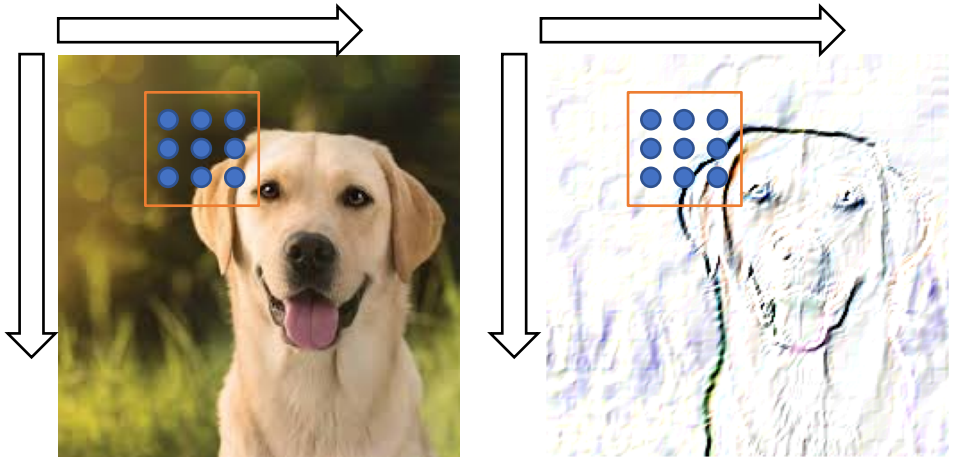
Attention in deep learning

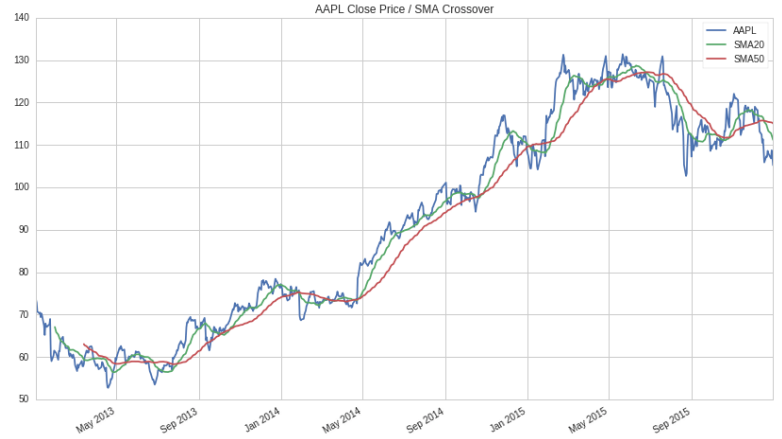
CNN



- Color, surface, posture
- Edge of the object against the background
- Edge of the eyes, noise, mouth...

→ a dog

RNN



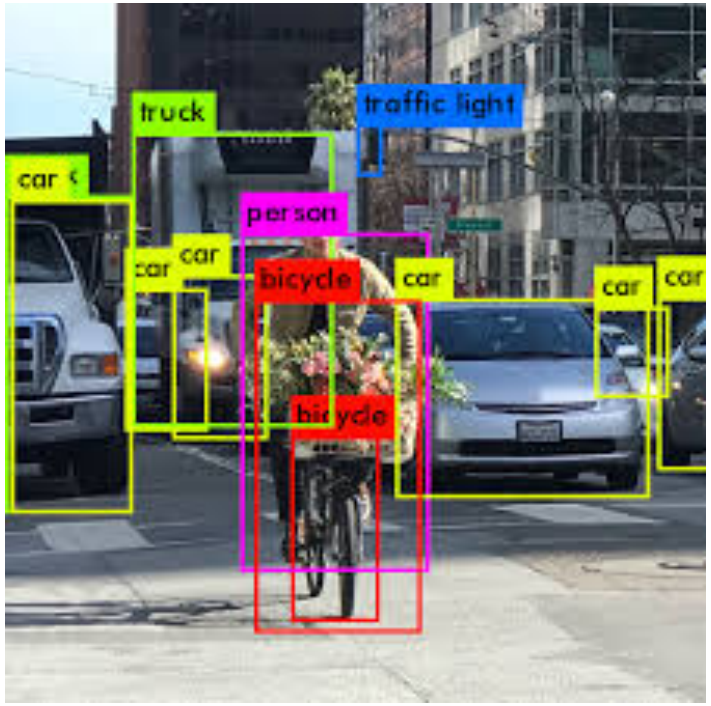
- Goes up at this state with this input
- Goes down at that state with that input

