**基本问题**

过拟合是指训练误差和测试误差之间的差距太大。换句换说，就是**模型在训练集上表现很好**，**但在测试集上却表现很差**。模型对训练集"死记硬背"（记住了不适用于测试集的训练集性质或特点），没有理解数据背后的规律，**泛化能力差**。

**为什么**会出现**过拟合**现象？造成原因主要有以下几种：  
1、**训练数据集样本单一，样本不足**。如果训练样本只有负样本，然后那生成的模型去预测正样本，这肯定预测不准。所以训练样本要尽可能的全面，覆盖所有的数据类型。  
2、**训练数据中噪声干扰过大**。噪声指训练数据中的干扰数据。过多的干扰会导致记录了很多噪声特征，忽略了真实输入和输出之间的关系。  
3、**模型过于复杂。**模型太复杂，已经能够“死记硬背”记下了训练数据的信息，但是遇到没有见过的数据的时候不能够变通，泛化能力太差。我们希望模型对不同的模型都有稳定的输出。

**解决过拟合的根本性方法——获取和使用更多的数据（数据集增强）**

让机器学习或深度学习模型泛化能力更好的办法就是使用更多的数据进行训练。但是，在实践中，我们拥有的数据量是有限的。解决这个问题的一种方法就是**创建“假数据”并添加到训练集中——数据集增强**。通过增加训练集的额外副本来增加训练集的大小，进而改进模型的泛化能力。

Hierarchical Data Augmentation and the Application in Text Classification论文内容

ABSTRACT

数据增强在自然语言处理中的应用受到限制。在本文中，针对文本分类任务我们提出了一种新的方法，称为层次数据增强（Hierarchical Data Augmentation，HDA）。

首先，受文本的层次属性启发，HDA通过文字层面和句子层面两个方面增强文本。

第二，受计算机视觉中裁切数据增强方法的启发，在每个层级上，HDA利用注意力机制来从文本中分层提取（裁切）重要内容。

具体来说，具体来说，我们使用经过训练的层次注意力网络（HAN）模型分别在两个层次上获取训练集中所有文本的注意力值，这些注意力值进一步用于提取单词/句子的最重要部分，并通过按顺序连接它们来生成新样本。然后我们得到两个级别的增强数据集，WordSet和SentSet。最后，用一定数量的HDA生成的样本扩展训练集，并用新的训练集评估模型的性能。结果表明，HDA可以在两个层级生成大量高质量的增强样本，使用这些样本的模型可以获得显著的改进。

与现有方法相比，HDA具有理论和实践上的简单性，并且它可以在两个层次上扩充文本，以实现数据的多样性。

I. INTRODUCTION

**作者提出HDA的动机：**

1.文本具有**层次结构**。对于一段文本而言，通常由若干个句子组成；而对于句子而言，其又是由若干个词构成。因此从结构上看，文本天然存在着词与句这两个层次分明的层级，这被称为**文本的层次属性**。

2.**裁切**是一种在**计算机视觉中常用的数据增强方法**，其原理是从一张图片中裁切出一块区域，这块区域要包含图片中的主体，然后将裁切下来的内容当作新的图片，用于模型的训练。由于裁切后的图片在尺寸上小于原始图片，因此需要对其进行一定的处理，例如填充像素等，保持新旧数据在格式上的统一性。此外，又因为具有一定的随机性（在保留主体的前提下，随机裁切一块区域），使得裁切可以生成相当数量的新数据。例如，以不同位置裁切图片，就可以得到大量的新数据。综上所述，裁切尽管实现简单，但可以生成大量且优质的数据，是一种优秀的数据增强方法。

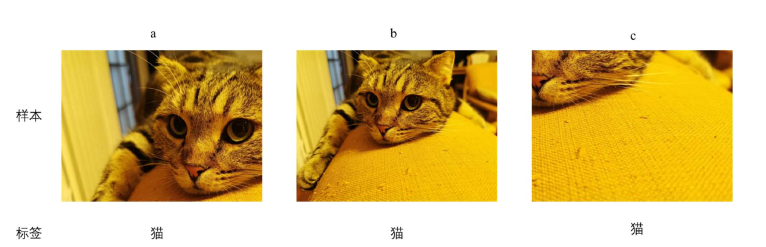


图1 裁切示例

受裁剪的启发，HDA通过使用注意力机制从文本中提取（裁切）一部分重要内容来生成新的样本。考虑到无论是对于图像还是文本，不同的部分对整体的重要性不同，引入**注意力机制**可以有效地衡量部分对总体的重要性。因此，我们可以使用它从文本中提取重要的单词、句子，以生成新的样本。

