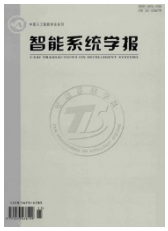

智能系统学报
CAAI Transactions on Intelligent Systems
ISSN 1673-4785,CN 23-1538/TP

《智能系统学报》网络首发论文

题目：一种建立在 GPT-2 模型上的数据增强方法
作者：张小川，陈盼盼，邢欣来，杨昌萌，滕达
收稿日期：2023-04-30
网络首发日期：2024-01-04
引用格式：张小川，陈盼盼，邢欣来，杨昌萌，滕达．一种建立在 GPT-2 模型上的数据增强方法[J/OL]．智能系统学报.
https://link.cnki.net/urlid/23.1538.TP.20240103.1021.004

智 能 系 统 学 报
CAAI Transactions on Intelligent Systems

DOI:10.11992/tis.202304055

一种建立在 GPT-2 模型上的数据增强方法

张小川，陈盼盼，邢欣来，杨昌萌，滕达
（重庆理工大学 两江人工智能学院，重庆 401135）

摘　要：针对句子分类任务常面临着训练数据不足，而且文本语言具有离散性，在语义保留的条件下进行数据增强具有一定困难，语义一致性和多样性难以平衡的问题，本文提出一种惩罚生成式预训练语言模型的数据增强方法（punishing generative pre-trained transformer for data augmentation,PunishGPT-DA）。设计了惩罚项和超参数 α，与负对数似然损失函数共同作用微调 GPT-2(generative pre-training 2.0)，鼓励模型关注那些预测概率较小但仍然合理的输出；使用基于双向编码器表征模型 (bidirectional encoder representation from transformers,BERT)的过滤器过滤语义偏差较大的生成样本。本文方法实现了对训练集 16 倍扩充，与 GPT-2 相比，在意图识别、问题分类以及情感分析 3 个任务上的准确率分别提升了 1.1%、4.9%和 8.7%。实验结果表明，本文提出的方法能够同时有效地控制一致性和多样性需求，提升下游任务模型的训练性能。

关键词：自然语言处理；人工智能；数据增强；句子分类；少样本；序列到序列；生成式预训练语言模型；双向编码器表征模型

中图分类号：TP391.1　　**文献标志码：**A

A data augmentation method built on GPT-2 Model

ZHANG Xiaochuan，CHEN Panpan，XING Xinlai，YANG Changmeng，TENG Da
(Liangjiang Artificial Intelligence College, Chongqing University of Technology, Chongqing　401135, China)

Abstract: Sentence classification tasks often face the problem of insufficient training data, Moreover,text language is discrete,and it is difficult to perform data augmentation under the condition of semantic preservation.Balancing semantic consistency and diversity is also challenging. To address this issues, this paper proposed a sequence-to-sequence generation-based data augmentation method called PunishGPT-DA. Firstly, the method designed a penalty term and hyperparameter α，which worked together with the negative log-likelihood loss function to fine-tune GPT-2(Generative Pre-Training 2.0) and encouraged the model to focus on the outputs with small predicted probabilities but still reasonable. Secondly, the method used a filter based on BERT (Bidirectional Encoder Representation from Transformers) to remove generated samples with significant semantic bias. The method achieves 16-fold expansion of the training set and improves accuracy by 1.1%,4.9%,and 8.7% in intent recognition,question classification,and sentiment analysis,respectively when compared with GPT-2.Experimental results demonstrate that the proposed method effectively balances the requirements for semantic consistency and diversity ,enhancing the training performance of downstream task models.

Keywords: Natural language processing; Artificial intelligence;Data augmentation;Sentence classification; Few samples;Sequence to sequence;Generative punishing generative pre-trained transformer for data augmentation;Bidirectional Encoder Representation from Transformers

句子分类[1]（sentence classification，SC）是最基本和常见的自然语言处理（natural language process，NLP）任务之一，广泛应用于 NLP 的很多子领域，如意图识别、情感分析、问题分类等。当给定一个句子作为输入时，其任务是将其分配给一个预定义标签。深度神经网络往往需要大规模的高质量标记的训练数据来实现高性能，然而在特定领域，由于人工标注数据集代价昂贵，常常只有少量样本可供使用。本文研究在数据匮乏情况下的句子分类任务准确率较低的问题，训练数据的不足使得句子分类任务模型无法得到有效的训练，从而导致

收稿日期：2023-04-30.
基金项目：国家自然科学基金项目（61702063）；重庆市技术创新与应用发展专项（cstc2021jscx-dxwtBX0019）.
通信作者：张小川. E-mail：zxc@cqut.edu.cn.

