



Bert和Transformer到底学到了什么

新浪微博 张俊林 2019.9.6





Bert和Transformer简介

探寻方法

Bert到底学到了什么



Bert及其影响

• 2018年10月提出

· 引发了NLP学术及工业界领域极大的反响

• 在各种应用中取得了各种突破性成果

• NLP领域里程碑的工作



Bert的应用效果

- QA问答领域
 - 应用Bert后,性能提升30%—70%
- 阅读理解领域
 - · 应用Bert后,性能提升30%—50%
- 信息检索领域
 - ·应用Bert后,短文档检索性能提升25%—106%
 - ·应用Bert后,长文档检索性能提升20%—30%



Bert的应用效果

- 对话机器人领域
 - 应用Bert后,性能提升5%—40%
- 文本摘要领域
 - ·应用Bert后,性能提升10%左右
- 其它应用领域
 - 中文分词
 - 文本分类
 - 文本生成......



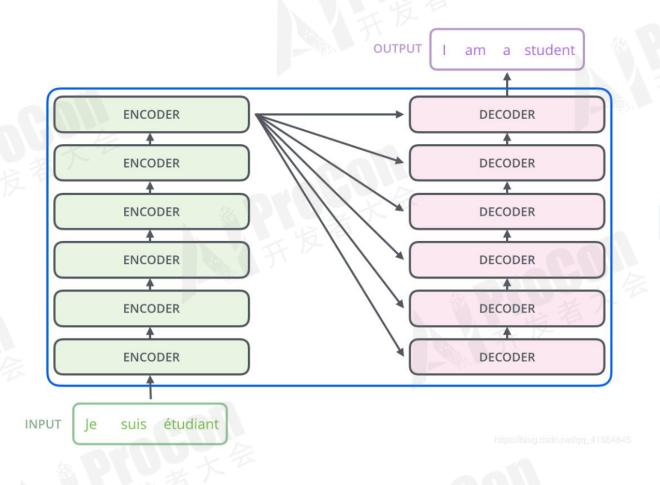
Bert的模型改进方向

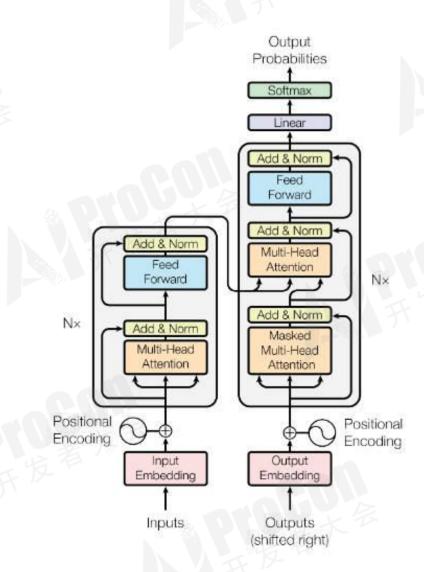
- 文本生成模型
- 结构化知识引入
- 多模态融合
- 更大更高质量的训练数据
- 更合适的训练目标和训练方法
- 多语言融合



