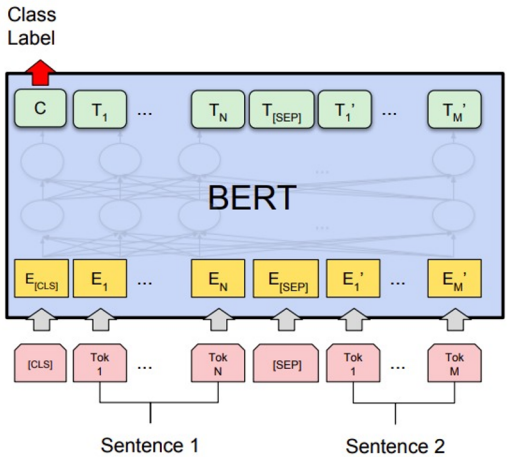
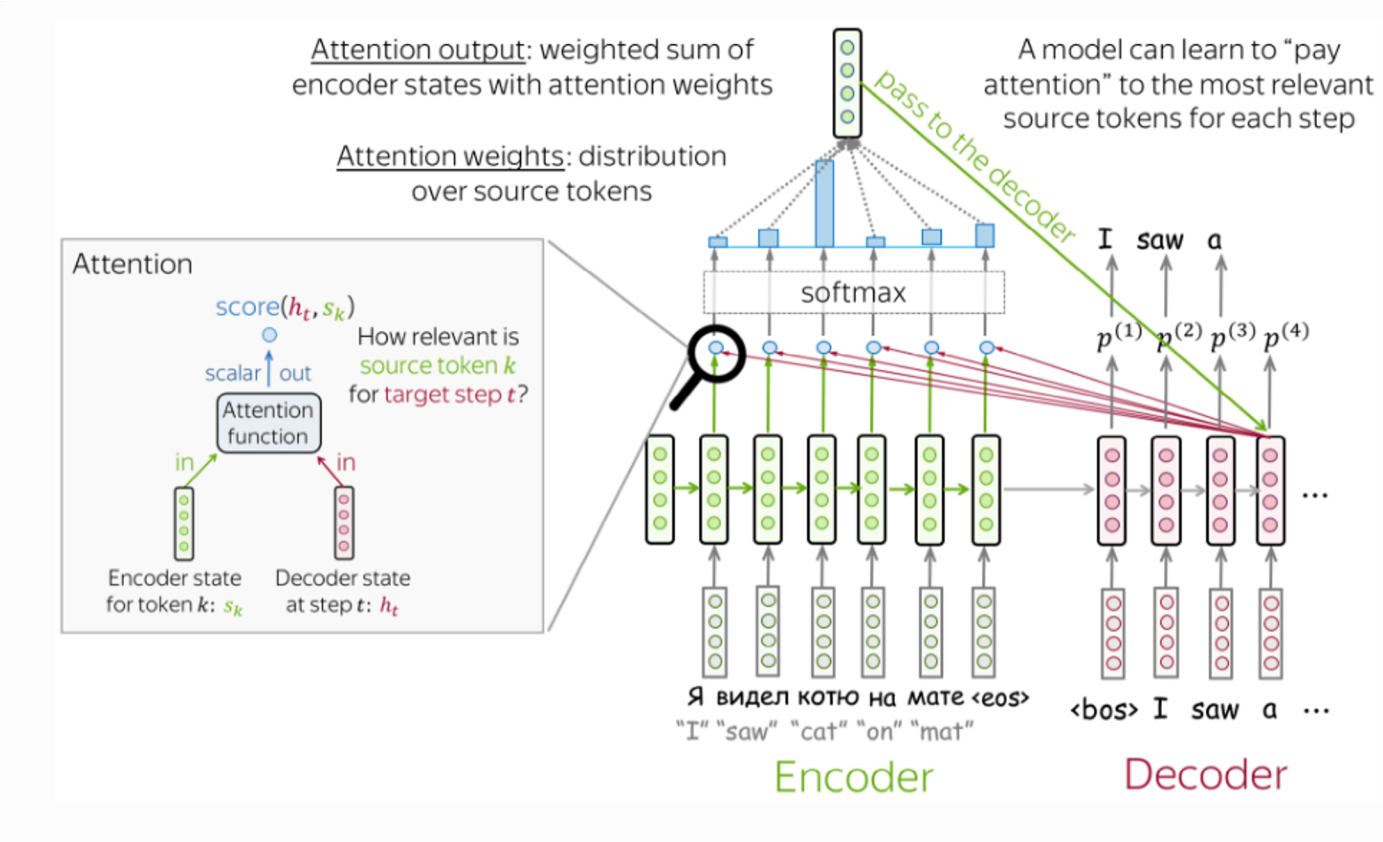
Seminar 9: transformers

Actually, AI does not exist. It is just ML.



