**阅读笔记：Adaptive Attention Span in Transformers**

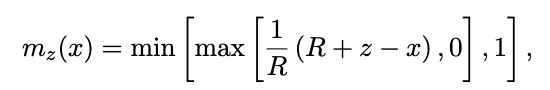
1. 解决的问题

Transformer在每一步都要计算之前所有的context的attention信息，复杂度为O(n^2)。这使得token的长度被限制在1k左右，基于这种方法的语言模型的序列长度很难被扩展。且计算长距离信息的时间和存储代价会随着序列长度的增加成倍增长。本文提出一种自适应宽度的方法，一可以扩展输入的长度，二可降低计算和存储开销。

1. 贡献
2. 提出一种降低Transformer的计算和存储开销的方法，对Transformer的自关注层进行改进，学习每层的最佳上下文窗口大小，每一层均整合与该层相关的上下文信息。
3. 引入相对位置嵌入以及加速训练和测试过程的缓存机制。
4. 方法

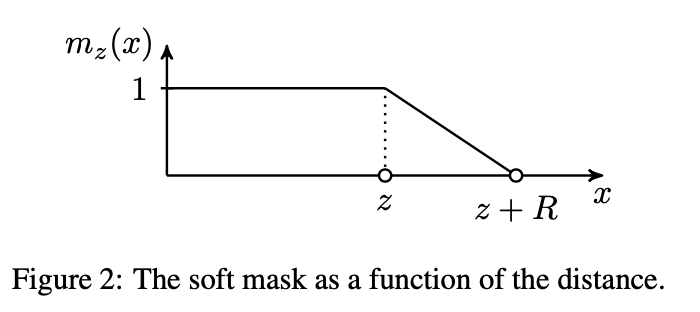
方法的灵感源自于对每个head的观察，Head A关注最近的attention信息，Head B关注几乎全部的token。故作者考虑一种忽略不相关的上下文信息的算法，即自适应注意力宽度算法。

对Transformer的每个head，都增加一个masking函数去控制注意力的宽度，注意力函数是一个非增函数，该函数将注意力的距离映射到[0, 1]之间，用masking 函数mz实现自适应宽度：

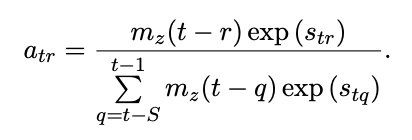


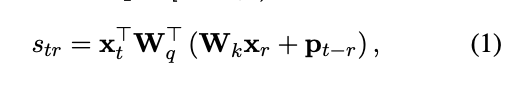
其中R是控制柔软度的超参数，z∈[0,S]

