自然语言处理系列（9）：DCNN

这篇文章翻译自A Convolutional Neural Network for Modelling Sentences

准确地表达句子的能力是语言理解的核心。我们描述了一个被称为动态卷积神经网络(DCNN)的卷积结构，我们采用的是句子的语义模型。该网络使用动态k-Max池，一个通过线性序列的全局池操作。该网络处理不同长度的输入句子，并在句子中归纳出一个特征图，可以明确地捕捉短和长期的关系。该网络不依赖于解析树，并且很容易适用于任何语言。我们在四个实验中对DCNN进行了测试:小尺度二进制和多类情绪预测，六道问题分类和远程监控的推特情绪预测。该网络在前三项任务中取得了出色的性能，并且在最后一项任务中，在最强大的Baseline上减少了25%的错误。

1 Introduction

句子模型的目的是为了分析和表征句子的语义内容，以达到分类或生成的目的。句子建模问题是许多涉及到一定程度的自然语言理解的任务的核心。这些任务包括情绪分析、意译、识别、总结、语篇分析、机器翻译、基础语言学习和图像检索。由于个别句子很少被观察到或根本没有被观察到，所以一个句子必须根据经常被观察到的句子中的单词和短的n-grams来表示一个句子。句子模型的核心是一个特征函数，它定义了从单词或n-grams的特征中提取句子特征的过程。

各种各样的模型方法已经被提出来了。基于组合的方法已经应用于从共现统计中得到的单词意义的向量表示，从而获得更长的短语的向量。在某些情况下，组合是由代数运算来定义的，而不是词义向量来生成句子的意思 (Erk and Pado´, 2008; Mitchell and Lapata, 2008; Mitchell and Lapata, 2010; Tur- ney, 2012; Erk, 2012; Clarke, 2012).。在其他情况下，一个复合函数被学习，或者与特定的句法关系有关(Guevara, 2010; Zanzotto et al., 2010) 或特定的词类型相关联(Baroni and Zamparelli, 2010; Coecke et al., 2010; Grefenstette and Sadrzadeh, 2011; Kart- saklis and Sadrzadeh, 2013; Grefenstette, 2013)。另一种方法是用自动提取的逻辑形式来表示句子的意义(Zettlemoyer and Collins, 2005)。

核心模型是基于神经网络的模型。这些范围从基本的神经词袋或n-grams模型到更结构化的递归神经网络，以及基于卷积运算的时滞神经网络(Collobert and Weston, 2008; Socher et al., 2011; Kalchbrenner and Blunsom, 2013b)。神经网络句模型有很多优点。他们可以通过预测，例如，利用单词和短语的上下文，来获得单词和短语的一般向量。通过监督训练，神经网络句模型可以将这些向量调整为特定于某项任务的信息。除了将强大的分类器作为其体系结构的一部分之外，还可以使用神经句模型来建立一个神经语言模型来逐字生成句子 (Schwenk, 2012; Mikolov and Zweig, 2012; Kalchbrenner and Blunsom, 2013a)。

我们定义了一个卷积神经网络结构，并将其应用于句子的语义建模。该网络处理不同长度的输入

序列。在网络中，层间层交织着一维的卷积层和动态k-max池化层。动态k-max池是最大池操作的一般化。最大池操作符是一个非线性的子采样函数，它返回一组值的最大值 (LeCun et al., 1998)。该算子在两个方面是广义的。首先，k-max在一个线性序列上集合，返回序列中k最大值的子序列，而不是单个最大值。其次，可以通过将k作为网络的另一种功能或输入来动态选择池参数k。

卷积层在句子矩阵的每一列特征上都应用了一个可编辑的过滤器。在句子中的每一个位置上，将相同的过滤器与n-gram进行卷积，使得这些特征可以独立于句子中的位置提取出来。一个卷积层接着是一个动态池层和一个非线性形式的feature map。就像在对象识别的卷积网络中，我们通过对输入语句应用不同的过滤器来计算多个特征映射来丰富第一层的表示。后续层还具有多个特征映射，这些特征映射通过与下面一层的所有映射进行卷积来计算。这些层的权值构成一个四维张量。由此产生的结构被称为动态卷积神经网络。

