[预训练语言模型专题] SpanBERT: 抽取式问答的利器

原创 管扬 朴素人工智能

来自专辑

预训练语言模型

本文为预训练语言模型专题的第16篇。

快速传送门

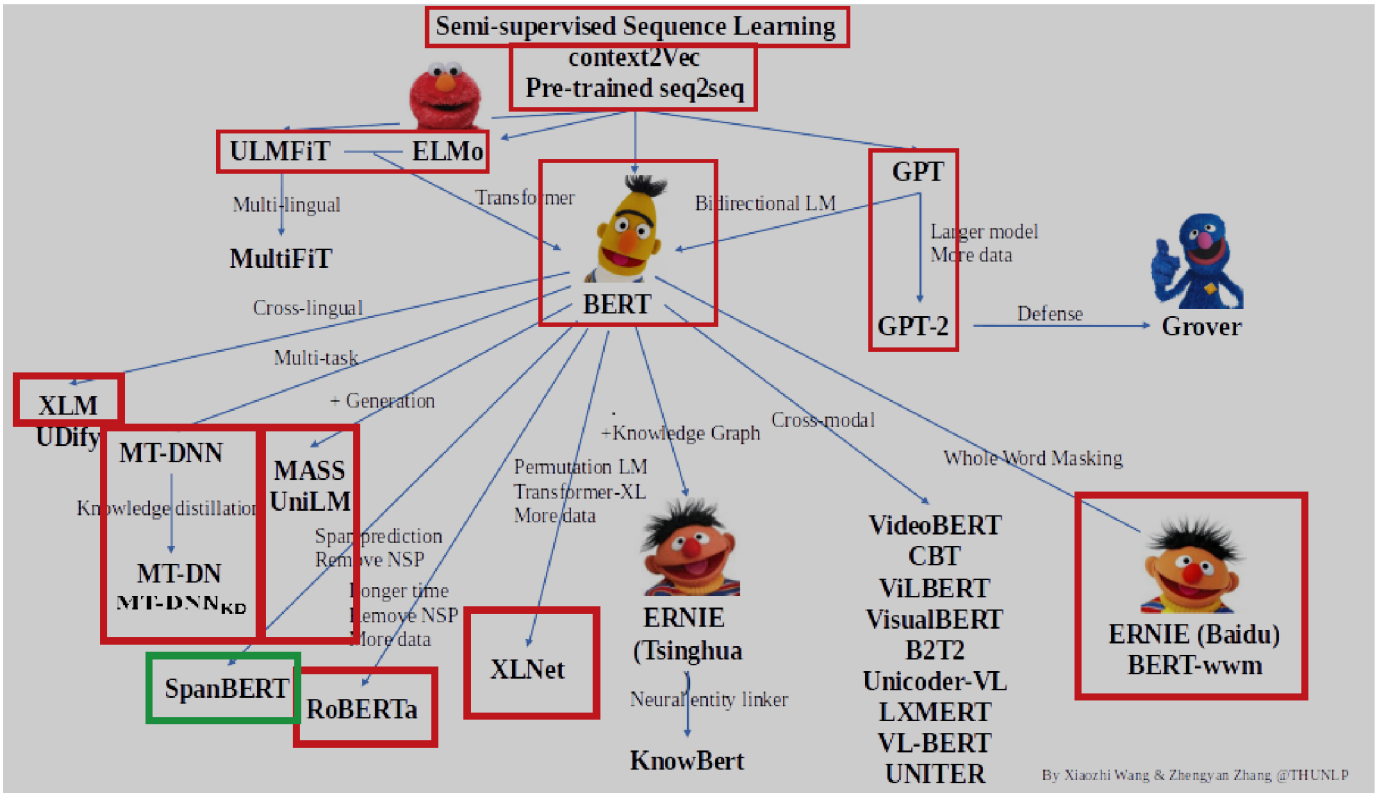
1-4:[萌芽时代]、[风起云涌]、[文本分类通用技巧] 、 [GPT家族]

5-8:[BERT来临]、[浅析BERT代码]、[ERNIE合集]、[MT-DNN(KD)]

9-12:[Transformer]、[Transformer-XL]、[UniLM]、[Mass-Bart]

13-14：[跨语种模型]、[XLNet]，[RoBERTa]

感谢清华大学自然语言处理实验室对**预训练语言模型**架构的梳理，我们将沿此脉络前行，探索预训练语言模型的前沿技术，红框中为已介绍的文章，绿框中为本期介绍的模型，欢迎大家留言讨论交流



