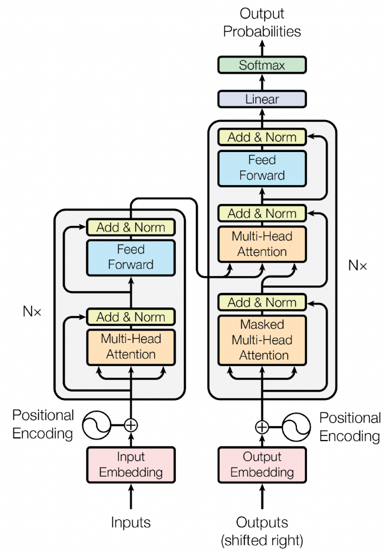
Attention Is All You Need

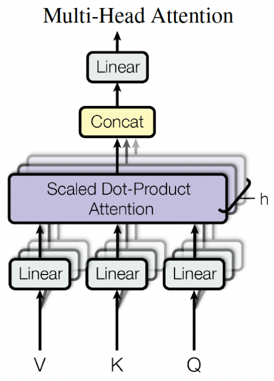
Status	Finished
Author	Ashish Vaswani et al.
Publishing/Release Date	@December 6, 2017
Publisher	
Link	https://arxiv.org/abs/1706.03762
Summary	Introduces a cross-modality method— Transformer
Score /5	⭐⭐⭐⭐⭐
Column	
Type	Academic Journal
Original Git Repo	https://github.com/tensorflow/tensor2tensor
Git Repo	
Tutorial	

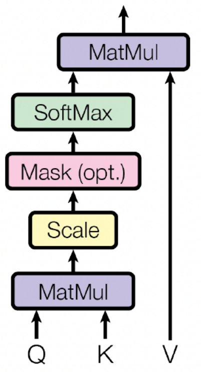
```
@inproceedings{NIPS2017_3f5ee243,
author = {Vaswani, Ashish and Shazeer, Noam and Parmar, Niki and Uszkoreit, Jakob and Jones, Llion and Gomez, Aidan N and Kaiser, \L ukasz
booktitle = {Advances in Neural Information Processing Systems},
editor = {I. Guyon and U. Von Luxburg and S. Bengio and H. Wallach and R. Fergus and S. Vishwanathan and R. Garnett},
pages = {},
publisher = {Curran Associates, Inc.},
title = {Attention is All you Need},
url = {https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf},
volume = {30},
year = {2017}
}
```

key concept behind the transformer:

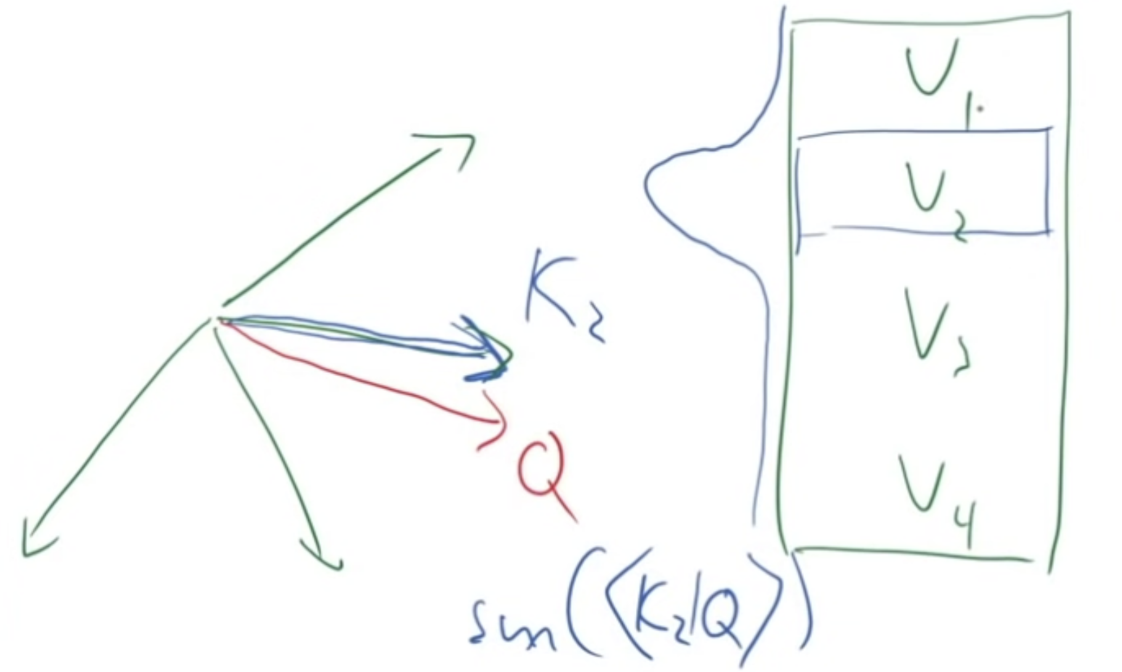


Unlike traditional NLP method (RNN), all the inputs and predicted outputs will be fed into the model at same time (PISO). RNN requires a sequential input (SISO), which rises a problem that the result is highly length dependent. If the input sequence gets longer and longer, the model will “forget” the contents which are far away from the prediction in terms of location, which results in a unsatisfied result.





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



Query, Key, and Value pairs work together to allow attention to specific feature. A good illustration is that key is the property of a person like name, weight, height. A value is the value attach to the property, for instance, if the property is name, then the value should be Yuqi. Query is asking the question which property of the person one wants to know. For instance, if one wants to know my height (Query), I will pick the height property (Key) from my info bank, and tell the person I am 180cm height (Value).

When Q dots K, we will mathematically select the Key which is closest to the Query, Softmax function exaggerates products. By normalizing, the desired product will be 1. Hence the desired Value of the Q-K pair is selected.

Problems with Transformer:

- There's an important limitation to the architecture of the Transformer: due to its **self-attention mechanism** , it scales **very poorly** in both compute and memory. In every layer, all inputs are used to produce queries and keys, for which a pairwise dot product is computed. Hence, it is not possible to apply self attention on high-dimensional data without some form of preprocessing.
- Time Complexity: O(M^2), in which M is num of K-Q pairs. Hence, it implies that once the inputs get huge, like image files, the time and space complexity will be a big issue