卷积和动态池操作的多层结构会在输入语句中引入结构化特征图。图1展示了这样一个图。在较高层次的小筛选器可以捕获在输入句中相隔很远的非连续短语之间的句法或语义关系。特征图引入了一种类似于语法分析树的层次结构。这种结构与纯句法关系无关，是神经网络的内部结构。

我们在四种环境中进行网络实验。前两个实验包括预测电影评论的情绪(Socher et al., 2013b)。该网络在二进制和多类实验中都优于其他方法。第三个实验涉及在TREC的数据集下的六个问题类型的分类 (Li and Roth, 2002)。该网络与其他基于大量工程特性和手工编码的知识资源的先进方法的准确性相一致。第四项实验是通过远程监控来预测Twitter帖子的人气(Go et al., 2009)。该网络接受了160万条推文，根据出现在他们身上的表情符号自动标记。在手工标记的测试集上，在Go et al.(2009)中所报告的最强的unigram和bigram基线的预测误差中，该网络减少了超过25%。

论文的提纲如下。第2节描述了DCNN的背景，包括中心概念和相关的神经句模型。第3节定义了相关的操作符和网络的层。第4节介绍了网络的诱导特征图和其他属性。第5节讨论了实验，并考察了学习的特征检测器。

2 Background

DCNN的层是由一个卷积操作和一个池化操作形成的。我们首先回顾一下相关的神经网络句模型。然后描述了一维卷积和经典时滞神经网络的运算(TDNN) (Hinton, 1989; Waibel et al., 1990)。通过在网络中加入一个最大池化层，TDNN可以被作为一个句子模型来使用(conbert and Weston, 2008)。

2.1 Related Neural Sentence Models

各种各样的神经网络语言模型已经被提出。基本句型模型的一般类型是神经网络词汇(NBoW)模型。这些通常包括一个投影层，将单词、子字单元或n-grams映射到高维的嵌入。然后将组件与

操作(如求和)组合在一起。所产生的组合向量通过一个或多个全连接层进行分类。

采用外部解析树提供的更一般结构的模型是递归神经网络 (RecNN) (Pollack, 1990; Ku¨chler and Goller, 1996; Socher et al., 2011; Hermann and Blunsom, 2013)。在树的每个节点上，节点的左右子节点都由一个经典层组合起来。该层的权值在树的所有节点上共享。在顶部节点计算的层给出了这个句子的表示。递归神经网络(RNN)是递归网络的一种特殊情况，它所遵循的结构是一个简单的线性链(Gers and Schmidhuber, 2001; Mikolov et al., 2011)。RNN主要作为一种语言模型，但也可以被看作是一个线性结构的句子模型。在最后一个单词中计算的层代表这个句子。最后，基于卷积运算和TDNN架构，进一步构建了一类神经语句模型 (Collobert and Weston, 2008; Kalchbrenner and Blunsom, 2013b)。这些模型中使用的某些概念是DCNN的核心，接下来我们将描述它们。

2.2 Convolution

一维卷积是一个权重向量$m \in R^m$和一个被视为序列的输入向量$s \in R^s$之间的运算。向量m是卷积的滤波器。具体地说，我们认为s是输入语句，s_i是与句子中第i个单词相关的单一特征值。一维卷积背后的思想是在句子中求每个m-gram与向量m的点积，从而获得另一个序列c：

$$c_j = m^T s_{j-m+1:j} \qquad (1)$$

根据j的取值范围，公式1给出了两种类型的卷积。窄卷积要求$s \geq m$，序列$c \in R^{s-m+1}$的j的范围是m到s。宽卷积对s或m没有大小要求，序列$c \in R^{s-m+1}$的索引j的范围是1到$s+m-1$。超出范围($i < 1$或者$i > s$)的输入值s_i会被设为零，狭义卷积的结果是广义卷积结果的子序列。两种一维卷积在图2中得到了说明。

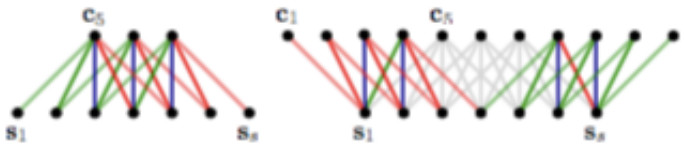

Figure 2: Narrow and wide types of convolution. The filter $\mathbf{m}$ has size $m = 5$.

