

清华 大学

综合 论文 训 练

题目：反馈卷积神经网络在语义分类
的应用

系 别：计算机科学与技术系

专 业：计算机科学与技术

姓 名：刘 晨

指导教师：胡晓林 副研究员

2015 年 6 月 15 日

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：学校有权保留学位论文的复印件，允许该论文被查阅和借阅；学校可以公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存该论文。

(涉密的学位论文在解密后应遵守此规定)

签 名：_____ 导师签名：_____ 日 期：_____

中文摘要

近些年来，卷积神经网络和反馈神经网络被广泛应用于计算机视觉领域并且在标准数据集上取得了优异的实验结果。然而，这两个模型在自然语言处理领域的应用相对较少，而将这两个模型结合的应用则更少。

我们认为图像中的相邻像素和自然语言中相邻的两个单词的局部相关机制有相似之处，所以通过将图像分类中表现优异的深层乃至带有反馈连接的卷积神经网络应用到自然语言中，我们提出了用于语义分类的模型。

通过实验和观察，我们发现目前的词向量表示技术并不能很好的适应卷积运算，通过词向量表示的句子矩阵并不像图片那样在长宽两个维度上都存在局部相关。我们需要通过新的词向量表示方法或者对输入句子进行有效的变换来使输入句子矩阵满足两个维度局部相关才能将卷积运算更好地用于自然语言处理问题中。

关键词：卷积神经网络;反馈神经网络;语义分类

ABSTRACT

In recent years, convolutional neural network and recurrent neural network have been widely used in the field of computer vision and achieved remarkable performance on standard datasets. However, these two models have few natural language processing applications, let alone the combination of them.

We think the adjacent pixels in the images have some similar local correlation mechanism with words in natural sentences. By applying deep neural networks and networks with recurrent connections, which have remarkable performance in images classification, to natural language processing problems, we put forward some models for semantic classification.

According to experiments and observations, we find that the current word2vec algorithms cannot match convolution operation perfectly. Matrixes representing sentences by word2vec do not have the property of local correlation in both dimensions as image do. We need to develop new word2vec algorithms or transform input sentences to matrixes to satisfy local correlation in both dimensions. Through this way, we can make better use of convolution operation in natural language processing problems.

Keywords : Convolutional Neural Network; Recurrent Neural Network;
Sentiment Classification

目 录

第 1 章 引言	1
1.1 人工神经网络	2
1.1.1 卷积神经网络 (CNN)	2
1.1.2 反馈神经网络 (RNN)	3
1.2 基于上下文预测的词向量表示技术	5
1.3 本文的创新点	6
1.4 论文的主要工作和组织结构	6
第 2 章 相关工作	8
2.1 反馈卷积神经网络在图像分类中的应用	8
2.2 卷积神经网络语义分类模型	10
2.2.1 浅层卷积神经网络模型	10
2.2.2 深层卷积神经网络模型	11
2.3 卷积神经网络在其它自然语言处理问题的应用	12
第 3 章 模型构建	13
3.1 深层卷积神经网络	13
3.2 反馈卷积神经网络	14
3.3 对输入进行变换的卷积神经网络	15
3.3.1 相似矩阵输入	15
3.3.2 线性变换输入	16
第 4 章 实验	18
4.1 实验数据	18
4.1.1 标准数据集	18
4.1.2 词向量	18
4.2 评价指标	18
4.3 硬件和软件配置	19
4.4 代码实现和超参数设定	19

4.5 实验设计与结果	19
4.5.1 浅层卷积神经网络和反馈卷积神经网络	19
4.5.2 浅层神经网络和深层神经网络	22
4.5.3 对输入进行变换的模型	24
4.5.4 非监督语料的作用	24
4.5.5 错误分析实验	26
4.6 实验总结与分析	29
第 5 章 总结与展望	30
插图索引	31
表格索引	32
参考文献	33
致谢	35
声明	36
附录 A 外文文献资料阅读报告或文献翻译	37

第1章 引言

语义分类是计算机自动将输入的句子按照情感进行分类的技术[1]，其基本任务是输入一句话，输出其情感倾向。该问题通常可以抽象为是二分类（正面/负面）或者五分类（正面/偏正面/中性/偏负面/负面）问题。

语义分类是自然语言处理技术中的重要组成部分，其核心问题是找出一个合适的模型刻画句子，从句子中单词的排列中提取特征。迄今为止，研究者们已经提出了很多不同的句子模型。一类模型是基于词语和词语之间的组合[1][2][3][4][5][6]，该模型利用函数来将词语向量不断组合成短语向量，最终合成代表整个句子的向量作为特征。还有一类方法是通过语法抽象出一组逻辑规则来处理句子[7]，并以此作为情感分类的依据。

另外一大类句子模型是基于人工神经网络的，从最初的词袋模型到后来的n-Gram模型都能通过人工神经网络刻画。此外，我们还可以通过现有工具提取中句子的句法树利用递归神经网路（Recursive Neural Network，简称 RecNN）来学习句子的特征表示[8][9][10]。神经网络模型有很多优点，首先它是端到端的（end-to-end）的，除了句法树之外（即使是句法树也是通过已有工具自动生成的），我们无需再借助其它人工提取的特征。除此之外，神经网络能够学习到更好的词向量表示并且这种性质几乎是与语种无关的。

另一方面，近些年来卷积神经网络（Convolutional Neural Network，简称 CNN）在计算机视觉技术、特别是图像分类问题上取得了突破性的进展[11]。通过一层一层卷积层的叠加，卷积网络能够抽象出层次性的高质量图像特征。例如 VGG 实验室就曾经利用 19 层深度的卷积神经网络在 1000 分类的 ImageNet 数据集上测试得到 Top5-92.5% 的准确率[12]。在下面的介绍中，我们可以看出卷积神经网络也是一个基于局部相关提取特征的模型，这一点上和自然语言类似，我们可以考虑将具有高性能处理图像分类的模型移植到自然语言处理领域中。

现在已经有一些将或浅或深的卷积神经网络模型运用到自然语言处理特别是语义分类问题中的工作。与他们不同的是，在本论文中我们将提出一个把反馈神经网络（Recurrent Neural Network，简称 RNN）[13]和卷积神经网络相结合的模型，我们希望这种特殊的网络结构能够使神经网络语义分类模型的性能得到提升。

本章是全文的引言，将介绍神经网络语义分类模型所用到的最基本的概念和技术，包括本文中所涉及的人工神经网络和基于上下文预测的词向量表示技术。

1.1 人工神经网络

人工神经网络（Artificial Neural Network，简称 ANN）是上个世纪末人工智能领域兴起的研究分支，它通过构造神经元组成的网络对人脑进行仿真来构建具有学习能力的模型。人工神经网络的最小组成单位是神经元，其数学模型如下，它接收若干输入，按照权值相加后通过一个激活函数产生输出。人工神经网络就是通过若干神经元组成层，层与层之间通过连接（权值）组成的一个复杂的网络体系。

$$y = \sigma \left(\sum_{i=0}^k w_i * x_i \right)$$

人工神经网络一般采用误差回传的方式，通过网络输出和教师信号的差异逐步回传调整每一层神经元连接的权值，最终使模型输出和教师信号尽可能一致。

在训练的过程中，我们可能会对某一层神经元的输出以一定的概率 P 进行舍弃，舍弃掉得神经元输出视为 0 并且不参加到这次误差回传的更新中；在测试的过程中，为了保证该层神经元输出值得分布不变，我们会对该层每一个神经元输出的值乘以 (1-P) 后再传入下一个神经元层的输入，这就是神经网络中常用的 Dropout 技术。Dropout 以随机的方式打断了同层神经元之间的联系，使它们的训练能够尽可能独立，有利于他们学到与其它神经元不一样的特征，从而能有效缓解整个模型 Over Fitting 的现象。

下面介绍本工作中主要用到的两种人工神经网络模型：卷积神经网络（Convolutional Neural Network，简称 CNN）[11]和反馈神经网络（Recurrent Neural Network，简称 RNN）[13]。

