1. 文献信息

论文题目：Effective Use of Word Order for Text Categorization

with Convolutional Neural Networks

作者：Rie Jhonson,Tong zhang

发表时间：2015

1. 问题背景和意义

文本分类是电脑自动为文档分配定义类别的任务，它在机器学习技术的支持下有许多实际有用的应用，例如进行情感识别、领域识别、垃圾邮件检测等等。本文主要是提出了一种另外的方法进行文本分类，即利用卷积神经网络CNN来进行文本分类。通常，CNN主要应用于对图像信息的处理，但是不难理解，将文本的词汇看作图像的像素点来应用CNN是作者的思路。

为了能够更好的理解本文，我查阅了一些资料来了解一些基础的机器学习的方法和思想或者名词，下面简单说明一部分背景知识。

传统的文本分类不是利用卷积神经网络CNN而是常用词向量包来表示文档，然后通过支持向量机SVM等模型来进行分类，但是这种方法的弊端是他会减弱语序对主题的影响。其原因是用词向量包来表示文本时，每个向量中只包含了文本中出现某一个词的次数，而并没有其出现的顺序。其最直接简单的解决方法是使用bi-gram或更多级的n-gram（我的理解是将多个连续的词在一个词向量中表示），但是其缺点是在主题分类中，这种方法是无效的。所以，作者提出了一种利用CNN的思想进行文本分类。

CNN是convolution neural networks缩写，其 主要原理是通过卷积层每一个计算单元对输入数据的一部分进行处理，然后再由池化层把每卷积层送上来的数据进行展平，通过池化层输出的数据是更抽象化但是提取出了原始数据的某些特点的，常用的合并方法有平均池和最大池，分别计算每个区域的平均/最大值。然后经过多层这样的处理后再进行分类。



1. 实验方法和结论

实验步骤

首先文章对文本进行用于CNN处理的向量化表示，其基本思想就是把每个词都当作一个像素点来处理，位置信息即句子的词序，这样，一个词库有V个词长度为D的文档就可以看成是一幅有D个像素每个像素有V维的向量。在此基础上文章提出了两种不同的表达方法——seq-CNN和bow-CNN，前者保留了词序信息，如下左图所示，但是这就导致维数可能会非常大，造成计算复杂和很大的稀疏；后者再每一个区域内只保留词频信息，如下右图所示，这类似于传统的表示方法，但是由于卷积网络的特性，它只在每一个小区域内失去词频，表现力介于二者之间。这两幅图所展示的都是区域大小为2，内容为“I love it”的文本向量。

文章中，作者还用到了一种平行CNN，即有两个或多个不同的卷积和层平行工作并向上传输结果，这样做的好处是可以有多种不同的嵌入（embedding）方式来提高模型精度，例如可以用不同的区域大小或者不同的表达方法。

模型用Relu激活函数、最小平方损失函数，词汇库中采用30k个常用词汇，若词不在这个集中用0向量表示。再bow-CNN中，采用可变区域步长来加快计算速度。同时，还采用了在图像处理中常用的两个技术，一个是“dropout”应用与顶层输出，另一个是“response normalization”来处理池化层的输出。

作为比较，文章将相同的问题通过多种模型来解决，首先是作为基础的SVM支持向量机和词向量包，它是线性层和全连接神经网络的应用。文章还用NB-LM进行了实验作为对照。

文章进行了两种文本分类的应用，情感分类和主题分类。在情感分类中，训练集和测试集数据来自IMDB电影评论和Elec电子产品评论，二者的训练集和测试集中都包含25k条评论并对这些评论做了标记表情符号和把所有字母变为小写的预处理。主题分类主要用RCV1的语料库，一个文档可以与多个主题相联系。

实验结果

文章以错误率作为指标，在得到的数据中，CNN的方法整体优于SVM方法，这证明方法是有效的。在情感分类中，seq-CNN方法优于bow-CNN方法和绝大多数SVN方法。在主题分类中，区域大小与情感分类相比更大，并采用平均池并且用多个池单元，相对应的因为更长的句段对主题的联系更密切，主题更体现在整体层面且文本的位置更加重要（例如新闻往往有起始位置的关键句），这里的bow-CNN表现优于seq-CNN，这表明在主题分类中词序的影响会小一些。

对于平行的卷积网络采用区域大小为2和3，最大池化，他的表现比单一seq-CNN要好。在实验中表现最好的方法是seq2-bown-CNN, 他是三层平行并列的结构，其中两层是seq卷积层，他们把整个文档看作一个区域；第三层层是bow卷积层，区域大小可变，以one-hot表示的n个词组成的词向量作为输入。这个模型在两个任务中的表现都比seq2-CNN更好，这表明不仅仅短的句段有用，全局的上下文在训练中也是有用的。

最终，文章总结了这种方法能够更有效的原因，在SVM方法中训练出的表示正面的和负面的权重最大的词汇中绝大部分是单个词，即uni-gram，即使表示中也有bi-gram和tri-gram，但是其选择的结果还是大多为单词，这意味着SVM并没有很好的利用n-gram的特点，之前提到的NB-LM方法则更倾向于n较大的n-gram，这样也不利于最佳性能。在seq-CNN中，经过训练得出的预测权重较大的文本长度更自然。另外，权重相近的词汇作者发现其含义也是相近的，这更有助于顶层的分类。



这个方法还展现出一个优势，在以往的bag-of-n-gram表示中，在测试集中用于预测的词汇必须是在训练集中出现过的，而在CNN的方法中，即在训练集中出现过的词汇的其他组合也可以作为测试集中判断的依据，这是对词序利用的体现，提高了模型的精确度。

四、心得体会

这次的论文内容对我来说是一个完全全新的领域，之前没有接触过机器学习，我为此在网上查阅了许多资料，看了许多讲解的视频，最后算是对其有了一个初步的了解，对其常用的一些名词有了认识。尽管如此，在这篇论文中仍有部分内容我不能很好理解，语言障碍是一方面，还有一方面是对其提到的一些模型不了解，对所说的一些操作不理解，这让我感觉我只能是看懂这篇文章主要讲述的内容。

通过这次学习我发现了机器学习有趣之处，感觉到他还有很大的发展空间。我之前对于编程方面了解不多，通过这次学习也让我深刻感受到编程的学习很需要进行实践操作，仅仅学习一些数学或者其他理论知识很难理解。我非常希望在之后的学习中更深入地去学习这方面的知识，去自己动手实现一些机器学习的功能。