第三十四次AAAI人工智能会议（AAAI-20）。

**机器阅读理解的鲁棒对抗式训练方法。**

**刘凯,**1\***刘鑫,**2\***杨安,**3**刘静,**1**苏劲松,**2**李苏建,**3**佘巧巧**1。

1百度公司，北京，中国 2厦门大学，厦门，中国

3教育部北京大学计算语言学重点实验室

{liukai20, liujing46, sheqiaoqiao}@百度网盘，[liuxin@stu.xmu.edu.cn。](mailto:liuxin@stu.xmu.edu.cn)

{yangan, lisujian}@pku.edu.cn, [jssu@xmu.edu.cn](mailto:jssu@xmu.edu.cn)

### 摘要

对于机器阅读理解（MRC）模型来说，缺乏鲁棒性是一个严重的问题。为了缓解这个问题，一个最有希望的方法是用复杂的设计的对抗性例子来增强训练数据集。一般来说，这些例子是根据观察到的成功对抗性攻击的模式通过规则创建的。由于对抗性例子的类型是不计其数的，因此手动设计和丰富训练数据以抵御所有类型的对抗性攻击是不够的。在本文中，我们提出了一种新型的鲁棒对抗性训练方法，以更通用的方式来im-证明MRC模型的鲁棒性。给定一个在原始数据集上训练良好的MRC模型，我们的方法根据当前模型的参数动态地生成对抗性实例，并通过使用生成的实例在迭代sched- ule中进一步训练模型。当应用于最先进的MRC模型，包括QANET、BERT和ERNIE2.0时，我们的方法在5个以不同方式构建的对抗性数据集上获得了明显和全面的改进，而没有牺牲原始SQuAD开发集的性能。更重要的是，当与其他数据增强策略相结合时，我们的方法进一步提升了广告对抗式数据集的整体性能，并优于最先进的方法。

# 介紹

机器阅读理解（MRC）已经成为近年来的热门研究课题。为了创建更好的MRC模型，人们投入了大量的努力（Seo等人，2016；Wang等人，2017；Yu等人，2018；Devlin等人，2018）。具体来说，最近的进展表明，几个MRC模型可以在几个数据集上实现人类的平等。然而，（Jia和Liang 2017）揭示，这些先进的模型容易受到特别设计的对抗性攻击。模型性能在由原始答案段落和生成的误导性文本组成的 adversar- ial 示例上明显下降。类似的问题有

中已经观察到（Goodfellow，Shlens，和Szegedy，2014；Zhang等人，2017）。如表1所示，不同类型的对抗性例子都能够分散MRC模型的注意力。因此，如何提高现有MRC模型的鲁棒性是一个关键问题。

为了处理上述的鲁棒性问题，再搜索者已经做了多种尝试。目前，最直接有效的方法是用对抗性例子增强训练数据集。Wang and Bansal 2018）通过纳入- ing adversarial examples that ﬁt a certain type of attacks来增强训练数据集，并在增强的数据集上训练一个MRC模型。这sig- niﬁcantly提高了已知某些类型攻击下的模型鲁棒性。然而，这样的增强数据集更能够模拟已知类型的广告实例，而忽略其他未观察到的类型。Ac- cording to our observation, the augmented training dataset of (Wang and Bansal 2018) helps defense adversarial attacks proposed in (Jia and Liang 2017) well, but still fails on other types of adversarial attacks ( shown in the experiment sec- tion).Hence, we render that rule-based data augmentation approach is not adequate since the types of adversarial ex- amples are innumerable.

为了应对上述挑战，我们提出了一种模型驱动的方法来生成可以攻击给定MRC模型的对抗性例子。然后，我们通过使用生成的对抗性例子来重新训练和加强MRC模型。我们的方法的主要优点是它不需要任何对抗性攻击类型的规范，而且我们期望我们的模型在一般的对抗性攻击下更加稳健。

具体来说，我们的方法可以分为三个步骤。

(1)我们将MRC模型作为一个黑箱，并从原始数据集中获取每个实例的per- turbation词嵌入序列。这些扰动词嵌入序列很可能导致MRC模型给出错误的预测。(2)从每一个扰动词嵌入序列中

Copyright *§*c 2020, Association for the Advancement of Artiﬁcial.

序，我们取样一个单词序列。然后，我们取

情报网（www.aaai.org）。保留所有权利。

∗这些作者对这项工作做出了同样的贡献。

†这项工作是在提交人在美国大学实习期间完成的。

百度公司

采样序列作为误导性文本，并将其插入到原始实例中以创建对抗性实例。(3)我们用步骤2生成的ad- versarial实例在原始数据集上重新训练MRC模型。然后，重复步骤

**问题 燃烧的**独特品质在相学理论中是不存在的？

**Passage** 木材或煤炭等高度可燃且几乎没有残留物的材料被认为主要 **空气**在phlogiston理论中并没有起到作用，最初也没有进行任何定量实验来检验这一观点，而是根据对某物燃烧时的观察，大多数常见的物体似乎变得更轻，似乎在这个过程中失去了什么。像木材这样的物质在燃烧中整体重量增加的事实，是被气体燃烧产物的浮力所掩盖的 ......

