ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.12, No.35, December 2016
E-mail: eduf@dnzs.net.cn
http://www.dnzs.net.cn
Tel:+86-551-65690963 65690964

基于多类型池化的卷积神经网络的文本分类算法

张添龙

（同济大学 计算机科学与技术系,上海 201804）

摘 要:准确地形式化表述句子对于语义理解是至关重要的。这篇文章中采用的卷积结构称为多类型池化的卷积神经网络,为句子建立语义模型。网络对线性序列进行多类型的动态池化操作。网络能够处理长度不等的句子并准确提取不同范围内词语之间的关系。网络不依赖于任何解析树,能够容易地应用于其他语言。对三组数据进行了测试:三组数据均取得较好的效果,优于或持平当前的最好文本分类算法。

关键词:深度学习;文本分类;多类型池化

中图分类号:TP393 文献标识码:A 文章编号:1009-3044(2016)35-0187-03

DOI:10.14004/j.cnki.ckt.2016.4974

1 引言

为了进行分类,我们建立句子模型来分析和表示句子的语义内容。句子模型问题的关键在于一定程度上的自然语言理解。很多类型的任务需要采用句子模型,包括情感分析、语义检测、对话分析、机器翻译等。既然单独的句子很少或基本不被采用,所以我们必须采用特征的方式来表述一个句子,而特征依赖于单词和词组。句子模型的核心是特征方程,特征方程定义了依据单词和词组提取特征的过程。求最大值的池化操作是一种非线性的二次抽样方法,它返回集合元素中的最大值。

各种类型的模型已经被提出。基于成分构成的方法被应用于向量表示,通过统计同时单词同时出现的概率来获取更长的词组。在有些情况下,通过对词向量进行代数操作生成句子层面的向量,从而构成成分。在另外一些情况下,特征方程和特定的句法或者单词类型相关。

一种核心模型是建立在神经网络的基础上。这种模型包含了单词包或者词组包的模型、更结构化的递归神经网络、延迟的基于卷积操作的神经网络。神经网络模型有很多优点。通过训练可以获得通用的词向量来预测一段上下文中单词是否会出现。通过有监督的训练,神经网络能够根据具体的任务进行良好的调节。除了作为强大的分类器,神经网络模型还能够被用来生成句子[6]。

我们定义了一种卷积神经网络结构并将它应用到句子语义模型中。这个网络可以处理长度不同的句子。网络中的一维卷积层和多类型动态池化层是相互交错的。多类型动态池化是一种对求最大值池化操作的范化,它返回集合中元素的最大值、最小值、平均值的集合[1]。操作的范化体现在两个方面。第一,多类型池化操作对一个线性的值序列进行操作,返回序列中的多个数值而不是单个最大的数值。第二,池化参数k可以被动态的选择,通过网络的其他参数来动态调整k的值。

卷积层的一维卷积窗口对句子特征矩阵的每一行进行卷积操作。相同的*n-gram*的卷积窗口在句子的每个位置进行卷积操作,这样可以根据位置独立地提取特征。一个卷积层后面是一个多类型动态池化层和一个非线性的特征映射表。和卷积神经网络在图像识别中的使用一样,为丰富第一层的表述,通过不同的卷积窗口应用到句子上计算出多重特征映射表。后续的层也通过下一层的卷积窗口的卷积操作计算出多重特征映射表。最终的结构我们叫它多类型池化的卷积神经网络。

在输入句子上的多层的卷积和动态池化操作产生一张结构化的特征图。高层的卷积窗口可以获取非连续的相距较远的词组的句法和语义关系。特征图会引导出一种层级结构,某种程度上类似于句法解析树。这种结构不仅仅是和句法相关,它是神经网络内部所有的。

我们将此网络在四种场景下进行了尝试。前两组实验是电影评论的情感预测[2],此网络在二分和多种类别的分类实验中的表现都优于其他方法。第三组实验在TREC数据集(Li and Roth, 2002)上的6类问题的分类问题。此网络的正确率和目前最好的方法的正确率持平。第四组实验是推特的情感预测,此网络将160万条微博根据表情符号自动打标来进行训练。在手工打标的测试数据集上,此网络将预测错误率降低了25%。