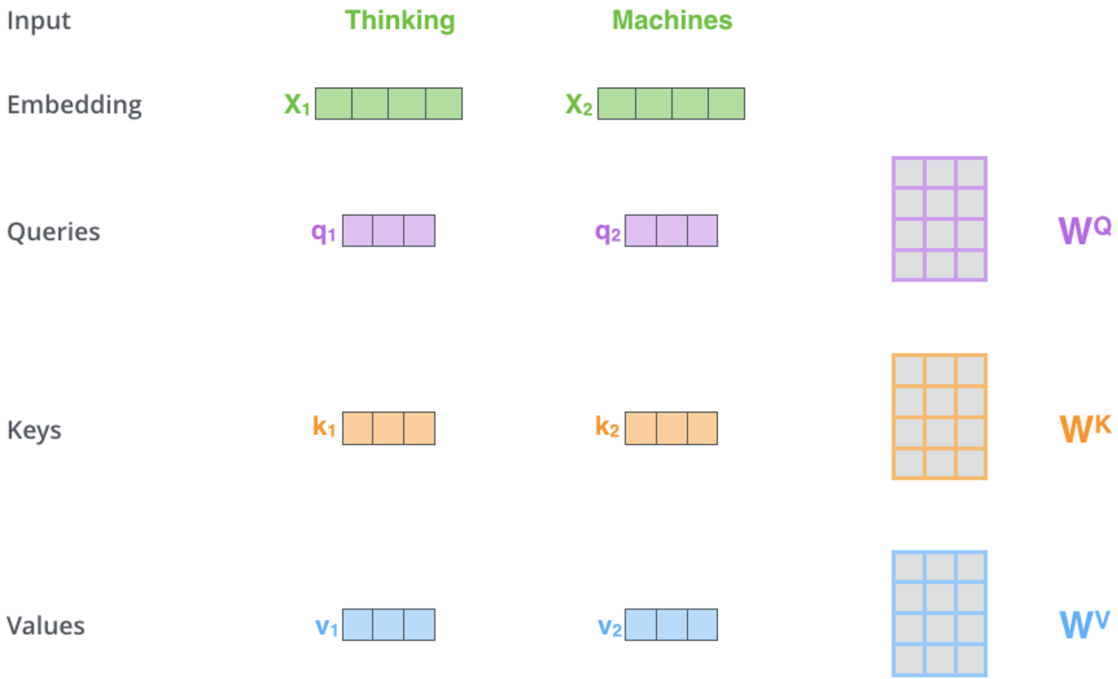
99% toxic comment

Attention

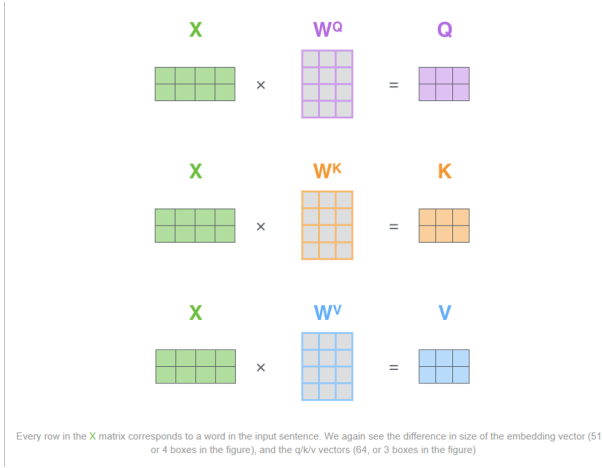


https://lena-voita.github.io/nlp_course/seq2seq_and_attention.html

Self-Attention



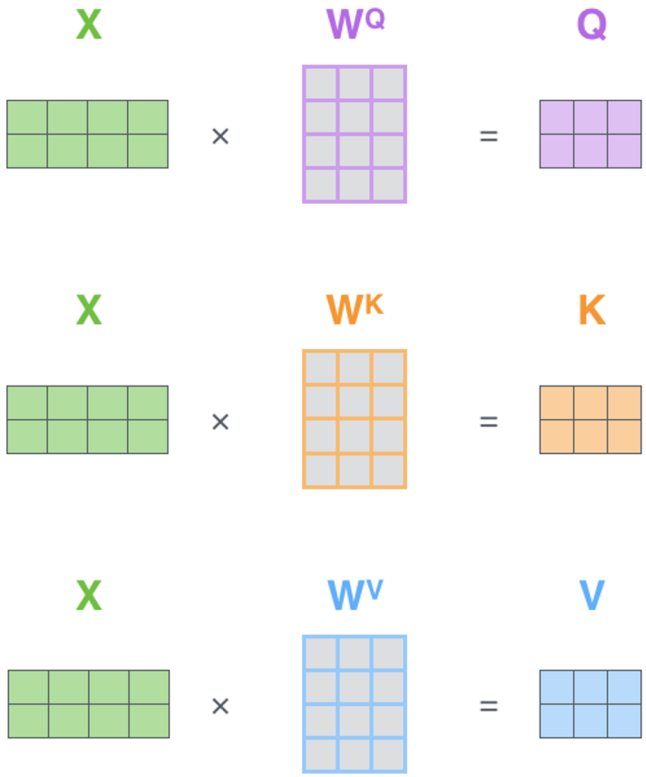
Multiplying x1 by the WQ weight matrix produces q1, the "query" vector associated with that word. We end up creating a "query", a "key", and a "value" projection of each word in the input sentence.



Every row in the X matrix corresponds to a word in the input sentence. We again see the difference in size of the embedding vector (51 or 4 boxes in the figure), and the q/k/v vectors (64, or 3 boxes in the figure)

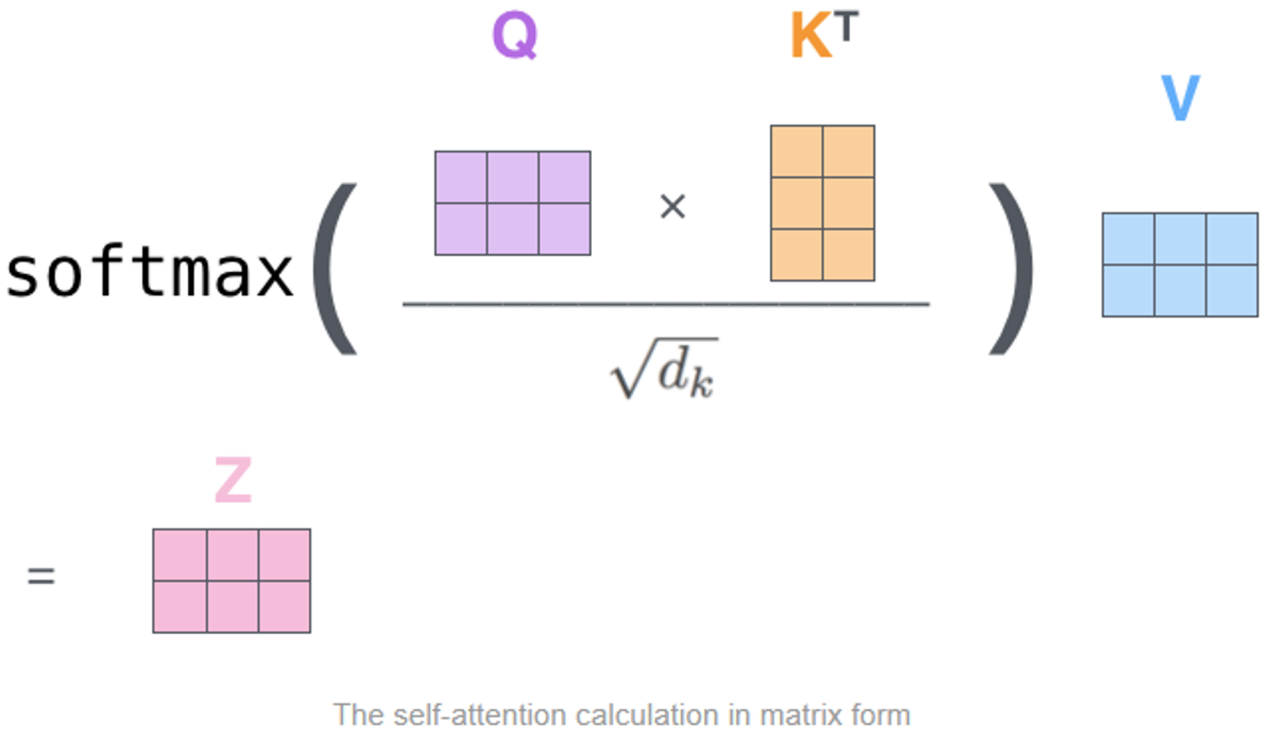
https://jalammar.github.io/illustrated-transformer/

Self-Attention



Every row in the X matrix corresponds to a word in the input sentence. We again see the difference in size of the embedding vector (512, or 4 boxes in the figure), and the q/k/v vectors (64, or 3 boxes in the figure)

