q_j, k_i) = v_\alpha^T tanh(W_\alpha [q_j; k_i])$$

Alignment (compatibility function)

General

$$score(q_j, k_i) = q_j^T W_\alpha k_i$$

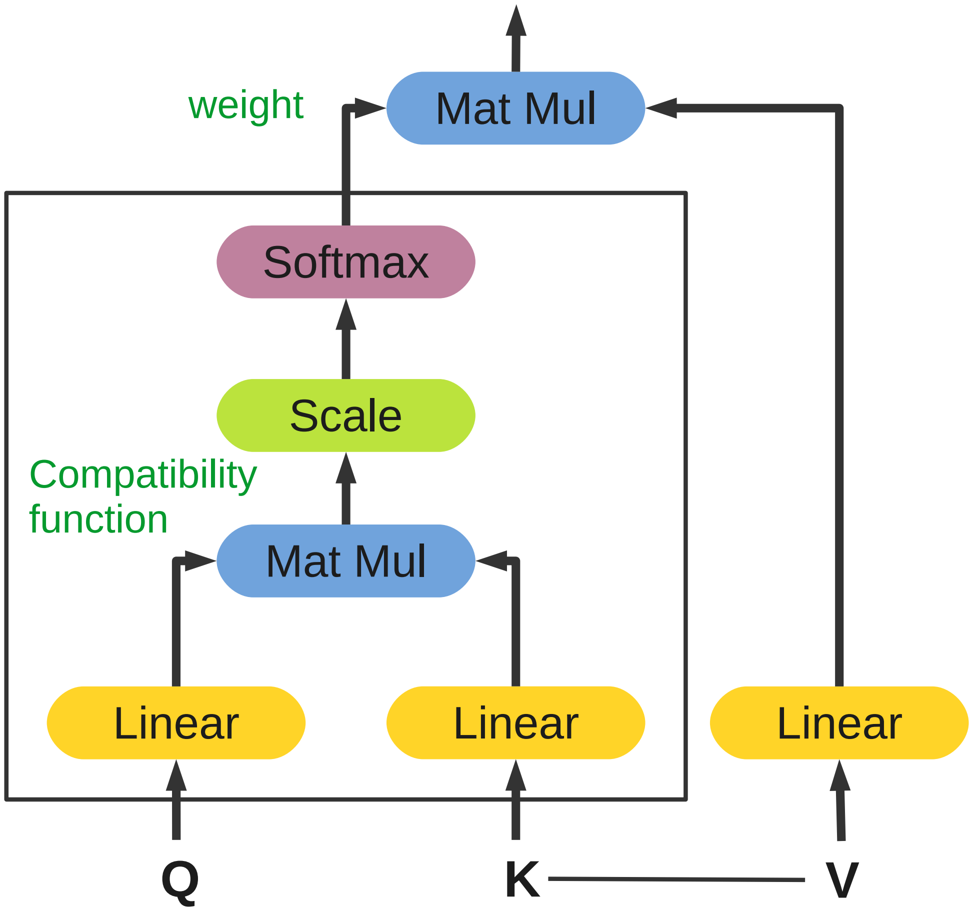
Dot-product

$$score(q_j, k_i) = q_j^T k_i$$

Scaled dot-product

$$score(q_j, k_i) = \frac{q_j^T k_i}{\sqrt{n}}$$

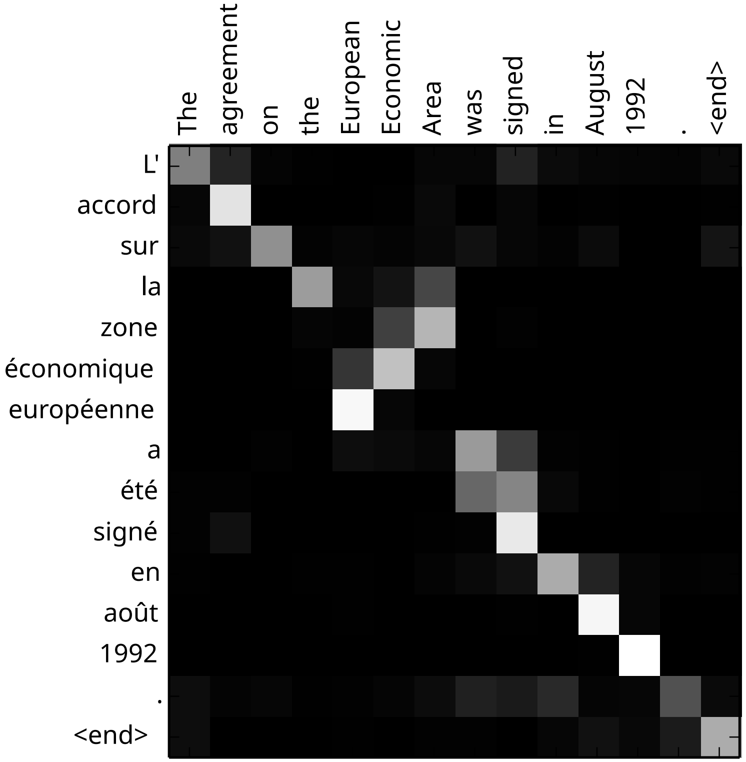
Scaled dot-product attention



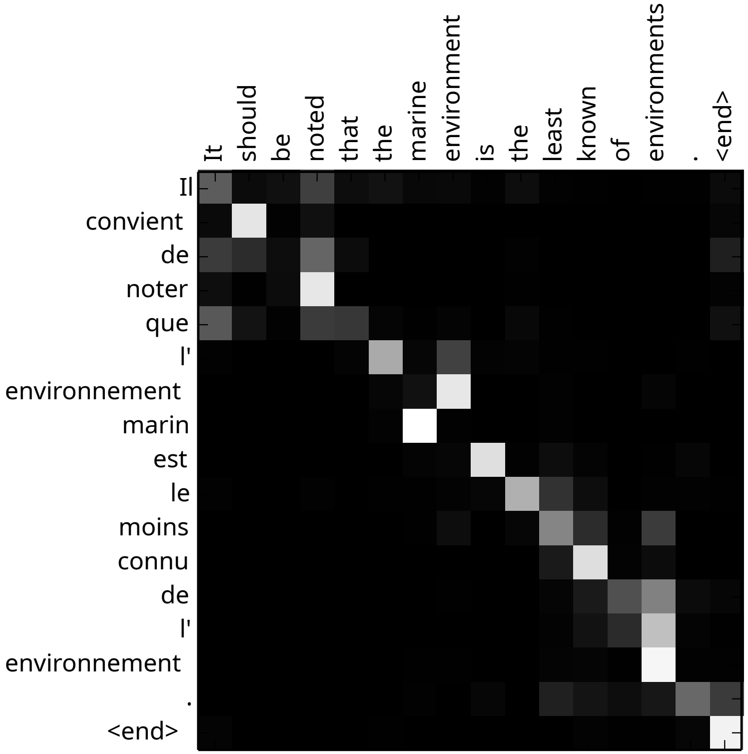
Applications of attention

- Summarization: [Rush 2015 (https://arxiv.org/abs/1509.00685)](https://arxiv.org/abs/1509.00685)
- Translation: [Bahdanau 2014 (https://arxiv.org/abs/1409.0473)](https://arxiv.org/abs/1409.0473), [Luong 2015 (https://arxiv.org/abs/1508.04025)](https://arxiv.org/abs/1508.04025)
- Image caption: [Xu 2015 (https://arxiv.org/abs/1502.03044)](https://arxiv.org/abs/1502.03044)
- ...

