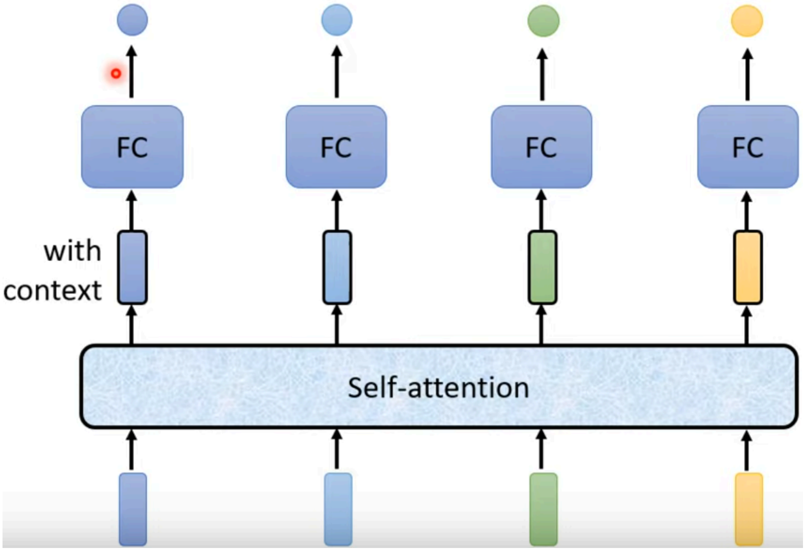
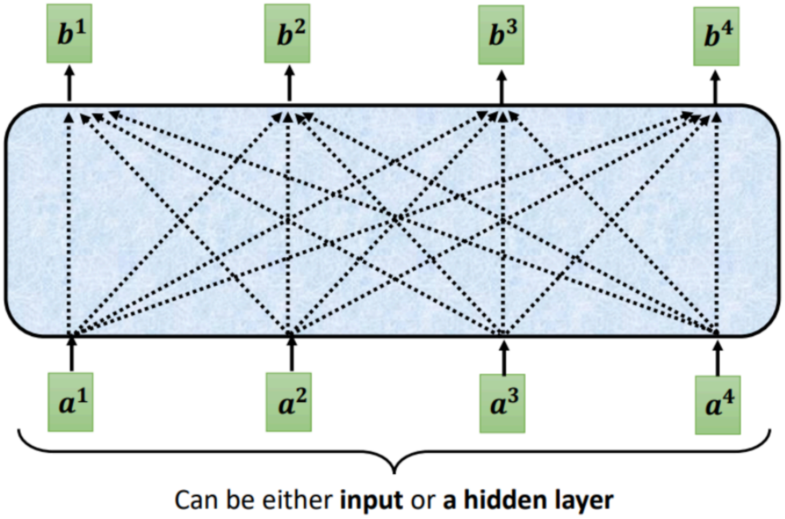
Self-Attention

vector sequence as input (e.g. text, voice, graph)
Self-Attention can process the original vectors to make meaningful within the context:



Self-Attention structure

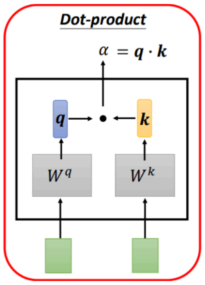
Self-Attention considers the whole sequence to determine how to construct the new vector



Focus on one vector e.g. a_2, how to get the corresponding b_2

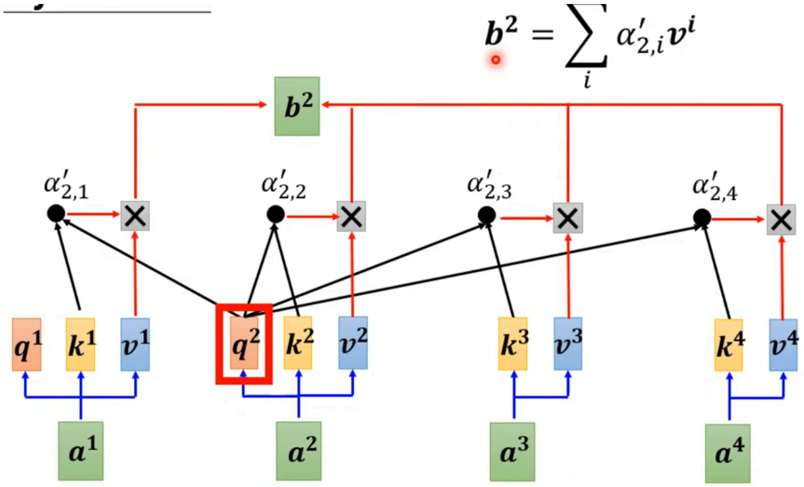
1. Find the relevant vectors in the sequence
   use **query** q of vector a_2 and **key** k_i of vector a_i to compute the **attention score** $\alpha_{2,i}$ which measures the relevance between vector a_i and vector a_2

    - Method 1: Dot product (used in Transformer)

