自然语言处理中的数据增强

——基于长短文本分类器的文本生成方法

杨磊鑫

云南师范大学东区 云南 昆明 650000

【摘　要】：本研究提出了一种复杂的基于生成的文本数据增强方法，旨在提升低数据量环境下自然语言处理（NLP）任务的性能。我们的方法结合使用新的语言模式，以增加语法多样性，并通过特定转换手段人工制造训练数据，从而改进分类器性能。本文采用了两种子方法，分别针对长文本和短文本任务，以保持标签质量的同时提高数据的创新性。研究结果表明，相比无增强基线和其他数据增强技术，我们的方法能够显著提高准确率和 F1 分数，特别是在数据稀缺的场景下。此外，本研究还从实证、实践和理论多个角度评估了方法的适用性，并讨论了其在不同类型数据集上的成功应用。

【关键词】：文本数据增强；深度学习；自然语言处理

DOI:10.12417/2705-0998.24.04.064

Data augmentation in natural language processing

——A Text Generation Method Based on Long Short Text Classifier

Leixin Yang

Yunnan Normal University East Area Yunnan Kunming 650000

Abstract: This study proposes a complex generative text data augmentation method aimed at improving the performance of natural language processing (NLP) tasks in low data environments. Our approach combines the use of new language patterns to increase syntactic diversity and artificially create training data through specific transformation techniques to improve classifier performance. This article adopts two sub methods, targeting long and short text tasks respectively, to maintain label quality while improving data innovation. The research results indicate that compared to non enhanced baselines and other data augmentation techniques, our method can significantly improve accuracy and F1 score, especially in scenarios with scarce data. In addition, this study evaluated the applicability of the method from multiple perspectives, including empirical, practical, and theoretical perspectives, and discussed its successful application on different types of datasets.

Keywords: text data augmentation; Deep learning; natural language processing

1 引言

随着计算能力的提升和训练数据的广泛获取，深度学习在众多领域受到了重视，尤其是在数据较少的学习任务中，发展训练数据的重要性往往超过了选择和建模分类器。数据增强方法因此被提出，旨在通过特定转换手段人工制造训练数据，以改进分类器性能。这些方法主要针对深度学习算法，虽然它们在许多分类任务中表现卓越，但在数据不足时性能仍不稳定。数据增强不仅可以作为正则化手段，偏好简单解决方案，还可以解决数据集不平衡问题，增强分类器的安全性，且有助于缓解大数据量难以获取的问题。

在自然语言处理（NLP）领域，特别是在数据稀缺或标注成本高昂的情况下，人工数据创建的研究能够带来实质性益处。例如，在紧急情况管理中，快速识别和分类信息至关重要，但资源限制可能影响效率。这种挑战同样适用于需要大量高质量标注数据的中小企业。尽管 NLP 中的预训练或迁移学习方法已部分解决了这一问题，但仅仅扰乱输入数据而不引入新的语言模式的增强手段，并未能显著提升模型性能。

因此，本文提出了一种复杂的基于生成的方法，通过结合使用新的语言模式（即提高语法多样性）来克服这些挑战，证明了其与预训练模型结合使用时的有效性。本文采用了两种子方法，分别适用于长文本和短文本，旨在保持标签质量的同时提高数据的创新性。本文的研究结果表明，这种方法在低数据情境下相比无增强基线和其他数据增强技术，能够显著提高准确率和 F1 分数，尽管在某些情况下可能不适用。本文从实证、实践和理论多个角度评估了方法的适用性，并讨论了其在不同类型数据集上成功应用的影响。

2 文献综述

数据增强是一种通过执行标签保留转换来人工扩充训练数据量的机器学习技术。最早可以追溯到 LeNet 模型中，通过对训练图片进行随机扭曲，使得 MNIST 数据集扩增了九倍，显著提升了手写数字识别的准确度。所谓的“标签保留”指的是在训练数据转换过程中保持类别信息不变，这是数据增强研究

中的一个关键概念，因为缺乏标签保留的转换可能导致生成的数据被错误地分类。例如，在情感分析任务中，改变句中的实体通常能够保持情感倾向不变，但随机添加单词可能会改变原文的情感色彩。随着研究的深入，对“标签保留”的定义逐渐放宽，即使是打破原有标签保留的转换，只要适当调整标签也被认为是可接受的。此外，一些转换可能以高概率保留正确类别，但并非绝对确定，这就引入了数据增强方法安全性的概念，即转换后正确标签被分配的概率。

