**BERT-ATTACK：利用BERT对BERT进行对抗性攻击。**

# 李林阳，马若天，郭启鹏，薛向阳，邱西鹏*\*。*

复旦大学计算机学院智能信息处理上海市重点实验室。

上海市张衡路825号

*(*linyangli19,rtma19,qpguo16,xyxue,xpqiu*}*@fudan.edu.cn。

# 摘要

离散数据(如文本)的对抗性攻击已被证明比连续数据(如图像)更具挑战性，因为用基于梯度的方法生成对抗性样本是很困难的。在文本的成功攻击方法中，通常会在字符或词的层面上采用启发式替换策略，这就需要在大量可能的替换组合空间中找到最优解，同时保持语义的一致性和语言的流畅性。在本文中，我们提出了**BERT-Attack**，这是一种高质量且有效的方法，可以使用BERT预先训练好的屏蔽语言模型来生成对抗性样本。我们将BERT与它的ﬁne-tuned模型和其他深度神经模式- els在下游任务中对立起来，这样我们就可以成功地完全误导目标模型来预测incor- rectly。我们的方法在成功率和扰动百分比上都优于最先进的攻击策略，同时生成的广告样本是开放的和语义预服务的。此外，计算成本低，因此可以进行大规模生成。代码可在[https://github.com/ LinyangLee/BERT-Attack](https://github.com/LinyangLee/BERT-Attack)。

# 介紹

尽管深度学习取得了成功，但最近的研究发现，这些神经网络容易受到对抗性样本的影响，这些样本是以对原始输入的微小扰动来制作的（[Goodfel- low等](#_bookmark25)，[2014](#_bookmark25)；[Kurakin等](#_bookmark28)，[2016](#_bookmark28)；[Chakraborty等](#_bookmark21)，[2018](#_bookmark21)）。也就是说，这些对抗性样本对人类法官来说是不可感知的，同时它们会误导神经网络做出错误的预测。因此，探索这些对抗性攻击方法是非常重要的，因为最终的目标是确保神经网络是高度可靠的。

*\** 通讯作者。

和鲁棒性。虽然在计算机视觉ﬁelds中，攻击策略及其防御对策都得到了很好的探索（[Chakraborty等](#_bookmark21)，[2018](#_bookmark21)），但由于语言的离散性，针对文本的对抗性攻击仍然具有挑战性。为文本生成对抗性样本需要具备这样的特质。(1)人类法官无法察觉，但对神经模型具有误导性；(2)在语法上是ﬂuent，在语义上与原始输入一致。

以前的方法主要基于speciﬁc规则制作对抗性样本（[Li等人](#_bookmark30)，[2018](#_bookmark30)；[Gao等人](#_bookmark23)，[2018](#_bookmark23)；[Yang等人](#_bookmark40)，[2018](#_bookmark40)；[Alzantot等人](#_bookmark18)，[2018](#_bookmark18)；[Ren等人](#_bookmark37)，[2019](#_bookmark37)；[Jin等人](#_bookmark27)，[2019](#_bookmark27)；[Zang等人](#_bookmark41)，[2020](#_bookmark41)）。因此，这些方法是difﬁcult guaran- tee ﬂuency和语义保存在生成的对抗性样本同时。另外，这些手工制作方法相当com- plicated。它们使用多个语言约束，如NER标记或POS标记。引入语境化的语言模型作为au- tomatic扰动生成器可以使这些规则设计变得更加容易。

最近兴起的预训练语言模型，如BERT（[Devlin等](#_bookmark22)，[2018](#_bookmark22)），将NLP任务的per- formances推向了一个新的水平。一方面，在下游任务上，ﬁne-tuned BERT 的强大能力使其更具挑战性，可以被敌意攻击（[Jin et al.](#_bookmark27) ，[2019](#_bookmark27)）。另一方面，BERT是在极大规模的无监督数据上预先训练的屏蔽语言模型，并且已经学会了通用的语言knowl- edge。因此，BERT有可能为输入文本生成更多ﬂuent和语义一致的substitu- tions。自然，BERT的这两个适当的关系促使我们探索用另一个BERT作为攻击者来攻击一个精确调整的BERT的可能性。

在本文中，我们提出了一种有效且高质量的对抗式样本生成方法。**BERT-Attack**，以BERT为语言模型。

6193

*Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6193-6202, November 16-20, 2020.Oc 2020计算语言学协会。

来生成对抗性样本。BERT-Attack的核心算法简单明了，包括两个阶段：在一个给定的目标模型的输入序列中寻找弱势词；然后以语义保护的方式应用BERT来生成弱势词的替代词。在BERT的能力下，扰动的产生要考虑周围的语境。在此前提下，扰动的生成是合理的。我们使用掩蔽的语言模型作为扰动生成器，并寻找能最大限度地降低错误预测风险的扰动([Goodfellow et al.](#_bookmark25) , [2014](#_bookmark25))。与以往需要传统的单方向语言模型作为约束的攻击策略不同，我们只需要in- ference一次语言模型作为扰动生成器，而不是在试错过程中反复使用语言模型对生成的对抗样本进行评分。

实验结果表明，与之前的方法相比，提出的BERT-Attack方法成功地骗过了其ﬁne-调整的下游模型，攻击成功率最高。同时，扰动百分比和查询次数大大降低，而语义保存率高。

