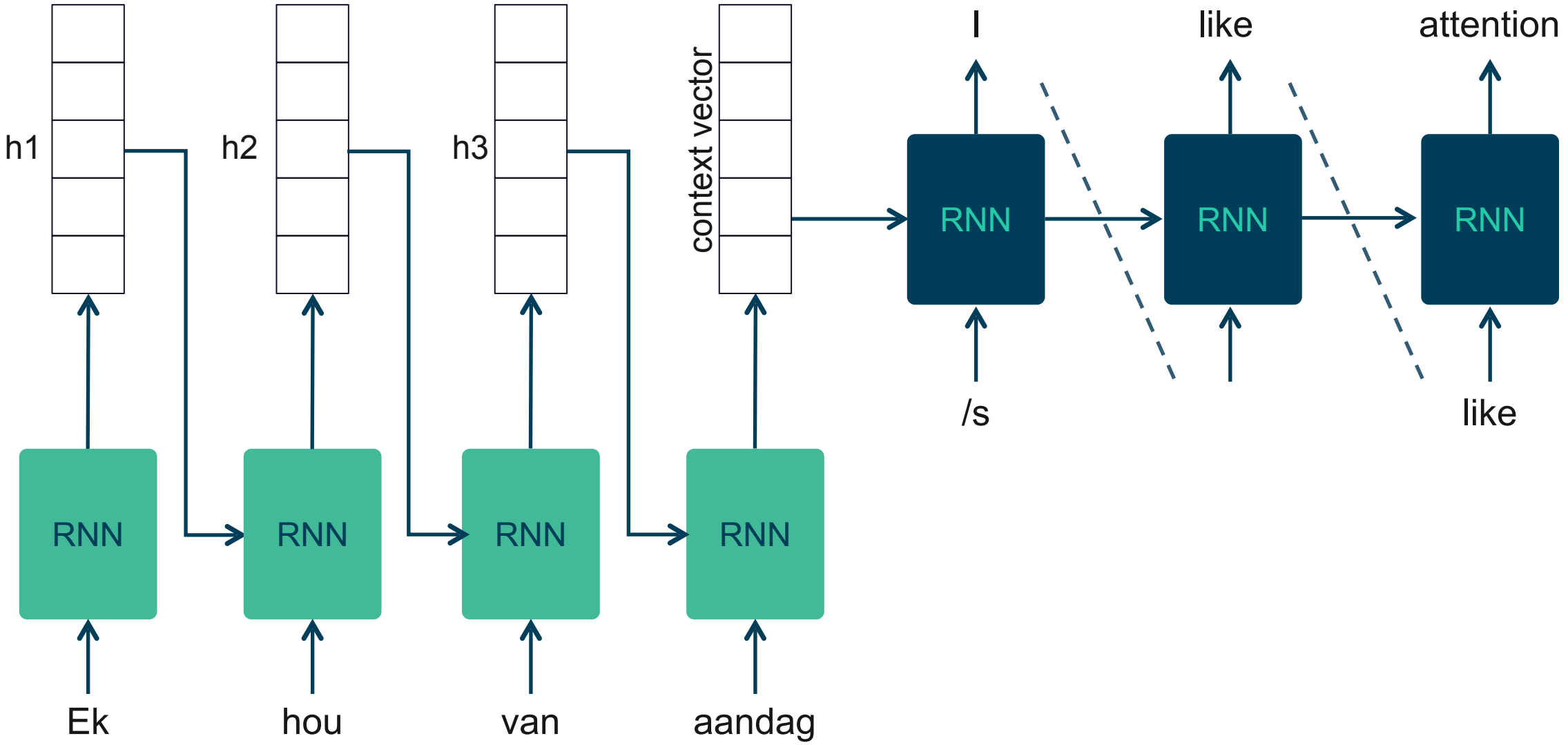
Hands on Generative AI

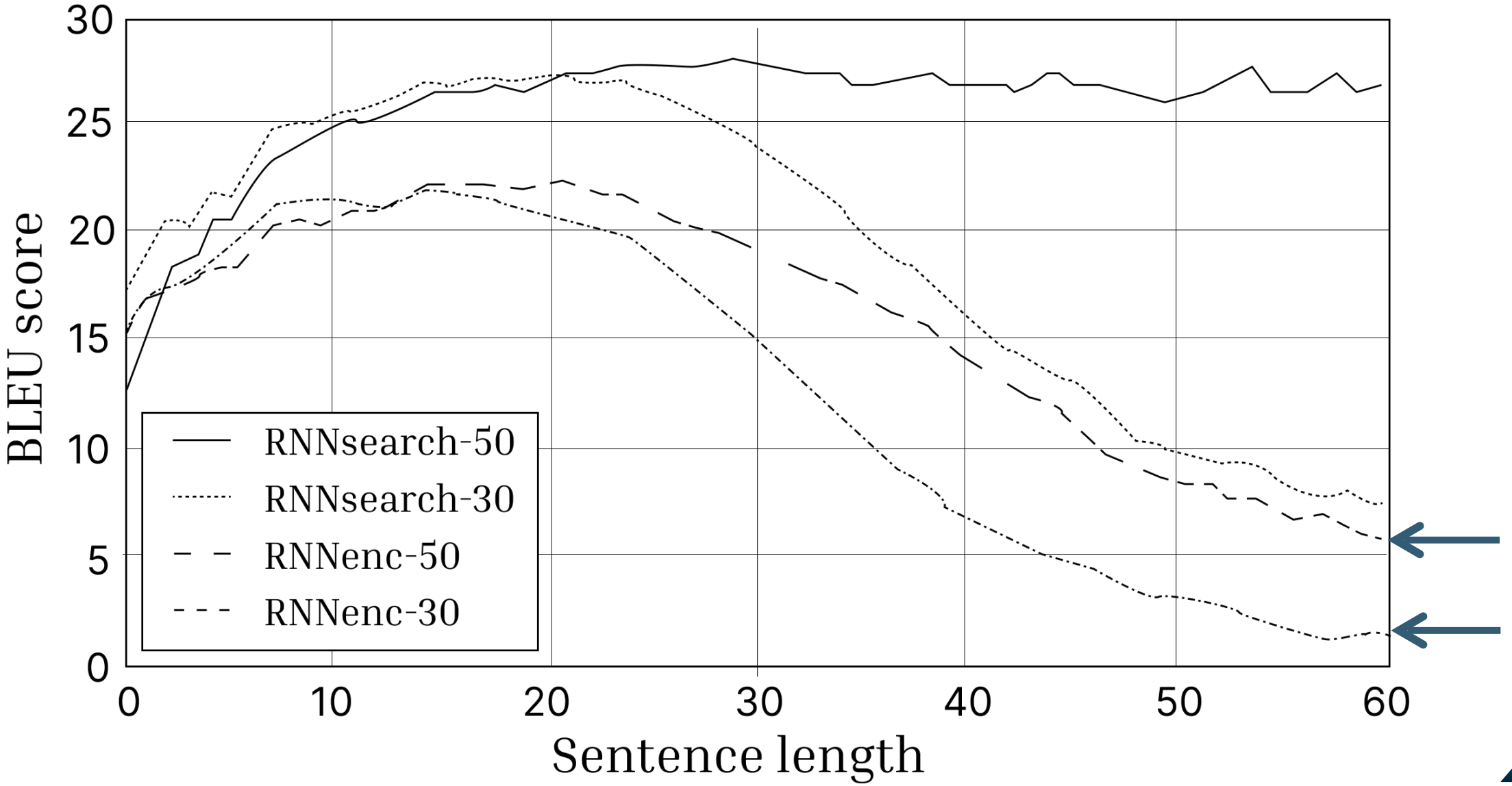
Introducing attention

Data Trainers

Attention

- We need to be able to provide good translations to an input text in, let’s say, Hungarian.


h1
h2
h3
context vector
RNN
RNN
RNN
RNN
Ek
hou
van
aandag
I
like
attention
RNN
RNN
RNN
/s
like


BLEU score
30
25
20
15
10
5
0
0
10
20
30
40
50
60
Sentence length
RNNsearch-50
RNNsearch-30
RNNenc-50
RNNenc-30

What is Attention?

