How attention was all we ever needed

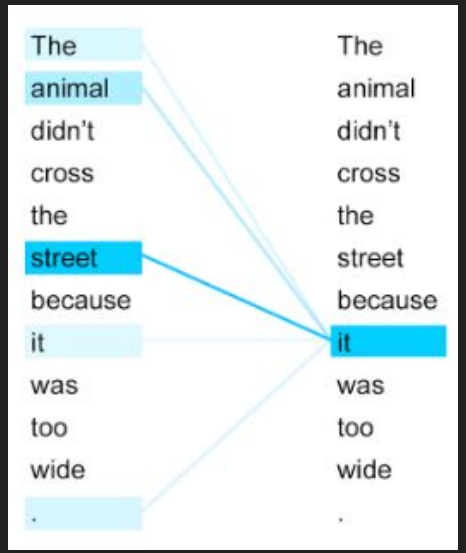
A report on classical self - attention models

Table of contents:

What Is Self-Attention?

- According to Google's paper Attention is All You Need, Self-attention, sometimes called intra-attention, is an attention mechanism relating different positions of a single sequence in order to compute a representation of the sequence.

Example of self attention



*Figures Taken from [2]

Self-Attention Contd.

- Self attention helps the model relate a word with other words of the sequence to gain an understanding of that word.
- It has been proven useful in a variety of tasks including text summarization, machine translation and much more.
- Self attention has been used in LSTMs [3], deep reinforced models [4], in extraction of sentence embeddings [5] and more.
- Transformers are the first transduction models which rely entirely on self - attention to compute representations of their input and output sequences.

Why self-attention instead of RNNs and previous SOTA*?

- Sequence-to-sequence (seq2seq) models convert a type of text sequence into another type. They find a huge range of applications such as Machine translation, text summarization, speech recognition, question answering, emotion detection, etc. They lie at the core of NLP.
- Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTMs) and Gated RNNs dominated NLP before transformers were introduced.
- The main issues were long sequence context losses and computation requirements.
- The sequential nature of previous seq2seq models prevented parallelization.

*State Of The Art

Transformers - An Example:

Transformers are transduction models which rely on self-attention to effectively compute representations. They take in a sequence of tokens and compute attention w.r.t to all other tokens to better understand the context of text. Let's take an example as shown

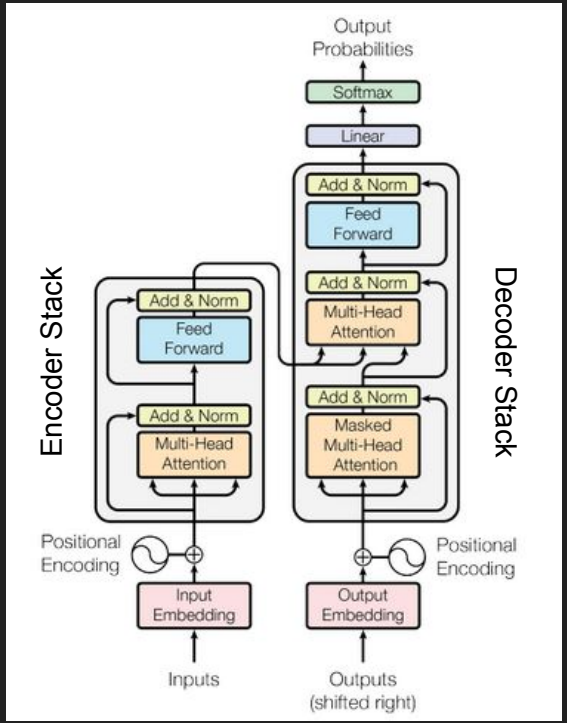
Example sequence: I like Muffins because of their texture and variety of flavours.

A transformer model would take Muffins and compute self attention w.r.t all other words and create embeddings. Similar context words would get nearby embeddings and dissimilar ones would get far away embeddings. This contextual information is used by the transformer to perform the required task for example, translating this sentence to another language.