总结一下我们的主要贡献：

我们提出了一种简单有效的方法，命名为**BERT-Attack**，以有效地生成能成功误导NLP中的最新模型的飞行和语义保存的对抗性样本，如用于各种下游任务的有限调谐BERT。

*●*

BERT-Attack与以往的攻击算法相比，在对目标模型的访问次数较少的情况下，具有更高的攻击成功率和更低的扰动百分比，同时不需要额外的评分模型ther- fore极其有效。

*●*

# 相关工作

为了探索神经网络的鲁棒性，人们广泛研究了针对连续数据（如图像）的逆向攻击（[Goodfellow等人](#_bookmark25)，[2014](#_bookmark25)；[Nguyen等人](#_bookmark34)，[2015](#_bookmark34)；[Chakraborty等人](#_bookmark21)，[2018](#_bookmark21)）。关键的想法是找到一个最小的扰动- bation，最大限度地提高做出错误预测的风险。这个最小化问题可以通过在图像的连续空间上应用梯度下降来轻松实现（[Miyato等](#_bookmark32)，[2017](#_bookmark32)）。

然而，针对文本等离散数据的对抗性攻击仍然具有挑战性。

## 文本的对抗性攻击

目前针对文本的成功攻击通常采用启发式规则来修改单词的字符（[Jin等](#_bookmark27)，[2019](#_bookmark27)），用同义词代替单词（[Ren等](#_bookmark37)，[2019](#_bookmark37)）。[Li 等](#_bookmark30)（[2018](#_bookmark30)）；[Gao 等](#_bookmark23)（[2018](#_bookmark23)）应用了基于单词 em- beddings 的扰动，如 Glove（[Pennington 等](#_bookmark35)，[2014](#_bookmark35)），它在语义和语法上并不严格协调。[Alzantot等](#_bookmark18)（[2018](#_bookmark18)）采用lan- guage模型对词嵌入空间中搜索近义词产生的扰动进行打分（[Mrksˇic´等](#_bookmark33)，[2016](#_bookmark33)），采用试错过程来寻找可能的扰动，然而产生的扰动仍然不具有语境意识，严重依赖词嵌入的余弦相似度mea- surement。手套嵌入不能保证相似向量空间与余弦相似性距离，因此扰动的语义一致性较差。[Jin等人](#_bookmark27)（[2019](#_bookmark27)）应用了一种语义增强的嵌入（[Mrksˇic´等人](#_bookmark33)，[2016](#_bookmark33)），它是不知道上下文的，因此与未扰动的输入不太一致。[Liang等人](#_bookmark31)（[2017](#_bookmark31)）使用短语级插入和删除，这就会产生与原始输入不一致的非自然句子，缺乏ﬂuency控制。为了预先服务于语义信息，[Glockner等](#_bookmark24)（[2018](#_bookmark24)）手动替换单词，以打破语言在推理系统（[Bowman等](#_bookmark19)，[2015](#_bookmark19)）。[Jia和Liang](#_bookmark26)([2017](#_bookmark26))提出了人工制作方法来At- tack机器阅读理解系统。[Lei等](#_bookmark29)（[2019](#_bookmark29)）引入使用嵌入过渡的替换策略。

虽然上述方法已经取得了良好的效果，但在扰动百分比、攻击成功率、语法正确性和语义一致性等方面仍有很大的改进空间。此外，这些方法的替换策略通常是非平凡的，导致它们被限制在speciﬁc任务中。

## 对BERT的对抗性攻击

预训练的语言模型已经成为许多NLP任务的主流。如（[Wal- lace等](#_bookmark38)，[2019](#_bookmark38)；[Jin等](#_bookmark27)，[2019](#_bookmark27)；[Pruthi等](#_bookmark36)，[2019](#_bookmark36)）等作品从很多不同的角度探讨了这些预训练语言模型。[Wallace等](#_bookmark38)（[2019](#_bookmark38)）探讨了预训练模型中学习知识可能存在的伦理问题。

# BERT-攻击

出于将BERT与BERT对立起来的有趣想法，我们提出了**BERT-Attack**，使用原始的BERT模型来制作对抗性的样本，以愚弄重新调整的BERT模型。

我们的方法包括两个步骤。(1)找出目标模型的易损词，然后...

(2)用语义相近、语法正确的词替换，直至攻击成功。

最脆弱的词是帮助目标模型做出判断的关键词。在这些词上的扰动可以在制作对抗性样本时发挥最大的作用。在确定了我们要替换的目标词后，我们使用掩蔽语言模型，根据掩蔽语言模型的top-K预测生成扰动。

## 寻找脆弱的词语

在黑盒方案下，目标模型（调整后的BERT或其他神经模型）输出的logit是我们能得到的唯一监督。我们首先选择序列中对最终输出logit有高signiﬁcance infuence的词。让*S* = [w0*, , wi* ]表示输入感性，*oy*(*S*)表示目标模型对正确标签*y的*对数输出，重要性为

*- · · · · ·*

分数*Iwi*被定义为

*Iwi* = *oy*(*S*) *. oy*(*S\wi* )*,* (1)

