Machine Learning to Language Model

Topic 03 - Self-Attention

Jaihua Yen
https://jaihuayen.github.io/homeweb/

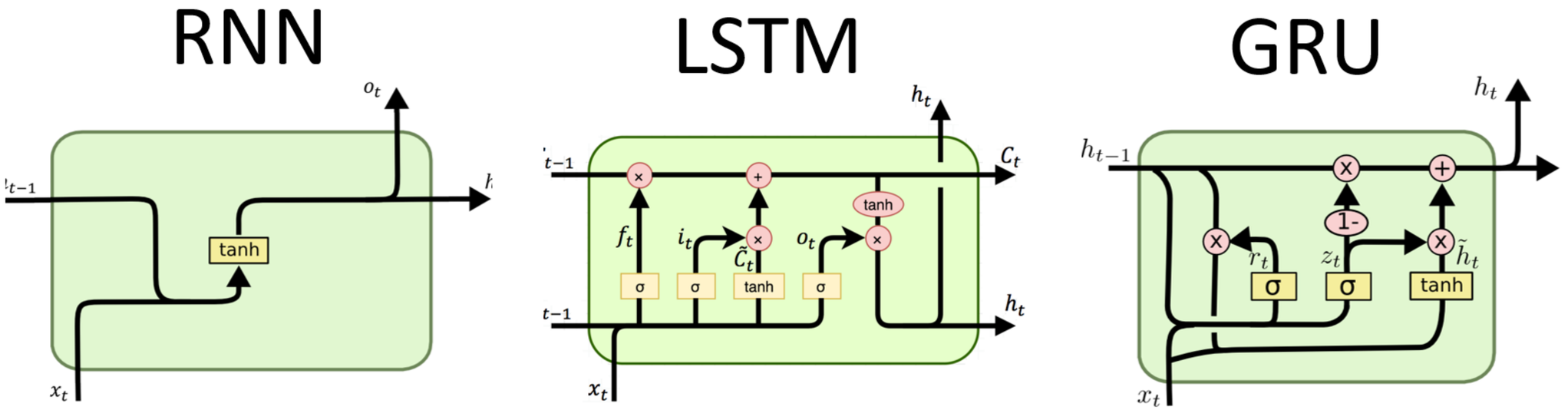
Contents

- Transformer Overview
- Self-Attention
- Wrap Up

Why Transformer?