由于HDA生成的样本保留了原始样本中最重要的部分，因此可以保持标签不变。

**HDA的工作流程：**

为了层次化地增强文本数据，HDA首先利用文本的层次属性，设计了两个独立的子算法，WordAtt（Word Attention Augmentation）和SentAtt（Sentence Attention Augmentation），从而在词级和句级这两个层级上分别对文本进行增强。

具体来说，在分层注意力网络（HAN）工作的基础上，我们首先用训练集训练HAN模型，并使用其句级注意机制（SentAtt）获得训练集中每个样本的句级注意力值。然后，我们根据注意力值从样本中提取最重要句子的前Ts%百分比，并通过将这些句子按顺序连接生成新样本。在整个训练集上执行SentAtt，可以获得句子级增强数据集SentSet。

类似地，我们使用HAN的词级注意机制（WordAtt）来获得所有样本中每个句子的词注意值。然后，我们在样本中随机删除50%属于长句中最不重要单词的前Tw%的单词，并保持其余句子不变，这样我们获得了一个扩充样本。通过对整个训练集执行WordAtt，还可以生成另一个新的扩展数据集WordSet。

此外，为了充分利用HAN的层次信息，通过从SentSet和WordSet中随机抽取等量的样本合并，进一步生成了一个名为HybridSet的增强数据集。最后，分别用来自WordSet、SentSet和HybirdSet的一定数量的样本扩展原始训练集，并用HDA扩展训练集评估模型的性能。

**重要性：**

1、与简单数据增强（EDA）相比，HDA不仅可以**更快**地扩充文本，而且可以生成**大量**的扩充样本；

2、HDA生成的样本经实验证明是高质量的，向训练集中添加的样本越多，模型的性能越好。

**创新点：**

1、考虑到文本的层次结构，本文提出的HDA首先分别在词级和句子级对文本进行扩充，以前的工作主要是在单个层次上。

2、基于注意力机制，HDA通过从文本中提取内容的最重要部分来生成新样本，从而保持标签不变。

3、实验结果表明，EDA在长文本数据集上模型的性能不够好，这证明了EDA不是一种令人满意的文本分类数据增强方法。

为了验证HDA的泛用性，作者选用了两个不同的HAN模型作为基础模型，分别是层次注意网络（HAN）与层次注意混合神经网络（HAHNN）；并在两个不同的长文本数据集IMDB与Yelp上测试模型。实验结果表明，HDA是一种很有前途的技术。

II. RELATED WORK（补充说明领域内其他研究）

近年来，人们对自然语言处理的**各种数据增强技术**进行了探索。

1. 同义词替换(synonym replacement，SR)；由于同义词的数量非常有限，SR不能产生各种数据，需要额外的语言知识，即同义词库。
2. EDA是SR的升级版；四种不同的子操作:同义词替换、随机插入、随机交换和随机删除。每个操作都会随机更改(添加、替换或删除)几个单词在一个句子中生成一个新的句子。然后，用四个eda生成的句子对原句子进行扩展，生成一个扩充样本。显然，EDA完全以一种随机的方式增加文本，这可能会导致不可靠和不可控的数据。此外，EDA增强样本的信息冗余严重，数据量是原始样本的近5倍(如图2)。
3. 上下文增强(context Augmentation，CA)；根据语境用语言模型预测单词的同义词，因此可以提供比SR更广泛的同义词范围。然而，为了实现语境增强，首先需要在没有标签的大型语料库上预训练语言模型，接下来，使用标签条件架构对语言模型进行进一步训练，以预测替代词和增强文本。因此，CA耗费巨大。上述三种方法都只能在单词层面增强文本。
4. 回译法（Back-translation）不能应用于文本分类等其他任务。