其中*Swi* = [w0*， ，wi*-l*，* [MASK]*，wi*＋l*，* ] 。

*- · · · · ·*

是将*wi*替换为[MASK]后的句子。

然后我们根据排名得分*Iwi从高到低对*所有单词进行排序，建立单词列表*L*，我们只取e%的最im-。

重要的词语，因为我们倾向于保持最小的扰动。

这个过程最大限度地提高了做出错误预测的风险，以前是通过在图像域中校准梯度来实现的。然后将问题表述为用语义一致的perturba- tions替换这些最脆弱的词。

## 通过BERT进行单词替换

在找出脆弱的词后，我们逐一迭代替换列表*L*中的词，以找出可能误导目标模型的per- turbations。前面的方法通常使用多个人工编写的

**生成的样品**

图1：我们替换策略的一步。

*w*

*w*

*w*



*pi*＋l

l

*pi*＋l

2

*pi*＋l

*k*

**Top-K预测的完全换位思考**

*w w* **...** cl **...**

*pli*

*p2i*

*pi*＋2

l

*pi*＋2

**排名**

2

*k*

*k*

云

c2

*ck*

**…**

*圆周率*

*pi*＋2

*wic2ck*

**BERT**

*w w* **.. ......** *wn wn*

*薇*的子词

**输入** *w w***...** *wi***...** *wn wn*

cl c2 **...** *ck*

**迭代**

云

**目标模型**

{

**…**

{

**…**

**…**

**…**

规则，以确保生成的例子与原例子在语义上一致，语法上正确，如同义词词典（[Ren等](#_bookmark37)，[2019](#_bookmark37)）、POS检查器（[Jin等](#_bookmark27)，[2019](#_bookmark27)）、语义相似性检查器（[Jin等](#_bookmark27)，[2019](#_bookmark27)）等。[Alzantot等](#_bookmark18)（[2018](#_bookmark18)）应用传统的语言模型，在每一次替换一个词的时候，都会对受扰句子进行打分。

这些生成替代物的策略不知道替代物po- sitions之间的语境（通常使用语言模型来测试替代物），因此在语义一致性方面是不够的。更重要的是，使用语言模型或POS检查器对扰动样本进行打分的成本很高，因为这个试错过程需要大量的推理时间。

为了克服在嵌入空间中使用同义词或同义词缺乏语义控制和语义保存的问题，我们利用BERT进行词语替换。掩码语言模型的真实性保证了生成的句子是相对流利和语法正确的，同时也保留了大部分的语义信息，这些语义信息将在以后被人类评估者所确认。此外，与之前使用基于规则的扰动策略的方法相比，掩蔽语言模型的预测是上下文感知的，因此动态地搜索扰动，而不是简单的同义词替换。

不同于以往的方法，上下文化的扰动生成器只用一个for- ward pass就能生成最小的扰动，而不需要运行额外的神经模组来给句子打分，耗时的部分是只访问目标模型。在没有运行额外的神经模态对句子进行打分的情况下，耗时的部分只是访问目标模型。因此，这个过程是非常有效的。

**算法1** BERT-攻击

1：**程序**词重要性排序。

2: *S* = [w0*,* wl*,* ]// 输入：标记化句子

*- · · ·*

3:*Y*金标

*-*

4:**对于***wi*在*S***做**

5:利用公式[1](#_bookmark1)计算重要性得分*Iwi。*

6: 选择词表*L* = [*wtop*-l*, wtop*-2*, - - -*]

7://用*Iwi*对*S*进行降序排序，并收集*最前面的 。K*字

8：使用BERT进行更换的**程序。**

9: *H* = [h0*, , hn*]//子字标记化的*S*序列

*- · · ·*

10:用BERT生成所有子词的top-K候选词，得到*Pen×K。*

11:**for** *wj* in *L* **do**

12:**如果***wj*是一个完整的单词**，那么**

13:得到候选人*C* = *F ilter*(*P j*)

14:替换单词*wj*

15:**其他**

16:利用PPL排名和Filter得到候选人*C。*

17: 17:替换子词[*hj, - - - , hj*＋*t*] 18:寻找可能的对抗性样本 19:**对于C**中的*ck，***做到**

20: *S/* = [w0*, , WJ L, CK,* ]// 尝试

*- · · · · · ·-*

21:**if** argmax(*oy*(*S/* ))!= *Y* **那么**

22:***return*** *Sadv* = *S/* //成功攻击。

23:**其他**

24:**如果***OY*(*S/* )*<OY(SADV)***，那么。**

25: *Sadv* = [w0*, - - - , wj* -l*, c, - - -* ] //做一次扰动。

26:***返回*无**

因此，使用屏蔽语言模型作为上下文化的扰动生成器，可以是一种可能的解决方案，以有效地制作高质量的对抗性样本。

## 词语替换策略

如图 [1 所示](#_bookmark2)，给定一个要被替换的选词 *w*，我们应用 BERT 来预测与 *w* 相似却能误导目标模型的 pos- sible 词。我们没有遵循屏蔽语言模型的设置，而是不屏蔽选择词*w*，并使用原始序列作为输入，这可以产生更多语义一致的子代（[Zhou et al.](#_bookmark43) ，[2019](#_bookmark43)）。例如，给定一个序列*"I like the cat."*，如果我们屏蔽了*猫*这个词，那么屏蔽后的语言模型将很难预测原词*cat*，因为如果序列是*"I like the dog."，那么*它也可能是一样的ﬂuent。此外，如果我们屏蔽掉给定的单词*w*，对于每一次迭代，我们将不得不重新运行屏蔽语言模型预测过程，这是很昂贵的。