Translation

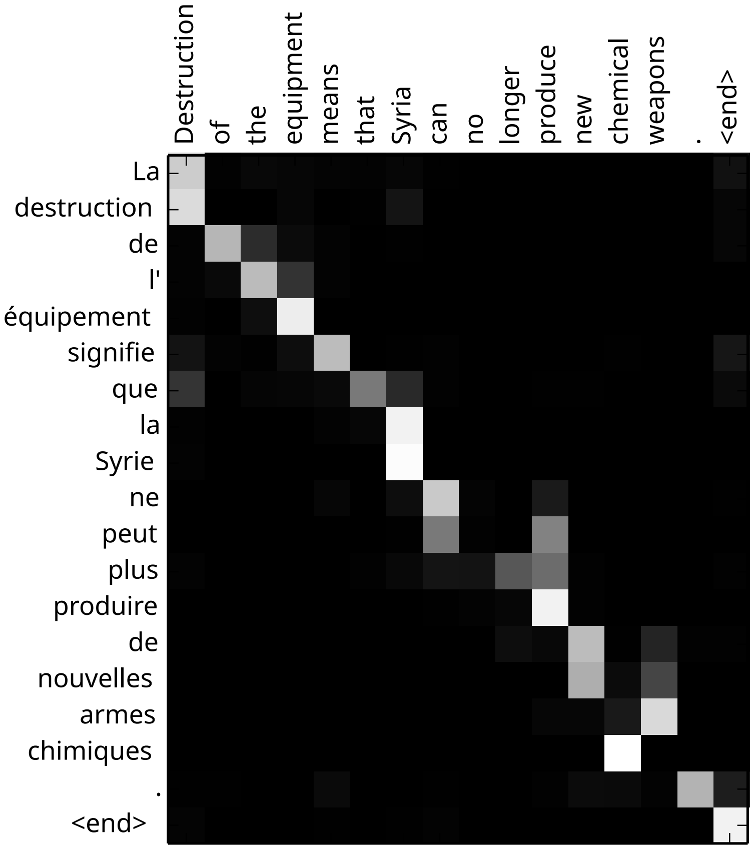


(a)

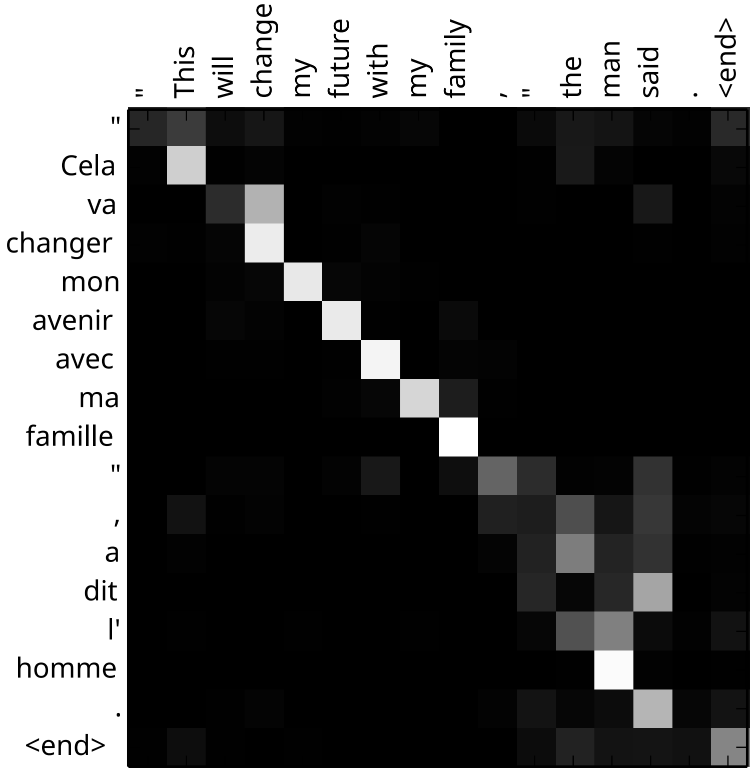


(b)

Translation



(c)



(d)