I love coding on my Apple computer

I code on my computer, while eating an apple

- In math terms, this means that there exists a Tensor

$$a_{i,j} = f(h_i, s_j)$$

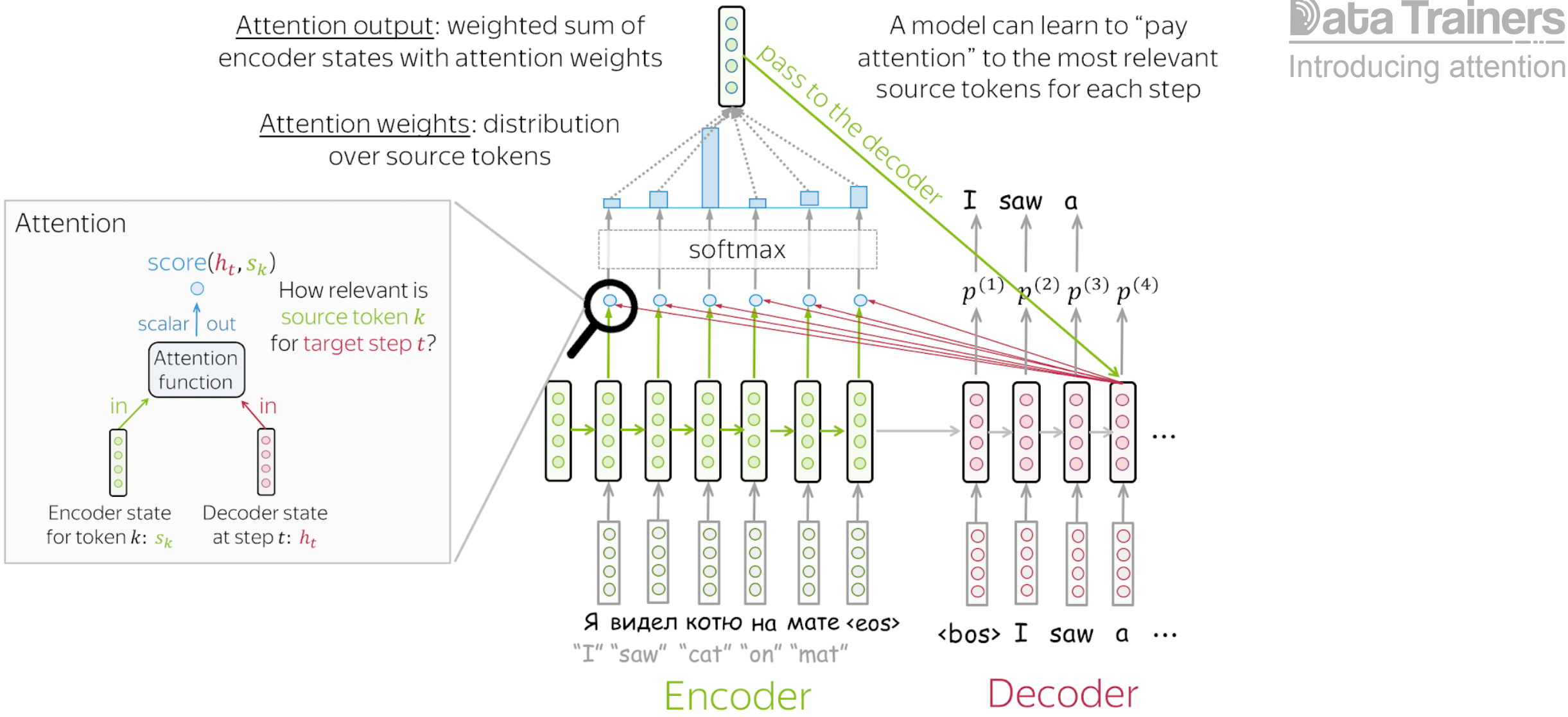
Neural Machine Translation by Jointly Learning to Align and Translate by **Bahdanau**