由于BERT使用字节对编码(BPE)，因此，我们可以通过以下方式来实现。

将序列*S*=[w0*， ，wi，*]标记成子词标记：*H*=[h0*，hl，*h2*，]*，我们需要将所选的词与其在BERT中对应的子词对齐。

让表示BERT模型，我们将标记化的序列*H*输入到BERT中，得到输出预测*P*= （*H*）。我们不使用argmax预测，而是在每个位置取最可能的*K个*预测，其中*K*是一个超参数。

*M*

*M*

*M*

*- · · ·*

*- · · · · ·*

我们通过单词impor- tance排序过程来迭代单词，以确定扰动。BERT模型使用BPE编码来构建词汇表。虽然大多数词仍然是单词，但罕见的词被标记为子词。因此，我们将单词和子词分开处理来生成替代词。

**单词** 对于一个单词*wj*，我们使用相应的top-K预测候选词*Pj进行尝试*。我们首先筛选出从NLTK中收集到的停顿词；对于情感分类，我们会使用相应的Top-K预测候选词Pj进行尝试。

tion任务，我们用同义词词典（[Mrksˇic´等](#_bookmark33)，[2016](#_bookmark33)）ﬁlter出反义词，因为BERT掩蔽语言模型不区分同义词和反义词。然后对于给定的candi- date *ck，*我们构建一个扰动序列*H/* = [h0*， ，hj*-l*，ck，hj*＋l ]。如果目标模型已经被愚弄得预测错误，我们就打破循环，得到最终的对抗样本*Hadv*；否则，我们就从ﬁltered候选人中选择一个最佳的扰动，然后转向词表*L*中的下一个词。

*- · · · · ·*

**子词** 对于一个在 BERT 中被标记为子词的词，我们无法直接获得其替代词。因此，我们利用子词组合的迷惑性，从子词层面的预测中寻找合适的替代词。给定单词*w*的子词[*h ，h ，------，h* ]，我们列出所有可能的

**Yelp** Review分类数据集，包含。继[Zhang等人](#_bookmark42)（[2015](#_bookmark42)）之后，我们对数据集进行处理，构建一个极性分类任务。

**IMDB**文档级的电影评论数据集，其中平均序列长度比Yelp数据集长。我们将数据集处理成极性分类任务[1](#_bookmark4)。

*●*

*●*

**AG的新闻**句子级别的新闻类型分类-ﬁcation数据集，包含4种类型的新闻。世界、体育、商业和科学。

*●*

**FAKE**假新闻分类数据集，从Kaggle假新闻挑战赛[2](#_bookmark5)中去检测一个新闻文档是否是假的。

*●*

## 自然语言推理

* + - * **SNLI**斯坦福语言推理任务（[Bow-。](#_bookmark19)

*lt*

预测*P*的组合

*et×K*

来自*M*,

[manetal.](#_bookmark19) , [2015](#_bookmark19)）。)给定一个前提和一个

假设，而目标是预测hy-。

即*Kt*子词组合，我们可以通过逆向BERT标记化过程将其还原为正常的词。我们将这些组合输入到BERT-MLM中，得到这些组合的困惑度。然后，我们对所有组合的困惑度进行排序，得到前K个组合，以确定合适的子词组合。

给定合适的扰动，我们用最有可能的扰动替换原词，并通过迭代重要性词排名列表来重复这个过程，从而找到最终的对抗性样本。通过这种方式，我们有效地获取了对抗性样本*Sadv*，因为我们只迭代了一次屏蔽语言模型，并且在没有其他检查策略的情况下使用屏蔽语言模型进行扰动。

我们在算法[1](#_bookmark3)中总结了两步BERT-Attack过程。

# 实验

## 数据集

我们将我们的方法应用于攻击不同类型的NLP任务，形式为文本分类和自然语言推理。按照[Jin等人](#_bookmark27)([2019](#_bookmark27))的要求，我们在给定任务的测试集中随机选取的1k个测试样本上评估我们的方法，这些样本与[Alzantot等人](#_bookmark18)([2018](#_bookmark18));[Jin等人](#_bookmark27)([2019](#_bookmark27))所使用的分割相同。GA方法只使用了FAKE、IMDB数据集中50个样本的子集。

**文本分类** 我们使用不同类型的文本分类任务来研究我们方法的有效性。

Pothesis是包含、神经或矛盾的前提。

**MNLI**语言推理数据集上的多体裁文本，涵盖了转录的语音、流行语和政府报告（[Williams等](#_bookmark39)，[2018](#_bookmark39)），这与SNLI数据集相比，包括与训练域匹配的评价数据和与训练域不匹配的评价数据，更加复杂，具有多样化的书面和口语风格文本。