Low Efficiency of Recurrent Neural Network

Step-by-Step is time-consuming

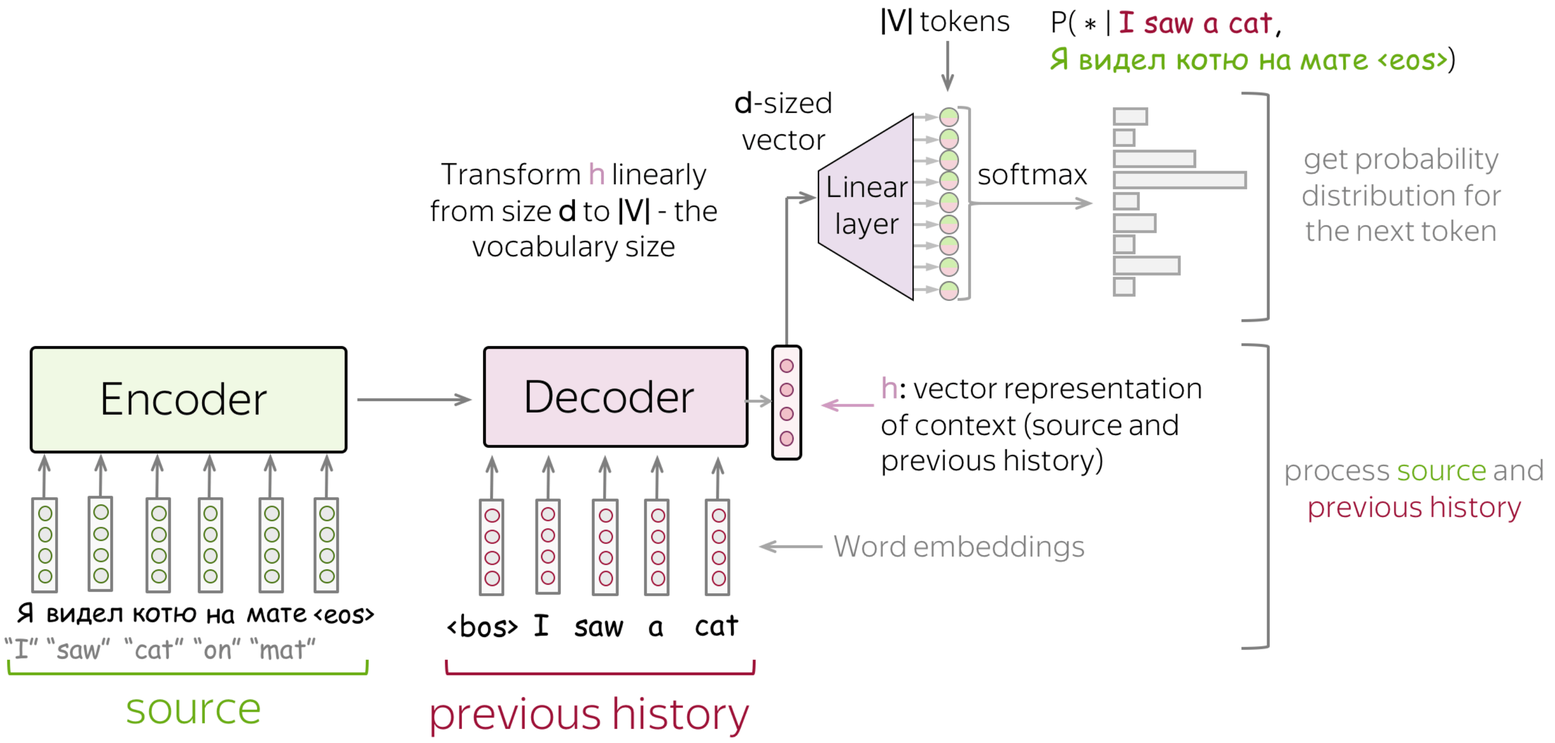


http://dprogrammer.org/rnn-lstm-gru

Lost Information in Recurrent Neural Network