1.1.1 卷积神经网络（CNN）

卷积神经网络是通过卷积运算连接神经元层和神经元层的一种神经网络，其主要结构如图 1.1 所示。

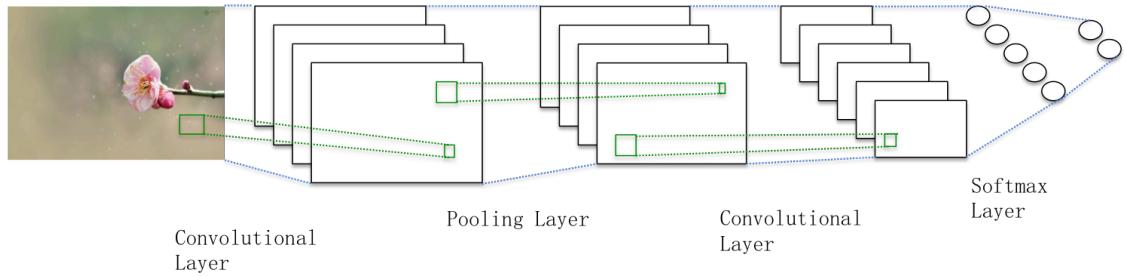


图 1.1 卷积神经网络

一般来说，卷积神经网络是由若干卷积层和若干 Pooling 层组成，每一个卷积层都有若干个特征映射空间（feature map）和若干个卷积核，每一个卷积核都会与该层的每一个特征空间进行卷积运算，将所得结果相加并作非线性后作为下一层某一个特征映射空间的输入，其数学表达式如下所示，其中 $Y_{k,j}$ 表示第 k 层的第 j 个特征映射的输出， $W_{i,j}$ 表示连接上一层第 i 个和下一层第 j 个特征映射的卷积核， f_k 表示第 k 层的特征映射空间数。

$$Y_{k,j} = \sigma \left(\sum_{i=1}^{f_{k-1}} W_{i,j} \otimes Y_{k-1,i} + B_i \right)$$

Pooling 层是讲上一层的局部区域的神经元映射到下层的一个神经元，起到降低模型复杂度、过滤噪声的作用。最常见的是 Max-Pooling 层，其功能是取上层的一个局部区域（通常是一个小方形区域）神经元输出的最大值作为下层的输入。

卷积运算的性质使卷积神经网络的底层神经元能够很好捕捉输入信号的局部特性，而通过一层一层的卷积层的级联，上层的神经元捕捉的底层信号区域越来越大，因此能很好反映一些抽象的、全局特性。也就是说，卷积神经网络的底层到高层是一个特征逐渐抽象的过程，底层更加偏向统计特征，而高层则能反映一些语义特征。

在实际应用中，卷积神经网络可以用来提取特征用作后续处理，也可以将高层的特征输出接到一个分类器中来进行分类。

1.1.2 反馈神经网络（RNN）

反馈神经网络是一种被设计用来处理序列数据的神经网络，其主要模型如图 1.2 所示。

与一般的神经网络模型相比，反馈神经网络的最大特征是添加了自连接（recurrent connection），该模型每一个时刻的状态均受到上一个时刻状态的影响，这使该模型具有一定记忆性。

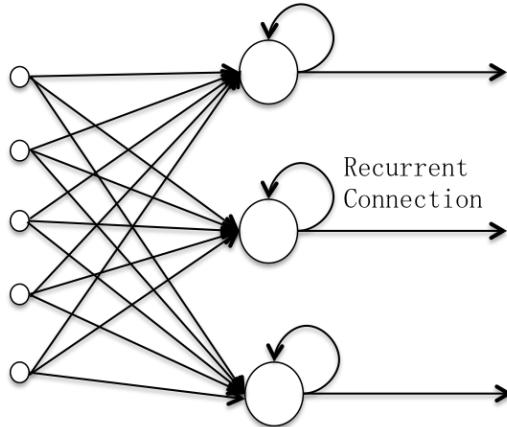


图 1.2 反馈神经网络

反馈神经网络的数学模型如下，其中 $Y_{t,k}$ 表示时刻 t 第 k 层神经元的输出， $W_{forward}$ 和 $W_{recurrent}$ 分别表示前馈连接和自连接的权重。

$$Y_{t,k} = \sigma(W_{forward} * Y_{t,k-1} + W_{recurrent} * Y_{t-1,k})$$

我们将一个带有自连接的神经元沿着时间展开（如图 1.3 所示）可以发现，反馈神经网络可以认为是一个深层的神经网络，它的深度体现在时间维度上。与空间维度不同的是，每一个时刻都有一个教师信号传入网络中更新其权值。

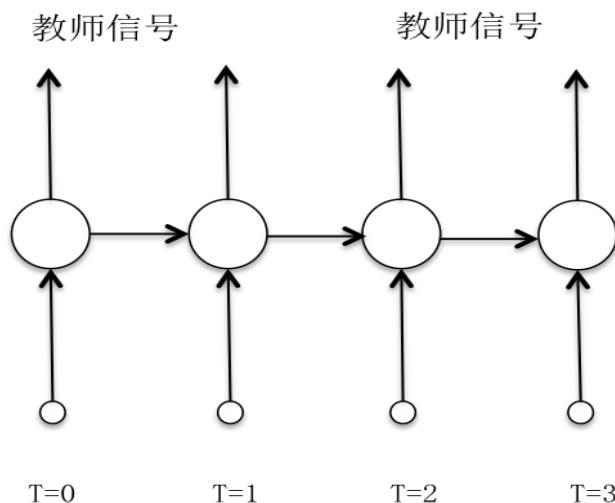


图 1.3 时间展开后的反馈神经网络

由于反馈神经网络具有较好的记忆性，所以在实际应用中，它经常用来处理诸如语音等不定长的序列数据。

1.2 基于上下文预测的词向量表示技术

词向量表示（word embedding）是把神经网络用语自然语言处理中的一种常见技术。因为将文字输入到神经网络之前必须将文字进行编码，需要一个向量来表示它，而最简单的 one-hot 编码方式既需要等同于单词个数的巨大存储空间，又忽略了单词与单词之间的语义关系。为了克服这两个问题，我们需要通过语料来学习一个“更好”的词向量表示，这种表示的维度比较低并且能够很好反映词语之间的语义关系。

本文工作中所用的词向量来自 Google 所训练好的词向量集合 GoogleNews-vectors-negative300.bin 文件，它们所用的词向量训练算法主要有 Skip-Gram 和 Continuous Bag-of-Word (CBOW) 两种[14][15]。这两种方法都是通过训练能够预测上下文的神经网络来得到词向量表示的。

如图 1.4 所示，Skip-Gram 是通过一个可见的词来最大化他周围词出现概率的模型，而 CBOW 是通过一个词周围的词来最大化该词出现概率的模型。也就是说 Skip-Gram 模型的优化目标是 $\frac{1}{T} \sum_{t=1}^T \sum_{-c \leq i \leq c, i \neq 0} \log(W_{t+i} | W_t)$ 而 CBOW 模型的优化目标是 $\frac{1}{T} \sum_{t=1}^T \log P(W_t | W_{t-c}, W_{t-c+1} \dots W_{t-1}, W_{t+1} \dots W_{t+c-1}, W_{t+c})$ ，其中 T 代表语料的单词数， W_i 代表语料的第 i 个单词。

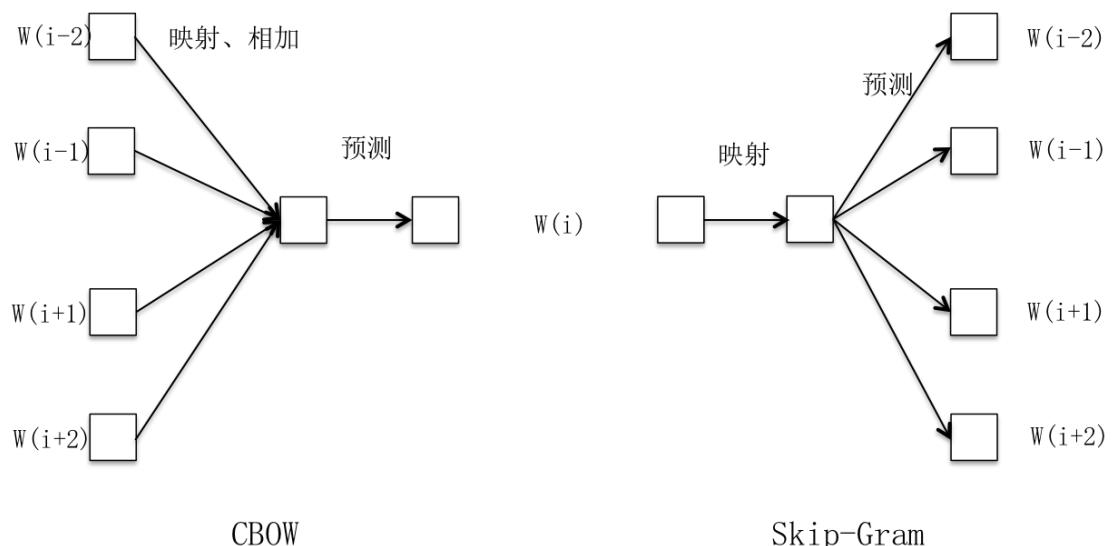


图 1.4 Skip-Gram 和 CBOW

在这两个模型中，每一个单词 W_i 都有一个输入向量 V_{W_i} 和一个输出向量 V'_{W_i} ，一个单词在另一个单词周边出现的概率定义为 $P(W_o|W_I) = \frac{\exp(V'^T_{W_o} V_{W_I})}{\sum_w \exp(V'^T_{W_o} V_{W_I})}$ ，对于CBOW 中 $P(W_t|W_{t-c}, W_{t-c+1} \dots W_{t-1}, W_{t+1} \dots W_{t+c-1}, W_{t+c})$ 一项，我们直接将可见的词向量进行相加作为上式的 V_{W_I} 进行概率的计算。

