RoBERTa

RoBERTa: A Robustly Optimized BERT Pretraining Approach. Yinhan Liu et al. 2019. [PDF]

---

1 总览

BERT提出之后，涌现出了XLNet、ALICE、XLM、MT-DNN等后续工作，成绩都在BERT的基础上得到了进一步的提升。然而本文认为那是因为**BERT其实根本没有得到充分的训练**（否则成绩和这些后来居上者相当），为此本文从模型设计选择(design choice)、训练策略、语料等方面入手，重新对BERT进行了预训练，得到RoBERTa。实验结果表明RoBERTa在GLUE、RACE和SQuAD都达到了SOTA.

2 主要工作

RoBERTa在每个部分都做了一点点修改，除了Text Encoding外其他的小修改都得到了轻微的提升，最后它将所有的修改合在一起得到最佳模型。具体修改包括：

- 修改了超参数：将adam的 β_2 参数从0.999改为0.98
- 加入了混合精度
- 加大batch size：**从BERT的batch_size从256改为2K甚至8K，训练步数从1M降到500K**
- 在更长的序列上训练，修改输入格式：FULL-SENTENCES+移除NSP任务
- **将BERT静态遮掩改为动态遮掩**
- 增加新的预训练数据集CC-NEWS，语料从16G文本到160G文本
- Text Encoding：采用更大的byte-level的BPE词典

接下来，将挑选出几个点进行详细阐述。

2.1 byte-level text encoding

BPE（Byte-Pair Encoding）是Sennrich在2016年提出的文本编码方法 ，BPE将字词统一成子词单元（subword units），通过在训练语料上的统计分析抽取得到。BPE的词汇表大小一般从1w到10w个子词不等，而其中unicode字符的占比很大。

Radford在GPT2 里提出了一种更巧妙的BPE实现版本，该方法使用bytes（字节）作为基础的子词单元，这样便把词汇表的大小控制到了5w。它可以**在不需要引入任何未知字符前提下对任意文本进行编码**。

BERT原始版本使用一个字级（character-level）的BPE词汇表，大小是3w，是用启发式分词规则对输入进行预处理学习得到的。

之前的实验结果表明，这些文本编码的实验性能区别不大，可能Radford BPE Encoding在某些任务上的终端性能略微差点，但是RoBerta作者坚信通用的编码模式比性能上的轻微损失更重要，所以在实验中采用了byte-level text encoding。

3 实验结果

在SQuAD、MNLI-m和SST-2上的实验结果：

Model	SQuAD 1.1/2.0	MNLI-m	SST-2	RACE
Our reimplementation (with NSP loss):				
SEGMENT-PAIR	90.4/78.7	84.0	92.9	64.2
SENTENCE-PAIR	88.7/76.2	82.9	92.1	63.0
Our reimplementation (without NSP loss):				
FULL-SENTENCES	90.4/79.1	84.7	92.5	64.8
DOC-SENTENCES	90.6/79.7	84.7	92.7	65.6
$\text{BERT}_{\text{BASE}}$	88.5/76.3	84.3	92.8	64.3
$\text{XLNet}_{\text{BASE}}$ (K = 7)	–/81.3	85.8	92.7	66.1
$\text{XLNet}_{\text{BASE}}$ (K = 6)	–/81.0	85.6	93.4	66.7

RoBERTa对比BERT有明显的提升，但是和XLNet差距不大。

在GLUE上的结果：

Model	data	bsz	steps	SQuAD (v1.1/2.0)	MNLI-m	SST-2
RoBERTa						
with BOOKS + WIKI	16GB	8K	100K	93.6/87.3	89.0	95.3
+ additional data (§3.2)	160GB	8K	100K	94.0/87.7	89.3	95.6
+ pretrain longer	160GB	8K	300K	94.4/88.7	90.0	96.1
+ pretrain even longer	160GB	8K	500K	**94.6/89.4**	**90.2**	**96.4**
$\text{BERT}_{\text{LARGE}}$						
with BOOKS + WIKI	13GB	256	1M	90.9/81.8	86.6	93.7
$\text{XLNet}_{\text{LARGE}}$						
with BOOKS + WIKI	13GB	256	1M	94.0/87.8	88.4	94.4
+ additional data	126GB	2K	500K	94.5/88.8	89.8	95.6

对于单任务单模型，RoBERTa九个任务均达到SOTA；

在SQuAD上的结果：

	MNLI	QNLI	QQP	RTE	SST	MRPC	CoLA	STS	WNLI	Avg
Single-task single models on dev										
$\text{BERT}_{\text{LARGE}}$	86.6/-	92.3	91.3	70.4	93.2	88.0	60.6	90.0	-	-
$\text{XLNet}_{\text{LARGE}}$	89.8/-	93.9	91.8	83.8	95.6	89.2	63.6	91.8	-	-
RoBERTa	**90.2/90.2**	**94.7**	**92.2**	**86.6**	**96.4**	**90.9**	**68.0**	**92.4**	**91.3**	-
Ensembles on test (from leaderboard as of July 25, 2019)										
ALICE	88.2/87.9	95.7	**90.7**	83.5	95.2	92.6	**68.6**	91.1	80.8	86.3
MT-DNN	87.9/87.4	96.0	89.9	86.3	96.5	92.7	68.4	91.1	89.0	87.6
XLNet	90.2/90.8	98.6	90.3	86.3	**96.8**	**93.0**	67.8	91.6	**90.4**	88.4

XLNet	90.2/89.8	98.6	90.3	86.3	**96.8**	**93.0**	67.8	91.6	**90.4**	88.4
RoBERTa	**90.8/90.2**	**98.9**	90.2	**88.2**	96.7	92.3	67.8	**92.2**	89.0	**88.5**

RoBERTa的成绩还可以。

在RACE上的对比结果：

Model	Accuracy	Middle	High
Single models on test (as of July 25, 2019)			
$\text{BERT}_{\text{LARGE}}$	72.0	76.6	70.1
$\text{XLNet}_{\text{LARGE}}$	81.7	85.4	80.2
RoBERTa	**83.2**	**86.5**	**81.3**

从实验结果上看，RoBERTa均达到了SOTA.

4 总结

RoBERTa其实本质上只是一个调参达到最优的BERT，和XLNet不相上下。

带给我们的意义就是：**RoBERTa再一次证明BERT才是众多预训练模型中的首选和扛鼎之作，也是那个真正引起质变的模型**。