Universal Transformers论文解读

1. introduction
2. 模型介绍：
3. Dynamic halting
4. 实验和分析

Introduction：

本文提出了一种基于transformer的新型神经网络结构Universal Transformer(UT)

UTs combine the parallelizability and global receptive field of feed-forward sequence models like the Transformer with the recurrent inductive bias of RNNs.

我们知道，RNNs由于其模型结构循环特性，序列无法并行计算，因此训练速度慢。而Transformer则很好地避免了这个缺点。但是，Transformer 也RNNs没有的弊端：**无法概括归纳模型轻松处理的许多简单任务，例如，当字符串或公式长度超过训练时观察到的长度时，复制字符串甚至进行简单的逻辑推断。**

像Transfomer卷积和全注意前馈网络结构已经成为RNNs的替代品，在许多语言序列建模任务中，其在兼具并行性的同时，也解决了梯度弥散问题。

但是，**具有固定不同层堆栈的Transformer放弃了RNN对学习迭代或递归转换的归纳偏见**。

本文中提出的Universal Transformer,作为transformer的衍生模型，在一系列具有挑战性的序列到序列任务中产生了更高的理论能力并改善了结果，似乎更适用于许多算法和自然语言理解序列间问题。在某种特定的假设下，UT可以认为是Turing-complete。

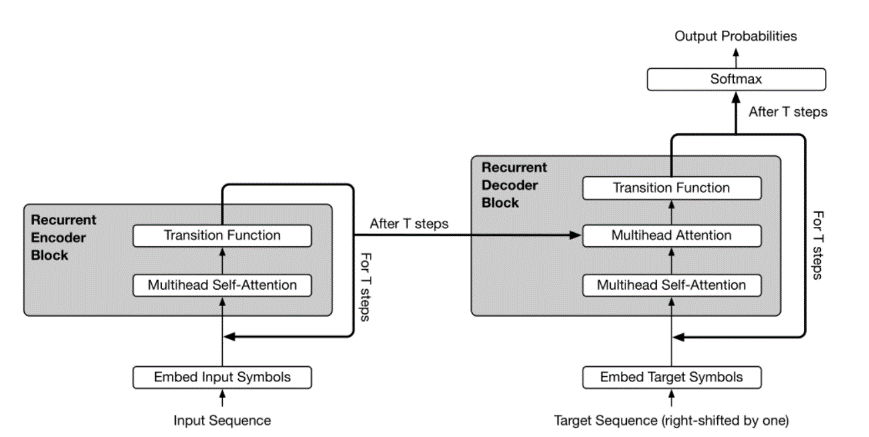
在每个循环step中，UT利用自我注意机制**并行地迭代**序列中所有symbols的表达。其中，模型中也在每个序列位置加入**动态暂停机制**，来动态地选择每个symbol的循环迭代次数。并且实验结果也表明，动态暂停机制的引入，提高了几个较小的结构化算法和语言推理任务的准确率。（但是MT任务的结果有所下降）

实验的结果表明，UT在一系列任务中都胜过Transformers和RNNs。循环机制的加入，使得UT在机器翻译任务中取得了比标准Transformer更好的结果。

**模型介绍：**

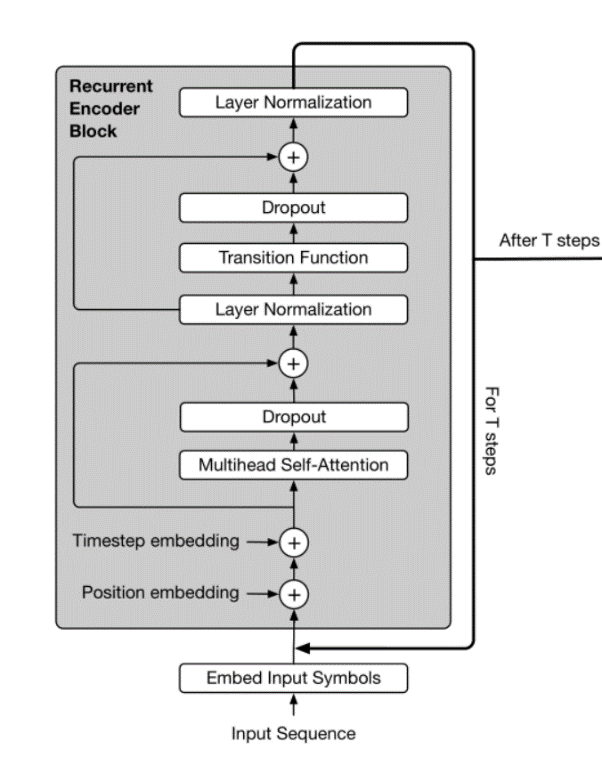
UT模型也是基于encoder-decoder结构的序列-序列神经网络模型，但是，UT模型在encoder和decoder中加入了循环机制来迭代输入和输出中每一个位置的向量表达。**但与RNNs对序列位置进行循环的方式不同，UT采取对序列每个位置的向量表达进行循环（深度）。**