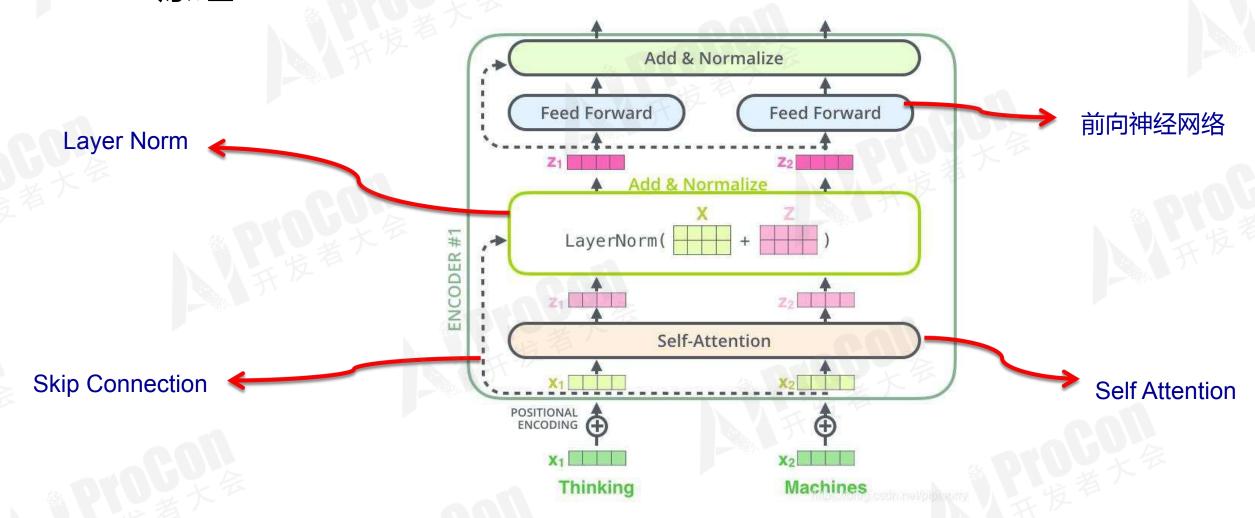


Bert原理—Transformer





Bert原理—Transformer Block

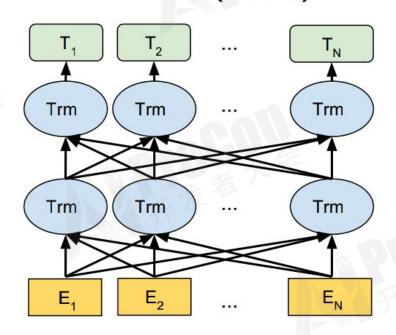






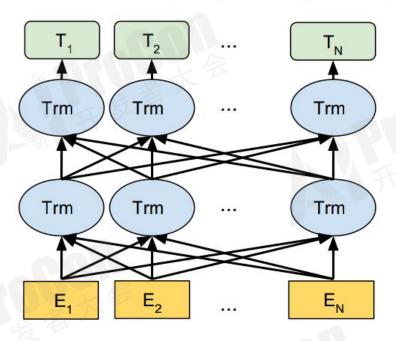
Bert原理—两阶段过程

BERT (Ours)



初始化参数

BERT (Ours)



预训练 (自监督学习语言知识)

Fine-Tuning (目标任务有监督调优)

我们的问题

为什么Bert效果这么好?

Bert里的Transformer到底学到了什么?





Bert和Transformer简介

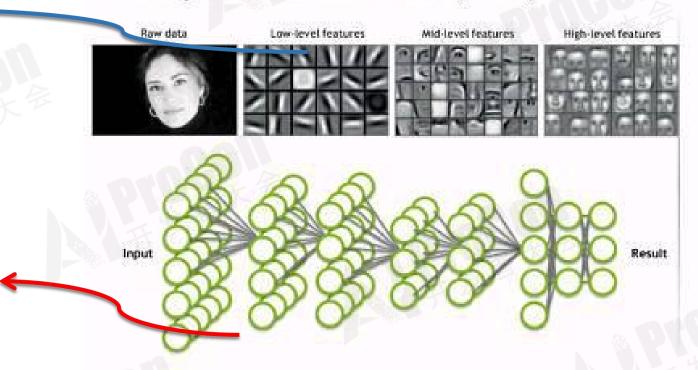
探寻方法

Bert到底学到了什么



DNN是个著名的黑盒系统

特征可视化: 典型的破解黑盒的方法 Deep neural network (DNN)



每个神经元到底学到了什么? 黑盒系统!

Application components:

Task objective

e.g. Identify face Training data

10-100M images

Network architecture

-10 layers

1B parameters Learning algorithm

- -30 Exaflops
- -30 GPU days



Bert和Transformer的探寻方法—可视化(2D t-SEN)

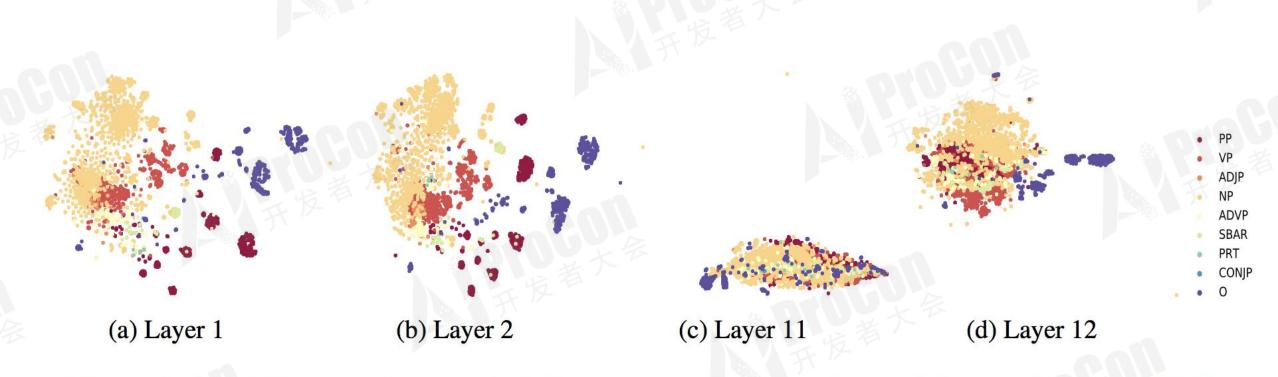


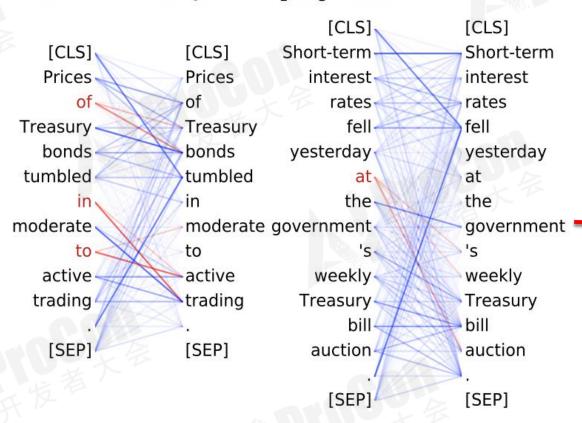
Figure 1: 2D t-SNE plot of span embeddings computed from the first and last two layers of BERT.



