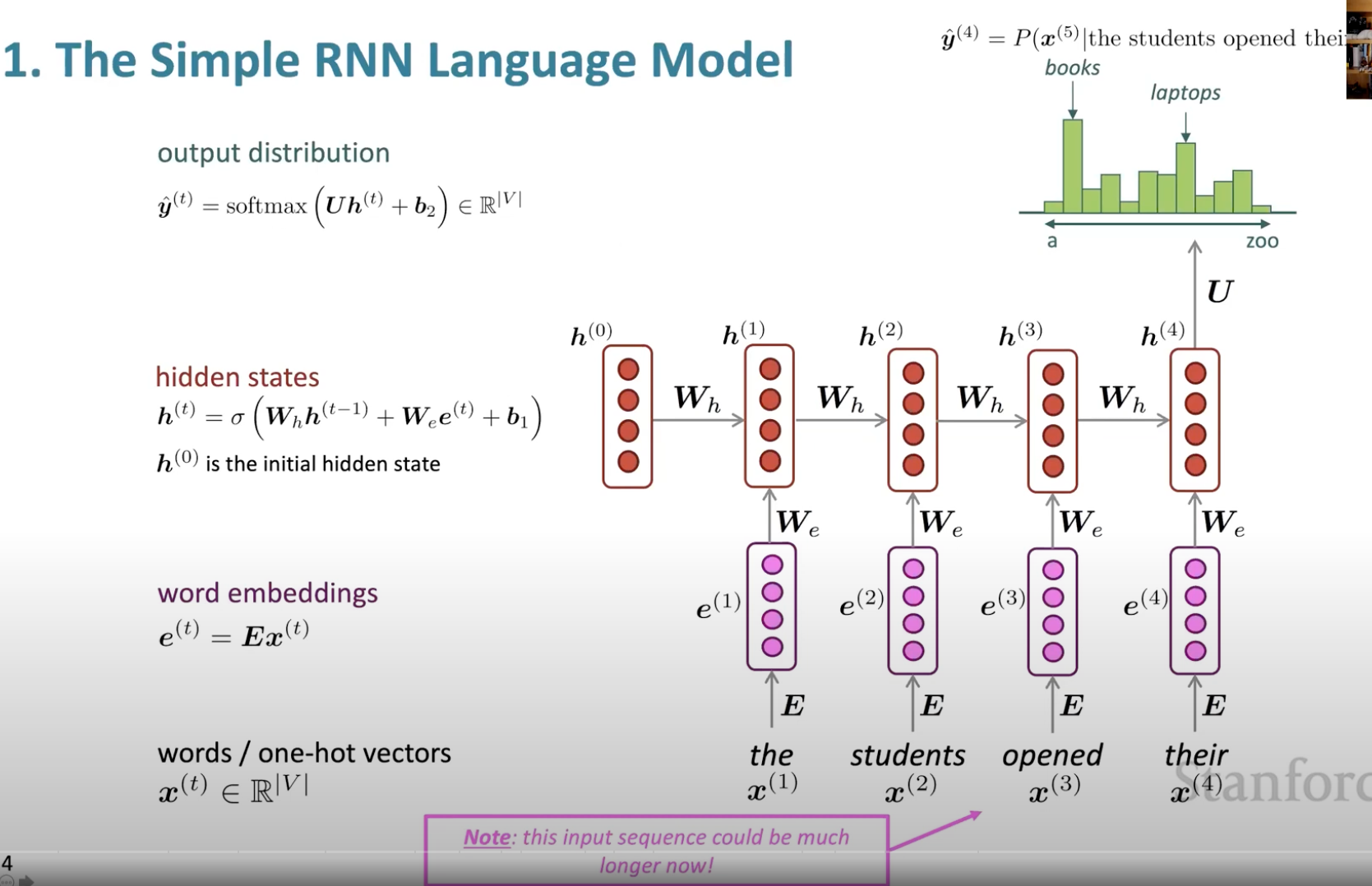
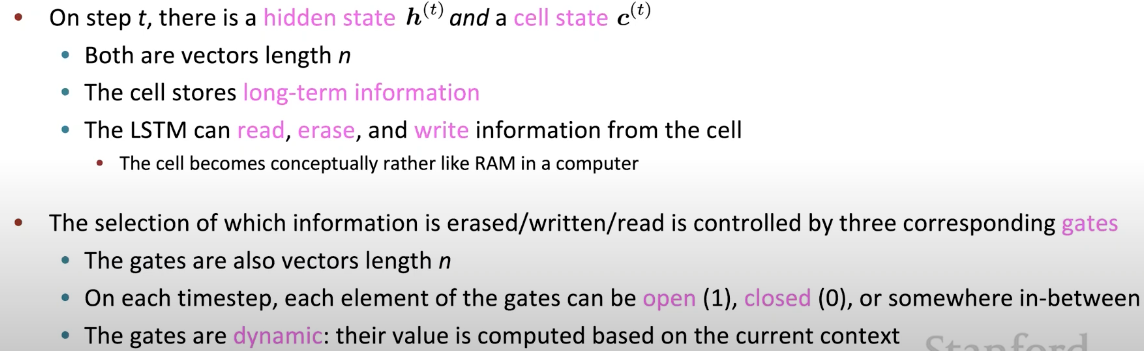
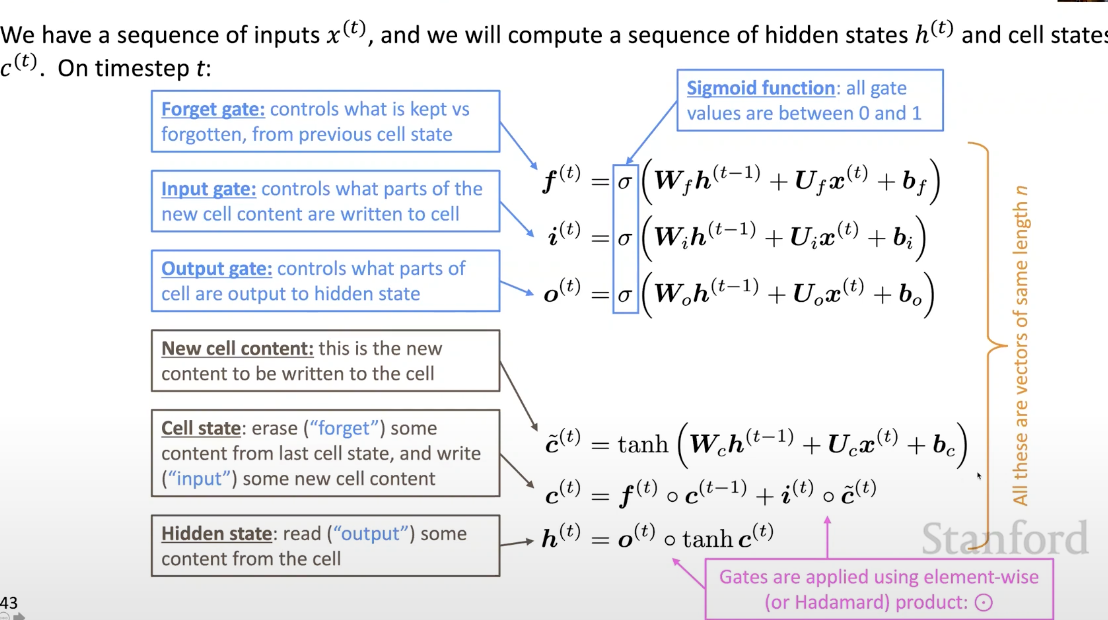
* **RNN**