### 敌意类型敌意例子与误导性文字预测。

AddSent 通道+*发动机燃烧的独特品质是存在于*philogistonengine*的。*

{}

*理论。*

AddAny{passage}+ *周跑了，我们不在，我们不在，我们不在，*我们*不在，我们不在，我们不在，我们不在，我们不在，我们不在。*

AddAnsCtx 通道+*并没有在phlogiston理论中发挥作用，也没有任何最初的定量经验*-浮力。

{}

*为测试这一想法而进行的各项活动。*

表1：通过在段落结尾处附加各种形式的误导性文本来攻击*BERTbase*(Devlin等人2018)的一个例子。在没有附加任何误导性文本的情况下，模型预测的结果是正确答案"空气"（黑体）。但是，当添加不同类型的误导性文字时，该模型就会分心，预测出错误的答案。误导性文本的创建方式不同。*AddSent：*(Jia and Liang 2017)通过按照一定的规则修改题目并人工校对生成误导性文本；*AddAny：*(Jia and Liang 2017)在各种MRC模型上自动逐字搜索误导性文本；*AddAnsCtx：*我们通过删除答案句中的答案词生成误导性文本。

1与重新训练的模型一起使用，直到它收敛。通过这种方式，我们期望经过良好训练的模型能够处理更普遍的攻击，而不是特定类型的对抗性例子。

实验结果表明，我们的方法可以在五种不同类型的对抗性例子上明显提高MRC模型的鲁棒性。基于最先进的MRC模型之一ERNIE2.0(Sun等人，2019)，我们的方法在所有类型的对抗性测试集上平均获得了8.4%的F1得分的明显改善。F1得分的整体改善表明，我们的训练- ing方法以更普遍的方式加强了模型的鲁棒性。此外，结合人工设计的训练数据(Wang and Bansal 2018)，我们的方法可以fur- ther提升平均F1得分，并且比不同的MRC模型获得2.3%的改进- ment。我们的贡献是con- cluded如下。

我们提出了一种模型驱动的方法，以提高MRC模型对各种类型的对抗性例子的防御能力，而不是指定对抗性例子的类型，它可以产生分散MRC模型注意力的对抗性例子。实验结果表明，我们的方法能够应对更多的一般性攻击，而不是规格化的攻击，我们的方法不需要指定任何类型的对抗性例子，它能够生成分散MRC模型注意力的对抗性例子。实验结果表明，我们的方法能够应对更普遍的攻击，而不是特定类型的对抗性例子。

-•

我们的方法是对其他数据增强方法的良好补充。通过使用我们的训练方法，可以进一步提高MRC模型的鲁棒性。

-•

# 相关工作

研究者们从很多方面致力于MRC系统的鲁棒性问题。其中大部分可以归结为两类。

**丰富数据。**抵御对抗性例子的一个直接有效的方法就是生成相应的例子，并对其进行训练。(贾和梁2017)设计了

一些类型的对抗性例子，并在各种MRC模型上研究它们。为了防御那些类型的ad- versarial例子，（Wang和Bansal 2018）根据规则自动创建了额外的训练样本，并用这些样本丰富了训练数据。基于spe- ciﬁc设计的训练数据，MRC模型能够在AddSent任务上实现最先进的性能。与上述工作不同的是，我们的工作并不是针对任何特定的对抗性数据集而设计的，我们试图寻求一种更通用的方式来加强模型的鲁棒性。

**模型改进。**许多研究人员试图更好地设计和训练MRC模型，以提高模型的鲁棒性，抵御对抗性例子。Salant和Berant 2018）通过使用预训练的语言模型嵌入作为输入，以收集丰富的语境化信息，提高模型的鲁棒性。Min等2018）提出了一种句子选择器，选择最小的句子集进行进一步的预测，以避免许多干扰。(Liu et al. 2018)设计了一个输出层，对多预测进行平均，提高模型的鲁棒性。(Hu等2018b)通过蒸馏训练的方法，在 ensem- ble ones的基础上训练了稳健的单模型。我们的训练方法不需要对MRC模型加入启发式设计，而是不需要修改任何模型结构，它可以应用于所有可派生模型。

除了致力于MRC系统之外，许多努力也致力于研究文本上的对抗性攻击方法。(Behjati et al. 2019)试图通过训练- ing扰动嵌入来分散文本分类器的注意力。(Iyyer et al. 2018)提出了一种语法控制的释义网络来生成语法上的对抗性例子。(Gong等人2018)和(Sato等人2018)通过训练per- turbation embeddings和搜索最近的tokens来生成误导性文本。而（Jia和Liang 2017；Alzantot等人2018）通过替换文本中的代币itera- tively，自动生成误导文本，直到获得成功攻击。区别于攻击meth-

ods上述，我们的工作不仅发现了更多的攻击方式，而且还试图以有效的方式提高模型的鲁棒性。

# 对抗式训练法

类似于生成式对抗网络（GAN），我们的广告-。

由于主流的MRC模型通常采用字em-垫层作为底层输入块，我们假设MRC模型的输入是嵌入，并将模型简化为。

*f* (*eq, ep*; *θ*)= *argmax Pr*(*s e q,e p*; *θ*)(2)

|

*s*

其中*eq*表示嵌入序列*eq eq ...eq的* 的

多功能训练法在以下几个方面进行最小-最大的博弈。

疑问

1 2*m*

*d*

对抗式实例生成器和相应的MRC

叮

1

2

*n*

*i*

*q*，其中*eqi*∈*R*表示第*i个*令牌嵌入-。

模型。由于离散代币的生成不是一个dif- ferentiable的过程，所以没有采用强化学习-------。

*q*的维度大小为*d，*同理，*ep*表示通道嵌入序列*ep ep ...ep* ，且*ep*∈*Rd*

ing方法（Kusner和Herna´ndez-Lobato，2016；Yu et

表示*p*的第*i*个token嵌入。

ulary嵌入表*V*∈*R*|*V* |×*d*，*eq*和*ep*都是

al.2017），我们选择了一个抽样策略来生成 adver- sarial 示例。我们的训练过程遵循三个步骤

检索结果

*i i*

和*tip* .