- They proposed the dot product

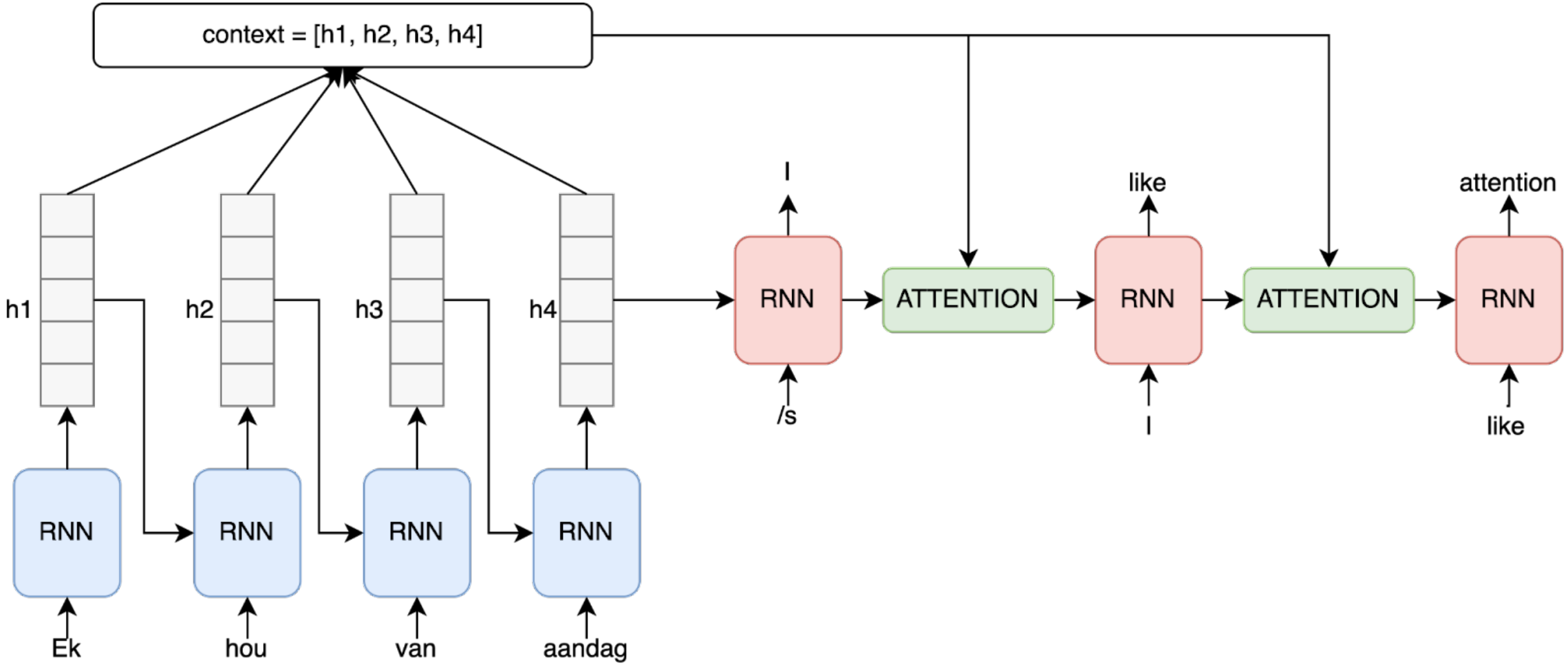
$$a_{i,j} = h_i^T * s_j$$

- We get a new tensor but the weights are not probabilities, so we get softmax

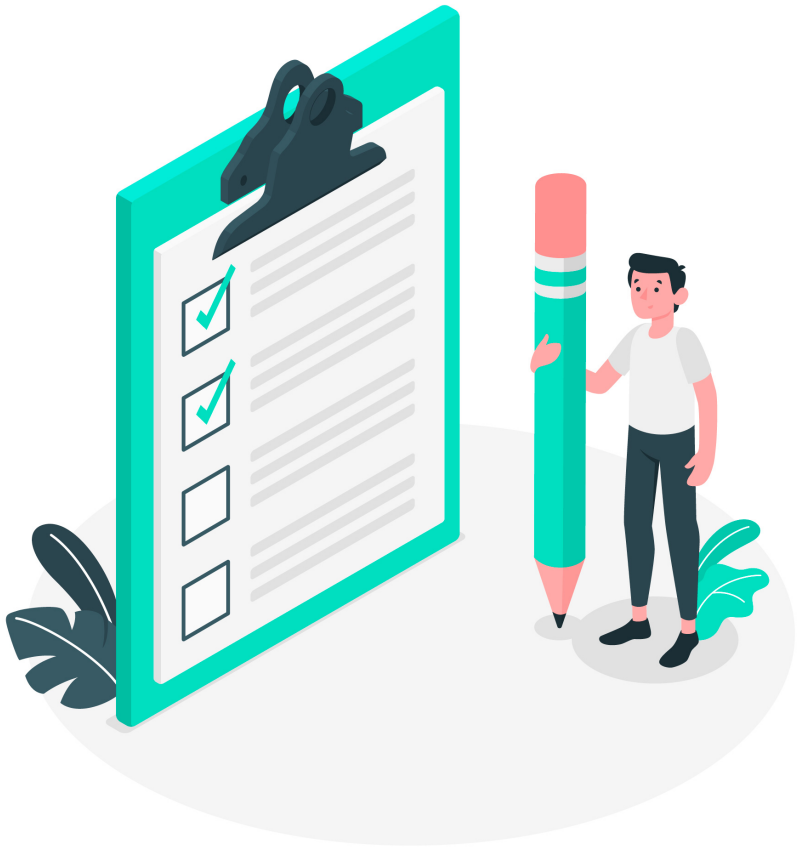
$$c_k = softmax(\sum_i a_{i,k} * h_i)$$



Source: https://lena-voita.github.io/nlp_course/seq2seq_and_attention.html#attention_intro


context = [h1, h2, h3, h4]
h1
h2
h3
h4
RNN
RNN
RNN
RNN
Ek
hou
van
aandag
RNN
I
/s
ATTENTION
RNN
like
I
ATTENTION
RNN
attention
like