泛化能力差。为解决这一问题，数据增强是一种有效的方法。

通常，数据生成的语义一致性和多样性对目标任务至关重要[2]，语义保留即前后语义保持一致是数据增强最基本的要求，训练样本的丰富表达能使神经网络更好地学习权重。一些学者的研究工作已经开始注重数据的多样性和质量。如在计算机视觉中，文献[3]使用代理网络来学习如何增强多样性。NLP 中的一些研究[4]对原句进行随机替换、随机交换、插入和删除操作实现增强数据的多样性，为了避免简单数据增强方法（easy data augmentation,EDA）方法引入过多噪声，一种更简单的数据增强方法(an easier data augmentation,AEDA) [5]将随机插入 token 改为随机插入标点符号，一定程度上缓解了噪声引起的语义偏差问题，然而随机插入标点符号可能会不恰当地断句，语义保留和多样性仍无法同时有效控制。随着大规模预训练语言模型的问世，一些研究将其应用于数据增强，Anaby 等[6]提出基于语言模型的数据增强方法（language-model-based data augmentation,LAMBADA），采用训练数据微调 GPT-2 模型[7]，在训练过程中将相应的标签拼接到每个样本，以便为该类生成新数据，在句子分类方面取得了显著的改进。然而，该方法采用 top-k 和 top-p 采样的方式增加多样性，这种方式很有可能会导致累计误差的产生，使得生成句子质量低下。

从本质上讲，语义一致性和多样性的目标其实是相互冲突的，即生成多样性高的样本更可能导致语义发生变化，因此，需要同时考虑多样性与语义一致性，对生成数据进行控制，得到较为平衡的数据。本文提出一种引入惩罚项的数据增强方法（punishing generative pre-trained transformer for data augmentation, PunishGPT-DA），用于生成增强数据来改进句子分类任务。此方法的数据增强过程建立在预训练语言模型 GPT-2 基础上，通过设计惩罚项、超参数，使用双向编码器表征模型（bidirectional encoder representations from transformers，BERT）[8]作为过滤器完成数据增强。实验结果表明了该方法的有效性。

1 数据增强相关工作

从增强数据的多样性来看，数据增强方法可以大致分为基于复述的方法、基于噪声的方法和基于采样的方法 3 类。

基于复述的方法包括在词汇、短语、句子层面的重写。Zhang 等[9]首先利用词库（a electronic lexical database,WordNet）替换句子中的同义词应用于数据增强；条件 BERT（conditional bert,CBERT）[10]掩盖句子的部分字符，由 BERT 生成替换词；Jiao 等[11]使用数据增强来获得特定任务的蒸馏训练数据，利用 BERT 将单词标记为多个单词片段，并形成候选集；回译以生成的方式重写整个句子，被应用于低资源句子分类[12]，使用不同的二级语言提高了分类精度，Hou 等[13]通过 L 层变换器对串联的多个输入话语进行编码，利用重复感知注意和面向多样性的正则化生成更多样化的句子。Kober 等[14]使用对抗生成网络（generative adversarial network,GAN）生成与原始数据非常相似的样本。

基于噪声的方法添加微弱噪声，使其适当偏离原始句子。EDA[4]通过随机插入、删除、替换、交换操作得到增强数据。Peng 等[15]通过删除对话语句中的槽值来获得更多的组合；Sahin 等[16]通过依赖树变形对句子进行旋转。Sun 等[17]将混合技术应用到基于 transformer 的预训练模型中进行数据增强（Mixup-Transformer），将 Mixup 与基于 Transformer 的预训练结构相结合，进行数据增强；Feng 等[18]在提示部分随机删除、交换和插入文本字符，用于微调文本生成器；Andreas[19]提出了一种简单的数据增强规则，通过采用出现在一个类似环境中的其他片段替换真实的训练样本的某个片段，来合成新的样本。Guo 等[20]提出一种序列到序列模型的混合方法（sequence-level mixed sample data augmentation，SeqMix），通过组合训练集中的输入输出序列来创建新的合成样本。

