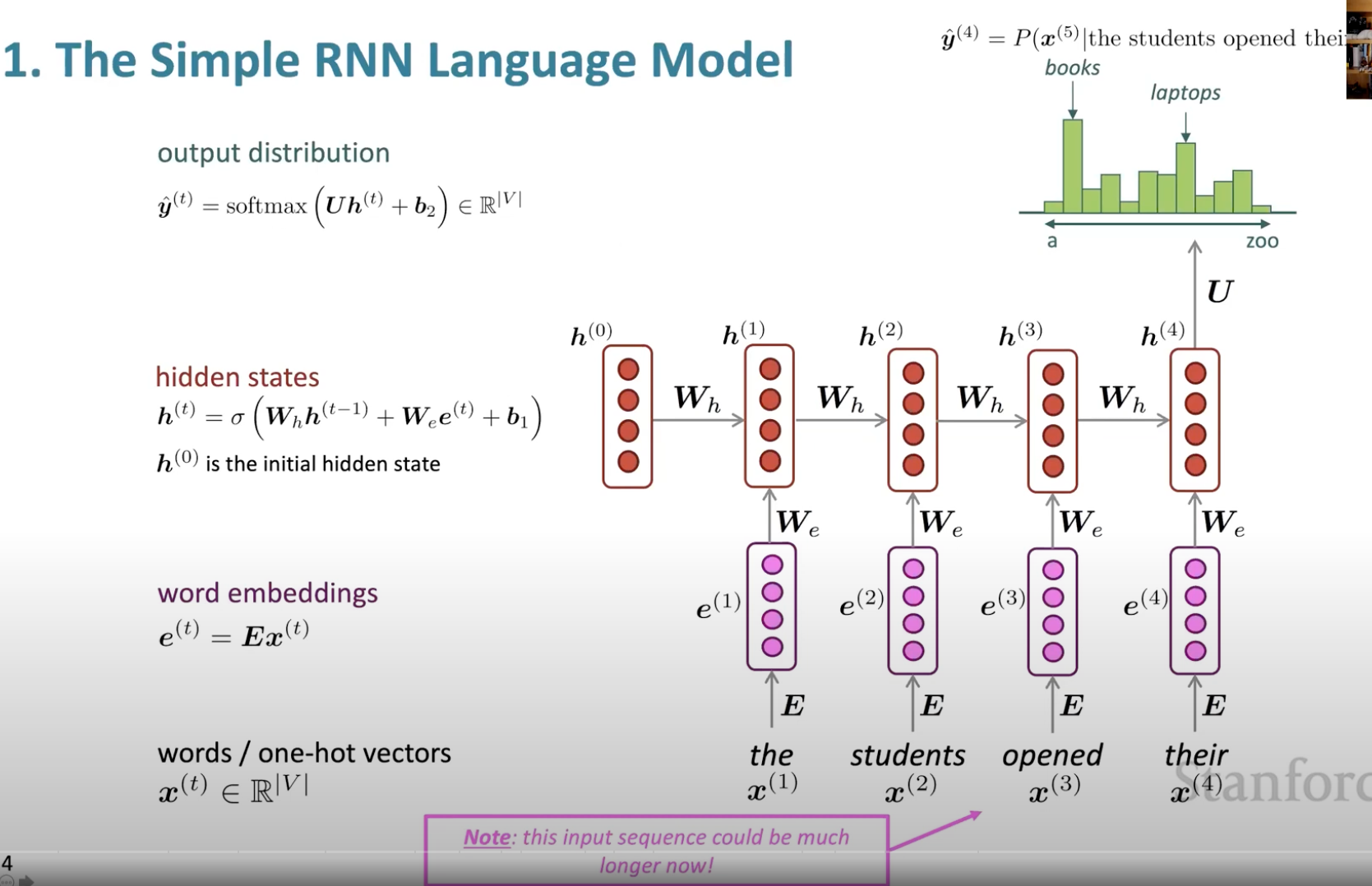
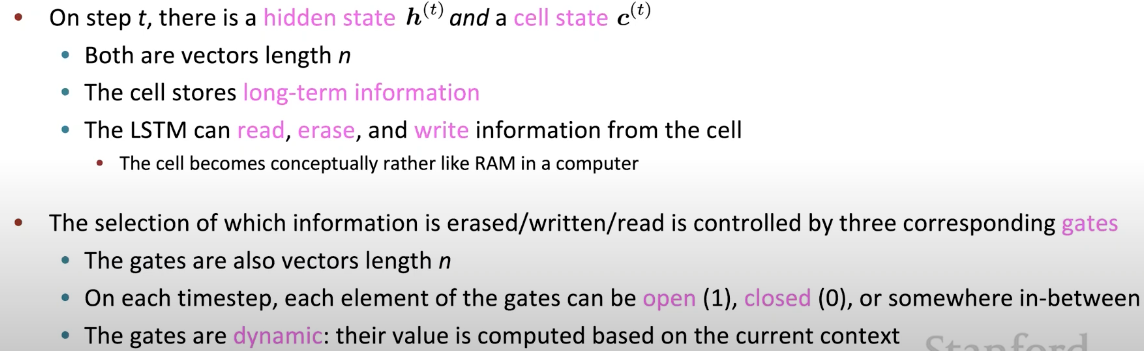
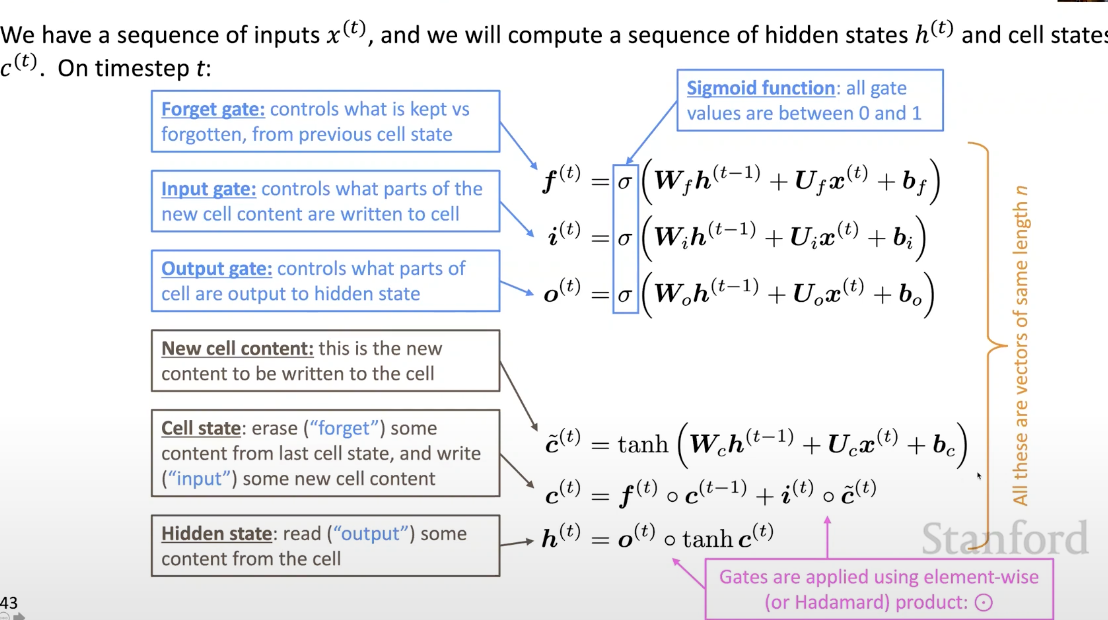
* **RNN**