SpanBERT: Improving Pre-training by Representing and Predicting Spans（2020）

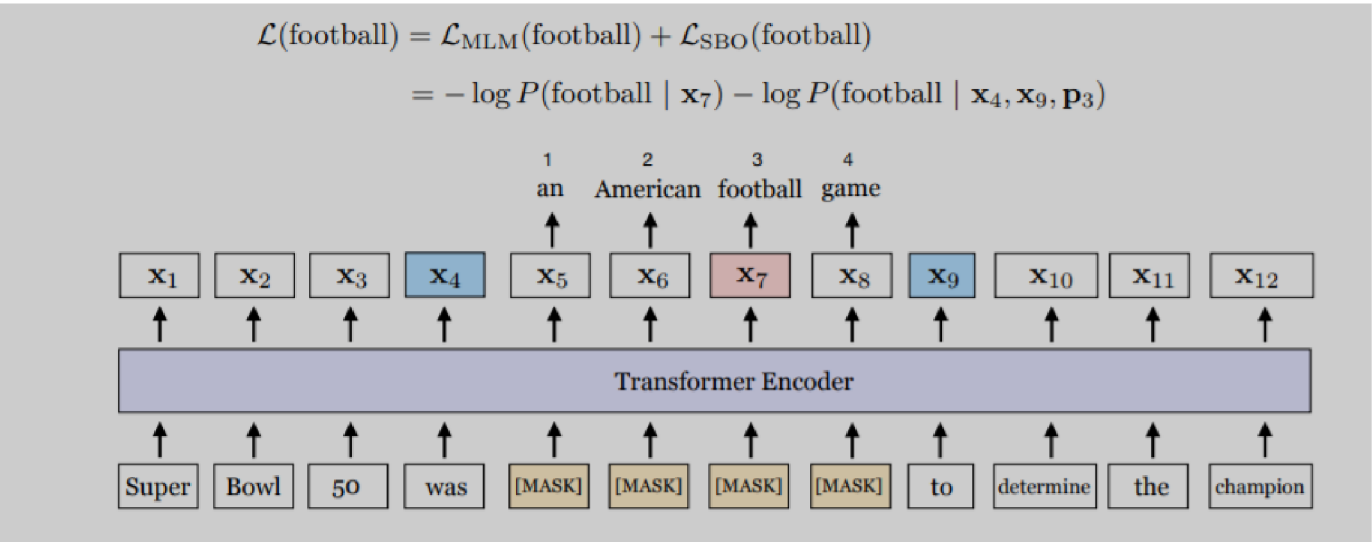
spanBERT是专门设计来更好地表示和预测文本的区间的，顾名思义它是BERT的一种扩展。相对于BERT，SpanBERT做了如下的改动。

1. 相比于BERT是对于随机的token进行masking，spanBERT会对连续的span进行masking。
2. spanBERT通过训练span的边界的表示，来对整个被mask的span来进行预测，而非其中的单个token。具体的说，文章引入了span-boundary objective(SPO)，来鼓励模型将span层面的信息存储在其边界的token表示上，以在finetune阶段获得更好的效果。

由于这两个机制，spanBERT在span选择的任务上明显地比BERT表现要好，包括问答匹配和指代消岐。同时在SQuAD和OntoNotes上都取得了state-of-the-art的效果。在SQuAD上，相比于BERT，SpanBERT降低了27%的误差，在各种抽取式问答benchmark(NewsQA, TriviaQA, SearchQA, HotpotQA, Natural Questions) 也观察到了类似的收益。

另外作者发现，在单个的片段上进行预训练，比两个一半长的片段预训练（带NSP）效果要好，所以，作者在预训练的时候选择在单个片段上进行，对比的BERT baseline也是在单个片段上预训练的BERT，效果是比原来的BERT baseline上是提高的。

