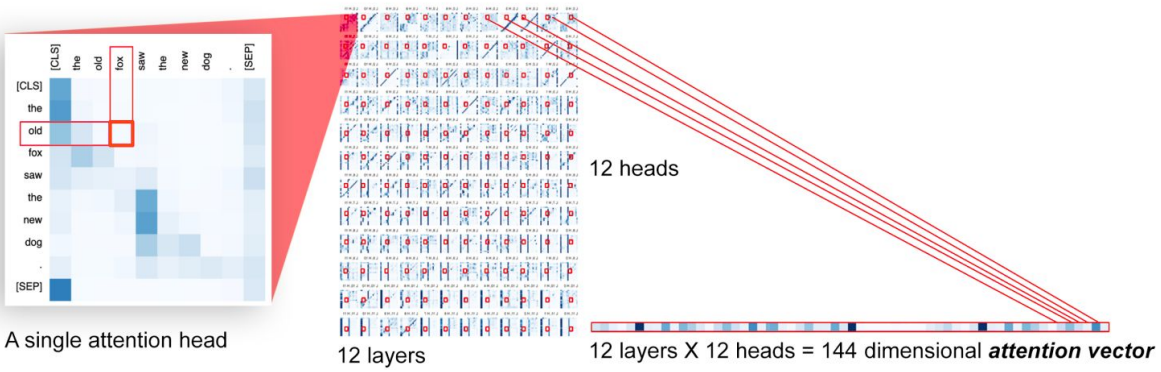
引言

BERT和类似的语言模型到底学会了何种语言的关系和模式，一直是大家想知道的，本文尝试从几个实验探索这一点。

假设：Attention学习到了词之间的依赖关系（依存关系）

Bert一共有12层，每层有12个head（multi-head attention），每个attention head相当于一个SxS的矩阵，这里S是句子长度（或者说包含的词数），也就是说在每个attention矩阵中，两个词包含两个关系，指向与被指向，我们只考虑指向的话就是一个浮点数。12层x12个head=144维度的attention vector



方法：

输入attention vector训练两个简单的linear model，分别是binary的和softmax的，代表词之间是否有依存树上的指向关系和具体的指向标签。