Transformer Architecture:

The architecture mainly consists of the following :

- Input Embedding
- Output Embedding
- The Encoder stack
- The decoder stack
- Linear and softmax activations



Transformer architecture

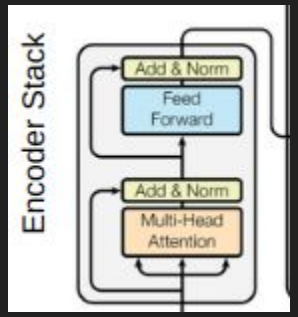
*Figure Taken from [1]

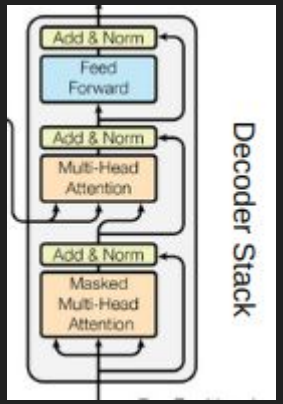
Positional Encoding :

- Positional encodings are used to inject information about relative or absolute position of tokens in the sequence.
- They are added to input embeddings at the bottom of encoder and decoder stacks. They have the same dimensions as the embeddings.
- The following functions are used :
  - P E(pos,2i) = sin(pos/100002i/dmodel)
  - P E(pos,2i+1) = cos(pos/100002i/dmodel)
  - Pos is the position and i is the dimension.

Encoder and decoder stacks:

- Both encoder and decoders use feed forward networks with multi head attention.
- Decoder uses an additional masked multi-head attention.
- Addition and normalization is done after every layer.





*Figures Taken from [1]

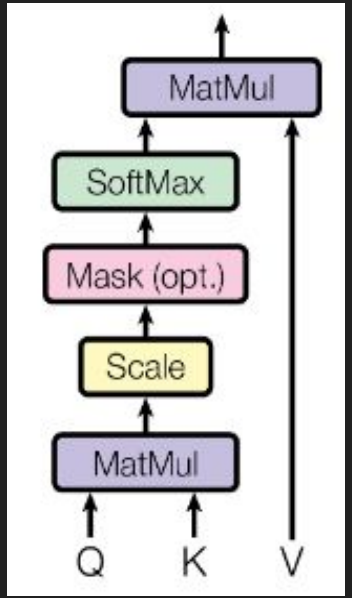
Attention Mechanisms - Scaled Dot-Product:

Inputs to attention mechanism:

- Query Vector
- Key Vector
- Value Vector

Outputs of attention mechanism:

- Scaled dot-product gives weighted values as output.



Scaled Dot-Product Attention

*Figure Taken from [1]

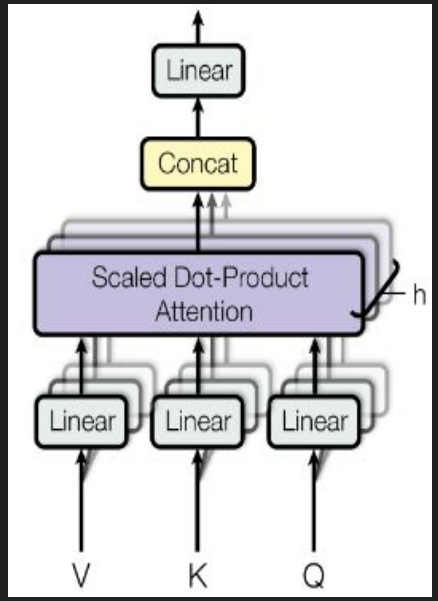
Attention Mechanisms - Multi-Head:

Inputs to attention mechanism:

- Query Vector
- Key Vector
- Value Vector

Outputs of attention mechanism:

- Multi-Head attention concatenates outputs of scaled dot-product to provide weighted sum of values as overall output of attention mechanism.
- The Query, Key and Value Vectors are basically used to calculate a score which is then used to determine the priority of words to be focussed on while processing a word.



Multi-Head Attention

*Figure Taken from [1]

Multi Head Attention contd.