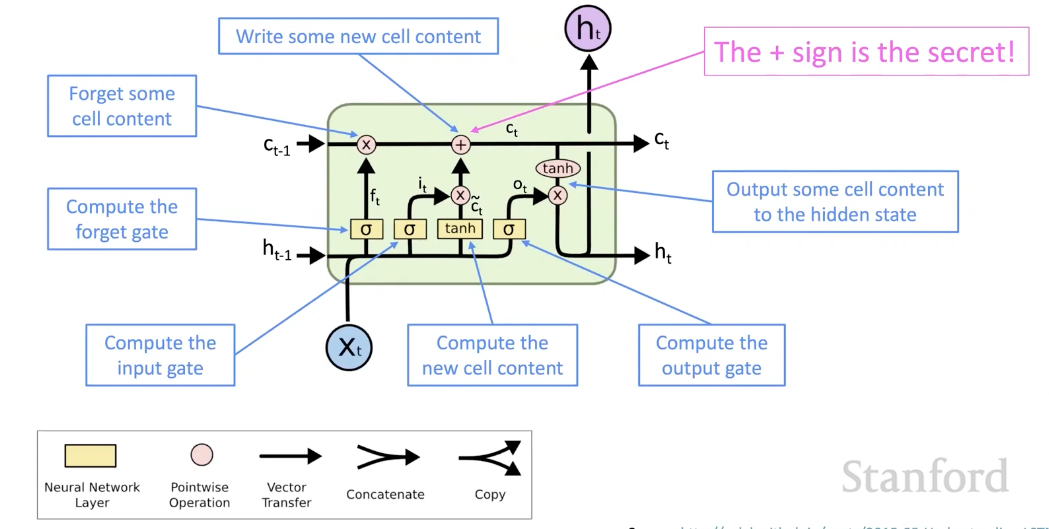
* **LSTM**



* An LSTM is similar to an RNN in that they both use hidden states to pass along information, but an LSTM also uses a cell state, which is like a long-term memory, to help deal with the issue of vanishing gradients
* An LSTM cell consists of a cell state, or long-term memory, a hidden state, or short-term memory, along with 3 gates that constantly update the relevancy of its inputs:
  + A **forget** gate, which decides which input units should be remembered and passed along. It's a tensor with values between 0 and 1.
    - If a unit has a value close to 0, the LSTM will "forget" the stored state in the previous cell state.
    - If it has a value close to 1, the LSTM will mostly remember the corresponding value.
  + An **update** gate, again a tensor containing values between 0 and 1. It decides on what information to throw away, and what new information to add.
    - When a unit in the update gate is close to 1, the value of its candidate is passed on to the hidden state.
    - When a unit in the update gate is close to 0, it's prevented from being passed onto the hidden state.
  + And an **output** gate, which decides what gets sent as the output of the time step