Bert和Transformer的探寻方法—可视化(Attention图)

Head 9-6

- Prepositions attend to their objects
- 76.3% accuracy at the pobj relation

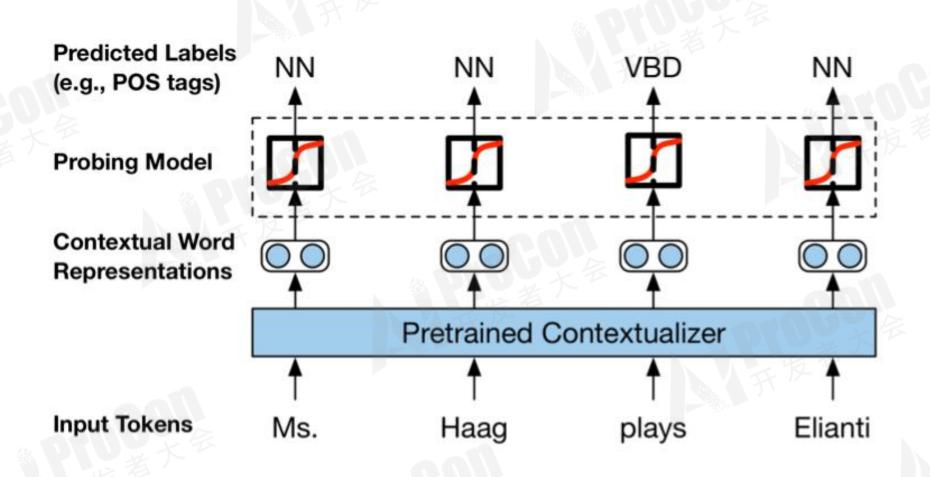


Attention图:通过可视化地表现句中单词之间的Attention强度,可以看出学到了什么知识

探寻介词与其它单词的关系

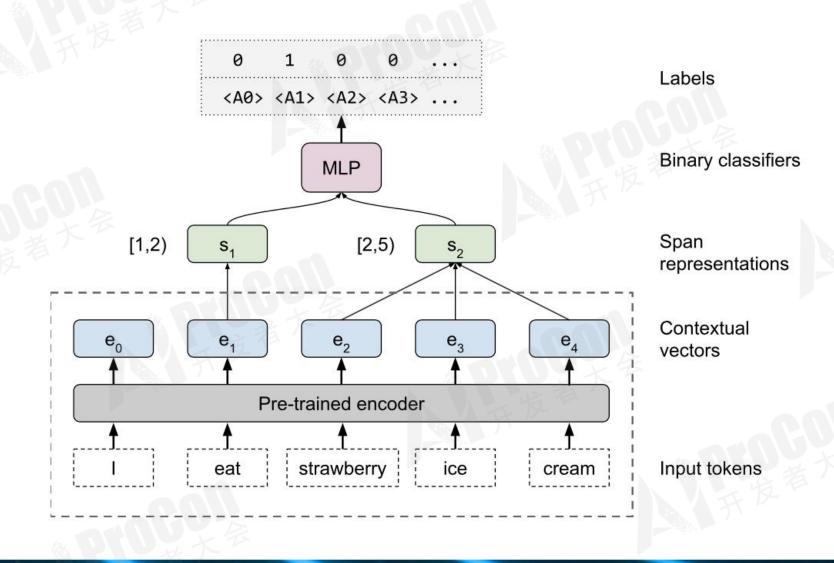


Bert和Transformer的探寻方法—Probing Classifer





