Chapter21 Transformer Seq2seq model with 'Self-attention'

1. Sequence
   a) RNN : hard to parallel
   b) CNN : Use CNN to replace RNN
      Filters in higher layer can consider longer sequence
      CNN can parallel
2. Self-Attention
   a) You can try to replace anything that has been done by RNN with self-attention
      q : query (to match others) $q^i = W^q a^i$
      k : key (to be matched) $k^i = W^k a^i$
      v : information to be extracted $v^i = W^v a^i$
   b) Use every q do attention on every k
      Scaled Dot-Product Attention $a_{i,j} = q^i \cdot k^j / \sqrt{d}$ d is the dim of q and k
   c) Softmax
      $\hat{a}_{i,j} = \exp(a_{i,j}) / \sum_k \exp(a_{i,k})$
   d) Considering the whole sequence
      $b^i = \sum_j \hat{a}_{i,j} v^j$
   e) Parallel Computation
      $Q = W^q I \quad K = W^k I \quad V = W^v I$
      $\hat{A} <= A = K^T Q$
      $O = V\hat{A}$
3. Multi-head Attention
4. Positional Encoding
   No position information in self-attention
   Origin paper : each position has a unique positional vector e^i
   x^i concatenate with p^i $\Leftrightarrow$ $a^i + e^i$