基于采样的方法掌握数据分布，并在其中采样新的样本。大型语言模型（large language models, LLMs）的出现为生成类似于人类标注的文本样本创造了新的条件。LLMs 的参数空间允许它们存储大量知识，大规模预训练使得 LLMs 能够编码用于文本生成的丰富知识。如生成式预训练语言模型(generative pre-trained transformer,GPT)系列，，GPT~GPT-3[7, 21-22]采用预训练+微调的方式，其中预训练阶段通过大规模的无标注数据对模型进行训练，使其学习到通用的语言表示和语义理解能力，微调阶段利用有标注数据进行监督学习，使模型能够适应特定的任务要求，提高性能和准确度。GPT 系列目前已经发展到 4.0，聊天生成预训练转换器（chat generative pre-trained transformer,ChatGPT）遵循指导生成预训练转换器（instruct generative pre-trained transformer，InstructGPT）[23]的训练方式，

利用带有人类反馈的强化学习(reinforcement learning from human feedback, RLHF),使其在对话领域能够对输入产生更丰富的响应。这些最先进的模型也被广泛地用来进行数据增强，Abonizio 等[24]通过连接样本中的 3 个随机 token 作为 GPT-2 模型生成阶段的前缀生成样本。Kumar 等[25]研究了不同类型的基于 Transformer 的预训练语言模型，表明将类标签处理到文本序列为微调预训练模型进行数据增强提供了一种简单有效的方法；Bayer 等[26]设计了一种基于 GPT-2 的方法，通过设计不同的前缀分别处理短文本和长文本的生成，在短文本任务和长文本任务上都取得了很好的改进。类似的，Claveau 等[27]使用特定于类的数据微调 GPT-2 模型，并从原始文本中输入一个随机单词进行生成。然后应用分类器对生成的数据样本进行过滤。Liu[28]冻结 GPT-2 模型 softmax 之前的层，采用强化学习对 softmax 之后的层进行微调。随着 ChatGPT 的问世，Dai 等[29]提出了 ChatAug，利用 ChatGPT 为文本生成增强数据，获得了显著提升。

引入噪声的方法可以有效提升数据的多样性，利用预训练语言模型的数据增强方法可以更好地学习到语言规律和语义信息，因此，基于上述工作，本文提出惩罚生成式预训练语言模型的数据增强方法（punishing generative pre-trained transformer for data augmentation, PunishGPT-DA），通过设计损失函数微调预训练语言模型 GPT-2，有效保证增强数据的质量。

2 PunishGPT-DA

2.1 方法概述

句子分类是一种基于句子数据进行分类的任务，属于监督学习问题的一个实例。给定训练集 $D_{train}=\{(x_i,l_i)\}_{i=1}^{N}$，包含 N 个训练样本，其中 x_i 是由 $\{x_i^1,x_i^2,...,x_i^p\}$ 组成的文本序列，包含 p 个字符，$l_i\in\{1,...,q\}$表示在含有 q 个标签的集合中，样本 x_i 对应的标签。$x_i\in X$，X 代表整个样本空间，假设对于所有 N，存在函数 f，使 $l_i=f(x_i)$，监督学习的目标是在仅给定数据集 D_{train} 的情况下在整个 X 上近似 f，从 D_{train} 的域推广到整个 X，即在 D_{train} 上训练分类算法 F，使其能够近似 f，然而如果 D_{train} 非常小，将显著地影响算法 F 的性能。数据增强试图通过合成额外的训练数据来解决这个问题，给定训练集 D_{train} 和算法 F，本文的目标是生成 $D_{aug}=\{(y_j,l_j)\}_{j=1}^{T}$，$D_{aug}=D_{train}\bigcup D_{filter}$，其中 D_{filter} 是方法每次迭代后生成的数据，D_{aug} 是最终数据集，包含 T 个样本，y_j 是由 $\{y_j^1,y_j^2,...,y_j^m\}$ 组成的文本序列，包含 m 个字符，对应标签为 l_j。

为此，本文提出了一种面向句子分类的数据增强方法 PunishGPT-DA。PunishGPT-DA 由生成器 G_θ 和过滤器 F 两个模块组成。图 1 说明了本方法的步骤：1)通过改进的损失函数（式(3)）微调生成器的语言模型，训练生成器学习在原始句子的基础上合成新样本，得到参数被微调之后的生成器 G_θ。2)对 D_{train} 进行处理作为 G_θ 的输入生成数据 D_{syn}，D_{syn} 相较于原损失函数训练出的生成器生成的数据拥有更高的多样性，但也不可避免地引入了噪声。3)针对此问题，采用原始数据 D_{train} 微调过滤器 F，将每次迭代生成的样本 D_{syn} 由 F 过滤，丢弃低质量的样本，得到过滤后的增强样本 D_{filter}，D_{filter} 并入原始数据集中作为新的 D_{train} 进行下一次迭代，经过一定次数的迭代后得到最终的数据集 D_{aug}。