通过上面的定义我们就能够将 Skip-Gram 和 CBOW 模型的优化目标写成词向量的函数了，不论是 Skip-Gram 还是 CBOW，我们都可以通过一个浅层的全连接网络进行词向量训练。

通过网络的训练，我们可以得到维度相对较低且携带有语义信息的词向量了，这样的词向量具有很好的性质。首先，语义相近的词的词向量欧氏距离相对比较近、夹角比较小。此外，对于具有类比关系的两个词对 A~B=C~D，它们的词向量也近似满足 $\text{Vec}(A)-\text{Vec}(B)=\text{Vec}(C)-\text{Vec}(D)$ ，其中 $\text{Vec}(w)$ 表示 w 的词向量，例如 $\text{Vec}(\text{'巴黎'})-\text{Vec}(\text{'法国'})=\text{Vec}(\text{'北京'})-\text{Vec}(\text{'中国'})$ 。使用语料比较多、质量比较好的词向量还具有概念叠加的性质，例如 $\text{Vec}(\text{'北京'})=\text{Vec}(\text{'中国'})+\text{Vec}(\text{'首都'})$ 。

1.3 本文的创新点

本文的创新点体现在以下两个个方面：一方面，我们改进了基于卷积的人工神经网络语义分类模型，加入反馈连接，提出了反馈卷积神经网络模型；另一方面，我们从其它自然语言处理问题中得到启发，将用词向量表示的句子矩阵进行一定变换后输入神经网络，而之前的语义分类模型往往是直接将句子矩阵作为神经网络的输入。

1.4 论文的主要工作和组织结构

本论文的主要工作是将人工神经网络中的卷积和反馈神经网络用到自然语言处理中的语义分类问题中，上面我们介绍了开展本论文工作所需的背景知识。我们将通过词向量表示的句子作为神经网络的输入，通过一层或多层卷积层或者带有自连接的卷积层提取特征，最后利用这些特征对句子进行分类。

论文一共分为五个章节，从第二章开始的各个章节的内容组织结构如下：

第二章重点介绍部分与本论文工作密切相关的前人工作。本章内容包括三节，第一节介绍反馈卷积神经网络在图像分类中的应用，由于我们认为图像和自然语言存在某种联系，所以反馈卷积神经网络在图像上所表现的性质值得我们指出；

第二节介绍卷积神经网络在语义分类中的应用，包括一个浅层模型和一个深层模型，本文提出的模型大多是在这个模型的基础上进行修改加强的；第三节简要介绍卷积神经网络模型在其它自然语言处理问题中的应用，这些模型对输入句子的处理方法给了我们启发。

第三章提出我们的模型，一共三个模型，分为三节进行阐述。第一节是深层卷积神经网络模型，该模型可以认为是第二章第二节模型的深层版本；第二节是浅层卷积反馈神经网络模型，该模型主要在第二章第二节模型的基础上加上自连接的反馈结构；第三节的模型对输入句子进行了变换，该模型主要受第二章第三节的模型启发，包括两种不同的对句子的处理方法（线性与非线性）。

第四章是实验部分，包括介绍我们使用的标准数据集的相关信息、所用机器的硬件和软件配置，还包括我们设计的实验对各个模型的比较以及对这些实验结果的分析。

第五章是对本论文工作的总结和展望，提出本论文中各个模型所存在的问题以及可能存在的改进手段。

第2章 相关工作

本章介绍前人所做的与本文密切相关的工作，这些模型的优势都是我们借鉴的依据，第四章的实验部分将会对其中部分模型与我们提出的模型进行对比。

2.1 反馈卷积神经网络在图像分类中的应用

近些年来，深层的卷积神经网络在图像分类等问题中取得了突破性的进展。通过卷积层的级联，我们能够在较低层得到图像基本的、主要具有统计意义的特征，而在较高层得到由底层特征通过非线性组合而成的抽象的、更具有语义意义的特征。我们把高层的特征接入到一个分类器中，利用高层的语义特征进行图像分类，这样往往能够得到较好的效果。

为了能够提取出更加抽象、更加富有语义色彩的特征，人们往往会构建更加深层的网络，例如 Google 在 2014 年推出用于 ImageNet 上分类的网络 GoogleNet 就有 22 层之深（不包括 Pooling 层）。更深的网络固然能够提升模型刻画非线性特征的能力，但是更深的网络意味着更多的待学习参数、更慢的收敛时间和更困难的超参选取，这些很多时候都是我们不能忍受的。

反馈卷积神经网络就能解决上面的一些问题，其模型如图 2.1 所示[16]。与一般的反馈神经网络相同，反馈卷积神经网络有一个自连接，但是该自连接和前馈连接一样都是卷积运算。值得一提的是，不同于反馈神经网络处理序列输入，这里所指的反馈卷积神经网络始终接受恒定的输入激励，不同的是另一个分量来自上一个时刻隐层的状态与自连接的卷积，这样的计算进行 T 次后（T 预先设定好）当前层的值才继续向后传。数学表达式如下所示：

$$X_{t,l} = \sigma(W_i \otimes Y_{l-1} + W_r \otimes X_{t-1,l}), Y_l = f(X_{T,l})$$

其中 W_i, W_r 分别表示前馈的卷积核和自连接的卷积核， $X_{t,l}$ 表示第 l 层在 t 这个时刻的状态，而 Y_l 表示第 l 层传向下一层的值。从上面的式子可以看出每次进行自连接迭代的时候，前馈的输入都是恒定，在进行 T 次循环迭代之后当前层神经元的值作为下一层的输入。

为了防止多次迭代出现的数值爆炸的现象，在每一次迭代之后，我们都使用了局部归一化措施 LRN（Local Response Normalization），以控制每一次迭代后神经元的数值。LRN 的计算公式如下，其中 $X_{i,j,k}$ 表示第 k 个特征映射空间的坐标为 (i,j) 的位置，N 为该层特征空间的个数， α, β 是人为设定的超参数。

$$LRN(X_{i,j,k}) = \frac{X_{i,j,k}}{\left(1 + \frac{\alpha}{N} \sum_{k'=\max(0, k-\frac{N}{2})}^{\min(N, k+\frac{N}{2})} X_{i,j,k'}^2\right)^\beta}$$

跟普通的反馈神经网络类似，反馈卷积神经网络也是可以按照时间进行展开的。图 2.1 左上部分展示了 $T=3$ 的情况下反馈卷积神经网络展开情况，其中所有虚线和实线部分所用的卷积核是分别相等的，这种思想叫做权值共享。通过图 2.1 左上部分，我们可以看出 $T=3$ 情况下的反馈卷积神经网络本质就是一个深度为 4 的卷积神经网络！

通过权值共享，我们能够在不增加参数规模、不显著增加收敛时间的情况下构建一个等效深层网络的模型。反馈卷积神经网络本质是一个更深的、带有短路的、权值共享的卷积神经网络。