算法。(1) 将一个训练有素的MRC模型作为广告语生成器 训练扰动嵌入序列以最小化给定问题和段落下真实答案的输出概率。(2)从扰动嵌入中贪婪地抽取单词序列作为误导性文本，以创建和丰富我们的对抗性示例集。

(3)训练MRC模型，使真实答案的概率最大化，以抵御那些对抗性的例子。然后回到步骤1，用重新训练的模型作为新的生成器，直到收敛。

为了充分覆盖潜在的对抗性ex-amples类型，我们的方法试图生成两种错误引导文本。

1. 误导性的答案文本：误导性的文本，它试图使MRC模型相信正确答案就在文本中。
2. 误导性语境文本：作为con-文本的误导性文本，试图分散MRC模型对正确答案的注意力，并将其引导到错误的答案上(在误导性文本中没有必要)。

其中，每一种类型在对抗性 训练中都拥有自己的损失函数。同时，为了增加多样性，我们的方法还试图控制误导性文本与问题ei- ther相似或不相似，而大多数已知的误导性文本主要是与问题相关的（Jia和Liang 2017；Wang和Bansal 2018）。

## MRC模型定义

我们将所有可推导的MRC模型都视为黑箱，不考虑网络内部细节，我们专注于模型的输入和输出。给定一对问题*q*和段落*p*作为in- puts，大多数MRC模型*f* (*q, p*; *θ*)试图搜索一个-。

## 扰动嵌入培训

与(Behjati等2019;Gong等2018;Sato等2018)类似，我们的扰动对抗式训练方法旨在在目标模型的监督下为每个实例训练一个扰动嵌入序列，以便分散其注意力。在训练过程中，我们将模型作为一个生成器，所有模型参数都是ﬁxed。在给定输入*eq*和*ep*的情况下，训练方法只尝试用一个额外的扰动嵌入序列来扰动每个通道输入*ep*。

根据MRC模型的定义，对于每个训练实例，我们首先将一个连续的扰动嵌入序列*e*×插入到通道嵌入序列*ep*中。因此，我们将模型输入*ep*替换为公式（2）。

*ept* = ep1 ep2 ...*epk* ⊕ *e*×*1e*×2..*.e*×*l* ⊕ *epk*+1 *...epn* (3)

其中为连词运算符，*k*为插入posi- tion指数，*l*为*e*×的长度，*e*×*i Rd*表示*e*×的第*i*个嵌入向量。

∈

⊕

为了限制*e*×的搜索空间，也为了方便进一步采样，我们将*e*×的*第i*个嵌入物，*e*×*i*定义为词汇嵌入物的加权和。

|

Σ

*e*×*i* = *wijvj* (4)

*j*=0

其中*wij*是第*i*-th位置的第*j*个词汇token的权重，而*vj*∈*Rd*是第*j*个token在*V*中的嵌入。我们将*w*∈*Rl*×|*V* |作为*e*×的权重矩阵，*wi R*|*V* |是*e*×*i*的权重向量。为了使*权重*向量*wi*在词汇空间中被归一化，我们将*wij*作为可训练参数*α*∈*Rl×*|*V* |的软最大值结果。

∈

獐子岛位于一个跨度*s*=[*ss*，*se*]中，该跨度能使模型最大化。

*exp*(*α* )

可能性。

*f* (*q, p*; *θ*)= *argmax Pr*(*s* |*q, p*; *θ*)

*w* = *ij ijexp*(*α* )

*k*

*ik*

(5)

*s*

= *argmax Pr*(*ss* |*q, p*; *θ*) *Pr*(*se* |*q, p*; *θ*)

Σ

(1)

其中*αij*是*wi*的可训练参数。这样一来，对于每个实例，训练一个扰动嵌入序列

*ss,se*

是为了找到一个合适的权重矩阵*w*的分布。

其中，*θ*表示MRC模型的参数，*ss*、*se*分别表示*s*的开始和结束位置。详细来说，*q*和*p*分别表示tq1 tq2 *...tqm*和tp1 *tp2 ...tpn的*令牌序列，其中*m*和*n*分别表示问题和段落的序列长度。

对每个实例单独训练一个*w*，用于后面的程序。基于上述设置，我们期望我们的训练方法能够生成误导性的答案文本或误导性的上下文文本。为了产生误导性的答案文本并分散MRC模型的注意力，我们设计了一个交叉熵损失------。

函数的目的是为了欺骗模型，使模型认为答案位于扰动嵌入序列中。



预测

sg 证件

梯度反推

MRC模型

e e

k

k+1

×

q

q

嵌入查询

词汇嵌入

e'

V

柔光

1 Σ

L = − log *Pr*(*y*|*e , e*× ; *θ*) (6)

*a*

2

*q*

*p*

*d*

*d*

*y*∈{*ss,se* }

其中，*sd*=[*ss*，*se*]是位于每w的分心答案跨度

*d*

*d*

湍流嵌入序列。为了产生误导性的con-文本，我们设计了一个损失函数，目的是最小化模型估计的地面真相跨度*sg*=[*ss*，*se*]，以达到

*g g*

以分心MRC模型。

1

Σ

*L* =

*c*

2

*y*∈{*ss ,se* }

*g*

*g*

log *Pr*(*y*|*e , e*× ; *θ*) (7) α

这样一来，我们将训练损失函数定义为。

*q*

*p*

*L* = *La* + *λcLc* + *Rs。* (8)