最后介绍HDA的优点

III. BACKGROUND（具体介绍算法的基础HAN与同类算法EDA）

**HAN**

HAN是基于三级文本分析的文本分类模型，它在句级与词级文本分析中还引入了注意力机制，从而更好地量化不同句子/词语对于文档/句子的重要性，引导模型对那些重要的内容给予更多关注，从而提升文本分类性能。

如图1所示，为了充分利用文本的层次属性，HAN模型显式地设计了两层结构，用以分别执行文本的词级分析与句级分析。每层结构设计相同，都由一个编码器和Attention机制组成。编码器有序地读取文本内容并学习文本的特征，Attention机制接受编码器习得的特征，并最终通过计算这些特征在文本中的权重（即这些特征对文本主题的影响程度），并以Attention值表示。最后将权重与对应的特征相乘，使得文本的重要性内容得以凸显。

**EDA**

EDA包含四个数据增强操作:同义词替换、随机插入、随机交换和随机删除。

（1）同义词替换(Synonym Replacement，SR)：从输入中随机选择N个不属于停用词集的单词，对选中的词，从它们的同义词中随机选择一个替换原词；

（2）随机插入(Random Insertion，RI)：随机的找出句中某个不属于停用词集的词，并求出其随机的同义词，将该同义词插入句子的一个随机位置。重复n次；

（3）随机交换(Random Swap，RS)：随机的选择句中两个单词并交换它们的位置。重复n次；

（4）随机删除(Random Deletion，RD)：以 p 的概率，随机的移除句中的每个单词；

为了充分利用长句相较于短句能吸收更多噪声的特性，基于句子长度来变化改变的单词数，即不同的句长，因增强而改变的单词数可能不同。

具体实现：对于SR、RI、RS，遵循公式：n=α\*l，l表示句长，α即增强因子表示一个句子中需要改变的单词数的比例。

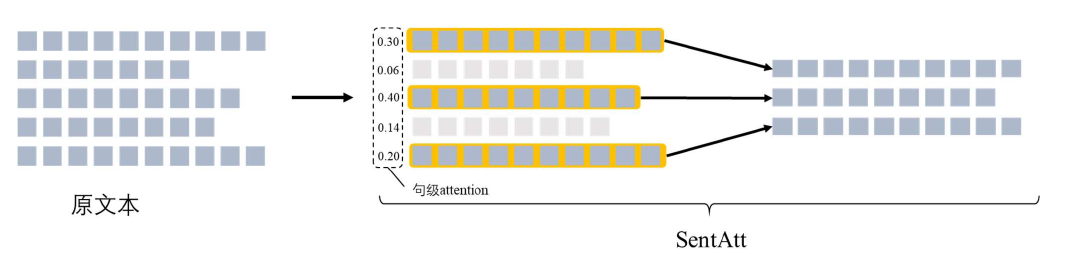
在RD中，让p和α相等。

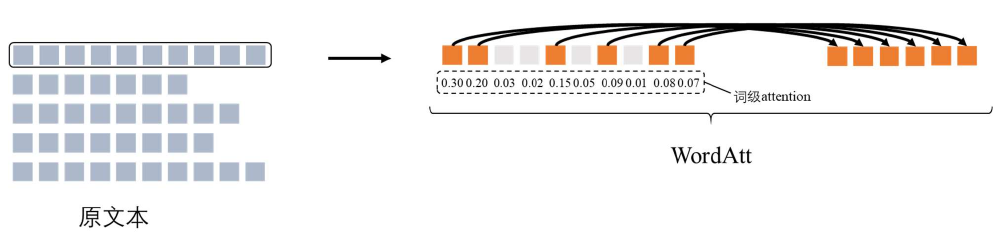
**缺点**：对短文本数据集有奇效，但对长文本大数据集效果不明显，有局限性。

IV. HDAHDA是一种高度基于HAN的数据增强算法。

**WordAtt与SentAtt**

HDA首先利用文本的层次属性，设计了两个独立的子算法，WordAtt和SentAtt，从而在词级和句级这两个层级上分别对文本进行增强，SentAtt和WordAtt的核心是通过删除文本中不重要的部分内容来突出文本中最重要的内容。其次，在每个层级上，HDA利用注意力机制，从文本中抽取出重要的词语/句子，然后将这些内容拼接新的样本。