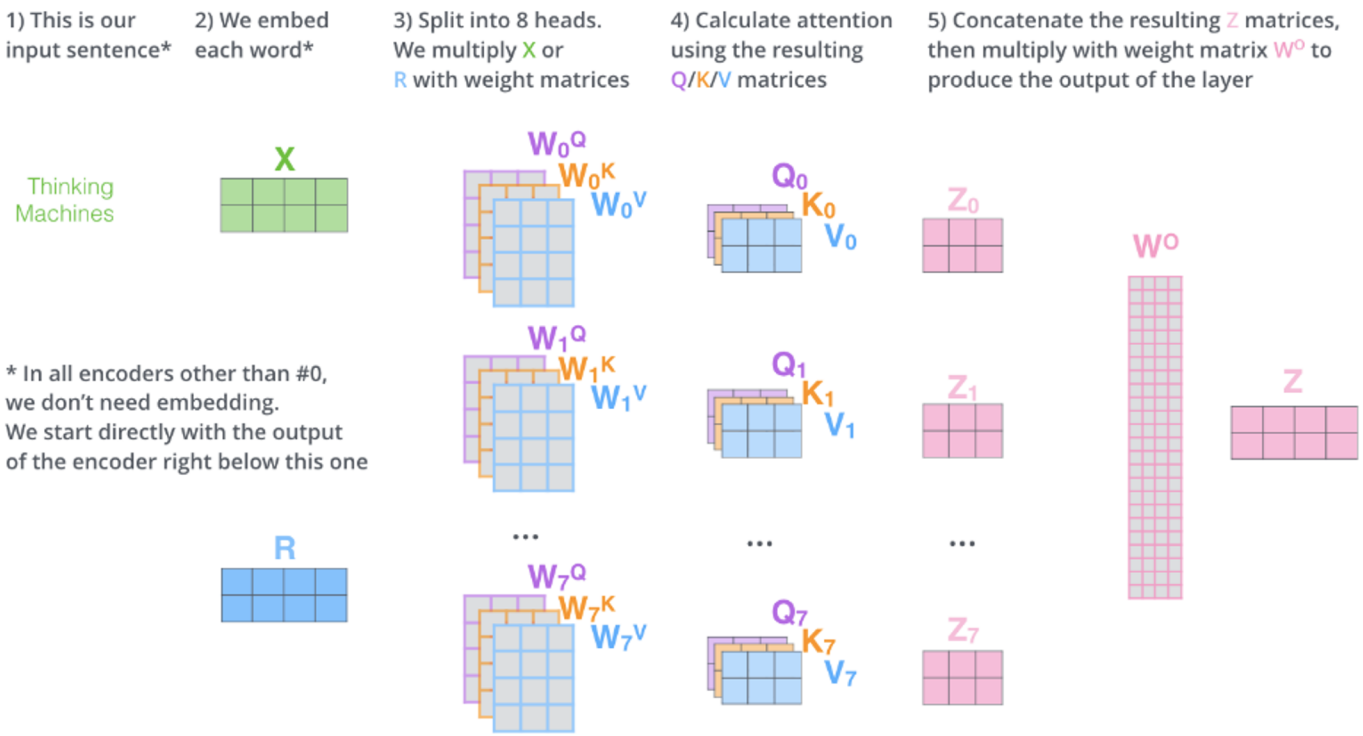
Self-Attention



The self-attention calculation in matrix form

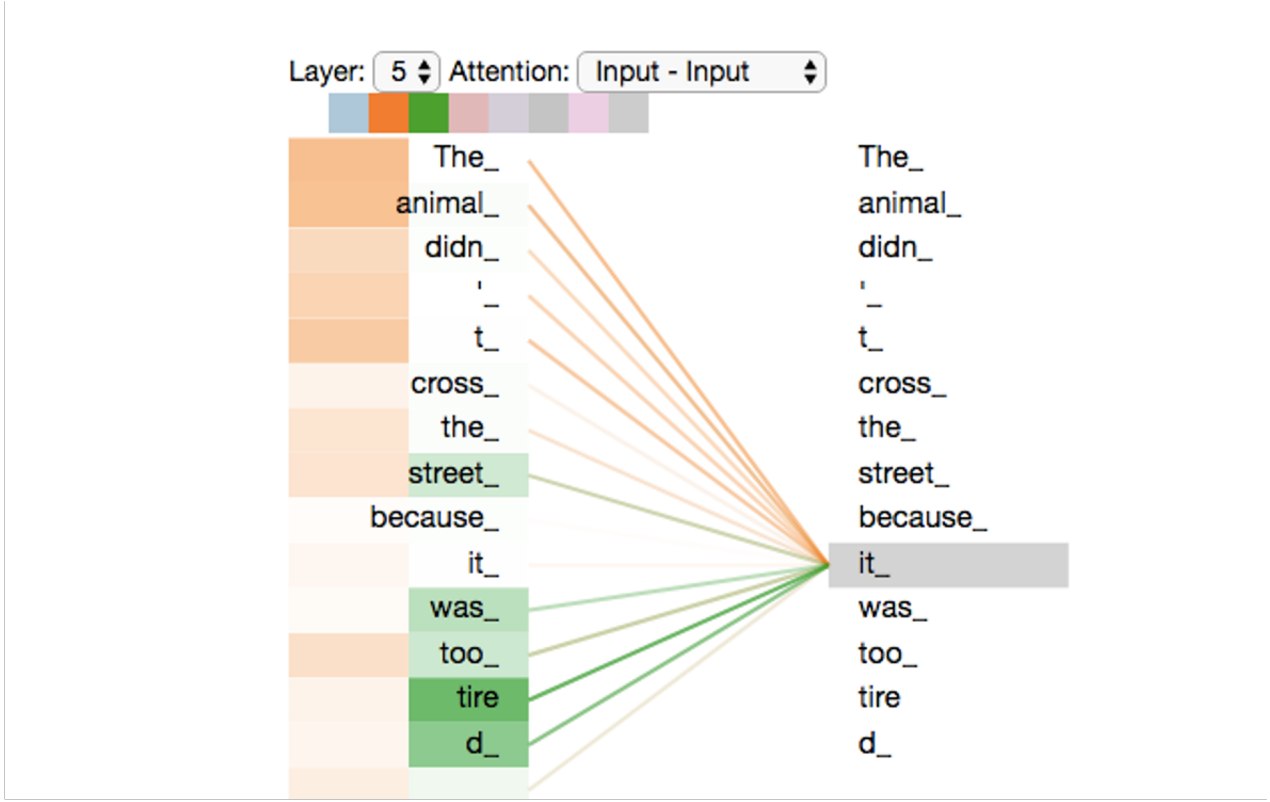
https://jalammar.github.io/illustrated-transformer/

