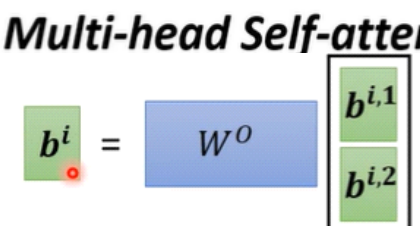
Self-Attention

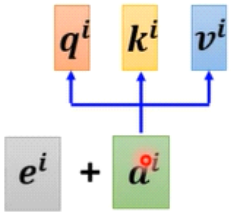
2021年4月1日 18:29



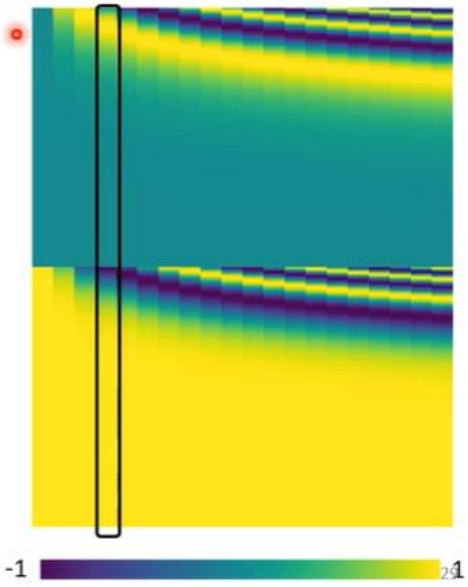
对于self-attention而言 少了一个位置的编码
所有的位置都是平权的
这样的设计是有问题的

Positional Encoding

- Each position has a unique positional vector e^i



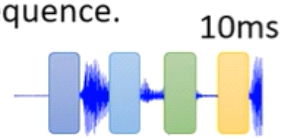
ng Each column represents a positional vector e^i



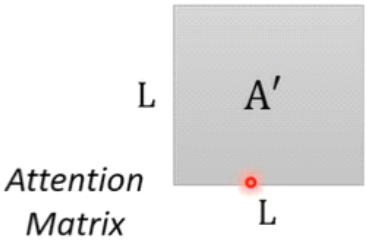
最早的vecotor（All you need is --）
通过sin和cos的方式产生的
Pos_en是一个尚待研究的问题，
可以当作参数learn出来，
目前尚不知道哪种方式最好

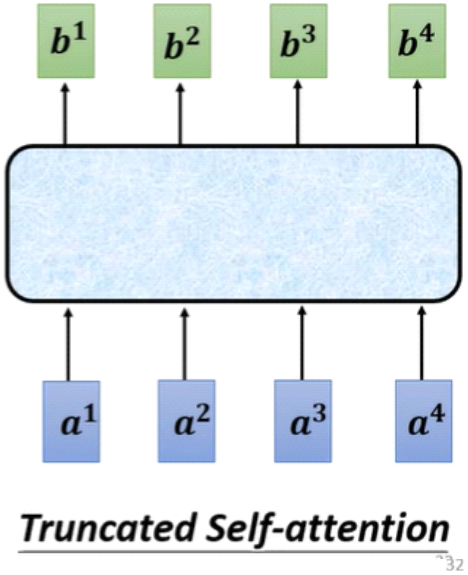
Self-attention for Speech

Speech is a very long vector sequence.