Bert和Transformer的探寻方法—Edge Probing Classifer





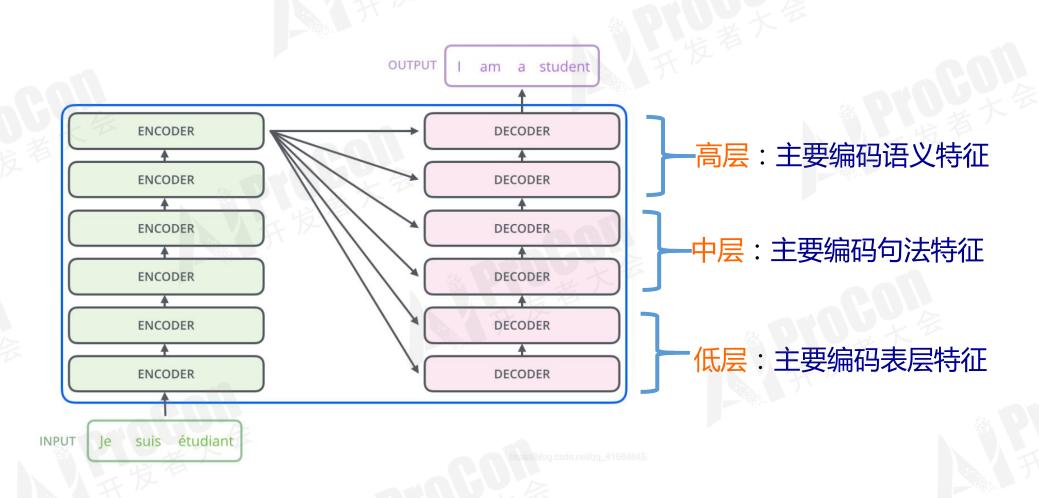
CSDN

Bert和Transformer简介

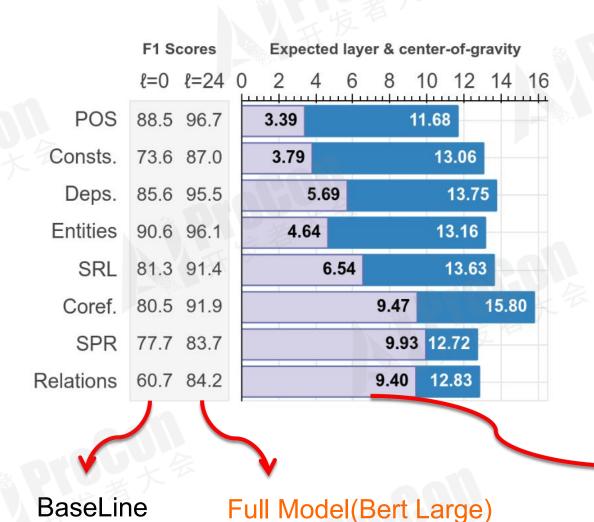
探寻方法

Bert到底学到了什么









高层:主要编码语义特征

中层:主要编码句法特征

低层:主要编码表层特征

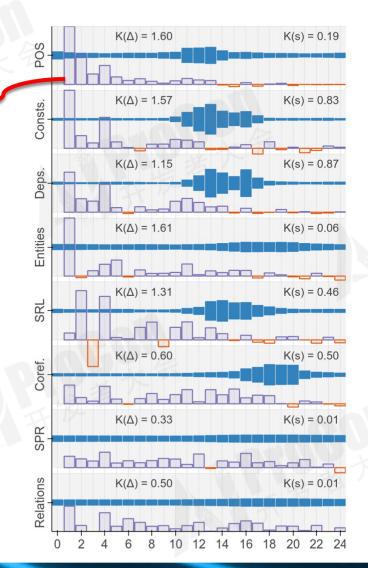
数值越大代表Transformer越高层级的作用越大



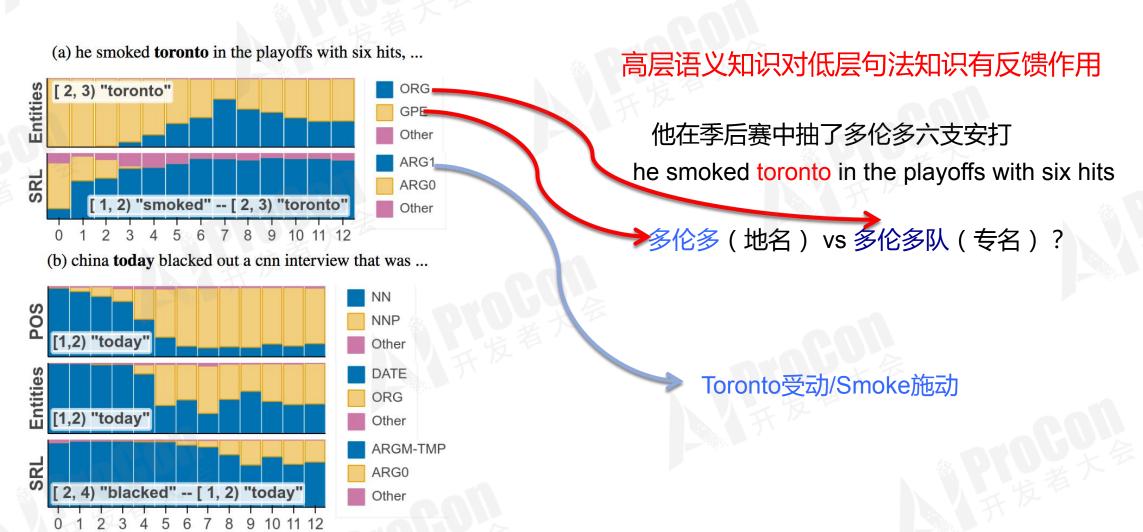
数值越大代表作用越大

句法:中低层,且具备Layer局部性