The information in the beginning will degrade in the future steps

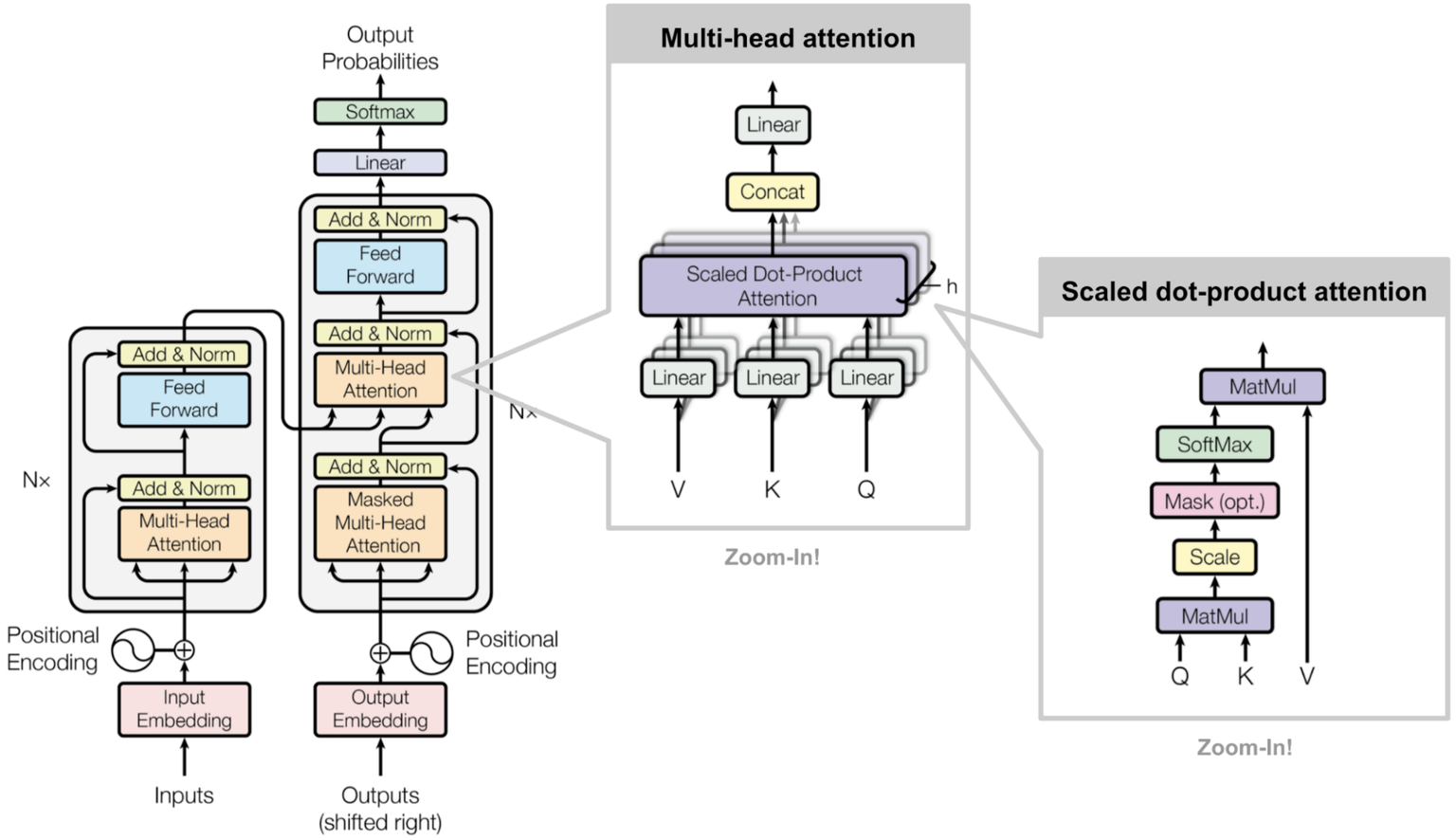
Is there a way to see the whole picture of the sentence at one time?