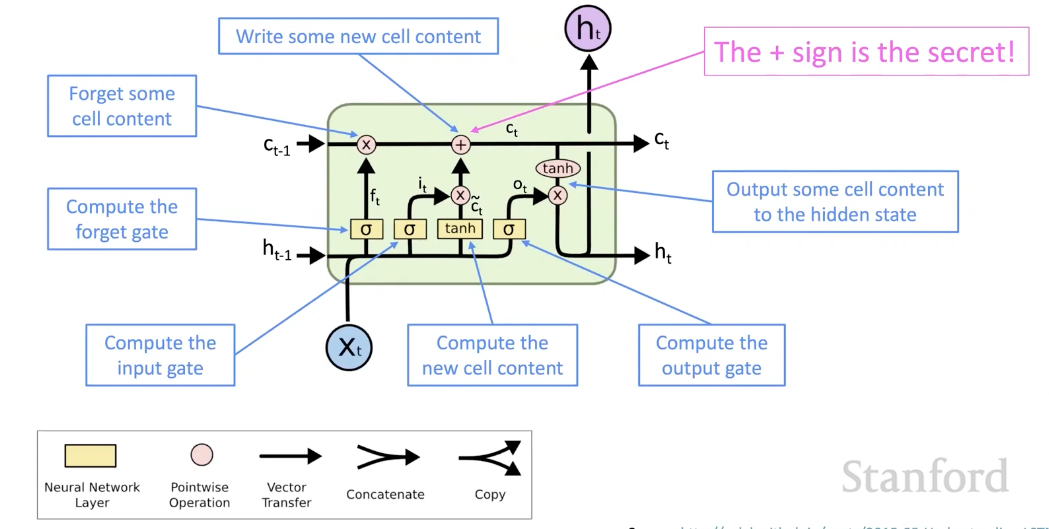


* **LSTM**



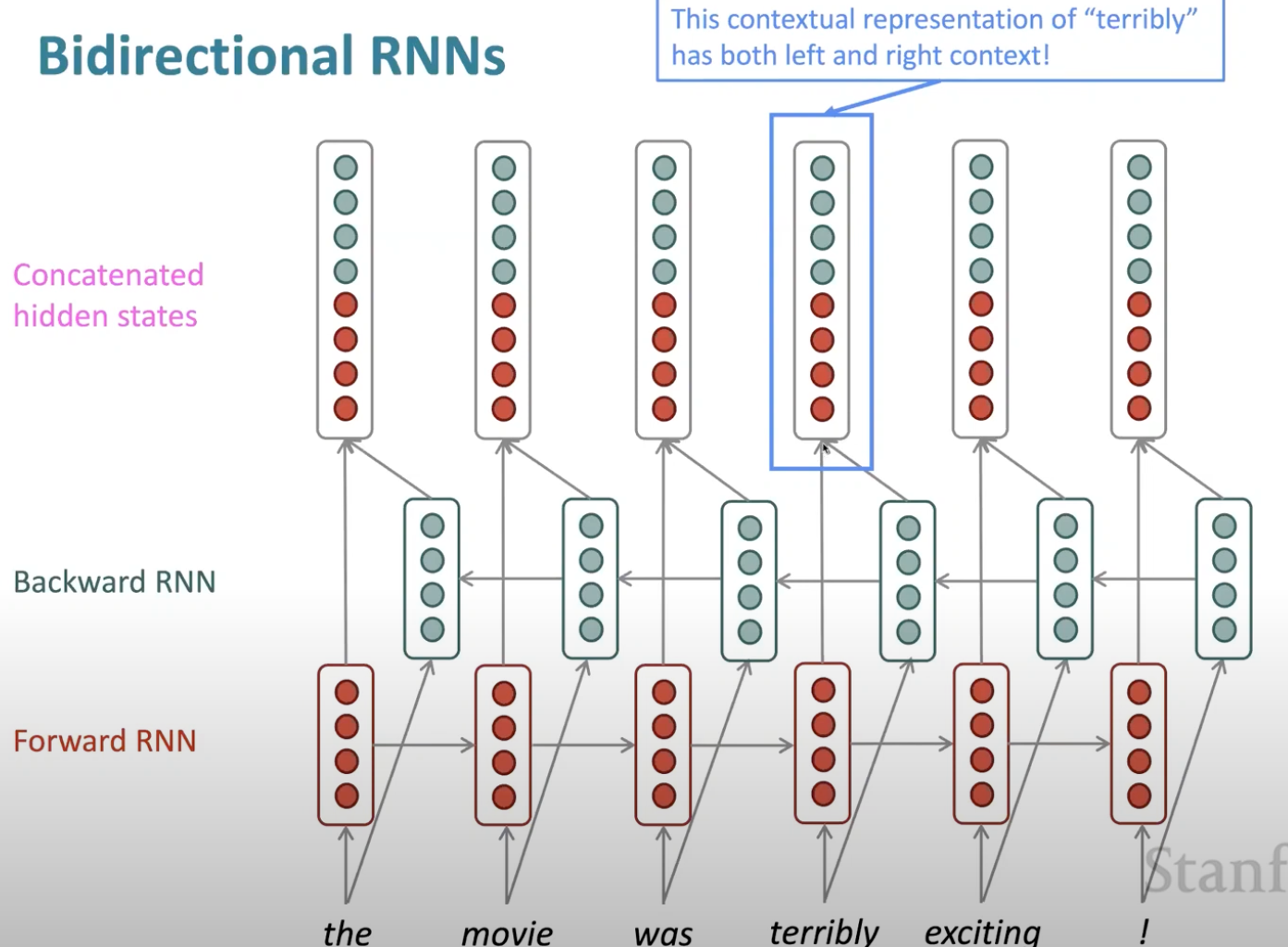


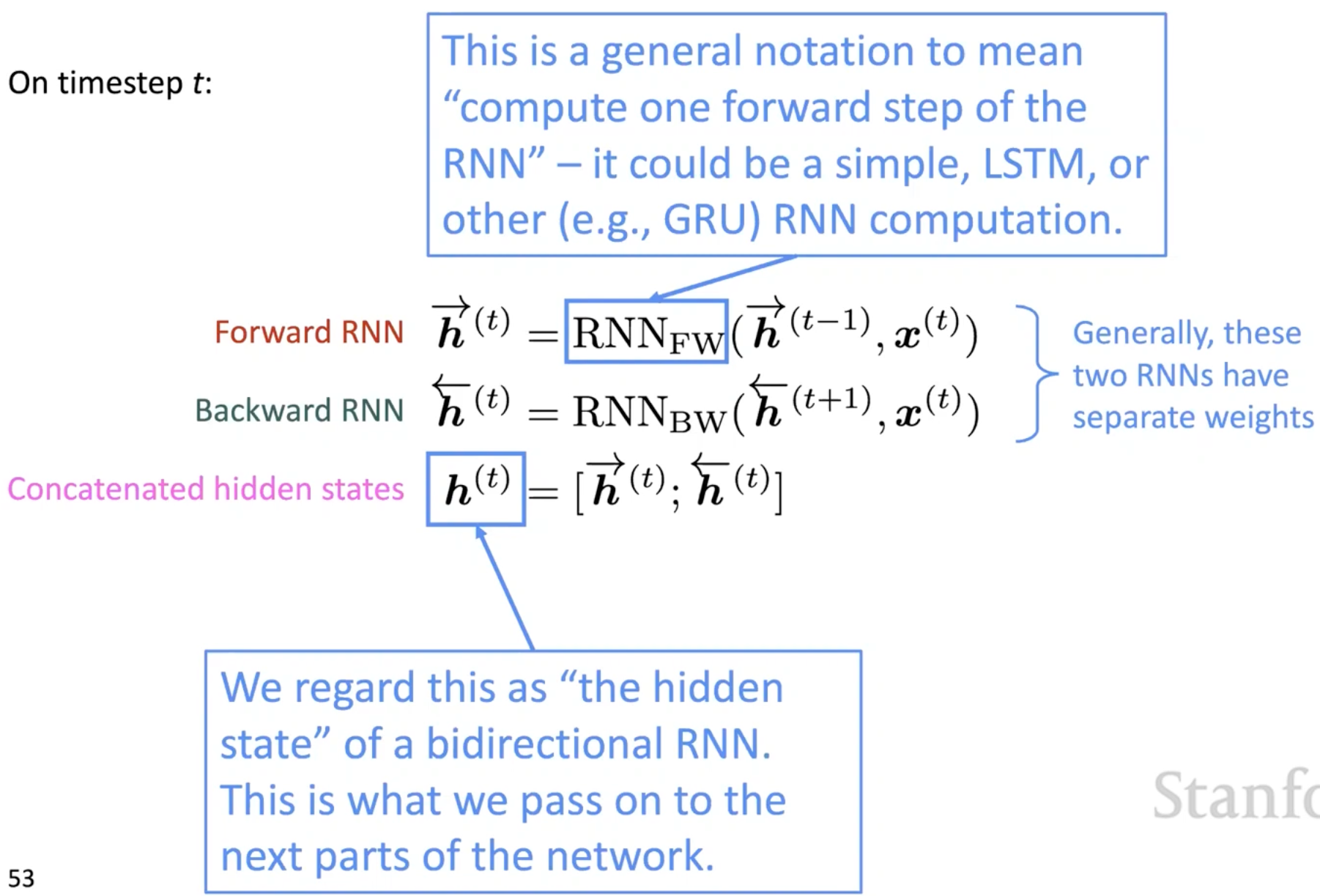
The prediction of classification for example, is **Y\_pred = softmax(W\_h@h\_t + b\_h)**



