论文阅读笔记一

Linformer: Self-Attention with Linear Complexity

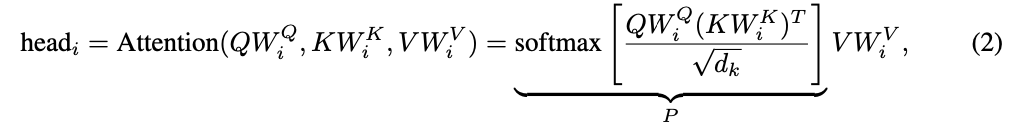
1. 贡献

用低秩矩阵逼近的方法实现注意力机制，将时间和空间复杂度由O(n2)降到了O(n)。

1. 方法

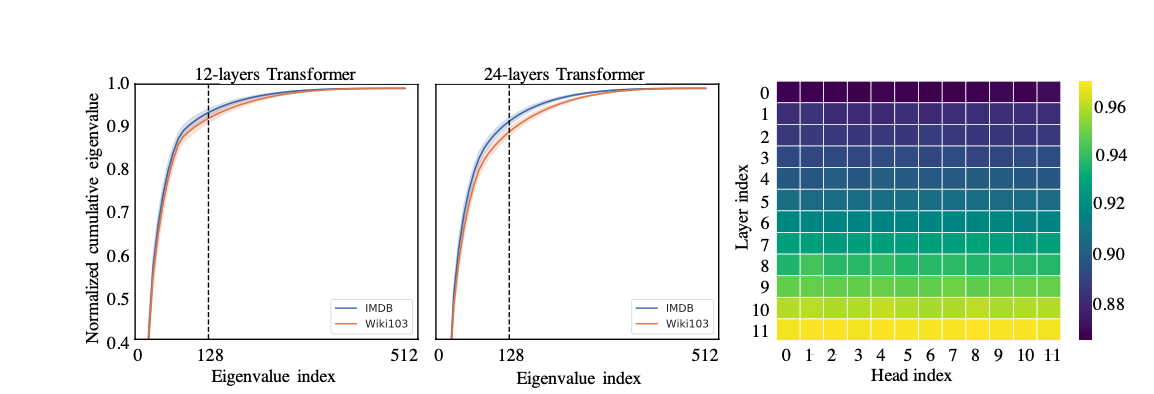
将Transformer中的点积注意力拆解成了多个更小的线性投影注意力，即对原注意力矩阵做低秩因式分解。







本文首先分析了注意力矩阵的奇异值分解频谱分析。



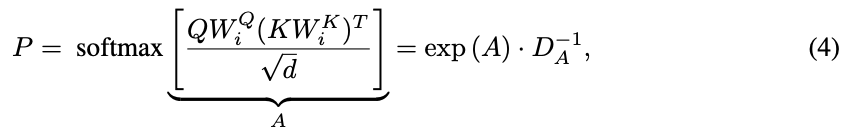
对模型的不同层、不同注意力头对应的注意力矩阵P都进行的奇异值分解（SVD），并将超过10K的句子上的归一化的累积奇异值做了平均，结果显示沿着不同层、不同注意力头和不同的任务，都呈现出一个清晰的长尾分布。这表明注意力矩阵中的大部分信息都可以由少量最大的奇异值 来恢复。如上面的热度图可以看到，最高层的Transformer中会有 比更底层 的Transformer有更大的偏度，这说明在更高层中有更多的信息存在于少量最大奇异值中，且注意力 矩阵的秩是更低的。