*●*

## 自动评估指标

为了衡量生成样本的质量，我们设置了各种自动评价指标。成功率，是攻击后准确性的反面，是衡量攻击方法成功与否的核心指标。同时，扰动百分比也是至关重要的，因为从基因上讲，较少的扰动会导致更多的语义一致性。此外，在黑盒设置下，目标模型的查询是唯一可获取的信息。一个样本的恒定查询不太适用。因此每个样本的查询次数也是一个关键指标。正如在TextFooler（[Jin等人](#_bookmark27)，[2019](#_bookmark27)）中使用的那样，我们还使用通用句子En- coder（[Cer等人](#_bookmark20)，[2018](#_bookmark20)）来衡量对抗样本和原始序列之间的语义一致性。为了在语义保存和攻击成功率之间取得平衡，我们设置了一个语义相似度分数的阈值来ﬁlter相似度较低的例子。

1 https://datasets.imdbws.com/。

2[https://www.kaggle.com/c/fake-news/data](http://www.kaggle.com/c/fake-news/data)

**Dataset Method 原有的 Acc 被攻击的 Acc 扰动% 查询数 Avg Len 语义模拟**

BERT-Attack(我方)

**15.5 1.1 1558**

**0.81**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **假的** | TextFooler([Jin et al.](#_bookmark27) , [2019](#_bookmark27)) | 97.819.3 | 11.7 | 44038850.76 | |
|  | GA([Alzantot等人](#_bookmark18)，[2018](#_bookmark18)) | 58.3 | 1.1 | 28508 | - |

BERT-Attack(我方)

**5.1 4.1 273**

**0.77**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Yelp** | TextFooler | 95.66.6 | 12.8 | 7431570.74 | |
|  | GA | 31.0 | 10.1 | 6137 | - |

BERT-Attack(我方)

**11.4 4.4 454**

**0.86**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **IMDB** | TextFooler | 90.913.6 | 6.1 | **11342150.86** | |
|  | GA | 45.7 | 4.9 | 6493 | - |

BERT-Attack(我方)

**10.6 15.4 213**

**0.63**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **AG** | TextFooler | 94.212.5 | 22.0 | 35743 | 0.57 |
|  | GA | 51 | 16.9 | 3495 | - |

BERT-Attack(我方)

7.4**16.1 /12.4/9.316/30**

0.40/**0.55**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **SNLI** | TextFooler | 89.4(H/P)**4.0**/20.8 | 18.5/33.4 | 60/1428/ | **180.45**/0.54 |
|  | GA | 14.7/- | 20.8/- | 613/- | - |

BERT-Attack(我方)

**7.9/11.9 8.8/7.919/44**

0.55/**0.68**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| mMatNchLeId | TextFooler | 85.1(H/P)9.6/25.3 | 15.2/26.5 | 78/15211/**210.57**/0.65 | |
|  | GA | 21.8/- | 18.2/- | 692/- | - |

BERT-Attack(我方)

**7/13.7 8.0/7.124/43**

0.53/**0.69**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 误区 | TextFooler | 82.1(H/P)8.3/22.9 | 14.6/24.7 | 86/16212/**220.58**/0.65 | |
|  | GA | 20.9/- | 19.0/- | 737/- | - |

表1.对各种ﬁne-tuned BERT模型的攻击结果。攻击不同的BERT模型的结果。TextFooler是最先进的基线。对于MNLI任务，我们分别攻击假设(H)或前提(P)。

## 攻击结果

如表[1](#_bookmark6)所示，BERT-Attack方法成功地愚弄了其下游的ﬁne-tuned模型。在文本分类和自然语言推理任务中，重新调整的BERT都不能正确地对生成的对抗性样本进行分类。

攻击后的平均准确率低于10%，说明大部分样本都成功地扰动了最先进的分类模型。同时，扰动百分比小于10%，这比以前的作品明显要少。

此外，**BERT-Attack**成功地攻击了所有列出的任务，这是在diverseﬁed领域，如新闻分类，审查分类，在不同领域的lan- guage推理。结果表明，该攻击方法在dif- ferent任务中是稳健的。与[Jin等人](#_bookmark27)([2019](#_bookmark27))[3](#_bookmark7)和[Alzantot等人](#_bookmark18)([2018](#_bookmark18))[4](#_bookmark8)介绍的强基线相比，BERT-Attack方法更efﬁcient

3https://github.com/jind11/TextFooler 4https://github.com/QData/TextAttack。

且更不易察觉。我们的方法的查询次数和扰动比例要小得多。

我们可以观察到，由于扰动百分比非常低，所以一般来说，在处理复习分类任务时比较容易。BERT-Atack可以通过只替换少量的单词来误导目标模型。由于平均序列长度相对较长，目标模型往往只通过序列中的几个词来进行判断，这不是人类预测的自然方式。因此，这些关键词的扰动会导致目标模型的预测不正确，暴露出它的脆弱性。

## 人类评价

为了进一步评估生成的对抗性样本，我们设置了人工评估，以衡量生成的样本在语气和语法以及语义保存方面的质量。

我们请人类评委对生成的广告混合句子的语法正确性进行打分。

versarial样本和原始序列，按照[Jin等人](#_bookmark27)（[2019](#_bookmark27)）的评分从1-5。然后，我们要求人类评判员在经过洗牌的原始文本和对抗性文本的混合中进行预测。我们使用IMDB数据集和MNLI数据集，对于每个任务，我们选择100个原始样本和对抗性样本供人类法官使用。我们请三位人类注释者来评估这些例子。对于标签预判，我们取多数类作为预测标签，对于语义和语法检查，我们使用注释者之间的平均得分。