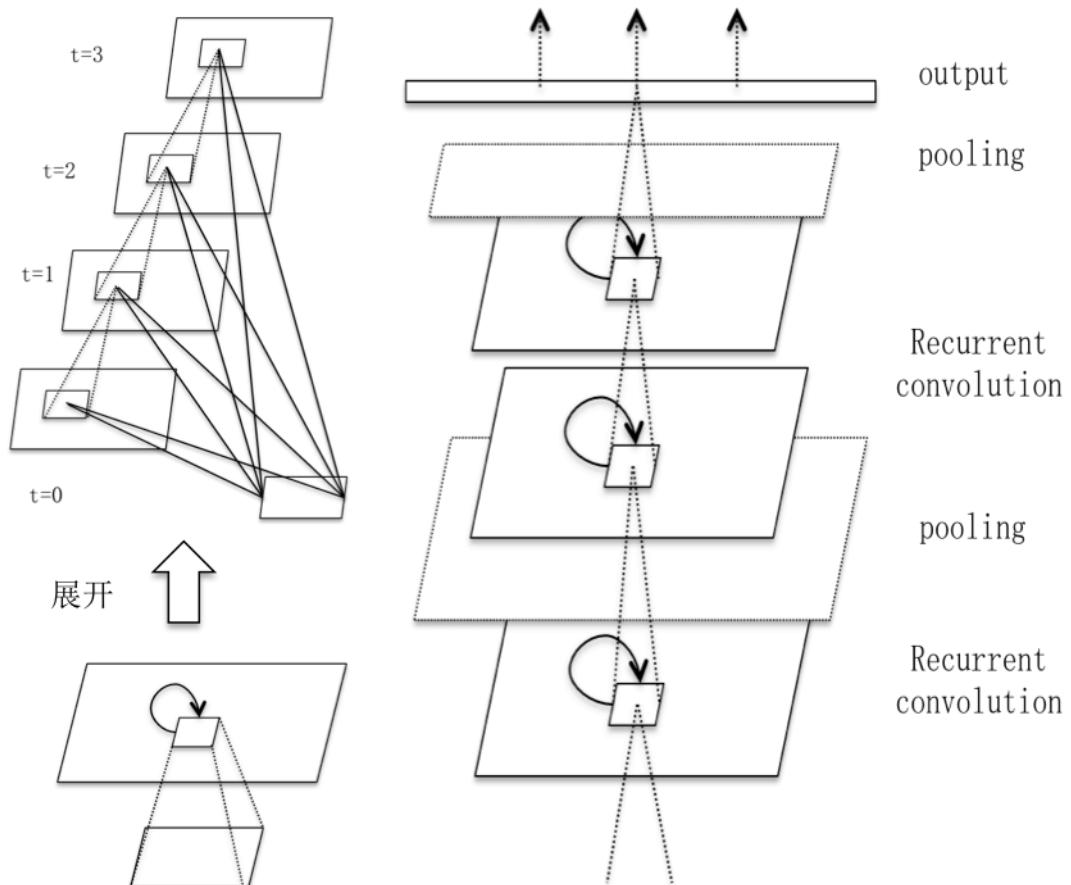


图 2.1 反馈卷积神经网络

我们只需要将几个这样的反馈卷积层连接起来（如 2.1 右半部分）就能得到一个等效十几甚至更深的卷积神经网络。正因为这种优势，反馈卷积在图片分类任务中取得了很好的效果。在 cifar-10 数据集上的准确率比同样参数规模的普通卷积神经网络高 2% 左右。

2.2 卷积神经网络语义分类模型

对于语义分类问题，我们可以利用词向量才能将输入的一句话转化为一个矩阵作为神经网络的输入。这个矩阵的一边长是句子的长度，另一边长是词向量的维度。在利用词向量对语言进行编码之后，我们便可以像处理图像那样将一句话输入到神经网络中去。

2.2.1 浅层卷积神经网络模型

浅层卷积神经网络的模型如图 2.2 所示[17]，它由一个多种卷积核组成的卷积层、一个全连接层和 softmax 分类器构成。

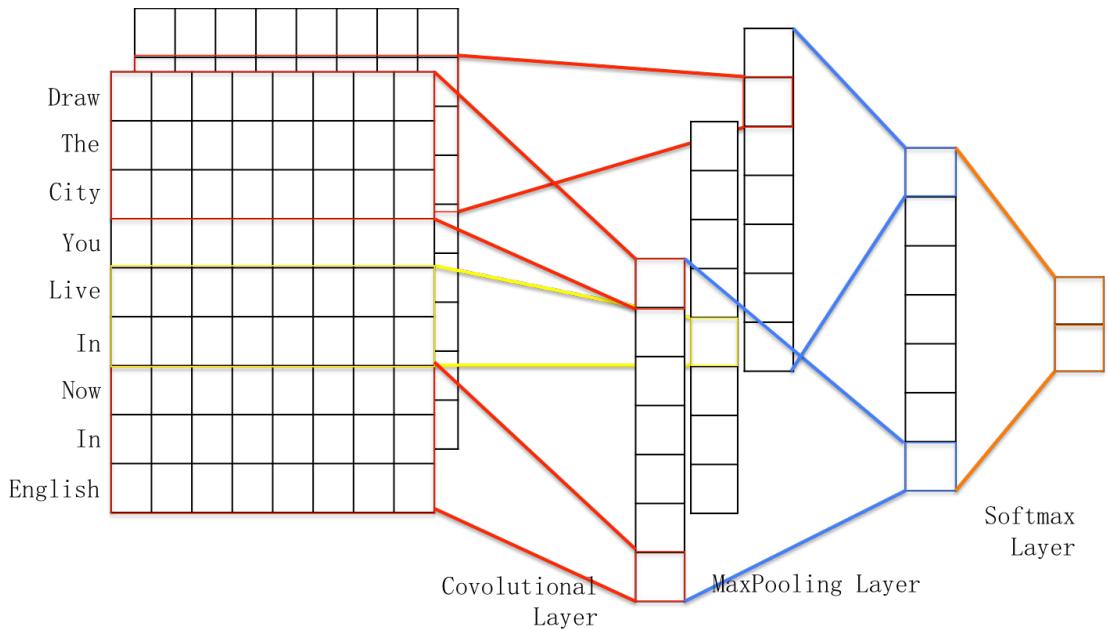


图 2.2 用于语义分类的浅层卷积神经网络

该模型的卷积层包括不同形状的卷积核，但是所有形状的卷积核的其中一维长度都是词向量的维度。在这种情况下，每一个句子经过不同的卷积核卷积之后都会变成若干个句子特征向量，这些特征向量经过中间的 Global Max Pooling 层

之后变成一个向量（该向量中的每一个值都是由原来对应的句子特征向量映射过来），该向量通过最后的一个 softmax 分类器进行分类。

对于单词词向量的不同初始化和不同训练方法，该模型一共有三种模式：第一种是单词词向量随机初始化，我们在神经网络训练的过程中，动态调整词向量；第二种是借助大规模语料（单词数在 billion 级，如 wikipedia）训练出的词向量作为单词词向量的初始化，在训练过程中词向量保持固定；第三种词向量的初始方式与第二种相同，不过词向量在训练中动态调整。

2.2.2 深层卷积神经网络模型

深层卷积神经网络的模型如图 2.3 所示[18]，它是由若干个卷积层、Pooling 层和最后的分类器构成。不同于上面的浅层模型，该模型卷积层的卷积核可以是任意方形形状（图中用红色表示），Pooling 层也从原来的 Global Pooling 改成了 Dynamic-K Pooling，即 Pooling 映射保留原来的 K 个最大值，这种 Pooling 方式用在句子中单词排列的那个方向上（图中用绿色表示），在词向量维度上，我们还采用一般的一维 Pooling 方式（图中用黄色表示）。对于不同的长度的句子，K 的大小也是不同的，因为卷积运算提取的是局部特征，这与输入数据的规模是没有关系的，我们只要保证最后分类器的输入数据规模恒定即可。最简单的取法是从底层到顶层，K 的取值线性减小，数学表达式为 $K_i = Input - (Input - Final) * \frac{i}{Layers}$ ，其中 $K_i, Input, Final, Layers$ 分别表示第 i 个 Pooling 层的 K 值、输入句子的单词数、分类器输入数据的规模和 Pooling 层的个数。

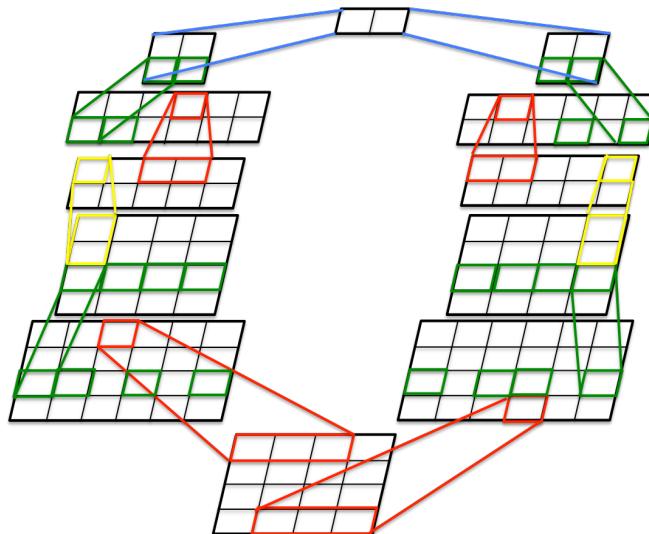


图 2.3 用于语义分类的深层卷积神经网络

与浅层神经网络相比，深度神经网络的卷积核和 Pooling 使用相对灵活，该模型的词向量初始化采用随机初始化的方法。

2.3 卷积神经网络在其它自然语言处理问题的应用

除了语义分析外，卷积神经网络在其它自然语言处理问题中也有应用，下面的模型是用于比较两个句子相似度的，该模型对句子的处理方式对我们的问题有很好的启发[19]。

如图 2.4 所示，该模型是处理句子相似度问题，输入是两个句子，输出是一个反映两个句子相似程度的标量。该模型的特别之处在于处理输入上，设输入句子的长度为 L，词向量的维度为 D，第一层的卷积核宽度为 F，则第一层将两个句子映射到一个 $(L-F) \times (L-F)$ 的矩阵上。具体计算方法如下：

$$Y_{i,j} = \sigma(W_1 \otimes S_{1(i,i+F)} + W_2 \otimes S_{2(j,j+F)})$$

其中 $Y_{i,j}$ 表示矩阵中位置为 (i,j) 的值， $S_{1(i,i+F)}$, $S_{2(j,j+F)}$ 分别表示第 1 句和第 2 句话中第 i 个到 $i+F$ 个词所对应的词向量构成的 $D \times F$ 的矩阵， W_1, W_2 分别表示卷积核中用于第 1 句和第 2 句的部分。