准确率达到了85.8%和71.9%，当然不是SOTA，不过也证明了这种关系应该包含在了attention表示中。

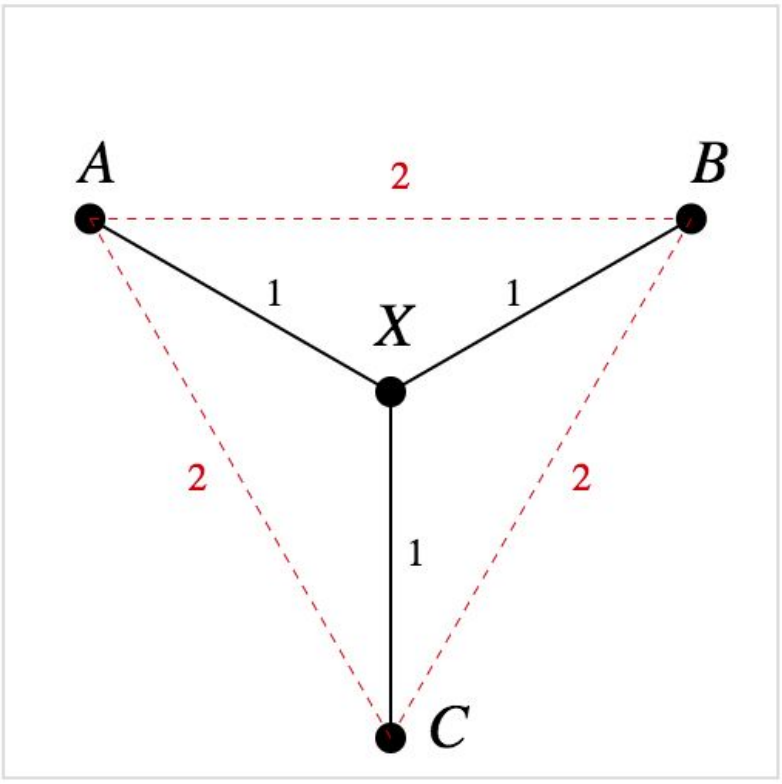
问题：如果从embedding的角度表示树？

基本讨论：我们不能使用可以同构（isometrically）的距离表示树

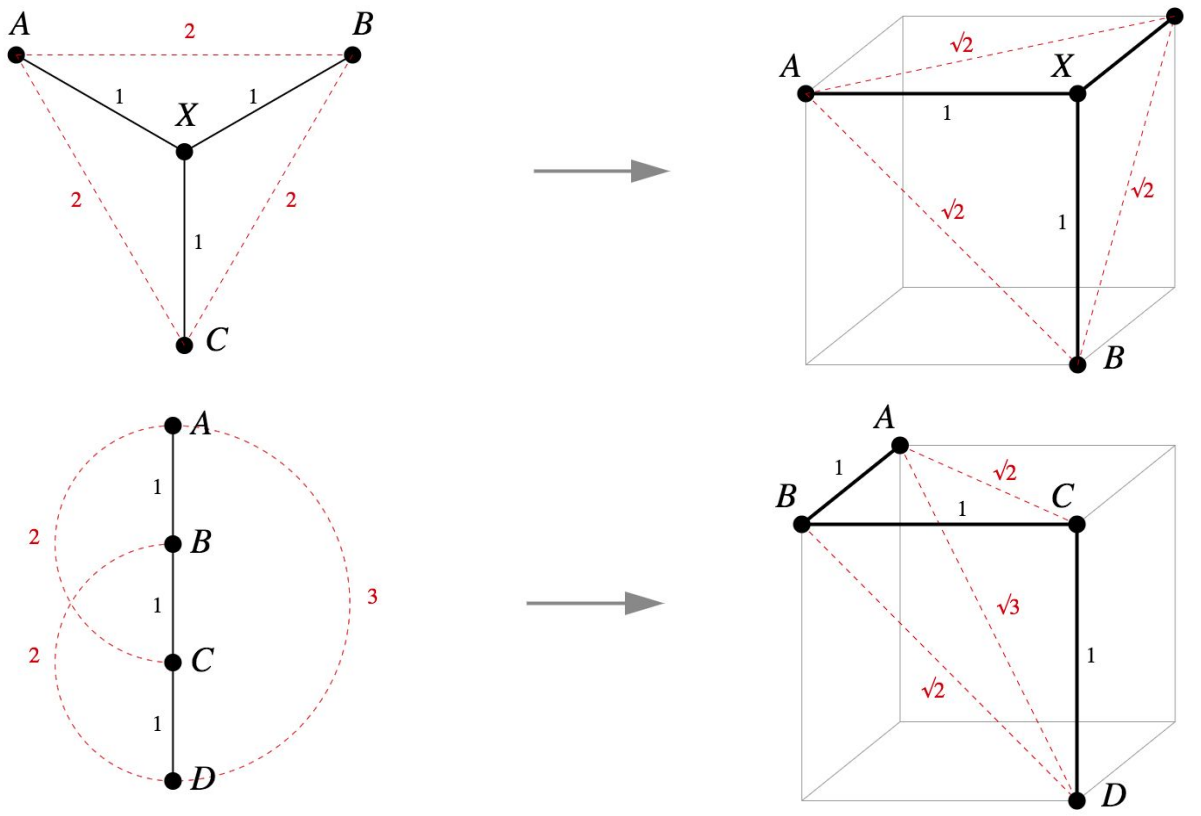
这里isometrically应该是指，假设ab和bc分别父子，ac则为爷孙，则 d(a, c) = d(a, b) + d(b, c) 不能成立

如果可以同构，假设表示一个树有三个几点A、B、C，假设A、B是父子距离为1，B、C是父子距离为1，则可以得出A到C的距离为2，则在欧基里德空间中ABC应共线。
那么假设B还有一个不同于C的孩子D，按照假设ABD也应该共线，但是这样就说明C=D，则不可能。