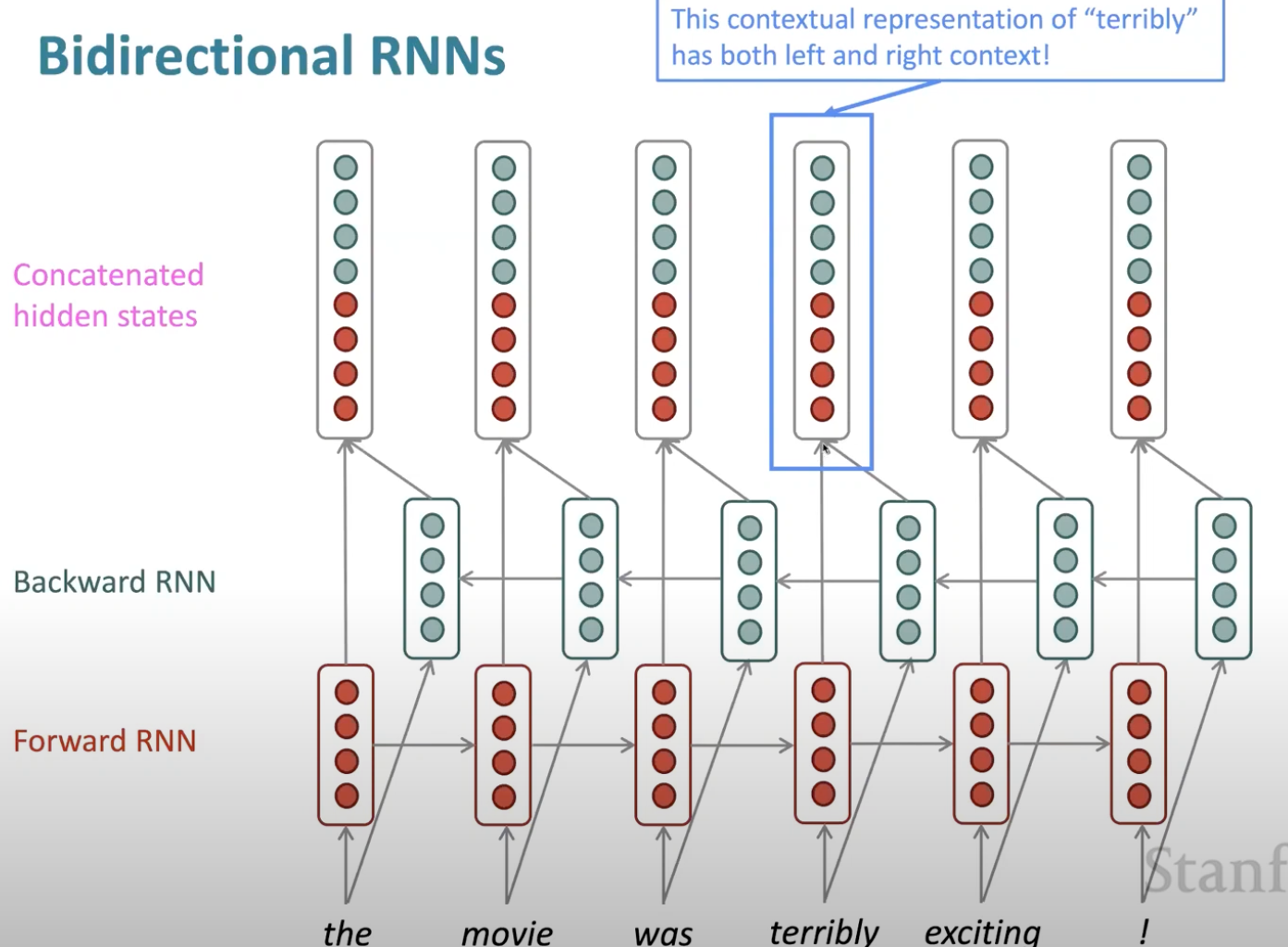


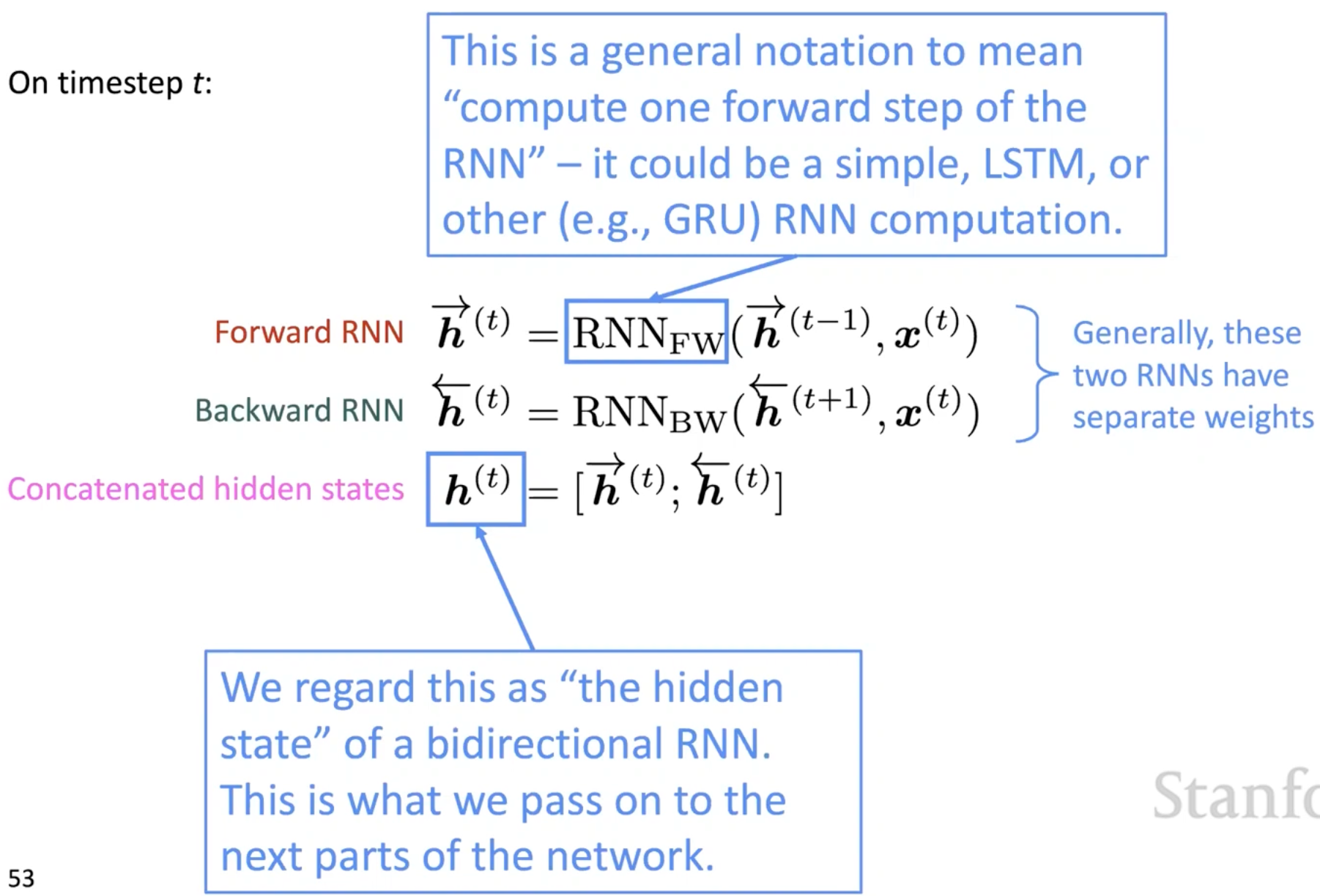
The prediction of classification for example, is **Y\_pred = softmax(W\_h@h\_t + b\_h)**