过滤器m中训练的权重相当于一个语言特征检测器，它学习对一个特定的n-gram进行编码，这些n-grams为大小n≤m,m是滤波器的宽度。在一个宽的卷积中应用权重m比在一个窄的卷积中应用它们有一些优势。一个宽卷积可以确保过滤器中的所有权重达到整个句子，包括边缘的单词。当m被设置为相对较大的值(比如8或10)时，这是非常重要的。此外，宽卷积保证了滤波器m对输入语句的应用总是产生一个有效的非空结果c，独立于宽度m和句子长度s。我们接下来描述一个TDNN的经典卷积层。

2.3 Time-Delay Neural Networks

TDNN将一系列的输入序列与一组权重m进行卷积。在语音识别的TDNN (Waibel et al.， 1990)中，序列s被视为具有时间维度，而卷积应用于时间维度。每一个s_j往往不只是一个单一的值，而是一个d个数字组成的向量，因此$s \in R^{d \times s}$ 。同样,m是一个大小为d×m权重矩阵。每一行m与相应的s行卷积，卷积通常是窄类型的。将生成的序列c作为输入到下一层，可以叠加多个卷积层。

Max-TDNN句子模型是基于TDNN的体系结构 (Collobert and Weston, 2008)。在模型中，窄卷积被应用到句子矩阵s上，其中每一列对应句子中的一个词的特征向量$w_i \in R^d$

$$s = [w_1 \dots w_s] \qquad (2)$$

为了解决不同的句子长度问题，Max-TDNN在生成的矩阵c中取每一行的最大值，从而产生一个d个数组成的向量。

$$C_{max} = \begin{bmatrix} max(c_{1,:}) \\ \cdots \\ max(c_{d,:}) \end{bmatrix} \qquad (3)$$

目标是捕捉最相关的特性，即对于生成的矩阵c的d行中的每一个具有最高值的特性。然后将固定大小的向量cmax作为输入，用全连接层进行分类。

Max-TDNN模型有许多可取的特性。它对句子中单词的顺序很敏感，它不依赖于外部语言特定的特性，如依赖关系或选区解析树。它对句子中每个单词的信号都有很大程度的统一的重要性，除了边缘上的单词，在计算窄卷积时次数更少。但该模型也有一些局限性。特征检测器的范围仅限于权重的跨度m。增加m或叠加多卷积层的窄类型使特征检测器的范围更大。与此同时，它也加剧了对句子边缘的忽视，增加了卷积所需要的输入语句的最小规格s。由于这个原因，高阶和远程特性检测器不能很容易地并入模型中。最大池操作也有一些缺点。它无法区分一个行中的相关特性是否只发生一次或多次，它会忘记特征发生的顺序。更普遍的是,池因子的减少信号矩阵的对应于s–m + 1。即使是对于中等的s，池因子也可能是过量的。下一节的目的是在保留优势的同时解决这些限制。

3 Convolutional Neural Networks with Dynamic k-Max Pooling

我们使用一个卷积结构来建模句子，它通过动态的k-max池来交替使用宽卷积层和动态池层。在网络中，中间层特征图的宽度取决于输入句子的长度。所得到的体系结构是动态卷积神经网络。图3表示一个DCNN。我们开始详细描述这个网络。