Multi-Head-Attention



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

Self-Attention



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

Transformer

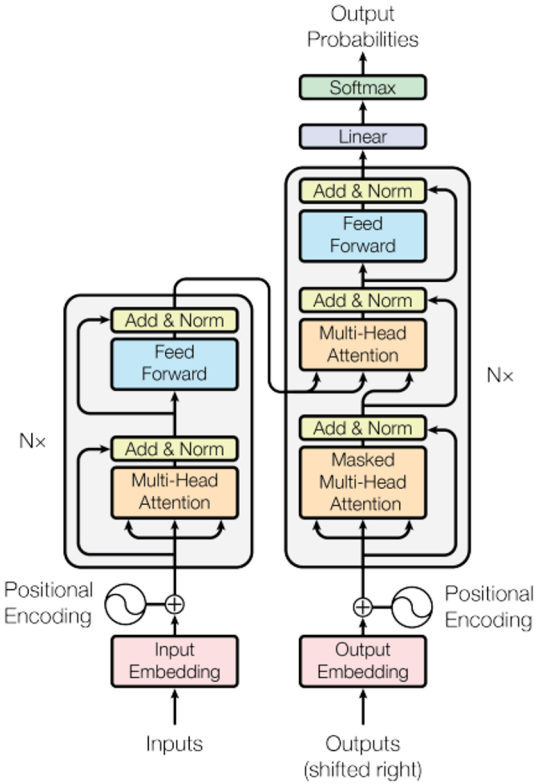
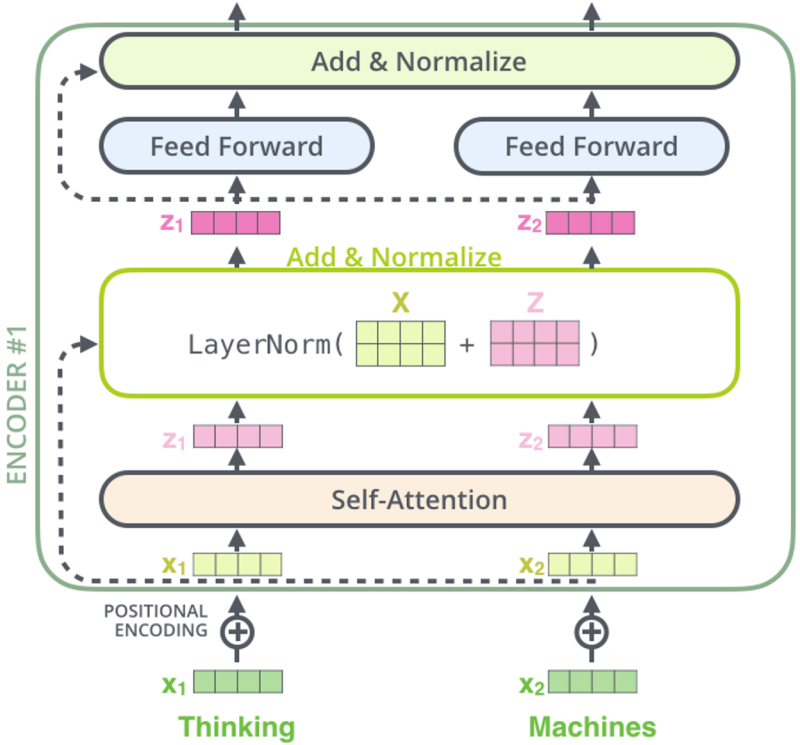
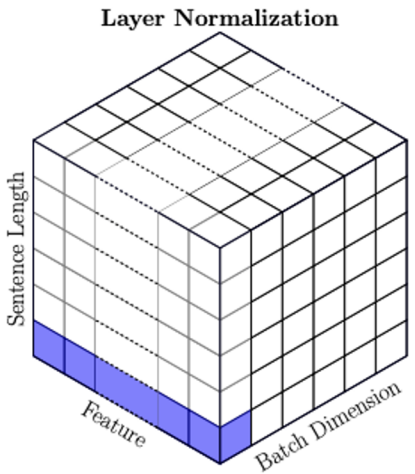


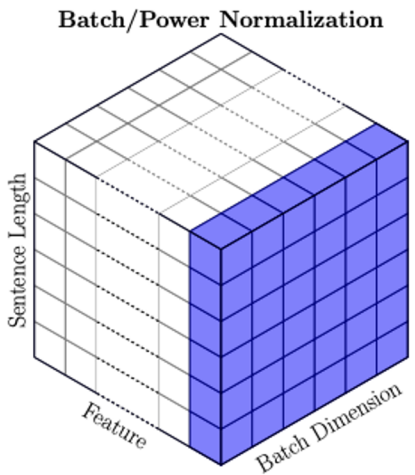
Figure 1: The Transformer - model architecture.

https://arxiv.org/pdf/1706.03762.pdf

Transformer (encoder block)

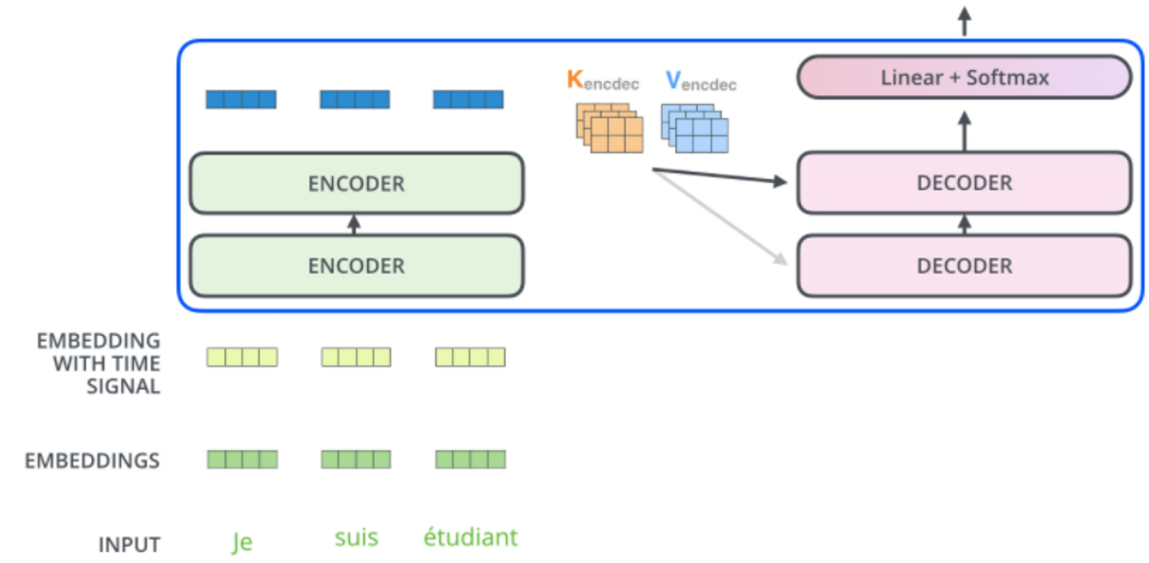






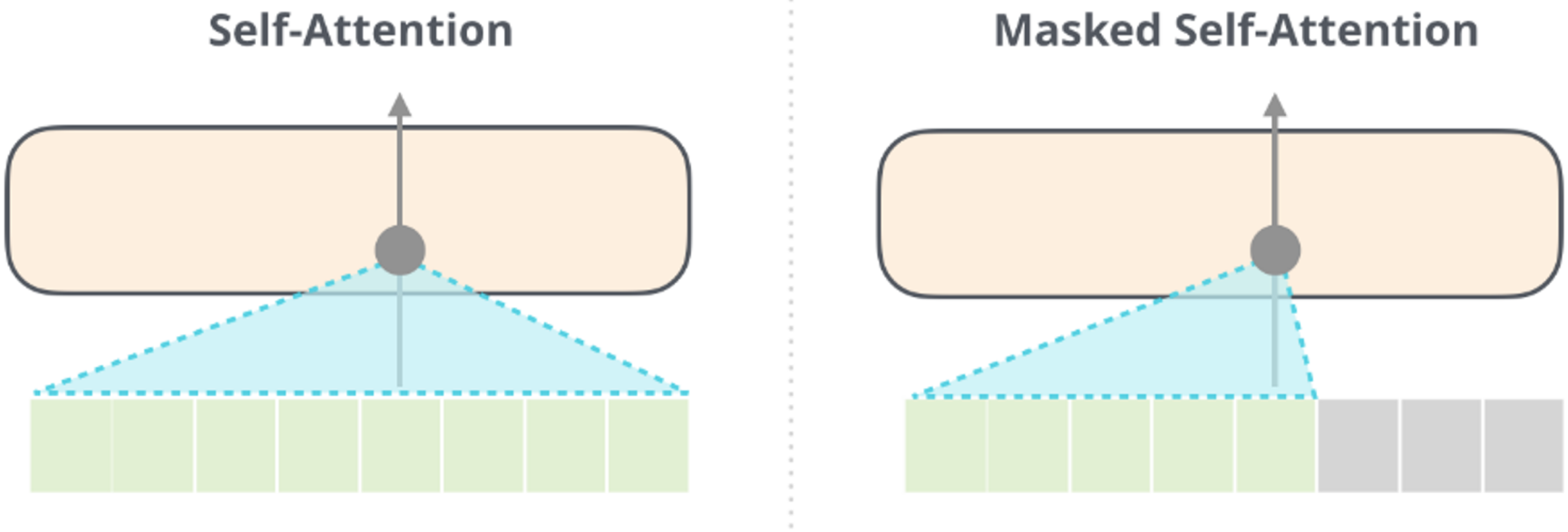
https://arxiv.org/pdf/1706.03762.pdf

Transformer (Decoder block)