作者为这样的频谱结果提供了一个理论分析，那就是，

定理一：自注意力是低秩的。在对于矩阵中任意满足的列向量，以及都存在一个低秩矩阵满足以下关系：



证明：

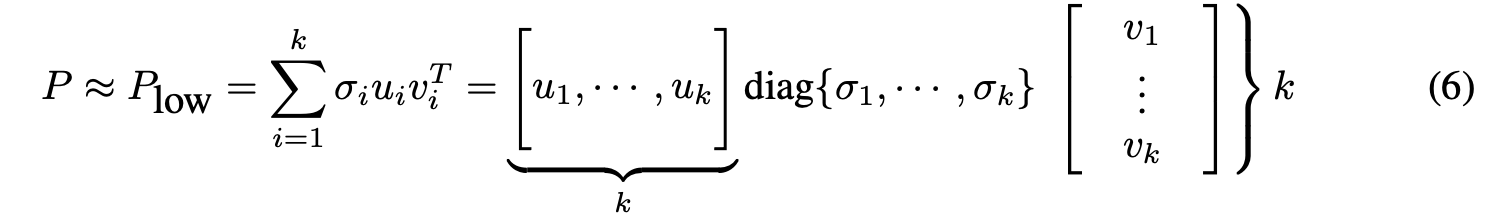


基于公式(3)，矩阵D是一个n×n的对角矩阵，该证明的主要思路是约翰逊-林登斯特劳斯定理。首先构建一个服从独立同分布的低秩矩阵



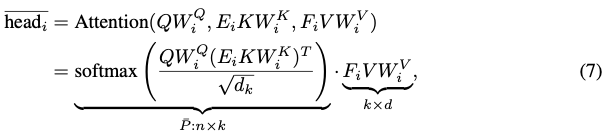
一个高维空间中的点集，可以被线性地镶嵌到低维空间中，且其空间结构只遭受较小的形变。JL 定理的证明说明了如何用随机投影法来明确求出这个变换，且该算法只需要多项式时间。降维是有代价的。如果要求尽可能地减少形变，被嵌入的低维空间则不能很低。反过来，如果要尽可能地压缩，则形变会不可避免地增加。最终JL定理给出的结论是，将维数下降到样本数的对数级，更兼容的变换是线性的，显式的，且可以被快速计算的。

当注意力矩阵P有了低秩性质之后，一个直观的想法就是用SVD来求解一个Plow来近似PP，该矩阵满足：



其中，σ，μ和v都为对应奇异向量的最大奇异值。基于定理1和低秩矩阵近似定理，可以用Plow去近似注意力矩阵。故时间和空间复杂度近似为O(nk)。

基于上面的矩阵低秩性的理论，本文提出了一种新的注意力机制，让模型能够以线性复杂度来计算注意力矩阵。模型核心思想就是用两个线性投射矩阵Ei，Fi来计sho算K矩阵和V矩阵。首先将征集的KW和VW映射到k\*d的上下文映射矩阵中，然后就点积注意力计算矩阵