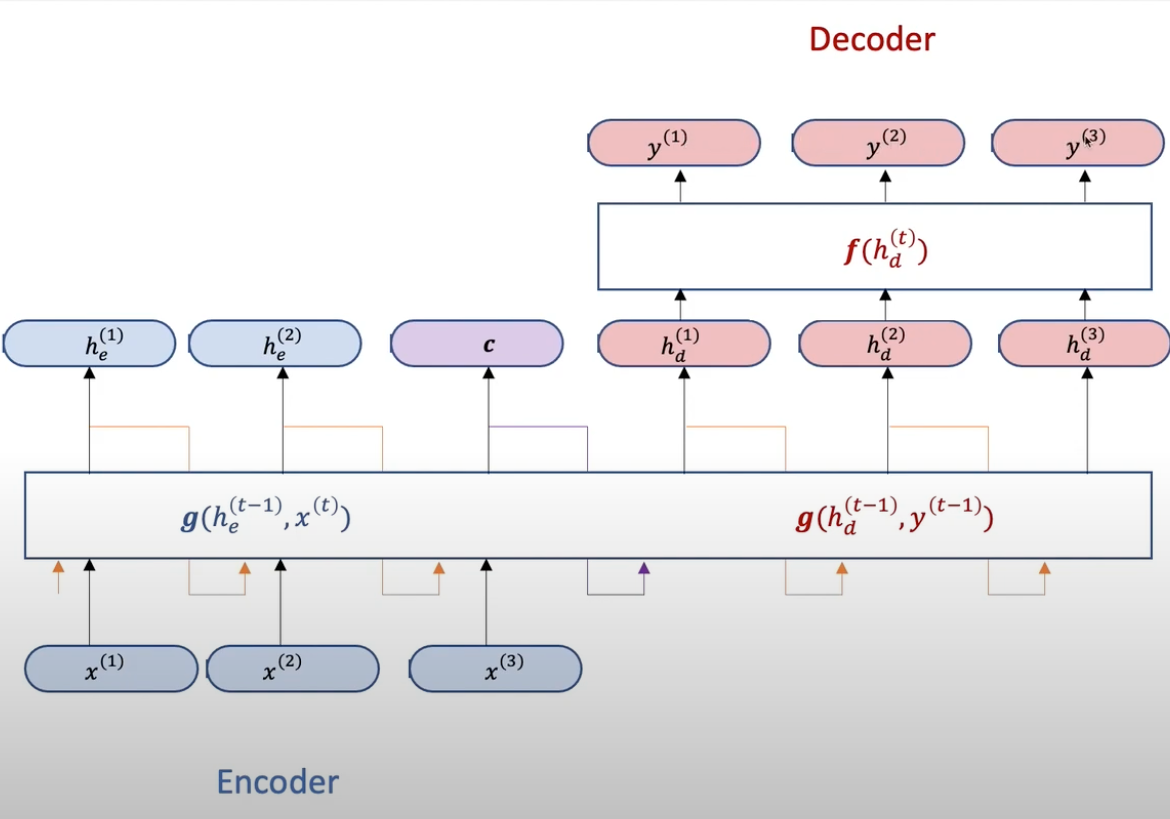
* **BIRDIRECTIONAL RNN**

In order to capture both ways of context

****

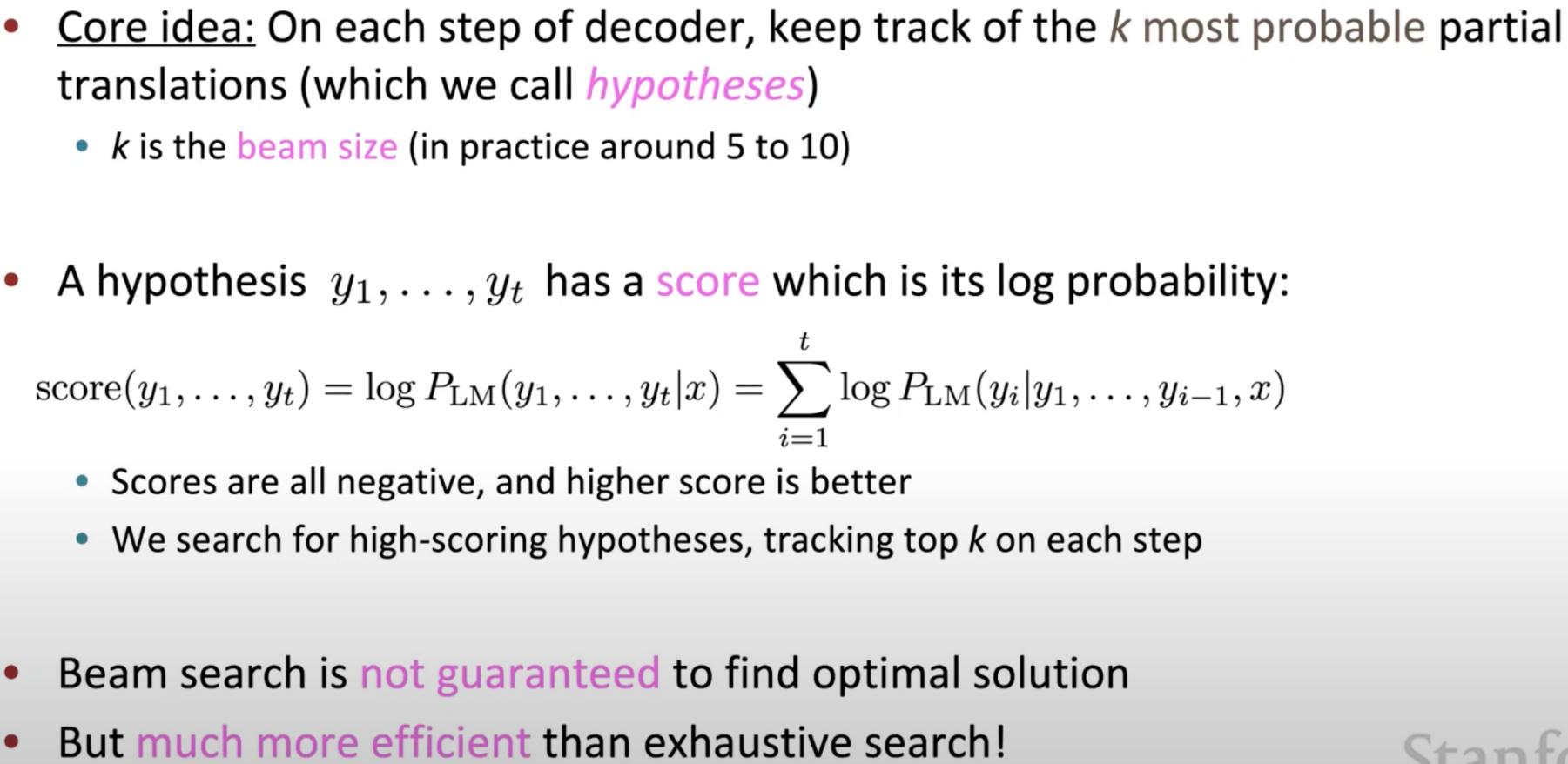
****

* **ENCODER DECODER**

****

C is the final hidden state in encoder part, which captures all the info from previous hidden state in the sentence; decoder takes in C as the first hidden state, and a special token (i.e `<END>`) of the previous output as the current input, CWh + <END>Wx + b to predict y(1), it unrolls C.

* **[BEAM SEARCH](https://youtu.be/wzfWHP6SXxY)**

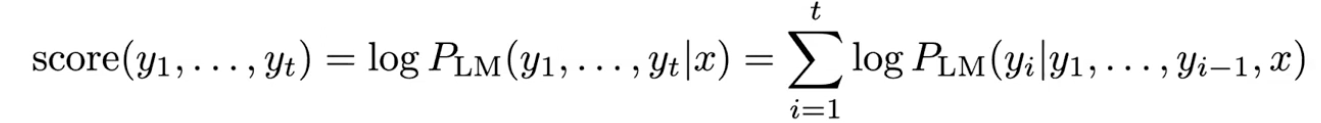
****

1. Stopping criterion

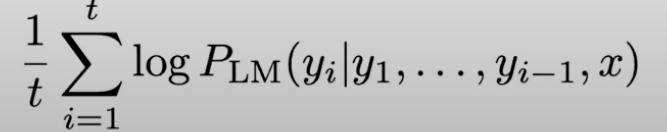


1. Problem

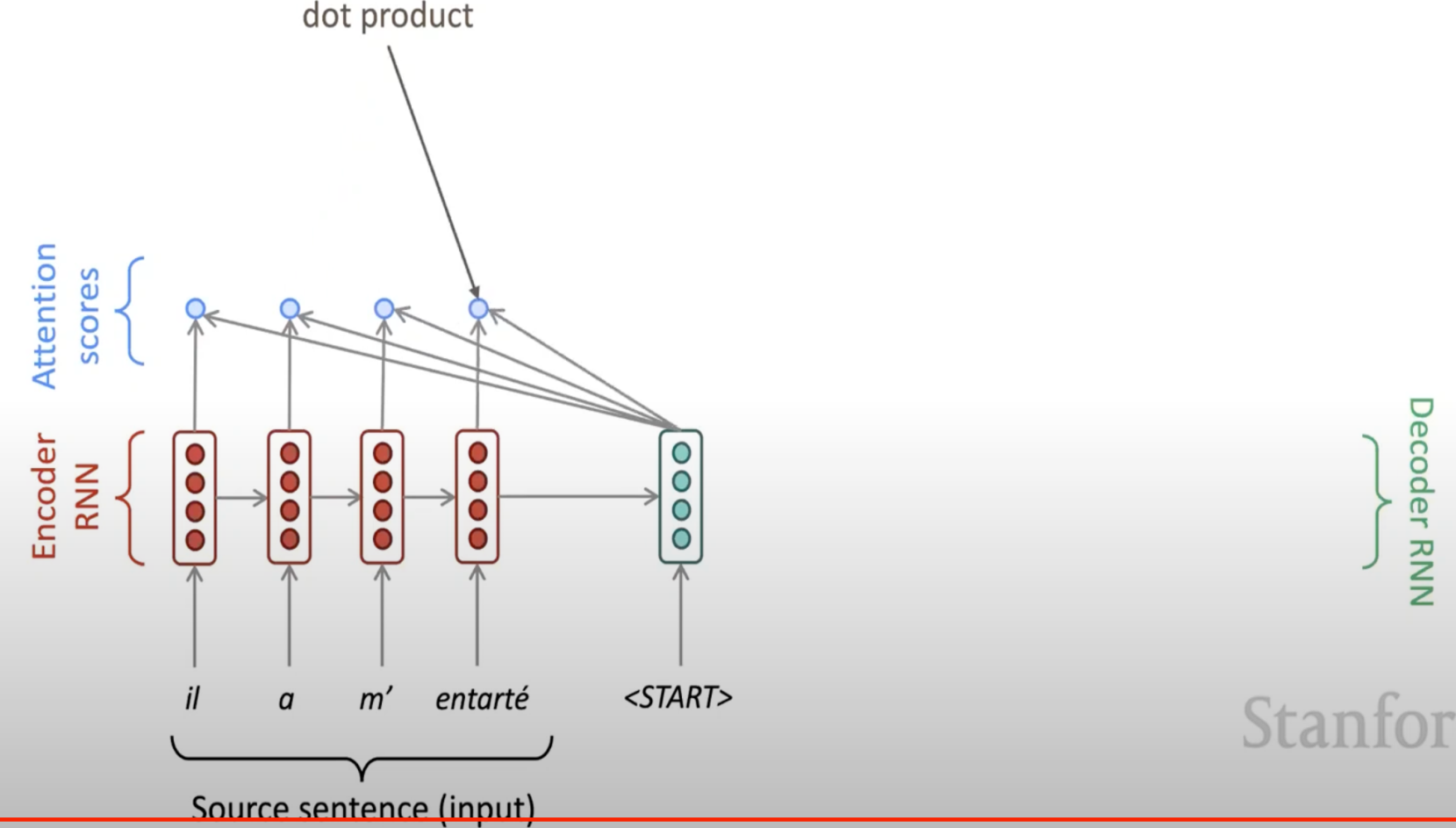
Longer sentences have lower score b/c accumulated multiplication