其中*λc*是*Lc*的权重。为了增加我们方法的多样性，我们在损失函数中添加了一个正则化项*Rs*，它被定义为一个相似性正则化器，以控制扰动嵌入和问题与答案之间的相似性。

*Rs* = *λqsim*(*e*×*，eq*)+*λasim*(*e*×*，ea*) (9)

其中，*ea*表示答案感的嵌入序列，它是*ep*的子序列。两种模拟的权重

图1：扰动嵌入训练方法的概述。绿色的网格是可训练变量，黄色的网格是由可训练变量和MRC模型决定的其他变量。

加权 *vj* 作为最具代表性的嵌入。结果，我们的贪婪采样方法可以简化为采样最大的加权token。§

相似性项表示为*λq*和*λa*。而*sim*( *，* )被去掉的是一个词袋余弦相似度函数。

- · ·

*i*

*sim*(e1*,* e2)=*cos*(Σ *e1i,* Σ *e2i*) (10)

*i*

*i*

*t*×*i* = *argmin EUC*(

*tti*

Σ

≈ *argmax witti*

*tti*

*wijvj, vtt* )

(13)

这样一来，我们的训练方法就会训练出一个能使损失*L*最小化的扰动em- bedding *e*× deﬁned by *w*。

*E*× = *argminL* (11)

*等*

其中*E*×是我们的目标扰动嵌入序列。

扰动嵌入训练过程概述

如图1简单说明。以给定的分心跨度*sd*和地面真相*sg*为监督信号，通过MRC模型从顶层到底层进行梯度反传播，*α*转为训练加权和嵌入*e*×，以分心模型与地面真相。而我们重复

每一个实例的训练过程，直到损失*L*是con-。

其中，我们也将*witt*的*t*×*i*视为词汇的对应索引，因为一个唯一的token在词汇中具有唯一的索引。因此，对于每个实例，生成一个

误导性文本是对最大加权令牌序列进行采样。

*i*

*A*=*t*×*1t*×2....*..t*×*l*从训练良好的*w*。

## 用对抗性例子进行再培训

为了训练一个更稳健的MRC模型，我们用采样的对抗性例子来丰富训练数据，并在丰富的数据上重新训练我们的模型。给定一个误导性文本和其相关的三重数据*q，p，sg* ，我们将误导性文本*A*重新插入到其段落的*k*位置。并且我们创建一个

()

对抗性例子(*q，p*×*，s*×*g* )与修改后的段落*p*×。

濒临或低于某个*阈值*，则返回权重矩阵*w*进行进一步的抽样。

## 贪婪的抽样

为了生成离散的误导性文本，对于*E*×的每一个位置，我们贪婪地抽取最具有代表性的*t*×*i*，他的嵌入*vtt*和*e*×*i*之间的欧氏距离最短。

*i*

*t*×*i* = *argmin EUC*(*e*×*i, vtt* )

*t i* (12)*t*

*i*

其中*s*×*g*表示插入错误引导文本后的新地真跨度。用这些逆向例子进行增强，我们在丰富的训练数据上重新训练MRC模型*θ*，得到新模型*θ*×。

## 培训战略

如算法1所示，我们的对抗性训练策略可以分为几个步骤。对于每次迭代，我们

从全训练数据集{(*q，p，sg*)}中随机抽取一个子训练集{(*q，p，sg)*}进行对抗性训练，通过

其中*t*×表示第*i*-th个嵌入物*e*×的采样令牌。

自然，我们可以把权重*wij*看成*ei*=。

*wijvj*

很容易证明，方程可以成立，当

采样函数*SubSample*(-)。对于给定的子训练

*i*×Σ*i*§

作为*vj*的重要性，并简单地取样最大的

*|v|*=1，且*witti≥*0*.*5。

**算法1：**对抗式训练策略

**输入。**三元组*q，p，sg*的训练集；对抗性例子长度*l*；一个训练良好的MRC模型*θ*；最大迭代时间*T*。

{()}

**输出。**对抗式训练模型*θt* ;

**1 当***火车损失<s且t<T时，做*

**2** {(*q，p，sg*)}×←- *SubSample*({(*q，p，sg*)})。

×

**3** {*w*}*，trainloss*←- *PertTrain*({(*q，p，sg*)} ; *θ*)。

**4** {*A*}←- *ttreedy*({*w*})。

**5** {(*q，p*×*，s*×*g* )}× ←- *Create*({(*q，p，sg*)}×*，*{*A*})。

**6** {(*q，p，sg*)}×× ←- {(*q，p，sg*)}。∪{(*q，p*×*，s*×*g* )}× 。

**7** *θt* ←- *Train*({(*q，p，sg*)}××)。

**8** *θ,* {(*q, p, s )*}。←- *θt ，*{(*q，p，s )*}××。

AddSentDiverse（ASD）（Wang和Bansal 2018）。基于对AddSent（Jia和Liang 2017）的观察，他们用相应设计的对抗性例子丰富了SQuAD训练数据。而数据集的大小达到109.4K。

### 测试组：

-•

SQuAD（DEV）（Rajpurkar等人，2016）。SQuAD v1.0的开发集，其中包含10K个用于评估的三倍*q、p、sg*实例。

()