从表[2](#_bookmark9)中可以看出，对抗样本的语义得分和语法得分都与原样本接近。MNLI任务是一个由人类精心设计的基于前提的hy- potheses构建的句子对预测任务，因此原始句子对共享相当多的相同单词。这些词的扰动会使人类法官无法正确预测，因此准确率低于简单的句子分类任务。

**Dataset 准确度 语义语法**

Original 0.90 3.9 4.0

**MNLI**

Adversarial 0.70 3.7 3.6

Original 0.91 4.1 3.9

**IMDB**

Adversarial 0.85 3.9 3.7

表2：人的评价结果。

## BERT-对其他型号的攻击

BERT-Attack方法也适用于攻击其他目标模型，而不仅仅局限于其整调模型。从表[3](#_bookmark10)中可以看出，该攻击对基于LSTM的模型是成功的，说明BERT-Attack对各种模型都是可行的。在BERT-Attack下，ESIM模型在MNLI数据集中更加稳健。我们认为，分别对两个句子进行编码可以得到更高的鲁棒性。在攻击 BERT-large 模型时，表现也很出色，说明 BERT-Attack 不仅仅是针对自己的ﬁne- tuned 下游模型，还能成功攻击不同的预训练模型。

# 消融和讨论

## 考生号的重要性

候选池范围是BERT-攻击算法中使用的主要超参数。从图[2](#_bookmark11)中可以看出，攻击率是随着候选池大小的增加而上升的。直观地讲，一个较大的

**Dataset Model Ori Acc Atk Acc Perturb %**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **IMDB** Word-LSTM 89.8 | | | 10.2 | 2.7 |
|  | BERT-大号 | 98.2 | 12.4 | 2.9 |
| **Yelp** | 字-LSTM | 96.0 | 1.1 | 4.7 |
|  | BERT-大号 | 97.9 | 8.2 | 4.1 |
| **MNLI** | ESIM | 76.2 | 9.6 | 21.7 |
| 匹配的BERT-大号 | | 86.4 | 13.2 | 7.4 |

表3：BERT-攻击与其他模型的对比。

100

IMDB

Yelp SNLI

假的

MNLI AG

90

攻击成功率

80

70

60

6 12 24 36

K值

图2：在攻击过程中使用不同的候选数*K*。

*K*会导致语义相似度降低。然而，通过通用句子编码器进行的语义测量却保持在一个稳定的范围内，（经验表明，语义相似度下降不到2%），这说明候选句子都是合理的，并且与原句在语义上是一致的。

此外，一个固定的候选数在实际使用中可能是僵化的，所以我们运行一个测试，使用一个阈值来切断可能性较小的候选数，作为一个可信的扰动。

从表[4](#_bookmark12)中可以看出，当使用可扩展阈值-老来切断不合适的候选者时，攻击过程中的查询次数较少，这说明被掩蔽的lan-guage模型预测的一些候选者的预测分数较低，可能没有意义，所以跳过这些候选者可以节省不必要的查询。这说明，被掩蔽的lan-guage模型预测的一些预测得分较低的候选者可能没有意义，所以跳过这些候选者可以节省不必要的查询。

**Dataset Method Ori Acc Atk Acc 查询 %** IMDBFixed*-*90.9 11.4 K454。

有Threshold 90.9 12.4 440人

表4：灵活的候选人 使用门槛来切断不合适的候选人。

## 5.2 序列长度的重要性

BERT-Attack方法是基于上下文u- alized masked language模型的。因此，序列长度在高质量的扰动过程中起着重要作用。可见，与之前的方法专注于攻击NLI任务的假设不同，我们的目标是其平均年龄长度较长的前提。这是因为我们认为，当处理极短的序列时，上下文替换将不太合理。为了避免这样的问题，我们认为很多词级的同义词替换策略可以与BERT-攻击相结合，使得BERT-攻击方法更加适用。

|  |  |  |
| --- | --- | --- |
| BERT-Atk85.1 | 7.9 | 8.8 |
| 匹配+Adv Train 84.6 | 23.1 | 10.5 |

**Dataset 方法 Ori Acc Atk Acc Perturb % MNLI (MNLI)**

表5：对抗性训练结果。

**Dataset 型号**LSTM BERT-底座 BERT-大号 Word-LSTM - 0.78 0.75

表明该模型更难被攻击。因此，生成的数据集可以作为额外的数据，用于进一步探索如何使神经模型更加稳健。

**Dataset Model 语义Yelp** BERT-5.1 4.1 Atk0.77

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 不含副词 | 7.1 | 4.3 | 0.74 |
| **MNLI** | BERT-Atk | 11.9 | 7.9 | 0.68 |
|  | 不含副词 | 14.7 | 9.3 | 0.63 |