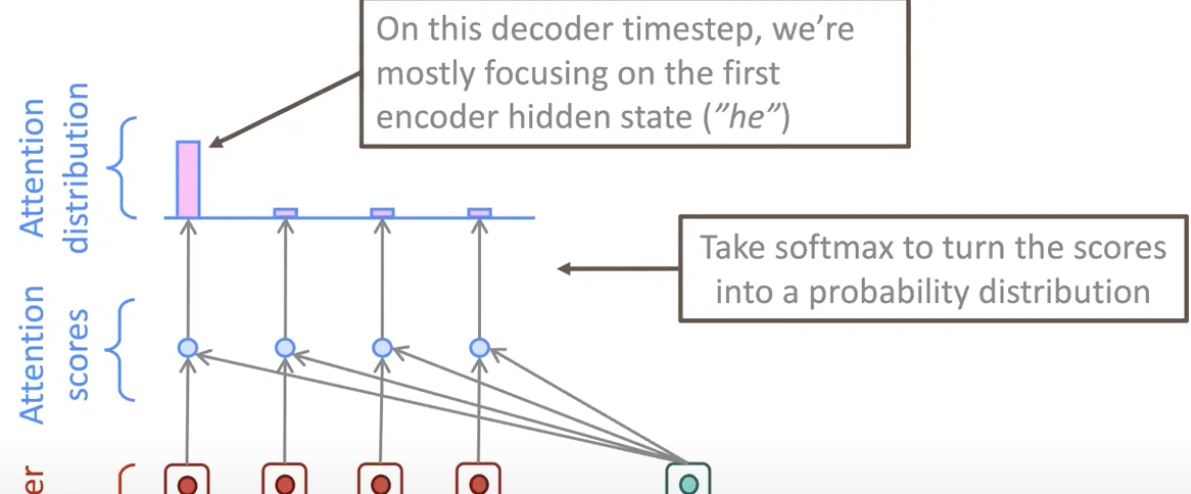
Fix: normalize be length, use this to select top one

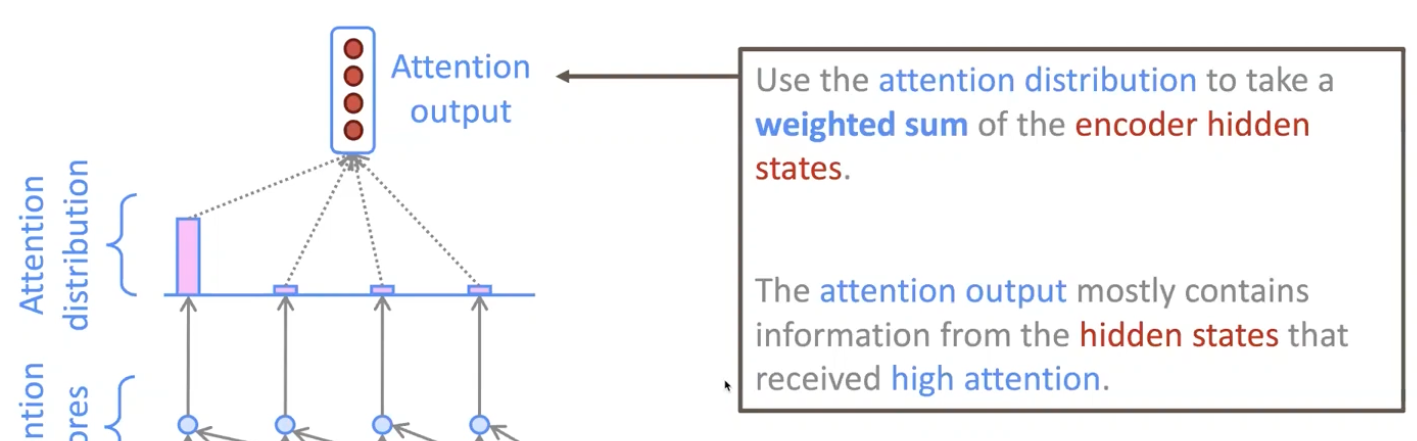


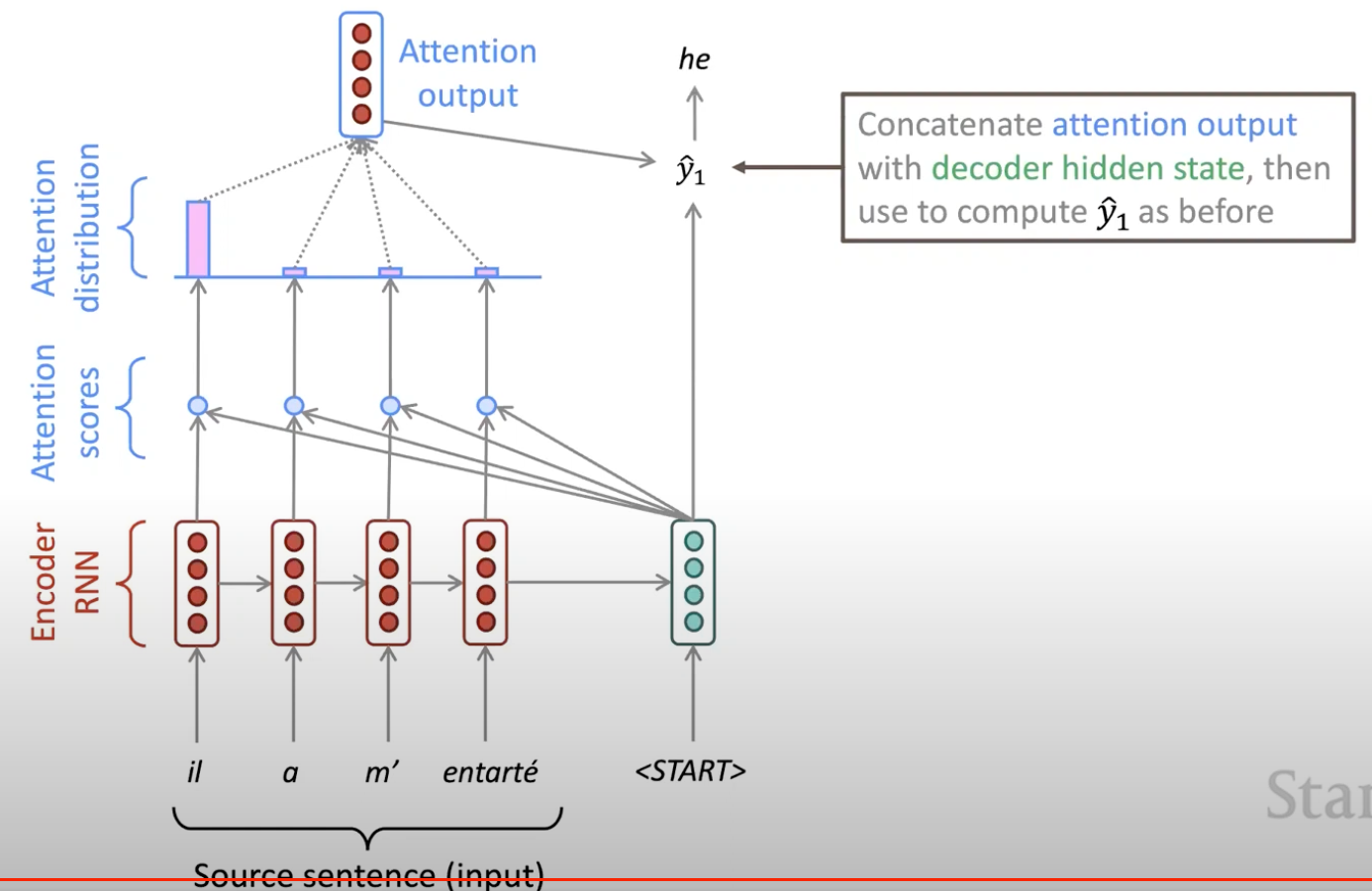
* **[ATTENTION](https://youtu.be/wzfWHP6SXxY)**