If input sequence is length L



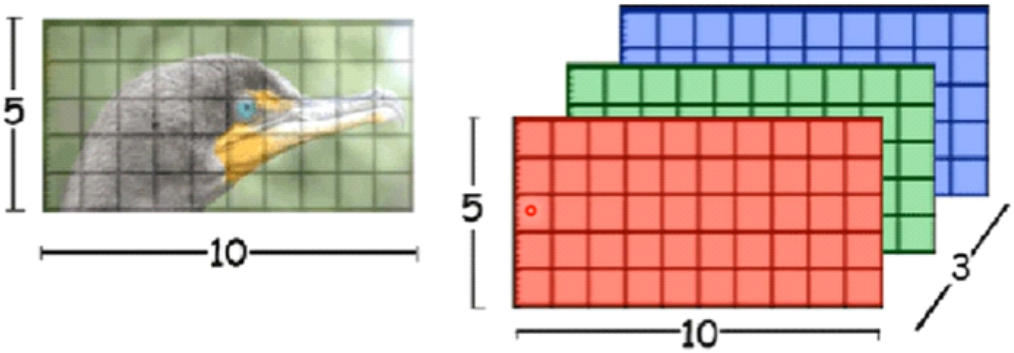


只看一个小的范围，不看全部的话

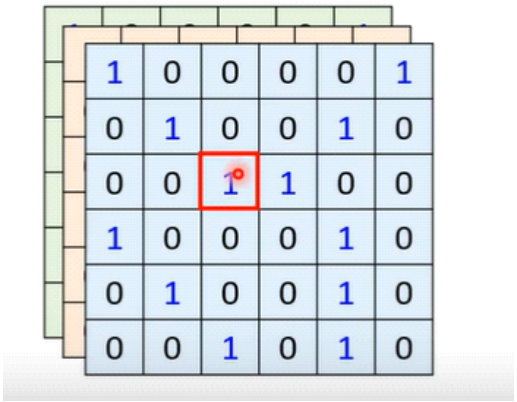
如果很长的话，矩阵会很大所以
Truncated-self attention

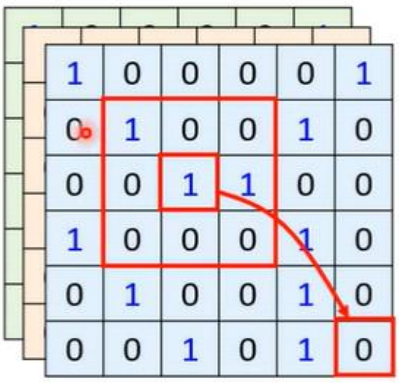
Self-attention for Image

An **image** can also be considered as a **vector set**.



Self-attention v.s. CNN

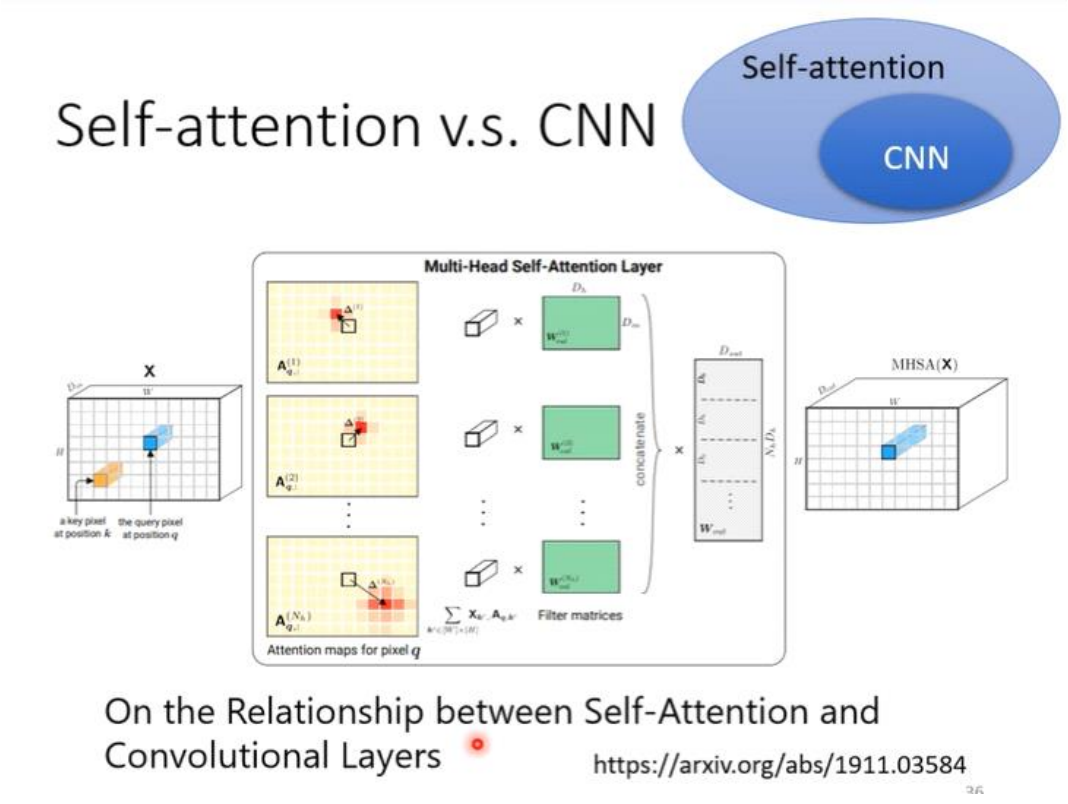




CNN 一种简化版的 Self- attention

CNN:self attention that can only attends in a receptive field.