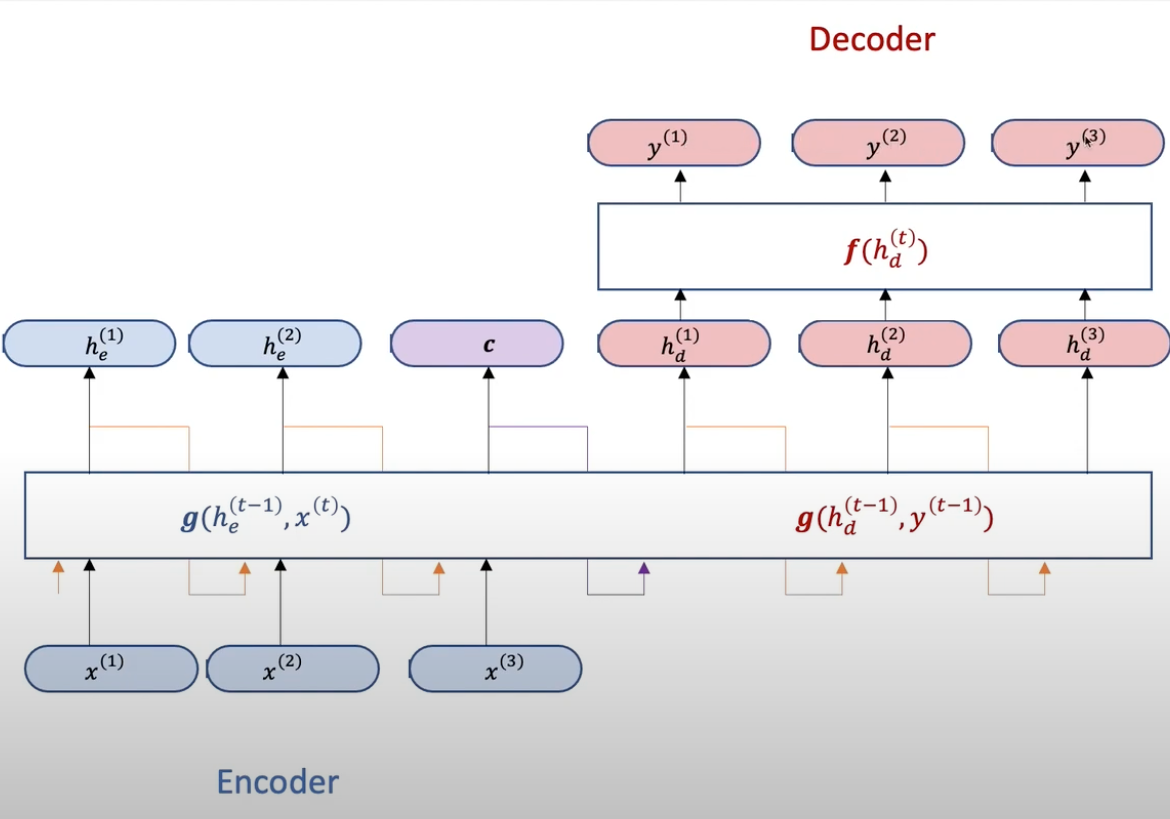
* **BIRDIRECTIONAL RNN**

In order to capture both ways of context

****

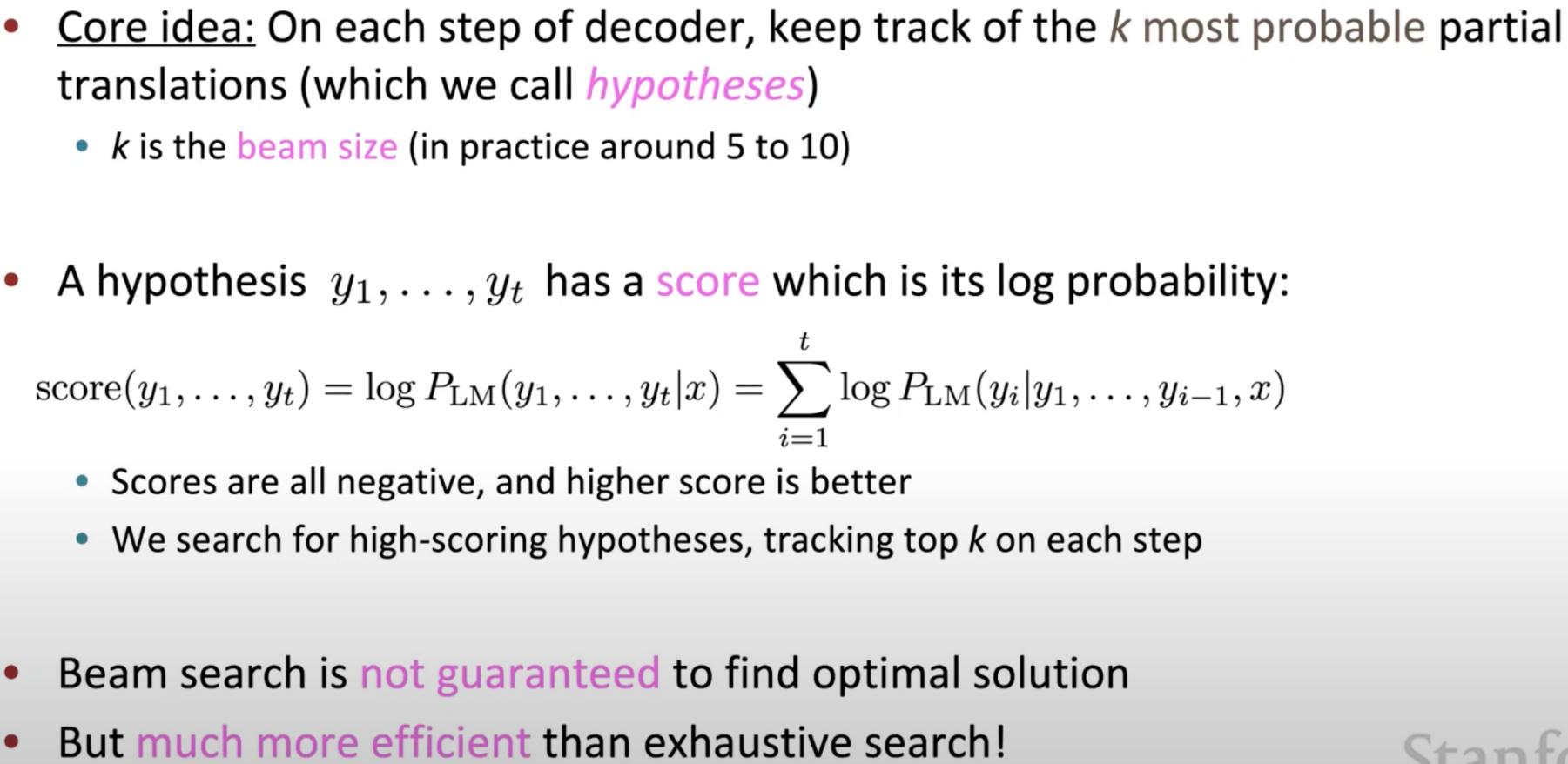
****

* **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)

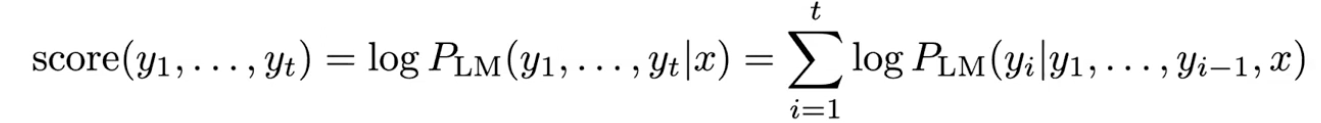
****

1. Stopping criterion

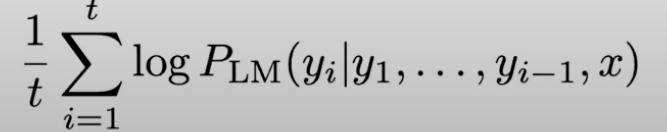


1. Problem

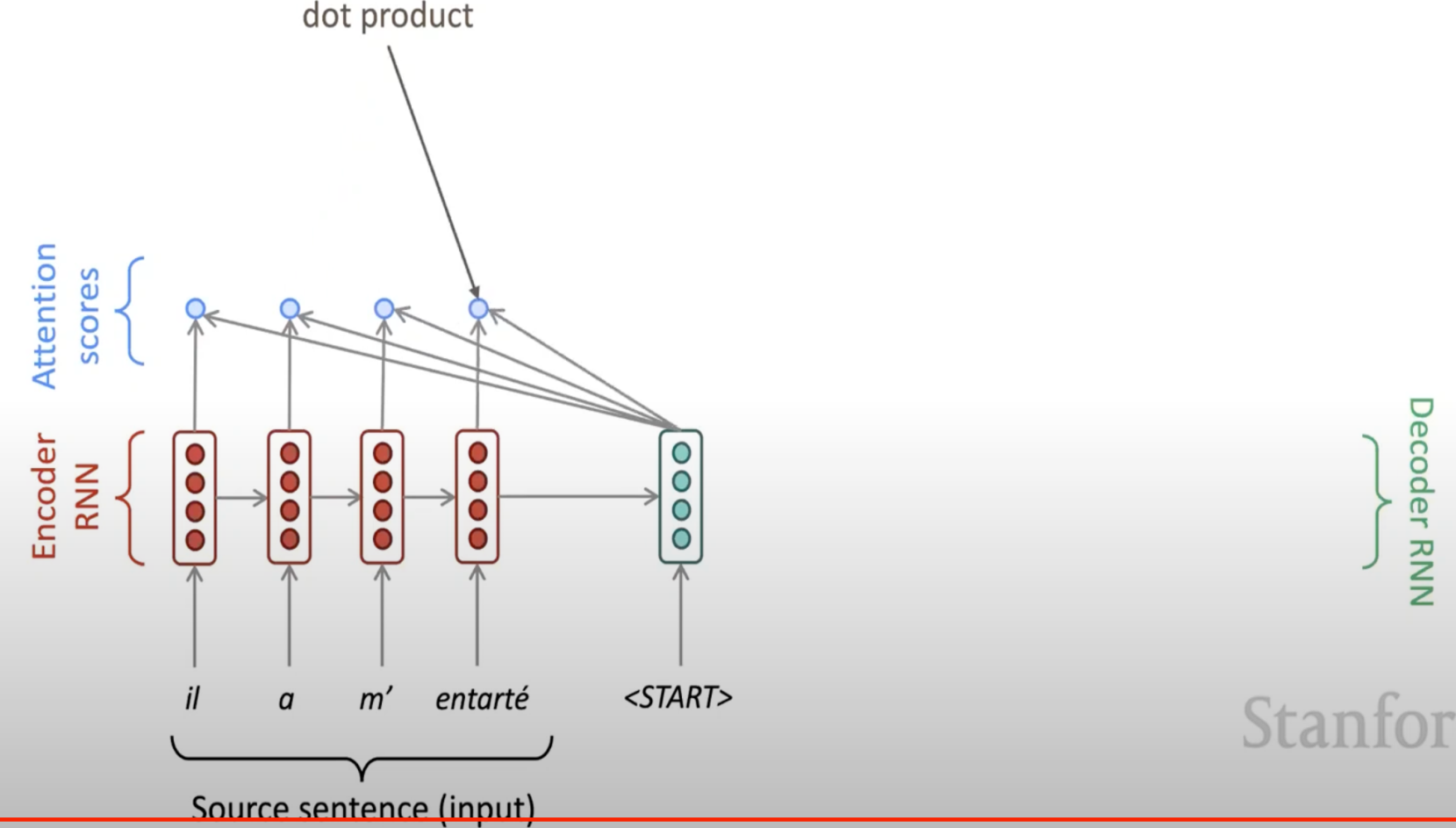
Longer sentences have lower score b/c accumulated multiplication