定理二：线性的自注意力。对于任意的Q、K、V及其权重矩阵，如果，那么一定存在矩阵Ei，Fi对于中任意的行向量w，会有：



该式的证明基于JL定理，首先证明中的每个行向量x∈，和矩阵中的每个列向量y∈，都满足

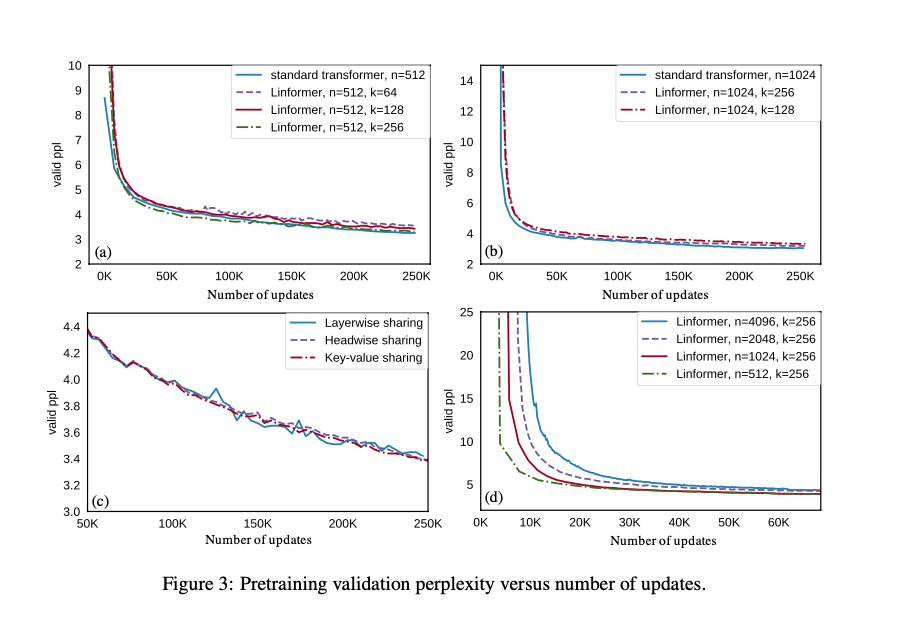


实验采用了 三种层级的参数共享，分别是注意力透共享E、F参数；键值参数共享、层参数共享。

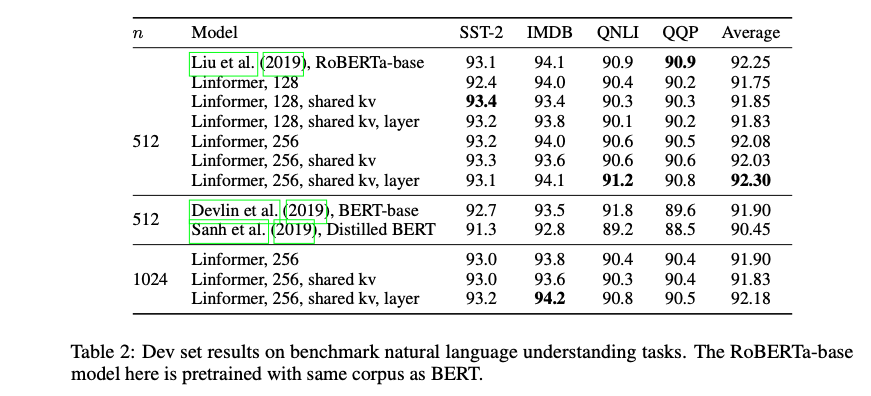
1. 实验结果

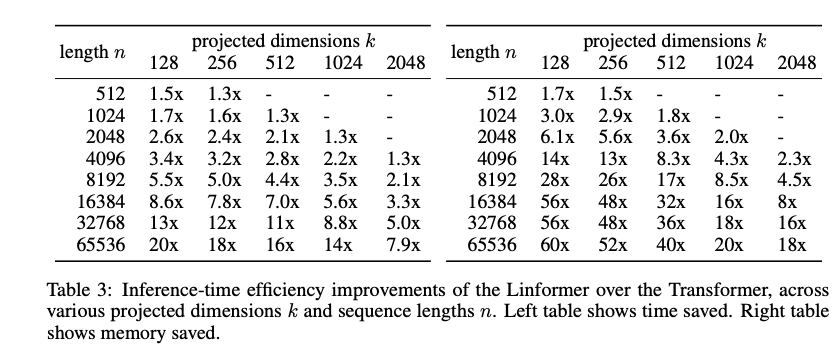
模型在已有的预训练模型RoBERTa架构下用BookCorpus和英文维基百科进行训练，共3300M单词，模型训练的膜表都是MLM，在TeslaV100上进行了250k次迭代。

预训练的实验目标主要验证k值的选择问题、参数共享策略和序列长度读模型的影响。



从图a可以看出，k值越大，困惑度越低，模型训练得越好；图b说明，k值从256下降到128，困惑度变化不大；图c展示了在不同共享参数下的效果；d比较了在固定k值的情况下，不同的序列长度对其困惑度的影响。结果显示随着长度增加，初始的困惑度长序列会比较大。但最终收敛之后，不同长度序列的困惑度都差不多。这验证了 Linformer 是线性复杂度。





