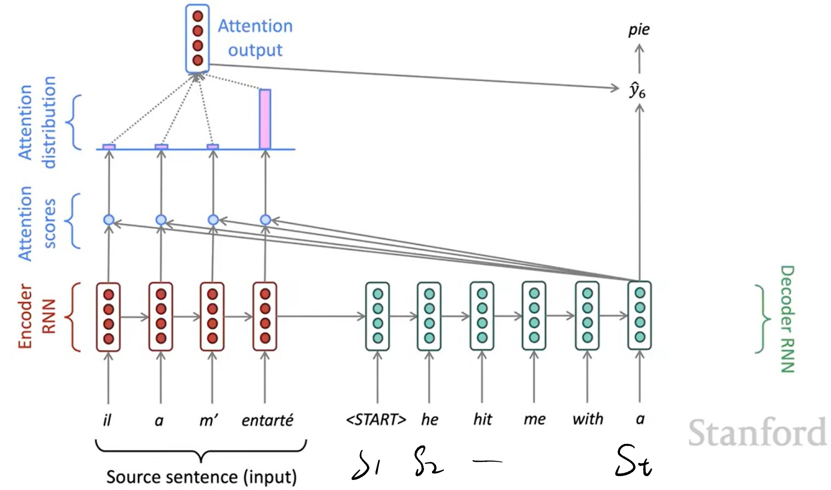
Attention Revisited

Seq2Seq with Attention



$s_1 \quad s_2 \quad - \quad s_t$

$h_1 \quad h_2 \quad - \quad h_N$

attention scores e^t: $\quad e^t = [s_t^T h_1 \; - \; s_t^T h_N] \in \mathbb{R}^N$

$\alpha^t = \text{softmax}(e^t) \in \mathbb{R}^N$ take softmax to get the attention distribution α^t for this step.

$$a_t = \sum_{i=1}^{N} \alpha_i^t h_i \in \mathbb{R}^h$$

concatenate $[a_t; s_t] \in \mathbb{R}^{2h}$

and proceed as in the non-attention seq2seq model

Attention: to focus on certain parts of the source

solves the bottleneck problem

helps with the vanishing gradient problem

provides some interpretability

get (soft) alignment for free

There are several attention variants

values $h_1, h_2, \cdots, h_N \in \mathbb{R}^{d_1}$ (similar to encode hidden state)

a query $s \in \mathbb{R}^{d_2}$ (similar to decoder hidden state)

Attention:

① computing attention scores $e \in \mathbb{R}^N$

② Taking softmax to get attention distribution α:

$$\alpha = \text{softmax}(e) \in \mathbb{R}^N$$

③ Using attention distribution to take weighted sum of values:

$$a = \sum_{i=1}^{N} \alpha_i h_i \in \mathbb{R}^{d_1}$$

$\Rightarrow$ attention output a

(sometimes called the context vector)

Several ways to compute $e \in \mathbb{R}^N$ from $h_1, \ldots, h_N \in \mathbb{R}^{d_1}$

$s \in \mathbb{R}^{d_2}$

① Basic dot product attention

$$e_i = s^T h_i \in \mathbb{R}$$

assumes $d_1 = d_2$

② multiplicative attention:

$$e_i = s^T W h_i \in \mathbb{R}$$

Where $W \in \mathbb{R}^{d_2 \times d_1}$ is a weight matrix

③ Reduced rank multiplicative attention:

$$e_i = s^T (U^T V) h_i$$

$$= (Us)^T (Vh_i)$$

low rank matrices $U \in \mathbb{R}^{k \times d_2}$, $V \in \mathbb{R}^{k \times d_1}$

$k << d_1, d_2$

④ Additive attention: $e_i = \underbrace{v^T}_{\in \mathbb{R}^{d_3}} \tanh(\underbrace{W_1}_{\in \mathbb{R}^{d_3 \times d_1}} h_i + \underbrace{W_2}_{\in \mathbb{R}^{d_3 \times d_2}} s)$

Given a set of vector values,

and a vector query, attention is a technique