- Calculating attention by hand



Before learning self-attention

ID	Item	Price
1	Shampoo	4,99
2	Conditioner	2,99

- First, we will calculate the similarity between Q and K

$$a_{i,j} = similarity(\text{Q,K})$$

- Here **we will have for each row many values** different from zero, so later the output context vector is:

$$c_i = \sum_i a_{i,k} * h_i$$

Two Peculiarities

- K = V

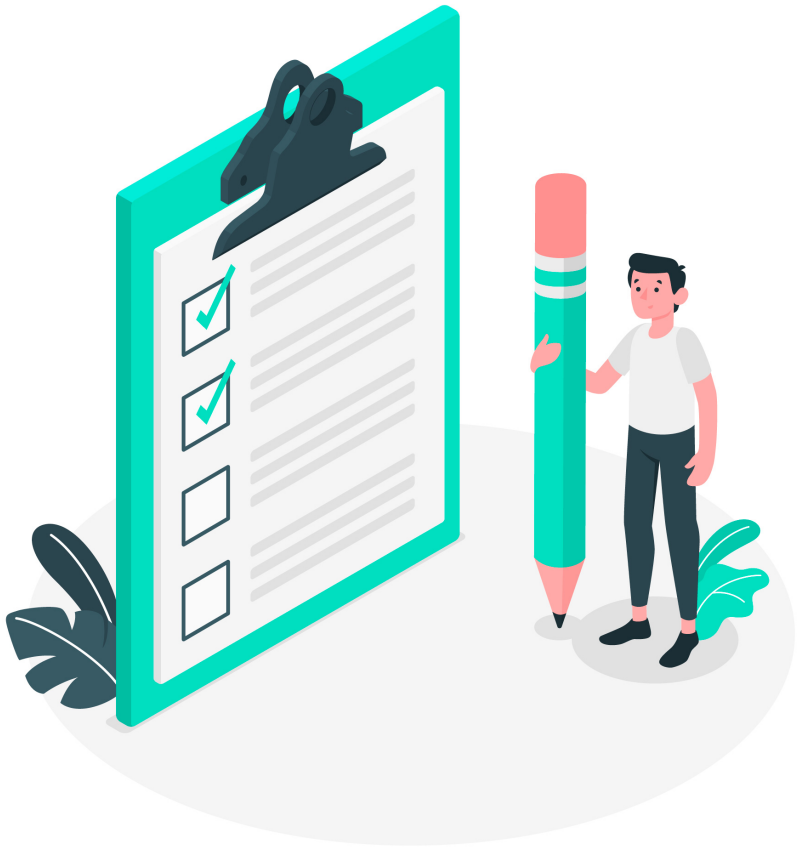
  This is why we say it is self-attention, because we do attention of the tensor with itself.