在每一个循环时间步中，每个位置的向量表达以**两个步骤并行**进行：利用自我注意机制（self-attention mechanism）在所有序列位置中获取信息，产生每一个位置的向量表达；然后，在通过一个转移函数（transition function）输出。

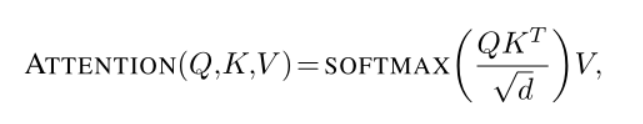


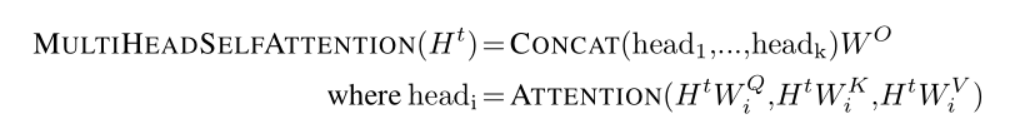
Encoder：

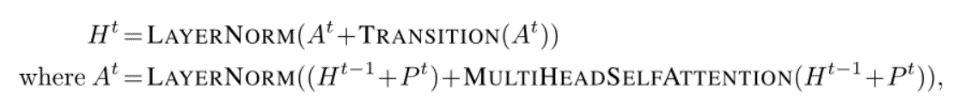
给定一个长度为*m*的输入序列，其每个位置的symbol embedding初始化为。在每一个时间步*t*中，运用多头自我注意机制和转移函数迭代计算所有*m*个位置的表达，在每块中还加入了残差连接，dropout和layer Normalization。如下图所示：



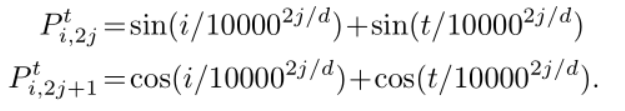
具体的计算公式如下：





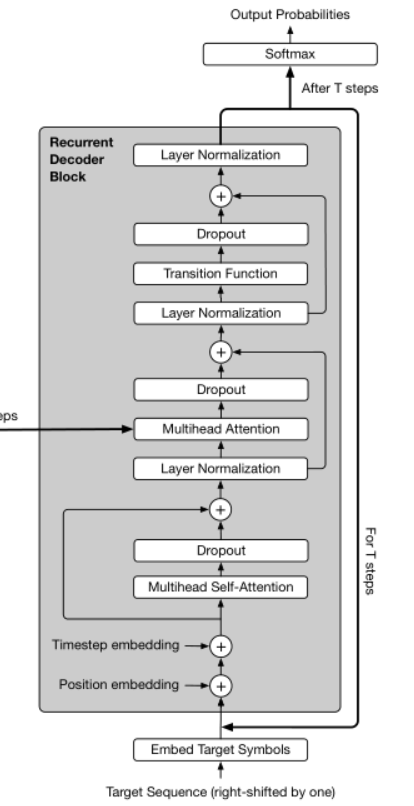


值得注意的是：是位置和时间维度的编码，编码函数均为*sin*函数，具体的编码计算过程如下：



Decoder：

Decoder的结构与encoder相似，只是在自我注意层和转移函数层之间额外加入了一个多头注意层（multi-head attention），多头注意层的输入*Q*来自上一层的输出，*K*和*V*则来自encoder的输出。同样，decoder也加入了循环机制，在经过*t*步循环后和Softmax层后输出预测概率。具体的网络结构如下图所是：