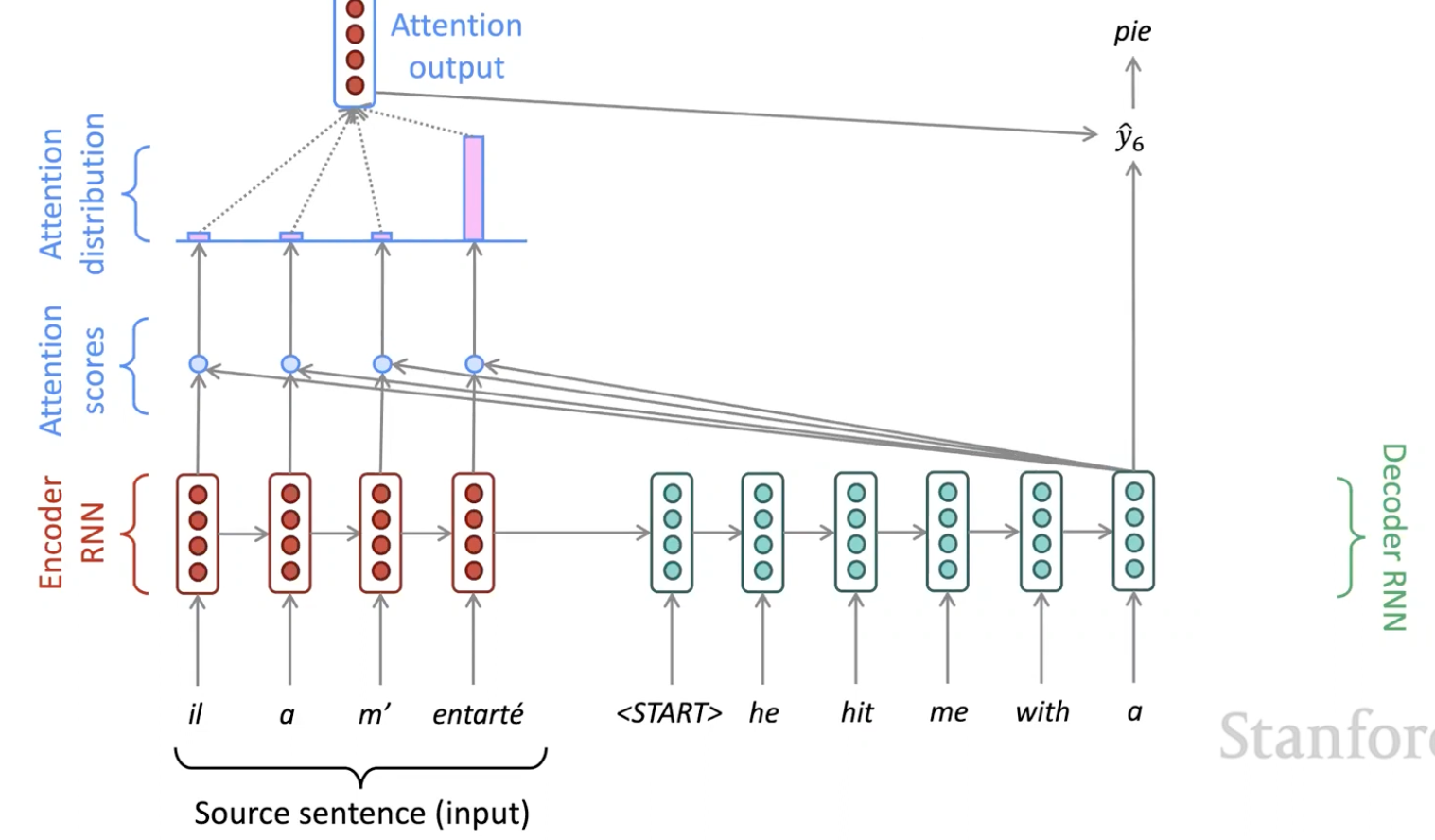


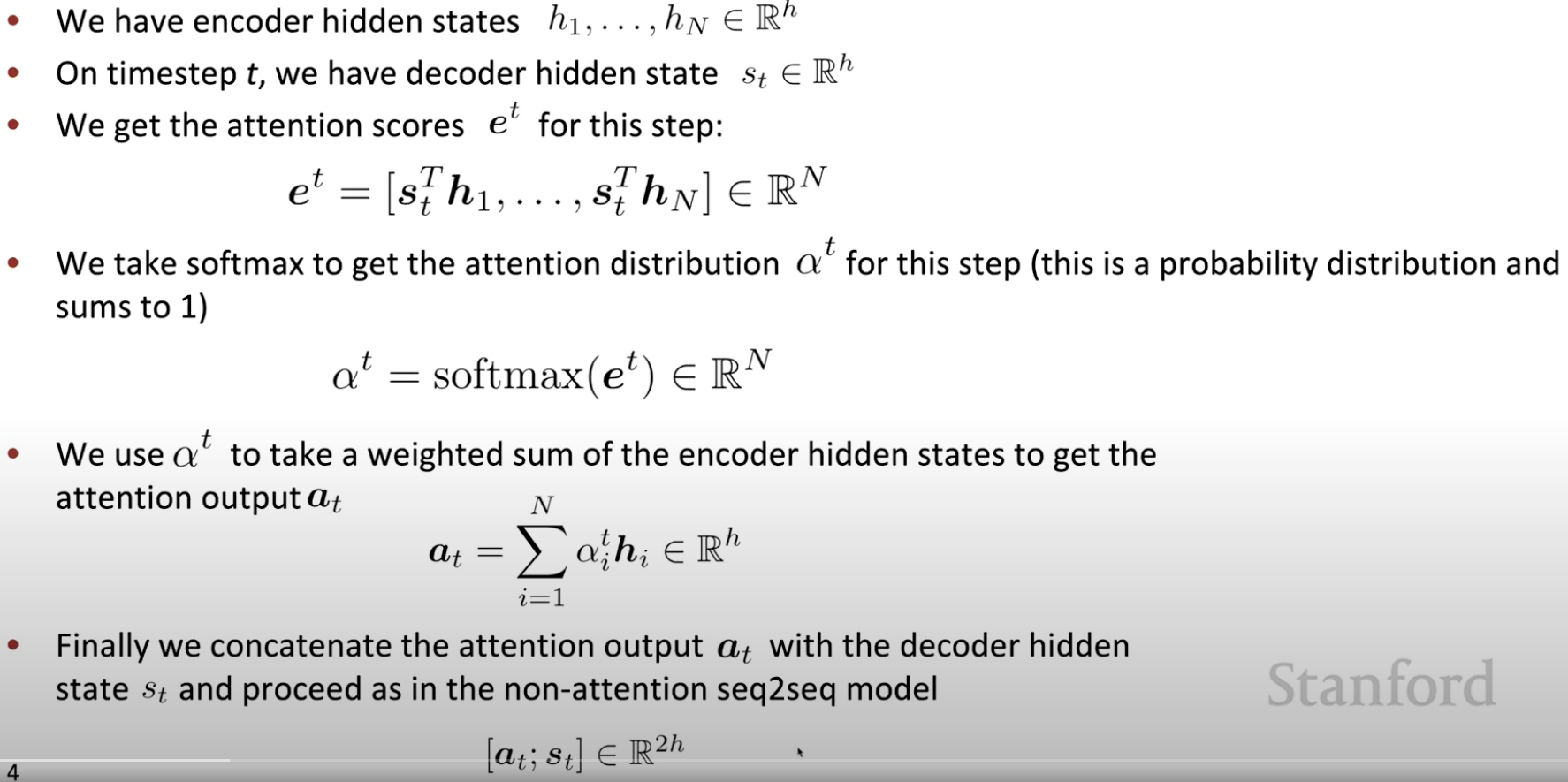
Find out which of the encoder state is most like the decoder state:





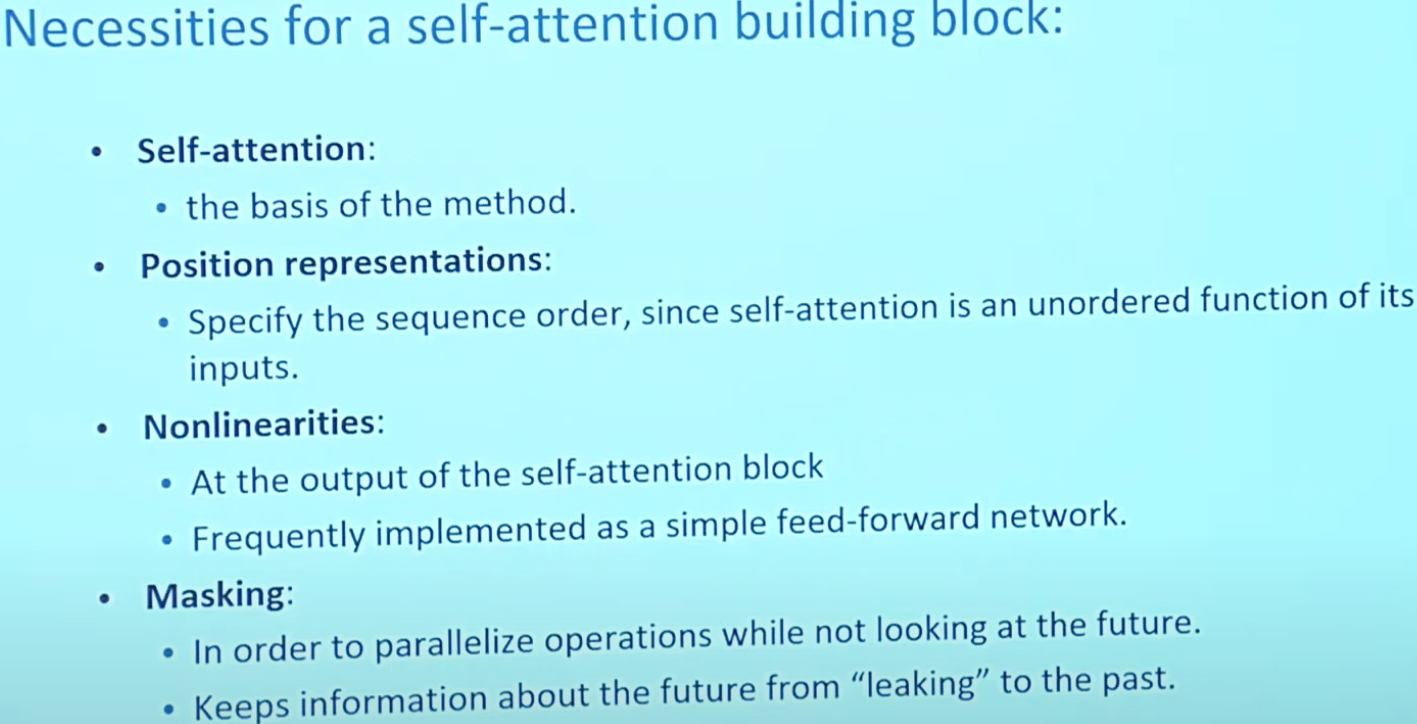




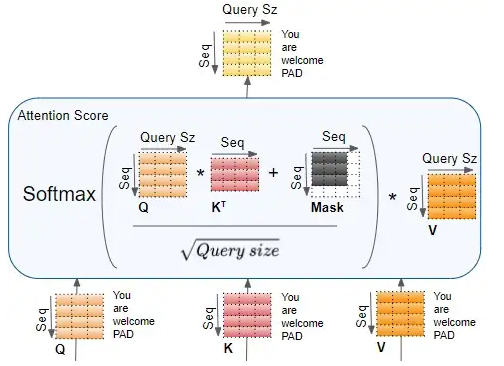


* **[TRANSFORMER](https://towardsdatascience.com/transformers-explained-visually-part-2-how-it-works-step-by-step-b49fa4a64f34)**

1. [Self-attention](https://youtu.be/ptuGllU5SQQ)