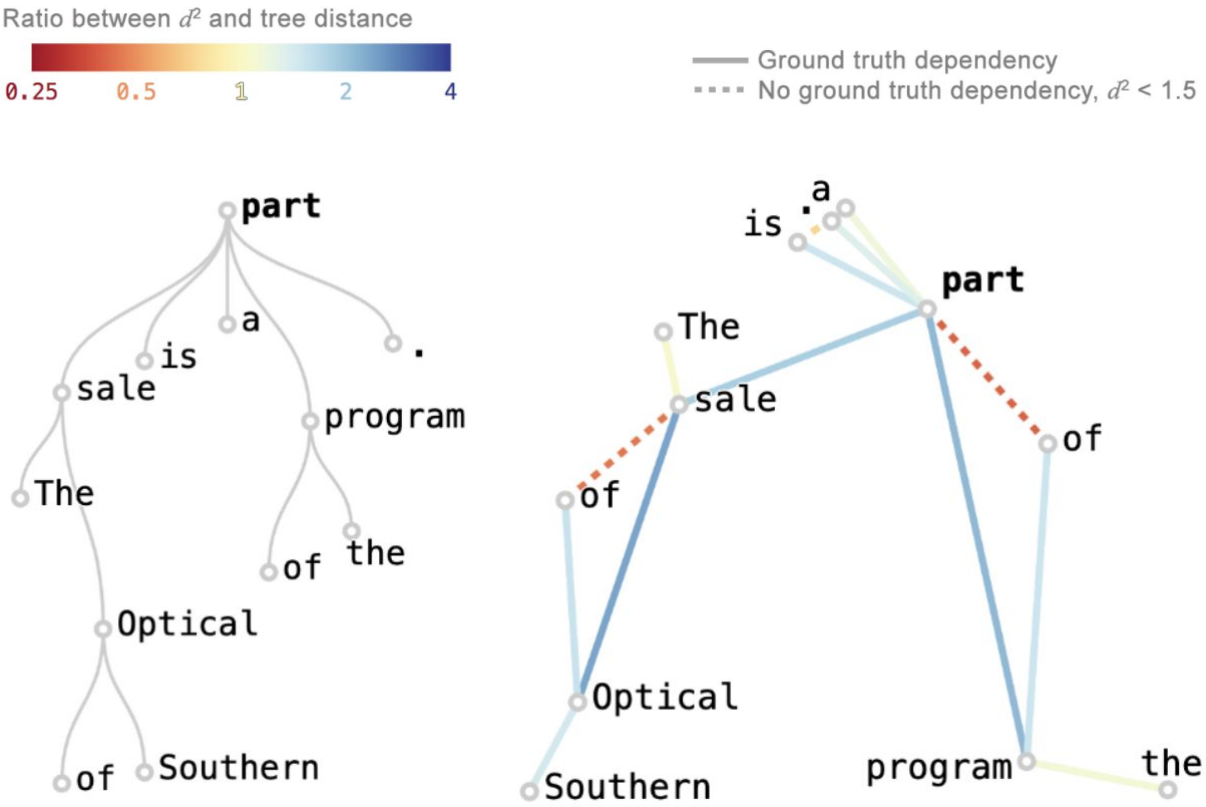
所以下图的同构情况不可能：

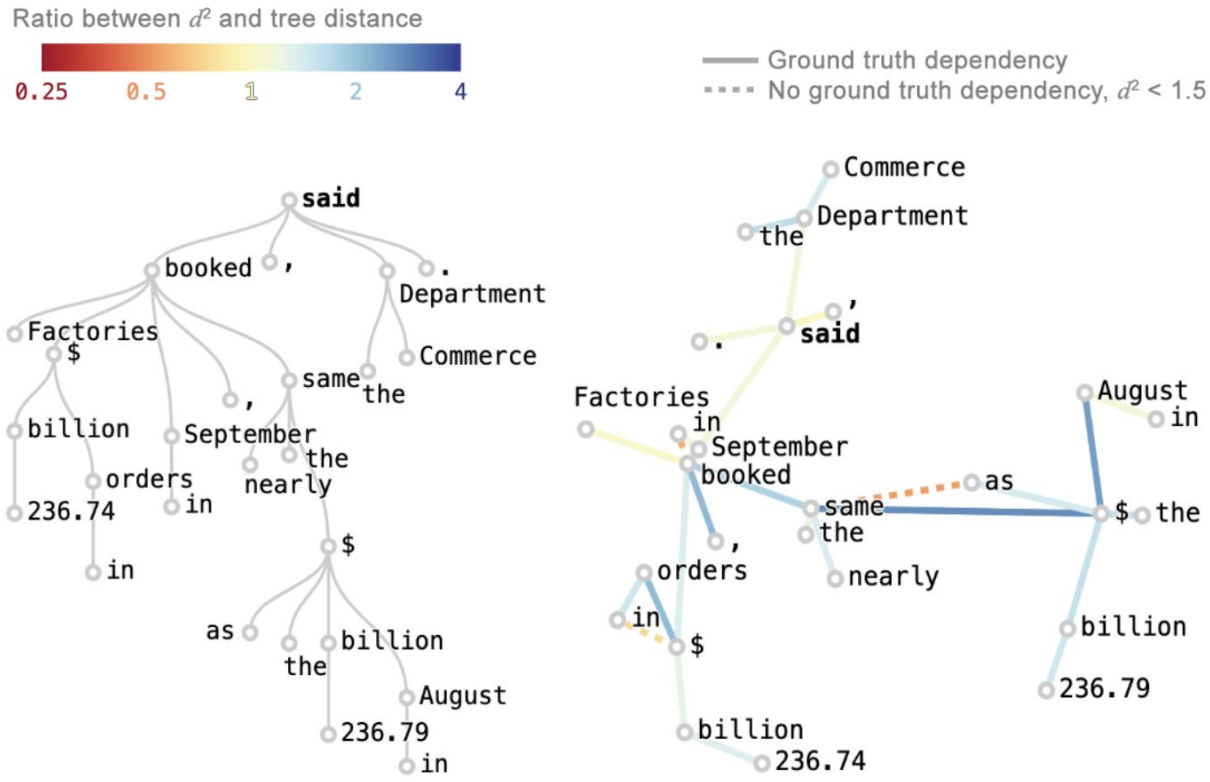


所以我们事实上是只能用下面的距离来表示一棵树的：


A
B
X
C
2
1
1
2
2
1
A
X
C
B
√2
1
√2
1
√2
A
B
C
D
1
2
3
2
1
1
1
A
B
C
D
1
√2
1
√3
1
√2

结果：树的可视化


Ratio between d^2 and tree distance
0.25
0.5
1
2
4
Ground truth dependency
No ground truth dependency, $d^2 < 1.5$
part
a
is
.
sale
program
The
of
the
Optical
of
Southern
a
.
is
part
The
sale
of
of
Optical
Southern
program
the



词义信息表达

这一部分主要的猜想是BERT是否能区分不同上下文情况下的语境

实验1：使用词的embedding简单构建Nearest Neighbour训练器

在词的消歧（word-sense disambiguation）任务中，只是简单的NN方法就可以得到很好的结果

Method	F1 score
Baseline (most frequent sense)	64.8
ELMo [20]	70.1
BERT	**71.1**
BERT (w/ probe)	**71.5**

这种语义表达是否包含一个子空间？

在BERT最后一层之上构建了一个线性映射矩阵B，通过对它进行一定的训练发现结果的达到/超过所有维度都使用的效果，则我们可以认为，肯定存在一个语义子空间，它就可以更好的表达部分的语义效果

m	Trained probe	Random probe
768 (full)	71.26	70.74
512	71.52	70.51
256	71.29	69.92
128	71.21	69.56
64	70.19	68.00
32	68.01	64.62
16	65.34	61.01