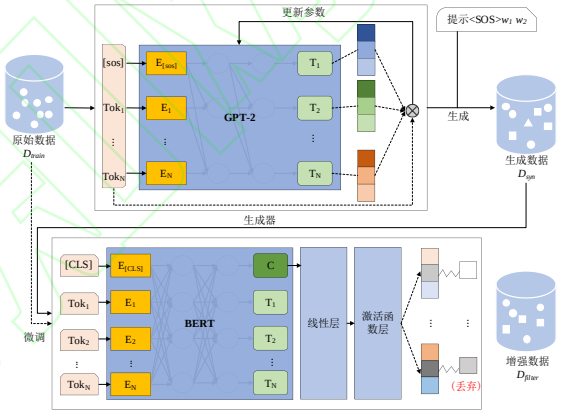

图 1 PunishGPT-DA 数据增强过程

Fig.1 PunishGPT-DA Data augmentation process

2.2 生成器

PunishGPT-DA 采用预训练语言模型 GPT-2 生成数据，GPT-2 是一个在海量数据集上训练的语言模型，采用“预训练+微调”的二段式训练策略，它利用庞大的语料库进行预训练，语料库被处理成由 token 组成的长序列，由 $\boldsymbol{U}=w^1,w^2,...w^j,...,w^T$ 表示，生成模型采用无监督自回归训练的方式，以最大化生成目标序列的概率为目标，根据极大似然估计，可以最大化目标序列 $\boldsymbol{U}$ 出现的概率,即最大化 $P(U)$，根据条件概率的链式法则，可以将生成目标序列的概率表示为条件概率的乘积：

$$P(U)=\prod_{j=1}^{T}\boldsymbol{G}_\theta(\boldsymbol{w}^j\mid\boldsymbol{w}^{j-k},...,\boldsymbol{w}^{j-1}) \quad (1)$$

将式(1)取对数并加上负号，得到负对数似然损失函数为

$$\begin{aligned}J_\theta&=-\log(\prod_{j=1}^{T}\boldsymbol{G}_\theta(\boldsymbol{w}^j\mid\boldsymbol{w}^{j-k},...,\boldsymbol{w}^{j-1}))\\&=-\sum_j\log(\boldsymbol{G}_\theta(\boldsymbol{w}^j\mid\boldsymbol{w}^{j-k},...,\boldsymbol{w}^{j-1}))\end{aligned} \quad (2)$$

在数据增强任务中，同预训练一致，以句子自身指导模型的微调，即以最大化生成目标序列的概率为目标，因此，以负对数似然函数作为损失函数的生成模型鼓励生成与原数据相似的句子，使生成的文本趋于重复和“枯燥”，当以此为目标训练得非常好时，甚至会生成与输入句子完全一致的样本数据。

为了关注生成数据的多样性，本文引入惩罚项来中和现有的损失函数，同时为了平衡多样性与语义一致性，引入超参数 α，改进后的损失函数为

$$L^{'}=\alpha\cdot J_{\theta}+(1-\alpha)\cdot\exp(-J_{\theta}) \quad (3)$$

式(3)是一种加权损失函数，由 J_θ 和 $\exp(-J_\theta)$ 2 部分组成。其中 J_θ，即式(2)是负对数似然损失，用于衡量生成的序列和目标序列之间的差距；$\exp(-J_\theta)$ 将其视为惩罚项，用于惩罚过度相似的生成结果，这意味着，如果生成器产生与目标序列中过于相似的 token，它将受到惩罚。本文拟通过添加 $\exp(-J_\theta)$，使模型会在给定上下文条件下，根据语言的语法和语义规则，更加关注可能性较小但仍然有一定意义与合理性的输出。这些输出可能是预测概率较小但仍然合理的单词、短语、句子结构等，在某些情况下可能会提供更有趣、更具创造性的文本。α 是一个用于控制 J_θ 和 $\exp(-J_\theta)$ 2 部分在损失函数中重要程度的超参数，当 α 较小时，$\exp(-J_\theta)$ 的影响更大，从而鼓励生成多样性更高的样本。相反，当 α 较大时，J_θ 的影响更大，从而鼓励生成语义一致性更高的样本。因此，式(3)可以看作在保证生成序列准确的基础上，通过惩罚过度自信的生成结果来鼓励生成更多的多样性，通过调整 α 的值，可以在一致性和多样性之间进行平衡，获得高质量的生成结果。