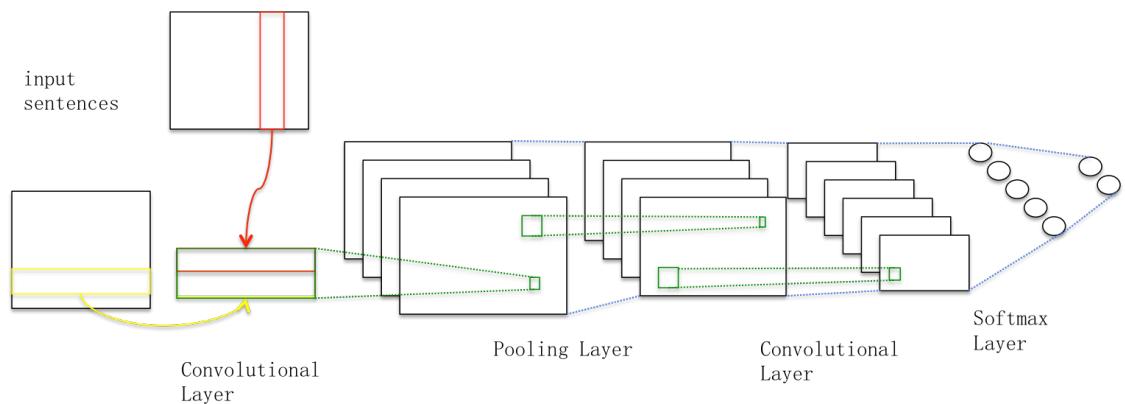


图 2.4 用于比较句子相似度的卷积神经网络

从第一层的计算方法我们可以看出，第一层输出的矩阵中的每一个元素都是分别从两句话中取出一个长度为 F 的片段之后的非线性组合。该模型在第一层就将两个句子融合提取相似特征，有利于比较句子相似度的任务。除此之外，第一层输出矩阵上相邻元素之间有比较大的相关关系（源于同一个句子上相邻词的相关关系），这给该模型后面进行卷积提取特征操作提供依据。

第3章 模型构建

本章介绍我们在第二章前人模型的基础上进行修改的模型，包括建立在 2.2.1 节模型基础上的深层卷积神经网络和反馈卷积神经网络，受 2.3 节模型的启发，我们改变了模型的输入，提出对输入进行变换的卷积神经网络模型。

3.1 深层卷积神经网络

该模型是 2.2.1 节中模型深层版本，与浅层模型不同的是，在同一层上只有一种大小的卷积核，我们希望通过多层卷积层的级联得到较单层神经网络更加抽象的特征，从而提升我们模型的性能。

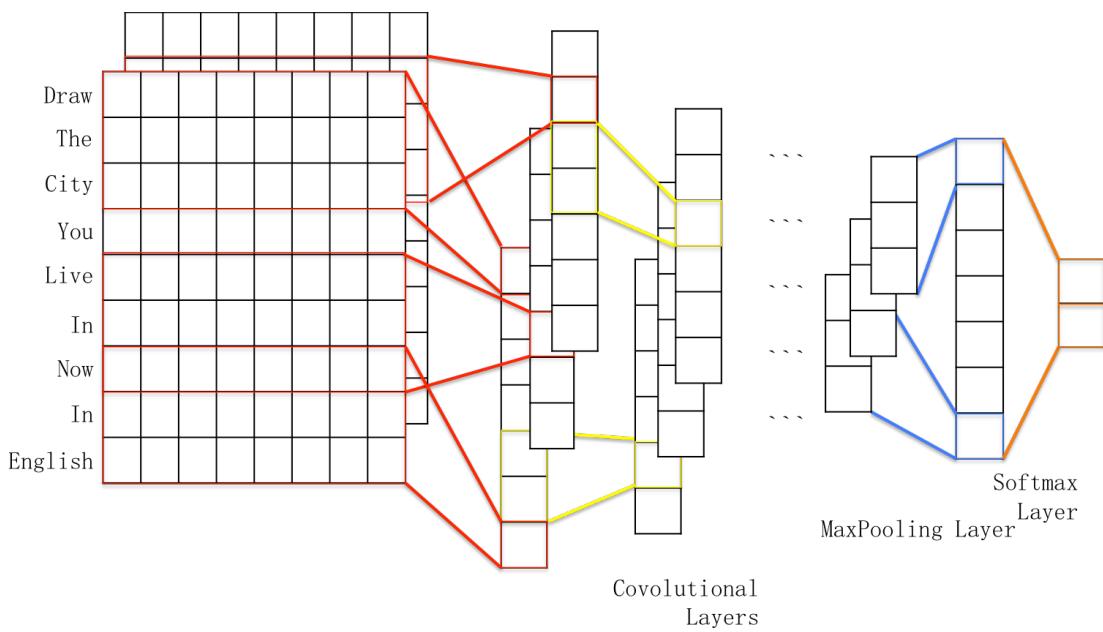


图 3.1 深层卷积神经网络

从引言中关于词向量的训练方法可以看出，作为一个单词语义表示的词向量，其各个维度之间是相互独立的。也就是说，对于已经训练好的一组词向量，如果我们将每个单词对应的向量的前两维进行对调，则新的一组向量作为词向量和原来的词向量在语义上是等效的。词向量的这个性质决定了在词向量的维度上没有局部相关的性质，也就是说词向量第一维和第二维的关系并不一定比第一维和第一百维的关系更加接近。词向量是一个整体，局部没有明显语义特征。从这个角

度上讲，词向量化的句子不能完全等同于一幅二维图像（图像在两个维度上都有局部相关性质），我们需要将词向量那个维度所有信息全部纳入到卷积核计算范围内才行。

基于以上考虑，正如图 3.1 所示，该模型的第一层卷积核在词向量方向的长度与词向量的维度相等，从第二层开始，每一句话对应一个向量，二维卷积层退化为一维向量之间的卷积。与浅层模型相同的是，在卷积计算完成之后，我们会通过一个 Global Max-Pooling 层将各个特征映射输入到最终的 Softmax 分类器中。

3.2 反馈卷积神经网络

该模型与 2.2.1 节模型的唯一区别在于我们在句子进行卷积之后的特征向量的那层加入自连接，形成了一个反馈卷积神经网络。

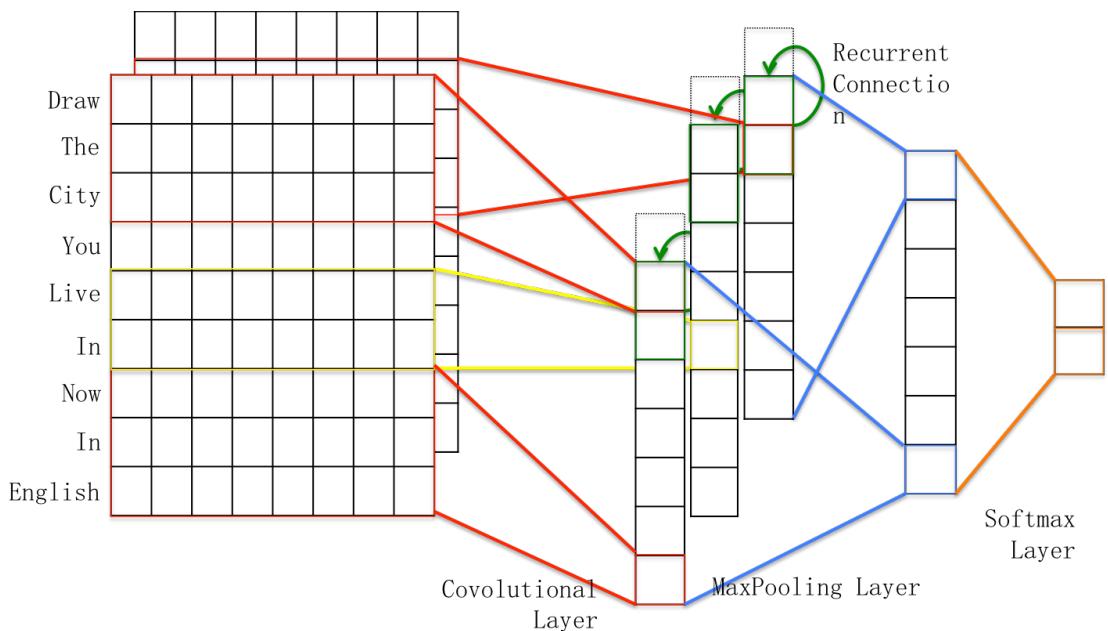


图 3.2 反馈卷积神经网络

正如 2.1 节在分析反馈卷积神经网络的时候沿着时间展开一样，该反馈卷积神经网络展开后的本质是一个共享权值的、带有短路连接的深层卷积神经网络。我们希望该网络能够提取比浅层卷积神经网络更加抽象的特征。