Self attention : CNN with learnable receptive field.

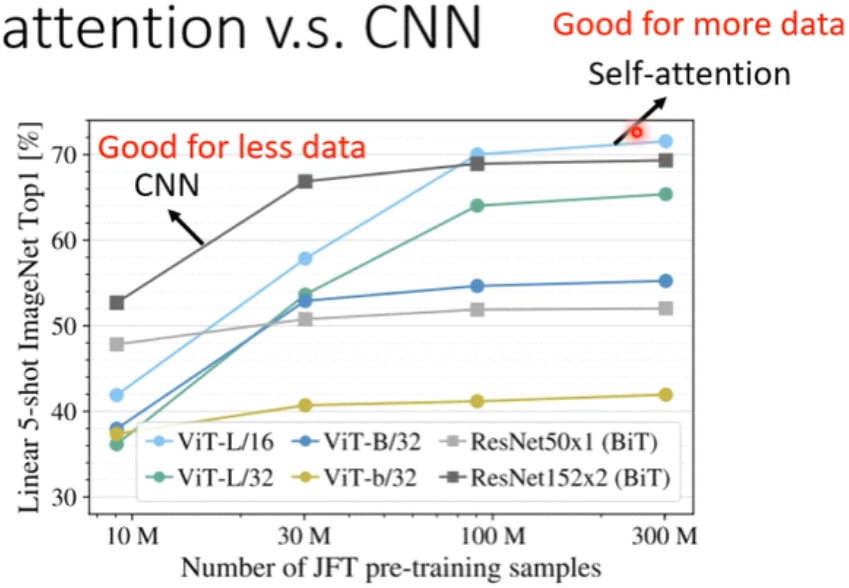


CNN是一种Self attention 的特例

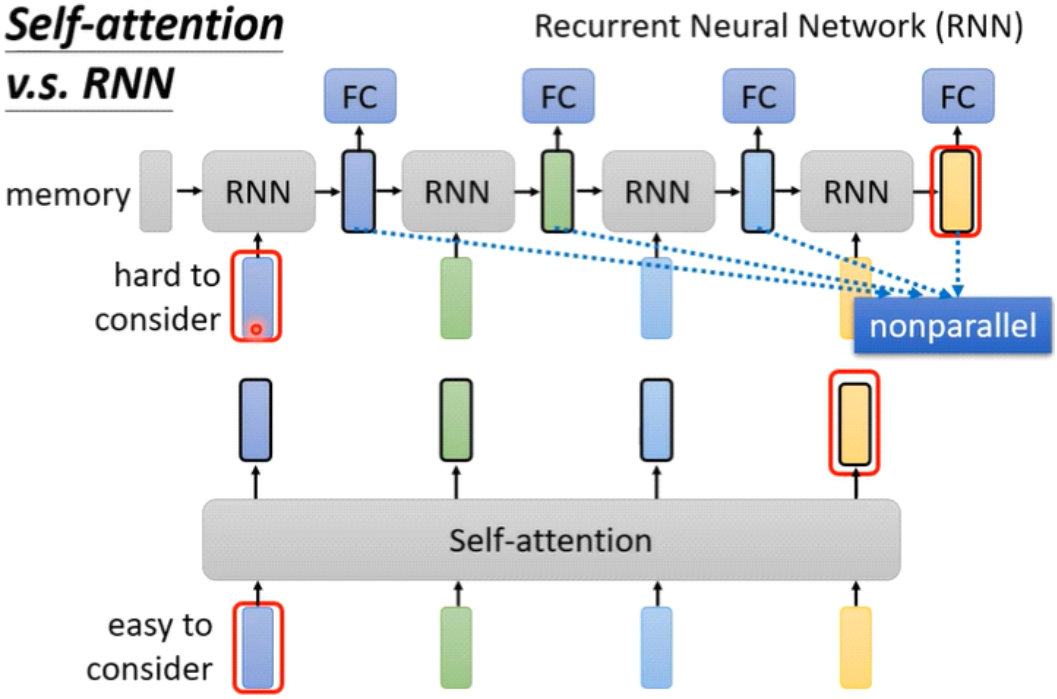
所以：Self attention 需要更多的 data

Self attention 的variance更大

Self-attention v.s. CNN



Self-attention v.s. RNN



RNN can't parallel so it requires more time

Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention https://arxiv.org/abs/2006.16236

RNN is included in self attention