表7：对分词级攻击的影响。

## 5.4对子词级攻击的影响

BPE方法是目前处理大量词的最有效的方法，如在BERT中使用。我们建立了一个不使用子词级攻击的对比实验。即我们跳过那些被标记了多个子词的词。

从表[7](#_bookmark15)中可以看出，采用分词级攻击可以获得更高的表现，不仅在攻击成功率上更高，而且在减少

**IMDB**

BERT 0.83 - .71

BERT 0.87 0.86 -

扰动百分比。

**数据集方法 Atk Acc Perturb % Semantic**

ModelESIM BERT-base BERT-large

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | ESIM | - | 0.59 | 0.60 |
| **MNLI** | BERT-base | 0.60 | - | 0.45 |
|  | BERT-大号 | 0.59 | 0.43 | - |

**MNLI**

匹配

MIR 7.9 8.8 0.68

Random 20.2 12.2 0.60

LIR 27.2 15.0 0.60

表6：以攻击的accu- racy作为评价指标的可转移性分析。栏为攻击中使用的目标模型，行为测试模型。

## 5.3 转移性和对抗性训练

为了测试生成的 adver- sarial 样本的可转移性，我们采取针对不同目标模型的样本来攻击其他目标模型。这里，我们使用BERT-base作为所有不同目标模型的屏蔽语言模型。从表[6](#_bookmark14)中可以看出，样本在NLI任务中是可转移的，而在文本分类中转移性较差。

同时，我们使用从训练集生成的对抗样本进一步对目标模型进行ﬁne- tuning，然后在之前使用的测试集上进行测试。如表[5](#_bookmark13)所示，在ﬁne- tuning中使用的生成样本帮助目标模型变得更加健壮，同时精度接近于用清洁数据集训练的模型。攻击变得更加difﬁcult。

表8：最重要排名(MIR)与最不重要排名(LIR)

## 对词语重要性排序的影响

字词重要性排名策略应该是为了筛选出对NN模型至关重要的关键，这很像计算FGSM算法中错误预测的最大风险([Good- fellow et al.](#_bookmark25) , [2014](#_bookmark25))。当不使用词重要性排名时，攻击算法的成功率就会降低。

**Dataset Method 运行时间(秒/样本)**

BERT-Attack(w/o BPE) 14.2)

IMDBBERT-Attack(w/BPE) 16.0)

Textfooler([Jin etal.](#_bookmark27) , ) 201942.4.

GA([Alzantot etal.](#_bookmark18) , ) 20182582.

表9：运行时间比较。

DatasetLabel

Ori 有些房间有阳台，. Hypothesis 所有的房间都有阳台。矛盾

**MNLI**

Adv许多房间都有阳台. Hypothesis 所有的房间都有阳台关闭。中性 这是很难的小说诺桑觉寺的爱好者坐在通过这个BBC改编和负

**IMDB**

东方

为什么有这么多关于Tilney家族和Tilney夫人的死亡的事实被不必要地改变了?’

这是很难的小说诺桑觉寺的情人坐在通过这个BBC的改编和积极的

为什么有这么多关于Tilney家族和Tilney夫人的死亡的事实被不必要地改变？’

我第一次看到这部电影在80年代初......它真的有很好的画质太......无论如何，我是积极的

**IMDB**

东方

很高兴我又找到了这部电影......我最喜欢的部分是当他从这个可怜的家伙劫持汽车......这是一部电影，我可以看一遍又一遍。 我强烈推荐它。

我第一次看到这部电影在80年代初......它真的有很好的画质太......无论如何，我是消极的。

我很高兴再次发现这部电影......我最喜欢的部分是当他劫持汽车从这个可怜的家伙......这是一部电影，我可以看一遍又一遍.我非常推荐它。

表10：一些生成的对抗性样本。原点标签是正确的预测，而标签是不利的预测- tion。只有红色部分受到扰动。我们只攻击MNLI任务中的处所。FAKE数据集和IMDB数据集中的文本被切割成表中的ﬁt。原始文本包含200多个单词。

## 运行时间比较

由于BERT-Attack在生成过程中不使用语言修改器或句子编码器来测量输出序列，同时，查询数量较少，因此运行时间比以前的方法更快。从表[9](#_bookmark16)中可以看出，BERT-Attack比generic algo- rithm([Alzantotetal.](#_bookmark18) , [2018](#_bookmark18))快得多，是Textfooler的3倍。

## 生成对抗性句子的例子

从表[10](#_bookmark17)中可以看出，生成的对抗性样本与它的原始输入在语义上是一致的，而目标模型却做出了不正确的预言。在审查分类样本和语言推理样本中，扰动不会误导人类法官。

# 结论

在这项工作中，我们提出了一种高质量且有效的方法**BERT-Attack**，利用BERT屏蔽语言模型生成对抗性样本。实验结果表明，所提出的方法在保持最小扰动的同时，实现了较高的成功率。尽管如此，由掩蔽语言模型生成的候选词有时可能是反义词或与原词不相关，造成语义损失。因此，增强语言模型以生成更多的语义再扰动，可以成为未来完善BERT-Attack的一个可能的解决方案。

# 鸣谢

We would like to thank anonymous review- ers for their valuable comments.We are thank- ful for the help of Demin Song, Hang Yan and Pengfei Liu.这项工作得到了国家自然科学基金委（编号：61751201，62022027和61976056）的支持，上海。

市级科技重大项目（编号：2018SHZDZX01）和ZJ实验室。

# 参考文献

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani B. Srivastava, and Kai-Wei Chang.2018.[Generating natural language adversar- ial examples](http://arxiv.org/abs/1804.07998).*CoRR*，abs/1804.07998.