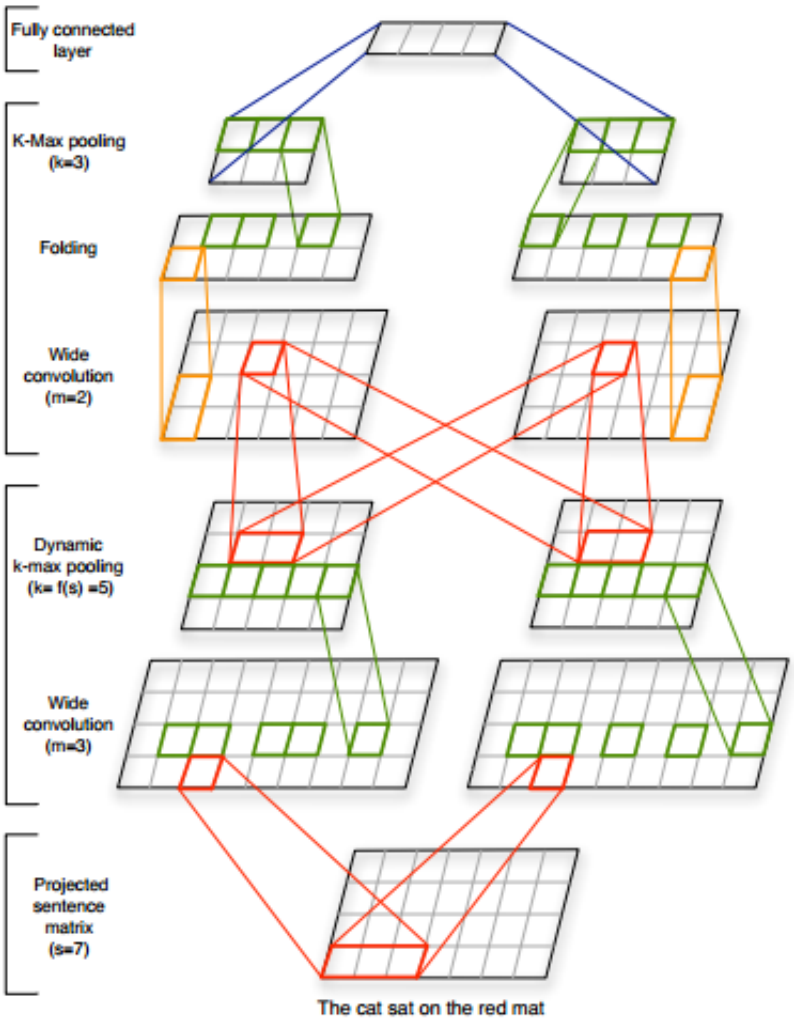

Figure 3: A DCNN for the seven word input sentence. Word embeddings have size $d = 4$. The network has two convolutional layers with two feature maps each. The widths of the filters at the two layers are respectively 3 and 2. The (dynamic) k-max pooling layers have values k of 5 and 3.

3.1 Wide Convolution

给定一个输入句子，获得DCNN的第一层，其中，我们得到每个句子的每个词的词嵌入表示$w_i \in R^d$，然后构造一个句子矩阵$s \in R^{d \times s}$，如（2）式所示。嵌入的值w_i是在训练期间优化的参数。网络中的卷积层是通过对权值矩阵m和下一层的激活矩阵进行卷积操作得到的。例如，第二层是通过将卷积应用到句子矩阵s本身来获得的。维度d和滤波器宽度m是网络的超参数。我们让这些操作是一维宽卷积，如第2.2节所述。由此产生的矩阵c的维度为$d \times (s + m - 1)$

3.2 k-Max Pooling

接下来我们描述的是一个池化操作，它是在Max-TDNN语言模型中的关于时间维度的最大池化操

作的一般化，不同于用于目标检测的卷积网络中使用的局部最大池化操作 (LeCun et al., 1998)。给定一个值k，和一个长度为p的序列$p \in R^p$ (p>k)，$k-maxpooling$选择了一个由向量p中最大的数值组成的子序列p^k_{max}。p^k_{max}中值的顺序对应于它们在p中的原始顺序。

k-max池操作使得选出在p中的k个最活跃特征成为可能，它们位于一些不同的位置。它保留了特征的顺序，但对它们的特定位置不敏感。它还可以更精细地分辨出在p中，特征被高度激活的次数，以及在p上的特征变化的被高度激活的状况。将k-max池运算符应用于最顶层卷积层后的网络中。这保证了全连接层的输入独立于输入句子的长度。但是，正如我们接下来看到的，在中间卷积层中，池参数k不是固定的，而是动态选择的，以便能够平滑地提取更高阶和更长的特性。