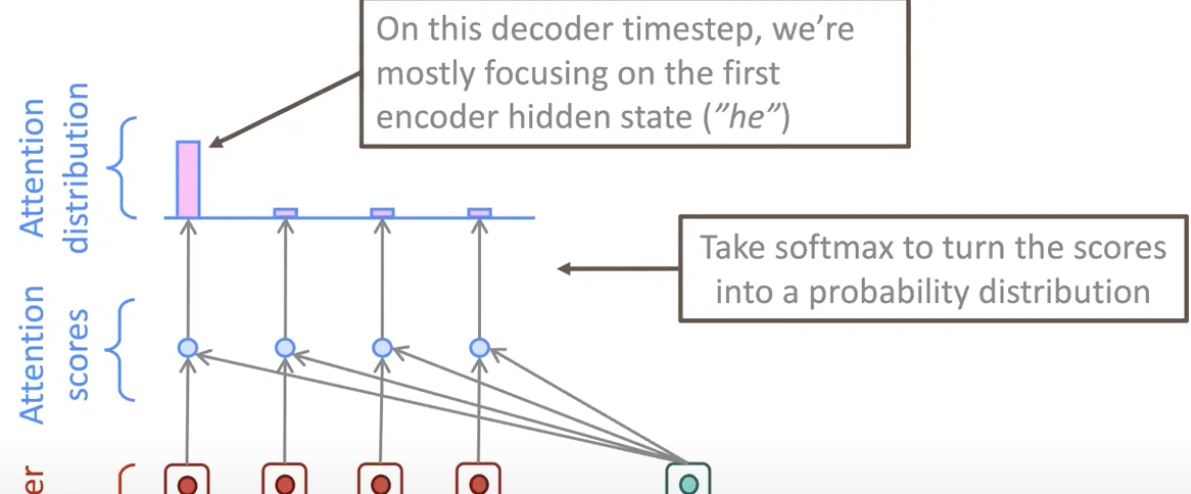
Fix: normalize be length, use this to select top one