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

Transformer

The Key to the AI Era

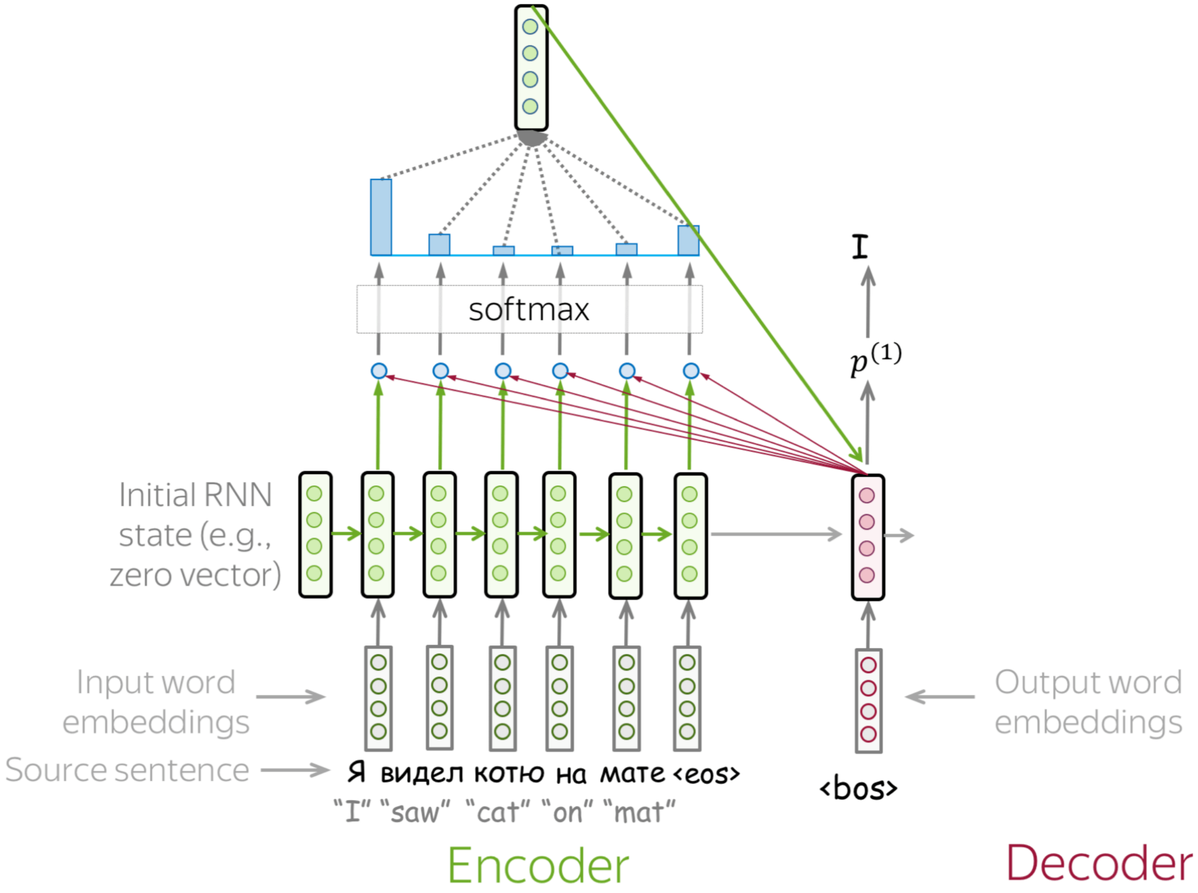


https://lilianweng.github.io/posts/2018-06-24-attention/

Self-Attention

Attention Mechanism

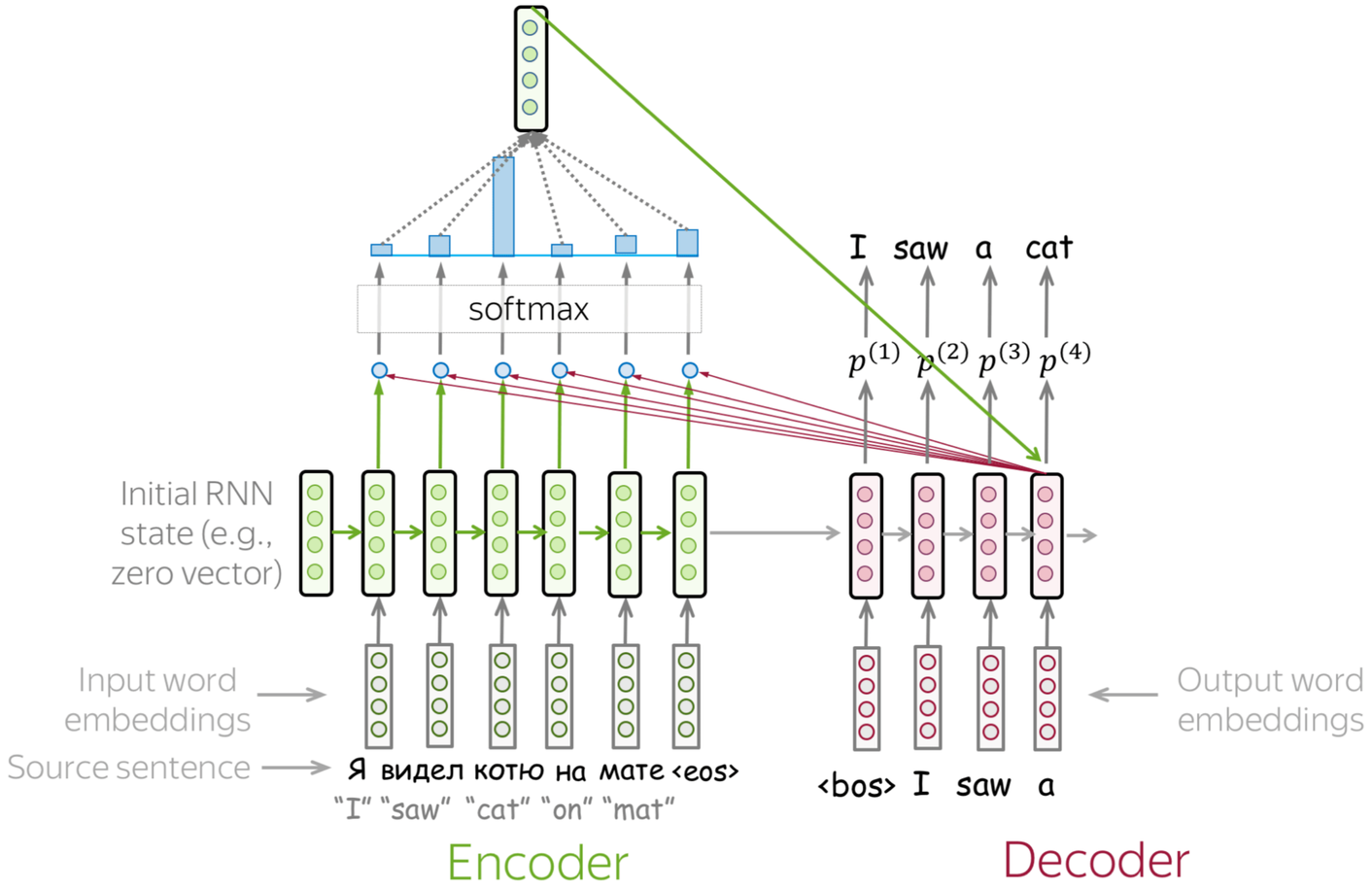
Remain the information of all words in all steps