根据 3.1 节关于词向量特性的分析，该模型的第二层仍然是一维向量，自连接中的卷积计算采用‘valid’卷积模式，我们在向量的头和尾填上若干零，保证卷积之后的向量大小与原来向量大小一致，能够与前馈输入直接相加。

3.3 对输入进行变换的卷积神经网络

词向量的性质决定了我们在 3.1 和 3.2 中模型的第一层必须使用长度等同于词向量维度的卷积核，这样的后果是整个神经网络系统的第一层和第二层之间的神经元个数差异太大，第二层的一维神经元排列不利于发挥卷积运算的优势，不利于构建更深的网络。

本节的两个模型是受到 2.3 节模型启发的，它们的共同点是不再简单地将句子的词向量表示作为网络的输入，而是利用词向量把句子进行变换之后，创造一个在两个维度方向都有局部相关特征的输入。

3.3.1 相似矩阵输入

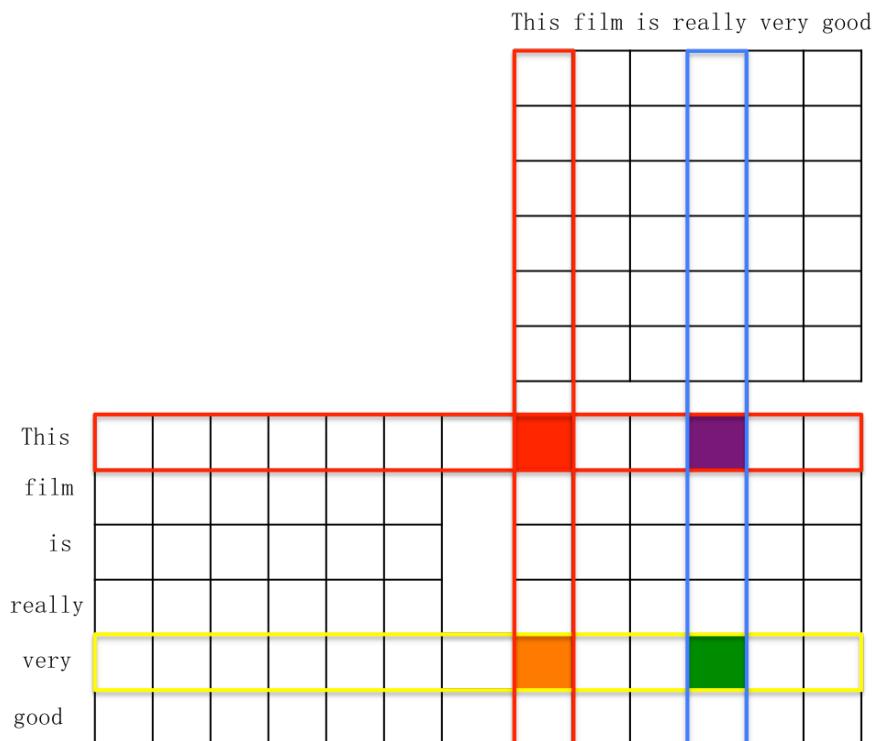


图 3.3 相似矩阵输入

在引言中我们提到，词向量包含了一个词的语义信息，一般说来，两个语义越接近的词，它们词向量之间的距离或者夹角就越小，反之亦然。由词向量的这条性质我们就可以计算词向量相似矩阵了。

如图 3.3 所示，向量相似矩阵 M 是一句话中各个词两两相似度所构成的矩阵。其计算方式如下，其中 W_i 表示第 i 个单词的词向量。

$$M_{i,j} = \frac{W_i^T \cdot W_j}{\|W_i\| \|W_j\|}$$

从上面的计算公式可以看出， M 是一个对称矩阵并且 M 的任何一个方形矩阵都能刻画连续的几个单词的相关关系，具备局部相关的特征。在将输入句子进行这样的处理之后，我们就可以用诸如 $2*2$ 的小型方块卷积核来提取特征了，这样就能基本解决本章前两节模型所遇到的问题。

3.3.2 线性变换输入

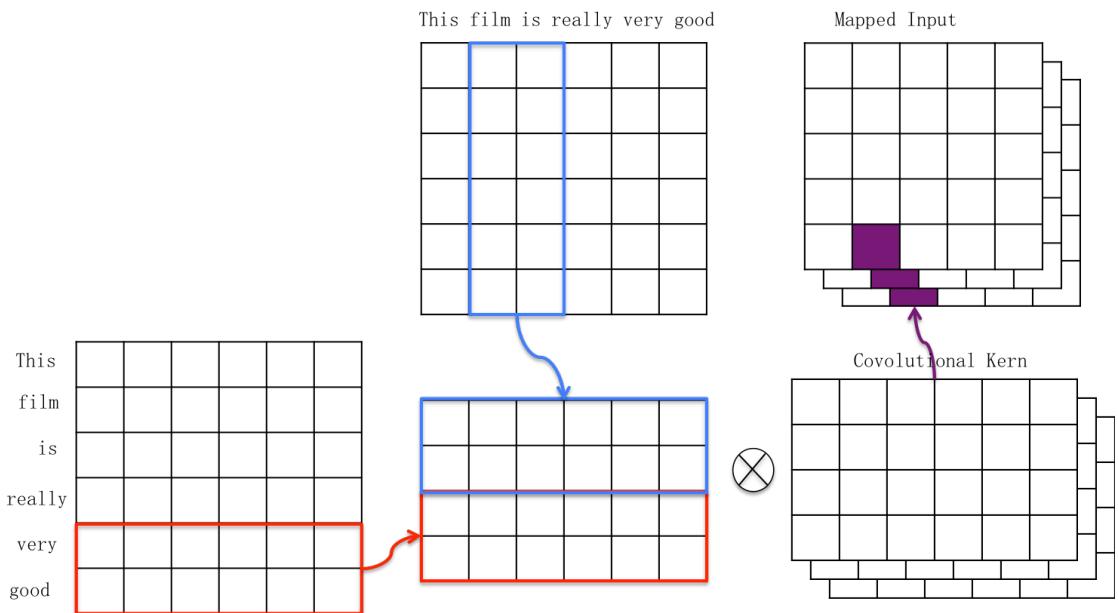


图 3.4 线性变换输入

该模型与 2.3 中的模型很相似，其第一层如图 3.4 所示，但是我们是需要解决一个语义分类而不是句子匹配问题，所以我们将 2.3 中两个维度输入的两个句子改成了同一个句子。这样，第一层的计算表达式变为：

$$Y_{i,j} = \sigma(W_1 \otimes S_{i,i+F} + W_2 \otimes S_{j,j+F})$$

从上面的式子可以看出，第一层输出矩阵中的每一个元素都由两部分构成：一部分只与行号 i 有关，另一部分只与列号 j 有关。在同一个特征映射空间输出的矩阵的每一行或者每一列的元素，它们都会有一个分量是相同的。这也是符合局部相关的特征，我们可以在这层的基础上叠上一般卷积层提取特征。

第4章 实验

4.1 实验数据

4.1.1 标准数据集

本论文所做的实验中一共用到两个数据集：Movie Review（简称 MR，下同）[20]和 Stanford Sentiment Treebank（简称 SST，下同）[21]，这两个数据集中的句子大多是人们对一些电影的影评，语言为英语。

MR 数据集是一个二分类数据集，每一句话都被标记为“正面”或者“负面”。MR 中一共包括 10662 个句子，其中正负例各一半。在对模型进行评价的时候，我们在该数据集上采用 10 次 10 折的交叉检测（cross validation）的方法。

SST 数据集是一个五分类数据集，每一句话和句子中每个短语被标记为“正面”、“偏正面”、“中性”、“偏负面”和“负面”5 种标签。SST 中一共包括 11855 个句子和 239231 个短语。在本文涉及到的实验中，我们均不利用 SST 中得短语信息进行训练或测试。SST 中的 11855 个句子大致按照 7:1:2 的比例划分训练集、验证集和测试集。

MR 和 SST 数据集中存在少量无法用 ASCII 码表示的非英语单词，我们对这些单词进行了过滤处理。

4.1.2 词向量

本论文所采用的词向量来源自 Google 所公布的 300 维词向量 GoogleNews-vectors-negative300.bin（下面简称 GoogleNews）。该文件包括 300 万个词或者短语的向量表示，这些词向量和短语向量是 Google 运用 1.2.1 节的方法从 GoogleNews 上大约 1000 亿个单词的语料集中训练出来的。

在比较非监督语料的规模对语义分类模型的效果实验中，我们还使用了一个从较少的语料中训练出来的词向量。该词向量集合的训练方法与 GoogleNews 相同，不过它是我们从网络上抓取的 10 个句子中训练出来的词向量。