→ correlations

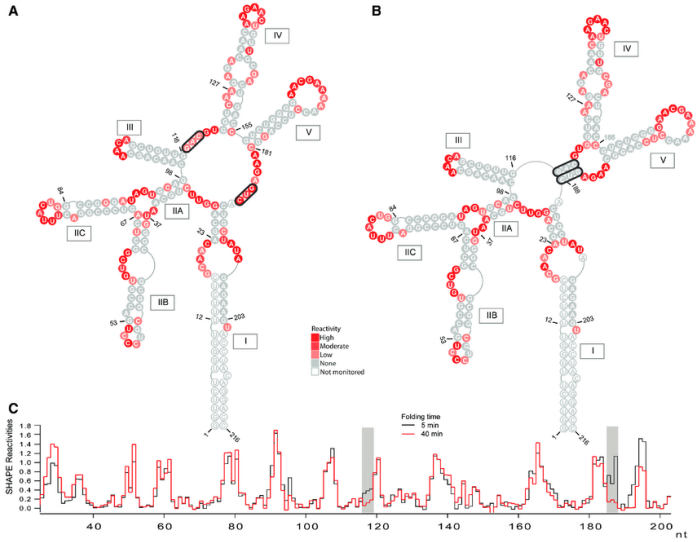
But these are the easy ones

Why we need attention

- Many objects:



- Convolved correlations



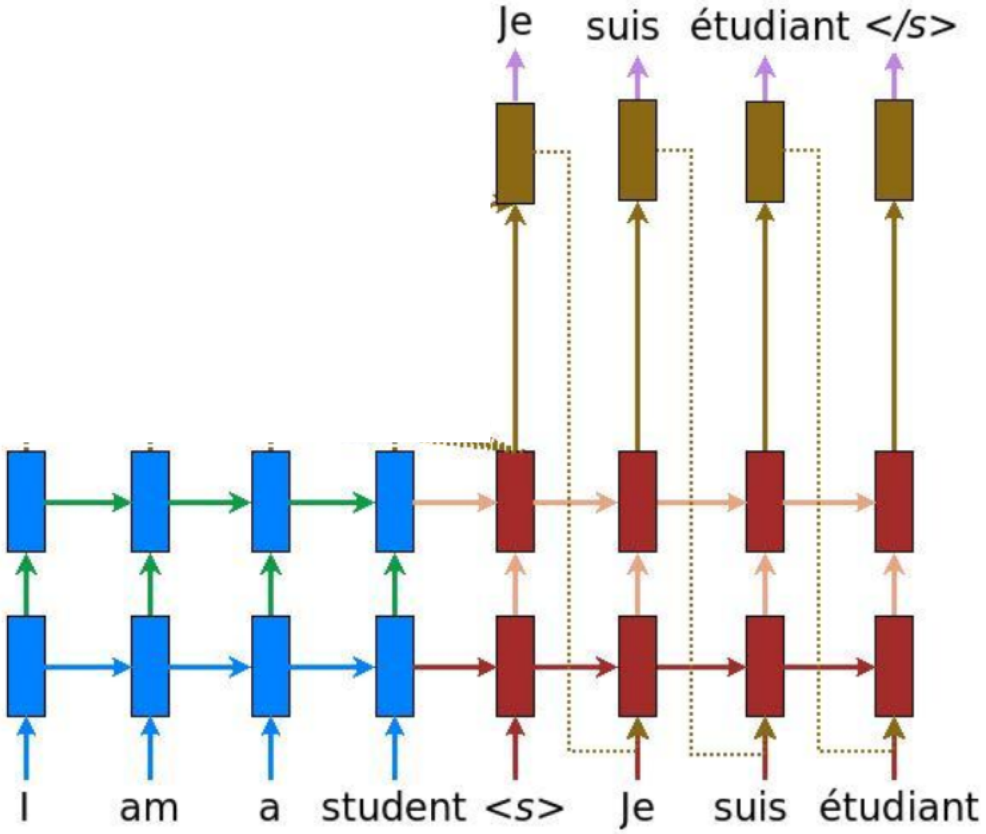
https://www.researchgate.net/figure/RRE2-structures-obtained-by-SHAPE-as-a-function-of-RNA-folding-time-RRE2-RNA-assumes_fig1_236608422

Summarize spatio-temporal information

Pooling/down-sampling/aggregation etc.