本文的概要如下。第二段主要阐述MCNN的背景知识,包括核心概念和相关的神将网络句子模型。第三章定义了相关的操作符和网络的层。第四章阐述生成的特征图的处理和网络的其他特点。第五章讨论实验和回顾特征学习探测器。

2 背景

MCNN的每一层的卷积操作之后都伴随一个池化操作。我们先回顾一下相关的神经网络句子模型。然后我们来阐述一维的卷积操作和经典的延迟的神经网络(TDNN)[3]。在加了一个最大池化层到网络后,TDNN也是一种句子模型[5]。

2.1 相关的神经网络句子模型

已经有很多的神经网络句子模型被描述过了。一种比较通用基本的模型是神经网络词包模型(NBoW)。其中包含了一

收稿日期:2016-11-10
基金项目:上海市科委科研计划项目(16JC1403000);上海市科技创新行动计划(14511108002)

个映射层将单词、词组等映射到更高的维度；然后会有一个比如求和之类的操作。结果向量通过一个或多个全连接层来进行分类。

有以外部的解析树为基础的递归神经网络，还有在此基础上更进一步的RNN网络。

最后一种是以卷积操作和TDNN结构为基础的神经网络句子模型。相关的概念是动态卷积神经网络的基础，我们接下来介绍的就是它。

2.2 卷积

一维卷积操作便是将权重向量 $m \in R^m$ 和输入向量 $s \in R^s$ 进行操作。向量 m 是卷积操作的过滤器。具体来说，我们将 s 作为输入句子，$s_i \in R$ 是与句子中第i个单词相关联的单独的特征值。一维卷积操作背后的思想是通过向量 m 和句子中的每个 $m-gram$ 的点积来获得另一个序列 c：

$$c_i = m^T s_{i-m+1:i} \quad (1)$$

根据下标 i 的范围的不同，等式1产生两种不同类型的卷积。窄类型的卷积中 $s >= m$ 并且会生成序列 $c \in R^{s-m+1}$，下标 i 的范围从 m 到 s。宽类型的卷积对 m 和 s 的大小没有限制，生成的序列 $c \in R^{s+m-1}$，下标 i 的范围从1到 $s+m-1$。超出下标范围的 s_i 窄（$i < 1$ 或者 $i > s$）置为0。窄类型的卷积结果是宽类型的卷积结果的子序列。

宽类型的卷积相比于窄类型的卷积有一些优点。宽类型的卷积可以确保所有的权重应用到整个句子，包括句子收尾的单词。当 m 被设为一个相对较大的值时，如8或者10，这一点尤其重要。另外，宽类型的卷积可以确保过滤器 m 应用于输入句子 s 始终会生成一个有效的非空结果集 c，与 m 的宽度和 s 句子的长度无关。接下来我们来阐述 $TDNN$ 的卷积层。

2.3 延迟的神经网络

$TDNN$ 包含一个输入序列 s 和一个权重向量 m 的集合。$TDNN$ 来进行音位识别[4]，序列 s 有一个时间维度，卷积操作应用于时间维度。每个 s_i 不是一个值，而是一个向量 d，所以 $s \in R^{d \times s}$。同样，m 是一个权重矩阵，大小为 $d \times m$。权重矩阵 m 的每一行与相应的输入矩阵 s 的每一行进行卷积操作并且采用窄类型的卷积操作。多重的卷积层被堆叠在一起，将产生的结果序列 c 作为下一层的输入。

$Max-TDNN$ 句子模型是建立在 $TDNN$ 结构的基础上[5]。在这个模型中，卷积层被应用于输入矩阵 s，矩阵的每一列的特征向量 $w_i \in R^d$ 代表句子中的一个单词：

$$s = \begin{bmatrix} | & | & | \\ w_1 & \dots & w_s \\ | & | & | \end{bmatrix} \quad (2)$$

为了处理句子长度会变化的问题，$Max-TDNN$ 取结果矩阵c每一行的最大值生成一个向量：

$$c_{\max} = \begin{bmatrix} \max(c_{1,:}) \\ \dots \\ \max(c_{d,:}) \end{bmatrix} \quad (3)$$