-•

AddSent（AS）（Jia and Liang 2017）。Grammatical adver- sarial test set，其中误导性文本通过规则和众包从问题中转换。该数据集

-•

*g g*

**9结束**

**10 返回** *θt*

包含1k个问题实例。

AddAny（AA）（Jia and Liang 2017）。Ungrammatical ad- versarial test set which misleading texts are automati- cally generated according to question words and common words.数据集包含1k个问题实例。

-•

集和模型*θ*，首先，我们训练权重矩阵*w*，并通过per- turbation嵌入训练过程*PertTrain*( ; )收集其平均训练损失（*trainloss*）。第二，我们从*w*中贪婪地抽取误导性文本*A的*样本(*ttreedy*( ))，并让它们相应地插入到段落中，以创建对抗性例子(*Create*( *，* ))。第三，在给定丰富的训练数据集的情况下 *q，p，sg*××。

-·

{()}

- · ·

-·

{ } { }

- · ·

我们通过*Train*( )重新训练模型*θ*。然后我们将模型

和训练数据与增强后的数据进行后期迭代。该算法重新开始，直到*训练损失*大于阈值*s*或达到最大迭代时间*T*。

# 实验

在本节中，我们评估了我们的训练方法在基于四种不同MRC模型的不同训练和测试集上的性能。然后，我们通过研究各种对抗性例子类型的分布，讨论为什么我们的方法可能是一个更通用的方法。此外，我们还研究了各种模型集的性能影响。

## 数据集和系统

在实验中，我们分别在SQuAD(Ra- jpurkar等人，2016)训练集和其增强的ver- sion AddSentDiverse(Wang和Bansal 2018)上训练MRC模型。我们在六个不同的测试集上测试模型，即标准的SQuAD开发集和五个不同类型的advers- sarial测试集。所有的对抗性测试集都在标准SQuAD开发集的基础上附加了误导性文本作为其段落的一部分。

### 训练数据集。

SQuAD（Rajpurkar等人，2016）。最受欢迎的MRC数据集之一。该数据集由87.5K个问题-答案训练对组成，其中文档是从维基百科上转储而来，问题-答案则是通过众包来注释的。而我们选择SQuAD v1.0作为训练数据集。

-•

AddAnyExtend（AAE）。这是我们的*AddAny*扩展词汇的实现，它不仅包含问题词，还包含高频词，pass- sage词和随机常见词。数据集con- tains 2.6k问题实例。

AddAnsCtx（AAC）。一个使用答案上下文作为误导性文本的测试集，其中包括10k个实例。误导性文本是答案句子与答案标记重新移动。

-•

-•

AddNegAns（ANA）。一个使用假答案的否定预言作为误导性文本的测试集，其中包含5千个实例。我们通过将答案句子从"A is *answer*"改为"A is not *fakeanswer*"（例如，"I like this *book*" *do not* like this *movie*"）来创建误导性文本。

-•

→−

基于上述数据集，我们测试了四个不同的MRC基线系统。QANet（Yu等人，2018）、BERT（Base & Large）（Devlin等人，2018）和ERNIE2.0（Sun等人，2019）。我们分别用我们的对抗式训练方法对它们进行了训练。我们采用F1得分作为主要的评估- ation指标。

## 实验设置

在扰动嵌入训练阶段，我们随机在句子之间进行扰动嵌入，并对嵌入进行随机初始化。在嵌入训练过程中，我们设置QANet的批次大小为32，*BERTbase为*12，*BERTlarge*/ERNIE 2.0为4。我们限制扰动序列长度*l*为10。对于每个批次，我们随机设置*λq*、*λp*为-10或10，并设置*λc*为0.5。并且我们在每个扰动em-铺垫的中间设置随机长度的*sd*。为了确定嵌入训练过程的收敛性，我们将*阈值*设为1.5，并且我们将最大训练步数设为200步，因为在200步左右，最训练- ing损失趋于稳定（差异小于1e-3）。

在训练迭代中，我们设置最大训练时间*T*为5，*trainloss*的停止阈值*s*为12.0。为了

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **型号** | **DEV** | **AS** | **AA** | **AAE** | **审咨委** | **ANA** | **平均** |

### 基线系统

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| QANet | 83.3 | 48.0 | 54.8 | 63.7 | **78.1** | **79.8** | 64.9 |
| *BERTbase* | 88.4 | 49.9 | 47.8 | 52.8 | 74.4 | 69.9 | 59.0 |
| *BERTlarge* | 90.6 | 60.2 | 60.4 | 64.1 | 81.1 | 80.1 | 69.2 |
| ERNIE2.0 | 92.5 | 64.8 | 68.2 | 68.7 | 86.8 | 89.1 | 75.5 |

**相关作品**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| DCN+MINIMAL(Min等人，2018) | 80.6 | 59.7 | - | - | - | - | - |
| R.M-Reader(Hu等人，2018a) | 86.3 | 58.5 | - | - | - | - | - |
| RMR+A2D(Hu等人，2018b) | 87.9 | 61.3 | - | - | - | - | - |
| QANet+AddSentDiverse | 80.7 | **73.4** | 54.3 | 61.2 | 74.5 | 75.7 | 67.8 (+2.9) |
| *BERTbase*+AddSentDiverse | 88.3 | 80.4 | 84.9 | 61.2 | 76.9 | 73.5 | 75.4 (+16.4) |
| *BERTlarge*+AddSentDiverse(*大号*+小号) | 90.4 | 86.3 | 87.3 | 70.5 | 82.2 | 82.4 | 81.7 (+12.5) |
| ERNIE2.0+AddSentDiverse------。 | 92.2 | 89.1 | 90.0 | 74.7 | 88.3 | 88.1 | 86.0 (+10.5) |