- Unparameterized:
  - reduce_max, reduce_mean, reduce_sum
  - Local pooling with sliding window
  - Slicing and scaling
  - etc.
- Weighted:
  - Attention
  - set2set
  - etc. and etc.

Vanilla attention



$$\alpha_{ts} = \frac{\exp\left(\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s)\right)}{\sum_{s'=1}^{S} \exp\left(\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_{s'})\right)} \quad (1)$$

$$\boldsymbol{c}_t = \sum_s \alpha_{ts} \bar{\boldsymbol{h}}_s \quad (2)$$

$$\boldsymbol{a}_t = f(\boldsymbol{c}_t, \boldsymbol{h}_t) = \tanh(\boldsymbol{W}_{\boldsymbol{c}}[\boldsymbol{c}_t; \boldsymbol{h}_t]) \quad (3)$$

[Attention weights] (1)

[Context vector] (2)

[Attention vector] (3)

Tensorflow seq2seq, Neural Machine Translation (Luong et al., 2015, Bahdanau et al., 2015)

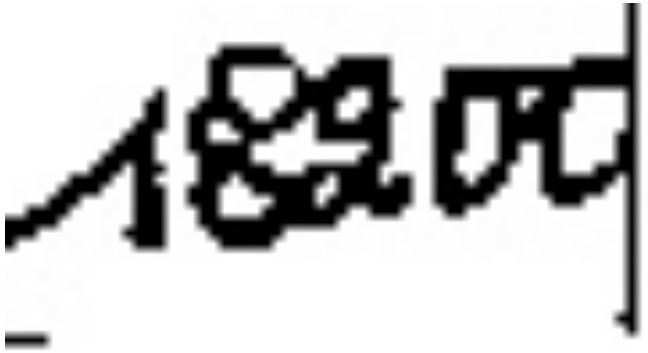
The gist of attention

$$\alpha_{ts} \propto score\,(h_t, h_s)$$

$$\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \begin{cases} \boldsymbol{h}_t^\top \boldsymbol{W} \bar{\boldsymbol{h}}_s & \text{[Luong's multiplicative style]} \\ \boldsymbol{v}_a^\top \tanh\left(\boldsymbol{W_1} \boldsymbol{h}_t + \boldsymbol{W_2} \bar{\boldsymbol{h}}_s\right) & \text{[Bahdanau's additive style]} \end{cases}$$

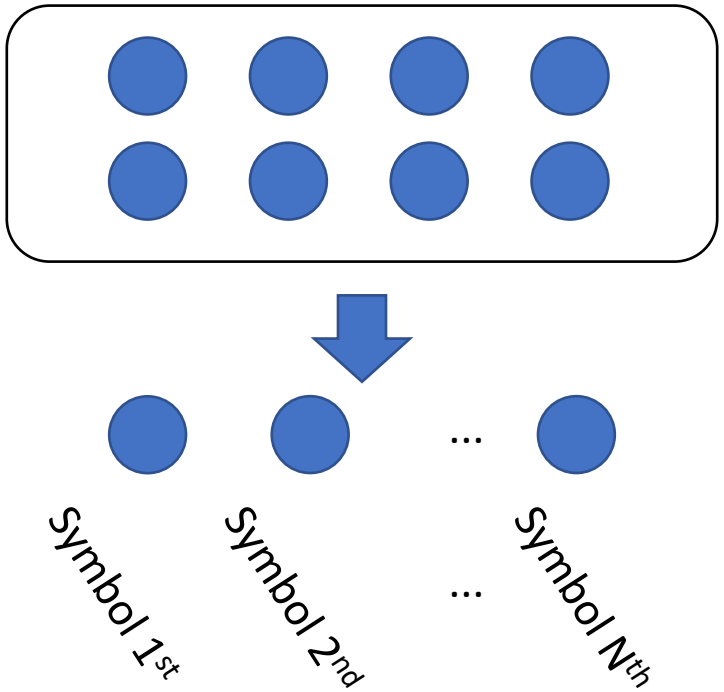
Attention in image to sequence translation

- OCR2Text



→ 18200

- Q. What's the problem of a naive mapping from the source to the target?