此外，在预测阶段，通常采用序列的前 i 个字符作为前缀提示后续词语的生成，然而，D_{train} 中存在多个序列前 i 个字符相同，以相同的前缀作为提示会导致原本不同标签的 2 个句子对应的增强样本可能相同，使得增强样本语义标签不明。因此，本文为每条训练数据添加了数字序号作为该数据的唯一标志，数字序号随训练数据一起参与训练。在预测阶段，数字序号与前 i 个字符一起作为前缀，确保了前缀的唯一性，并为生成器提供了额外的上下文，形式为(<SOS>,w_1,w_2,...,w_i)，其中<SOS>是数字序号，(w_1,w_2,...,w_i)是样本的前 i 个字符。这种操作确保了增强样本彼此不同，但仍然基于实际数据。

2.3 过滤器

使用增强样本的一个障碍是它可能引入的噪声和误差。虽然在微调生成器时同时考虑了语义保留和丰富表达，避免了模型过度生成低频词，但自然语言具有复杂性，有可能微小的改动便会影响句子的语义，导致增强数据集中的低质量样本对下游任务模型的性能产生影响。为此，如图 1 所示，本文使用基于 BERT 的过滤器 F 对其进行过滤选择，过滤器 F 包括 BERT 层、线性层、ReLU 激活函数层。输入数据首先经过 BERT 层获取特征表示，其次通过 Dropout 技术进行正则化处理，以减少过拟合风险，然后将 Dropout 层的输出输入到一个具有 786 个输入特征和类别数量输出特征的线性变换层，将特征表示映射到分类标签的空间，最后经过 ReLU 激活函数得到最终的分类结果。对于生成的样本 (y,l)，验证是否 $F(y)=l$，若分类正确则保留，不正确舍弃。因此，每一次完整的迭代后会得到增强数据集 D_{filter}，D_{filter} 并入原始集作为新的训练集。

3 实验结果与分析

3.1 数据集

本文共使用了三个公开的句子分类数据集，分别是由法国公司 Snips 在人机交互过程中收集的数据集 Snips，包含 7 个意图类别共 14 484 条数据。由文本检索会议（text retrieval conference,TERC）标注的细粒度问题分类数据集 TREC,包含 6 种问题类型共 5 952 条数据。由斯坦福大学自然语言处理组标注的情感分析数据集（stanford sentiment treebank v2,SST-2）,SST-2 属于电影评论情感分类的数据集，用 2 个标签（positive 和 negative）标注，共 8 741 条数据。

3.2 实验设置

根据先前工作[25]模拟用于句子分类少样本场景的设置，本文针对每个任务的训练集进行子采样，每个类随机选择 10 个样本，每个数据增强模型均对其进行 16 倍扩充。为避免数据集的随机性带来误差，本文一个任务下的对比实验均采用相同的子数据集。为更好地测试模型的性能，本文的验证集和测试集采用完整的数据集。

在微调 GPT-2 阶段，设置批量大小为 2，迭代次数为 100，学习率设定为 1e-5，样本最大长度为 20，超过则截断；生成数据时每条句子的提示为“i w_1 w_2”。BERT 在大量数据上进行预训练，并在几个句子分类任务上表现出最先进的性能。因此，本文使用 BERT 模型构建过滤器及句子分类器，本文使用“BERT-Base-Uncased” 模型，该模型有 12 层，768 个隐藏状态和 12 个头。PunishGPT-DA 使用 BERT 模型第一个特殊字符([CLS])的输出作为句子

的特征表示，在传入下一层进行分类之前，以 0.1 的 dropout 设置应用于句子表示。训练过程采用自适应矩估计算法（adaptive moment estimation,Adam）进行优化，学习率设置为 4e-5，本文对模型进行 100 个 epoch 的训练，并在验证集上选择表现最好的模型进行评估。

所有的实验均在 Intel Core i5-9500 3.00GHz 处理器，GeForce RTX 2028 SUPER 显卡，Ubuntu 20.04.4 LTS，python 3.8.0 下进行。

