基于卷积神经网络的自适应权重multi-gram语句建模系统

张春云[1] 秦鹏达[2] 尹义龙[3]
(山东财经大学计算机科学与技术学院 济南 250014)[1] (北京邮电大学信息与通信工程学院 北京 100876)[2]
(山东大学计算机科学与技术学院 济南 250101)[3]

摘 要 如今信息量呈爆炸式增长,自然语言处理得到了越来越广泛的重视。传统的自然语言处理系统过多地依赖昂贵的人工标注特征和语言分析工具的语法信息,导致预处理中语法信息的错误传递到系统训练和预测过程中。因此,深度学习的应用受到了学者们的关注。因为它能实现端对端预测并尽可能少地依赖外部信息。自然语言处理领域流行的深度学习框架为了更好地获取句子信息,采用multi-gram策略。但不同任务和不同数据集的信息分布状况不尽相同,而且这种策略并没有考虑到不同n-gram的重要性分布。针对该问题,提出了一种基于深度学习的自适应学习multi-gram权重的策略,从而根据各n-gram特征的贡献为其分配相应的权重;并且还提出了一种新的multi-gram特征向量结合方法,大大降低了系统复杂度。将该模型应用到电影评论正负倾向判断和关系分类两种分类任务中,实验结果证明采用的自适应multi-gram权重策略能够大大改善模型的分类效果。

关键词 深度学习,自然语言处理,自适应权重,multi-gram
中图法分类号 TP391.1 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.01.011

Self-adaptation Multi-gram Weight Learning Strategy for Sentence Representation Based on Convolutional Neural Network

ZHANG Chun-yun[1] QIN Peng-da[2] YIN Yi-long[3]
(School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan 250014, China)[1]
(School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China)[2]
(School of Computer Science and Technology, Shandong University, Jinan 250101, China)[3]

Abstract Nowadays, with the explosive growth of the information, nature language processing has been paid more attention. The traditional nature language processing systems are overly dependent on the expensive handcrafted features annotated by experts and synatx information of language analysis tools. Deep neural network can achieve end-to-end learning even without costly features. In order to extract more information from input sentences, most neural networks of nature language processing combines with multi-gram strategy. However, due to various tasks or various datasets, the information distribution of diverse n-gram is different. With this consideration, this paper proposed a self-adaptation weight learning strategy of multi-gram, which generates the importance order of multi-gram by the training procedure of neural network. Moreover, a novel combination method of multi-gram feature vectors was exploited. Experimental results show that such method can not only reduce the complexity of network, but also can improve performances of positive and negative tendency classification of movie criticism, and relation classification.

Keywords Deep learning, Natural language processing, Self-adaptation, Multi-gram

1 引言

早期的自然语言处理(Nature Language Processing)基本上都是将基于语言学专家制定的一些语言学规则和模板进行匹配完成的。这样的方法在当时已经取得了突破性的进展。但其局限性也比较明显,当有新的领域出现时,就需要不断地完善,其成本非常昂贵。随着统计学理论的发展,学者们开始将机器学习的方法有效地与语言学知识进行结合,完成自然语言方面的工作[1]。这也就导致了其本质上还是一定程度依赖于语言学知识的准确性和完备性。当前有很多比较成熟的自然语言标注系统,例如stanford parser[1),其中包括词性标注(part of speech)、实体识别(Named Entity Recognizer)、依存分析(Dependency parser)等,虽然其准确率已经达到了相对较高的程度,但是仍存在一定程度上的错误,这种错误传递

1) http://nlp.stanford.edu/software/lex-parser.shtml

到稿日期:2015-08-01 返修日期:2015-10-11 本文受国家自然科学基金重点项目:基于机器学习的多模态医学影像信息处理与分析(U1201258),山东省自然科学杰出青年基金项目:基于机器学习的生物特征识别研究(JQ201316)资助。

张春云(1986-),女,博士,讲师,主要研究方向为信息抽取、机器学习,E-mail:zhangchunyun1009@126.com;秦鹏达(1991-),男,博士生,主要研究方向为信息抽取、自然语言处理,E-mail:qinpengda0406@163.com;尹义龙(1972-),男,博士,教授,主要研究方向为机器学习、生物信息识别,E-mail:ylyin@sdu.edu.cn。