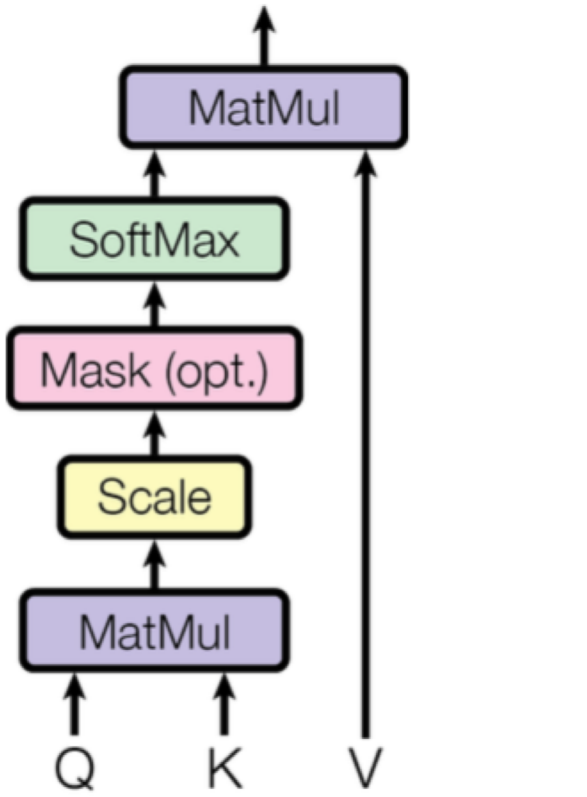
- A. It may not align well.
- Therefore use attention.
- Demo

Self-attention: coping with the loss of reference

- Generalized attention function:
- Q — querys (target hidden state h_t)
  - [batch_size, d_k]
- K — keys (source hidden states $h_s, s = 1..N$)
  - [batch_size, N, d_k]
- V — memory (source hidden states $h_s, s = 1..N$)
  - [batch_size, N, d_v]
- Note: multiplicative style
- Scaling factor accounts for gradient saturation

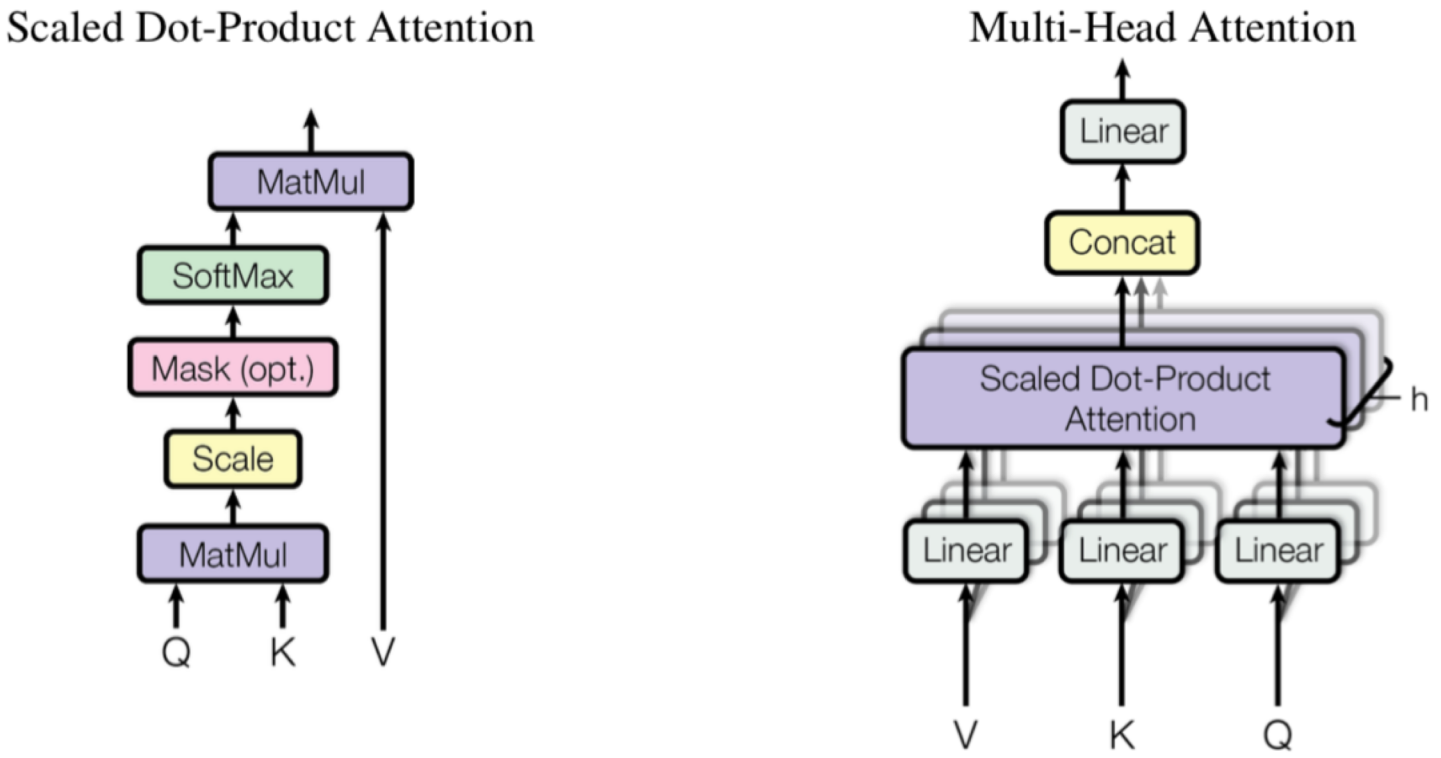
$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