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

Attention Mechanism

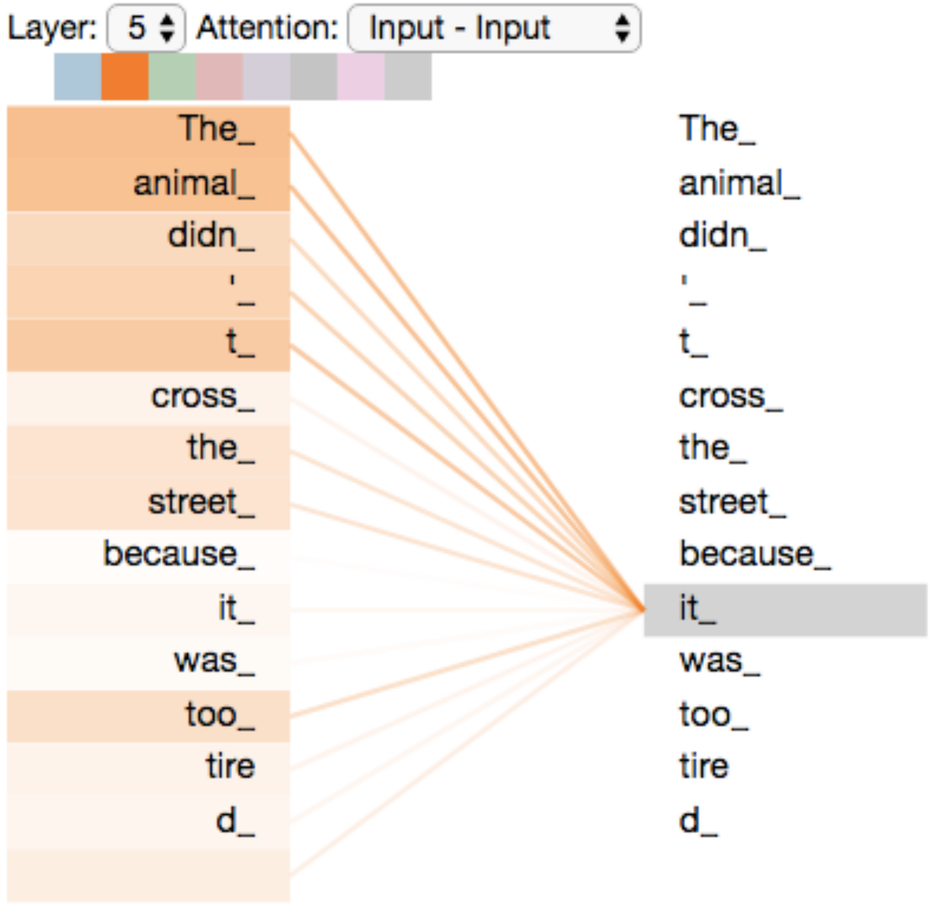
Remain the information of all words in all steps