到随后的系统中会被放大，并制约着系统的效果。因此，学者们开始思考，是否可以不利用这些标注工具，而直接从文本中学习出有价值的特征，直接实现端对端(end-to-end)[2]的学习。

深度学习[3]的兴起使这种想法的实践成为了可能。由于深度学习复杂的多层结构，使其具有优异的拟合能力，最初在计算机视觉(Computer Vision)[4]和语音识别(Speech Recognition)[5]领域均取得了显著的效果。随后，学者们开始将其引入到自然语言处理领域，例如关系抽取(Relation Extraction)[6]、信息检索(Information Retrieval)、问答系统(Question-answering)[7]等领域，而且取得了不错的进展。

目前，在自然语言处理领域常用的深度学习框架主要包括卷积神经网络(Convolutional Neural Network)[8]、循环神经网络(Recurrent Neural Network)以及 LSTM(Long Short-Term Memory)。循环神经网络的层级结构虽然并不复杂，但是由于其层数较深，因此优化问题仍是一个难题。BPTT[9]是当前学术界比较认可的一种循环神经网络的误差传递策略，然而其仍存在传播距离短和训练不足的问题，这些弊端也限制了循环神经网络的应用。LSTM 是一种循环神经网络的改进，其缓解了优化难题，但网络结构却十分复杂，参数较多，训练相对较困难。相比之下，卷积神经网络的网络结构更加清晰，其也被称作当前人工智能领域界最成功的一种深度学习框架，并且它给计算机视觉领域的发展带来了质的飞跃。所以，近两年，关于利用卷积神经网络做自然语言处理的工作涌现出来，并且也有很多科技公司的研究院(华为诺亚方舟研究院等)将其列为重点研究对象并取得了大量的成果[1)]。因此，本文以卷积神经网络为研究对象，针对其当前的应用现状进行改进和优化。

卷积神经网络在自然语言处理方面的早期的、比较经典的工作呈现在 Collobert 的论文[10]中。他结合自然语言的特点，将计算机视觉中的卷积神经网络的思想移植到文本处理中，并在词性标注、命名实体识别等任务中均取得了显著的效果。随后便涌现出很多利用卷积神经网络做不同的自然语言处理任务的工作。为了结合更多的信息，文献[11]结合 multi-channel 和 multi-gram 的思想使得卷积神经网络结合了更丰富的信息，并在多项句子分类任务的效果上得到了提高。文献[12]认为 max-pooling 可能会丢失一些重要信息，于是利用动态 multi-pooling 的策略进行优化，并在事件抽取任务上验证了效果。文献[13]将卷积神经网络利用在关系分类任务上，实现了将人工标注的语言学信息与神经网络自动抽取的信息相结合。文献[6]在文献[13]工作的基础上，结合 multi-gram 的策略，出色地完成了关系抽取任务。文献[14]总结了当前自然语言处理领域比较流行的卷积神经网络框架，分析了 multi-gram(该文中称作 region)对系统表现的影响，并认为 multi-gram 策略是提高卷积神经网络处理自然语言任务效果的有效途径。

当前基于 multi-gram 策略的卷积神经网络系统均是利用 multi-window 予以实现。有几种 n-gram，则有几种不同长度的滑动窗口(window)，它们并行地进行卷积和 pooling 操作，运算出不同 n-gram 下的句子特征向量的表示。然后，在输入最后的全连接层(full-connected layer)前，将这些特征向量首尾相连组成一个较长的特征向量表示，称其为句子向量。这个句子向量的维度决定着全连接层的权值矩阵大小，而矩阵的大小又决定着需要学习的参数个数。参数越多，网络拟合能力越好，但过拟合的问题也会越严重[15]，而且训练时间会增加，得到最优参数集的难度也会变大。因此，在不影响深度学习系统表现的前提下，减少参数的数量是个明智的选择。

基于上面的阐述，针对自然语言处理领域的卷积神经网络框架，本文提出了一种参数更少但效果更好的自适应学习 multi-gram 权重的策略。在并行的 n-gram 输出特征向量表示之后，给每一种 n-gram 特征向量定义一个权重，并将这些向量在权重的作用下进行加和，得到一个与每个 n-gram 特征向量维度相同的向量作为 multi-gram 策略下的句子向量。这种改进使得最终生成的句子向量的维度并不随n-gram的个数而改变，而且减少了全连接层的参数个数，提高了整个深度学习框架的学习训练速度，并且在一定程度上降低了过拟合的风险。通过在电影评论正负倾向分类和关系分类两个任务上的一系列实验发现，在自适应学习 multi-gram 权重的策略帮助下，系统的训练速度和分类效果都得到了明显提升。这说明卷积神经网络的自动学习能力得到了优化，从而证明了所提方法的有效性。