3.3 Dynamic k-Max Pooling

一个动态k-max池操作是一个k-max池操作，我们让k成为句子长度和网络深度的函数。虽然可能有很多函数，但我们只是简单地将池参数建模如下：

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

其中l表示当前卷积的层数（即第几个卷积层），L是网络中总共卷积层的层数；k_{top}为最顶层的卷积层pooling对应的k值，是一个固定的值。举个例子，例如网络中有三个卷积层，$k_{top}=3$，输入的句子长度$s=18$；那么，对于第一层卷积层下面的pooling参数$k_1=12$，而第二层卷积层的pooling参数$k_2=6$，第三层有固定的池化参数$k_3=k_{top}=3$。方程4是描述第l部分的序列的相关部分在长度为s的句子上的相关部分所需要的数值数量的模型。对于情绪预测中的一个例子，根据方程，在长度s的句子中，一阶特征，例如一个正数，大多数出现k_1次，而另一个二阶特征，如否定句或子句，最多出现k_2次。

3.4 Non-linear Feature Function

在动态池化应用于卷积的结果之后，一个偏置项$b \in R^d$和一个非线性函数g被应用于被池化的矩阵的每一个元素。每个集合矩阵的每一行都有一个单偏差值。如果我们暂时忽略池化层，我们可以说明在卷积层和非线性层后如何计算矩阵a中的d维的列a。定义M为对角矩阵。

$$M = [diag(m:,1), \cdots, diag(m:,m)] \qquad (5)$$

其中m为宽卷积的d维滤波器的权值。然后，在第一对卷积和一个非线性层之后，矩阵a中的每一列a得到如下所示，对于一些索引j：

$$a = g\left(M\begin{bmatrix} w_j \\ \cdots \\ w_{j+m-1} \end{bmatrix} + b\right) \qquad (6)$$

这里a是一列的一级特征。二阶特征也类似于将公式（6）应用于一个序列的一阶特性$a_j, \ldots a_{j+m'-1}$与另一个权重矩阵m'。除了池，公式（6）是特征提取函数的核心，它有一个一般形式。特征函数与池化才操作引入了位置不变性，使高阶特征变量的范围变大。

3.5 Multiple Feature Maps

到目前为止，我们已经描述了如何应用一个宽卷积，一个(动态的)k-max池化层和一个非线性函数到输入句子的矩阵来获得一阶特征图。这三种操作可以重复，以生成递增顺序的特征图和增加深度的网络。我们用F_i来表示第i阶的特征图。在对目标识别的卷积网络中，为了增加特定顺序的学习特征检测器的数量，可以在同一层并行计算多个特征映射。每一个特征映射F_j^i是通过不同的滤波器来对矩阵$m_{j,k}^i$和第$i-1$阶的特征映射F_k^{i-1}进行卷积操作，并把结果加和：

$$F_j^i = \sum_{k=1}^{n} m_{j,k}^i * F_k^{i-1} \qquad (7)$$

其中$*$表示宽卷积。权重$m_{j,k}^i$构成一个4阶张量。在宽卷积之后，首先进行动态k-max池化，然后将非线性函数分别应用于每个映射。

3.6 Folding

到目前为止，在网络的形成过程中，应用于句子矩阵的单个行的特征检测器可以有多种顺序，并在多个特征映射的同一行中创建复杂的依赖关系。然而，不同行的特征检测器相互独立，直到顶部的全连接层。可以通过在Eq. 5中使M变成满矩阵来实现完全依赖于不同的行，而不是一个稀疏矩阵的对角线。在这里，我们探索一种更简单的方法，称为折叠，它不引入任何其他参数。在一个卷积层之后，动态k-Max池化之前，将一个feature map组件中的每两行相加。对于d行的映射，折叠返回d/2行的映射，从而使表示的大小减半。使用折叠层之后，第i层的特征探测器现在依赖于更低的第$i-1$层的两行特征值。这样我们结束了对DCNN的描述。