语义:不具备Layer局部性

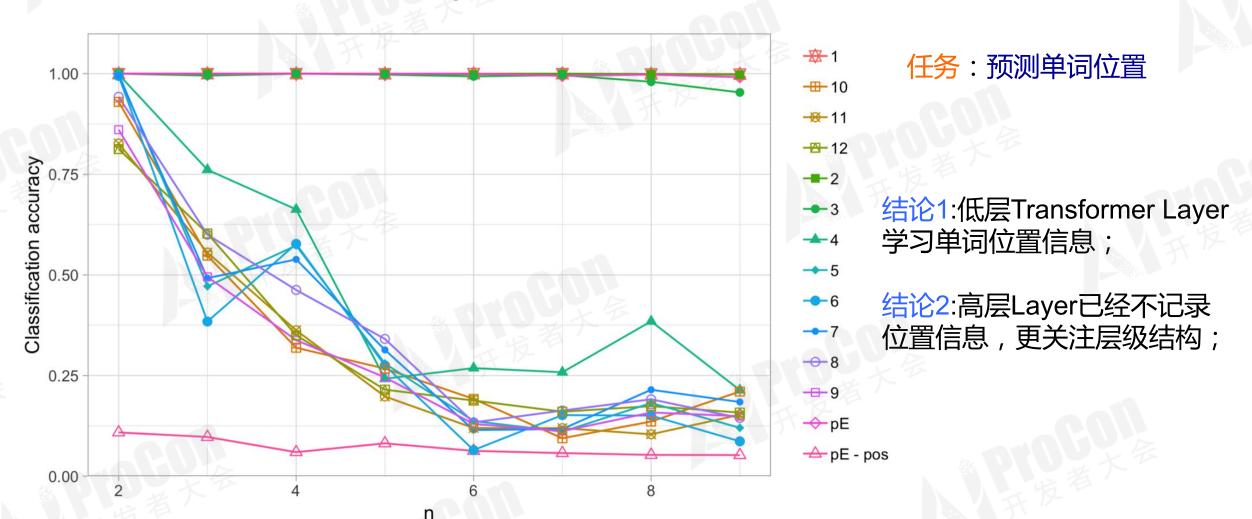








低层Transformer Layer-对单词位置信息的编码





低层Transformer Layer-对短语信息的编码

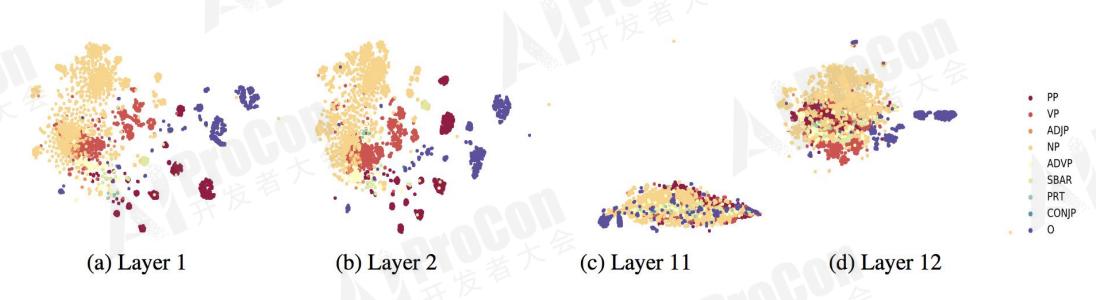
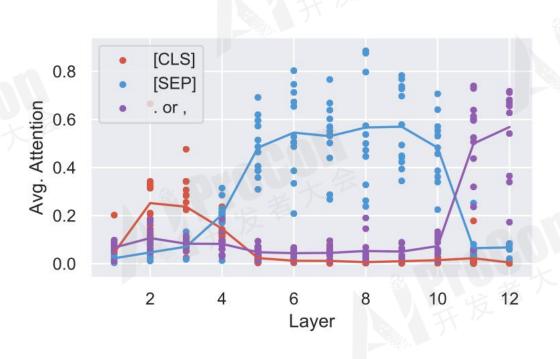


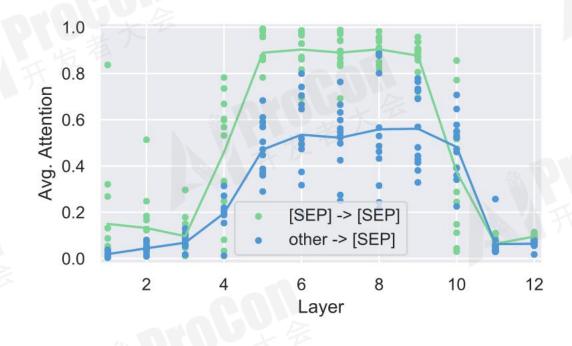
Figure 1: 2D t-SNE plot of span embeddings computed from the first and last two layers of BERT.