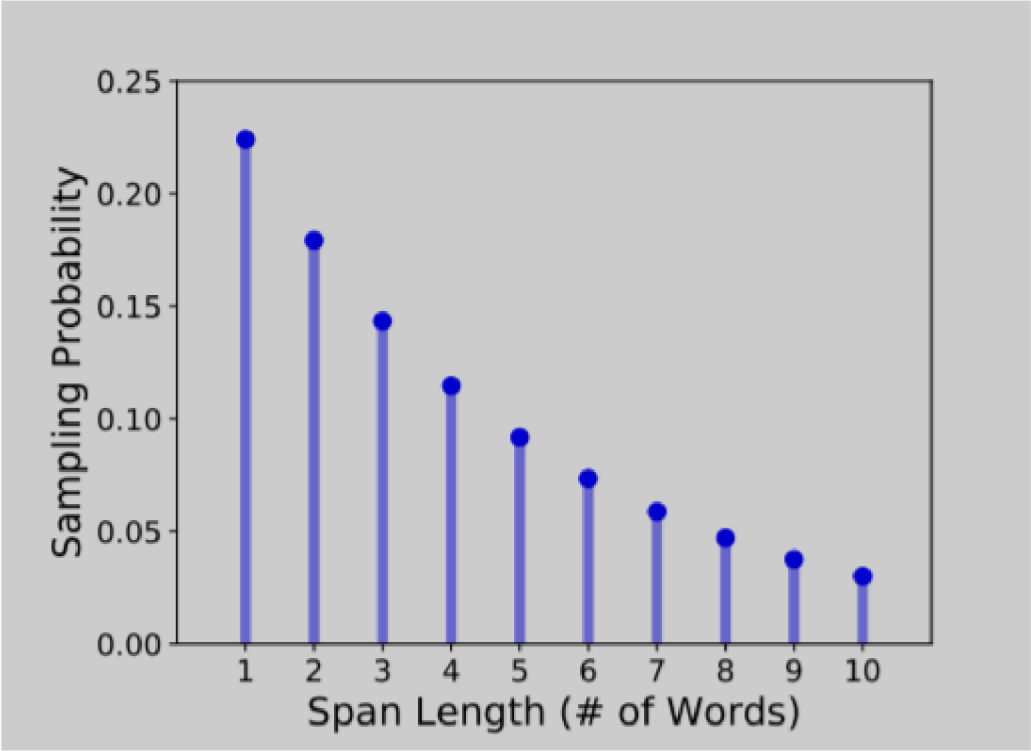
相比于其他approach通过增加训练的数据和加大模型的尺寸来提升性能。SpanBERT通过优化预训练模型的训练目标来提升效果。比如下图中展现了一个实例，被mask的是“an American football game"。SBO目标，就是用蓝色的x4(was)和x9(to)作为边界表示来预测被mask的每个token。从式子中可以看出MLM与SPO loss的区别，MLM利用x7的编码表示来预测football这个token，而SBO会利用，这个span(x5-x8)的边界 -> x4，x9和p3(从x4开始第三个token的位置embedding) 来进行预测。



Span Masking

对一个序列X = (x1, x2, . . . , xn)， 我们选取它的一个子集Y进行mask。通过不断地选取span，直到选够了X中15%的token。选取的方法是，首先通过几何分布选取span的长度L，会微倾向于选取较短的span，然后，半随机地选取span的起始位置。我们经常会去sample一个完整的

单词或短语，所以开始的位置一般选在单词的开始。选取长度时，官方的设置是 L ~Geo(0.2)，同时裁剪L使Lmax=10，于是span长度的分布如下，平均值为3.8。span masking，指的式对span中的每一个token都替换成[MASK]。



Span Boundary Objective(SBO)

SBO引入了目标函数来预测span内被mask的每一个token，只通过在span边界上被观察到的token的表示和其在span内的位置来进行预测。如果span内的token为 (xs, . . . , xe)，(s, e)是其起始的位置和结束的位置，那么我们就会使用其外部的边界表示 xs−1 和 xe+1，以及目标token的位置embedding即Pi−s+1来对其进行表示。一般会用两层的线性层加非线性激活来得到logits以计算其cross-entropy loss。

$$
\begin{aligned}
\mathbf{h}_0 &= [\mathbf{x}_{s-1}; \mathbf{x}_{e+1}; \mathbf{p}_{i-s+1}] \\
\mathbf{h}_1 &= \text{LayerNorm}(\text{GeLU}(\mathbf{W}_1\mathbf{h}_0)) \\
\mathbf{y}_i &= \text{LayerNorm}(\text{GeLU}(\mathbf{W}_2\mathbf{h}_1))
\end{aligned}
$$

Single-Sequence Training

BERT的预训练方法是每个样本包含两个片段序列 (XA, XB)，训练目标包括MLM(masked language model) 以及NSP (next sentence prediction)。训练中可以用NSP去预测XA，XB是否是连接的。作者发现，这种训练方法几乎总是比单纯只用MLM去训练单个片段序列样本更差。他认为有如下原因