在自然语言处理（NLP）中，定义保留标签的文本转换尤为困难，因此研究者尝试了诸多方法，包括词汇级的交换、删除、引入拼写错误、同义词替换、使用语言模型预测的单词等，以及更为复杂的操作如改变依存树结构、往返翻译或实例插值等。近期研究还探索了利用文本生成方法进行数据增强，包括应用循环神经网络、生成对抗网络和从变分自编码器中采样实例进行短文本增强，以及使用 GPT-2 模型进行文本生成等。

然而，这个领域面临的挑战在于，文本数据增强被认为只在生成数据包含对任务有帮助且预训练模型未曾见过的新语言模式时才有效。基于此，我们的方法受到文本生成技术的启发，旨在解决三个主要的研究缺口：同时考虑短文本和长文本的连贯性与高度新颖性；维护增强方法的标签保真度与数据质量；以及克服文本数据增强与预训练模型结合使用时遇到的限制。

3 概念与实现

3.1 概念设计

文本生成过程依托于具备卓越文本生成能力的语言模型，该模型定义了词序列的概率分布如下：

$$P_\theta(\omega_t|\omega_{t-k},...,\omega_{t-1})\forall t$$

其中，模型P_θ预测给定上文$\omega_{t-k},...,\omega_{t-1}$的情况下，当前词$\omega_t$出现的概率。这一能力使得$P_\theta$能够生成文本。通过使用短语前缀作为上文，模型可以在遵循特定主题的同时，抽象化采样方法的细节。此外，通过引入温度参数来调整 softmax 函数中对数几率的缩放，从而控制生成文本的随机性。为确保语言模型在数据增强中的合理应用，关键在于生成的文本不仅与训练数据相似，而且还能反映出相应的类别信息（即标签保留或安全性）。

我们提出的增强方法通过以下三个步骤实现上述行为建模：

第一，使用特定类别 c 的训练数据X_c对预训练模型P_θ进行进一步训练（微调），旨在丰富该类别的表示。这不仅使模型掌握训练数据的词汇、拼写和结构特征，还为选定的类别生成偏向，明确地保留类别信息。此过程区分了适用于长文本的上下文依赖增强过程与适合短训练实例的上下文独立过程。

第二，为增强安全性和标签保留，每个训练数据的微调输入中均添加特殊的“文本开始”-token。在文本生成阶段，这些 token 作为生成前缀使用，引导模型产生与特定训练数据相似的文本，确保增强示例既具有差异性又基于真实数据。对于长文本实例，通过添加实例开头的几个词或标题作为上下文 token（例如，“$<|startoftext|>(\omega_1,...,\omega_k)$”），以提升生成数据的多样性。对于短文本，采用上下文独立变体，连接训练集中实例的出现次数（例如，“|i|”）。模型经过微调后，能够将唯一 token 与相应实例关联，从而在基于记忆的前提下完成文本生成，同时通过调整温度参数引入采样不确定性，避免完全复制。

第三，通过生成数据的过滤来增强标签保留，为此，为每个类的生成文本和训练数据实例创建文档嵌入，反映各自内容。若生成的数据实例X_{gen}与要增强的类的实际训练数据X_c在潜在空间中相距过远，则认为两者在语义和/或句法上存在差异，因此将这些数据剔除：

$$X_{filtered}=\left\{x_i\in X_{gen}|\mathrm{dist}\left(\mathrm{Emb}(x_i),\mathrm{Centroid}\left(\mathrm{Emb}(X_c)\right)\right)<\delta\right\}$$

3.2 实现

图 1 详细展示并总结了增强的三个步骤，按照算法顺序排列。采用此方法可以显著提高过程的类安全性，虽然无法完全排除错误标签的可能性。实施过程中，我们采用了含有 3.55 亿参数的 GPT-2 模型，因其多样化的生成能力特别适合小数据集分析。模型通过执行概念设计部分讨论的三种不同扩展进行丰富。

首先，导入 GPT-2 模型并提取特定类别数据。所有类别实例均添加前缀 token 和后缀 token，如果实例足够长以嵌入上下文，则去除“|{num}|”字段。对模型进行数百至数千次微调，以降低损失并确保生成中的训练数据优先级。

接下来，根据实例长度，为每个类别生成文本，并设置适当的温度参数以调节随机性/创造性。最后，通过使用 Sentence-BERT 为生成数据创建文档嵌入并过滤，根据与训练数据质心的距离过滤不适合的实例。通过手动调节阈值并展示最远的 10 个实例，以找到合适的参数设置，直到达到满意的标签保留水平。