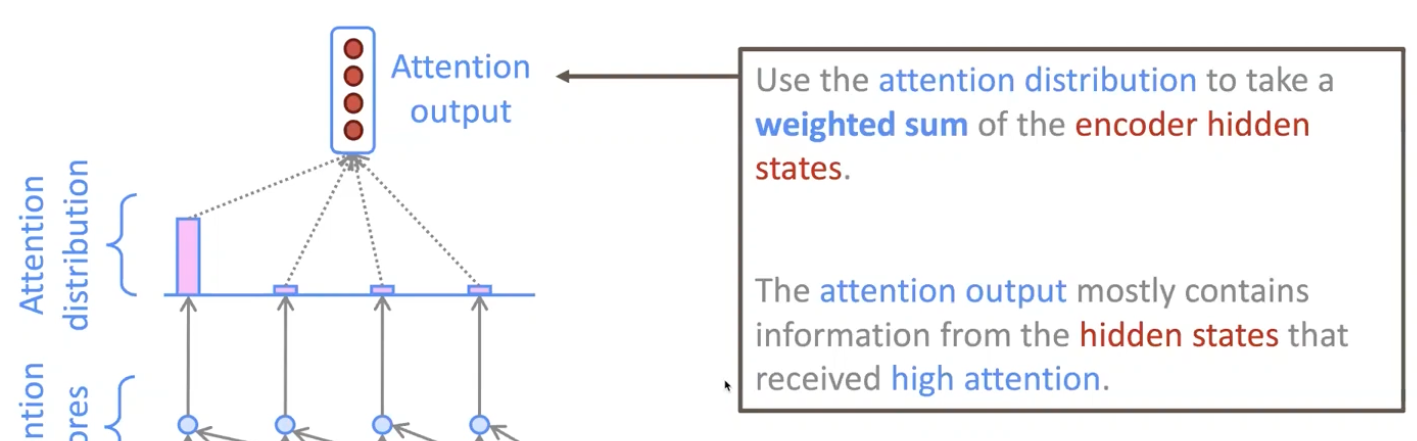


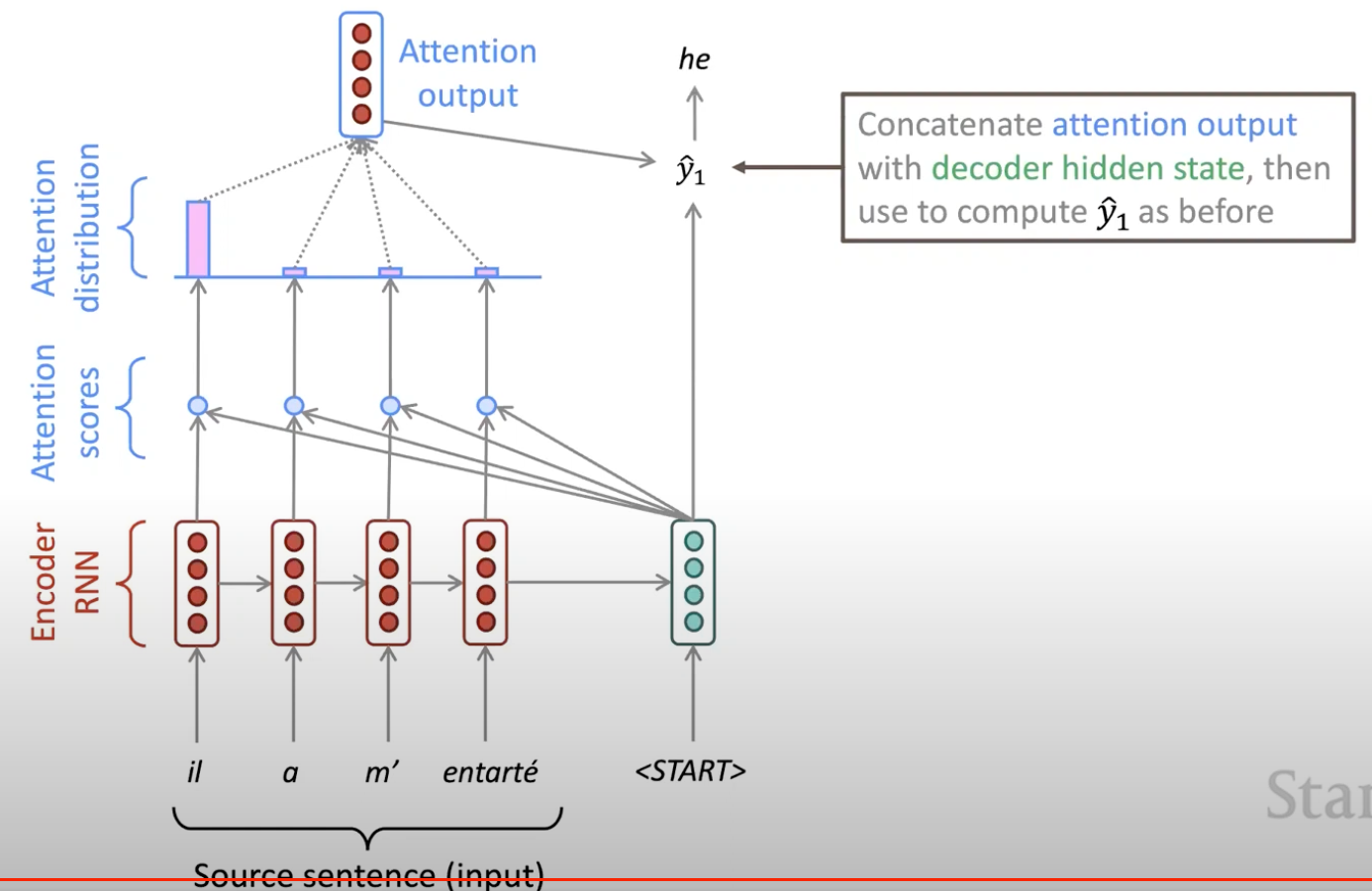
* [**ATTENTION**](https://youtu.be/wzfWHP6SXxY)