2　卷积神经网络结构

图像和文本在组成成分和表示形式方面存在的差别，决定着其卷积神经网络结构也存在差异。图像是由像素点构成，而文本则是由词组成。相对而言，像素点是一种低级的表示，而词语本身就是一种较高级的抽象表示，其携带的信息量远大于像素点。因此，利用词语为基本单元作为输入的卷积神经网络结构的深度较图像识别得要浅层一些。本文阐述的卷积神经网络结构总体上由词向量映射层、卷积层、max-pooling 层、全连接层以及最后的 softmax 组成，如图 1 所示。

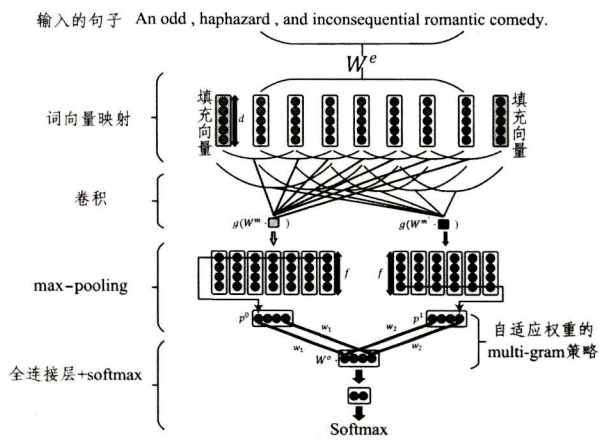

图 1　卷积神经网络结构图

2.1　词向量映射层

文本的基本组成单元是词语。词语是一种较高级的抽象表示，其本身包含丰富的信息。作为卷积神经网络的输入，如何表示词语能使其包含更好的语义和语法信息是非常重要的。词向量(word embedding)[16]是神经网络在自然语言领域应用的成功产物。词向量是利用神经网络的结构基于大规模语料集的无监督学习得到的词语的低维表示，并且能很好地表示词语间的语义相似性和语法特征。词向量的这种优秀特性决定了其被广泛应用于自然语言深度学习框架的输入的

1) http://www.csdn.net/article/2015-12-16/2826498

可行性和实效性。本文以句子为单位进行建模，每个句子可表示为由 N 个词语 x_n 组成的序列 $\{x_1, x_2, \cdots, x_n, \cdots, x_N\}$。$x_n$ 在整个词表中的 one-hot 表示向量 $h_n \in \{0,1\}^{|V|}$，其中 $|V|$ 表示词表的大小。利用文献[16]提出的方法生成词向量矩阵 $W^e \in \mathbb{R}^{d \times |V|}$。如图 1 所示，输入的句子中第 n 个词的词向量 v_n 可表示为：

$$v_n = W^e \cdot h_n \tag{1}$$

其中，$v_n \in \mathbb{R}^d$，d 代表词向量的维度。由于词向量是由大规模数据集无监督学习得到的，因此其语义相对于所做的任务针对性不强。常用的策略是将词向量也作为系统的参数进行调整，实践证明这样的处理会得到更好的实验效果。

2.2 卷积层

利用卷积神经网络做自然语言处理任务最重要的优势不仅在于其不依赖于人工标注特征和语言分析工具的语法信息，而且其可以通过网络训练的过程自动抽取需要的特征。自动抽取特征的工作主要由卷积层的卷积操作完成。图像是二维的输入，而文本则可以理解为一种一维的输入。那么卷积操作区域的概念需要重新定义。当前最被认可也最成功的语言模型是 n-gram，因此本文选择 n-gram 作为卷积的基本单元。为了便于理解，下面的描述均以一种窗口长度(n-gram)为例进行阐述，如图 1 所示，从卷积层开始，并行的两列代表不同的窗口长度。窗口的长度定义为 l，l 即代表 n-gram中的 n。每次卷积操作的输入则为连续的 l 个词的词向量组成的 n-gram 向量 c_j。