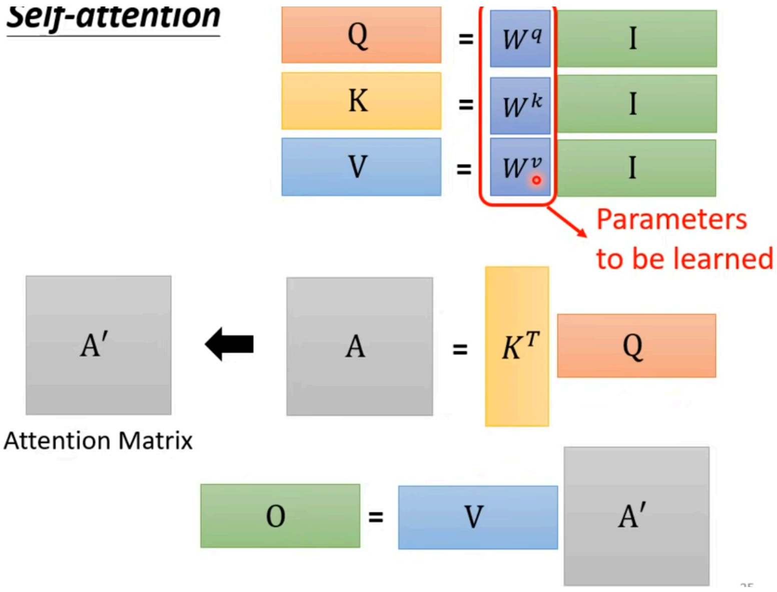


    - Method 2: Additive

2. Translate the attention scores into weights $\alpha'_{2,i}$
   use Softmax
3. Compute the wieghted sum of v_i, then we get b_2



The parallel process of computing b_i can be represented by matrix multiplications



- I: inputs, all a_i
- O: ouputs, all b_i

Multi-head Self-attention

A single Attention Matrix may not be to capture the various aspects of relationships present within a sequence.
To address this, we use multiple independent Self-Attention and introduce another matrix W_o to measure the weight of every relevance factor.

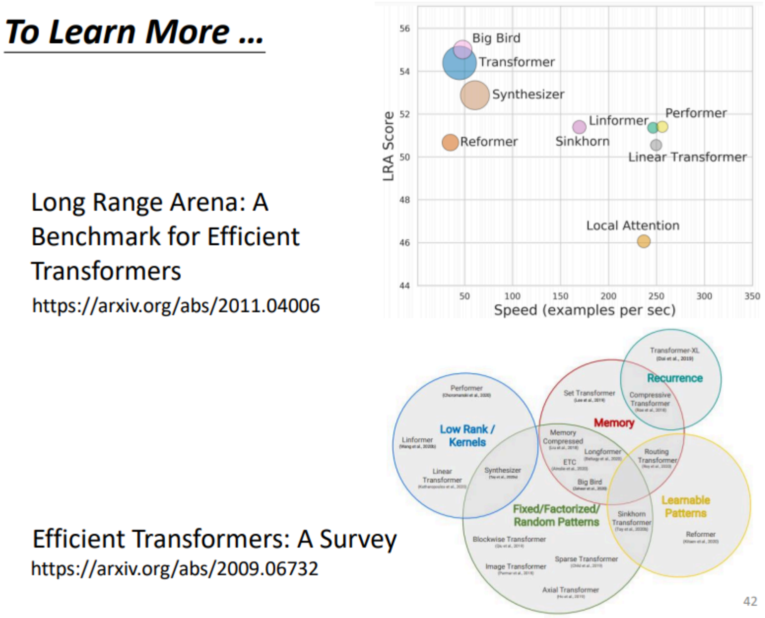
Positional Encoding

To add positional information into Self-Attention, enabling the model to understand the sequential natures (e.g. part of speech)
...

Advanced Self-Attentions

Self-Attention dominates computation in models like Transformer because

- complexity of the attention mechanism, $O(n^2)$ for a sequence of length n
- demands for Multi-head
- large sequence lengths
- ...
  To make it more efficiency, here are some varients of Self-Attention, usually called "xx-former"



To Learn More ...

Long Range Arena: A Benchmark for Efficient Transformers
https://arxiv.org/abs/2011.04006

Efficient Transformers: A Survey
https://arxiv.org/abs/2009.06732

42

- Local Attention / Truncated Attention
  Only pay attention to the closest tokens in the sequence, similiear with CNN
- Stride Attention
- Global Attention
- Longformer: Local + Stride + Global
- Big Bird: Longerformer + Random Attention

- ...

Self-Attention applications

1. NLP
2. Speech
   - Truncated Self-attention. Your understanding on the data determines the scope of **context**, i.e. the range of keys k
3. Image
   - In CNN, Image --> a long vector
   - In Self-attention, Image --> a set of vector, also reasonable!
   - Self-Attention GAN, DEtection Transformer (DETR)...
4. Graph
   - becomes one type of Graph Neural Network (GNN)