**我们的方法**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| QANet+Ours | 82.1 | 49.2 | **66.9** | **67.3** | 77.2 | 79.3 | 68.0 (+3.1) |
| QANet+AddSentDiverse+Ours------。 | 80.8 | 72.2 | 61.9 | 63.2 | 75.3 | 76.0 | **69.7 (+4.8)** |
| BERT*基地*+我们 | 87.9 | 57.2 | 83.5 | **70.7** | 76.9 | 76.9 | 73.0 (+14.0) |
| *BERTbase*+AddSentDiverse+Ours。 | 87.8 | **81.7** | **85.2** | 70.0 | **79.2** | **79.2** | **79.1 (+20.1)** |
| *BERTLarge*+Ours | 90.4 | 63.1 | 88.3 | **76.6** | 81.6 | 81.7 | 78.3 (+9.1) |
| *BERTlarge*+AddSentDiverse+Ours(大号+小号) | 90.1 | **86.7** | **88.5** | 75.7 | **84.3** | **84.5** | **84.0 (+14.8)** |
| ERNIE2.0+我们 | 92.5 | 70.8 | **92.0** | **79.6** | 88.2 | **89.1** | 83.9(+8.4) |
| ERNIE2.0+AddSentDiverse+Ours------。 | 91.9 | **89.6** | 91.2 | 79.4 | **88.5** | 88.2 | **87.4(+11.9)** |

表2：SQuAD开发集和对抗性测试集的实验结果。所有分数均为F1分数，单位为百分比。

寻求有效性和效率之间的平衡，我们跑- domly抽样5%的训练数据进行对抗性训练，根据我们的早期实验，较大的比例将无法在单次迭代中提供满意的性能。在采样阶段，我们保存所有成功的嵌入贪婪的采样。抽样后，我们重新训练MRC模型遵循早期停止策略（Bassler等人，2010）。

为了有效地生成误导性文本，对于每一个训练实例，我们利用一个本地词汇*V*，其中的标记主要与问题和段落相关。我们通过不同的方式来收集每一个批次的本地词汇：高频词；在嵌入空间中与*q* & a的标记有短距离的前10个相似词（余弦相似度）；同义词和反义词与问题和段落的关系，从WordNet中收集（Miller 1995）；超义词和次义词同样从WordNet中收集。为了使模型更容易收敛，词汇量限制在200个。

## 结果和讨论

**实验结果** 所有的基线结果都显示在Ta- ble 2中。结果表明，基线系统在面对语法或非语法的对抗性测试集时，都很脆弱。虽然基于变压器的系统在所有测试集上都获得了较好的性能，但在所有对抗性测试集上都有很大的降低。与*BERTbase相比*，QANet的读取性能较弱，但由于其模型更简单，通常会带来更好的鲁棒性性能。

基于相同的系统，我们的训练方法在所有对抗性测试集上都显示出它的有效性。它获得了

在ERNIE2.0上的F1改进最多为23.8%，平均为8.4%。此外，它在*BERTbase*/*BERTlarge上*获得了14.0%/9.1%的更好的平均改进。由于QANet的模型设计比较简单，可能限制了它的模型能力，因此它获得的改进并不那么一致。然而，我们的方法仍然促进了它的平均表现- mance 3.1%的F1得分。在所有 adver- sarial 测试集和系统上的改进表明，它能够增强 MRC 模型的鲁棒性。

实验结果也表明，增强型训练数据ASD（Wang和Bansal 2018）在一些对抗性测试集上获得了很大的im- provement。由于该训练数据是针对问题相关的测试集（AS，AA），所以它在这些测试集上的效果非常好，符合预期。然而，它在其他测试集(AAE, AAC, ANA)上的改进，并不像它 在问题相关测试集上那样大。对于所有的基线系统，它在AAE、AAC、ANA的F1上只获得了1.7%的平均改进。这种差异表明，基于规则的方法可能无法处理所有可能的对抗性例子类型。相比之下，我们的方法在未观察到的对抗性例子类型上表现出更好的鲁棒性，它在AAE、AAC、ANA测试集上获得了平均4.7%的F1得分改进。我们的方法并没有在测试集上设置任何强如和，因此我们认为我们的方法具有更强的能力，能够以更普遍的方式加强模型的稳健性。

此外，表2所示的性能表明，我们的方法加上ASD数据集，可以进一步提高性能，在*BERTbase*/*BERTlarge*/ERNIE2.0再的F1平均提高3.7%/2.3%/1.4%。

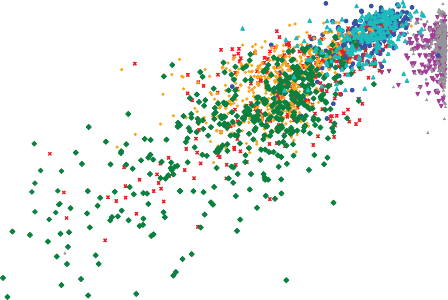


图2：不同测试集和训练样本的误导性文本分布。



tenvensiess使自己能够覆盖更多类型的对抗性例子，这可以部分解释表2中各种测试集的普遍改善。

**不同迭代的**结果 为了研究训练迭代次数对性能的影响，我们使用迭代时间在[1，14]的范围内对*BERTbase*模型进行测试，当迭代时间等于0时为基准性能，如图3所示，不断增长的性能曲线表明我们的迭代训练算法是有效的，在迭代4或5次左右就可以获得良好的性能。但在后面的迭代中，如曲线所提示的那样，可以获得有限的改进。有一个可能的解释是，由于扰动嵌入训练损失越来越大，在以后的迭代中，扰动嵌入更难欺骗模型。由于我们的误导性文本不够"真实"，不足以迷惑人类，我们相信我们的方法可以通过更好地搜索扰动嵌入和更好地采样有意义的误导性文本来进一步改进。