实验2：BERT是否真的编码了上下文信息？

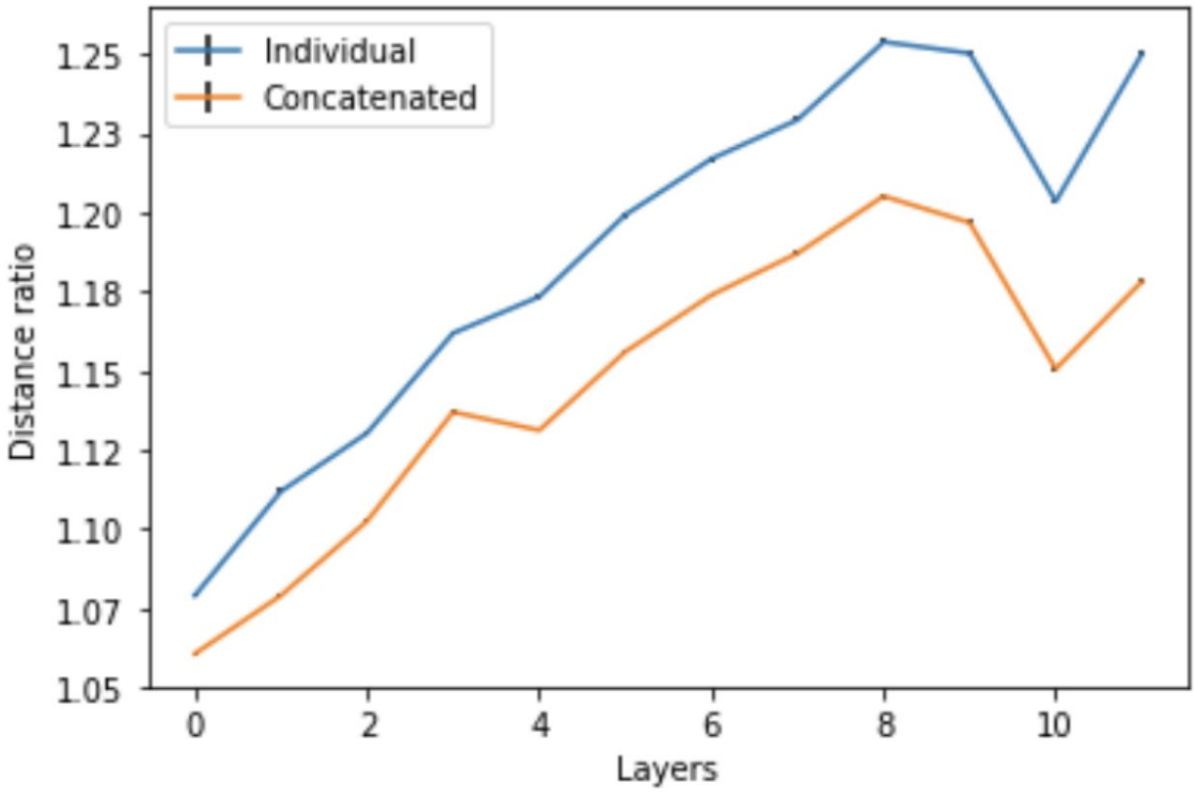
假设一个词有两个词义A和B，并且所有词义取平均值为A-center和B-center，那么从词义A中的词的表示，到A-center的距离应该比到B-center的更近（如果是欧基里德距离的话更近就是越小，cosine距离更近就是越从0接近1）

实验2.1：如果在包含词义A的句子中，随机加入一个句子，用and连接，取A中的词的embedding，对比到A-center和B-center的结果比例（比例越高就代表到A-center更接近1，到B-center更接近0，则分歧差距越大，是我们想得到的）

与

实验2.2：如果在包含词义A的句子中，加入一个词义B中的句子，用and连接，取A中的词的embedding，对比到A-center和B-center的结果比例

那么2.1的结果应该显著好于2.2的结果，从结果来看也是这样：



结论