A. Vaswani et al. Attention is all you need. NIPS. 2017



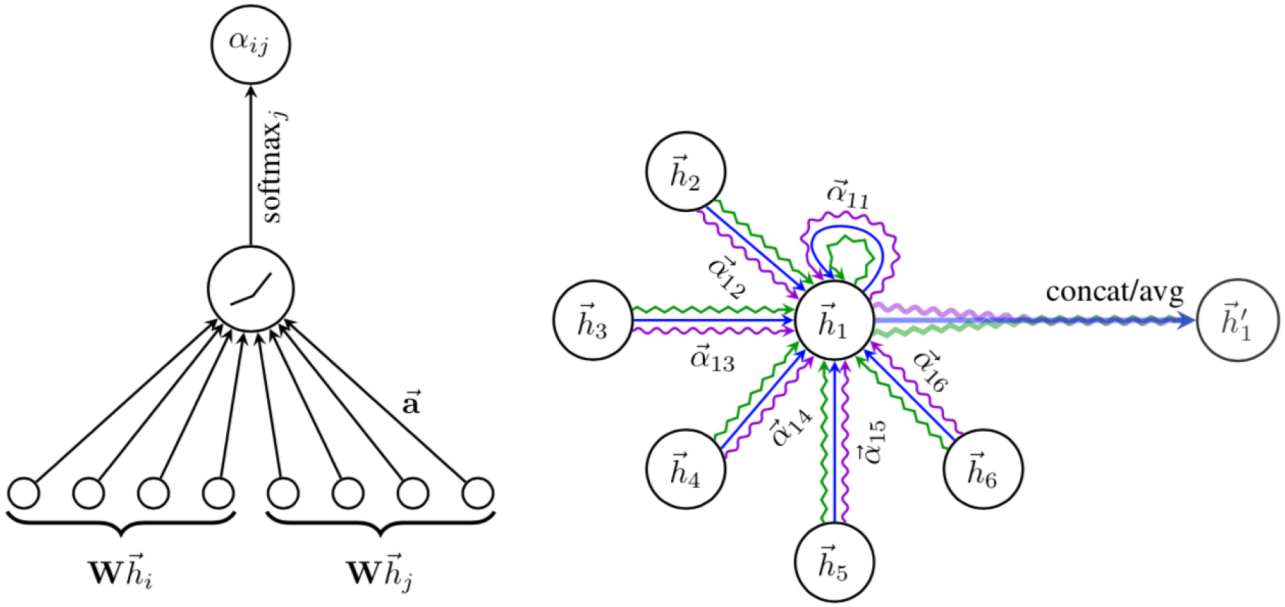
- Self-attention is when Q, K, V are the same → comparing to itself

Mutli-head attention



* have h parallel&indepedent self-attention modules from the last slide
* then concatenate/average their features

Attention is pervasive



Q: What's the style of this attention?

A: additive

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(\vec{\mathbf{a}}^T[\mathbf{W}\vec{h}_i \| \mathbf{W}\vec{h}_j]\right)\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(\text{LeakyReLU}\left(\vec{\mathbf{a}}^T[\mathbf{W}\vec{h}_i \| \mathbf{W}\vec{h}_k]\right)\right)}$$

where $\cdot^T$ represents transposition and $\|$ is the concatenation operation.

P. Velickovic et al. Graph attention network. ICLR. 2018