注意力机制

**计算attention score的模组：**

dot product (常用) / additive

图示

描述已自动生成

**计算attention score:**

具体步骤，q与k，计算a1与其他每个向量（包括自己）的attention score(a)。再过一遍softmax层 (与分类模型一样)。

手机屏幕截图

低可信度描述已自动生成

**根据attention score抽取information – 计算b**

用Wv和第一步的a相乘获得v，再将v和第二步的a相乘，最后再相加。

图示

描述已自动生成

**结果的矩阵表达：**

图表, 瀑布图

描述已自动生成图示

描述已自动生成

只有Wq, Wk, Wv这三者是未知的。（透过training data找出）

**Self-attention的进阶版本： Multi-head self-attention**

图片包含 箱线图

描述已自动生成图形用户界面, 应用程序

描述已自动生成

**Position encoding:** (有各式各样的方法)

现在的self-attention没有位置信息，各个input后续的操作都是一样的。

为此，每一个位置，都应设有一个唯一的position vector (ei)

图示, 示意图

描述已自动生成

**Self-attention与CNN的关系：**

On the relationship between Self-attention and convolutional layers

**Self-attention与RNN的关系：**

Fast Autogressive Transformers with Linear attention

RNN没有办法平行计算，前者可以

Transformer

是一个sequence to sequence(seq2seq)的model, 可以做的工作很多。

**基础结构：**

Encoder处理数据，给Decoder，之后决定输出。

图示

描述已自动生成

**Transformer Encoder:**

给一排向量，输出另外一排向量。

图示

描述已自动生成

图示

描述已自动生成

最初，对input进行position encoding。

之后，从左下角开始，做self-attention，得到的a与input相加（残差结构）。

之后，过一遍layer Norm。得到一个值temp。

再temp过一遍Fully connected network，并将结果与temp相加（残差结构）。

最后再过一遍norm，得到最终的输出结果。

注意，这只是一个block的计算操作，在encoder中会有N个block。

**Transformer Decoder: Auto-regressive**

图示

描述已自动生成

**Masked self-attention:**

考虑b1时，只能看a1。考虑b2时，只能考虑a1, a2。

图示

描述已自动生成

**Encoder与Decoder之间的连接：**

图示

描述已自动生成

红色标注出来的地方是二者连接的桥梁，可以被称为cross attention。

Encoder提供左边两个箭头，Decoder提供另一个。

q来自Decoder，k与v来自encoder。

图示

描述已自动生成

**Model Training:**

图示

描述已自动生成

根据声音讯号得到中文的过程：

Decoder每得到一个字，相当于进行一次分类。Distribution可以看做一个带有所有中文字的字典。

在训练过程中，我们会给Decoder正确答案。(Teacher Forceing: 用正确的答案当作Decoder的input)

图示

描述已自动生成

有Begin的时候，输出机。有Begin有机的时候，输出器。有Begin，机，器的时候，输出学……