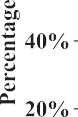


图3：不同迭代上的性能.实线表示对抗性测试集的平均性能.虚线表示在perturba- tion嵌入训练阶段的平均训练损失。虚线表示在perturba- tion embedding训练阶段的平均训练损失。

谱上。考虑到de- velopment集的上限性能约为92%，我们的训练方法可以将平均性能从86.0%提升到87.4%，带来超过20%的错误率下降。这些观察表明，我们的一般方法可以成为（Wang and Bansal 2018）的一个很好的补充，它是基于规则设计的。

误导**性例句分布分析** 为了更深入地了解不同类型的误导性例句，我们在二维坐标系中研究误导性例句的分布，其中X轴为答案句的相似性，Y轴为问题的相似性。两种相似度都是误导性文本与答案&问题之间的bag-of-words嵌入余弦模拟相似度。结果如图2所示，正如预期，ASD数据集与AS和AA有更多的重合。因此，它在这些测试集上获得了更好的改进。相比之下，我们的数据在空间上有更广泛的分布。其ex-

图4：错误预测与误导性文本之间的距离分布（以字符为单位）。如果错误预测的一部分位于误导性文本中，则距离为0。

**错误预测的距离分布** 为了研究错误预测与误导性文本之间的距离相关性，我们根据有效的对抗性例子，研究预测与误导性文本之间的距离。如图4所示，超过三分之二的错误预测定位在误导性文本中。而作为误导性语境，误导性文本更有可能将模型引导到错误答案附近。这一观察结果表明，产生误导性答案要比误导性语境容易得多，局部化表明误导性文本能够将模式的注意力吸引到附近的欺骗性"答案"上。

**消融研究** 为了研究我们模型中不同部分的有效性，我们在*BERTbase*上测试我们的方法，并去除相应的部分，结果如表4所示。根据结果，在两个损失函数之间，*La*在扰动嵌入-丁训练中起着更重要的作用。它表明，"答案"包含误导性文本在我们的方法中可能更有效。而上述实验中的Simi- lar观察也可以支持

|  |  |  |
| --- | --- | --- |
| **型号** QANet *BERTbase* | **生成的误导性文本**  *确切地说，球员也许是烧伤倾向于认为确切地关闭最好的。*  *和独特的更多的燃烧##m没有可能理论共同的路程。* | **预测**  烧毁的燃烧##m没有 |
| *BERTlarge* | *燃烧的 ##on ，因为没有理论定量的 ##t 的。* | 量化 |
| ERNIE2.0 | *不是没有理论，也不是没有想法，而是质量较轻。* | 轻质 |

表3：生成的不同模式的误导性文字。对应的问题和段落如表1所示。变换模型的例子以子词为单位（Sennrich，Haddow和Birch，2015）。

**Model 平均值**

*BERTbase* 73.0

-回答损失*拉* 64.1 -8.9

-上下文损失*Lc* 70.9 -2.1

-相似性规整器*Rs* 69.7 -3.3

表4：消融实验结果。

这个说法。对于相似性正则化器*Rs*，还原法认为，令牌分布的广泛性增强了模型的鲁棒性。

**误导性文本分析** 为了研究误导性文本的质量，我们对不同模型的误导性文本结果进行抽样分析，结果如表3所示。我们可以注意到，所有的文本都是无语法和无意义的序列，但它们成功地分散了基线系统的注意力。这说明目前的模型如*BERTlarge*和ERNIE2.0虽然有很好的语言建模能力，但仍然容易受到不可预知噪音的影响。因此，我们的生成方法仍有很大的改进空间，以使生成的文本误导性更强。

# 结论

目前最先进的MRC模型在面对不同类型的对抗性例子时是脆弱的，缺乏稳健性成为一个严重的问题。由于列举和研究所有可能的对抗性例子是不切实际的，通过投资- tigating其弱点和加强它们来提高MRC模型的鲁棒性是一个更有效的方法。因此，我们提出了一种简单而有效的对抗性训练方法，通过在生成的对抗性实例上训练这些模型来提高MRC模型的鲁棒性。通过三步算法，我们的方法通过在对抗性例子生成器和模型训练器之间进行最小-最大游戏来训练一个健壮的MRC模型，其中生成器就是模型本身。基于强大的基线，在各种测试集上的实验表明，我们的新型ap- proach可以以更普遍的方式大幅提升MRC模型的鲁棒性per- formance。此外，结合基于规则设计的增强型训练数据，我们的方法能够进一步提高模型的鲁棒性，并超越起步阶段的性能。

# 鸣谢

这项工作得到了国家重点研发计划项目（编号：2018AAA0101900）、国家自然科学基金项目（编号：61533018和。

No.61672440）和百度-北京大学联合项目。同时感谢吕胜和匿名审稿人对稿件的建设性批评。

# 参考文献

Alzantot，M.；Sharma，Y.；Elgohary，A.；Ho，B.-J.；Sri- vastava，M.；和Chang，K.-W.2018.Generating natural language adversarial examples.*arXiv preprint arXiv:1804.07998*.