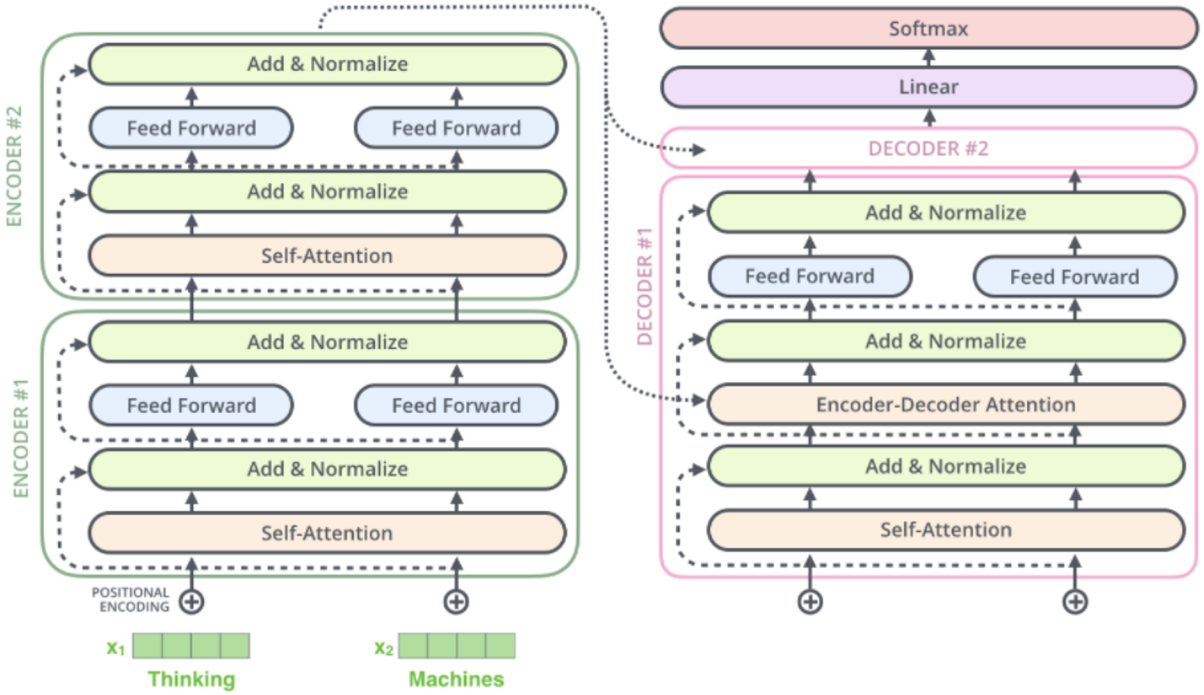


https://arxiv.org/pdf/1706.03762.pdf

Self-Attention

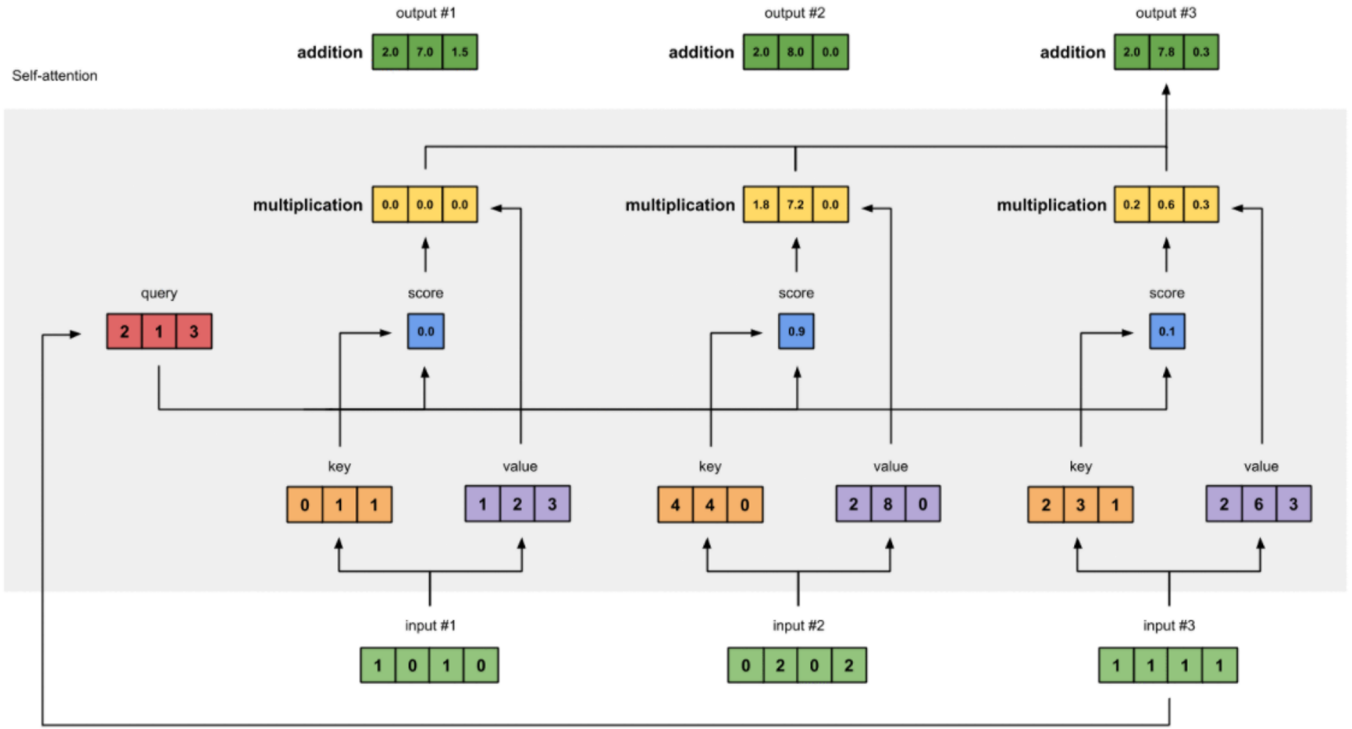


Transformer (Full)