目的是为了获取最重要的特征，比如，取结果矩阵 c 中每行的最大值。固定大小的向量 c_{max} 作为全连接层的输入来进行分类。

$Max-TDNN$ 模型有许多优秀的特性。它对句子中单词的顺序非常敏感，并且不依赖于外部语言特定的特诊，比如依赖解析树。此模型给予句子中的每个单词同样的重要性，为了防止句中边缘的单词在窄类型的卷积操作被计算更少的次数。但是这个模型也有一些方面的限制。它会使忽略句子边缘的特性以及输入 s 的长度最小值变大的特性更差。由于这个原因，更长的特征探测器不能很好地适应这种模型。取最大值的池化操作也有一些缺点。它不能区分一个相关的特征在某一行中出现一次还是多次，它也不能记下特征出现的次序。下一章节主要是为了解决模型的这些限制同时保留模型原有的优点。

3 多类型池化的卷积神经网络

我们的句子模型使用卷积结构，其中卷积层和多类型池化层相交替。网络中间层的宽度根据输入句子的长度而变化；最后的结构就是多类型池化的卷积神经网络。接下来我们详细的阐述它。

3.1 宽类型的卷积

根据输入的句子获取 $MCNN$ 的第一层，我们用向量 $w_i \in R^d$ 来表示句子中的每个单词，从而如Eq. 2构建句子矩阵 $s \in R^{d \times s}$。参数 w_i 的值在训练的时候会被优化。通过卷积权重矩阵 $m \in R^{d \times m}$ 和下面的激活矩阵，可以得到一个卷积层。比如，第二层是通过对句子矩阵 s 进行卷积操作得到的。维度 d 和过滤器的宽度 m 是网络的参数。我使用2.2节中描述的宽类型的卷积操作。结果矩阵 c 的维度即为 $d \times (s+m-1)$。

3.2 多类型池化

我们接下来描述的池化操作是 $Max-TDNN$ 模型中最大池化操作的范化版本，它和图像识别[1]中卷积神经网络的最大池化操作也不同。给出一个参数 k 和一个长度为 p 的序列 $p \in R^p$，并且 $p >= k$，$k-max$ 池化选择序列 p 中的最大的 k 个值作为子序列 $p^k_{\max}$。$p^k_{\max}$ 中值的顺序和它们原来在 p 中的顺序相同。$k-min$ 池化选择序列 p 中最小的 k 个值作为子序列 $p^k_{\min}$。$k-average$ 池化将序列 p 均分成 k 段，取每一段的平均值构成子序列 p^k_{ave}。

多类型动态池化操作可以池化 p 中最显著的 $3k$ 个特征，而这 $3k$ 个特征在 p 中可能不是连续的，它对特征的具体位置并不敏感。在最高层的卷积层之后应用k多类型池化操作。这就能保证全连接层的输入与句子的长度无关。但在中间卷积层的池化参数k不是固定的而是动态生成的，为的是能够更加平滑地提取高层的特征。

3.3 动态多类型池化

动态多类型池化操作中k是句子长度和网络深度的函数。尽管有许多函数是可行，我可以按如下方式简单地定义池化参数：

$$k_l = \max(k_{top}, \left\lceil \frac{L-l}{L} s \right\rceil) \quad (4)$$

其中 l 是当前应用池化操作的卷积层的层数，L 是网络中总共的卷积层层数；k_{top} 是为最顶层的卷积层设置的固定参数。比如，在一个有3层卷积层的网络中，$k_{top} = 3$，输入句子的长度 $s = 18$，所以第一层的池化参数 $k_1 = 12$，第二层的池化参数 $k_2 = 6$，第三层有固定的池化参数 $k_3 = k_{top} = 3$。