任务:短语聚类 结论:低层Layer能很好区分短语聚类,高层混杂在一起



低层Transformer Layer-对特殊符号的编码



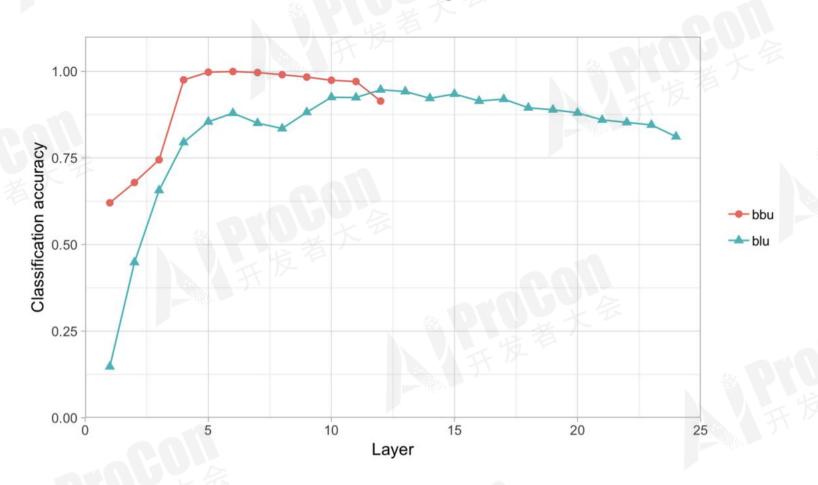


低层级Head: 注重[CLS]起始符号; 中间层级Head:注重句子分割符号[SEP] 高层级的Head:注重句号和逗号标点符号;

[SEP]也特别注重[SEP]起始符号;



中层Transformer Layer-对句法信息的编码



句法预测任务

◆ 中间层编码了更多的句法信息



中层Transformer Layer-对句法信息的编码

-	0.000						
Relation	Head	Accuracy	Baseline				
All	7-6	34.5	26.3 (1)				
prep	7-4	66.7	61.8 (-1)				
pobj	9-6	76.3	34.6 (-2)				
det	8-11	94.3	51.7 (1)				
nn	4-10	70.4	70.2(1)				
nsubj	8-2	58.5	45.5 (1)				
amod	4-10	75.6	68.3 (1)				
dobj	8-10	86.8	40.0 (-2)				
advmod	7-6	48.8	40.2 (1)				
aux	4-10	81.1	71.5 (1)				
poss	7-6	80.5	47.7 (1)				
auxpass	4-10	82.5	40.5 (1)				
ccomp	8-1	48.8	12.4 (-2)				
mark	8-2	50.7	14.5 (2)				
prt	6-7	99.1	91.4 (-1)				

句法任务:

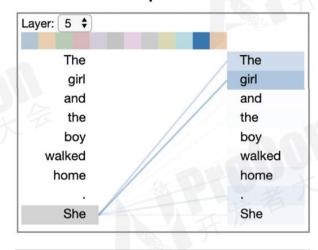
语料: Wall Street Journal portion of the Penn Treebank

- ◆ 探测Attention Head的句法能力
- ◆ 没有Attention Head具备全面的句法特征表达能力
- ◆ Attention Head具备句法特征能力特异性
- ◆ 句法能力集中在Transformer中间层
- ◆ Bert捕获的句法特征有些和人定义的标准不同

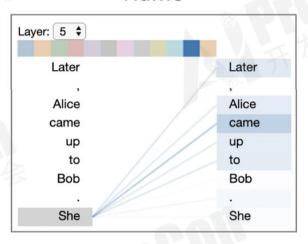


高层Transformer Layer-指代消解

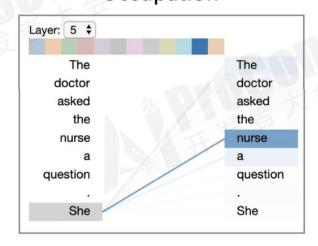
Gender-specific term



Name

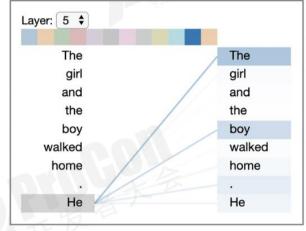


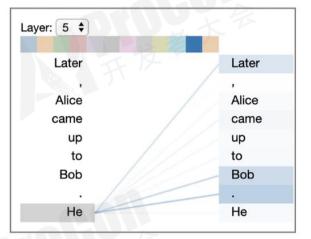
Occupation

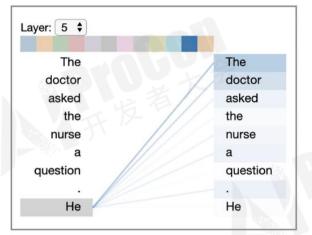


Bert学习指代消解

类特征: He vs She



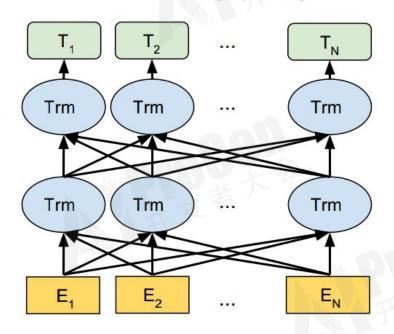






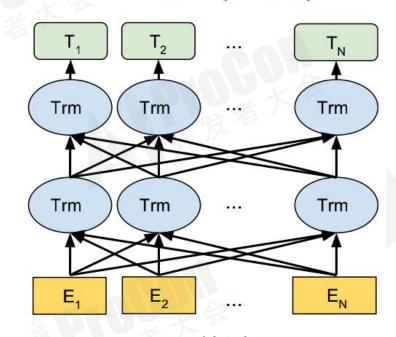
Bert的预训练过程比非预训练过程多学了什么

BERT (Ours)



有预训练过程

BERT (Ours)



无预训练过程

做监督任务Finetuning后,有什么差异?