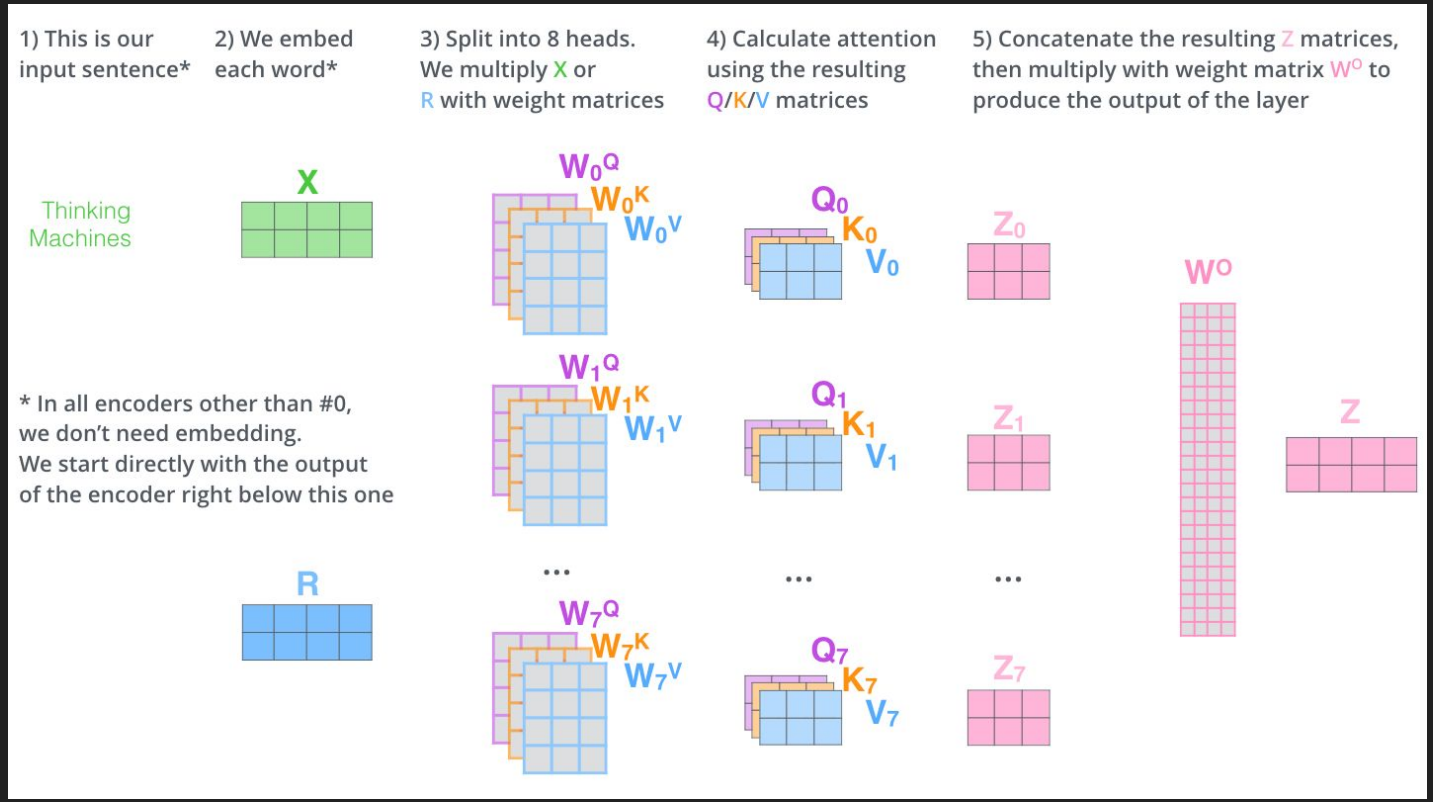
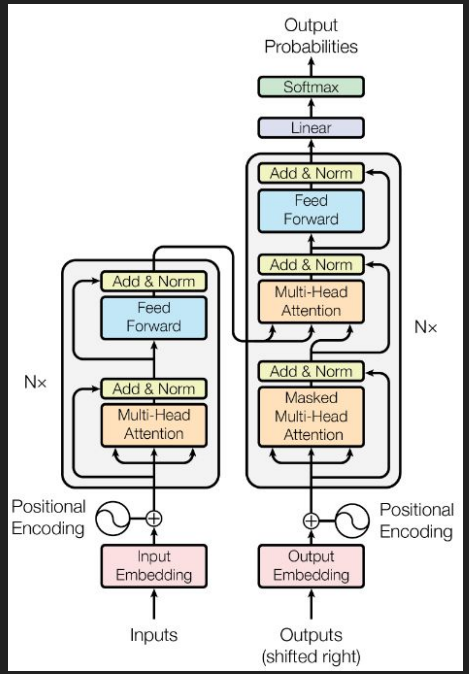
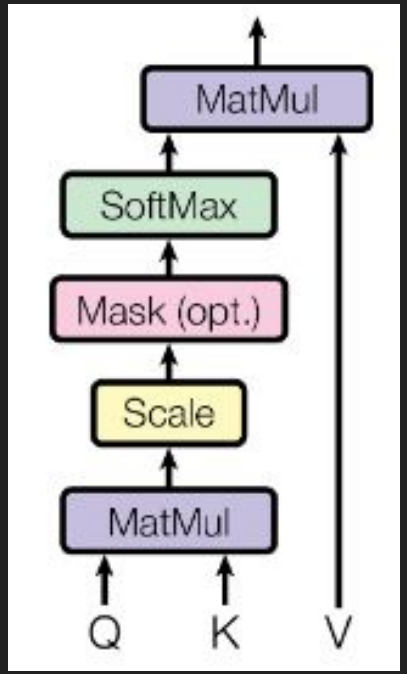