阅读笔记二

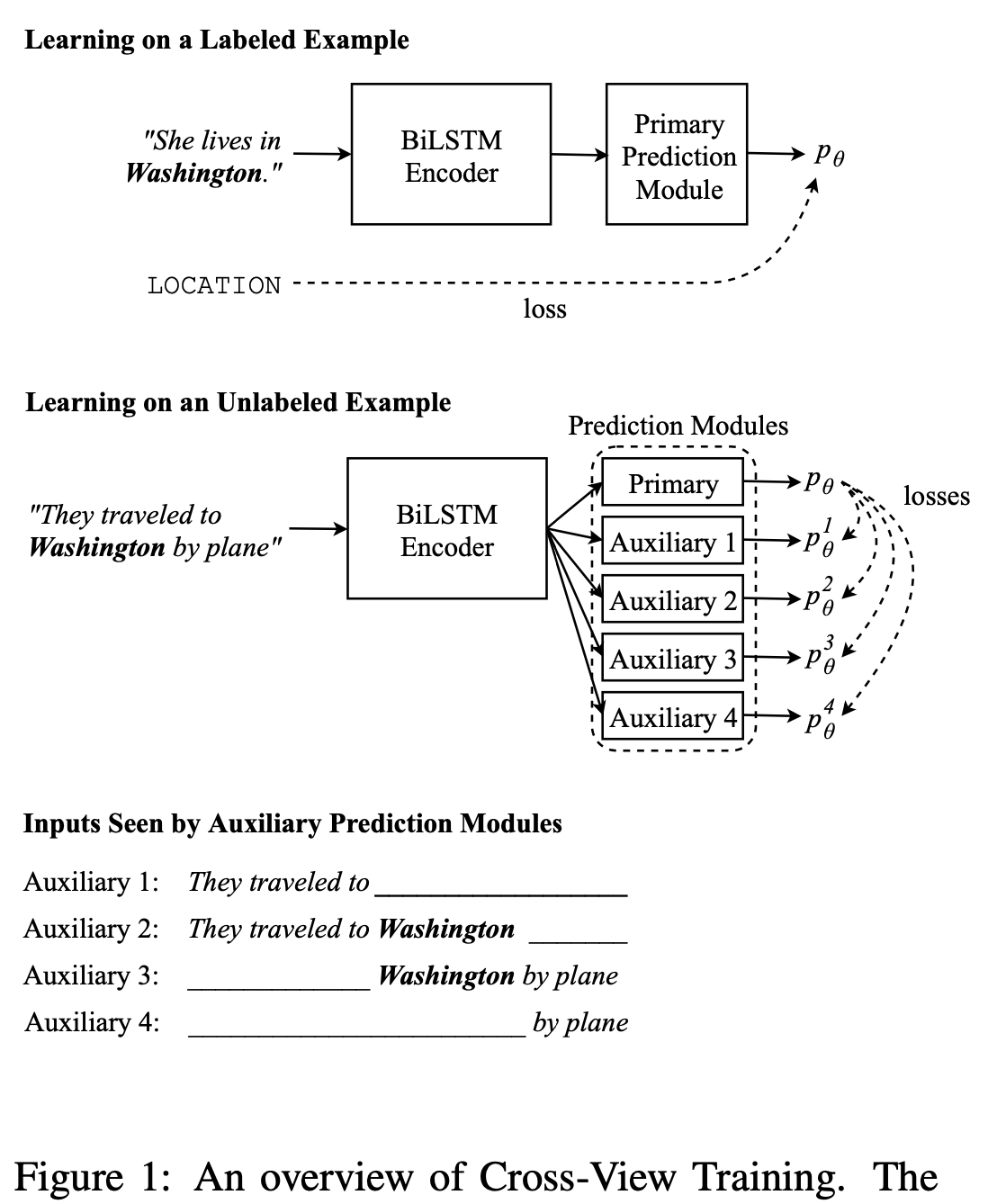
Semi-supervised sequence modeling with cross-view training

1. 思想

有监督+无监督学习相结合，无监督学习可以充分利用未标注的文本，有监督学习在训练阶段学习标记数据之间的约束。

本文提出一种交叉视角训练（Cross-View Training, CVT）的半监督学习方法，混合使用标注和无标注数据来改进数据的表示。用self-training训练无标注数据，在有标注数据上训练好模型，给无标注数据一个pseudo-label，并将这个标签加到有标注的数据中重新训练。如果这个pseudo-label是错的，在训练过程中可以进行修正，使得结果更加有效。