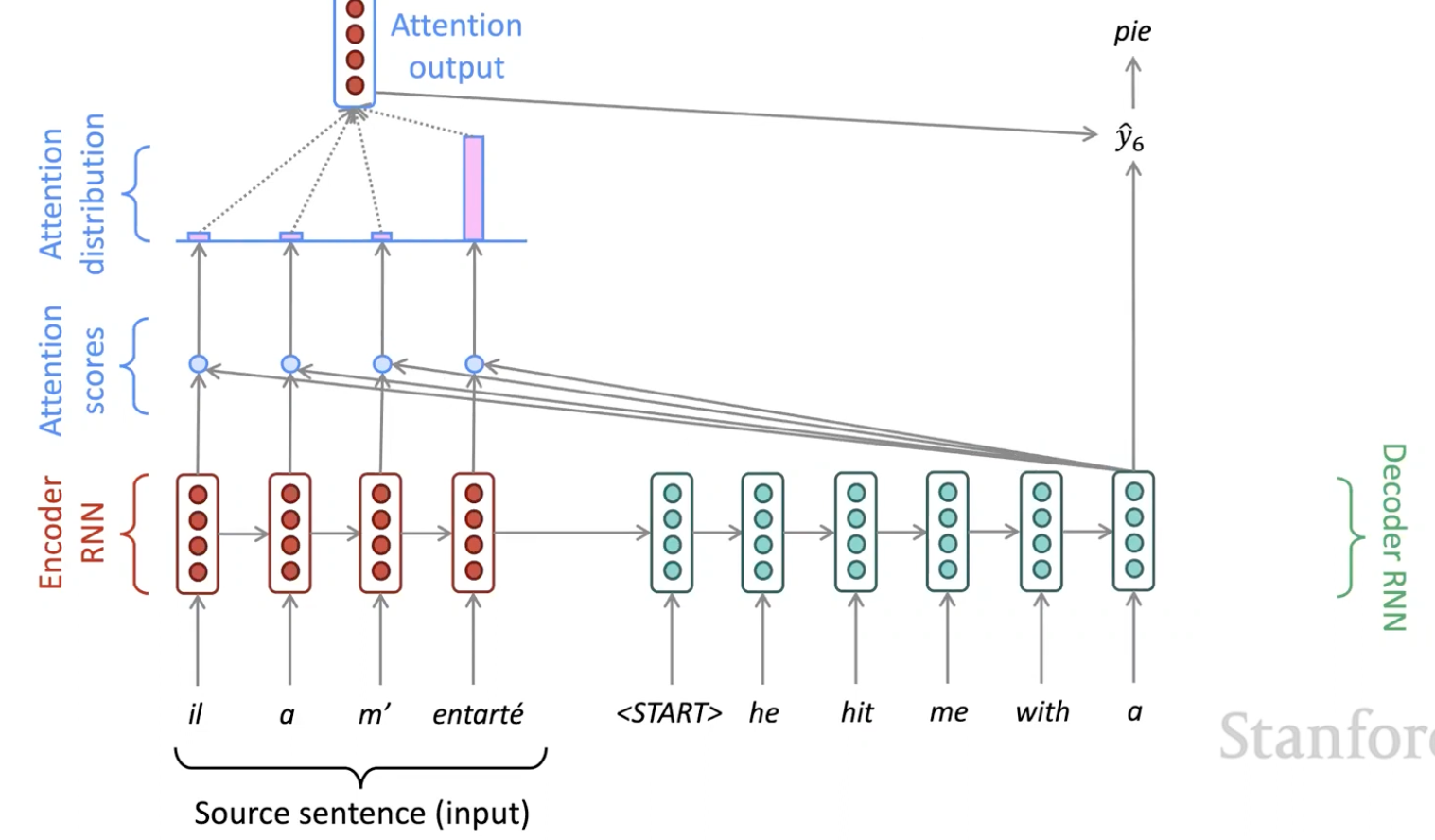


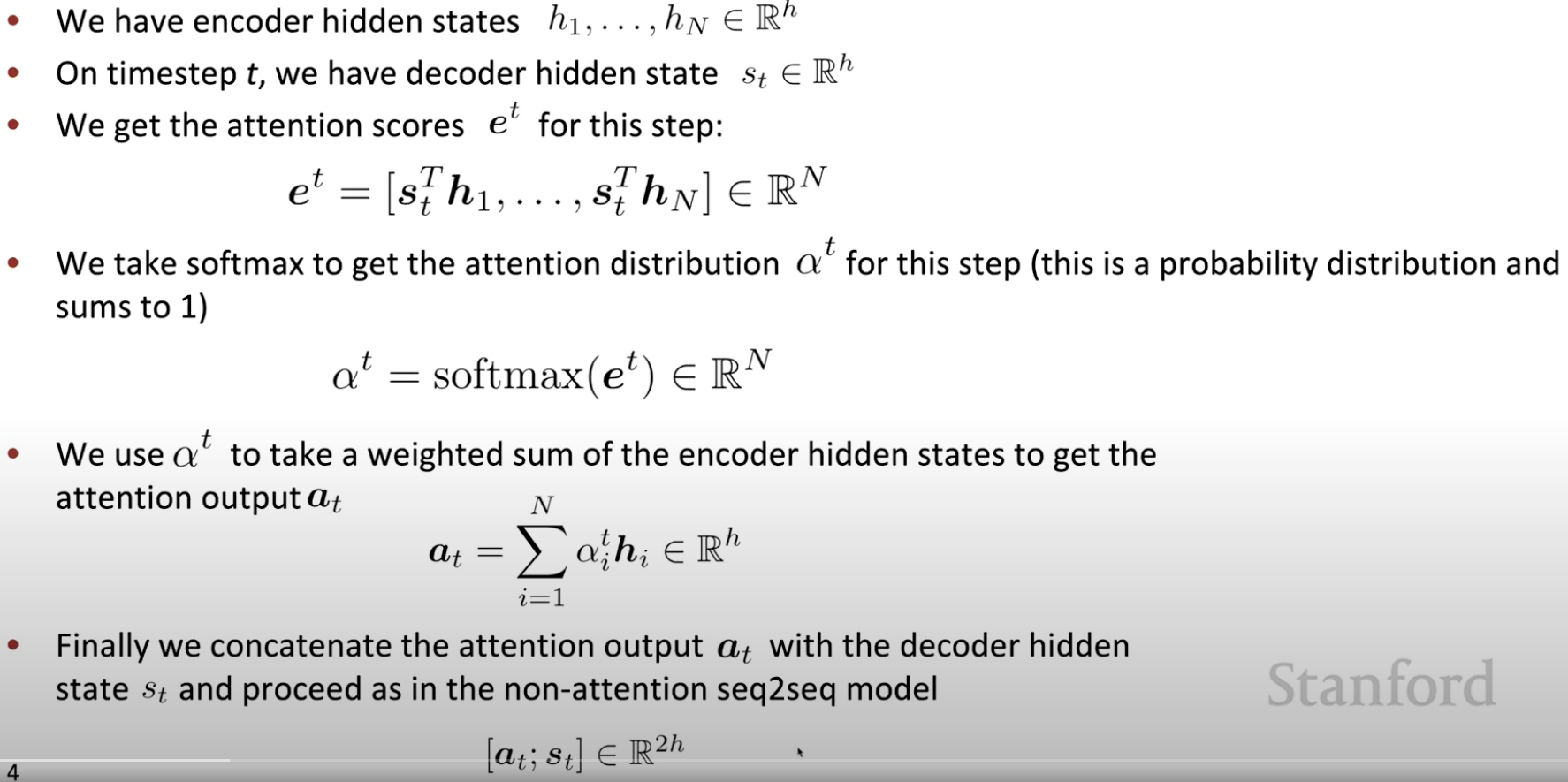
Find out which of the encoder state is most like the decoder state:





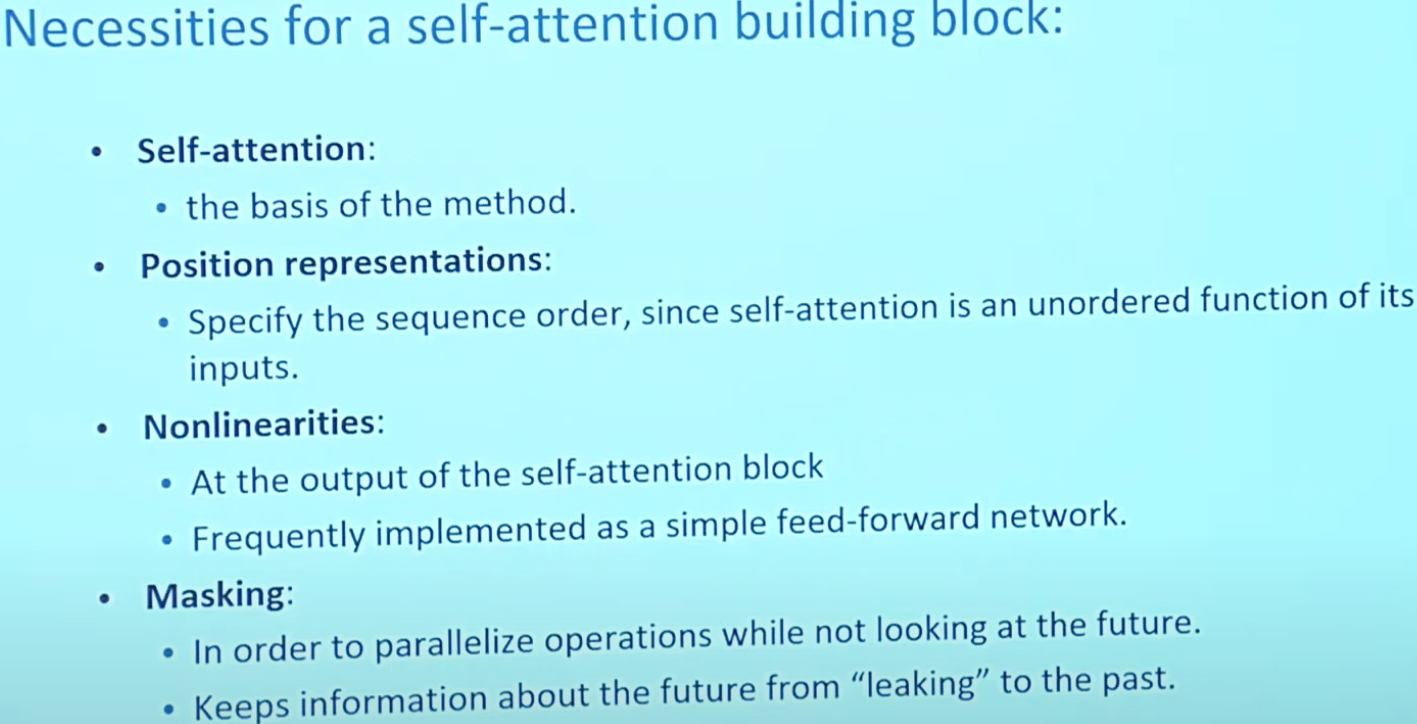




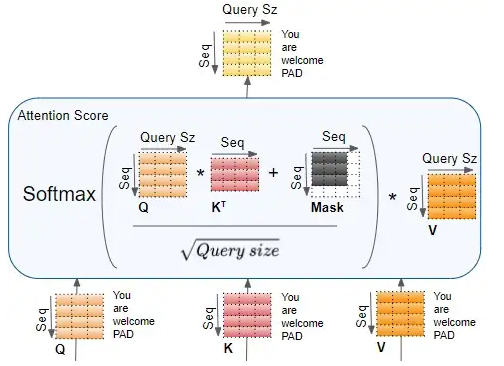


* [**TRANSFORMER**](https://towardsdatascience.com/transformers-explained-visually-part-2-how-it-works-step-by-step-b49fa4a64f34)