- 用单个片段序列时，片段序列会更长，模型可以从中收益。
- 有时候，XA，XB来自不同文档，会将噪声加入到masked language model中

因此，在SpanBERT的训练中，只用了单个长512的序列片段。

总结一下这三点就是，SpanBERT通过

1. 通过几何分布抽样span长度，均匀分布寻找span的起始位置，并将其中的token全部mask。
2. 除了MLM目标外，新添了SPO目标进行优化，使抽取能力更强。
3. 使用了单个片段序列进行训练和预测，稳定提升。

模型细节

大部分设置和BERT相同，不同点有如下

- 相比于BERT对每个样本进行十种不同的mask，然后进行40个epoch的训练，SpanBERT会在每个epoch都对序列进行不同的masking。
- 移除了BERT的短句训练的策略，原来BERT会以0.1的概率用短文本进行训练，同时前90%的训练会以长128的序列来训练。
- 模型会跑2.4M步， AdamW的epsilon 使用 1e-8，因为这样收敛得更好。
- SBOloss中使用的位置编码为200维，对应离span最左侧的距离。
- 预训练使用了32张V100，花了15天完成。

接着作者在一些数据集上进行了一些比较。

	SQuAD 1.1		SQuAD 2.0	
	EM	F1	EM	F1
Human Perf.	82.3	91.2	86.8	89.4
Google BERT	84.3	91.3	80.0	83.3
Our BERT	86.5	92.6	82.8	85.9
Our BERT-1seq	87.5	93.3	83.8	86.6
SpanBERT	**88.8**	**94.6**	**85.7**	**88.7**

Table 1: Test results on SQuAD 1.1 and SQuAD 2.0.

	MUC			B^3			$CEAF_{\phi_4}$			
	P	R	F1	P	R	F1	P	R	F1	Avg. F1
Prev. SotA: (Lee et al., 2018)	81.4	79.5	80.4	72.2	69.5	70.8	68.2	67.1	67.6	73.0
Google BERT	84.9	82.5	83.7	76.7	74.2	75.4	74.6	70.1	72.3	77.1
Our BERT	85.1	83.5	84.3	77.3	75.5	76.4	75.0	71.9	73.9	78.3
Our BERT-1seq	85.5	84.1	84.8	77.8	76.7	77.2	75.3	73.5	74.4	78.8
SpanBERT	**85.8**	**84.8**	**85.3**	**78.3**	**77.9**	**78.1**	**76.4**	**74.2**	**75.3**	**79.6**

Table 3: Performance on the OntoNotes coreference resolution benchmark. The main evaluation is the average F1 of three metrics: MUC, B^3, and $CEAF_{\phi_4}$ on the test set.

结论

作者在17个任务上比较了SpanBERT和BERT，几乎所有任务上SpanBERT都更好。并且，SpanBERT由于其目标的机制，其更擅长抽取式问答。最后，使用单个片段序列的预测和训练好过，带有NSP的双片段样本进行训练，这让作者有点惊讶，因为BERT论文中的对比实验，显示NSP是有收益的。不过编者介绍的前篇推送中也讲到了，RoBERTa的研究中也认为NSP会降低模型的性能。

	SQuAD 2.0	NewsQA	TriviaQA	Coref	MNLI-m	QNLI	GLUE (Avg)
Span Masking (2seq) + NSP	85.4	73.0	78.8	76.4	87.0	93.3	83.4
Span Masking (1seq)	86.7	73.4	80.0	76.3	87.3	93.8	83.8
Span Masking (1seq) + SBO	**86.8**	**74.1**	**80.3**	**79.0**	**87.6**	**93.9**	**84.0**

未完待续

本期的论文就给大家分享到这里，感谢大家的阅读和支持，下期我们会给大家带来其他预训练语言模型的介绍，敬请大家期待！

欢迎关注朴素人工智能，这里有很多最新最热的论文阅读分享，有问题或建议可以在公众号下留言。

往期推荐

- [预训练语言模型专题] RoBERTa: 捍卫BERT的尊严
- [预训练语言模型专题] Transformer-XL 超长上下文注意力模型
- [预训练语言模型专题] XLNet：公平一战！多项任务效果超越BERT
- [预训练语言模型专题] Huggingface简介及BERT代码浅析
- [预训练语言模型专题] 结合HuggingFace代码浅析Transformer