Summarization

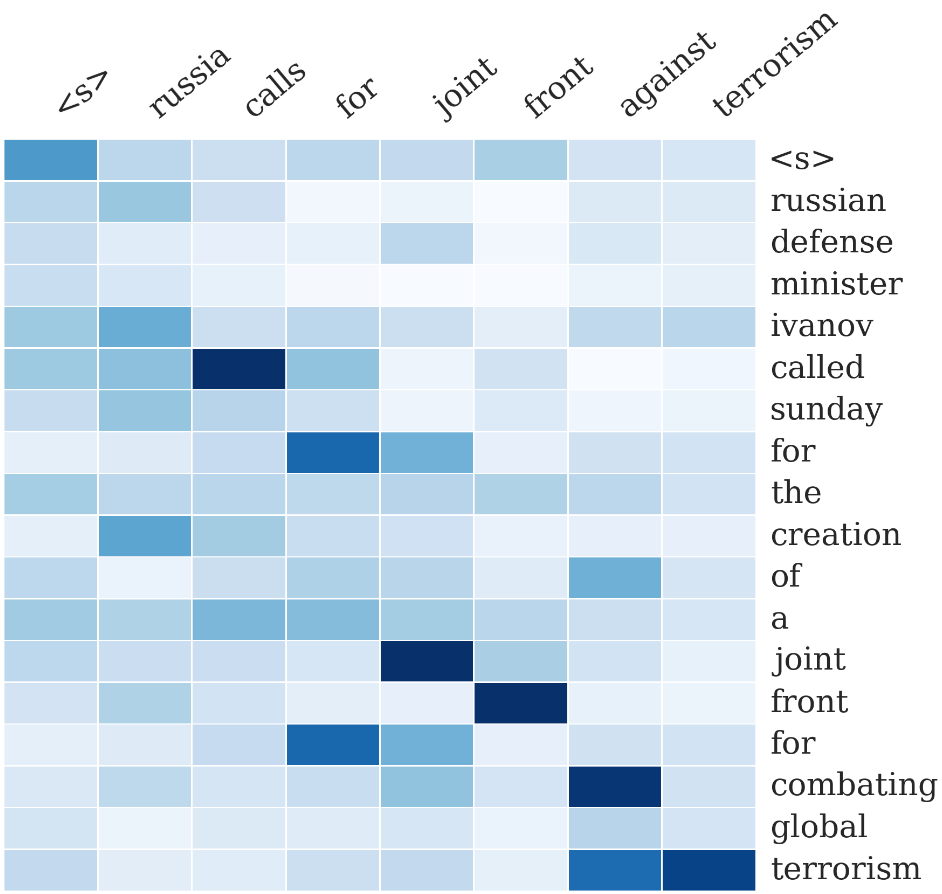
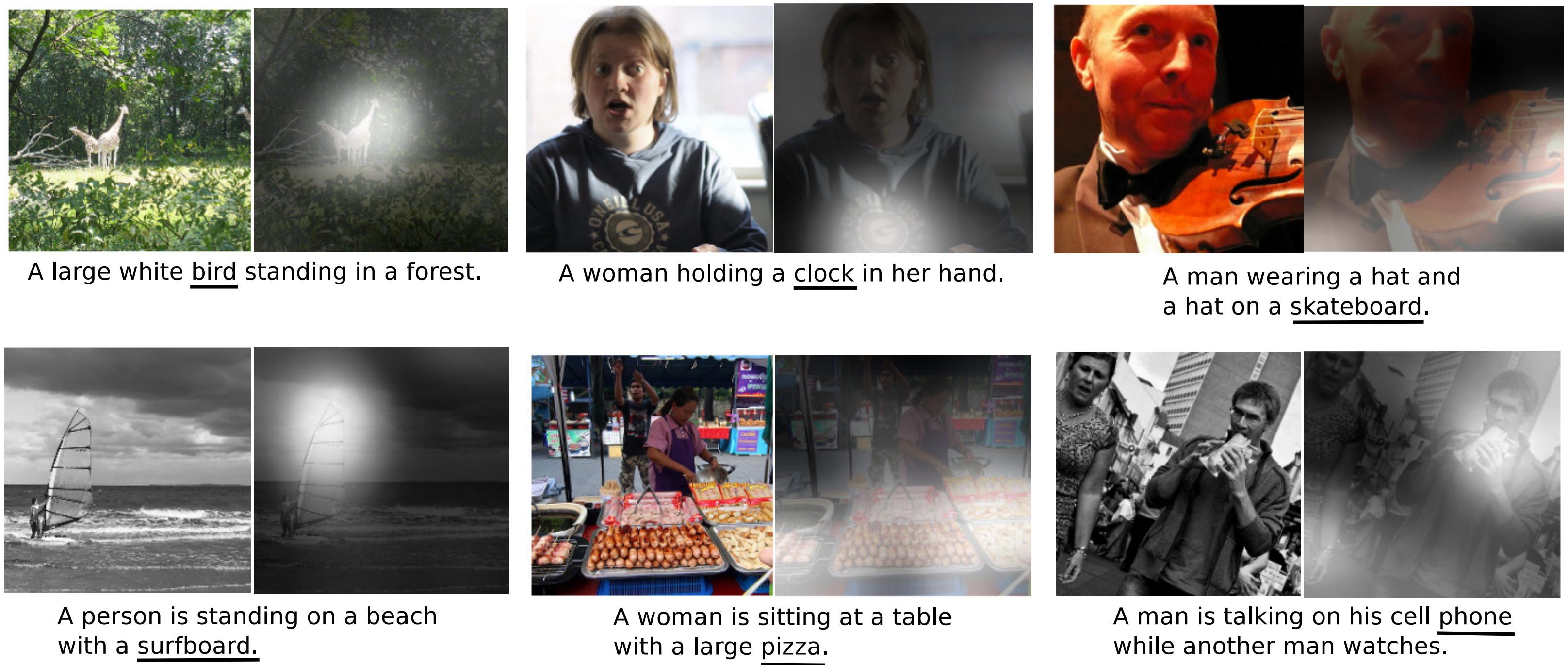
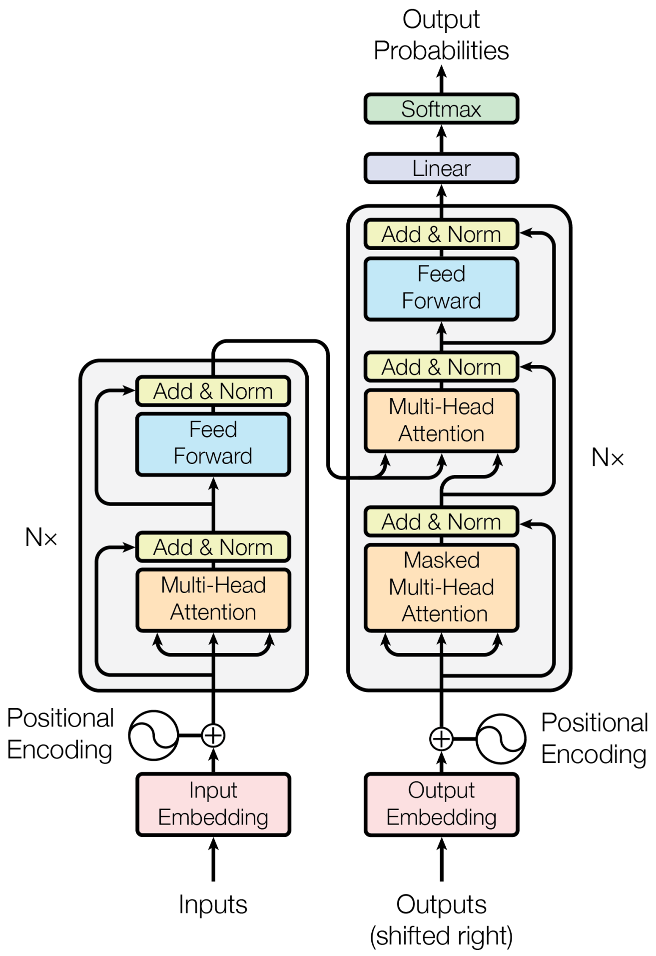


Image caption

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



Transformer



The era of Transformer

Why self-attention?

- More or equal efficient than RNN/CNN

Integration of RNN/CNN into Transformer

Thank you for attention.

References

- [Attention? Attention! (https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html)](https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html)
- [放棄幻想，全面擁抱 Transformer：自然語言處理三大特徵抽取器（CNN/RNN/TF）比較 (http://bangqu.com/f7t7X5.html)](http://bangqu.com/f7t7X5.html)

Papers