Image taken from http://jalammar.github.io/illustrated-transformer/

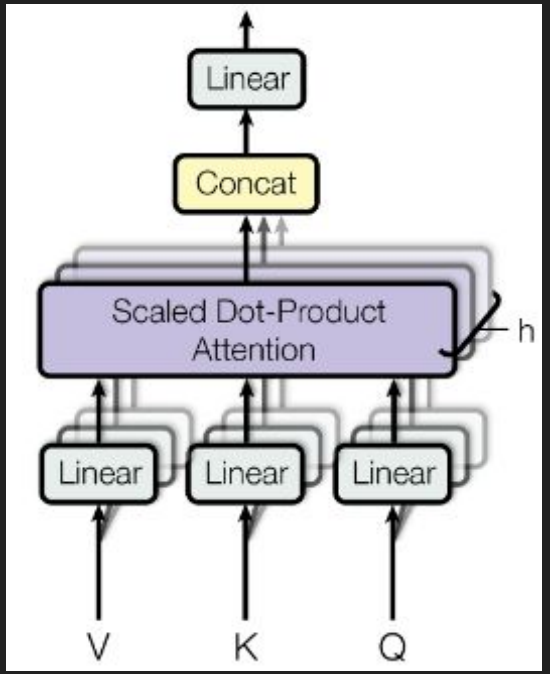
The transformer architecture and attention mechanisms:



Transformer architecture



Scaled Dot-Product Attention



Multi-Head Attention

*Figures Taken from [1]

Conclusion:

- Transformers have taken over the NLP domain and many variations have become current state-of-art models in many tasks.
- Currently, transformers such as Generative Pretrained Transformers (GPT) models have made big waves in the industry for providing human like interaction capabilities.
- The biggest drawbacks transformers faced was that they only work with fixed length sequences and context fragmentation. These were resolved with the introduction of Transformer XL [6].
- Transformers use a huge amount of computation during training. Quantum Hybrid transformers have been proposed to solve some intractable classical problems and also provide computational benefits by replacing some layers where it is possible to utilize quantum computing to speed up and enhance transformer models' training and inference.

References:

[1] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[2] Joshi, Prateek. “How Do Transformers Work in NLP? A Guide to the Latest State-of-the-Art Models.” Analytics Vidhya, 19 June 2019, How do Transformers Work in NLP? A Guide to the Latest State-of-the-Art Models.

[3] Junyoung Chung, Çaglar Gülçehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR, abs/1412.3555, 2014.

[4] Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model. In Empirical Methods in Natural Language Processing, 2016.

[5] Sainbayar Sukhbaatar, arthur szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, Advances in Neural Information Processing Systems 28, pages 2440–2448. Curran Associates, Inc., 2015.

[6] Dai, Zihang, et al. "Transformer-xl: Attentive language models beyond a fixed-length context." arXiv preprint arXiv:1901.02860 (2019).

THANK YOU!