**增强过程**

**①**若要对某一数据集执行HDA，首先使用这个数据集对应的训练集训练一个HAN，然后再次将训练集中的所有样本逐一喂入已训练好的HAN中，从而获得样本的词级Attention值和句级Attention值。

假设样本D中包含个句子， 表示D中第个句子；是中的总词数，是中的第个词语。

**②**对于**SentAtt**而言，首先只有当样本D中句子总数不少于5，即L≥5时，SentAtt才对其进行增强。若D的L<5，则SentAtt将直接跳过这个样本，不对其进行增强。对于符合要求的样本D，当其被喂入已训练好的HAN后，HAN中的句级Attention机制便可生成对应的句级Attention值。SentAtt 则将读取这些句级Attention值，将其按值从大到小排序，并将值最大的前 %部分对应的句子抽取出来，然后按照其在原文本中出现的顺序，将这些句子有序地拼接起来，从而得到D在句级上的增强文本。对训练集中的所有样本遍历执行SentAtt后，将得到一个句级增强数据集，记为SentSet。

**③**对于D中任意句子而言，**WordAtt**只对那些总词数不少于10（即T≥10）的句子进行增强，不满足条件的句子则保持原样。对于符合要求的句子，WordAtt利用HAN中的词级Attention机制，按照Attention值从大到小排序，然后对于那些Attention值最小的后%所对应的词语，WordAtt不直接将其删去，而是以50%的概率将其随机删去。换言之，最终WordAtt将从句子中删除%的词语，这些词语从那些Attention值最小的后%所对应的词语中随机地筛选出来。当对D中所有句子遍历执行WordAtt后，将得到一个关于原文本的词级增强文本。对于训练集中所有样本执行WordAtt后，将得到一个词级增强数据集WordSet。

④为了更充分地利用文本的层次属性，在 SentSet 和 WordSet 的基础上，通过从这两个数据集中**抽取等量**的样本，从而得到一个新的数据集HybridSet，HybridSet中包含两种层级的增强样本。

1. SentSet、WordSet 和 HybridSet都可作为数据增强结果使用。

**如何确定****与值**

作者以在IMDb数据集上寻找最优值为例，首先假定对于所有数据集而言，与的最优值都是固定的。作者分别使用增强数据集SentSet和 WordSet扩充到原训练集中，重新训练HAN模型并记录性能得到图3。

**结果**：=50%和=30%时分别达到最优。

V. EXPERIEMNTS**实验内容**

实验选用了两个长文本数据集分类任务，详细情况如下所列：

1、**IMDb Reviews**（IMDb）：电影评论数据集，包含 50000 条带标签的电影评论，所有电影评论都被归为消极情绪或者积极情绪，因此，该数据集为二分类数据集。

2、**Yelp 2018**（Yelp）：2018年耶鲁数据集，包含150万条数据。该数据集为多分类数据集，并被分为1-5之间的任意整数值。由于原始数据集较大，本章实验

从150万条数据中随机挑选50万条数据，作为实验所用。

实验时从上述数据集中随机抽取 90%的样本用作训练集，剩余的10%用作测试集。

**实验结果**

增强数据集数量如表2所示。

作者首先分别从SentSet、WordSet和HybridSet中随机抽取相当于训练集10%的样本，使用这些样本扩展训练集。然后，我们使用扩展训练集评估模型的性能，结果如表3所示。

**分析结果**

1.在四组实验中，模型使用 EDA 的表现总是5个方法中最差的，并且相较于仅使用原训练集的性能更差。例如在Yelp数据集上，HAHNN使用EDA数据集后的性能反而比使用原始训练集降低了0.07%，而HAN的性能降低了0.06%。在IMDb数据集上，HAHNN 性能降低了0.16%，HAN性能降低了0.12%。这表明，虽然 EDA 在短文本数据集上被证实对模型性能的提升效果明显，但是在长文本数据集中却反而对模型的性能有负面影响。由此表明EDA有一定的局限性。