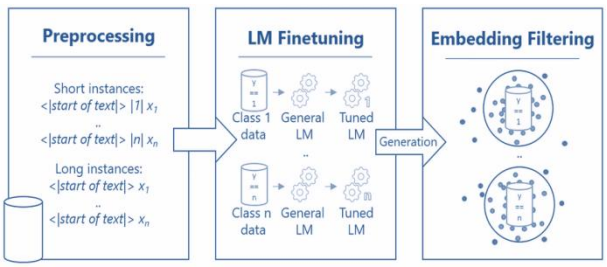

图 1 增强步骤

4 研究结果

4.1 应用领域

情感分析涉及对个体、事件或主题的观点、态度和情绪的分析。作为自然语言处理领域一个极为常见的任务，情感分析在众多应用场景中扮演着重要角色。例如，无论是组织还是个人，决策过程越来越依赖于公众意见。由于情感分析在多项文本机器学习基准测试中占有一席之地，并在学术研究中频繁被采用，它成为了本研究实验的理想对象。

4.2 模型和数据集

在模型和数据集方面，本研究基于构建的低数据量环境以及针对长短文本主题分类的实际低数据量场景，评估了先前概念化和实施的数据增强方法，特别采用了 SST-2 数据集（结果 I）。我们选用了 Howard 和 Ruder（2018）提出的 ULMFit 模型，该模型整合了预训练的编码器、线性池化网络和 softmax 输出层。在数据增强阶段，编码器针对所有可用的特定任务数据（包括增强数据）进行了微调，之后在监督学习任务中训练整个网络。

针对短文本情感分析和危机相关 Twitter 数据分类的上下文独立方法评估，我们采用了 SST-2 数据集的子集来模拟标准数据集在低数据量环境下的情形。鉴于这些构建条件在实际应用中的局限性，我们对实际低数据量环境进行了更广泛的评估。

4.3 评估设置和超参数的预评估

在评估设置和超参数预评估方面，所有的数据增强方法都通过与基线的 10 折交叉验证比较进行了评估。情感分析的结果进一步与 Wei 和 Zou（2019）中的 EDA 数据增强方法进行了对比。对于情感分析任务，我们对两个类别进行了增强，而对于其他任务，则针对少数类别进行了增强。文本生成过程提供了利用多种超参数进行优化的可能性，我们通过使用不同数据集来避免模型的过拟合。模型的温度参数决定了生成文本的创造性，合适的温度范围被认为是 0.7 到 0.9 之间。在管理变更主题的评估中，温度设置为 0.7 被认为最适合当前的应用案例。

4.4 结果

本节的研究成果表明，从人类的视角评价为优秀的结果同样在量化评估中得到了体现，具体数据见表 1。我们提出的数据增强方法在几乎所有情况下均优于基线方法及 Wei 和 Zou（2019）中的 EDA 方法。值得注意的是，在数据量较少的情境下，我们的方法展现出显著的性能提升，相较于基线方法和 EDA 方法分别提高了 15.53%和 3.56%。而在数据量最大的场景中，即便如此，我们的增强策略在最佳状态下也分别实现了相对于基线和 EDA 方法 0.49%和 1.22%的性能提高。

表 1 各子样本上的准确性评估结果

Dataset	Run	Baseline	EDA	Text Gen
SST-2 100	AVG(SD)	0.5581(0.0463)	0.6934(0.0124)	0.7134(0.0207)
	Best	0.6226	0.7139	0.7495
SST-2 300	AVG(SD)	0.7241(0.0119)	0.7217(0.0047)	0.7402(0.0067)
	Best	0.7417	0.7295	0.7534
SST-2 500	AVG(SD)	0.7505(0.0077)	0.7534(0.0074)	0.7598(0.0126)
	Best	0.7651	0.7671	0.7754
SST-2 700	AVG(SD)	0.7646(0.0054)	0.7578(0.0038)	0.7627(0.0066)
	Best	0.7705	0.7632	0.7754

当数据量较少时，本方法展现出的改进幅度最大，这归因于 GPT-2 模型的先验知识在这种情况下发挥了最有效的作用。与 EDA 方法有时未能改善基线性能不同，本模型还生成了质量上乘的实例。此外，Longpre 等人（2020）对低效数据增强方法的分析也在本研究中得到了体现。与 EDA 方法不同，我们提出的增强算法通过引入编码器之前未曾见过的新语言模式，从而丰富了训练数据（如图 2 所示）。