4.2 评价指标

本文实验采用朴素的测试准确率来衡量模型的性能，对于 10 次 10 折的交叉检测，我们取 10 次准确率的平均值作为最终准确率。

4.3 硬件和软件配置

服务器拥有 32G 公用内存、24 个主频为 2GHZ 的 CPU 和显存 6G 的 NVIDIA 显卡 TITAN BLACK。

服务器的操作系统为 Linux 主流发行版本 Ubuntu。

4.4 代码实现和超参数设定

所有程序均采用 python 用于开发神经网络的工具包 Theano 编写，Theano 版本为 0.7.0。浮点数的精度为 32 位，以便在 GPU 上能够高速并行运转。

我们分别实现了神经网络常见的神经元层：全连接层、卷积层、反馈卷积层（包括 LRN）和 Softmax 层，再利用这些层之间的串联或者并联来构建下面实验中所涉及到得所有模型。我们实现的更新方法有两种：一阶的随机梯度下降（SGD）和二阶的 Adadelta 方法[22]。在没有验证集的情况下，我们将最终训练好的模型进行一次测试，测试结果作为最终结果；在有验证集的情况下，我们每使用一定数量的训练数据就会在验证集和测试集上进行测试，我们取在验证集上性能最好的模型在测试集上的结果作为最终结果。训练的 Epoch 数定为 10。

为了调试的需要，我们会讲每次运行完的核心代码、最终训练好的模型和中间训练测试集的准确率保存下来。方便我们直观地查看结果、反复测试同意模型和保存模型的细节实现。我们还使用 HighChartJS 工具包进行结果绘图并且开发简单地实验结果和错误分析命令行工具。

4.5 实验设计与结果

4.5.1 浅层卷积神经网络和反馈卷积神经网络

在这两个模型中，我们都使用 GoogleNews 的词向量作为初始化并且在训练的过程中词向量保持恒定，两个模型第一层的卷积核的大小都是 $3*300$ 、 $4*300$ 和 $5*300$ 三种，为了充分发挥自连接的作用，所有自连接卷积核的大小为 $5*1$ 。为了保证两个模型的参数规模保持一致，普通的卷积神经网络每一个卷积核的特征空间数为 100，而反馈卷积神经网络的卷积核的特征空间数为 80。两个模型的最后都是 Softmax 层，区别仅仅在于输出节点的不同。这样两个模型的参数总数相当，都为约 35 万上下。（如下表）

表格 4.1 浅层卷积和反馈卷积神经网络参数规模对比

问题	参数规模
卷积神经网络二分类	$(3+4+5)*300*100+100*3*2=360600$
卷积神经网络五分类	$(3+4+5)*300*100+100*3*5=361500$
反馈卷积神经网络二分类	$(3+4+5)*300*80+3*3*1*80*80+80*3*2=346080$
反馈卷积神经网络五分类	$(3+4+5)*300*80+3*3*1*80*80+80*3*5=346800$

对于这两个模型，我们都可以使用 Dropout 技术来防止模型出现过拟合的现象。对于反馈卷积神经网络，我们可以改变迭代次数来测试不同的参数配置。在训练网络的过程中，我们都采用 Adadelta 的方法来进行参数更新，较传统的 SGD 优化方法，作为二阶优化方法的 Adadelta 具有收敛速度快、稳定且不需要手动调整学习率的优点。

表格 4.2 不同配置的卷积和反馈卷积神经网络在 MR 数据集上的表现

名称	类型	Dropout	循环次数	准确率
BaselineCNN	CNN	/	/	80.80 %
DropoutCNN	CNN	0.5	/	80.46%
RCNN-1	RCNN	/	1	75.07%
RCNN-2	RCNN	/	2	77.02%
RCNN-3	RCNN	/	3	79.91%
RCNN-5	RCNN	/	5	79.89%
DropoutRCNN-1	RCNN	0.5	1	77.83%
DropoutRCNN-2	RCNN	0.5	2	77.01%
DropoutRCNN-3	RCNN	0.5	3	79.36%
DropoutRCNN-5	RCNN	0.5	5	79.61%

表格 4.3 不同配置的卷积和反馈卷积神经网络在 SST 数据集上的表现

名称	类型	Dropout	循环次数	准确率
BaselineCNN	CNN	/	/	46.64%
DropoutCNN	CNN	0.5	/	47.37%
RCNN-1	RCNN	/	1	44.52%
RCNN-2	RCNN	/	2	45.61%
RCNN-3	RCNN	/	3	45.88%
RCNN-5	RCNN	/	5	47.87%
DropoutRCNN-1	RCNN	0.5	1	46.29%
DropoutRCNN-2	RCNN	0.5	2	46.87%
DropoutRCNN-3	RCNN	0.5	3	46.42%
DropoutRCNN-5	RCNN	0.5	5	47.46%

表 4.2 是不同模型在 MR 数据集上的表现，表 4.3 是它们在 SST 数据集上得表现。其中浅层卷积神经网络记为 BaselineCNN，循环次数为 T 的反馈卷积神经网络记为 RCNN-T，从上表可以看出，在参数一定的情况下，反馈卷积神经网络和普通的卷积神经网络表现不相上下，对于反馈卷积神经网络，循环的次数越多，分类效果往往更好些。

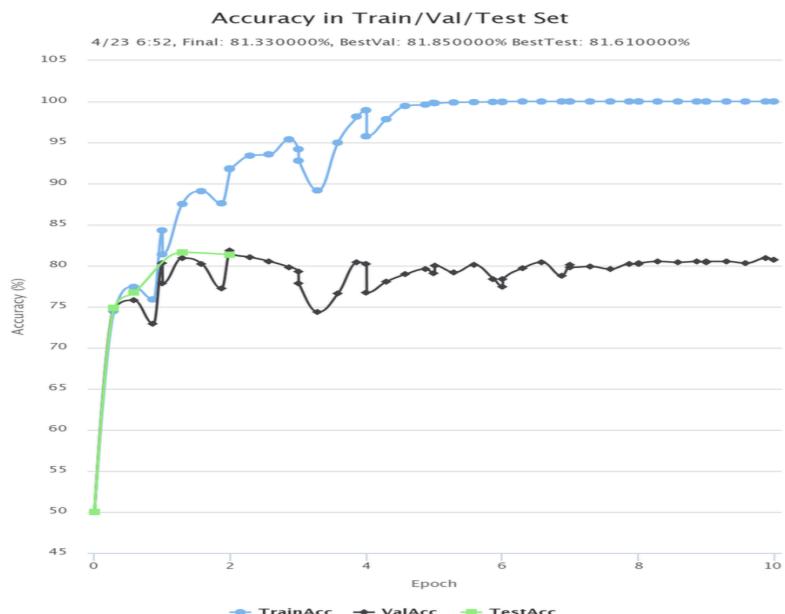


图 4.1 卷积神经网络在 MR 数据集上的训练情况

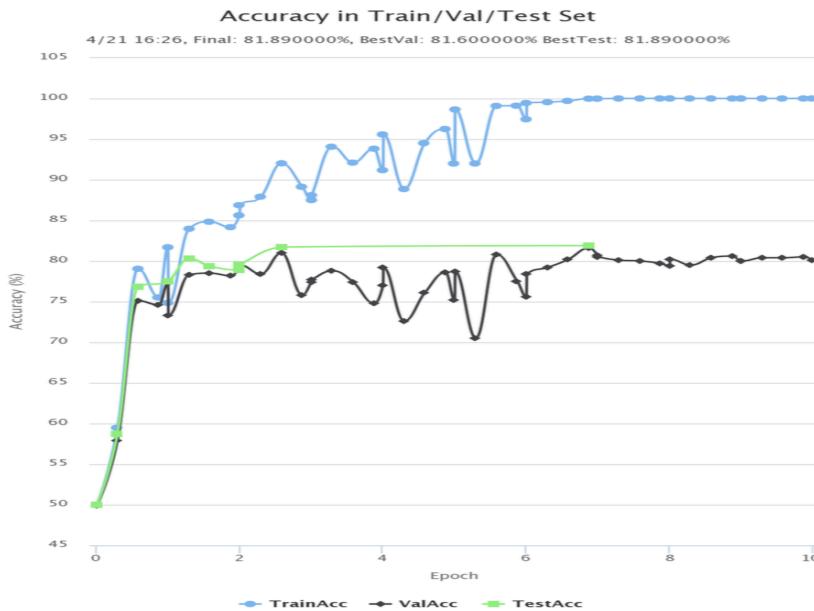


图 4.2 循环次数为 1 的反馈卷积神经网络在 MR 数据集上的训练情况

收敛性方面, 图 4.1 和图 4.2 展示了普通神经网络和循环次数为 1 的反馈卷积神经网络的训练进程, 蓝色、黑色和绿色折线分别表示训练集、验证集和测试集的准确率随训练进程的变化。从这两幅图我们可以清楚看出, 与普通的卷积神经网络相比, 反馈卷积神经网络的收敛速度更慢而在训练的过程中也表现地更加不稳定, 在实际的评测中, 我们会多次测试取平均值。