$$c_j = v_j \oplus v_{j+1} \oplus \cdots \oplus v_{j+l-1} \tag{2}$$

其中，$\oplus$ 表示连接操作(concatenate)，即将窗口内的词向量首尾相接组成一个更长的 n-gram 向量表示 $c_j \in \mathbb{R}^{dl}$。为了训练平衡，给输入句子中每个词语相同的训练次数，一般在输入句子的前后各加 $l-1$ 个 d 维零向量作为填充向量，如图 1 所示。设 N 为包含填充向量的句子长度，则输入句子的 n-gram 矩阵表示为 $C \in \mathbb{R}^{dl \times (N-l+1)}$。

$$C = [c_1, c_2, \cdots, c_j, \cdots, c_{N-l+1}] \tag{3}$$

卷积操作的本质可以理解为权重矩阵与 n-gram 矩阵的矩阵乘积运算。定义卷积矩阵为 $W^m \in \mathbb{R}^{f \times dl}$。矩阵的每一行代表一种特征映射层(feature map)，所以 f 代表特征映射层的个数。参数 f 是人工设定的，f 过小，代表提取的特征较少，可能会导致欠拟合；而 f 过大，则可能会造成提取的特征之间的冗余性大，而且可能导致过拟合。所以合适的 f 值的选取对系统效果的意义很重要。卷积操作[1]可以表示为：

$$Q = g(W^m \cdot C) \tag{4}$$

其中，$Q \in \mathbb{R}^{f \times (N-l+1)}$，$g(\cdot)$ 代表非线性激活函数。非线性激活函数的使用会增强网络的拟合能力，常用的有 Tanh 和 ReLU[17]。ReLU 函数是当前比较公认的适合深度学习的非线性激活函数，它具有强非线性和负激活值无差异性的特点，这种特性能让深度网络的稀疏性变大，从而降低冗余性。ReLU 可以表示为：

$$ReLU(x) = \begin{cases} x, & x \geqslant 0 \\ 0, & x < 0 \end{cases} \tag{5}$$

从 ReLU 的表达式也可以看出其求导计算也非常简单。

2.3 Max-pooling

从式(4)可以看出，卷积操作生成的特征矩阵 Q 的规模依赖于句子的长度 N，但句子的长度是变化的。一个规模不统一的特征矩阵是没有办法训练网络参数的，因此 max-pooling 起到了统一特征表示规格的作用，如图 1 所示。max-pooling 操作可以表示为：

$$p_i = \max_j \{q_{ij}\}, \forall j = 1, \cdots, (N-l+1) \tag{6}$$

特征矩阵 Q 经过 max-pooling 操作之后转换成了长度统一的向量 $p \in \mathbb{R}^f$。不仅如此，max-pooling 操作也可以理解为抽取在每种特征映射下表现最突出的 n-gram，这与传统的方法思路相似，而且降低了信息的冗余性，对效果的提升也起到重要的作用。

如图 1 所示，本文采用的是 multi-gram 策略，每种n-gram 生成一种特征表示向量。设一共有 K 种 n-gram，则最终生成的句子向量表示[2]为：

$$S = p^0 \oplus p^1 \oplus \cdots \oplus p^K \tag{7}$$

2.4 全连接层和 softmax

最后的全连接层和 softmax 起到的是分类作用，由句子向量 S 经过矩阵 $W^o \in \mathbb{R}^{L \times fK}$ 的线性映射成[3]类别向量 $o \in \mathbb{R}^L$，其中 L 代表任务所需要分类的类别总数。

$$o = W^o \cdot S \tag{8}$$

生成的类别向量 o 中的每一维的值代表输入的句子在这个类别上的信心指数。最后，通过 softmax 决策出概率最大的类。

$$p(i|x,\theta) = \frac{e^{o_i}}{\sum_{k=1}^{L} e^{o_k}} \tag{9}$$

其中，θ 代表深度网络需要学习的参数集合，即 $\theta = [W^e, W^m, W^o]$。

3 自适应权重的 multi-gram 策略

在自然语言处理领域，n-gram 是目前公认的对自然语言最合理的近似。第 1 节所述的卷积神经网络结合多种n-gram 进行特征提取，有利于抽取到更有价值的信息。那么，抽取到不同 n-gram 的向量表示后，如何将这些向量结合才能更好地利用这些信息则是一个需要考虑的问题。1.3 节中使用的是连接(concatenate)的方法，将多种的 n-gram 向量首尾相连成一个长的句子向量。这种方法虽然保留了不同n-gram的信息，但是同时也存在着冗余性，而且生成的句子向量长度取决于 n-gram 的种类数。句子向量越长，全连接层需要训练的参数也就越多，造成过拟合现象的可能性也就越大。因此，基于对这些问题的思考，本文提出了自适应权重的 multi-gram 策略。

