## Transformer

### Architecture：

整体结构：

图示

描述已自动生成

**注意这里下面的outputs是上一次计算的outputs**

Attention结构：

图示

描述已自动生成

self-attention公式：

文本

描述已自动生成

### Train&Evaluation：

1.总数据量4.5million，每个batch 25000

2.optimizer：Adam，β1=0.9，β2=0.98，ε=10-9，warm\_up step=4000

3.使用residual dropout，p=0.1

4.使用label smoothing，提升了准确率和BLEU score

### Skill&Knowledge：

1.提出self-attention，针对序列问题，不再需要使用CNN和RNN，显式使用encoder-decoder。

2.提出dot-product attention，速度比additive attention更快，对空间利用率更高

3.使用多头attention，类似于卷积时使用多个卷积核，然后最后合并，文中使用8个注意力层。实验证明，在增加注意力层时，计算耗时基本没有增加。

文本

描述已自动生成

4.BLEU score：Bilingual Evaluation Understudy（双语评估替换），用于评估评估自动机器翻译系统的预测结果而开发的。

当待评价译文同任意一个参考译文长度相等或超过参考译文长度时，BP值为1，当待评价译文的长度较短时，则用一个算法得出BP值。 以c来表示待评价译文的长度，r来表示参考译文的文字长度，则

文本

描述已自动生成

BLEU值(Bilingual Evaluation Understudy)计算为

图片包含 文本

描述已自动生成

在对数情况下，计算变得更加简便

手机屏幕截图

描述已自动生成

通常这个N取4，wn=1/4，这就是很多论文里面的一个经典指标Bleu4，其中，pn:

文本, 信件

中度可信度描述已自动生成

### Questions:

1.QKV分别是什么

### Improve&Validation：

## Extension