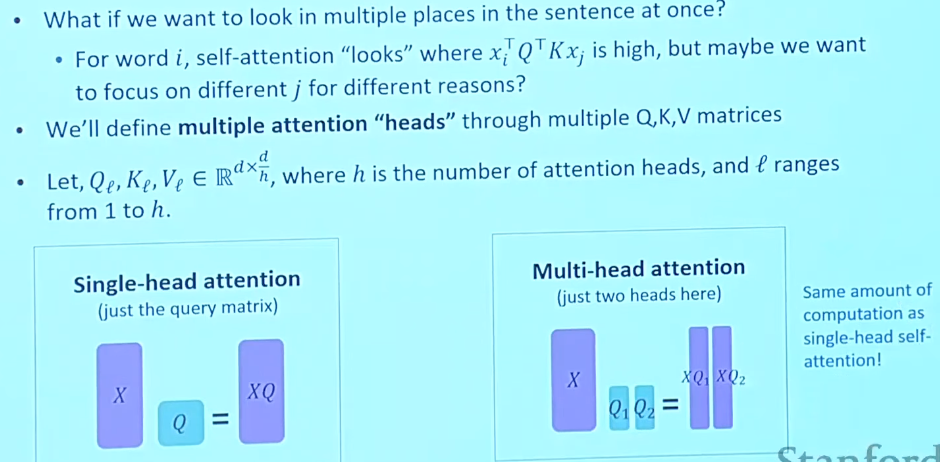


<https://towardsdatascience.com/transformers-explained-visually-not-just-how-but-why-they-work-so-well-d840bd61a9d3>

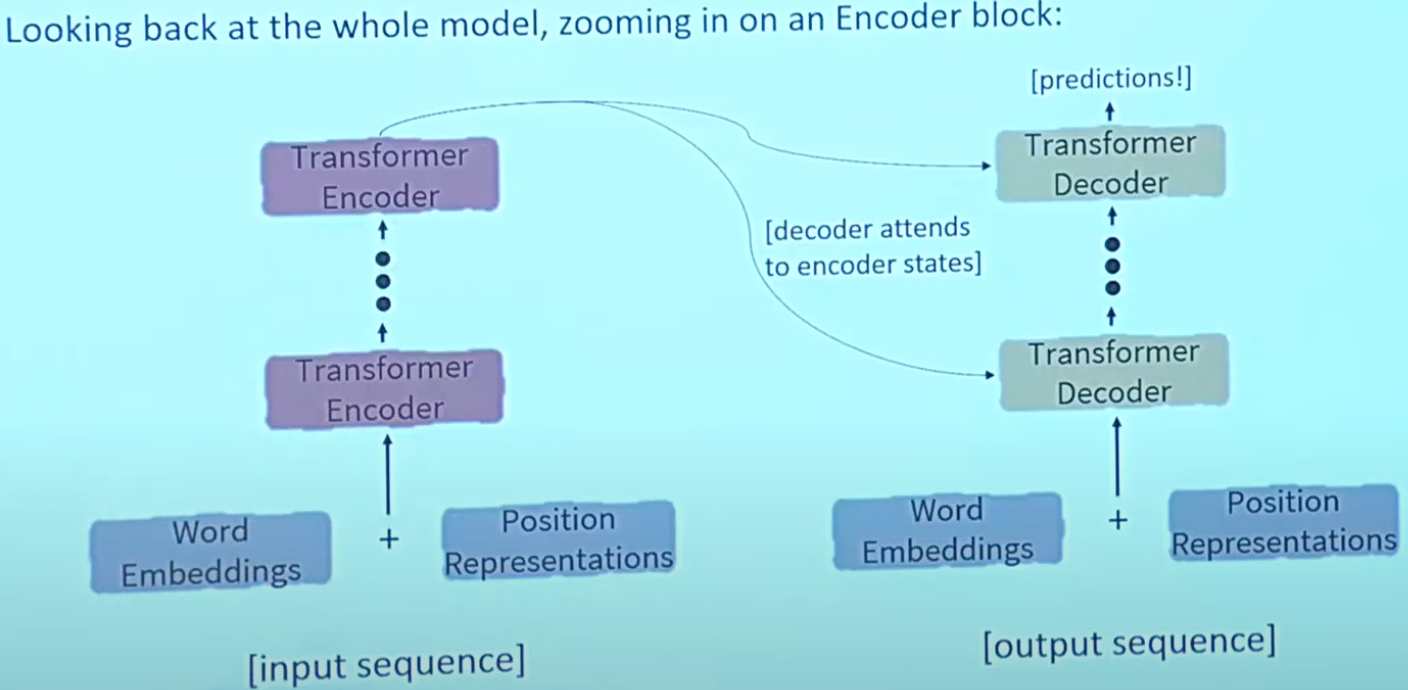


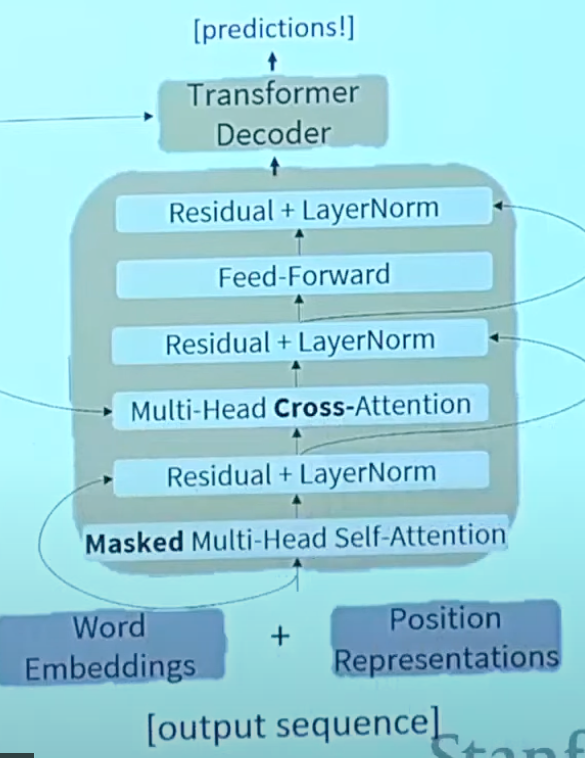
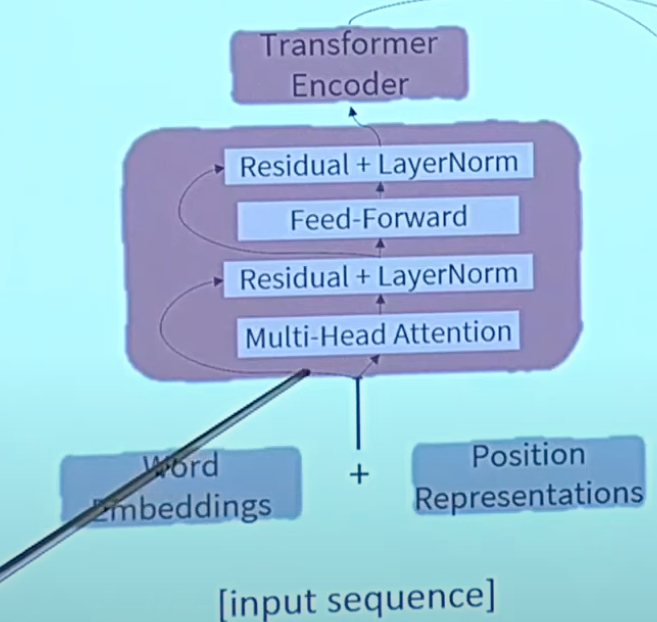
1. Multi-head attention

To look at a certain word’s interactions with different position in the sentence.



1. Transformer





1. Pretraining