Samuel R Bowman、Gabor Angeli、Christopher Potts和Christopher D Manning。2015.A large anno- tated corpus for learning natural language inference.*arXiv preprint arXiv:1508.05326*.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018.Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Anirban Chakraborty、Manaar Alam、Vishal Dey、Anu- pam Chattopadhyay和Debdeep Mukhopadhyay。2018.Adversarial attacks and defences:A survey.*arXiv preprint arXiv:1810.00069*.

Jacob Devlin，Ming-Wei Chang，Kenton Lee，and Kristina Toutanova.2018.[BERT：pre-training of deep bidirectional transformers for languageunder- standing](http://arxiv.org/abs/1810.04805).*CoRR*，abs/1810.04805.

Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yan-jun Qi.2018.Black-box generation of adversarial text sequences to evade deep learning classiﬁers.In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 50-56.

Max Glockner、Vered Shwartz和Yoav Goldberg。2018.Breaking nli systems with sentences that require simple lexical inferences.*arXiv preprint arXiv:1805.02266*.

Ian J Goodfellow、Jonathon Shlens和Christian Szegedy。2014.Explaining and harnessing adversar- ial examples.*arXiv preprint arXiv:1412.6572*.

贾彦宏和梁珀西。2017.Adversarial exam- ples for evaluating reading comprehension systems.*arXiv preprint arXiv:1707.07328*.

金迪、金志敬、周天一和Peter Szolovits.2019.[IsBERTreally robust? natural languageattackontextclassiﬁcationandentailment](http://arxiv.org/abs/1907.11932).*CoRR*，abs/1907.11932.

Alexey Kurakin、Ian Goodfellow和Samy Bengio。2016.Adversarial examples in the physical world.*arXiv preprint arXiv:1607.02533*.

Qi Lei，Lingfei Wu，Pin-Yu Chen，Alexandros G Di- makis，Inderjit S Dhillon，and Michael Witbrock.2019.Discrete adversarial attacks and submodular optimization with applications to text classiﬁcation.*Systems and Machine Learning（SysML）*.

李金凤，纪守玲，杜天宇，李博，王婷.2018.Textbugger:Generating adversarial text against real-world applications. *ArXiv preprint arXiv:1812.05271*.

梁斌，李洪成，苏苗强，边潘，李锡荣，史文昌。2017. Deep *arXiv preprint arXiv:1704.08006*.

Takeru Miyato, Shin ichi Maeda, Masanori Koyama, and Shin Ishii.2017.Virtual adversarial training:A regularization method for supervised and semi-supervised learning.第41卷，1979-1993页。

Nikola Mrksˇic´，Diarmuid O Se´aghdha，Blaise Thom- son，Milica Gasˇic´，Lina Rojas-Barahona，Pei- Hao Su，David Vandyke，Tsung-Hsien Wen，and Steve Young.2016.Counter-ﬁtting word vec- tors to linguistic constraints.*arXiv preprint arXiv:1603.00892*.

Anh Nguyen, Jason Yosinski, and Jeff Clune.2015.Deep neural networks are easily fooled: High con- ﬁdence predictions for unrecognizable images.In *Proceedings of the IEEE conference on computer vi- sion and pattern recognition*, pages 427-436.

Jeffrey Pennington, Richard Socher, and Christopher Manning.2014.Glove:Glove: Global vectors for word representation.In *Proceedings of the conference on empirical methods in natural language processing*, pages 1532-1543.

Danish Pruthi, Bhuwan Dhingra, and Zachary C Lip- ton.2019.Combating adversarial misspellings with robust word recognition.*arXiv preprint arXiv:1905.11268*.

任书怀、邓义和、何坤、车万祥。2019.Generating natural language adversarial ex- amples through probability weighted word saliency.In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1085-1097.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh.2019.Universal adversarial trig- gers for attacking and analyzing NLP.*Empirical Methods in Natural Language Processing*.

Adina Williams、Nikita Nangia和Samuel Bowman。2018.A broad-coverage challenge corpus for sen- tence understanding through inference.In *Proceed- ings of the Conference of the North American Chap- ter of the Association for Computational Linguistics:Human Language Technologies*, pages 1112-1122.

Puyudi Yang, Jianbo Chen, Cho-Jui Hsieh, Jane-Ling Wang, and Michael I Jordan.2018.Greedy attack and gumbel attack:Generating adversarial examples for discrete data. *ArXiv preprint arXiv:1805.12316*.

袁藏，齐凡超，杨承浩，刘志远，张萌，刘群，孙茂松。2020.Word-level textual adversarial attacking as combina- torial optimization.In *Proceedings of the 58th An- nual Meeting of the Association for Computational Linguistics*, pages 6066-6080.

张翔，赵俊波，晏乐村。2015.Character-level convolutional networks for text clas- siﬁcation.In *Advances in neural information pro- cessing systems*, pages 649-657.

周望春书，葛涛，徐珂，魏芙蓉，周明.2019.[BERT-based lexical substitution](https://doi.org/10.18653/v1/P19-1328).In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3368-3373, Florence, Italy.Association for Compu- tational Linguistics.