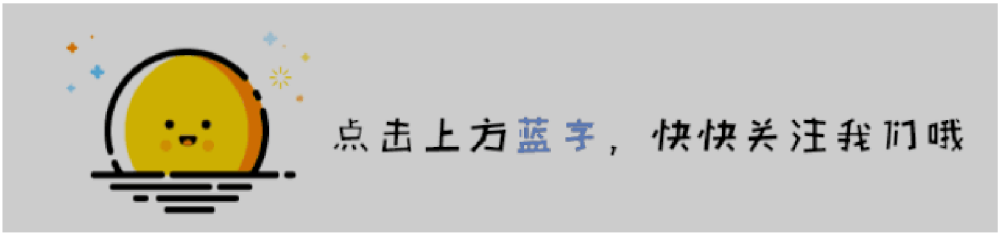
性能媲美BERT却只有其1/10参数量？ | 近期最火模型ELECTRA解析

原创 Lulu 朴素人工智能

来自专辑

预训练语言模型



快速传送门

论文链接

https://openreview.net/forum?id=r1xMH1BtvB

Google开源地址

https://github.com/google-research/electra

中文ELECTRA开源地址

https://github.com/ymcui/Chinese-ELECTRA

要说近期NLP领域最吸引眼球的模型之一，恐怕非ELECTRA莫属了。Electra全称Efficiently Learning an Encoder that Classifies Token Replacements Accurately，由斯坦福Manning大神与谷歌联合发布。

要问这个模型厉害在哪，先亮个结果图镇一下楼（右图是左图虚线框内的放大版）

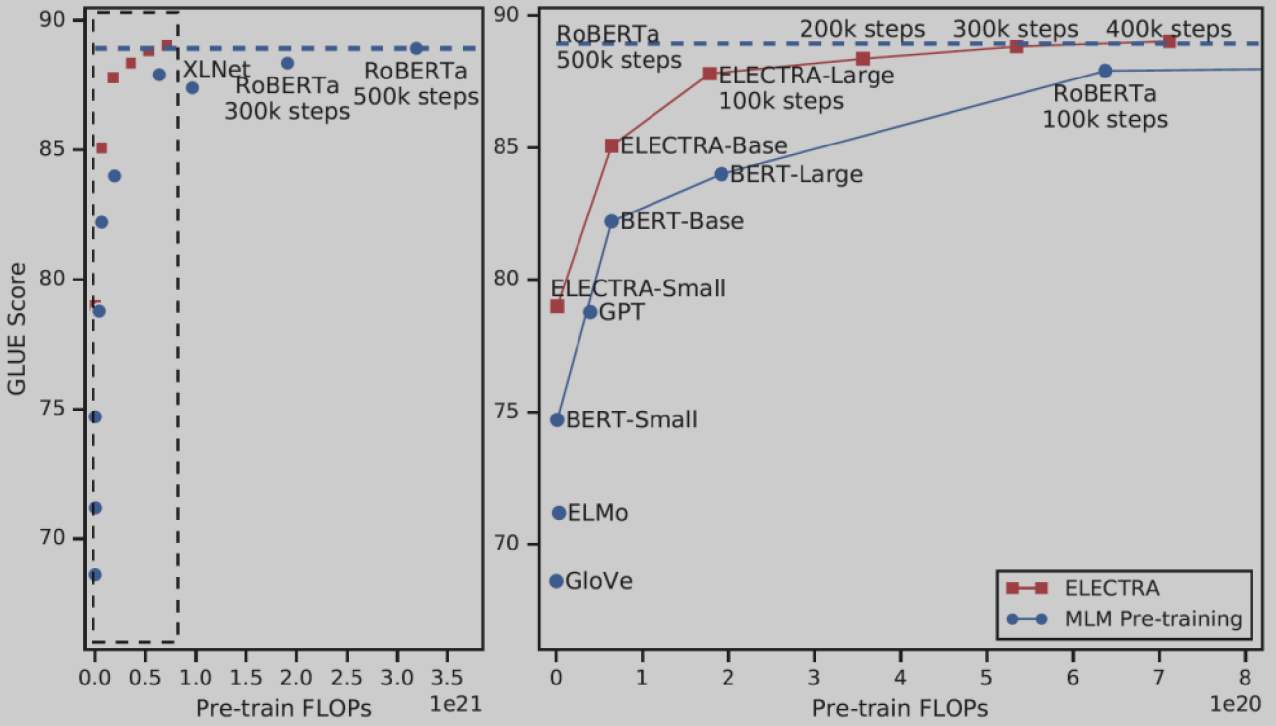


Figure 1: ELECTRA consistently outperforms MLM pre-training given the same compute budget. The right figure is a zoomed-in view of the dashed box.