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

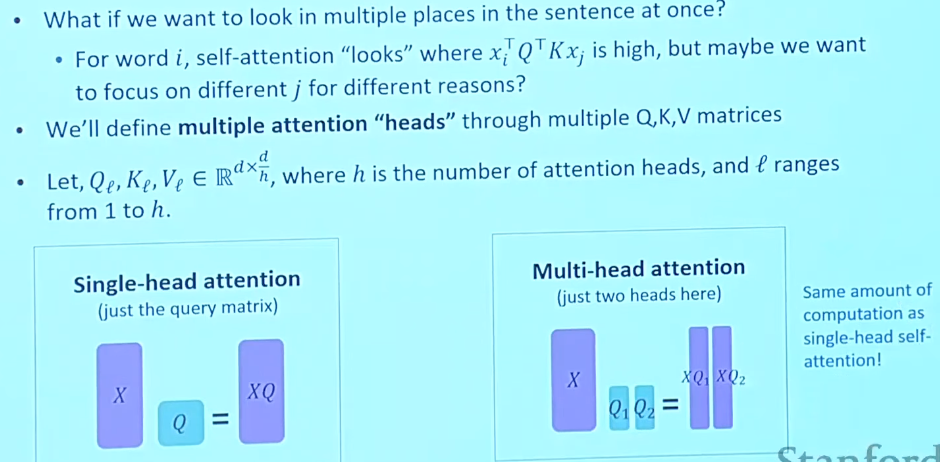


<https://towardsdatascience.com/transformers-explained-visually-not-just-how-but-why-they-work-so-well-d840bd61a9d3>

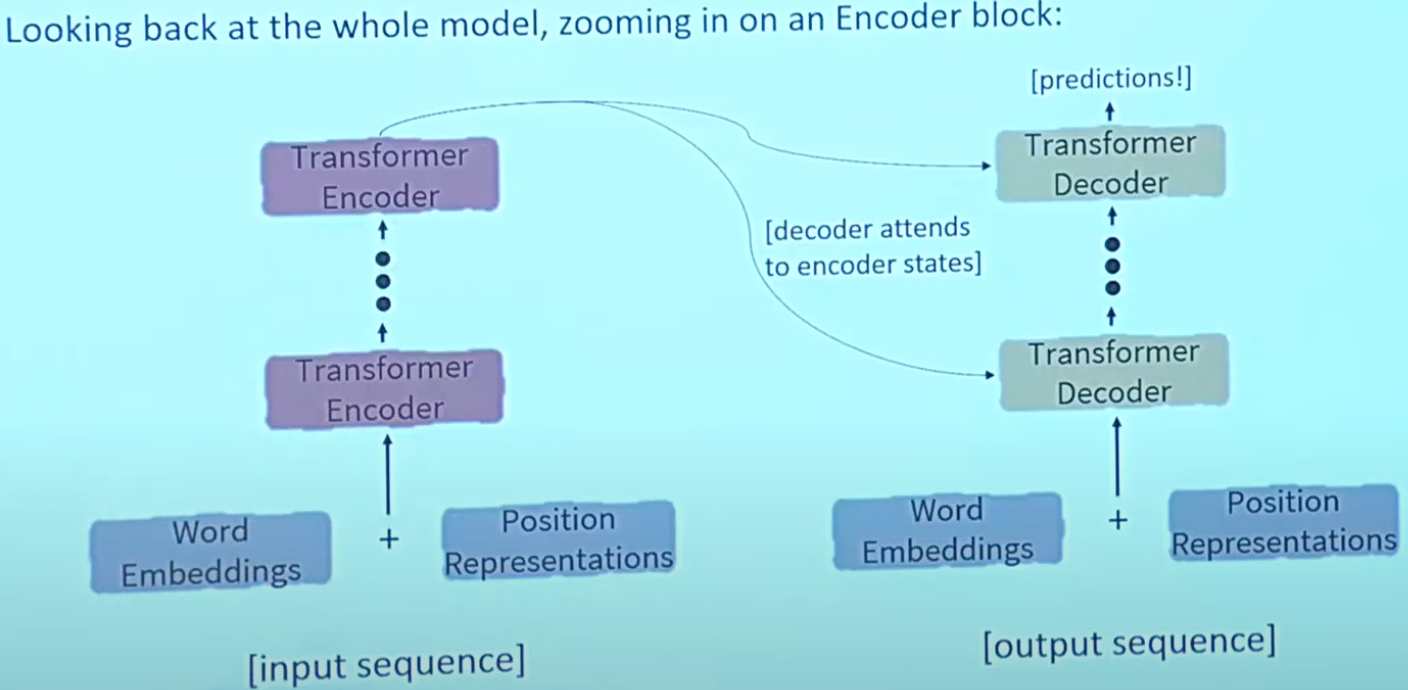


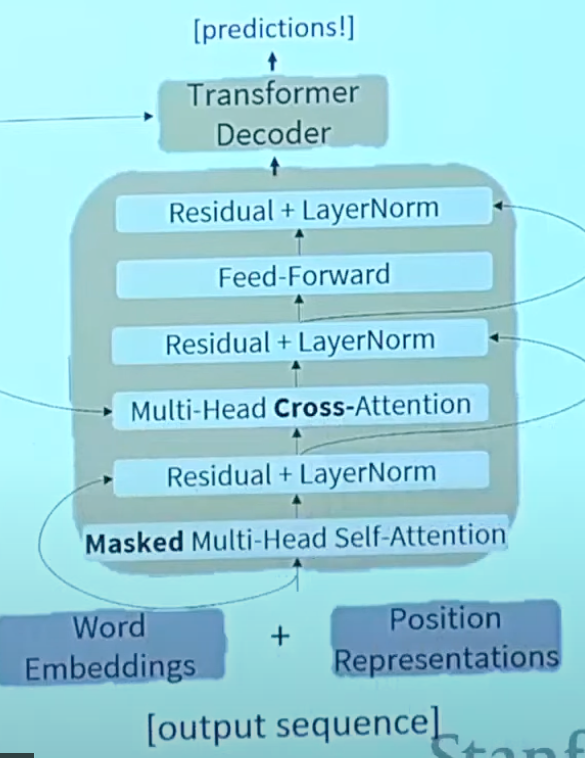
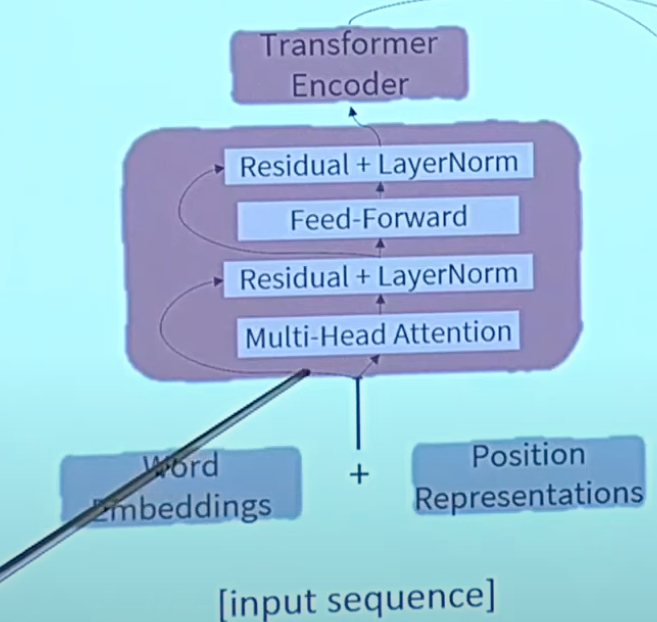
1. Multi-head attention

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



1. Transformer





1. Pretraining