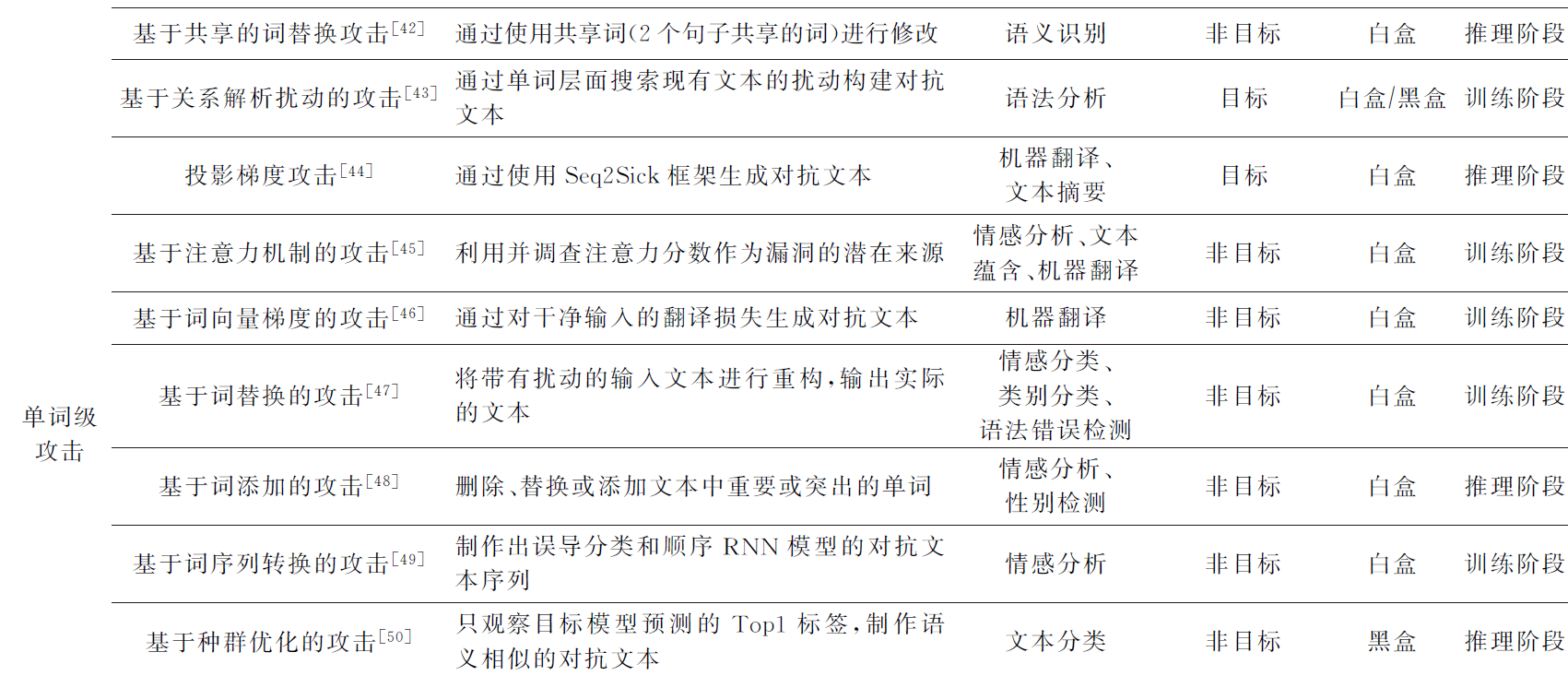
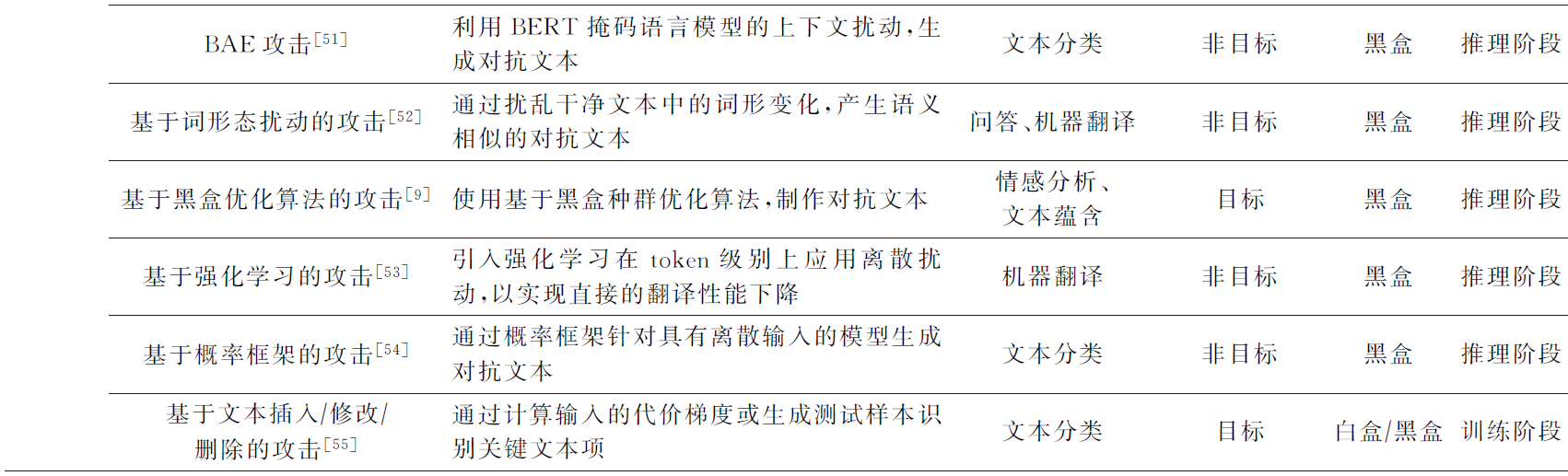
NLP攻防技术

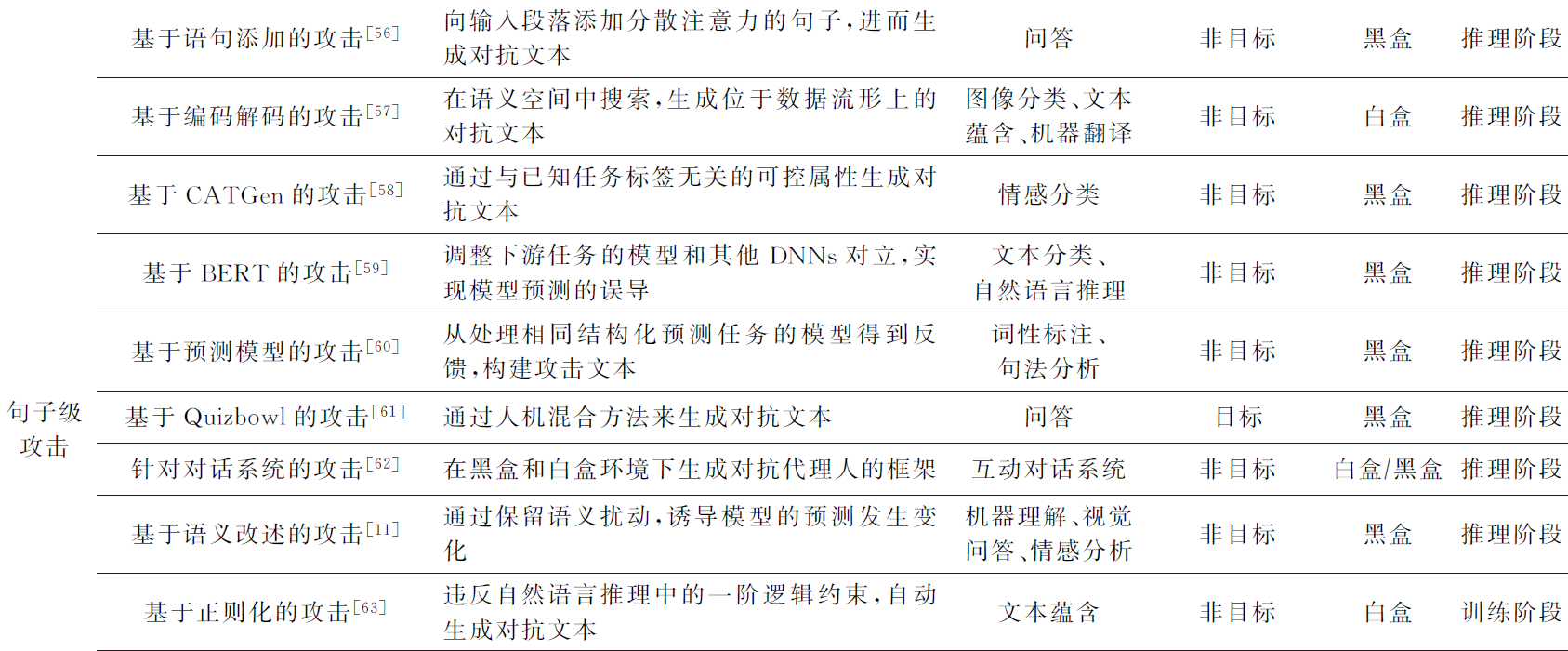
1. 针对NLP的攻击和防御技术:

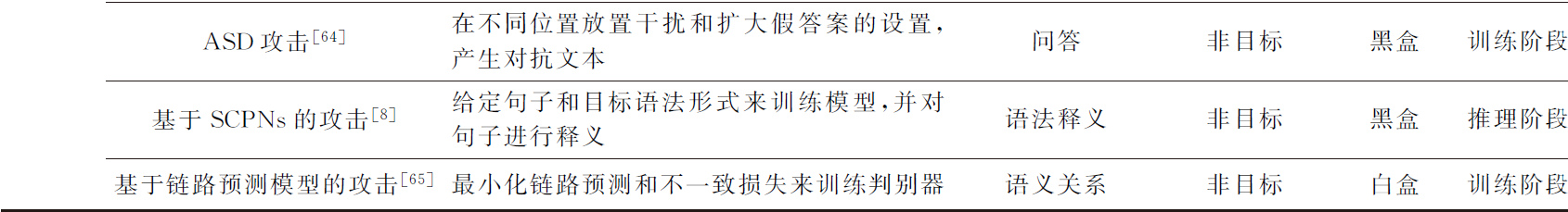
针对NLP的攻击技术如下表所示。



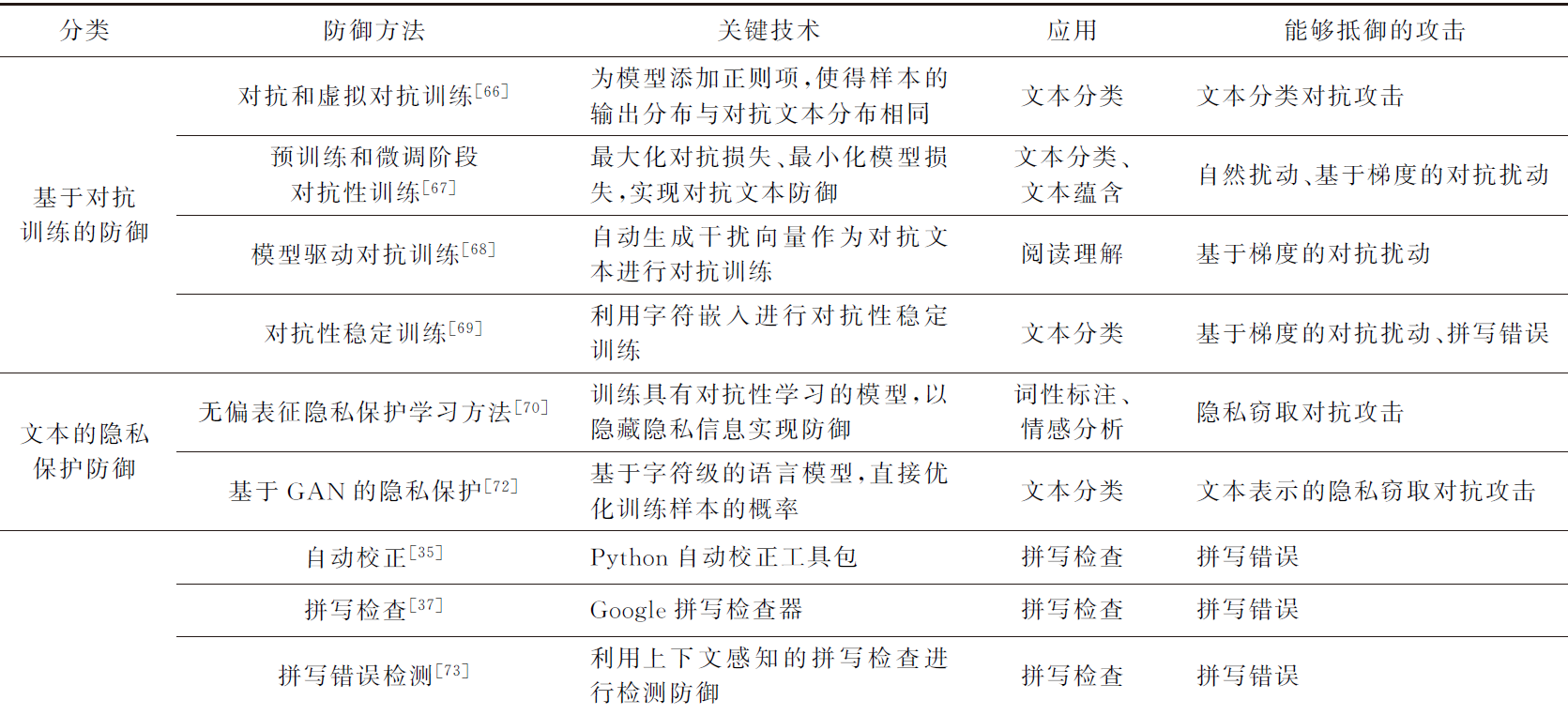


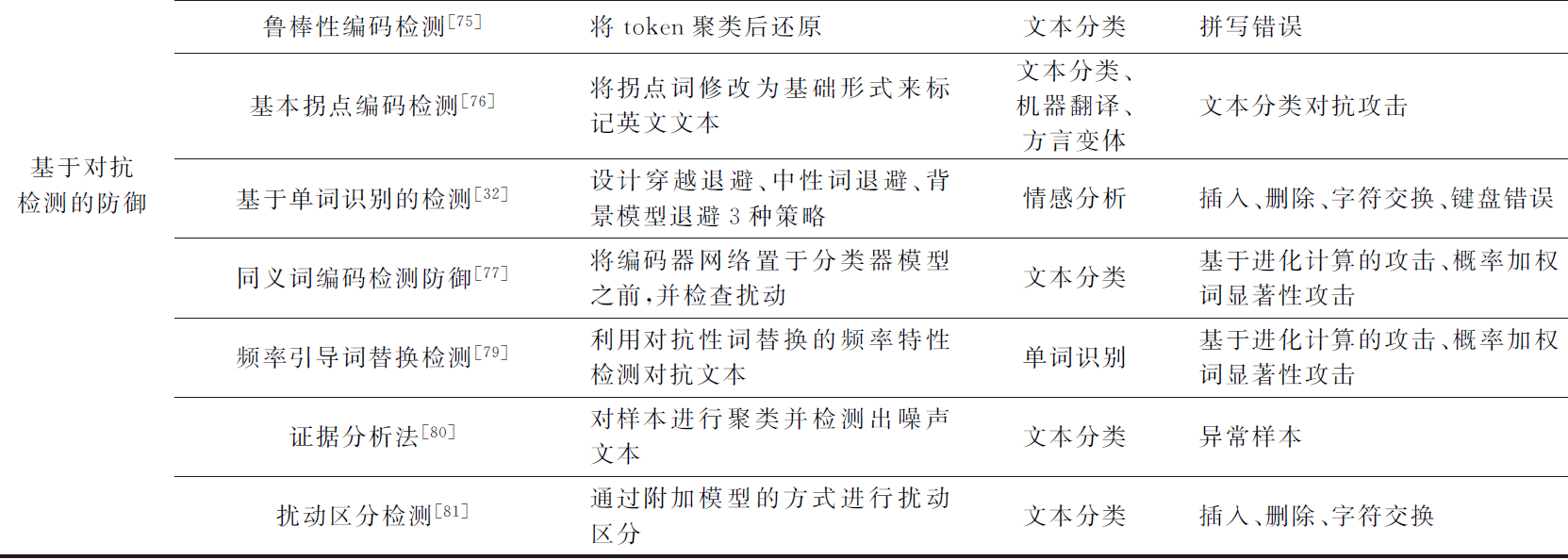






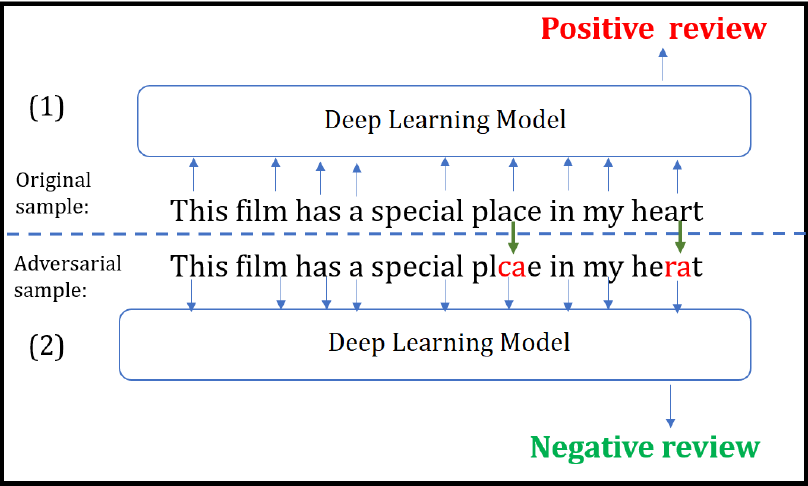
面向NLP的防御技术如下表所示。





1. 面向NLP的攻击技术:
2. DeepWordBug攻击[2]:

DeepWordBug可以在黑盒环境中有效生成微小的文本扰动，从而迫使深度学习分类器对文本输入进行错误分类。文中采用新的评分策略来识别关键标记，如果发生修改就会导致分类器做出错误的预测。通过在排名最高的标记上运用简单的字符级转换以最小化编辑距离，同时更改原始的分类结果。由于在离散输入中使用范数距离来限制修改的大小存在以下两种限制: (1)文本输入是符号，从而很难定义在上的扰动; (2)没有度量文本差异的指标，虽然范数对连续像素值有意义，但是对离散的文本是没有意义的。由于这些原因，对抗性文本攻击不同于连续输入情况下的对抗性样本攻击。要使用文本对抗性修改，需要定义一个类似于范数的距离来限制对离散文本的修改幅度。一个可行的方法是定义文本和文本之间的编辑距离，该距离被定义为将更改为所需要的最小编辑操作数。通过对文本序列进行微小的编辑操作来设计对抗函数序列的评分函数，会使人们认为它们和原始序列很相似，但微小的变化其实会产生与原来的词语有席位差别的对立词语，即对抗样本。因此，文中首先针对文本序列中的重要词语进行标记，然后对这些标记进行修改，这就可以有效的误导深层分类器做出错误的决策。文中定义的对抗序列的示例如下图所示。通过深度RNN模型将原始文本输入值正确分类为积极评论，然而，通过只改变几个字符的位置，生成的对抗样本序列就让整个深层分类器做出了错误的分类(图中被误分类为负面评论)。



由于在符号文本输入上很难定义梯度的概念而且由于模型参数不可见也很难计算梯度，因此文中该设计了一种方法来直接在输入序列上生成对抗性修改，而不需要梯度的指导。考虑到复杂多变的搜索空间，文中设计了在黑盒环境下生成对抗样本的两步法: (1)判断需要改变的重要序列; (2)轻微修改这些序列，然后创建不易察觉的改动可以避开目标深度学习分类器。

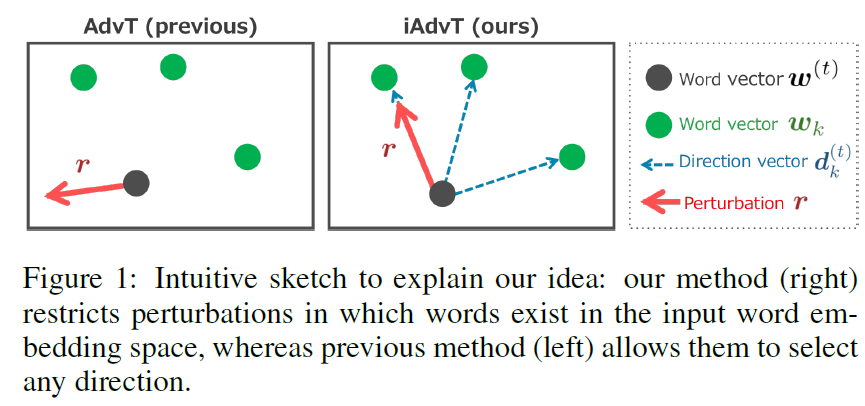
为了找到重要序列，文中设计了评分函数来评估哪些序列对目标模型做出决策是重要的。这些评分函数用于确定每个单词对最终预测的重要性。在确定了这些重要策略过后，在使用一个简单强大的算法来修改这些序列，从而生成一个对抗样本，即: 交换、替换、删除、插入方法。其中，出交换操作的编辑距离为2外，其余操作的编辑距离为1。

DeepWordBug算法有以下特性: (1)黑盒: 以前的方法需要了解单词的嵌入层的模型结构和参数，而文中的算法可以在黑盒环境下进行攻击; (2)有效: 该算法可以比当前最先进的基准算法更能成功地误导深度RNN模型，而且，通过误导一个深度模型产生的对抗性示例还可以成功避开类似的模型; (3)简单: WordBug使用简单的字符级替换来生成对抗序列，这比以前使用投影梯度等方法操作更简单易行; (4)对人类而言观察到的干扰很微小: WordBug可以生成看起来很像种子序列的对抗样本。

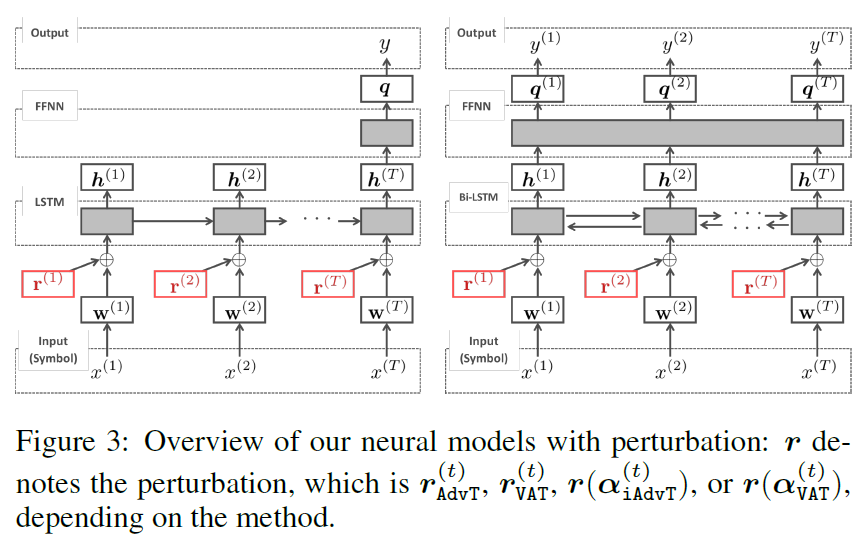
1. 基于词替换的攻击[3]:

由于文本输入是由符号组成的，离散符号不属于图像处理中使用的连续空间，因此无法计算NLP字段任务中任务的扰动输入，所以不能将对抗训练方法(AdvT)直接应用于NLP任务之中。最近提出一种改进NLP任务的AdvT策略: 将AdvT应用于连续的单词嵌入空间，而不是输入文本的离散空间。这种方法保留了图像处理领域发展起来的理论背景知识，显著提高了性能。另一个优点是体系结构简洁，只需要使用损失函数的梯度来获得对抗性扰动，无序加入任何额外的复杂体系结构，计算成本非常小。但是，这种方法存在的主要缺点是: 由于需要在实际文本的输入词嵌入空间中适当地重构输入的干扰，因此它放肆了生成可解释性的对抗性示例。因此，只能在形成良好和低成本方法与NLP领域的AdvT方法的可解释性之间进行权衡。

文中的主要目标是缩小这种权衡差距，在保持正则化的良好性能的同时还要恢复模型的可解释性。主要思想是只限制扰动的方向，即在单词嵌入空间中现有单词的位置。通过这种方法，可以直接将带有扰动的每个输入解释为一个实际的句子，将扰动视为句子中单词的替换。一个示例如下图所示。