表现:未训练Bert在句子长度预测任务好于预训练过程Bert

结论:预训练模型牺牲部分表层特征表达能力,获得更多更丰富的复杂特征



预训练模型 (Bert/ELMO/GPT)学到了些什么

		CoVe			ELMo			GPT		
	Lex.	Full	Abs. Δ	Lex.	Full	Abs. Δ	Lex.	cat	mix	
Part-of-Speech	85.7	94.0	8.4	90.4	96.7	6.3	88.2	94.9	95.0	
Constituents	56.1	81.6	25.4	69.1	84.6	15.4	65.1	81.3	84.6	Cove/ELMO/GPT
Dependencies	75.0	83.6	8.6	80.4	93.9	13.6	77.7	92.1	94.1	相对基准方法,多
Entities	88.4	90.3	1.9	92.0	95.6	3.5	88.6	92.9	92.5	编码了更多句法信息
SRL (all)	59.7	80.4	20.7	74.1	90.1	16.0	67.7	86.0	89.7	
Core roles	56.2	81.0	24.7	73.6	92.6	19.0	65.1	88.0	92.0	语义相对少些;
Non-core roles	67.7	78.8	11.1	75.4	84.1	8.8	73.9	81.3	84.1	
OntoNotes coref.	72.9	79.2	6.3	75.3	84.0	8.7	71.8	83.6	86.3	
SPR1	73.7	77.1	3.4	80.1	84.8	4.7	79.2	83.5	83.1	
SPR2	76.6	80.2	3.6	82.1	83.1	1.0	82.2	83.8	83.5	
Winograd coref.	52.1	54.3	2.2	54.3	53.5	-0.8	51.7	52.6	53.8	
Rel. (SemEval)	51.0	60.6	9.6	55.7	77.8	22.1	58.2	81.3	81.0	
Macro Average	69.1	78.1	9.0	75.4	84.4	9.1	73.0	83.2	84.4	