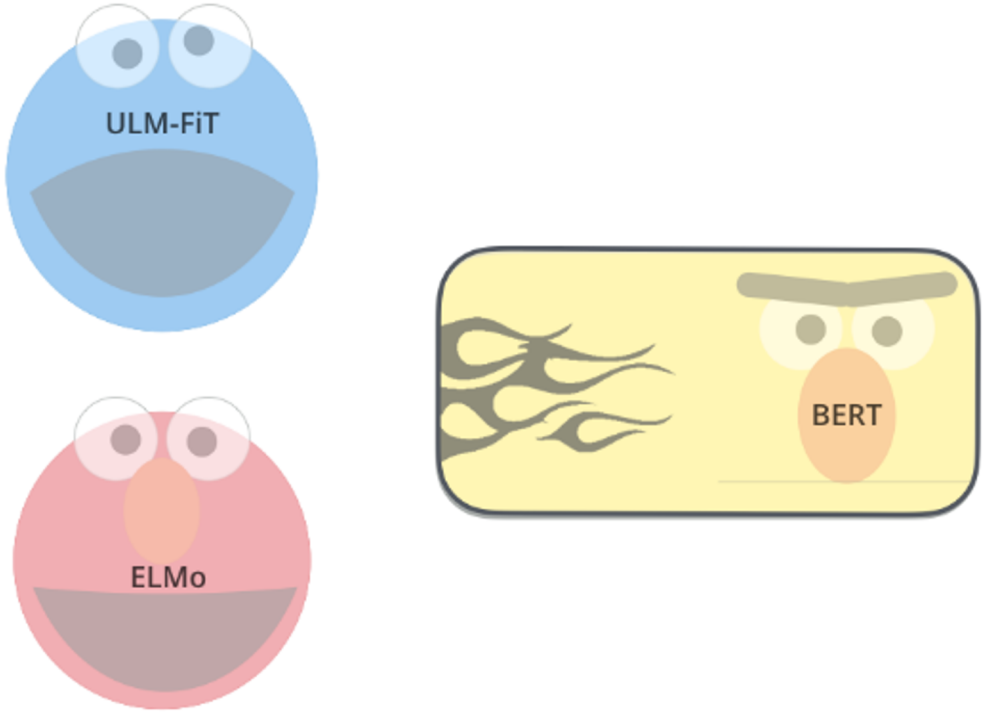
https://arxiv.org/pdf/1706.03762.pdf

Self-Attention



https://towardsdatascience.com/illustrated-self-attention-2d627e33b20a

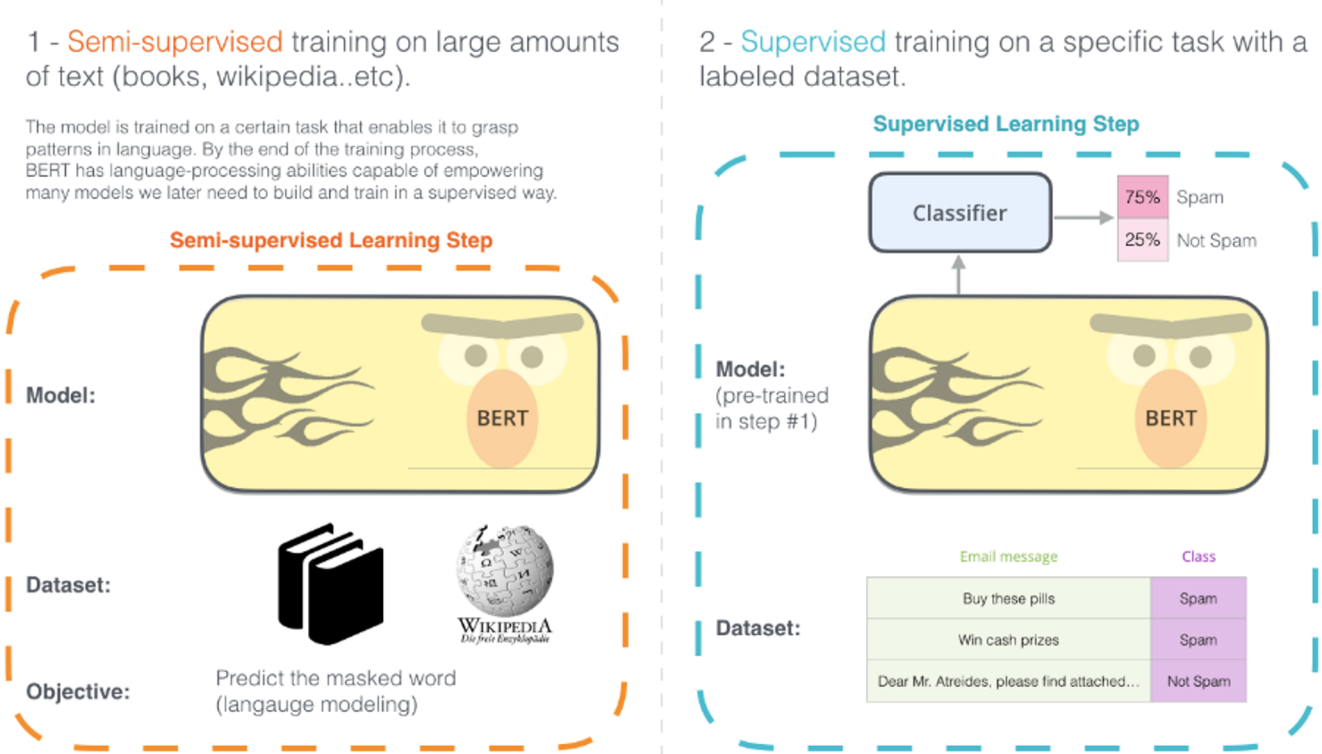
Better Embeddings



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

Bert

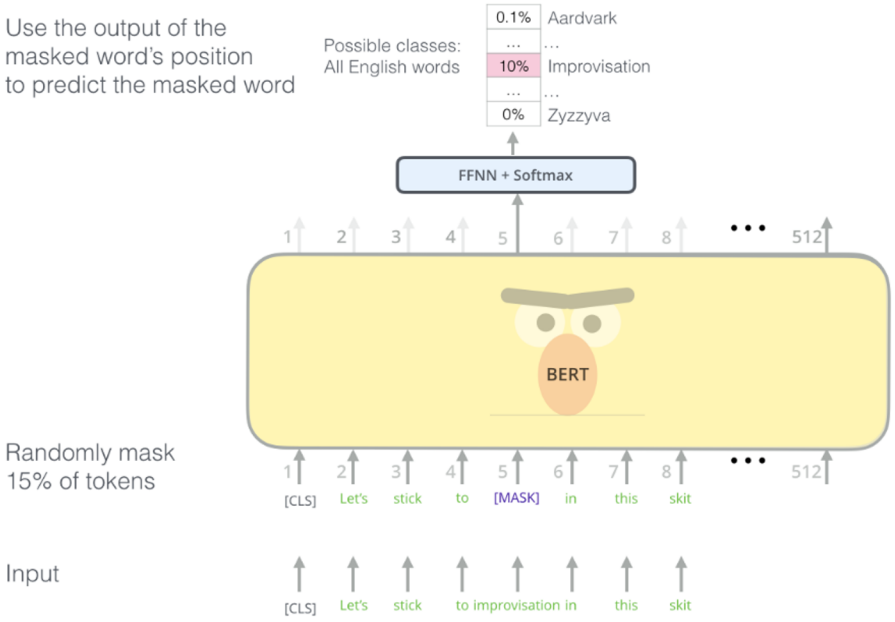
Bidirectional Encoder Representations from Transformers