本文实验将与以下模型进行对比：

1) GPT-2[7]：为验证本文提出损失函数的有效性，本文以 GPT-2 作为基准模型，该模型以式(1)为损失函数，其余条件与 PunishGPT-DA 保持一致。

2) EDA[4]：以词替换、交换、插入和删除为基础的数据增强方法。

3) AEDA[5]：在句子中随机插入标点符号实现数据增强。

4) $GPT_{context}$[25]：采用文献[6]中的方式，将标签与序列连接起来构造训练集：$y_1SEPx_1EOSy_2...y_nSEPx_nEOS$。在此基础上以 $y_iSEPw_1...w_k$ 作为生成阶段的提示，生成增强数据。

3.3 实验结果与分析

本文对比了在意图识别、问题分类及情感分析任务少样本情景下的数据增强策略，表 1 总结了多种数据增强方法下同一模型在不同数据集中的分类准确率。

表 1 不同增强策略下的模型准确率

Table 1 Model accuracy under different augmentation strategies

方法	SNIPS	TREC	SST-2
无增强	84.4	61.7	52.5
GPT-2*	86.3	63.2	52.5
EDA	86.0	53.5	50.1
AEDA	86.9	64.8	53.2
$GPT_{context}$	79.3	61.0	58.7
PunishGPT-DA*	**87.4**	**68.1**	**61.2**

注：“*”表示按照本文提出的方法得出的模型结果，其余是按照原文献方法复现得出的模型结果，对于数据增强模型，每个类均采用 10 个样本进行训练，最佳结果用粗体标出。

如表 1 所示，与基线模型 GPT-2 相比，本文提出的数据增强方法在 3 个数据集上的准确率相对提升了 1.1%、4.9%和 8.7%，这说明本文提出的损失函数能有效提升增强数据的质量；相较于 EDA、AEDA 和 GPT_{contex} 方法，本文提出的数据增强方法在 3 个数据集上的准确率均有提升，表明了本文增强方法的普遍性。

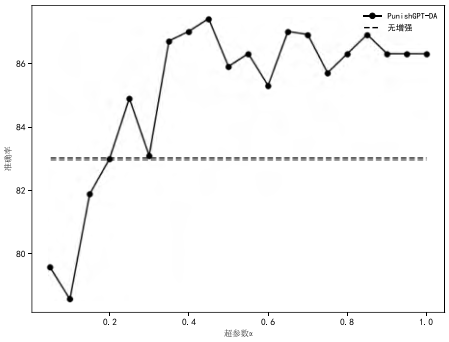

图 2 不同超参数下模型准确率

Fig. 2 Model accuracy under different hyperparameters

本文对比了不同超参数 α 设置下 PunishGPT-DA 的性能，采用 snips 的子采样后的数据集，每个类别包含 10 个样本，对其进行 16 倍扩充。如图 2 所示，α=0.3 之前模型准确率较低，这是因为在超参数控制下增强数据多样性较强，为数据集引入了过多的噪声；随着 α 增大，曲线逐渐上升，直到 α=0.45 时下游任务模型准确率达到最高，此时生成模型能够很好地控制数据多样性和一致性之间的平衡，使模型准确率达到最好的效果；随着 α 继续增大，一致性占据优势，使得生成数据相较于原数据只有微小的改动，致使模型准确率下降，趋于平缓。这表明，本文提出的损失函数能够同时控制语义和多样化的表达，有效平衡数据的一致性和多样性。

本文研究了过滤机制对 PunishGPT-DA 性能的影响，分别在 3 个子采样后的数据集上进行了消融实验。实验结果如表 2 所示，删除了过滤机制后，模型准确率均有下降。这表明过滤器对整个增强过程至关重要。

表 2 过滤机制对 PunishGPT-DA 的影响

Table 2 Influence of filtering mechanism on PunishGPT-DA

方法	Snips	TREC	SST-2
PunishGPT-DA	87.4	68.1	61.2
-过滤	87.0	59.3	52.5

表 3 PunishGPT-DA 在不同数据集大小下的准确率

Table 3 Accuracy of PunishGPT-DA under different dataset sizes

方法	5	10	20	50	100
无增强	75.9	84.4	88.9	94.6	97.0

PunishGPT-DA	78.9	87.4	91.3	95.3	97.1

此外，本文还研究了在不同数据集大小情况下 PunishGPT-DA 对下游任务模型性能的影响。表 3 为模型在 snips 数据集上进行实验的结果，每种意图类别分别取为 5、10、20、50、100 条数据作为训练样本，构成少样本数据集，并进行 16 倍扩充。如表 3 所示，随着训练数据的增多，本文的数据增强方法对下游任务模型性能的提升作用越来越弱。这表明在少样本情境下，本文所提出的数据增强方法可以有效提升句子分类任务模型性能，当训练数据较为充足时，已经能为下游任务模型提供较为丰富的信息，数据增强带来的效益也就随之减弱。