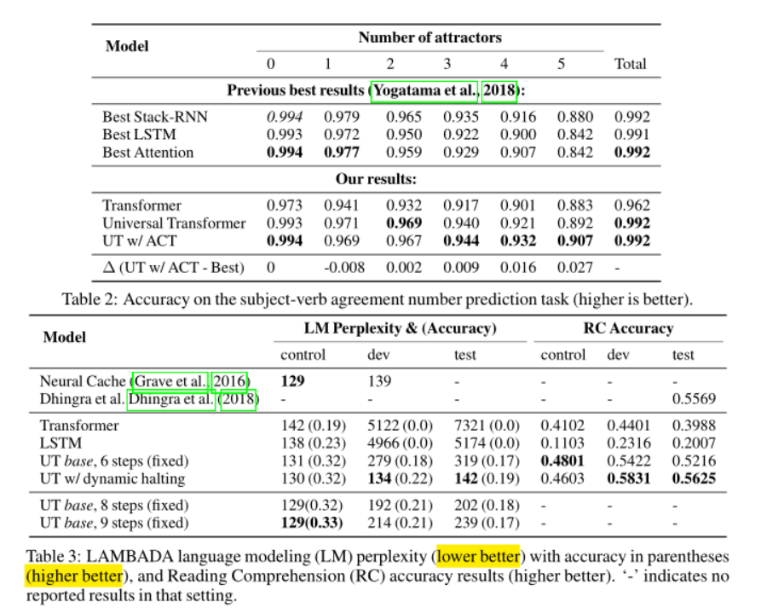


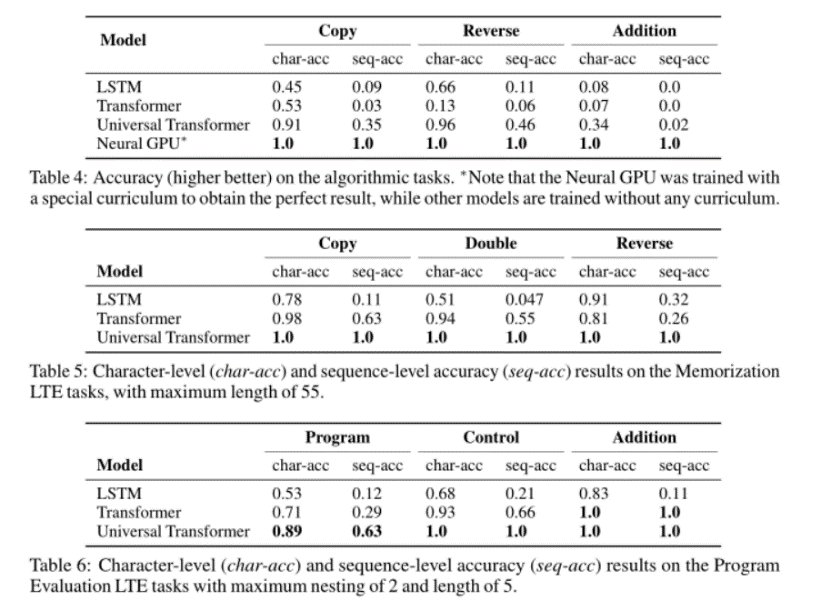
**Dynamic halting:**

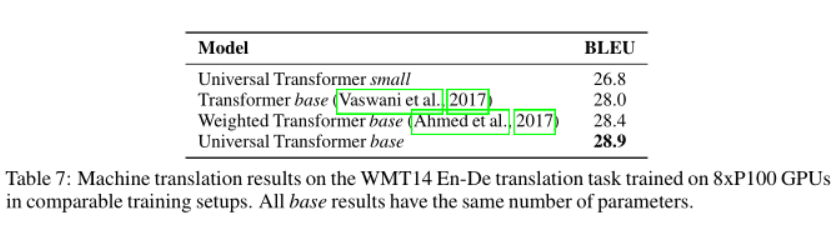
在处理序列过程中，存在某些symbol的含义比其他的更模糊或者模棱两可，因此有必要对这些symbol进行更多的处理。**Adaptive Computation Time (ACT)是一种在RNNs中用于动态计算循环次数的机制**，本文在对每个位置的循环迭代中加入了动态ACT暂停机制，来动态地循环处理不同的字符。

**实验和分析：**

本文在一系列的算法和语言理解任务的数据库中测试了UT网络，例如bAbi问答，subject-verb agreement ，LAMBADA。其结果如下图所示：

****

****

****

从实验的结果来看：UTs在许多算法和语言理解任务中展现出了比Transformer更高的准确率，尤其是在LAMBADA语言建模任务中，UTs取得了SoTA效果。

**个人总结：**

我觉得本文有以下几点创新：在Transformer结构中对每个位置加入了循环机制，除此以外，在循环中引入了动态暂停机制，使得UTs可以根据不同的输入选择合适的循环次数。