- The similarity function is the following:

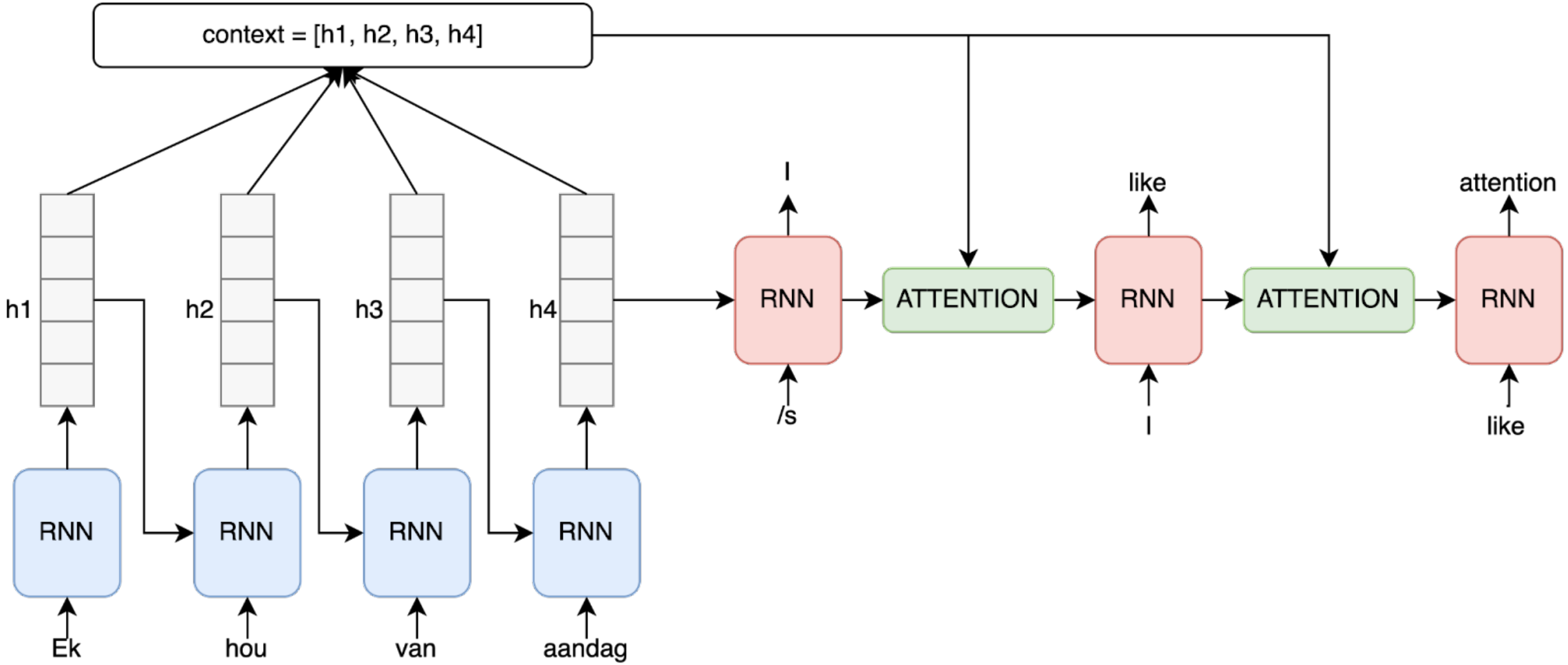
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Lab