3.4 非线性特征函数

在卷积操作之后进行动态多类型池化操作得到一个池化矩阵，我们对其每一个分量进行偏置和非线性化操作。我们接下来陈述如何矩阵 a 中的每一列a。我们定义 M 为对角矩阵：

$$M = [\text{diag}(m_{m:,1}), \dots, \text{diag}(m_{:,m})] \quad (5)$$

$$a=\left(M\begin{bmatrix}w_j\\ \dots\\ w_{j+m-1}\end{bmatrix}+b\right)$$(6) m是卷积操作中的权重。经过卷积层和非线性层之后，矩阵a的第j列可用如下公式获得：

a是矩阵中的一列。等式6表述了特征提取函数的一种方式并且有更加普遍的形式。结合池化操作，可以保持位置不变性。

3.5 多重特征映射

所以到目前为止，我们已经学会了对输入矩阵进行卷积、动态多类型池化、非线性化操作来获得第一层的特征映射。重复这3个操作来获得更高层的特征映射，网络的层数也更深。我们用F^i来第i层的特征映射。正如用卷积神经网络来进行图像识别，在同一层的多重的特征映射$F^i_1,\cdots,F^i_n$会被并行计算。通过对$i-1$层的每个特征映射F^{i-1}_k和过滤矩阵$m^i_{j,k}$进行卷积并求和得到F^i_j。

$$F^i_j=\sum_{k=1}^{n}m^i_{j,k}*F^{i-1}_k$$(7) 公式中 * 表示宽类型的卷积操作。

3.6 折叠

所特征探测器被应用于句子矩阵s的每一行，这会在同一行的多重特征映射中产生复杂的依赖关系。在全连接层之前的网络层中，不同行的特征探测器是相互独立的。通过等式5的对角矩阵M的转化可以实现不同行的全依赖。这里我们介绍一种更加方法叫做折叠，而且这种方法并不会引入任何其他参数。在卷积层和动态多类型池化层之间，我们将每两行相加。因此，对于一个d行的矩阵，折叠会使这个矩阵变为$d/2$行，将其大小变为原来的一半。有了折叠层以后，第i层的任一特征探测器将依赖于$i-1$层矩阵的两行的特征值。这就是对*MCNN*所有的描述。

4 实验与结果分析

我们对此网络进行了4组不同的实验。

4.1 电影评论的情感预测

前两组实验是关于电影评论的情感预测的，数据集是*Stanford Sentiment Treebank*.实验输出的结果在一个实验中是分为2类，在另一种试验中分为5类：消极、略微消极、中性、略微积极、积极。而实验总的词汇量为15448。

表 1

分类器	5类(%)	2类(%)
NB	41.0	81.8
BINB	41.9	83.1
SVM	40.7	79.4
MAX-TDNN	37.4	77.1
NBoW	42.4	80.5
MCNN	48.5	86.8

表1表示的是电影评论数据集情感预测准确率。*NB*和*BINB*分别表示一元和二元朴素贝叶斯分类器。*SVM*是一元和二元特征的支撑向量机。在三种神经网络模型里——*Max-TDNN*、*NBoW*和*DCNN*——模型中的词向量是随机初始化的；它们的维度d被设为48。*Max-TDNN*在第一层中滤波窗口的大小为6。卷积层后面紧跟一个非线性化层、最大池化层和*softmax*分类层。*NBoW*会将词向量相加，并对词向量进行非线性化操作，最后用*softmax*进行分类。2类分类的*MCNN*的参数如下，卷积层之后折叠层、动态多类型池化层、非线性化层。滤波窗口的大小分别7和5。最顶层动态多类型池化层的k的值为4。网络的最顶层是*softmax*层。5类分类的*MCNN*有相同的结构，但是滤波窗口的大小分别为10和7，k的值为5。

我们可以看到*MCNN*的分类效果远超其他算法。*NBoW*的分类效果和非神经网络算法差不多。而*Max-TDNN*的效果要比*NBoW*的差，可能是因为过度池化的原因，丢弃了句子太多重要的特征。除了*RecNN*需要依赖外部的解析树来生成结构化特征，其他模型都不需要依赖外部资源。

4.2 问题分类

问题分类在问答系统中应用非常广泛，一个问题可能属于一个或者多个问题类别。所用的数据集是*TREC*数据集，*TREC*数据集包含6种不同类别的问题，比如一个问题是否关于地点、人或者数字信息。训练集包含5452个打标的问题和500个测试集。