从上图右边可以看到，在同等量级（同样的数据规模，模型大小）下，ELECTRA一直优于BERT，Roberta等模型。另外，从上图左边可以看出，ELECTRA模型能够仅用1/4的计算量就达到Roberta的效果，并且随着模型量级的增加，ELECTRA的性能还能进一步提升。

1 背景

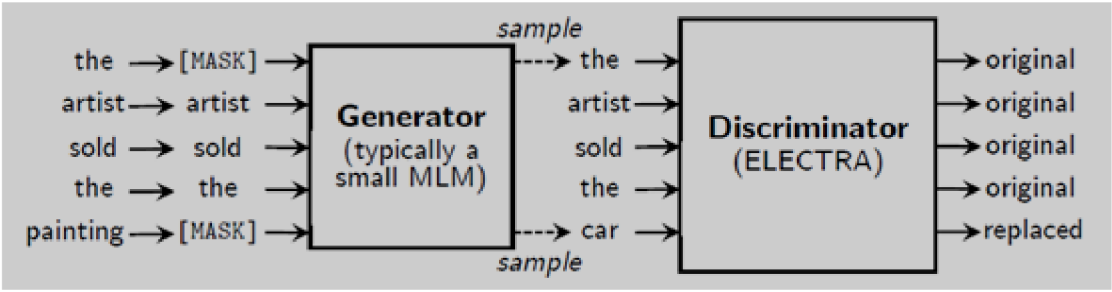
自Google于2018年提出BERT以来，NLP领域涌现出了一大堆基于BERT的改进，包括关注masked token之间的依赖关系，使用动态masking等等。但是，越来越优秀的性能往往伴随着越来越长的训练时间，越来越大的数据规模以及越来越高的硬件要求。
在此背景下，ELECTRA另辟蹊径，抛弃传统的MLM（masked language model）任务（详情见BERT介绍），提出了全新的replaced token detection任务，使得模型在保持性能的前提下大大降低了模型参数量，提高了模型的运算速度。

2 方法

语言模型预训练任务replaced token detection包含两个步骤：

1. mask一些input tokens，然后使用一个生成式网络预测被mask的token
2. 训练一个判别式网络来判断每一个token是否“虚假”

下图清晰展示了replaced token detection预训练的整体架构：



Replaced token detection的优化目标函数为：

$$\min_{\theta_G,\theta_D} \sum_{\boldsymbol{x}\in\mathcal{X}} \mathcal{L}_{\mathrm{MLM}}(\boldsymbol{x},\theta_G) + \lambda\mathcal{L}_{\mathrm{Disc}}(\boldsymbol{x},\theta_D)$$

加号左边代表MLM的loss，右边代表discriminator的loss。在预训练的时候，generator和discrimiator同时训练。

Generator网络其实就是一个小型MLM，discriminator就是论文所说的ELECTRA模型。在预训练完成之后，generator被丢弃，而判别式网络会被保留用来做下游任务的基础模型。

论文指出，replaced token detection之所以比MLM优秀，主要原因在于：

1. MLM仅从15%被mask的tokens学习，而replaced token detection要辨别inputs的所有tokens的“真假”，因而可以学习到所有tokens；
2. MLM任务中[mask]的存在导致了预训练和fine-tuning数据分布不匹配的问题，而这个问题在ELECTRA模型中不存在。尽管MLM做了一些措施来弥补，但是并没有完全解决这个问题。

尽管与GAN的训练目标很像，replaced token detection任务还是与GAN存在一些关键性差异：

1. 如果generator正确还原了一些token，这些正确还原的token在discriminator部分会算作真实token。而在GAN中，只要是generator生成的token，就会被当作“虚假”token；
2. Generator的训练目标与MLM一样，而不是像GAN一样尽力去“迷惑”discriminator。对抗地训练generator是困难的，因为对文本来说，字词是离散的，无法用反向传播把discriminator的梯度传给generator。针对这一问题，作者尝试过使用强化学习来训练generator，但是效果并没有MLM的效果好；
3. GAN的输入是随机噪声，而ELECTRA的输入是真实文本

