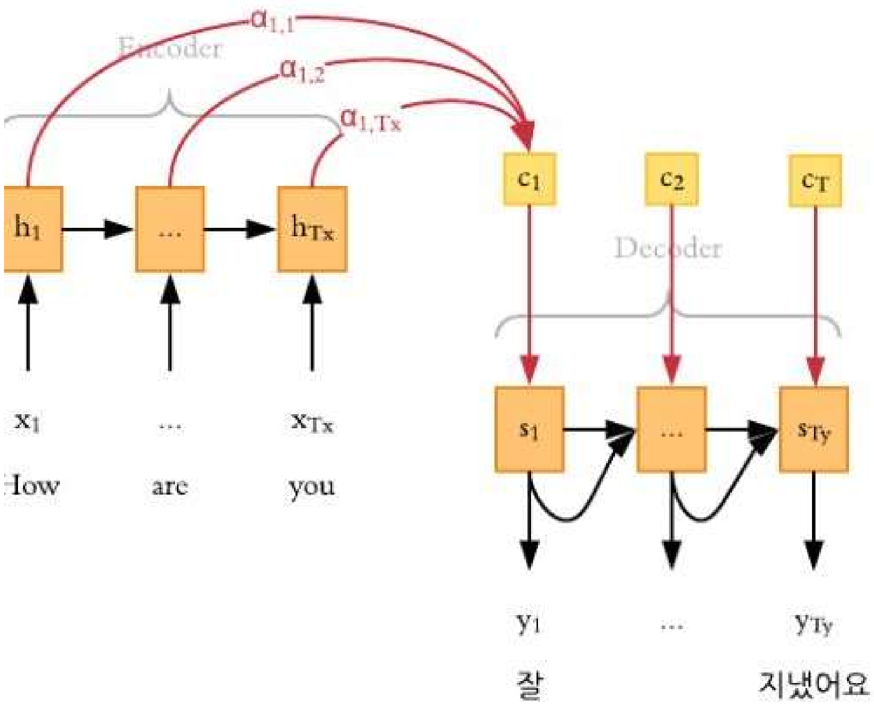
Transformers and Attention Mechanisms

Harsha (Harsh) Gandikota



Attention Mechanisms in Seq2Seq Models

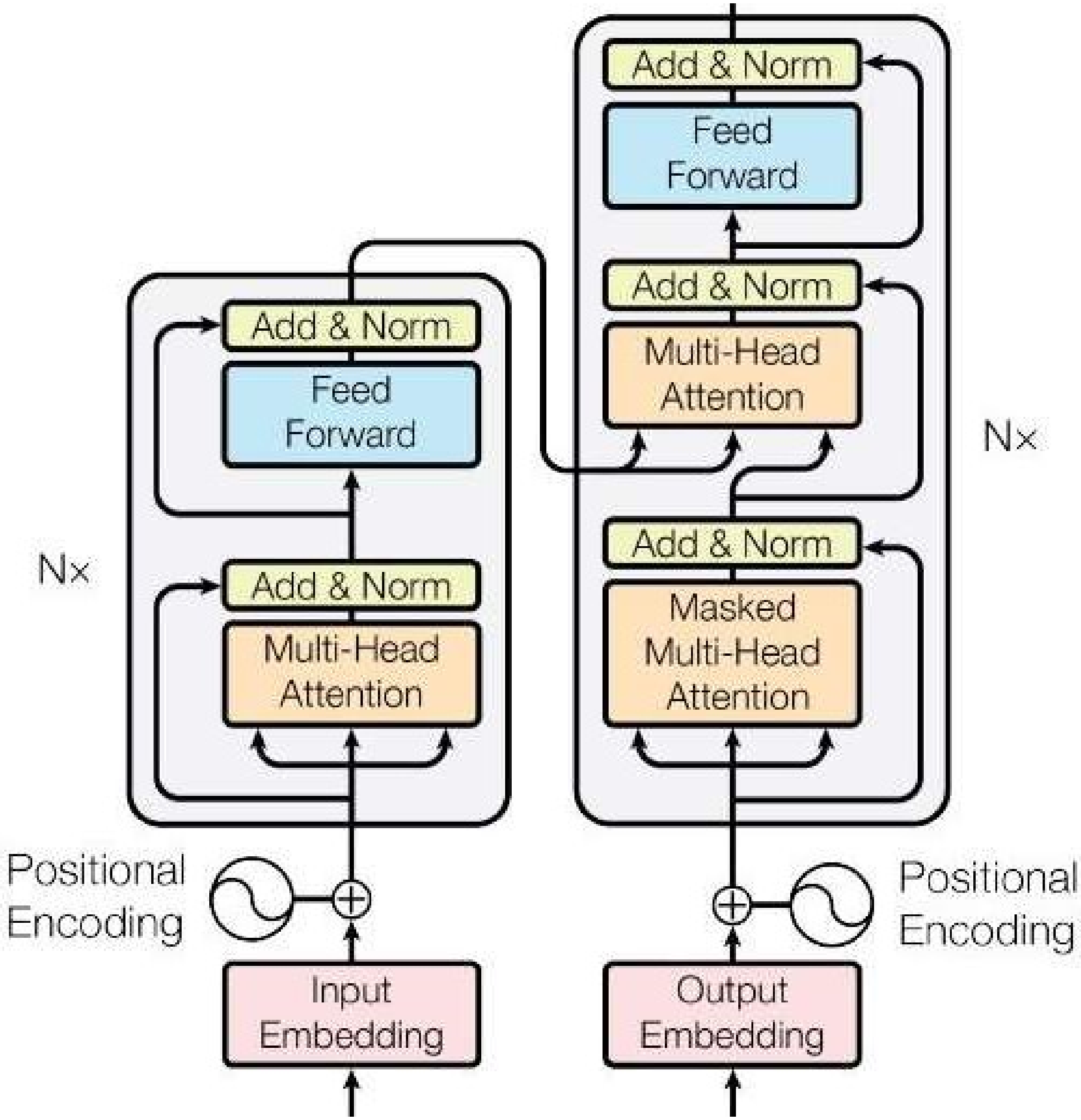
A way for the decoders of Seq2Seq Models to consider only the relevant parts of the input when generating an output.