- Adding attention layers with frameworks

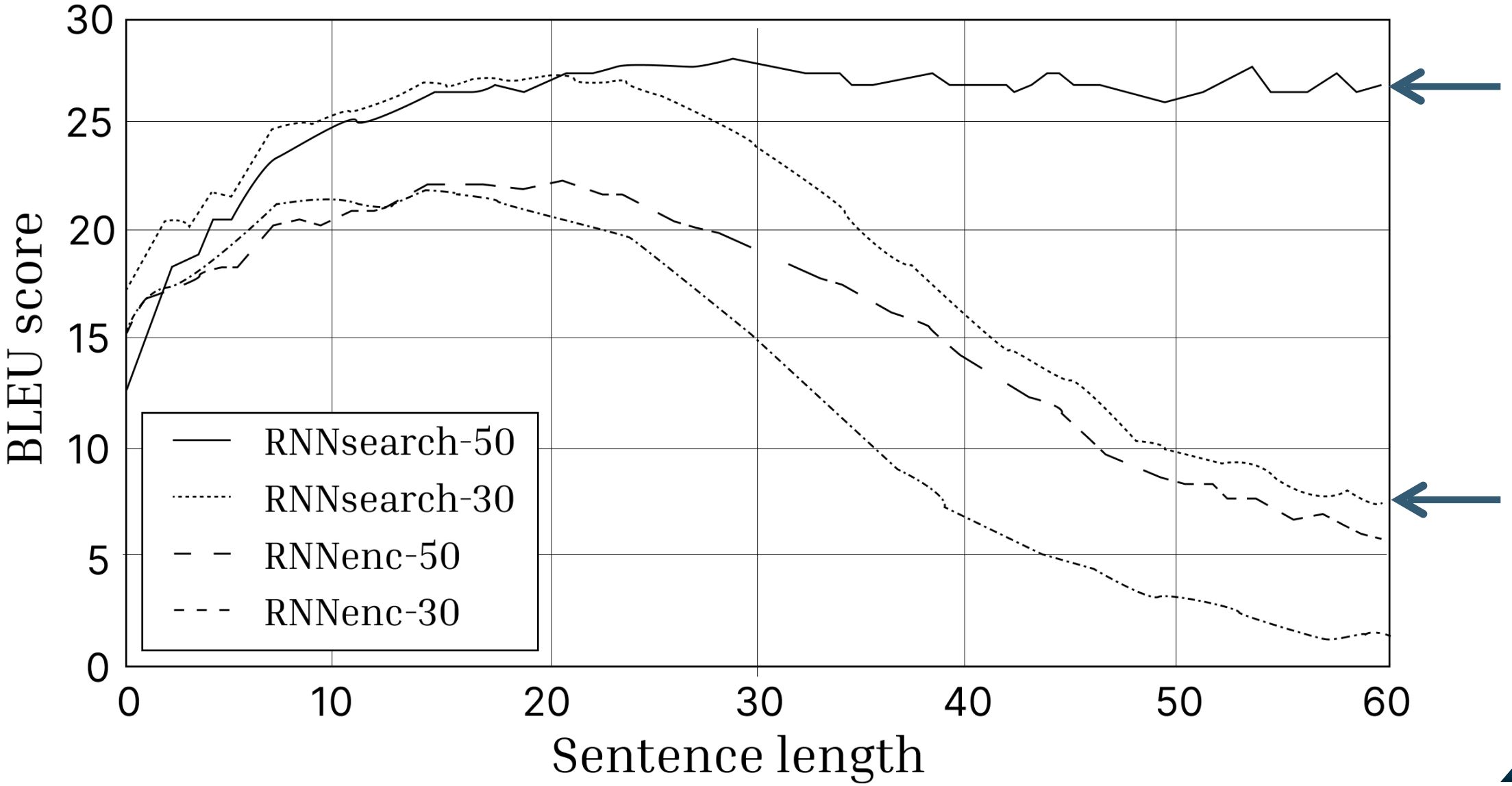


Recap: Seq2Seq with attention


context = [h1, h2, h3, h4]
h1
h2
h3
h4
RNN
RNN
RNN
RNN
Ek
hou
van
aandag
RNN
I
/s
ATTENTION
RNN
like
I
ATTENTION
RNN
attention
like

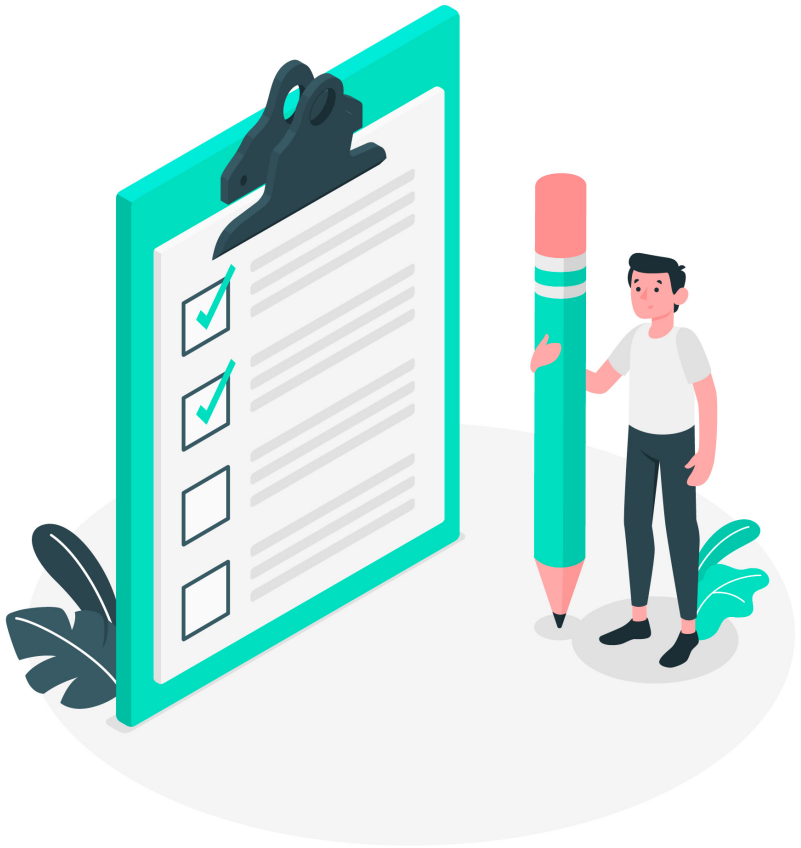
There are many ways

- One is to pick the máximum probability – greedy search it is called and has one issue: it over picks simple words like the, a, and such
- Another one is called beam search, which you can take as a homework to learn.
- Finally, the one we will use is picked according to the distribution of probability, so maybe we don’t pick the highest scored word at a given point in time.


BLEU score
30
25
20
15
10
5
0
0
10
20
30
40
50
60
Sentence length
RNNsearch-50
RNNsearch-30
RNNenc-50
RNNenc-30

Lab

- Implementing Neural Machine Translation with attention



Summary

- Attention is just a mechanism that enables for each output token to treat, get different personalized context
- Dot product attention, being as simple as it is, is still able to detect relationships of words and is used in models nowadays
- We have built a full encoder decoder model, for that we needed to create the necessary layers
- These ideas actually generalise to any NLP task and pretty much everywhere. However after the next module that will become clearer