为了更加明确损失函数的作用机制，本文分别对采用 2 种损失函数生成的数据进行了探索，如表 4 所示，本文分别摘取了部分数据。通过观察损失函数式(3)生成的数据及过滤后的数据可以发现，数据较原始数据有较大的多样性，但大体上符合标签语义；采用损失函数式(2)生成的数据较原始数据只有个别单词的变化，多样性引入不足。由此可以发现本文提出损失函数的有效性。

表 4 生成数据示例

Table 4 Generate data samples

操作	数据示例
原始数据	1.book a brasserie(BookRestaurant)
	2.play a melody by colin blunstone(PlayMusic)
	3.can you add some disco to my playlist called genuine r&b(AddToPlaylist)
损失函数式(3)生成数据	1.book a **restaurant where i can get a burrito**
	2.play a melody **by kaori utatsuki off the album that has top-twenty hits**
	3.can you **play a song off a disc that has a top ten hit rate of 2%**
损失函数式(3)过滤后数据	1.book a **restaurant where i can get a burrito**
	2.play a melody **by kaori utatsuki off the album that has top-twenty hits**
损失函数式(2)生成数据	1.book a **restaurant**
	2.play a melody by colin blunstone **please**
	3.can you **please** add some disco **music** to my playlist called genuine r&b
损失函数式(2)过滤后数据	1.book a **restaurant**
	2.play a melody by colin blunstone **please**
	3.can you **please** add some disco to my playlist called genuine r&b

注：表中“原始数据”行括号内单词代表该句子的标签；表格中加粗部分代表相对于原始数据的不同之处

4 结束语

针对少样本句子分类任务中训练数据不足的问题，本文提出一种平衡语义一致性和多样性的数据增强方法 PunishGPT-DA，与当前主流方法相同，此方法建立在大规模的预训练语言模型的基础上，同时又区别于当前主流方法修改提示指导生成模型生成阶段的做法，本文提出的方法从训练角度指导模型生成数据。实验结果表明，在小样本情景下，本文方法可以更有效地保证数据质量，有效提高句子分类模型的分类准确率。尽管本文解决了增强样本质量不高的问题，然而通过损失函数控制数据的生成，可能会导致语法不可控地变化，不符合人类正常的阅读习惯，因此，在句子结构多样性方面还有一定的提升空间。下一步将探索句子结构方面的改进，使其更加自然流畅。

参考文献:

[1] AGGARWAL C C, ZHAI Chengxiang. A survey of text classification algorithms[M]. Boston, MA: Springer, 2012: 163-222.

[2] ASH J T, ZHANG Chicheng, KRISHNAMURTHY A, et al. Deep batch active learning by diverse, uncertain gradient lower bounds[EB/OL].(2020-02-24)[2023-04-30]. https://arxiv.org/abs/1906.03671.pdf.

[3] CUBUK E D, ZOPH B, MANÉ D, et al. AutoAugment: learning augmentation strategies from data[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 113-123.

[4] WEI J, ZOU Kai. EDA: easy data augmentation techniques for boosting performance on text classification tasks[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 6382-6388.