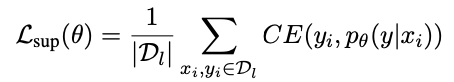
1. 方法



1. 编码器

有标注和无标注数据均用LSTM编码，将字符级和单词级的词嵌入表示直接相加，得到编码器的输出。

1. 解码器

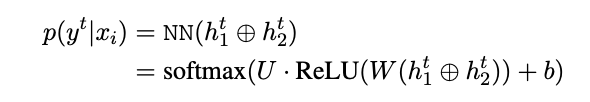
解码器部分根据任务的不同单独设计。通过对传入解码器的词嵌入信息加入不同的限制来获取多个view的输出。以有标注（不受限解码器）和无标注（受限解码器）的输出之间的差异作为监督信号进行训练。

其中 k表示辅助模块的个数，也就是view的个数

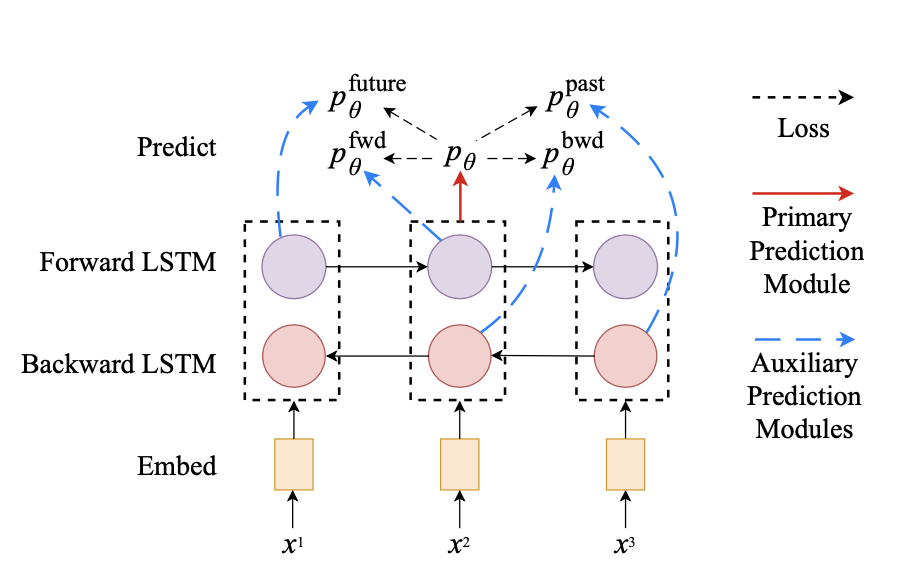
编码器共享，在降低loss的时候，促使编码器抽取出更好的特征表示用来提高有标注数据的解码器的性能。因为 监督信号不需要标注信息，故无标注的数据可以得到利用。