3 实验结果与分析

3.1 模型扩展

为了提升ELECTRA模型的效果，论文尝试对模型做了多种扩展，包括：

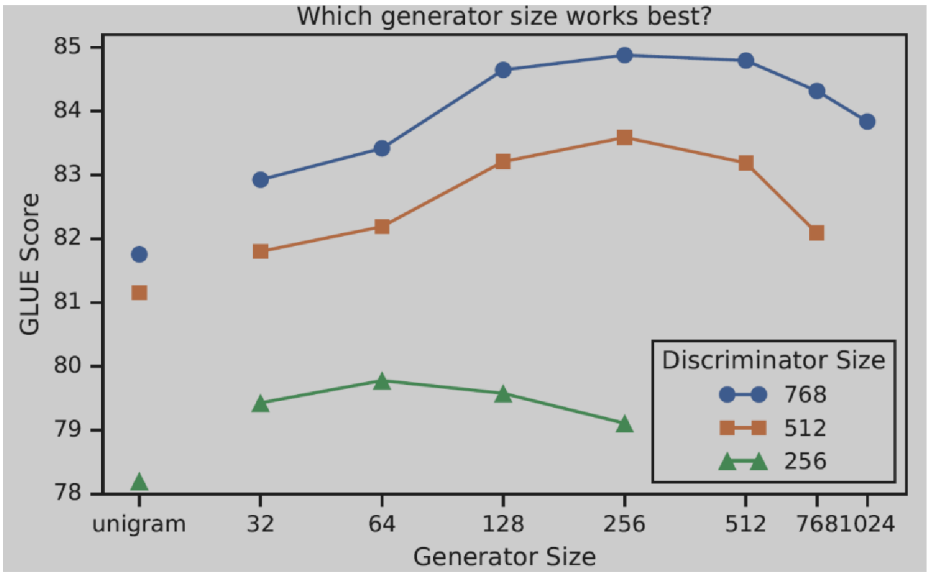
1. 权重共享

论文尝试对generator和discriminator做了两种权重共享：token embeddings共享，以及所有权重共享。

实验得到，在保持generator和discriminator大小相同的情况下，不共享权重的GLUE score是83.6，共享token embeddings的GLUE score是84.3，共享所有权重的score是84.4。论文分析，这是因为generator对token embedding有着更好的学习能力，因此共享token embeddings后discriminator也能获得更好的token embeddings。

2. 更小的Generator

如果保持generator和discriminator模型大小一样，ELECTRA大约要花费MLM预训练的两倍计算时间，因此论文提出使用小size的generator。GLUE score随generator和discriminator大小的变化如下图所示。



根据结果可得到，generator的大小为discriminator的1/4-1/2时效果最好。另外，过大的generator会给discriminator带来过大的挑战，结果反而会变差。

3. 不同的训练策略

论文额外尝试了两种训练策略

1. 使用强化学习的方式来进行对抗性训练
2. 两步式训练：先训练generator，训练好之后固定住generator，把generator的权重赋给discriminator，然后开始discriminator的训练

实验结果表明，最初提出的训练策略是最优的，其次是对抗性训练，最差的是两步式训练。但是即使是两步式训练，结果也比bert要好。

3.2 大小ELECTRA模型的性能

在这一部分，论文用实验结果证明了ELECTRA模型的核心improvements：

1. 小ELECTRA使用很少的参数量，就能够在提升训练速度的同时保证训练效果
2. 大ELECTRA用1/4的计算量就可以达到Roberta的结果

值得一提的是，就在本月6号，ELECTRA刷新了SQUAD2.0单模型的榜单，相信过不了几天，ELECTRA能在更多的地方证明它的实力。

4 小结

ELECTRA不仅提出了全新的语言模型预训练任务，还提出了类似GAN的全新预训练框架，让我们从另一个视角看到了预训练语言模型更多的可能性。从实验结果来看，用仅1/10的参数量达到BERT的效果也让我们对ELECTRA的广泛应用表示期待。目前，英文ELECTRA和中文ELECTRA预训练模型都已开源，大家可以直接上手感受下ELECTRA的魅力了。我们建了一个微信群，欢迎大家进群一起交流。

往期文章推荐

- 统一预训练语言模型UniLM2.0
- BERT--开启NLP新时代的王者
- 十分钟了解文本分类通用训练技巧
- Kaggle宠物收养比赛亚军复盘

听说点的人都是大神