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

Self-Attention

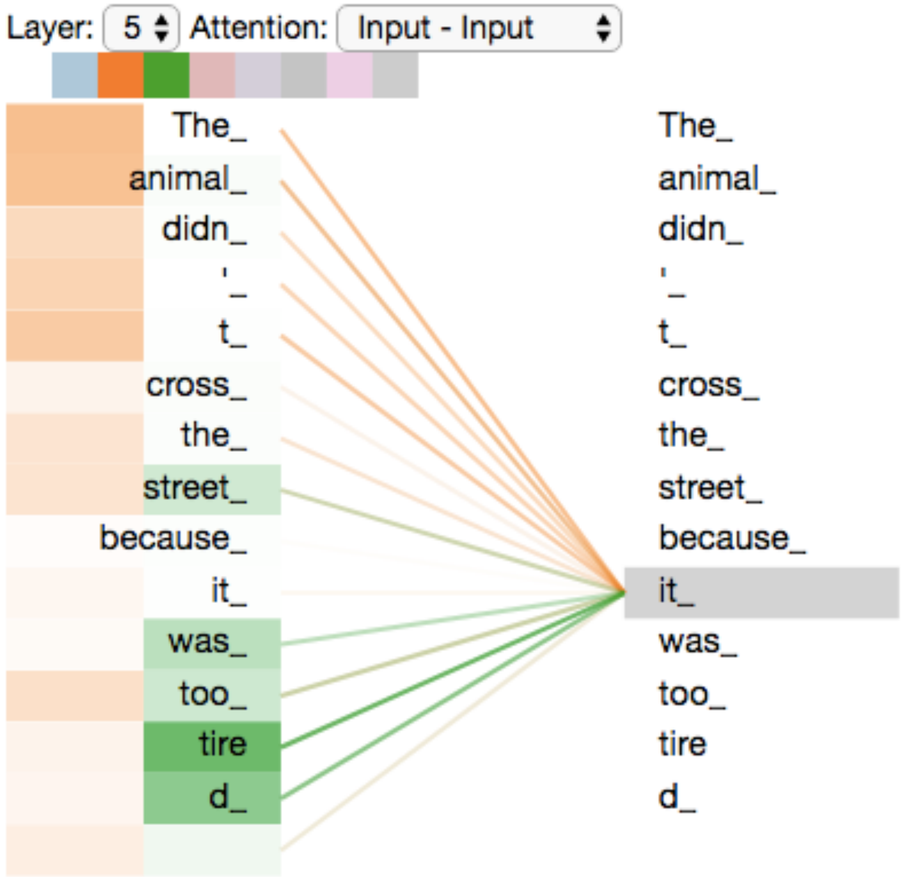
Attention to the same sentence



https://jalammar.github.io/illustrated-transformer/