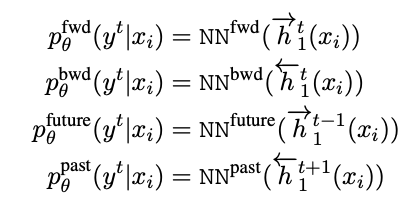
1. 序列标注任务

Primary prediction module：一个含有一个隐藏层的神经网络。



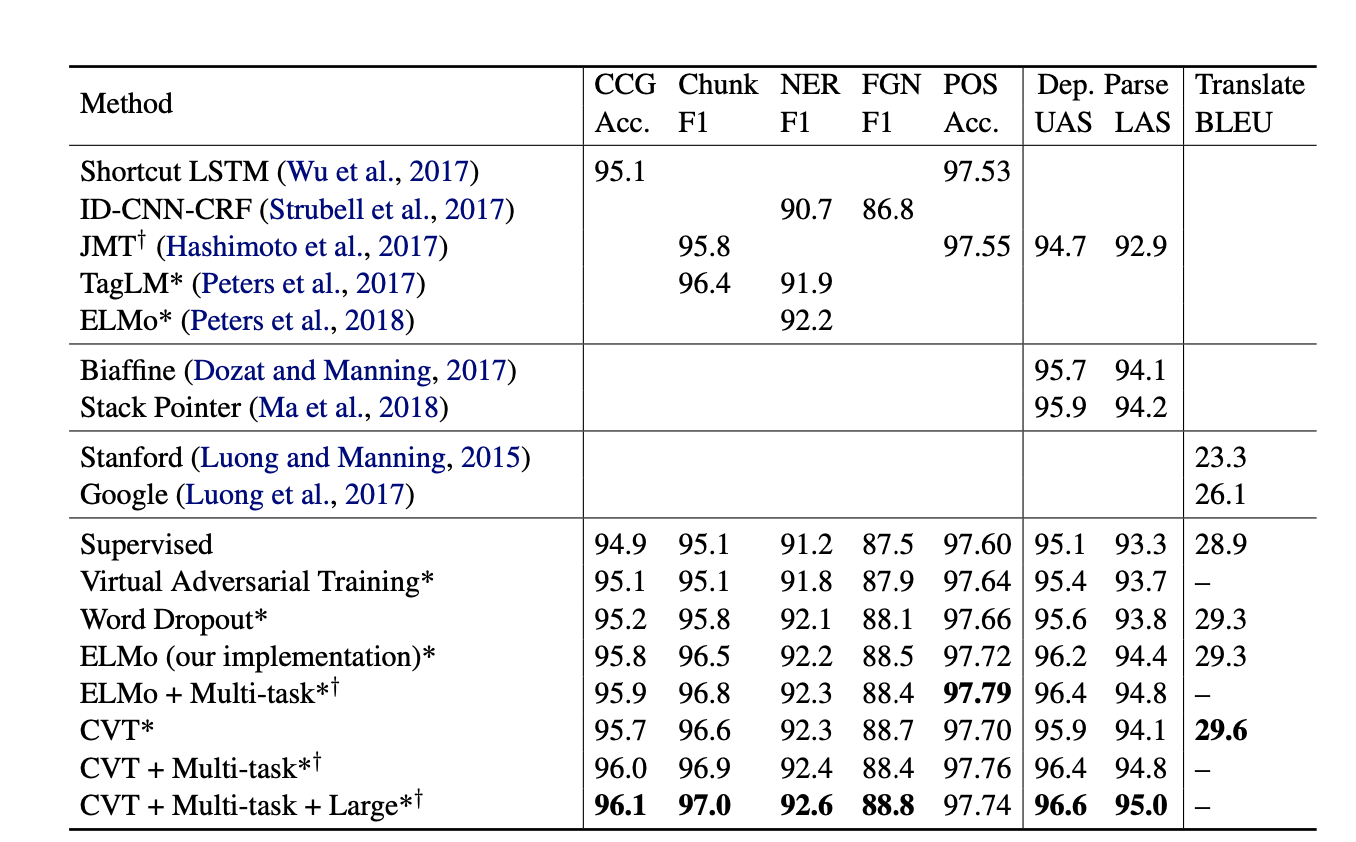
Auxiliary prediction module：也是一个含有一个隐藏层的神经网络，作者只用了LSTM的第一个隐藏层。



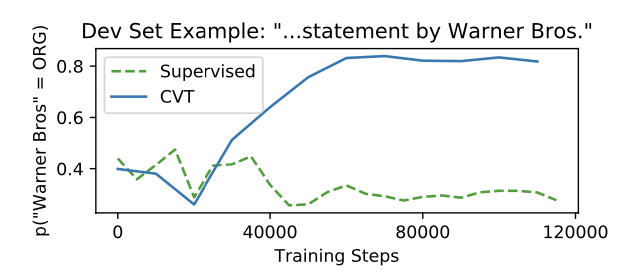


1. 实验结果

作者在7个序列任务上做了实验，并与3种半监督方法进行对比。下表是模型在所有测试集上的结果：



作者还做了一个使用了CVT的无监督学习的NER的实验



在数据集中，Warner仅作为姓氏在训练集中出现，有监督学习会将Warner分类为人名