该masking函数为分段函数，Fig 2为函数图像：

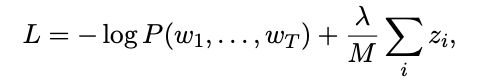


然后在masking函数基础上计算attention权重，使用上下文和当前位置的距离作为mask函数的输入，attention的权重如下：



其中str为第r个token在[t-S, t]范围内与其之前的token的相似度，

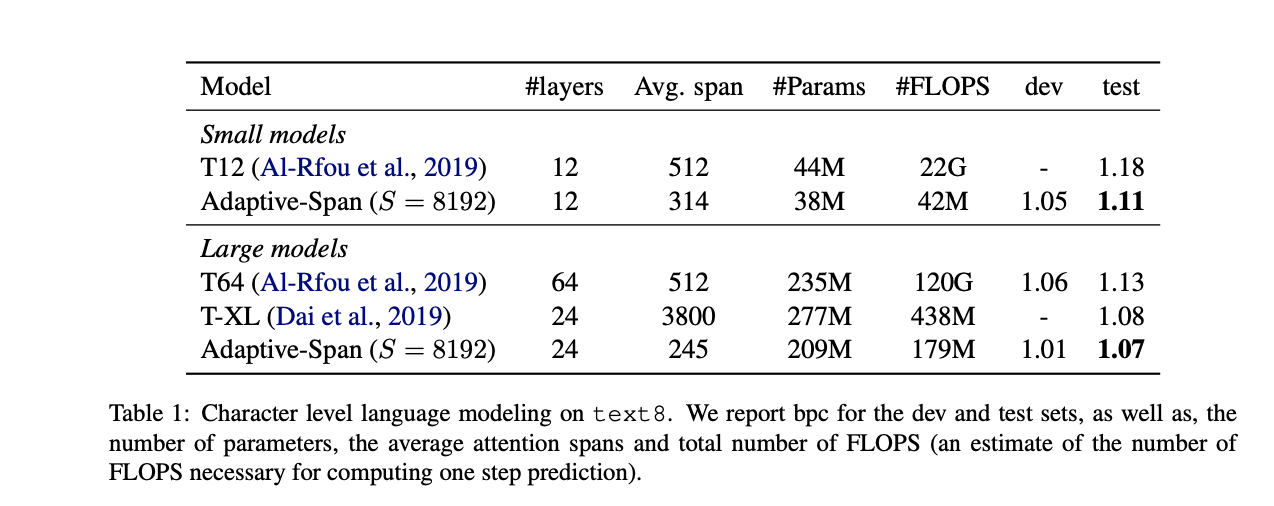
在损失函数中给z添加一个L1正则项：

，其中λ是一个正则化超参数，M是每层head的个数。

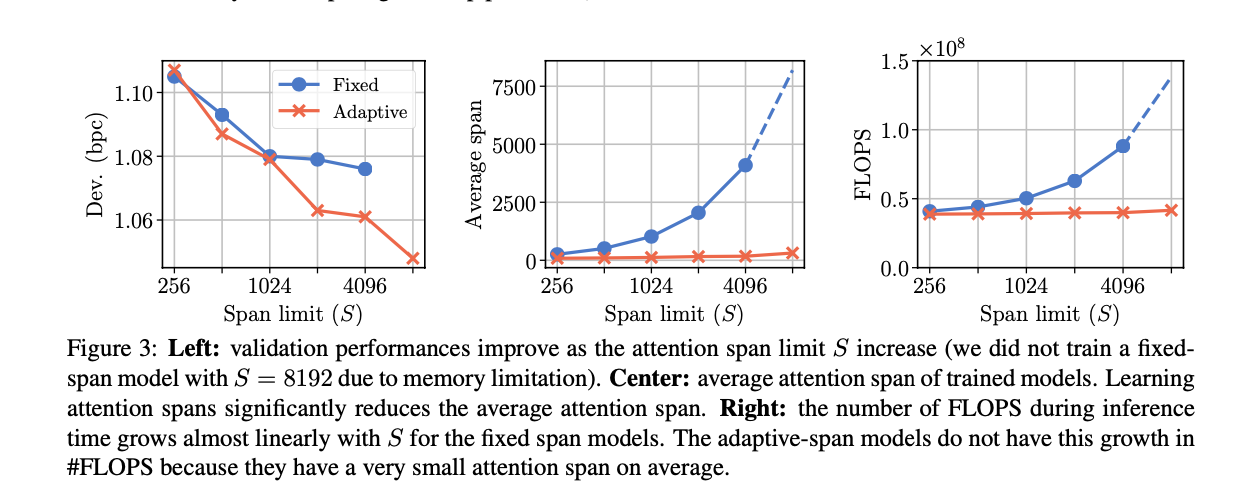
1. 实验

数据集：text8 & enwik8

模型及其参数设置对计算量的影响：



实验结果：



随着输入长度的增长，注意力的平均宽度和计算量的对比，自适应宽度算法的注意力宽度变化不是很大。

**实验：HMM、CRF、BILSTM、BILSTM+CRF for NER**

**中文数据集：**2018中文NER Lattice简历数据，数据来自新浪财经，包括中国股票市场上市公司高管的简历。采用BIOES标注法，每个句子之间用空行隔开，实体被分为NAME、COUNT、RACE、EDU、TITLE、ORG、共3×6+1（O）=19种不同标签。

**中文数据实验结果**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **模型** | **HMM** | **CRF** | **BiLSTM** | **BiLSTM+CRF** | **Ensemble** |
| **Precision** | 0.9149 | 0.9543 | 0.9560 | 0.9561 | 0.9567 |
| **Recall** | 0.9122 | 0.9543 | 0.9556 | 0.9561 | 0.9566 |
| **F1** | 0.9130 | 0.9542 | 0.9555 | 0.9558 | 0.9565 |

**英语数据集：**CoNLL 2003中1393篇英语新闻，数据采用BIO标注法，实体被分为LOC、PER、ORG、MISC四种类型，共2×4+1（O）=9种不同标签。

**CoNLL2003英语新闻实验结果**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **模型** | **HMM** | **CRF** | **BiLSTM** | **BiLSTM+CRF** | **Ensemble** |
| **Precision** | 0.9290 | 0.9301 | 0.9118 | 0.9227 | 0.9419 |
| **Recall** | 0.9312 | 0.9321 | 0.9127 | 0.9095 | 0.9441 |
| **F1** | 0.9252 | 0.9265 | 0.9111 | 0.9135 | 0.9404 |

**实验结果分析**

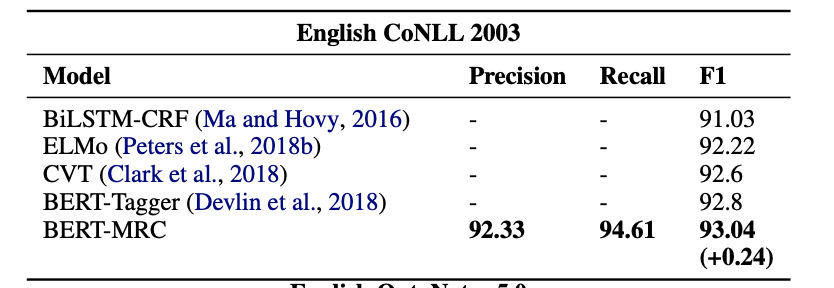
对于同一中文数据集，各个模型的表现，

HMM<CRF<BiLSTM<BiLSTM<BiLSTM<Ensemble；

而在同一英语数据集中，BILSTM<BiLSTM<HMM<CRF<Ensemble，与预计不一致的是在英语数据集CoNLL上的结果。

猜测英语数据集上结果可能的原因：

1. 嵌套实体，CoNLL中含有一定比例的嵌套实体（具体比例暂时没查到），论文A Unified MRC Framework for Named Entity Recognition提出了一种基于机器阅读理解(MRC)的方法去解决嵌套实体问题的方法，作者在CoNLL2003上做了实验，BERT-MRC的方法结果要优于BILSTM-CRF的实验结果。



其中BILSTM-CRF结果比本次实验结果低的原因可能是本次实验采用了更小的词向量维度和隐藏层层数。

1. LSTM参数的影响：尝试降低词向量维度和提高隐藏层层数之后实验结果有很大提升。词向量维度从128降到100，隐藏层层数从100提升到200。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **模型** | **HMM** | **CRF** | **Bilstm** | **BiLSTM+CRF** | **Ensemble** |
| **Precision** | 0.9149 | 0.9543 | 0.9537 | 0.9574 | 0.9569 |
| **Recall** | 0.9122 | 0.9543 | 0.9532 | 0.9574 | 0.9565 |
| **F1** | 0.9130 | 0.9542 | 0.9532 | 0.9570 | 0.9564 |