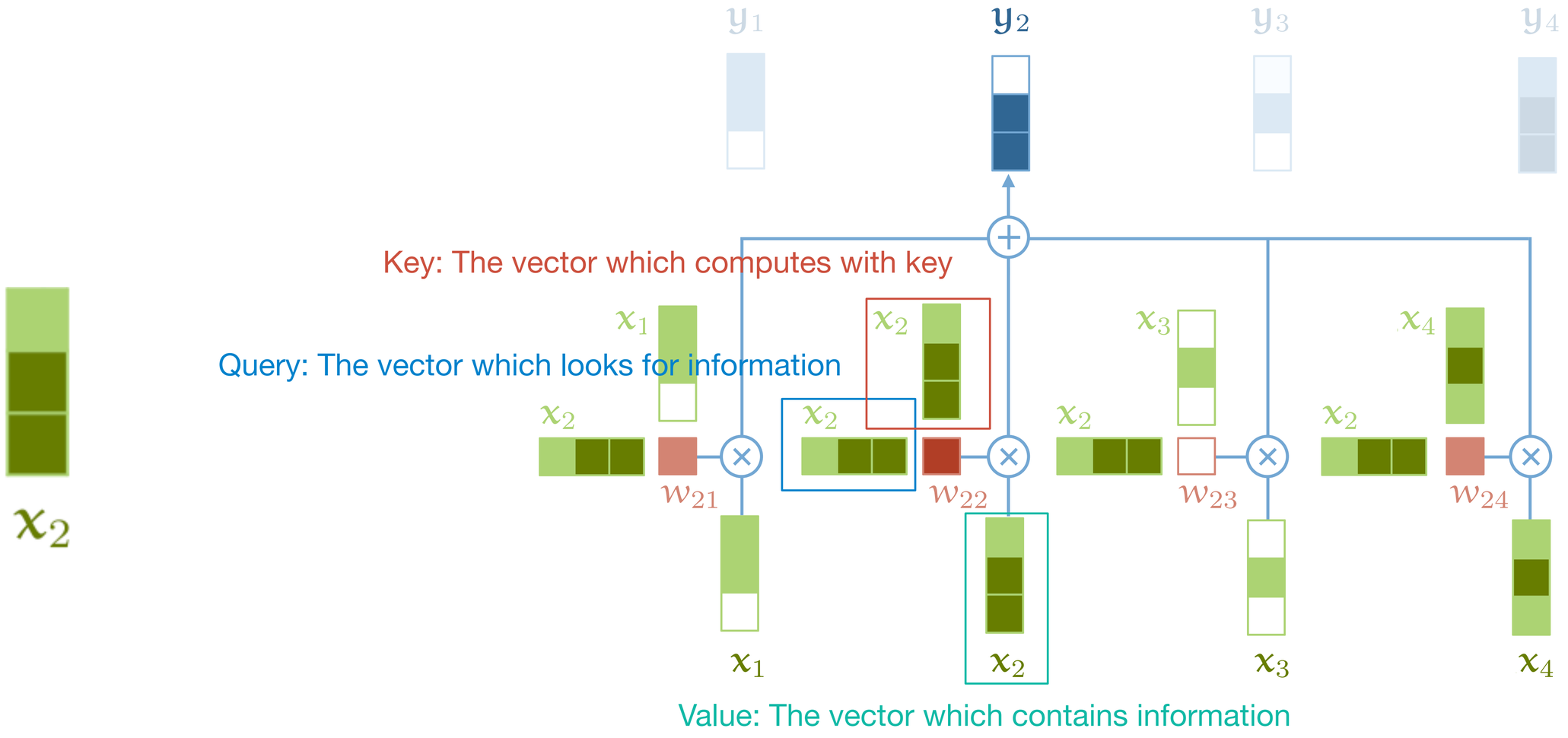
Multi-Head Self-Attention

Attention to the same sentence



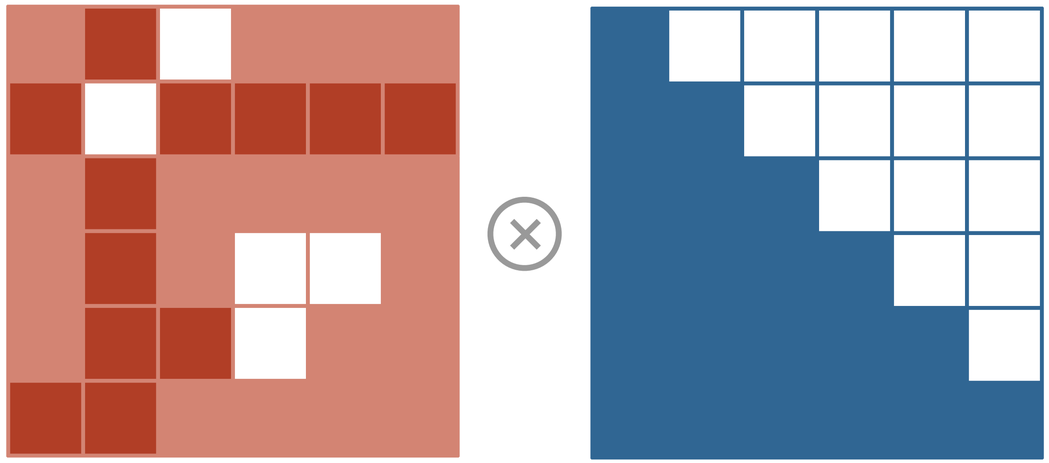
https://jalammar.github.io/illustrated-transformer/