Instead of using a single vector to represent the entire input, attention forms multiple ‘context’ vectors to help in decoding.

- Seq2Seq Models have traditionally followed an encoder-decoder architecture coupled with RNNs, LSTMs, GRUs etc.
- Calculations are still done sequentially and non-concurrently.
- This setup still executes sequentially, and is not completely parallelizable, which becomes critical when training long sequences.

The Transformer

- The Transformer is an Encoder-Decoder architecture that ditches recurrence and relies entirely on attention mechanisms.
- This feature allows the transformer to achieve significantly more parallelization and achieve far better results.
- The Transformer can calculate the number of operations required to relate signals from input and output in a constant number of operations.



Transformer Architecture

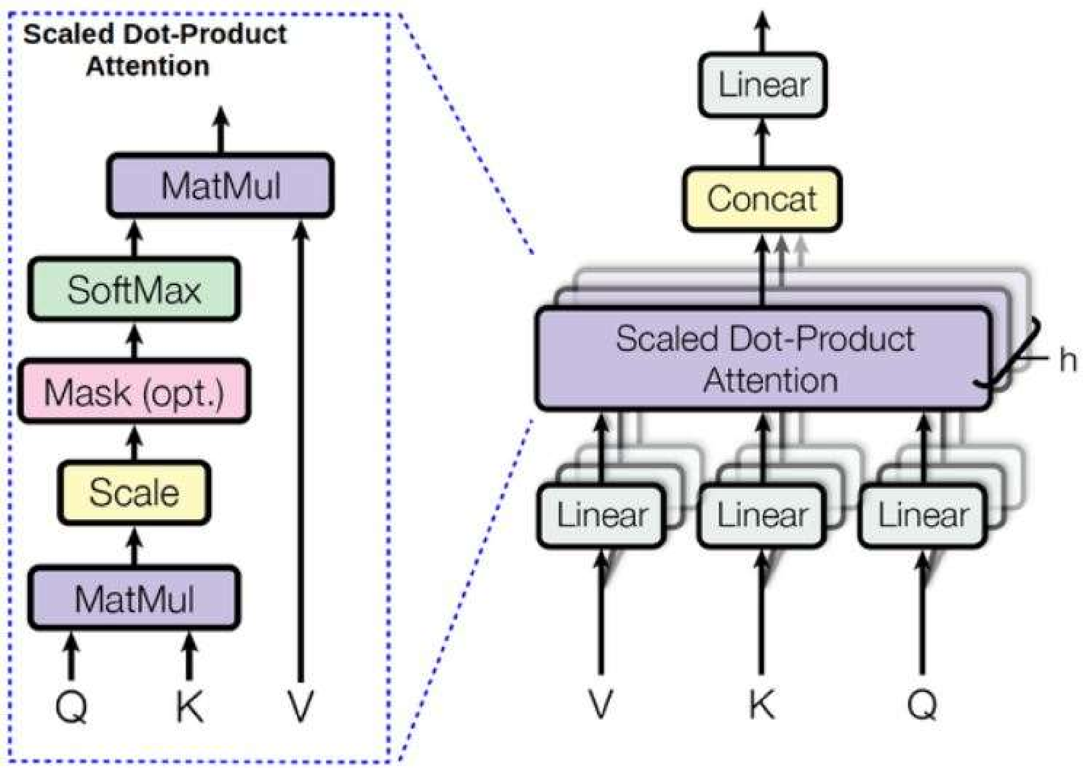
$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Scaled Dot Attention and Multi- Head Attention