Bassler, D.; Briel, M.; Montori, V. M.; Lane, M.; Glasziou, P.; Zhou, Q.; Heelsansdell, D.; Walter, S. D.; Guyatt, G. H.; and Group, A. S. 2010.尽早停止随机试验以获得效益和估计治疗效果。系统性审查和元回归分析。*Jama the Journal of the American Medical Association* 303(12):1180.

Behjati, M.; Moosavi Dezfooli, S. M.; Soleymani Baghshah, M.; and Frossard, P. 2019.Universal adversarial attacks on text classiﬁers.

Devlin，J.；Chang，M.；Lee，K.；和Toutanova，K. 2018。BERT：用于语言理解的深度双向变换器的预训练。*CoRR* abs/1810.04805.

Gong，Z.；Wang，W.；Li，B.；Song，D.；和Ku，W.-S.2018.Adversarial texts with gradient methods.*arXiv preprint arXiv:1801.07175*.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014.Explain- ing and harnessing adversarial examples.*arXiv preprint arXiv:1412.6572*.

Hu, M.; Peng, Y.; Huang, Z.; Qiu, X.; Wei, F.; and Zhou, M. 2018a.Reinforced memonic reader for machine reading comprehension.In *Proceedings of the Twenty-Seventh Inter- national Joint Conference on Artiﬁcial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, 4099–4106.

Hu, M.; Peng, Y.; Wei, F.; Huang, Z.; Li, D.; Yang, N.; and Zhou, M. 2018b.Attention-guided answer distilla- tion for machine reading comprehension.In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2077-2086.Brussels, Belgium:As- sociation for Computational Linguistics.

Iyyer, M.; Wieting, J.; Gimpel, K.; and Zettlemoyer,

L.2018.Adversarial example generation with syn- tactically controlled paraphrase networks.*arXiv preprint arXiv:1804.06059*.

Jia, R., and Liang, P. 2017.Adversarial examples for eval- uating reading comprehension systems.In *Proceedings of the 2017 Conference on Empirical Methods in Natural Lan- guage Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, 2021-2031.

Kusner, M. J., and Herna´ndez-Lobato, J. M. 2016. Gans for sequences of discrete elements with the gumbel-softmax distribution.*arXiv preprint arXiv:1611.04051*.

Liu, X.; Shen, Y.; Duh, K.; and Gao, J. 2018.Stochastic an- swer networks for machine reading comprehension.In *Pro- ceedings of the 56th Annual Meeting of the Association for Computational Linguistics（Volume 1：Long Papers）*，1694- 1704.Melbourne, Australia:Association for Computational Linguistics.

Miller, G. A. 1995。Wordnet。英语词汇数据库。

*Commun.ACM* 38（11）：39-41。

Min, S.; Zhong, V.; Socher, R.; and Xiong, C. 2018.Efﬁ- cient and robust question answering from minimal context over documents.In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Pa- pers*, 1725-1735.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016.Squad:100，000+问题的机器理解文本。In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, 2383-2392.

Salant, S., and Berant, J. 2018.Contextualized word repre- sentations for reading comprehension.In *Proceedings of the 2018 Conference of the North American Chapter of the As- sociation for Computational Linguistics:Human Language Technologies，Volume 2（Short Papers）*，554-559.New Or- leans, Louisiana:Association for Computational Linguis- tics.

Sato, M.; Suzuki, J.; Shindo, H.; and Matsumoto, Y. 2018.输入嵌入空间中文本的可解释对抗性扰动.*arXiv preprint arXiv:1805.02917*.

Sennrich, R.; Haddow, B.; and Birch, A. 2015.Neural ma- chine translation of rare words with subword units.*arXiv preprint arXiv:1508.07909*.

Seo, M. J.; Kembhavi, A.; Farhadi, A.; and Hajishirzi, H. 2016.Bidirectional attention ﬂow for machine comprehen- sion.*CoRR* abs/1611.01603.

Sun, Y.; Wang, S.; Li, Y.; Feng, S.; Tian, H.; Hua, W. and Wang, H. 2019.Ernie 2.0:A continual pre-training frame-work for language understanding.*ArXiv* abs/1907.12412.

Wang, Y., and Bansal, M. 2018.Robust machine com- prehension models via adversarial training.In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics:Human Lan- guage Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, 575-581.

Wang, W.; Yang, N.; Wei, F.; Chang, B.; and Zhou, M. 2017.Gated self-matching networks for reading comprehension and question answering.In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 189-198.

Yu, L.; Zhang, W.; Wang, J.; and Yu, Y. 2017.Seqgan:Sequence generative adversarial nets with policy gradient.In *Thirty-First AAAI Conference on Artiﬁcial Intelligence*.

Yu, A. W.; Dohan, D.; Luong, M.; Zhao, R.; Chen, K.; Norouzi, M.; and Le, Q. V. 2018.Qanet:Combining lo- cal convolution with global self-attention for reading com- prehension.In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

张刚;严志明;季新华;张婷婷;徐。

W. 2017.Dolphinattack。Inaudible voice commands.In *Proceedings of the 2017 ACM SIGSAC Conference on Com- puter and Communications Security*, 103-117.ACM.SIGSAC Conference on Com- puter and Communications Security, 103-117.