Query Key Value in Self-Attention



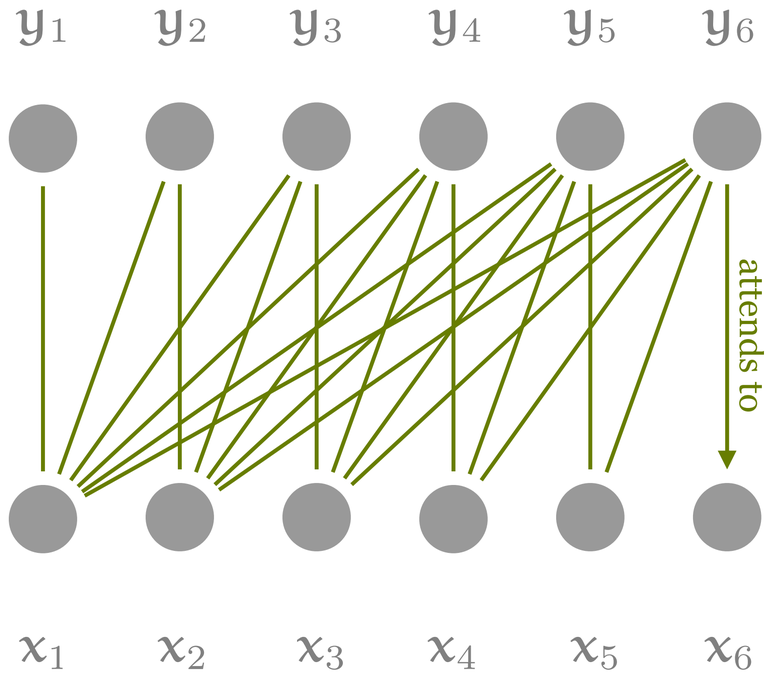
https://peterbloem.nl/blog/transformers

Masked Self-Attention



raw attention weights

mask



https://peterbloem.nl/blog/transformers

Let’s do this in Colab!

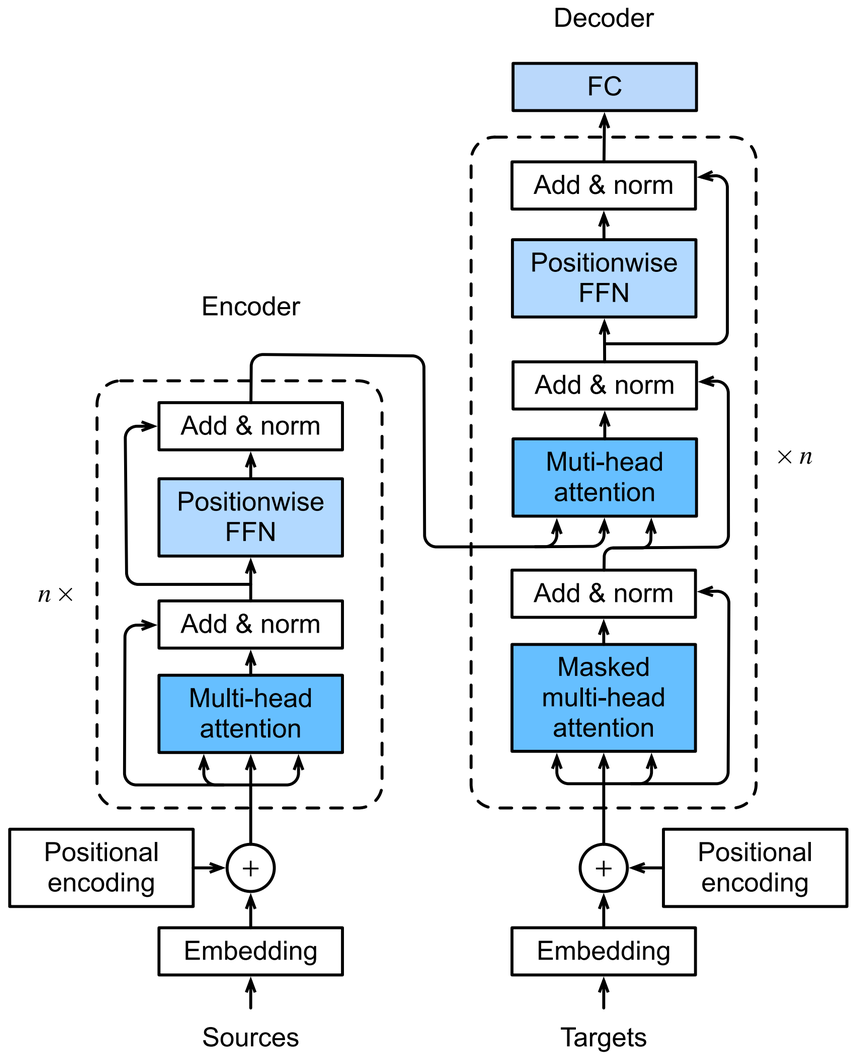
Wrap Up

What We Have Gone Through

- Overview of Transformer
- Self-Attention Mechanism

What's Next

- Transformer Encoder
- Transformer Decoder
- Positional Encoding
- Wrap Up Transformer



Q & A