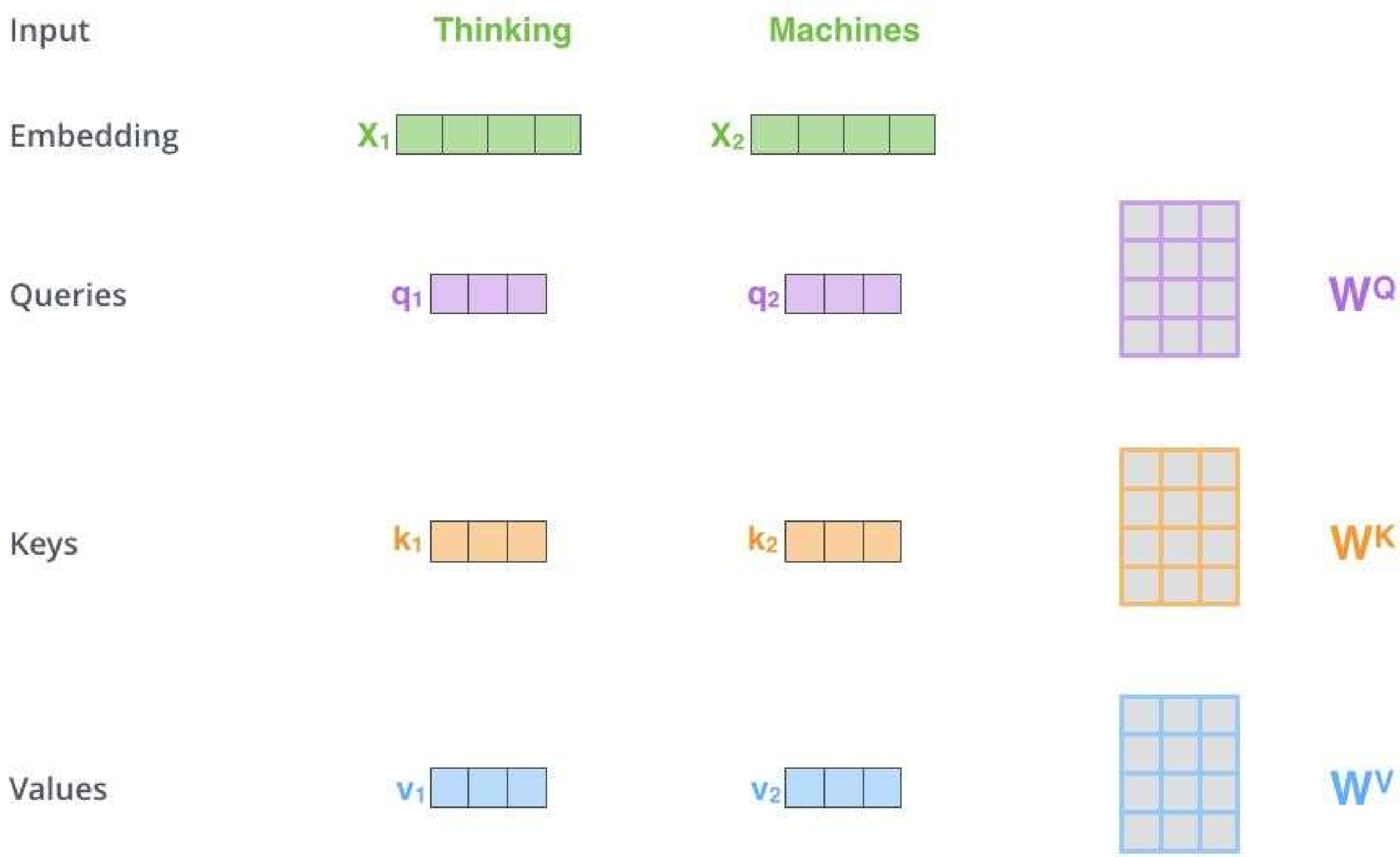
Q -> Query Vector

K -> Key Vector

V -> Value Vector

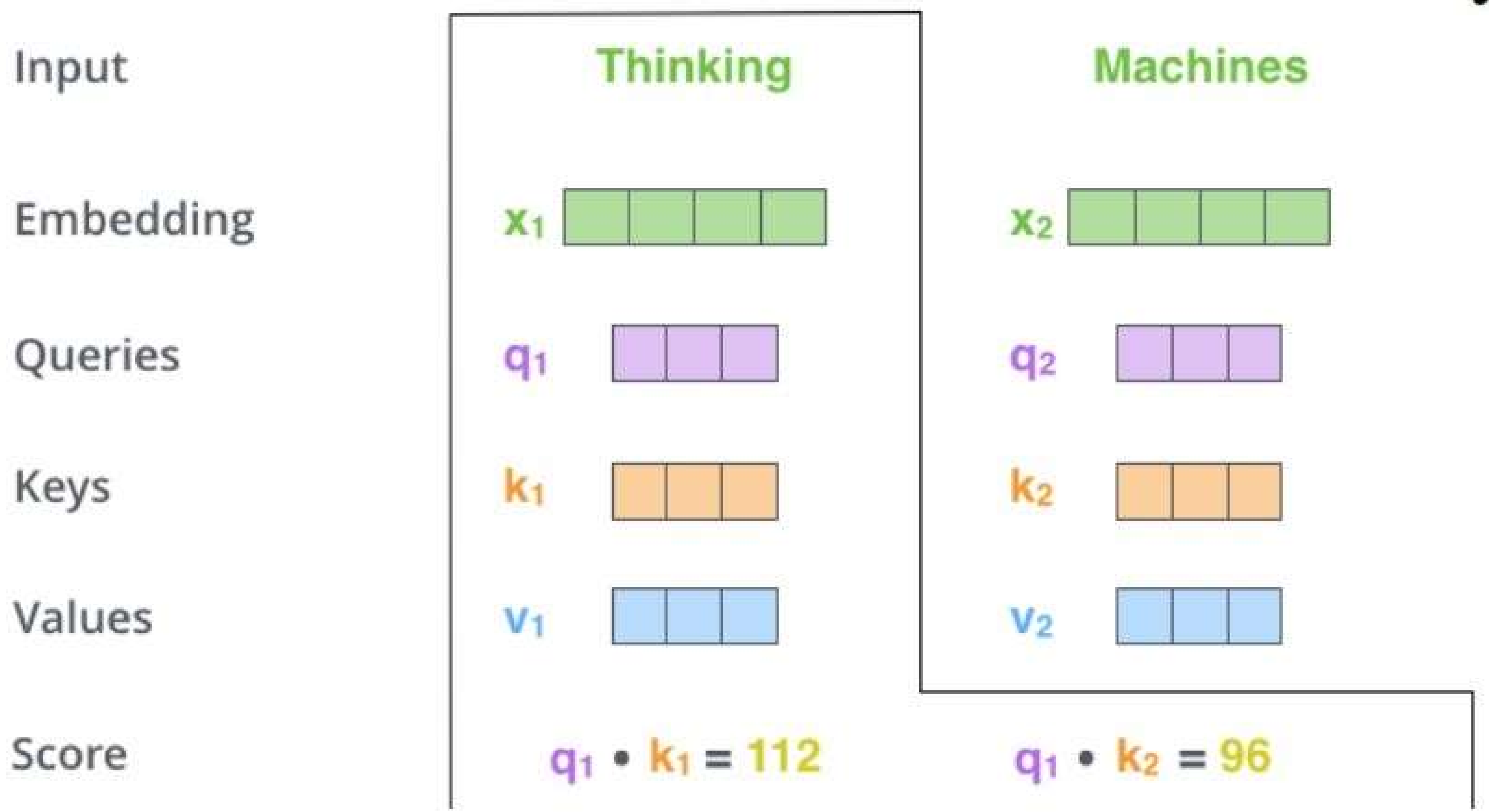


Self Attention in Detail : Setting up our Vectors

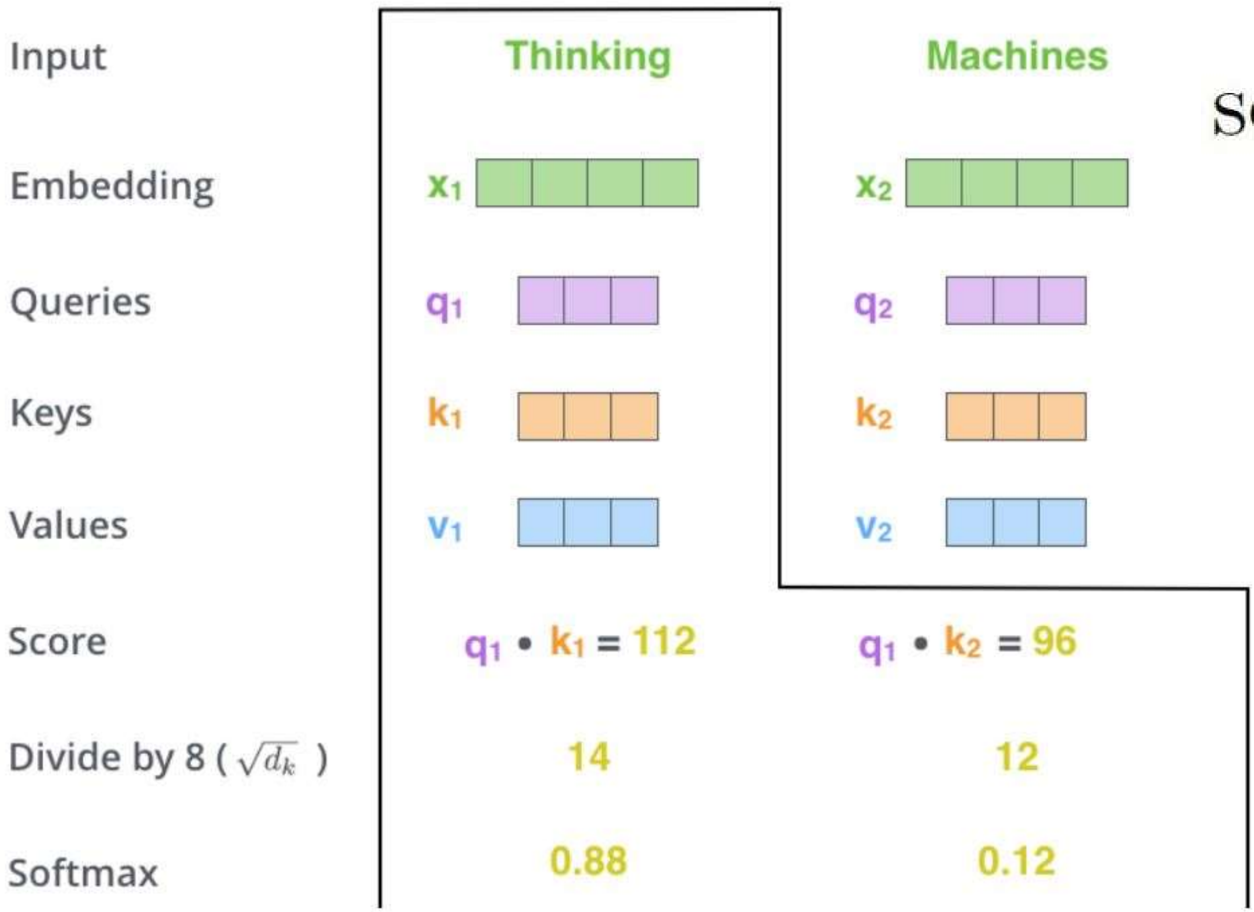


Self Attention in Detail: Query * Key

$$QK^T$$

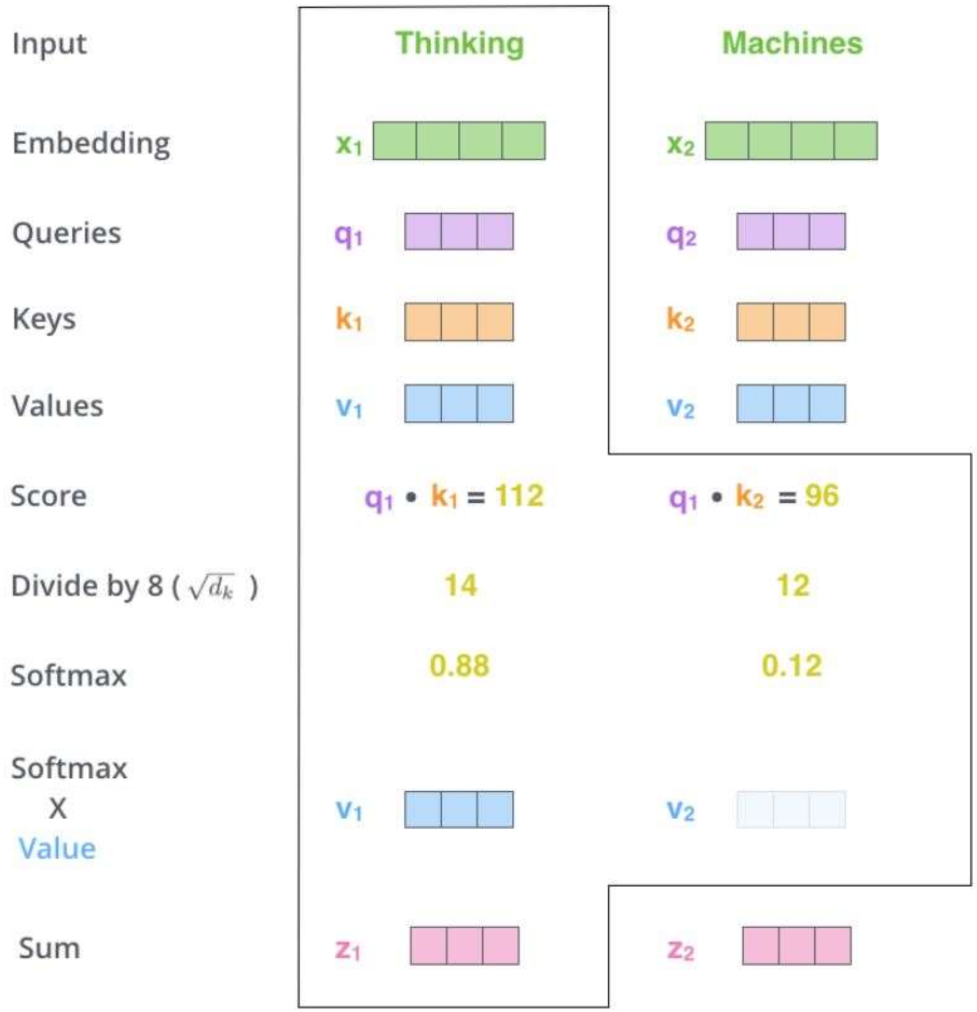


Self Attention in Detail: Apply Softmax



$$\mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)$$

Self Attention in Detail : Multiply by Value



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

Transformers BLEU Score

Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [15]	23.75			
Deep-Att + PosUnk [32]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [31]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [8]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [26]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [32]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [31]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [8]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	$\mathbf{3.3 \cdot 10^{18}}$	
Transformer (big)	**28.4**	**41.0**	$2.3 \cdot 10^{19}$	

References

Attention Is All You Need

https://arxiv.org/abs/1706.03762

The Transformer Attention is All you Need

https://mchromiak.github.io/articles/2017/Sep/12/Transformer-Attention-is-all-you-need/#.XYIj8ChKikw

Paper Dissected: "Attention is All you Need" Explained

http://mlexplained.com/2017/12/29/attention-is-all-you-need-explained/