3.1 基于自适应权重的 n-gram 向量加和

图 1 描述了本文提出的自适应权重的 multi-gram 策略。该图使用了 3-gram 和 4-gram 做特征提取。这两种 n-gram 分别生成了两种向量表示 p^0 和 p^1。如图 1 所示，本文使用的策略是将 p^0 和 p^1 进行向量相加运算。由于 p^0 和 p^1 的向量维度均与卷积层的特征映射的个数相同，因此满足了向量相加的前提条件。但由于相加运算可能会在一定程度上造成

[1] 本文的省略了变量偏置 b，在该卷积操作和后面的全连接层的线性变换均可以加偏置项，偏置项的引入可以加快神经网络的收敛速度。

[2] 图 1 中并没有给出这种表示的生成过程，详细图解可参考文献[14]。

[3] 此处也可以增加非线性激活函数变成非线性映射。

信息的丢失，因此本文引入权重 w_1 和 w_2 对信息进行合理的分配，则句子向量可以表示为：

$$S=w_1 p^0+w_2 p^1 \tag{10}$$

对于 K 种 n-gram 向量的情况，定义 K 个权重 $w=[w_1, w_2, \cdots, w_K]$。

句子向量的维度降低，从而也会减小 W^c 的规模。不仅减少了网络的参数个数，缩短网络的训练时间，而且有效地缓解了过拟合的可能性。

不同的任务和不同的数据集可能都会导致权重值分配的不同。所以在权重值的选取方面，采用由深度学习网络结构自适应学习的策略。权重值由损失函数进行梯度回溯学习得到，因此现在的参数集合为 $\theta=[W^e, W^m, W^c, w]$。

3.2　正则化操作

深度学习由于其突出的拟合能力，也给其带来了过拟合问题。如果深度网络的参数对于训练数据过度拟合，会让神经网络记住一些训练数据集特有的特征，这样会影响网络的预测效果。因此，正则化的应用也是必不可少[18]的。正则化的核心思想是限制参数值的量级，使其在一定范围内变化。所以，为了避免过拟合的发生，对定义的权重 w 也进行了正则化操作。

假设有 K 种 n-gram 向量 $\{p^0, p^1, \cdots, p^K\}$，则需要 K 个权重构成的权重向量 $w=[w_1, w_2, \cdots, w_K]$，因此定义正则化操作如下：

$$\| w \|_2 \leqslant \sqrt{r} \tag{11}$$

其中，r 是限定值，其值的大小由用户决定。可以将这个表达式理解为将权重向量 w 限定在一个半径为 $\sqrt{r}$ 的超球体内部。

4　实验

4.1　实验数据

本文分别采用电影评论正负倾向分类和关系分类两种语句建模的任务验证自适应权重的 multi-gram 策略的有效性。

所使用的数据集 MR 英文电影评论数据集[19]以句子为基本单位，该数据集包括 5331 句正倾向评论和 5331 句负倾向评论，所以类别数为 2，在图 1 中也有体现。整个数据集中，评论的最大句长为 56 个词，平均句长为 20 个词。本次实验分别从正负倾向评论集合中各抽取 4317 句作为训练集，480 句作为验证集，剩下的 534 句作为测试集。

关系分类任务的数据集为 SemEval-2010 Task 8[20]数据集。该数据集以句子为基本单位，包含 9 个明确的关系类别和 1 个“Other”类，因此类别数为 10。整个数据集包含 10717 个句子，每个类别的样本数比较平均，数据集中最大句长为 100 个词。本次取前 7000 句作为训练集，1000 句作为验证集，2717 句作为测试集。

4.2　参数设置

2.2 节阐述了针对本文提出的自适应权重的 multi-gram 策略而制定的正则化约束条件。其中参数 γ 为超参数，需要通过一系列实验来确定。图 2 给出了实验结果准确率随 r 的变化趋势。可以看出，随着 r 值的增大，准确率整体呈现先上升后下降的趋势。这说明，如果 r 值过小，可能造成权重值过小，这样不利于最后分类的决策；相反，如果 r 值过大，说明对权值量级的限制小，在学习率较大的情况下可能会造成权值爆炸(blow-up of weight)[15]的现象。因此，合适的 r 值有利于得到更好的实验结果。由图 2 可以看出，当 $r=3$ 时，准确率达到最高。