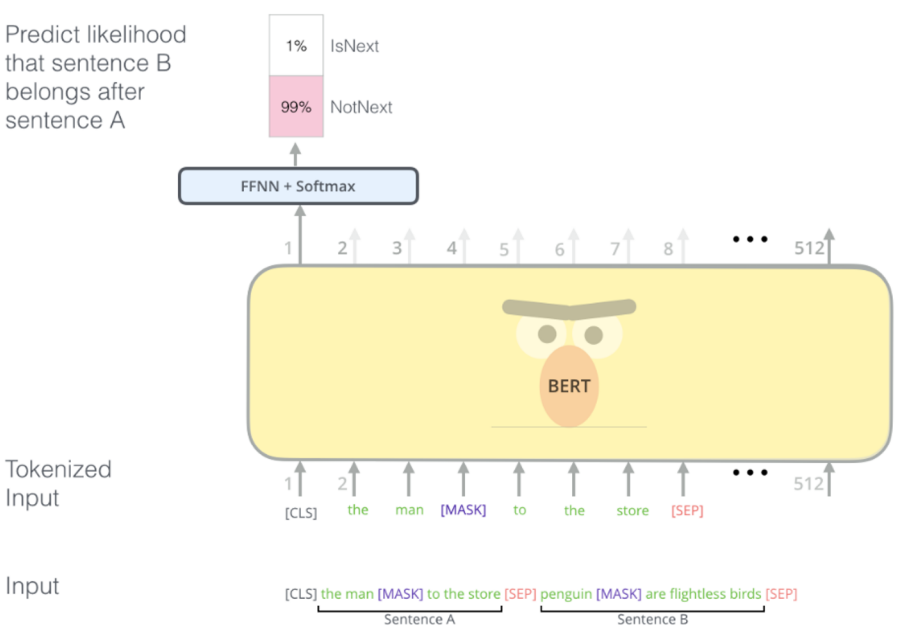


The two steps of how BERT is developed. You can download the model pre-trained in step 1 (trained on un-annotated data), and only worry about fine-tuning it for step 2. [Source for book icon].

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

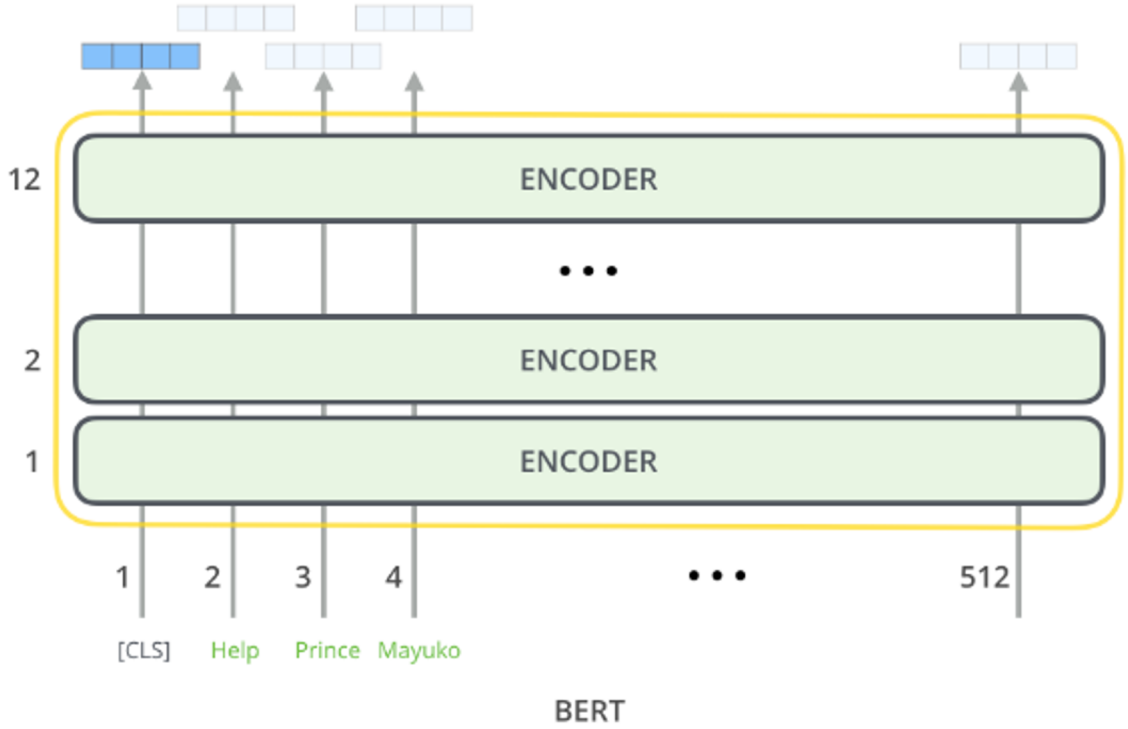
Bert





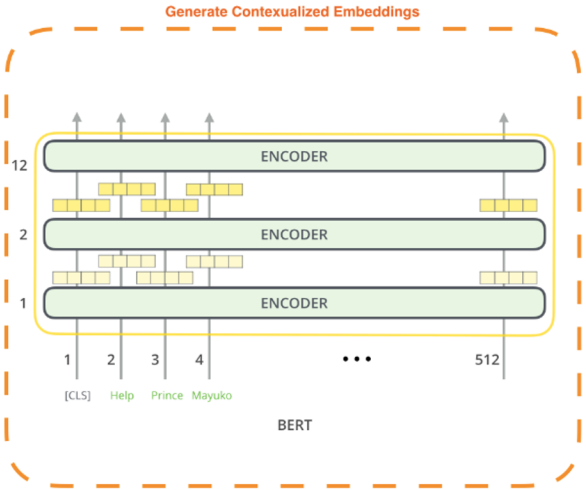
https://jalammar.github.io/illustrated-bert/