文本对抗训练(AdvT-Text)的模型结构如下图所示。

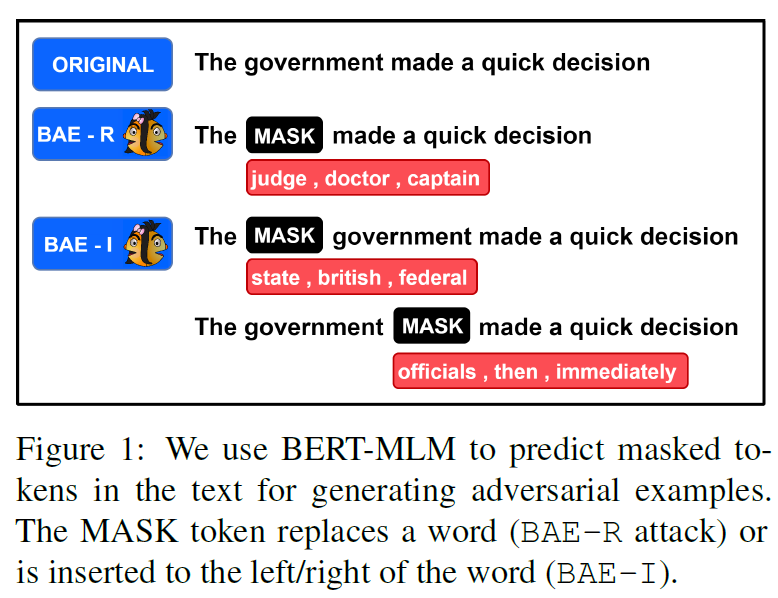


文中通过扩展AdvT-Text模型，期望在保持模型性能的同时恢复生成可解释性对抗性样本的能力。文中仅限制嵌入空间中的扰动方向指向输入单词嵌入空间中的现有单词，指向其他单词的方向可以解释为替换句子中的另一个单词，这可能会重构对抗样本。将这种扩展的AdvT-Text称为可解释的AdvT-Text或iAdvT-Text。文中采用半监督学习的方式来完成这个任务，通过扩展已有的虚拟对抗训练(VAT)方法，将其应用到NLP领域中，称为iVAT-Text方法。iAdvt和iVAT这两种方法可以成功地生成合理的对抗性文本和输入嵌入空间中扰动的可解释性。

1. BAE攻击[4]:

NLP最近工作使用的基于规则的同义词替换策略生成对抗性示例，这些策略可能导致脱离上下文和非自然地短语替换，很容易被人类识别。文中介绍的BAE是一种黑盒模型攻击策略，使用Bert模型的蒙板策略产生上下文干扰从而生成对抗性样本。BAE通过屏蔽部分文本并利用BERT-MLM来生成屏蔽词语的替代方案，在原始文本中替换和插入词语。BAE除了生成的对抗样本有更好的语法性和语义连贯性外，还具有更强的攻击能力。

由于(1)输入空间的离散型和(2)需要确保与原始文本的语义一致性，NLP中的对抗性样本生成比计算机视觉任务中的更具挑战性。基于梯度或基于生成器模型的方法在NLP中生成对抗性示例的一个主要瓶颈是扰动从连续词嵌入空间向后传播到离散词语空间。文本攻击模型的初始工作依赖于在字符级别引入错误或添加和删除单词，以创建对抗性样本。这些方法通常会导致看起来就不自然的对抗性样本，却反语法正确性，因此很容易被人类识别。虽然同义词替换可以缓解这个问题，但是依然存在语法错误的问题，而且无法处理一词多义的情况。文中提出的BAE（基于Bert的对抗性攻击），是一种使用Bert蒙板的语言模型(MLM)进行单词替换的新技术，可以更好地适应英语中的整体语境。除了替换单词外，文中还建议在句子中插入新的标记以提高BAE的攻击强度。输入句子中的这些干扰时通过屏蔽输入中的一部分词语并使用LM来填充蒙板实现的。如下图所示。

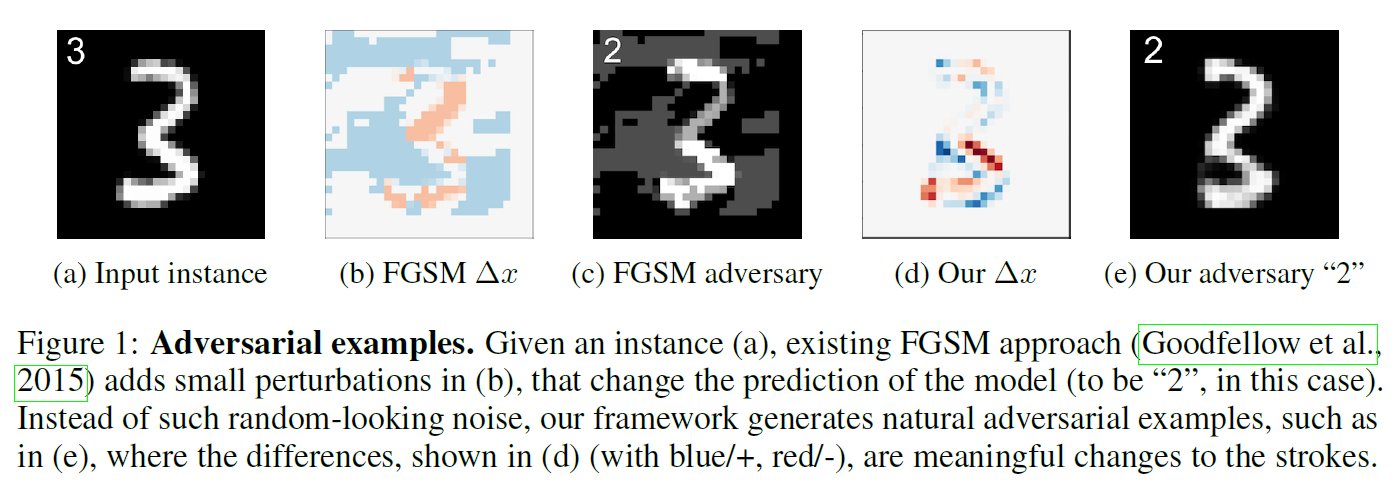


基于替换(R)和插入(I)操作，文中为BAE提供了4种攻击模式，其中对于输入中的每个词语，可以(1)替换词语: BAE\_R; (2)在标记的左侧或右侧插入词语: BAE-I; (3)替换标记或在标记的左侧或右侧插入标记: BAE-R/I; (4)首先替换标记，然后再插入标记到左侧或右侧位置: BAE-R+I。

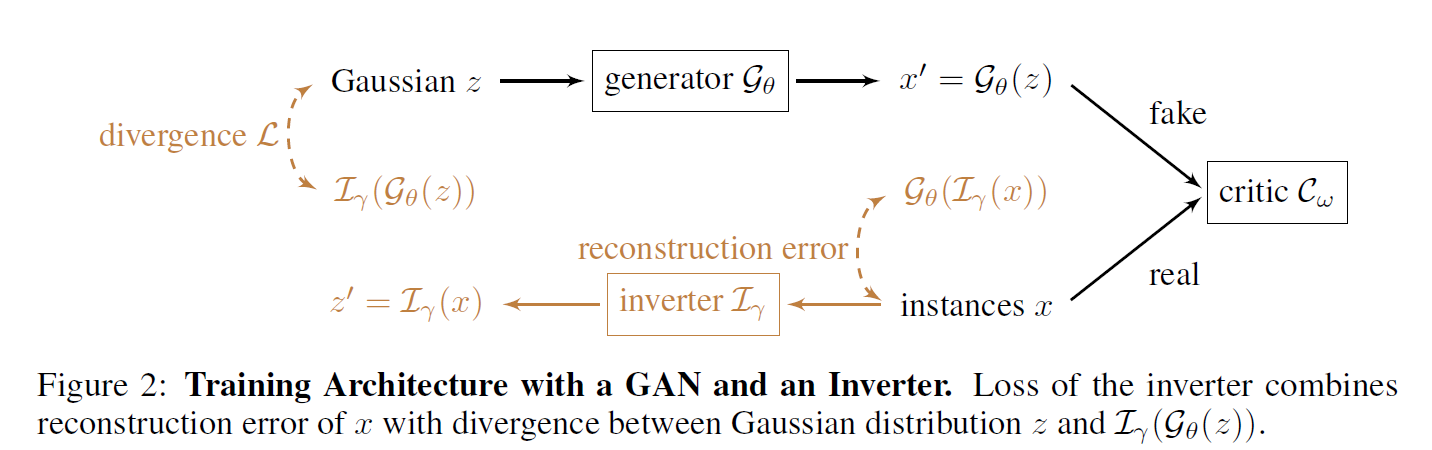
1. 基于编码解码的攻击[5]:

对抗性样本，即带有微小扰动的输入，导致了实质性不同的模型预测，有助于通过暴露失败的对抗性场景来评估这些模型的稳健性。然而，这些恶意干扰通常是不自然的，再语义上没有意义，而且不适合复杂的领域，如语言。文中提出一个框架，通过在密集连续数据表示的语义空间中搜索，利用生成性对抗网络的最新进展，生成位于数据流形上的自然易读的对抗性样本。

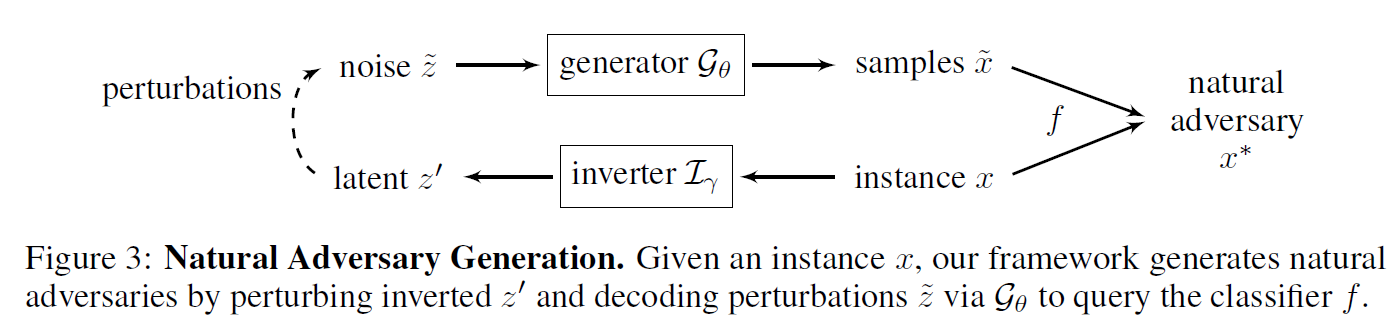
尽管已有的一些对抗样本生成方法暴露了机器学习中的“盲点”，但是，这些样本是不自然的，也就是说，这些最坏情况下的扰动实例不是分类器部署时可能遇到的，因此，很难对黑箱分类器内的基本决策行为获得有用的见解。文中的框架旨在生成自然的对抗性样本，即有意义的，有效的，易读的并且有助于解释的对抗样本。这种方法是在密集且连续的数据表示中寻找敌手，而不是直接在输入数据空间中搜索。文中使用生成性对抗网络(GANs)学习将正态分布的固定长度的向量映射为数据实例的投影。给定一个输入实例，模型通过在递归收紧的范围内采样，在其对应的隐式空间表示的邻域之中搜索敌手。一个例子如下图所示，通过增加差异(图1d)，例如将底部的笔划稍微加厚(蓝色)和让上面的笔划变薄(红色)会误导分类器将数字3误分类为数字2，但是数字3(图1e)看起来仍然像一个手写数字3，且与原始输入看起来很像。



这种方法可以用于图像和文本领域，生成更自然、语法更准确、语义更接近输入、有助于解释黑盒模型的局部行为的敌手。给定一个黑盒分类器和一个未标记的数据集，模型的目标是生成对抗样本来误导实例的预测结果，即。同通常，实例可以不属于中，但需要来自相同的分布中，同时，也需要满足这个分布，这样可以保证的多样性。不同于其他直接在输入空间中搜索敌手的方法，文中的方法提出在一个对应的密集表征空间中搜索。也就是说，不是直接寻找，而是在一个满足分布的隐式的密集向量空间中寻找一个对抗样本，然后通过生成对抗网络模型将映射为。通过在隐式低维向量空间中搜索对抗样本并将它们映射到空间中，这些对抗样本是有效的且在语义上接近于原始输入。模型的训练框架如下图所示。



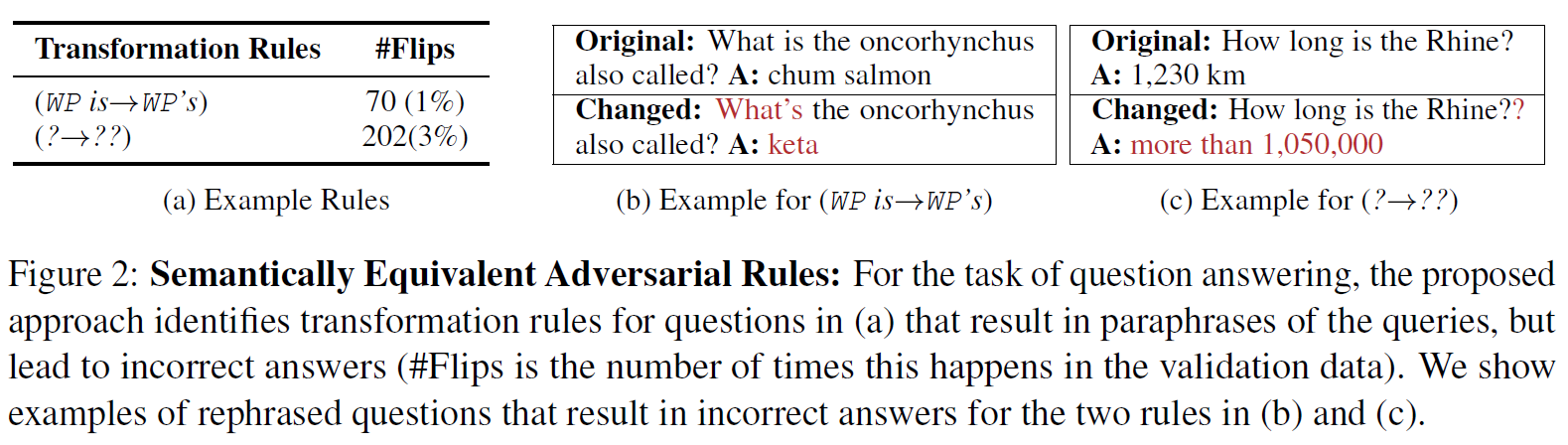
通过使用搜索算法，生成对抗样本的流程如下图所示。



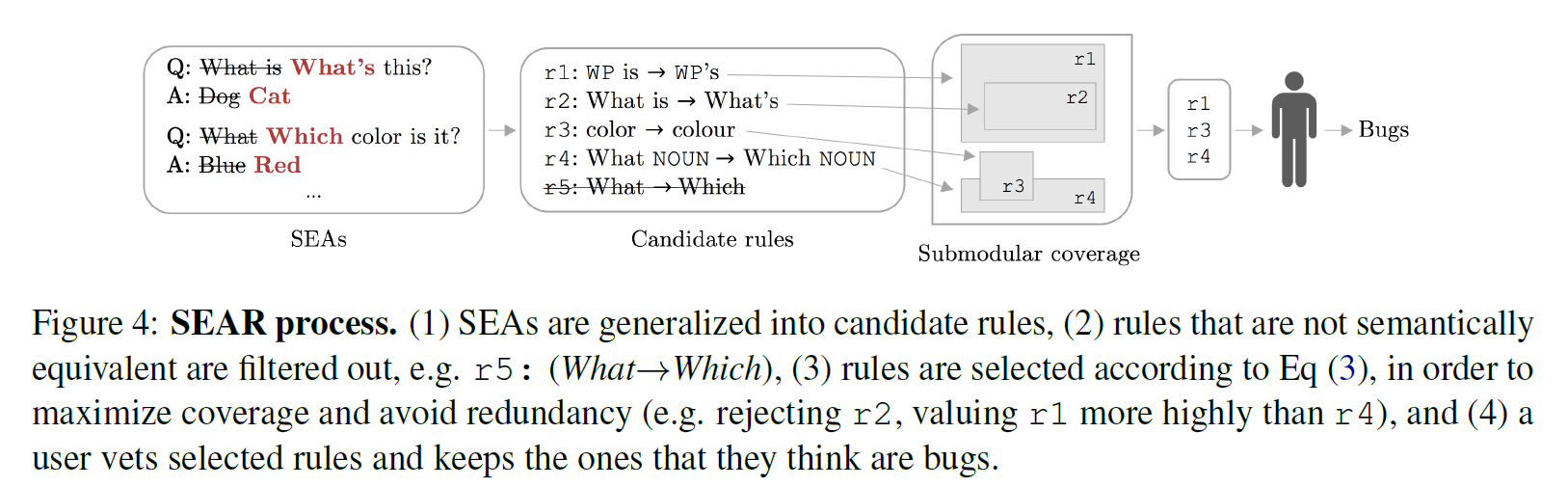
1. 基于语义改进的攻击[6]:

复杂的NLP机器学习模型通常很脆弱，对语义极其相似的输入实例会做出不同的预测。为了自动检测单个实例的这种行为，文中提出语义等价敌手(SEA)，可以保留语义扰动，导致模型预测的变化。文中将这些敌手概括为语义等价的敌手规则(SEARs)，是一种简单通用的替换规则，很多情况下会诱导敌手。而且，SEARs也是可操作的，可以使用数据扩充模型进行再训练来显著减少失败率，提高准确性。

受到图像对抗性样本的启发，文中引入语义等价敌手(SEA)。让文本输入以保持语义的方式受到干扰，但会导致黑盒模型的预测发生变化。系统地生成此类对抗性样本可以极大地帮助调试ML模型，因为它允许用户检测现实世界中发生的问题，而不是过于敏感地恶意攻击。SEA描述了针对特定预测的局部脆弱性，当考虑模型的全局缺陷时，可以通过简单的替换规则来表示这些缺陷，这些规则会在多个预测中产生SEA，文中称这种规则为语义等价的对抗性规则(SEAR)。文中介绍的SEAs和SEARs旨在揭示NLP模型中的局部和全局过度敏感缺陷。文中首先提出一种基于语义生成技术生成语义等价敌手的方法，即模型不可知时的方法，适用于任何黑箱模型。然后，将SEA推广到语义等价的规则中，并概述了最优规则集的性质：语义等价、高敌手数和非冗余。将这种寻找此类规则集的问题定义为子模块优化问题，从而得到一个准确且高效的算法。同时，SEARs可以帮助用户发现不同任务的各种最先进模型的重要缺陷。规则集的形式即示例如下图所示。



SEAR的过程如下图所示。



最优规则集的性质为: (1)语义等价: 应用集合中的规则应产生语义等价的实例，这相当于考虑应用时有很高概率产生语义等价实例的规则; (2)高敌手数: 集合中的规则应在验证数据中归纳出尽可能多的SEA，而且，每个归纳出的SEA应该具有尽可能高的语义相似性得分; (3)非冗余: 集合中的不同规则可能会导致相同的SEA，或者可能会导致相同实例的不同SEA，理想情况下，集合中的规则应该在验证中覆盖尽可能多的实例，而不是关注一小部分脆弱的预测，而且规则不应该对用户重复。

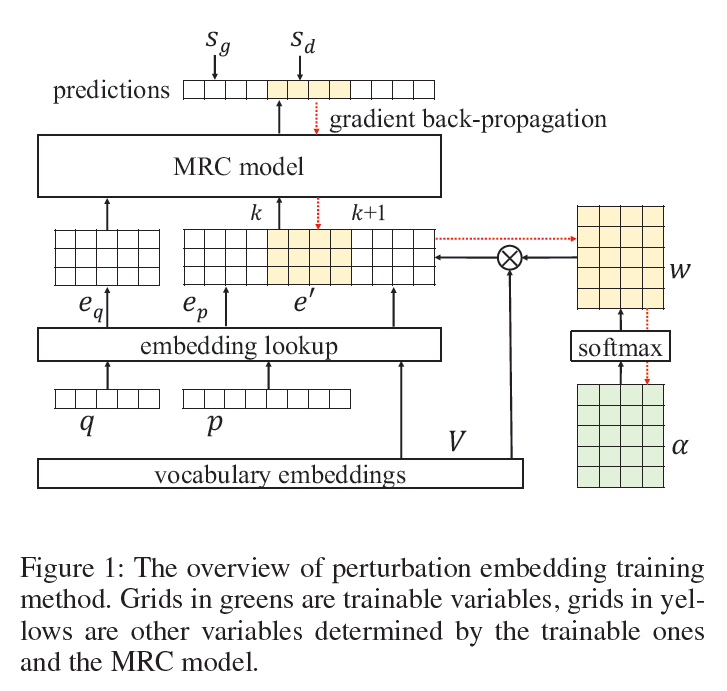
1. 面向NLP的防御技术:
2. 模型驱动对抗训练[7]:

缺乏鲁棒性是机器阅读理解(MRC)模型的一个严重问题，为了缓解这个问题，最有希望的方法之一是使用复杂设计的对抗性样本来扩充数据集。通常，这些样本是根据观察到的成功的对抗性攻击模式按规则创建的。由于对抗样本的类型数不胜数，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现实的。因此，文中提出了一种新的鲁棒性对抗训练方法，以更通用的方式提高MRC模型的鲁棒性。给定在原始数据集上训练良好的MRC模型，文中的方法基于模型的参数动态生成对抗性样本，并通过在迭代计划中使用生成的样本进一步训练模型。

为了解决模型的稳健性问题，目前最直接和有效的方法是使用对抗性样本扩充训练数据集，但这些增强的数据集虽然能够模拟已知类型的对抗性样本攻击，但忽略了其他未观察到的攻击类型。也就是说，虽然增强训练数据集有助于防御特定的对抗性攻击，但是在其他类型的对抗性攻击中仍然失败。因此，文中认为基于规则的数据扩充方法是不够有效的，因为对抗性样本的类型数不胜数。为了解决这个问题，文中提出了一种模型驱动的方法来生成能够攻击给定MRC模型的对抗性样本。然后，使用生成的对抗性样本来对MRC模型进行重新训练和增强。该方法的主要优点是不需要任何对抗性攻击类型的范式，但可以在一般的对抗性攻击下更加健壮。

这种模型驱动的方法主要分为三个步骤: (1)将MRC模型作为一个黑箱，从原始数据集中逐次抽取实例，得到一个扰动词嵌入序列。扰动词嵌入序列很可能会导致MRC模型给出错误的预测。(2)从每个扰动词嵌入序列中抽取一个单词的序列，然后将抽取的序列作为误导性样本，插入到原始样本中，创建一个对抗性样本。(3)训练MRC模型来最大化真实结果的概率，从而抵抗这些对抗样本的攻击。然后，重复步骤(1)~(3)，再次训练模型直到模型收敛。通过这种方式，最终的模型将可以抵御更多的攻击，而不是几种特定类型的攻击。

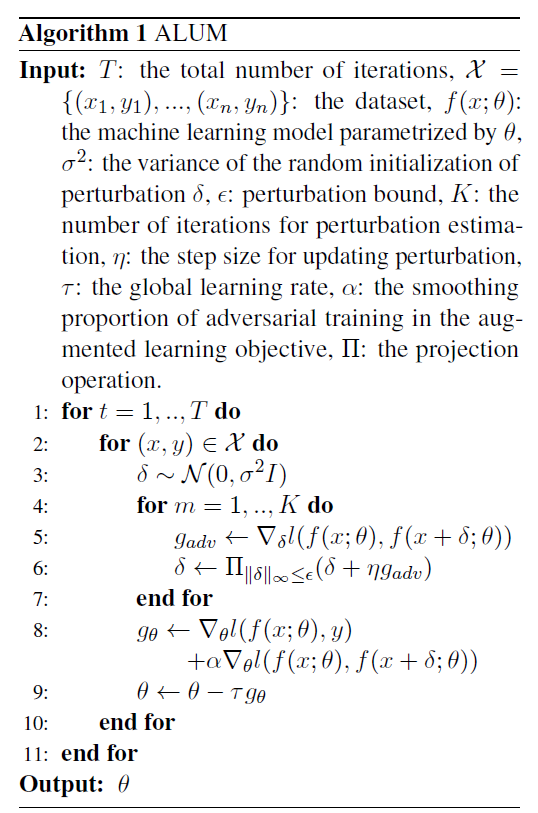
与生成对抗网络(GAN)类似，文中提出的对抗性训练方法在对抗性样本生成器和相应的MRC模型之间进行min-max博弈。由于离散词语的生成是一个不可微过程，因此选择一种抽样策略来生成对抗性样本。训练过程为: (1)使用训练好的MRC模型作为对抗性生成器，训练扰动嵌入序列使其最小化给定问题和段落下真实答案的输出概率。(2)贪婪地从扰动嵌入中抽取单词序列作为误导性文本，以创建和丰富我们的对抗性样本集。(3)对MRC模型进行训练，是真实答案的概率最大化，以抵御这些对抗性样本。然后，返回步骤(1)，重新训练模型成为新的生成器，直到收敛。为了充分涵盖潜在类型的对抗性样本，模型试图生成两种误导性文本: (1)误导性答案文本，试图证明正确答案位于文本中的MRC模型; (2)误导性上下文文本，试图分散MRC模型对正确答案的注意力，并将其引导至错误答案。扰动词嵌入训练过程如下图所示。



1. 预训练与微调阶段对抗性训练[8]:

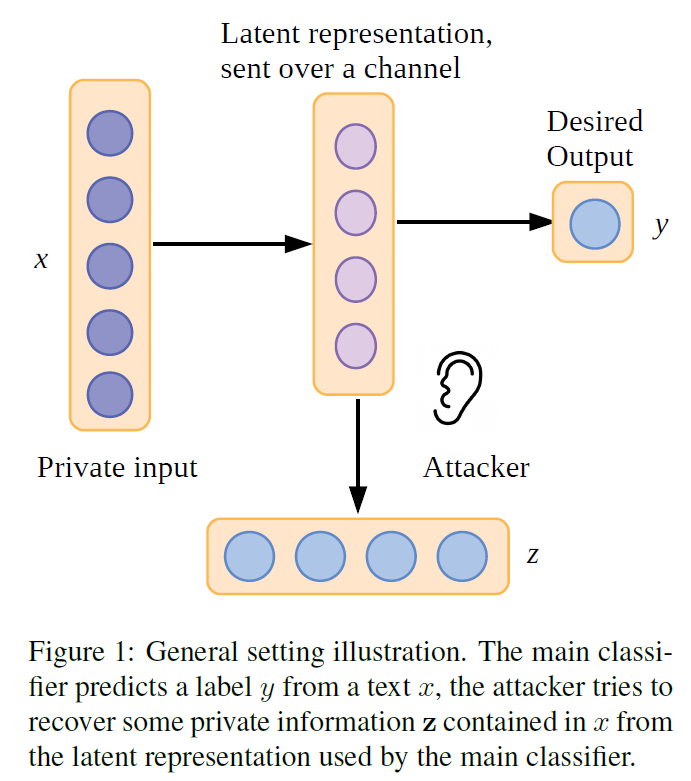
泛化和鲁棒性都是设计机器学习方法的关键要求，对抗性训练可以增强鲁棒性，但是过去的工作发现这会损害泛化能力。在自然语言处理(NLP)中，预训练大型神经网络在各种任务的泛化能力很好，并可以从对抗性微调中得到进一步的改进。然而，这些模型仍然容易受到敌对攻击。文中证明了对抗性预训练可以提高模型的泛化能力和鲁棒性，因此提取一种通用算法ALUM(大型神经语言摩西那个的对抗训练)，该算法可以通过在嵌入空间中应用最大对抗损失函数的扰动来调整训练目标。ALUM可以进一步与特定任务的微调结合，从而干扰获得额外的收益。

文中首次对对抗性预训练进行了全面的研究，表明其可以提高NLP任务的泛化能力和鲁棒性，在此基础上提出一种统一的算法ALUM，增加了一个额外的项，通过在嵌入空间中应用扰动来最大化对抗损失。ALUM通常适用于任何基于Transformer的语言模型之上的预训练和微调。它基于几个在先前工作中证明有效的关键思想。首先，不直接对输入文本应用扰动，而是扰动嵌入空间。其次，使用虚拟对抗性训练来规范标准目标。文中发现虚拟对抗训练优于传统的对抗训练，尤其是当标签中可能有噪声时。受ERNIE和其他持续训练方法的影响，本文采用课程学习的方法: 首先使用标准目标来训练模型，然后使用虚拟对抗训练进行训练。ALUM算法如下所示。



1. 基于GAN的隐私保护[9]:

文中刚讨论了在隐私保护的背景下，针对自然语言处理(NLP)深度学习系统的对抗性攻击，并研究一种特定类型的攻击: 攻击者窃听神经文本分类器的隐藏表征，并试图恢复有关输入文本的信息。当神经网络的计算在多个设备之间共享时，例如一些隐藏表示由用户的设备计算并发送到基于云的模型时，可能会出现这种情况，如下图所示。文中通过攻击者准确预测特性私有信息的能力来衡量隐藏表示的隐私信息，并描述隐私和神经网络表示的效益之间的权衡。



一些私有信息与输出标签相关，将有网络学习。这种情况下，网络的准确性与隐私之间就存在权衡。为了满足隐私需求，可能需要牺牲一些准确性。但并不是所有的私有信息都是如此，因为其中有一些信息与文本标签的预测无关。不过，私人信息可能时偶然获得的。这种非接触性和偶然性的学习也会引起隐私问题，因为攻击者能够访问隐式表征并利用它们来恢复有关输入的信息。文中讨论了以下两种情况: (1)主分类器使用深度神经网络从文本数据预测标签; (2)攻击者窃听网络的隐藏层，并试图恢复有关未看到样本的输入文本的信息。文中提出不保护训练集中样本的隐私，只保护用户提供的不可见样本的隐私。

文中提出了针对以上对抗性攻击的三种防御方法，前两种方法基于两个具有竞争目标函数的神经网络，最后一种方法旨在阻止模型将具有类似私有变量的训练样本聚集在一起。文中将著分类器的训练框架分为两部分: 著代理和对抗生成器。生成器从隐式表征中重构实例，而主代理学习主要任务并使生成器的任务变得困难。三种方法具体如下:

1. 预测二值属性的分类器: 为了不使输出结果成为输入重建的良好表示，文中对主模型的训练设置进行了两次修改。首先，使用一个重复的对抗性分类器，它与主分类器同时训练，训练样本是动态生成的，并且随着主分类器更新自己的参数而改变。该分类器用来模拟训练期间的攻击。然后，修改主分类器的目标函数，以便在敌对分类器擅长重建时加入惩罚，也就是说，主分类器试图更新其参数来混淆重复攻击者。
2. 直接优化训练样本可能性的基于字符的语言模型: 基于字符的LSTM语言模型，该模型通过训练以重构完整的训练样本。这种防御方法的一个特性是其并不知道私有变量是什么，因此，有可能保护神经网络的表征免受任何私有信息的攻击。这种方法的目标是将隐藏表征作为主要任务，并避免学习到任何与之无关的任务。
3. 分散策略: 基于这样一个直觉：当主模型隐式地学习到在表征空间的相同区域中聚集的具有相似属性的样本时，私有变量更容易预测。为了避免这种隐式的聚类，文中挂在主模型的训练目标中添加了一项，用于惩罚具有相似重构特性的成对实例以及在同一个空间区域中的隐式表征。
4. 基于单词识别的检测[10]:

为了对抗对抗性样本攻击中的拼写错误问题，文中建议在下游分类器前面放置一个单词识别模型，这个单词识别模型建立在RNN半字符体系结构的基础上，引入了几种新的退避策略来处理罕见的单词。对抗性的拼写错误攻击一直是一个长期存在的现实问题。文中主要关注文本分类任务中的逆向选择拼写错误问题，解决以下攻击类型: 删除、添加、交换单词内部字符。这些干扰源于心理语言学研究，该研究表明，只要每个单词的第一个和最后一个字符保持不变，人们就恶意理解由混杂的内部字符改变的文本。文中提出了一种任务不可知的防御，附加一个单词识别模型，该模型可以预测给定完整输入序列(可能存在拼写错误)的句子中的每个单词。单词识别模型的输出构成下游分类模型的输入。这个单词识别模型时基于RNN的半字符单词识别模型。虽然单词识别器是针对手头任务中特定领域的文本进行训练的，但由于特定领域的词汇量较小，它们通常在测试时会遇到UNK(不在单词库中的词汇)。为了处理未观察到的和罕见的单词，文中提出几种退避策略，包括使用在较大语料库上训练的通用单词识别器。

为了解决字符级对抗性攻击，文中引入了一个简单的两阶段解决方案，从而将单词识别模型置于下游分类器之前。在这种方案下，所有的输入均按照组合模型进行分类。这种模块化方法有几个好处: (1)可以为多个下游分类任务或者模型部署相同的单词识别模型; (2)可以使用较大的未标记语料库来训练单词识别模型。对于对抗性错误来说，有两个重要因素决定了这种组合模型的鲁棒性: 识别拼写错误单词的准确性和对相同输入的对抗性干扰的敏感性。

用于单词识别的半字符RNN模型只对子集进行攻击和评估，在这种情况下，单词识别性能不合理地依赖于所选词汇量。在敌对环境中，罕见的词语尤其重要，因为它们为攻击者提供了机会。进行回退的方法有: (1)单词识别器按原样传递可能拼写错误的单词; (2)将单词回退为一个中性单词，这个中性单词在类中相似的分布; (3)回退到一个更大更泛化的单词识别模型。