2.在四组实验中，虽然仅仅使用HDA生成的数据（SentSet，WordSet 和 HybridSet）为原始训练集扩充了10%的样本，但是模型的性能都得到了明显的提高。例如，HAHNN使用 HybridSet 后，性能在Yelp数据集和 IMDb 数据集上分别提高了 1.05%和 0.41%，HAN 使用 SentSet后，在IMDb上性能提高了 0.31%，使用WordSet后，在Yelp上性能提升了 0.40%。由此可知，HAD 所生成样本的质量很高。

**实验中不同参数对性能的影响**

TABLE4中P值大小表示抽取的样本数占原训练集样本数的百分比，0%表示模型使用原始训练集的性能。

1. 随着加入训练的样本数不断增加，HAN模型使用 SentSet，WordSet 和 HybridSet 的性能也持续提高。表中所列的六组实验均在最大 P 值时获得最优性能。
2. 从表中还可以获得另一个重要的现象，以P=25%为分水岭，在P<=25%时，HybridSet在三者之间的性能表现通常是最差的，而在P>25%时，HybridSet 在三者之间的性能稳定地表现最佳。

VI. DISCUSSION

本节将HDA与现有几种具有代表性的数据增强法做一个简要的对比，从中讨论HDA的优势与不足。

**HDA与EDA的主要差异**：

1.HDA是一种高度基于Attention机制的算法，而EDA是一种高度基于随机机制的算法。

2.对于生成的数据，HDA将其视为一个独立的样本，而EDA则将其视为原数据的补充。因此，从最终对数据集的影响来说，HDA扩充了样本数量，而EDA扩充了样本的长度。

3.HDA可以从词级与句级这两个层级上分别对文本进行增强，而EDA只能从词级这一个层级上对文本进行增强。由此，HDA通过充分利用文本的层次属性，生成了大量且优质的数据，极大丰富了数据的多样性。

**EDA高度基于随机机制的主要弊端**：

1. 随机机制非常容易破坏文本的信息。
2. 随机机制的存在使得α值不能设定为一个较大的值，因为α越大，则代表文本中被随机修改的词语越多，显然文本信息被破坏得越严重，故α只能设一个较小的值，而较小的值又会使得EDA生成的数据之间相似度太高，这对数据增强而言是不可取的。

**与EDA相比，HDA的优势明显，主要可以概括为以下三点**：

1.提取文本的重要内容避免信息冗余。

2.引入Attention机制避免随机化处理。

3.HDA复杂度更低。

**图表说明HDA与现有几种算法的比较（TABLE 5）**

|  |  |  |  |
| --- | --- | --- | --- |
| 算法 | 增强层级 | 模型复杂度 | 额外的支持 |
| HDA | 词级与句级 | 非常简单 | 不需要 |
| Vae+Dis | 仅句级 | 复杂 | 不需要 |
| CA | 仅词级 | 复杂 | 额外需要语料库 |
| SR | 仅词级 | 简单 | 额外需要同义词库 |
| EDA | 仅词级 | 简单 | 额外需要同义词库 |

VII. CONCLUSION AND FUTURE WORK（**总结**）在本文中，我们首先提出了一种新的用于文本分类的数据增强算法HDA，它可以分层地增强文本，并通过基于注意机制从文本中提取最重要的内容来生成新样本。HDA可以创建两个不同的增强数据集，WordSet和SentSet。结果表明，HDA可以生成大量高质量的训练数据。此外，当使用大量新样本进行训练时，WordSet和SentSet的组合HybridSet可以进一步提高模型的性能。

（**展望**）对于未来的工作，虽然我们使用WordAtt和SentAtt来分层增强文本，但到目前为止，这两种方法是独立工作的。因此，我们计划联合使用SentAtt和WordAtt在两个层次上同时增强文本。（**给出具体思路**）例如，我们可以首先使用SentAtt提取最重要的句子，然后使用WordAtt进一步扩充这些句子。此外，考虑到本文只对文本分类任务进行了HDA评估，我们还计划将HDA应用于其他NLP任务，如机器翻译和问答。

REFERENCES有30条

IEEE中参考文献编号可以直接作为名词使用