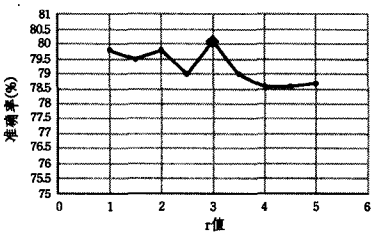

图 2　准确率随 γ 值的变化趋势

当不使用权重向量 w 时，可以将 w 理解为全 1 向量。对于本文使用{3,4,5} 3 种 n-gram 做特征抽取的情况，$\| w \|_2=\sqrt{3}$，即 $r=3$ 时，可以得到最好的实验结果。

另外，在词向量映射层，本文利用 Mikolov 的 word2vec[1)]工具，基于 2.5G 的 wiki 语料训练得到 W^e，词向量的维度设定为 300。参照文献[11]的工作，选取的 n-gram 集合为{3,4,5}。卷积层的特征映射个数为 100，即 W^m 的行数为 100。这个值的选取是根据一系列实验确定的，其值过小会导致欠拟合，其值多大会导致过拟合。迭代次数为 20。本文所用的卷积神经网络的非线性激活函数 $g(\cdot)$ 为 ReLU，并且结合 Adadelta 和 dropout 来提高实验效果。参数学习方面选择批量随机梯度下降算法，批度的选择为 50。

4.3　实验结果及讨论

本节通过在数据集 MR 和 SemEval-2010 Task 8 上的实验，给出了自适应权重的 multi-gram 策略在电影评论倾向分类任务和关系分类任务上的实验效果和分析。

表 1 给出了电影评论正负倾向分类的准确率对比结果和本文提出的自适应权重的 multi-gram 策略学习得到的权值分配。其中“向量连接”表示 1.3 节中提到的原始的 multi-gram 处理方法，“向量相加”为本文提出的方法。对于向量相加的方法，给出两组实验。一组不加权重；另一组加权重，并在卷积神经网络的训练过程中进行自适应调整。准确率分别在表 1 的第三行和第四行中给出。总体来说，对比结果显示，本文提出的自适应权重的 multi-gram 策略对整个系统的效果有提升；并且，通过有无权重的对比实验可以看出，自适应权重的使用对准确率的提高是非常明显的。值得注意的是，通过 3 种 n-gram 的权值分配结果可以发现，4-gram 的权值较其他两种 n-gram 的权值要大，可以理解为 4-gram 的信息对该任务的贡献是最大的。这个显现具有非常重要的实际意义，因为不同的任务，甚至不同的数据集，各异的 n-gram 提供的信息量的比例都会是不同的，那么通过系统的参数学习过程得到最适合的权重分配对实验效果的提升是非常有意义的。

表 1　电影评论正负倾向分类准确率及权值分配结合方法

结合方法	权重	准确率(%)	权值大小		
			3-gram	4-gram	5-gram
向量连接	无	79.5	—	—	—
向量相加	无	78.6	—	—	—
向量相加	有	**80.0**	0.7441	1.1240	1.0876

表 2 的形式与表 1 相同，给出了自适应权重的 multi-gram 策略在关系分类任务上的对比结果。可以看出，在无权重项情况下，本文提出的基于向量相加的 n-gram 特征向量结

1) https://code.google.com/p/word2vec/

合方法已经带来了 F1 值的提高；并且，增加自适应权重向量以后，关系分类的 F1 值有了进一步的提升。因此，结合电影评论正负倾向分类任务的结果，可以证明本文提出的自适应权重的 multi-gram 策略对卷积神经网络在语句建模方面的表现具有明显的改进和优化。不同的是，关系分类的 n-gram 权值分配中，5-gram 的权重最大，可以认为 5-gram 对 SemEval-2010 Task 8 数据集的关系分类任务的贡献最大。同时，进一步证明了自适应的 multi-gram 策略会自动找到更适合的权重分配方案。

表 2 关系分类任务 F1 值及权值分配

结合方法	权重	F1 值	权值大小		
			3-gram	4-gram	5-gram
向量连接	无	81.7	—	—	—
向量相加	无	82.3	—	—	—
向量相加	有	**82.7**	0.8015	1.0238	1.1443