4 Properties of the Sentence Model

我们描述了基于DCNN的句子模型的一些属性。我们描述了由卷积和池化层的继承而导致的特征图的概念。我们简单地将这些性质与其他神经句模型的性质联系起来。

4.1 Word and n-Gram Order

其中一个基本属性是对输入句子中单词顺序的敏感性。对于大多数应用，为了学习细粒度的特性检测器，对于模型来说，能够区分输入中是否存在特定的n-gram是有益的。同样，对于模型来说，能够分辨出最相关的n-grams的相对位置是有益的。该网络旨在捕捉这两个方面。第一层的宽卷积的滤波器m可以学习识别具有小于或等于滤波器宽度m的特定n-grams。正如我们在实验中看到的，在第一层中，m通常被设置为一个相对较大的值，比如10。由广义池操作提取的n-grams子序列，将不变性引入到绝对位置，但维持了它们的顺序和相对位置。至于其他的神经句模型，NBoW模型的类根据定义对词序不敏感的。一个基于递归神经网络的句子模型对词序很敏感，但它倾向于将其作为输入的最新单词 (Mikolov et al., 2011)。这使得RNN在语言建模方面表现出色，但在输入句中，它是在记忆n-grams时是最不理想的。类似地，递归神经网络对词序很敏感，但对树中最顶层的节点有偏倚。浅树可以在一定程度上缓解这种影响 (Socher et al., 2013a)。在第2.3节中，max - tdnn对词序是敏感的，但是max池只在句子矩阵的每一行中选出一个单一的n-grams特征。

4.2 Induced Feature Graph

模型不需要任何的先验知识，例如句法依存树等，并且模型考虑了句子中相隔较远的词语之间的语义信息；

5 Experiments

> 模型训练及参数

- 输出层是一个类别概率分布（即softmax），与倒数第二层全连接；
- 代价函数为交叉熵，训练目标是最小化代价函数；
- L2正则化；
- 优化方法：mini-batch + gradient-based (使用Adagrad update rule, Duchi et al., 2011)

> 实验结果

在三个数据集上进行了实验，分别是(1)电影评论数据集上的情感识别，(2)TREC问题分类，以及(3)Twitter数据集上的情感识别。结果如下图：

Classifier	Fine-grained (%)	Binary (%)
NB	41.0	81.8
BINB	41.9	83.1
SVM	40.7	79.4
RECNTN	45.7	85.4
MAX-TDNN	37.4	77.1
NBOW	42.4	80.5
DCNN	**48.5**	**86.8**

Table 1: Accuracy of sentiment prediction in the movie reviews dataset. The first four results are reported from Socher et al. (2013b). The baselines NB and BINB are Naive Bayes classifiers with, respectively, unigram features and unigram and bigram features. SVM is a support vector machine with unigram and bigram features. RECNTN is a recursive neural network with a tensor-based feature function, which relies on external structural features given by a parse tree and performs best among the RecNNs.

Classifier	Features	Acc. (%)
HIER	unigram, POS, head chunks NE, semantic relations	91.0
MAXENT	unigram, bigram, trigram POS, chunks, NE, supertags CCG parser, WordNet	92.6
MAXENT	unigram, bigram, trigram POS, wh-word, head word word shape, parser hypernyms, WordNet	93.6
SVM	unigram, POS, wh-word head word, parser hypernyms, WordNet 60 hand-coded rules	95.0
MAX-TDNN	unsupervised vectors	84.4
NBOW	unsupervised vectors	88.2
DCNN	unsupervised vectors	93.0

Table 2: Accuracy of six-way question classification on the TREC questions dataset. The second column details the external features used in the various approaches. The first four results are respectively from Li and Roth (2002), Blunsom et al. (2006), Huang et al. (2008) and Silva et al. (2011).

Classifier	Accuracy (%)
SVM	81.6
BINB	82.7
MAXENT	83.0
MAX-TDNN	78.8
NBOW	80.9
DCNN	**87.4**

Table 3: Accuracy on the Twitter sentiment dataset. The three non-neural classifiers are based on unigram and bigram features; the results are reported from (Go et al., 2009).

可以看出，DCNN的性能非常好，几乎不逊色于传统的模型；而且，DCNN的好处在于不需要任何的先验信息输入，也不需要构造非常复杂的人工特征。

6 Conclusion

我们描述了一个动态的卷积神经网络，它使用动态k-max池化运算符作为非线性的子采样函数。网络所引起的特征图能够捕获不同大小的单词关系。该网络在问题和情绪分类上获得了很高的性能，而不需要由解析器或其他资源提供外部特性。