可以通过四种字符级编辑来干扰: (1)交换，交换一个单词的两个相邻内部字符; (2)删除，删除单词的内部字符; (3)键盘，用QWERTY键盘的相邻字符替换内部字符; (4)添加，在单词内部插入新字符。为了保证干扰不会影响人类理解句子的能力，只允许敌手编辑单词的内部单词，且不允许编辑停止词或短语4个字符的单词。

对于单字符的攻击，尝试所有可能的干扰，直到找到一个推翻模型预测的敌手。对于2字符攻击，贪婪地修复1字符攻击中置信度最低的编辑，然后对剩余单词尝试所有允许的干扰。高阶攻击可以以类似的方式执行。贪婪策略减少了获得高阶攻击所需要的计算量，但也以为者鲁棒性得分是分类器真实鲁棒性的上界。

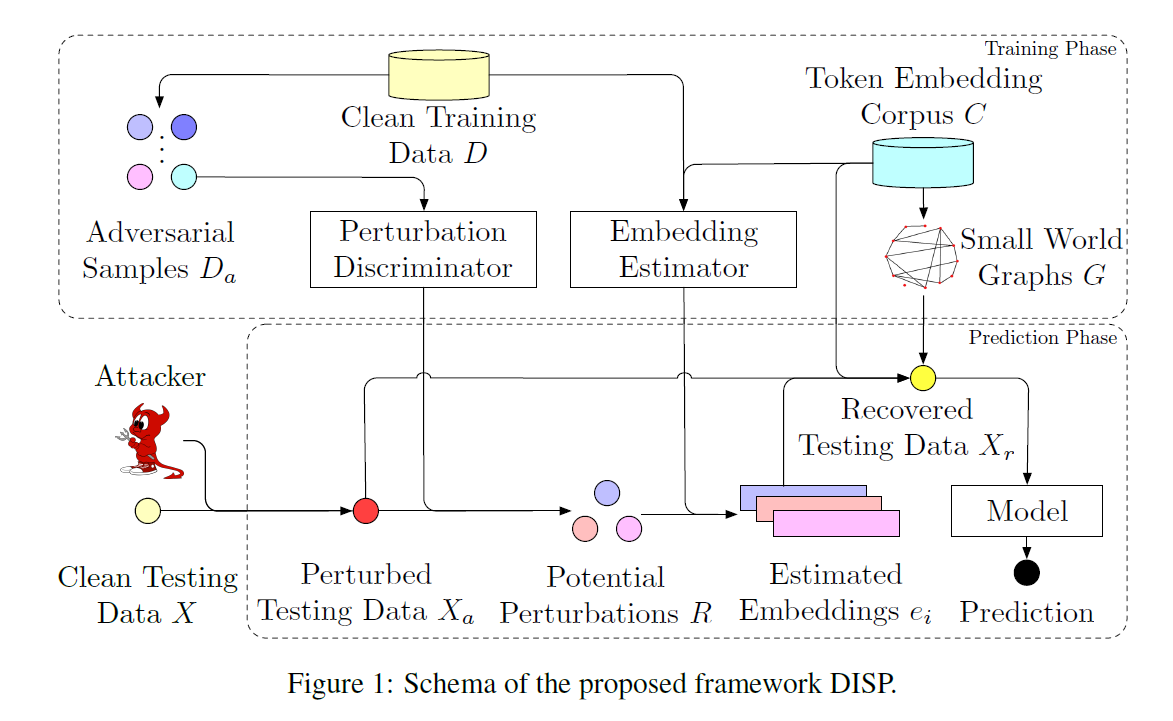
随着字符和词条输入在现代NLP算法中变得司空见惯，因此有必要强调其中的漏洞。文中证明了最小篡改攻击可以将分类器的准确率降低到随机猜测。因此，文中建议将单词识别作为一种防范措施，并以基于RNN的半字符单词识别器为基础。当作为一种防御机制使用时，最准确的单词识别模型并不总是能抵御敌对攻击。文中强调了需要控制这些模型的灵敏度以提高鲁棒性。

1. 扰动区域检测[11]:

针对机器学习的对抗性攻击已经威胁到各种实际应用，如垃圾邮件过滤和情感分析。文中提出一个新的框架，学习辨别扰动框架(DISP)，用来调整和识别恶意干扰，从而阻止针对文本的分类模型的对抗性攻击。为了识别对抗性攻击，扰动鉴别器验证文本中的词语被扰动的可能性，并提供一组潜在扰动。对于每个潜在的扰动，嵌入估计器根据上下文恢复原始单词的嵌入，并根据近似神经网络搜索选择替代标记。DISP可以组织任何NLP模型的对抗性攻击，而无需修改模型结构和训练过程。识别和消除不明显的干扰是对抗性攻击防御的核心。由于图像由连续的像素组成，因此与捕获相邻像素强度之间不寻常的差异方法相比，自然外观中的扰动不容易被识别。其次，句子由词汇量巨大的单词组成，所以很难列举所有可能的句子。因此，现有的计算机视觉防御方法不能直接用于NLP任务。

识别到受干扰的词语后，消除干扰以阻止对抗性攻击的最简单方法是移除这些受干扰的词语，然而，从句子中删除单词会导致句子断裂，导致NLP模型的性能下降。因此，恢复移除的词语是很重要的。然而，训练一个令人满意的语言模型需要大量多样的训练数据，而这些数据往往是不可用的。不准确的语言模型会不连贯地修补丢失的标记，从而进一步恶化预测性能。为了解决这个问题，文中建议通过一个带有上下文语言建模的蒙板语言模型目标从有区别的扰动中恢复标记。文中使用DISP(识别扰动)作为NLP中阻止对抗性攻击的框架。模型的目标是在不修改模型结构和训练过程的情况下保护模型免受对抗性攻击。DISP由三部分组成: 扰动鉴别器、嵌入估计器和具有相应小世界图的标记嵌入语料库。给定扰动测试数据，扰动鉴别器首先识别一组扰动词语。对于每个受干扰的词语，使用词嵌入语料库优化的嵌入估计器推断出一个嵌入向量来表示其语义。最后，在层次分类法上进行有效的神经网络搜索，将每个嵌入向量转化为适当的标记，以替换相关的扰动词。

DISP在训练阶段，从语料库中构造一个语料库，用于训练扰动鉴别器，以便能够识别扰动标记。然后再建立嵌入语料库的小世界图后，利用词嵌入语料库训练嵌入估计器恢复溢出的词语。再预测阶段，对于测试数据中的每个词语，扰动鉴别器能够识别词语是否受到了扰动。对于每个可能受到干扰的潜在词语，嵌入估计器生成一个近似的嵌入向量，并检索嵌入空间中距离最近的词语以进行恢复。最后，恢复的测试数据用于预测。预测模型可以是任何NLP模型。DISP是阻止对抗性攻击的通用性框架，因此鉴别器和估计器也可以灵活选择。DISP的框架如下图所示。



参考文献:

[1]郑海斌,陈晋音,章燕,等.面向自然语言处理的对抗攻防与鲁棒性分析综述[J].计算机研究与发展,2021,58(8):1727.

[2]Gao J, Lanchantin J, Soffa M L, et al. Black-box generation of adversarial text sequences to evade deep learning classifiers[C]//2018 IEEE Security and Privacy Workshops (SPW). IEEE, 2018: 50-56.

[3]Sato M, Suzuki J, Shindo H, et al. Interpretable adversarial perturbation in input embedding space for text[J]. arXiv preprint arXiv:1805.02917, 2018.

[4]Garg S, Ramakrishnan G. Bae: Bert-based adversarial examples for text classification[J]. arXiv preprint arXiv:2004.01970, 2020.

[5]Zhao Z, Dua D, Singh S. Generating natural adversarial examples[J]. arXiv preprint arXiv:1710.11342, 2017.

[6]Ribeiro M T, Singh S, Guestrin C. Semantically equivalent adversarial rules for debugging nlp models[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 856-865.

[7]Liu K, Liu X, Yang A, et al. A robust adversarial training approach to machine reading comprehension[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(05): 8392-8400.

[8]Liu X, Cheng H, He P, et al. Adversarial training for large neural language models[J]. arXiv preprint arXiv:2004.08994, 2020.

[9]Coavoux M, Narayan S, Cohen S B. Privacy-preserving neural representations of text[J]. arXiv preprint arXiv:1808.09408, 2018.

[10]Pruthi D, Dhingra B, Lipton Z C. Combating adversarial misspellings with robust word recognition[J]. arXiv preprint arXiv:1905.11268, 2019.

[11]Zhou Y, Jiang J Y, Chang K W, et al. Learning to discriminate perturbations for blocking adversarial attacks in text classification[J]. arXiv preprint arXiv:1909.03084, 2019.