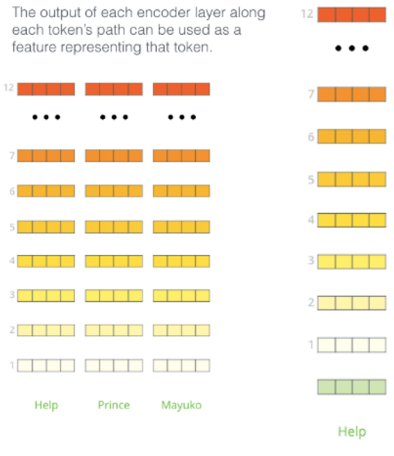
Bert

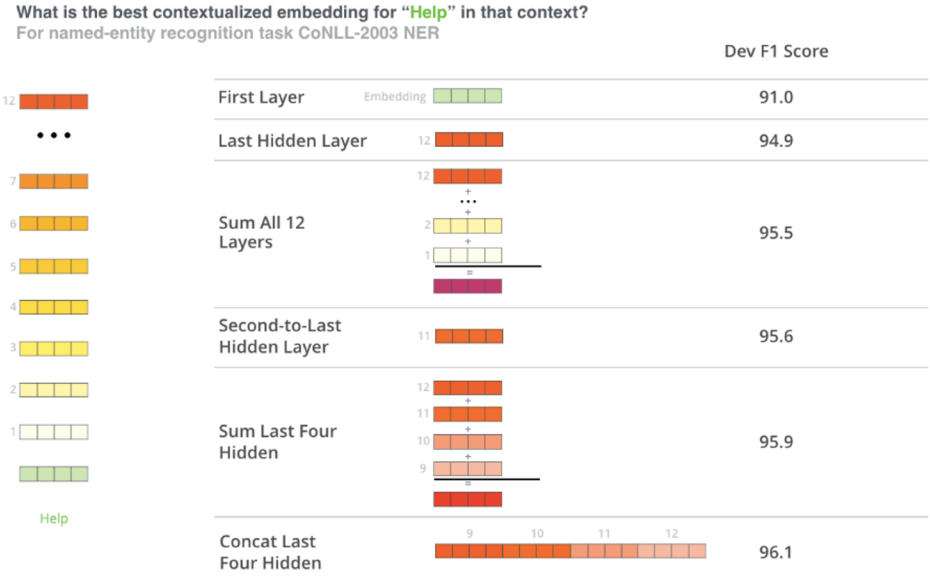


https://jalammar.github.io/illustrated-bert/

Bert

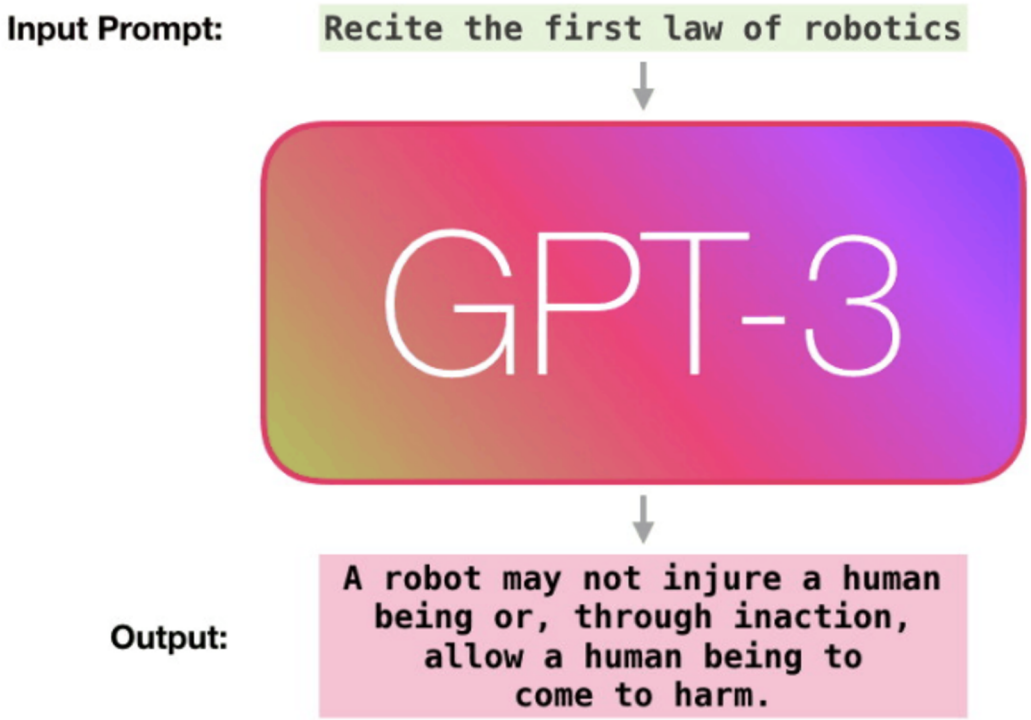






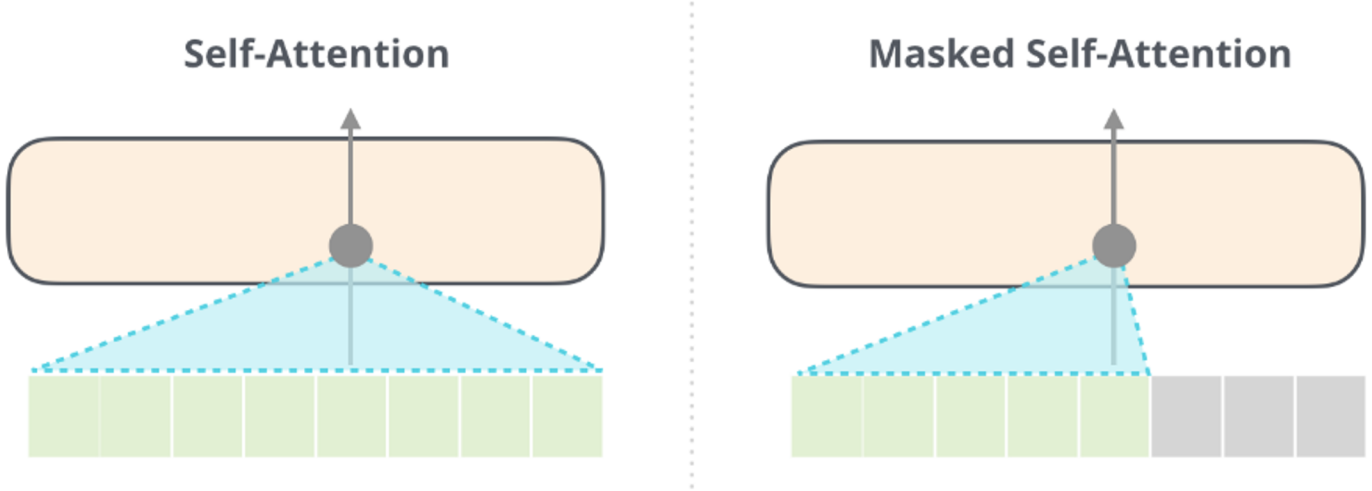
https://jalammar.github.io/illustrated-bert/

GPT-3



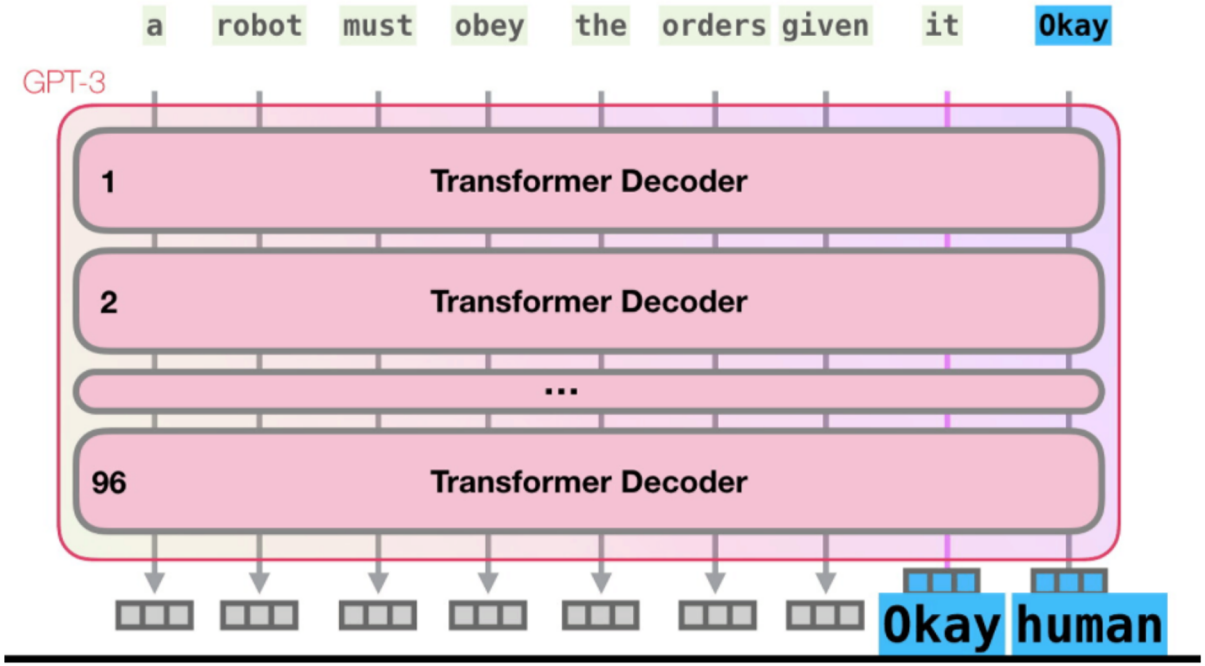
https://jalammar.github.io/how-gpt3-works-visualizations-animations/

GPT-3



https://jalammar.github.io/how-gpt3-works-visualizations-animations/

GPT-3



https://jalammar.github.io/how-gpt3-works-visualizations-animations/