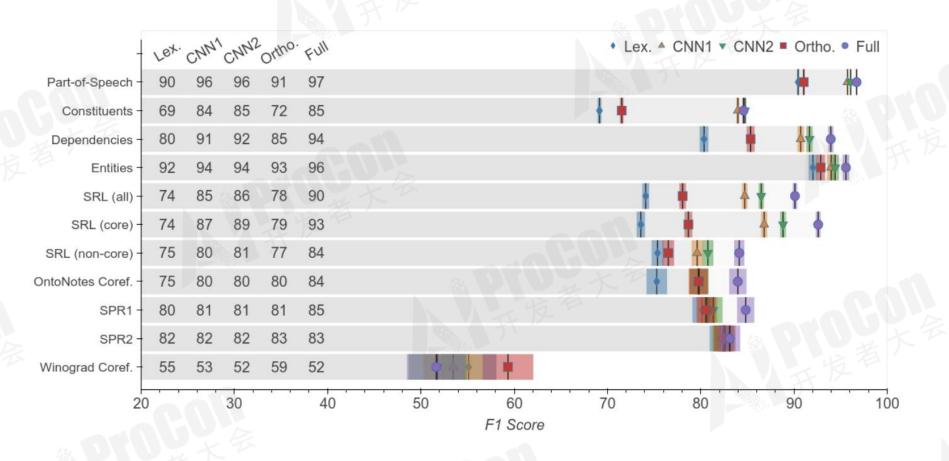
预训练模型(Bert/ELMO/GPT)学到了些什么

BERT-base			BERT-large					
F1 Score		Abs. Δ	F1 Score			Abs. Δ		
Lex.	cat	mix	ELMo	Lex.	cat	mix	(base)	ELMo
88.4	97.0	96.7	0.0	88.1	96.5	96.9	0.2	0.2
68.4	83.7	86.7	2.1	69.0	80.1	87.0	0.4	2.5
80.1	93.0	95.1	1.1	80.2	91.5	95.4	0.3	1.4
90.9	96.1	96.2	0.6	91.8	96.2	96.5	0.3	0.9
75.4	89.4	91.3	1.2	76.5	88.2	92.3	1.0	2.2
74.9	91.4	93.6	1.0	76.3	89.9	94.6	1.0	2.0
76.4	84.7	85.9	1.8	76.9	84.1	86.9	1.0	2.8
74.9	88.7	90.2	6.3	75.7	89.6	91.4	1.2	7.4
79.2	84.7	86.1	1.3	79.6	85.1	85.8	-0.3	1.0
81.7	83.0	83.8	0.7	81.6	83.2	84.1	0.3	1.0
54.3	53.6	54.9	1.4	53.0	53.8	61.4	6.5	7.8
57.4	78.3	82.0	4.2	56.2	77.6	82.4	0.5	4.6
75.1	84.8	86.3	1.9	75.2	84.2	87.3	1.0	2.9
	88.4 68.4 80.1 90.9 75.4 74.9 76.4 74.9 79.2 81.7 54.3 57.4	F1 Score Lex. cat 88.4 97.0 68.4 83.7 80.1 93.0 90.9 96.1 75.4 89.4 74.9 91.4 76.4 84.7 74.9 88.7 79.2 84.7 81.7 83.0 54.3 53.6 57.4 78.3	F1 Score Lex. cat mix 88.4 97.0 96.7 68.4 83.7 86.7 80.1 93.0 95.1 90.9 96.1 96.2 75.4 89.4 91.3 74.9 91.4 93.6 76.4 84.7 85.9 74.9 88.7 90.2 79.2 84.7 86.1 81.7 83.0 83.8 54.3 53.6 54.9 57.4 78.3 82.0	F1 Score Abs. Δ Lex. cat mix ELMo 88.4 97.0 96.7 0.0 68.4 83.7 86.7 2.1 80.1 93.0 95.1 1.1 90.9 96.1 96.2 0.6 75.4 89.4 91.3 1.2 74.9 91.4 93.6 1.0 76.4 84.7 85.9 1.8 74.9 88.7 90.2 6.3 79.2 84.7 86.1 1.3 81.7 83.0 83.8 0.7 54.3 53.6 54.9 1.4 57.4 78.3 82.0 4.2	F1 Score Abs. Δ I Lex. cat mix ELMo Lex. 88.4 97.0 96.7 0.0 88.1 68.4 83.7 86.7 2.1 69.0 80.1 93.0 95.1 1.1 80.2 90.9 96.1 96.2 0.6 91.8 75.4 89.4 91.3 1.2 76.5 74.9 91.4 93.6 1.0 76.3 76.4 84.7 85.9 1.8 76.9 74.9 88.7 90.2 6.3 75.7 79.2 84.7 86.1 1.3 79.6 81.7 83.0 83.8 0.7 81.6 54.3 53.6 54.9 1.4 53.0 57.4 78.3 82.0 4.2 56.2	F1 Score Abs. Δ F1 Score Lex. cat mix ELMo Lex. cat 88.4 97.0 96.7 0.0 88.1 96.5 68.4 83.7 86.7 2.1 69.0 80.1 80.1 93.0 95.1 1.1 80.2 91.5 90.9 96.1 96.2 0.6 91.8 96.2 75.4 89.4 91.3 1.2 76.5 88.2 74.9 91.4 93.6 1.0 76.3 89.9 76.4 84.7 85.9 1.8 76.9 84.1 74.9 88.7 90.2 6.3 75.7 89.6 79.2 84.7 86.1 1.3 79.6 85.1 81.7 83.0 83.8 0.7 81.6 83.2 54.3 53.6 54.9 1.4 53.0 53.8 57.4 78.3 82.0 4.2 56.2 77.6	Lex. cat mix ELMo Lex. cat mix 88.4 97.0 96.7 0.0 88.1 96.5 96.9 68.4 83.7 86.7 2.1 69.0 80.1 87.0 80.1 93.0 95.1 1.1 80.2 91.5 95.4 90.9 96.1 96.2 0.6 91.8 96.2 96.5 75.4 89.4 91.3 1.2 76.5 88.2 92.3 74.9 91.4 93.6 1.0 76.3 89.9 94.6 76.4 84.7 85.9 1.8 76.9 84.1 86.9 74.9 88.7 90.2 6.3 75.7 89.6 91.4 79.2 84.7 86.1 1.3 79.6 85.1 85.8 81.7 83.0 83.8 0.7 81.6 83.2 84.1 54.3 53.6 54.9 1.4 53.0 53.8 61	F1 Score Abs. Δ F1 Score Abs Lex. cat mix ELMo Lex. cat mix (base) 88.4 97.0 96.7 0.0 88.1 96.5 96.9 0.2 68.4 83.7 86.7 2.1 69.0 80.1 87.0 0.4 80.1 93.0 95.1 1.1 80.2 91.5 95.4 0.3 90.9 96.1 96.2 0.6 91.8 96.2 96.5 0.3 75.4 89.4 91.3 1.2 76.5 88.2 92.3 1.0 74.9 91.4 93.6 1.0 76.3 89.9 94.6 1.0 76.4 84.7 85.9 1.8 76.9 84.1 86.9 1.0 74.9 88.7 90.2 6.3 75.7 89.6 91.4 1.2 79.2 84.7 86.1 1.3 79.6 85.1 85.8 -0.3 </td

Bert vs. ELMO 更深的结构有利于 复杂语义任务;



预训练模型(Bert/ELMO/GPT)学到了些什么



Lex:基准

CNN1:左右窗口大小1

CNN2:左右窗口大小2

Ortho:未预训练ELMO

Full: 预训练ELMO

ELMO:捕获了更长的

上下文