表 3 给出了正负倾向各 3 个代表性的 4-gram 样例。在预测阶段，通过训练好的卷积神经网络，由输出回溯网络找到这些 4-gram。根据 1.3 节介绍的 max-pooling，取在 100 个特征映射上被 max 操作取出次数最多且非零的位置作为对输入的电影评论倾向激活最大的位置，然后找出这个位置代表的 4-gram。表 3 的结果直观地表明了卷积神经网络在做句子分类任务方面是非常有效的。

表 3 代表性的 4-gram 样例

类别	4-gram
正倾向	good fun good action an excellent romp that make it more interesting
负倾向	i did not laugh it is hardly watchable but it grows tedious

结束语 基于自然语言处理领域的卷积神经网络结合多种 n-gram 进行特征提取的特点，本文提出了一种基于自适应权重的 multi-gram 策略。这种策略不仅减少了网络结构的参数个数，降低了过拟合的风险，而且通过网络自主学习出不同 n-gram 的权重分配，自动找出对于任务具有最大的激活能力的 n-gram，并赋予其较大的权重。实验结果也表明，本文提出的改进方法对电影评论正负倾向性分类和关系分类的分类效果有明显的提升。虽然在当前的自然语言处理领域，卷积神经网络取得了非常显著的效果，但其也存在着一些不足。由于句子的长度是不统一的，因此 max-pooling 在句子长度这个维度上进行操作时会导致生成的句子向量丢失了语序信息。这也是卷积神经网络相对于循环神经网络和 LSTM 的劣势。而且，目前自然语言领域的深度学习方法多数情况下仅被作为工具使用，但语言是一种相对比较高级的、具有结构层次的表示，因此如何将语言的先验知识更好地融入到深度学习的网络学习中是一个非常值得思考的问题，也是我们下一步努力的方向。

参 考 文 献

[1] GRISHMAN R. Information extraction: Capabilities and challenges[Z]. Lecture Notes of 2012 International Winter School in Language and Speech Technologies, Roviral Virgili, 2012.

[2] WANG T, et al. End-to-end text recognition with convolutional neural networks[C]// 2012 21st International Conference on Pattern Recognition (ICPR). 2012.

[3] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[4] HINTON G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.

[5] HINTON G, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups [J]. Signal Processing Magazine, IEEE, 2012, 29(6): 82-97.

[6] NGUYEN T H, GRISHMAN R. Relation Extraction: Perspective from Convolutional Neural Networks[C]// Workshop on Vector Modeling for NLP. 2015: 39-48.

[7] IYYER M, et al. A neural network for factoid question answering over paragraphs[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014.

[8] LECUN Y, BENGIO Y. Convolutional networks for images, speech, and time series[M]// The Handbook of Brain Theory and Neural Networks. MIT Press, 1995.

[9] MOZER M C. A Focused Backpropagation Algorithm for Temporal Patern Recognition[M]. Hillsdale, 1995: 137-169.

[10] COLLOBERT R, et al. Natural language processing (almost) from scratch[J]. The Journal of Machine Learning Research, 2011(12): 2493-2537.

[11] KIM Y. Convolutional neural networks for sentence classification[J]. arXiv preprint arXiv: 1408.5882, 2014.

[12] CHEN Y, et al. Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015.

[13] ZENG D, et al. Relation classification via convolutional deep neural network[C]// Proceedings of COLING. 2014.

[14] ZHANG Y, Wallace B. A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification[J]. arXiv preprint arXiv: 1510.03820, 2015.

[15] HINTON G E, et al. Improving neural networks by preventing o-adaptation of feature detectors[J]. arXiv preprint arXiv: 1207.0580, 2012.

[16] MIKOLOV T, YIH W T, ZWEIG G. Linguistic Regularities in Continuous Space Word Representations[C]// HLT-NAACL. 2013.

[17] GLOROT X, BORDES A, BENGIO Y. Deep sparse rectifier neural networks[C]// International Conference on Artificial Intelligence and Statistics. 2011.

[18] ZEILER M D. ADADELTA: An adaptive learning rate method [J]. arXiv preprint arXiv: 1212.5701, 2012.

[19] PANG B, LEE L. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales[C]// Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2005.

[20] HENDRICKX I, et al. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals[C]// Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions. Association for Computational Linguistics, 2009.