进一步地，我们展示了对我们消融评估结果的摘要（见图 2）。首先，我们探讨了不同增强规模对 SST-2 100 数据集上结果的影响。由于计算时间的考虑，我们将每个实例的增强样本数量限制在 10 个以内。结果表明，增强样本数量越多，获得的结果越好。然而，人类评审者指出，较高的样本数可能并非总是有益的，因为随着每个实例内样本的增加，样本重复的情况也会上升。

我们还去除了增强过程中的某些步骤，以评估每一步骤的贡献度。在第一项测试中，我们没有在微调和生成阶段包含实例编号（如图 3 所示，“w/o n.”）。结果显示，平均准确率下降了 5.42 个百分点，这表明该组件对整个增强过程至关重要。对于增强方法的最后一步（图 3 中表示为“w/o f.”）同样适用。没有进行手动过滤，平均准确率下降了 2.64 个百分点。在选择每个任务的过滤参数时，已经观察到了这一点，因为某些实例看起来与类别不相关。

总体而言，这表明为了达到最优效果，必须包含增强过程中的所有步骤。基于这一最佳组合，我们进行了下一个评估研究。此外，通过检查我们技术生成的实例并将其与 EDA 方法比较，我们进行了错误分析。对于我们的方法，由于生成可能是多个原始实例的插值，因此难以明确知道生成内容来源于哪个原始实例。尽管存在这一挑战，我们仍然尝试通过计算莱文斯坦距离来识别与之最为接近的原始实例。

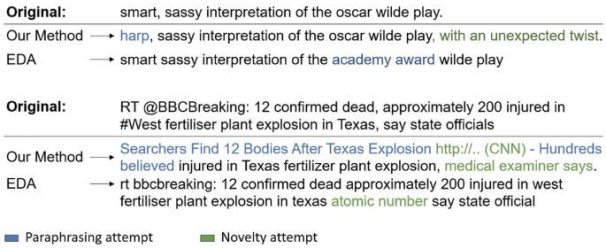

图 2 实例及其转换

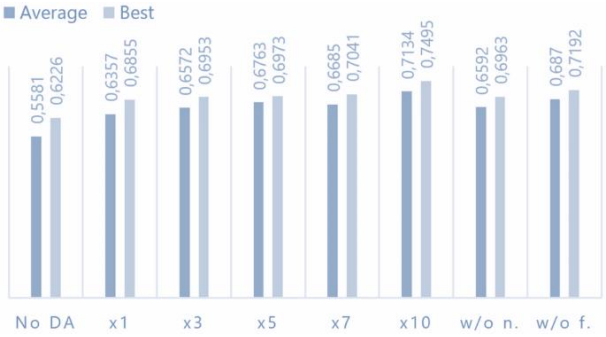

图 3 评估结果

5 研究结论

本研究成功地开发并评估了一种基于生成的文本数据增强方法，证明了其在低数据量 NLP 任务中的有效性。通过结合使用新的语言模式和针对性的增强策略，我们的方法不仅提高了模型的准确率和 F1 分数，还增强了分类器的安全性并解决了数据集不平衡问题。尽管在某些情况下可能不适用，但本研究从多个角度展示了其广泛的适用性和潜在的实用价值。特别是，在数据稀缺或标注成本高昂的情况下，我们的方法能够为中小企业和紧急情况管理提供实质性帮助。未来的工作将进一步探索不同 NLP 任务和数据集上的应用，以及方法的进一步优化和定制化。

参考文献：

[1] 周治.基于自然语言处理的社交媒体情感分析在公益慈善中的应用[J].科技传播,2024,16(01):1-4.

[2] 张小川,陈盼盼,邢欣来等.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,1-8[2024-03-02].

[3] 龚倩.自然语言处理技术在特检文本中的应用前景分析[J].西部特种设备,2023,6(06):49-52.

[4] 苗育华,李格格,线岩团.融合标签关联的隐空间数据增强多标签文本分类方法[J].现代电子技术,2023,46(24):159-164.

[5] 桂韬,奚志恒,郑锐等.基于深度学习的自然语言处理鲁棒性研究综述[J].计算机学报,2024,47(01):90-112.

[6] 崔振新,张卓言.基于文本增强的民航安全信息自动分类[J].中国民航大学学报,2022,40(03):47-53+64.

[7] Howard J,Ruder S.Universal language model fine-tuning for text classification[J].arXiv preprint arXiv:1801.06146,2018.

[8] Wei J,Zou K.Eda:Easy data augmentation techniques for boosting performance on text classification tasks[J].arXiv preprint arXiv:1901.11196,2019.

[9] Longpre S,Wang Y,DuBois C.How effective is task-agnostic data augmentation for pretrained transformers?[J].arXiv preprint arXiv:2010.01764,2020.