表 2

分类器	准确率(%)
MAXENT	92.6
SVM	95.0
MAX-TDNN	84.4
NBoW	88.2
MCNN	93.0

表2显示的是问题分类的结果。三种神经网络的结构和参数与4.1节的结构参数基本相同。由于数据集相对较小，我们将词向量的维度设小一些$d=32$。可以看出，*MCNN*的分类效果和依赖其他外部数据的最好的分类器的效果基本相同。

4.3 *Twitter*情感预测

在我们最后的实验里，我们用*tweets*的大数据集进行训练，我们根据*tweet*中出现的表情符号自动地给文本进行打标签，积极的或是消极的。整个数据集包含160万条根据表情符号打标的*tweet*以及400条手工标注的测试集。整个数据集包含76643个单词。*MCNN*的结构和4.1节中结构相同。随机初始化词向量且维度d设为60。表3表示的是实验结果：

表 3

分类器	准确率(%)
SVM	81.6
BINB	82.7
MAX-TDNN	78.8
NBoW	80.9
MCNN	87.4

（下转第211页）

网络模型为例，构建小型猫的数据集，提取猫的图象特征信息，最后和目标猫图像进行预测，并和传统的图像分类算法进行对比，预测的准确率有很大的提升。

参考文献：

[1] 杨铮, 吴陈沭, 刘云浩. 位置计算: 无线网络定位与可定位性[M]. 北京: 清华大学出版社, 2014.

[2] 丁士折. 人工神经网络基础[M]. 哈尔滨: 哈尔滨工程大学出版社, 2008.

[3] McClelland J L, Rumelhart D E, PDP Research Group. Parallel distributedprocessing[J]. Explorations in the microstructure of cognition, 1986, 2.

[4] Hopfield J J. Neural networks and physical systems with emergent collectivecomputational abilities[J]. Proceedings of the national academy of sciences, 1982, 79(8): 2554-2558.

[5] Ackley D H, Hinton G E, Sejnowski T J. A learning algorithm for boltzmannmachines[J]. Cognitive science, 1985, 9(1): 147-169.

[6] Kohonenmaps T. Self-Organized Formation of Topologically Correct Feature Maps[J]. Biological Cybernetics,1982, 43(1): 59-69.

[7] Carpenter G A, Grossberg S. A massively parallel architecture for aself-organizing neural pattern recognition machine[J]. Computer visiongraphics, and image processing, 1987, 37(1): 54-115.

[8] Jia Y, Shelhamer E, Donahue J, et al. Caffe: Convolutional architecture for fast feature embedding[C]//Proceedings of the ACM International Conference on Multimedia. ACM, 2014: 675-678.

（上接第189页）

我们发现*MCNN*的分类效果和其他非神经网络的算法相比有极大的提高。*MCNN*和*NBoW*在分类效果上的差别显示了*MCNN*有极强的特征提取能力。

5 结语

在本文中我们阐述了一种动态的卷积神经网络，它使用动态的多类型池化操作作为非线性化取样函数。此网络在问题分类和情感预测方面取得了很好的效果，并且不依赖于外部特征如解析树或其他外部资源。

参考文献

[1]. Yann LeCun, Le´on Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278-2324, November.

[2]. Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013b. Recursive deep mod- els for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Process- ing*, pages 1631 - 1642, Stroudsburg, PA, October. Association for Computational Linguistics.

[3]. Geoffrey E. Hinton. 1989. Connectionist learning procedures. *Artif. Intell.*, 40(1-3):185 - 234.

[4]. Alexander Waibel, Toshiyuki Hanazawa, Geofrey Hinton, Kiyohiro Shikano, and Kevin J. Lang. 1990. Readings in speech recognition. chapter Phoneme Recognition Using Time-delay Neural Networks, pages 393 - 404. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[5]. Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *International Conference on Machine Learning, ICML*.

[6]. Holger Schwenk. 2012. Continuous space translation models for phrase-based statistical machine translation. In *COLING (Posters)*, pages 1071-1080.