[5] KARIMI A, ROSSI L, PRATI A. AEDA: an easier data augmentation technique for text classification[C]//Findings of the Association for Computational Linguistics: EMNLP 2021. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021: 2748-2754.
[6] ANABY-TAVOR A, CARMELI B, GOLDBRAICH E, et al. Do not have enough data? deep learning to the rescue![J]. Proceedings of the AAAI conference on artificial intelligence, 2020, 34(5): 7383-7390.
[7] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.
[8] DEVLIN J, CHANG Mingwei, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).Minneapolis, Minnesota: Association for Computational Linguistics, 2019:4171-4186.
[9] ZHANG Xiang, ZHAO Junbo, LECUN Y. Character-level convolutional networks for text classification[C]//NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1. Cambridge: MIT Press, 2015:649-657.
[10] WU Xing, LV Shangwen, ZANG Liangjun, et al. Conditional BERT contextual augmentation[C]//International Conference on Computational Science. Cham: Springer, 2019: 84-95.
[11] JIAO Xiaoqi, YIN Yichun, SHANG Lifeng, et al. TinyBERT: distilling BERT for natural language understanding[C]//Findings of the Association for Computational Linguistics: EMNLP 2020. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 4163-4174.
[12] NG N, YEE K, BAEVSKI A, et al. Facebook FAIR's WMT19 news translation task submission[C]//Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1). Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 314-319.
[13] HOU Yutai, CHEN Sanyuan, CHE Wanxiang, et al. C2C-GenDA: cluster-to-cluster generation for data augmentation of slot filling[J]. Proceedings of the AAAI conference on artificial intelligence, 2021, 35(14): 13027-13035.
[14] KOBER T, WEEDS J, BERTOLINI L, et al. Data augmentation for hypernymy detection[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021: 1034-1048.
[15] PENG Baolin, ZHU Chenguang, ZENG M, et al. Data augmentation for spoken language understanding via pretrained language models[EB/OL].(2021-03-11)[2023-04-30]. https://arxiv.org/abs/2004.13952.pdf.
[16] SAHIN G G, STEEDMAN M. Data augmentation via dependency tree morphing for low-resource languages[EB/OL]. (2019-03-22)[2023-04-30].https://arxiv.org/abs/1903.09460.pdf .
[17] SUN Lichao, XIA Congying, YIN Wenpeng, et al. Mixup-transformer: dynamic data augmentation for NLP tasks[C]//Proceedings of the 28th International Conference on Computational Linguistics. Stroudsburg, PA, USA: International Committee on Computational Linguistics, 2020: 3436-3440.
[18] FENG S Y, GANGAL V, KANG D, et al. GenAug: data augmentation for finetuning text generators[C]//Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 29-42.
[19] ANDREAS J. Good-enough compositional data augmentation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 7556-7566.
[20] GUO Demi, KIM Y, RUSH A. Sequence-level mixed sample data augmentation[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 5547-5552.
[21] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[EB/OL].(2018)[2023-04-26].https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.
[22] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. New York: ACM, 2020: 1877-1901.
[23] OUYANG Long, WU J, XU Jiang, et al. Training language models to follow instructions with human feedback[C]//Advances in Neural Information Processing Systems. New Orleans: Curran Associates, Inc. ,2022: 27730-27744.
[24] ABONIZIO Q H, JUNIOR B S. Pre-trained data augmentation for text classification[C]//Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I. Cham: Springer International Publishing, 2020: 551-565.
[25] KUMAR V, CHOUDHARY A, CHO E. Data augmentation using pre-trained transformer models[C]//Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems. Suzhou, China:Association for Computational Linguistics,2020:18-26.
[26] BAYER M, KAUFHOLD M A, BUCHHOLD B, et al. Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers[J]. International journal of machine learning and cybernetics, 2023, 14(1): 135-150.
[27] CLAVEAU V, CHAFFIN A, KIJAK E. Generating artificial texts as substitution or complement of training data[C]//Proceedings of the Thirteenth Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, 2022:4260-4269.
[28] LIU Ruibo, XU Guangxuan, JIA Chenyan, et al. Data boost: text data augmentation through reinforcement learning guided conditional generation[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language

Processing (EMNLP). Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 9031–9041.
[29] DAI Haixing, LIU Zhengliang, LIAO Wenxiong, et al. AugGPT: leveraging ChatGPT for text data augmentation[EB/OL].(2023-03-20)[2023-04-30]. https://arxiv.org/abs/2302.13007.pdf.

作者简介：

张小川，教授，重庆理工大学两江人工智能学院副院长、人工智能系统研究所所长、兼任中国人工智能学会常务理事、机器博弈专委会主任委员、重庆市人工智能学会常务理事、副秘书长,主要研究方向为计算机博弈、智能机器人、软件工程。主持和参与纵向科研项目 30 余项、横向科研项目 50 余项，获省部级科技类奖 2 项、教学类成果奖 2 项。发表学术论文 100 余篇，主编专著或教材 6 部。

陈盼盼，硕士研究生，主要研究方向为自然语言处理、问答服务机器人。

邢欣来，讲师 ，博士，主要研究方向为自然语言处理、对话系统，主持和参与科研项目 10 余项，发表学术论文 10 余篇。