4.5.2 浅层神经网络和深层神经网络

这个实验对比 2.2.1 节的模型和 3.1 节的模型。我们比较了不同深度的卷积神经网络在语义分类问题上的表现。

实验结果如表 4.4 和表 4.5 所示, 其中每层的规模写成了卷积长度*卷积宽度*输入特征映射空间数*输出特征映射空间数的形式。不论是那种模型, 词向量均采用 GoogleNews 的词向量且在训练过程中不更新, 参数的更新方式均采用二阶的 Adadelta 方法。在实验中, 我们发现在没有预训练的情况下, 模型的深度超过一定深度, 模型将不再收敛。(对于二分类的 MR, 层数大于 4 将不再收敛; 对于五分类的 SST, 层数大于 3 将不再收敛)

表格 4.4 不同配置的深度卷积神经网络在 MR 数据集上得表现

名称	2LPOOL	3LPOOL	4LPOOL	4LThin
Layer0	3*300*1*200	3*300*1*250	3*300*1*128	3*300*1*32
Pool0-1	2*1	1*1	1*1	1*1
Layer1	5*1*200*250	3*1*250*150	3*1*128*128	3*1*32*64
Pool1-2		1*1	1*1	1*1
Layer2		5*1*150*150	3*1*128*128	3*1*64*128
			2*1	2*1
			3*1*128*256	3*1*128*256
FinalPool	Global Max-Pooling			
Dropout	/	/	0.1,0.2,0.3,0.5	0.1,0.2,0.3,0.5
参数	43 万	45 万	31 万	16 万
准确率	79.22%	78.24%	79.23%	79.23%

表格 4.5 不同配置的深度卷积神经网络在 SST 数据集上的表现

名称	WCNN	2LDropout	2LPOOL	3LDropout	3LPOOL
Layer0	3*300*1*512	3*300*1*256	3*300*1*256	3*300*1*256	3*300*1*256
Pool0-1		1*1	2*1	1*1	2*1
Layer1		5*1*256*128	5*1*256*128	3*1*256*128	5*1*256*128
Pool1-2				1*1	1*1
Layer2				3*1*128*128	5*1*128*128
FinalPool	Global Max-Pooling				
Dropout	/	0.1,0.2	/	0.1,0.2,0.3	/
参数	46 万	40 万	40 万	38 万	48 万
准确率	47.01%	45.06%	45.15%	47.01%	43.61%

从表 4.5 的结果可以看出，深度的卷积神经网络用于语义分类的效果并不如简单地单层卷积神经网络。相反地由于模型深度的增加，把模型训练到收敛就更难了，当网络的深度大于一定值得时候，我们必须借助预训练等方法来训练网络，尽管如此，3.1 节所提出的深度卷积模型仍然得不到理想的结果。

4.5.3 对输入进行变换的模型

我们尝试过 3.3 节所提出的两个模型，然而在深度神经网络不加预训练的设定下，我们暂未能使这两个模型收敛。

图 4.3 是我们训练某一个句子所得到的相似矩阵，图中越亮的地方代表它所代表的值越大，我们可以明显看到这是一个对称矩阵且对角线所有元素都是 1。从一个句子映射到一个相对比较特殊的矩阵，其中信息的损失量是很大的，这可能就是这两个模型没有收敛的原因。

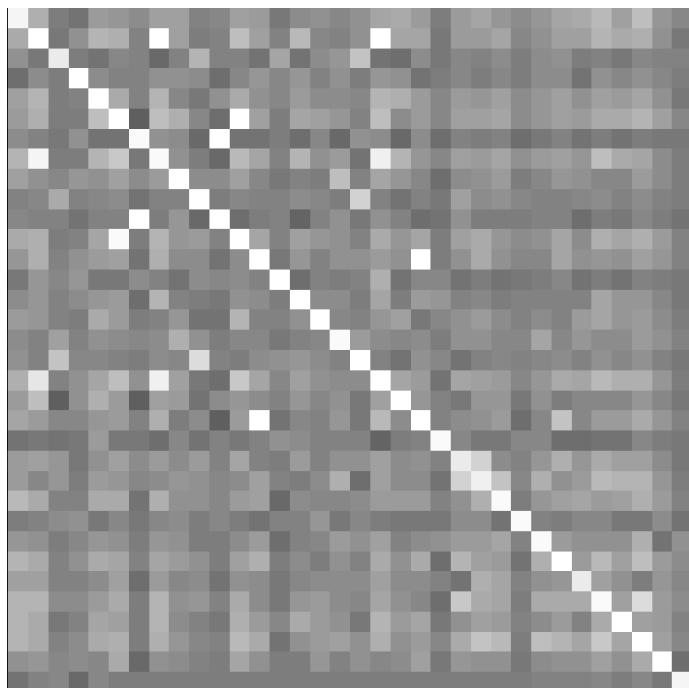


图 4.3 数据集中某一个句子的相似矩阵

通过实验，我们注意到并非出现在同一个句子中的词的词向量的夹角更小。词向量的夹角小只能说明词语的词义类似存在某种替代关系，并不一定存在组合关系。我们从 GoogleNews 中的词随机采样，它们词向量夹角 Cosine 绝对值的平均为 0.64，而在我们语料集中采样同时出现在一个句子中的词语，它们词向量夹角 Cosine 绝对值的平均只有 0.11。

4.5.4 非监督语料的作用

本实验的目的在于明确非监督语料所训练出来的词向量对整个语义分类系统的作用。词向量的质量对于语义分类系统的影响大小决定了我们只能利用大量非

监督语料所训练出来的词向量还是可以随机初始化然后再训练分类器的时候调整词向量，由于后者是直接利用卷积网络训练词向量的，所以这种情况下词向量的各个维度之间不一定是相互独立的，所以我们可以使用像图像那样的正方形小卷积核进行卷积计算，避免出现 3.3 节所说的 3.1 和 3.2 节模型存在的问题。

我们对非监督语料的多少设定了三个标准：第一个是非常多的非监督语料，这里是 GoogleNews 中上千万个单词，由此训练出来的词向量是 Google 提供的 300 维词向量（记为 GoogleNews）；第二个是少量的相关语料，这是我们从网络上抓取的大约 10 万个评论，由此我们利用 CBOW 和 Skip-Gram 算法训练出 300 维、100 维和 50 维三组词向量（分别记为 Vec300、Vec100 和 Vec50）；第三个是完全没有非监督语料，我们随机生成 300 维、100 维和 50 维三组词向量（分别记为 Random300、Random100 和 Random50）。对于在数据集中出现但是在非监督语料中没有出现的词，它们的词向量随机生成，所有单词的词向量在训练过程中都在不断调整。

我们使用 SST 作为测试数据，尝试上面七组词向量作为输入，每一种配置测试 10 次，最终结果取 10 次的平均值。实验结果表 4.6 所示，其中未出现单词比例指在数据集中出现但是没有在非监督语料中出现的单词占数据集单词总数的比例。

表格 4.6 非监督语料对语义分类模型性能的影响

词向量	未出现单词比例	10 次准确率平均值	10 次准确率方差(10^{-5})
GoogleNews	13%	47.26%	2.42
Vec300	37%	42.08%	18.83
Vec100	37%	43.07%	3.54
Vec50	37%	42.77%	0.98
Random300	100%	41.26%	13.12
Random100	100%	41.60%	10.46
Random50	100%	41.64%	8.08

从上面的结果可以看出，非监督语料的多少对整个分类模型的性能影响是很大的。更多的非监督语料训练出来的词向量能够使整个模型得到更好、更稳定的

结果，非监督语料一定的情况下，一定范围内词向量的维数越小，整个模型的表现更稳定。

4.5.5 错误分析实验

错误分析是分析一个模型的必要环节，通过错误分析，我们能够直观地看出模型的缺陷，更有利于我们对模型进行分析和改进。

在本实验中，我们使用五分类的 SST 作为数据集，研究卷积神经网络模型和反馈卷积神经网络的分类结果，分析哪些类型的句子容易分类正确而哪些类型的句子容易分错。

对于卷积神经网络模型，我们将最简单的浅层卷积神经网络（4.4.1 节中 BaselineCNN 模型）测试 5 次，各次结果差异不大，表 4.7 是其中一次结果，横向表示句子的真实属性，纵向表示句子的分类结果。

表格 4.7 浅层卷积神经网络分类结果

真实\预测	负面	偏负面	中性	偏正面	正面
负面	98	137	6	36	2
偏负面	90	352	36	145	9
中性	27	130	30	193	10
偏正面	4	43	13	386	62
正面	5	7	3	223	161

表 4.8 列举出了一些分类结果的句子，从这些分类结果的句子特征中，我们可以看出浅层卷积神经网络在用于语义分类时候的一些特性。

表格 4.8 一些分类结果的句子举例

真实情况是负面，被分类为正面（一共 2 句）
even die-hard fans of japanese animation ... will find this one a challenge .
it feels like a community theater production of a great broadway play : even at its best , it will never hold a candle to the original .
真实情况是正面，被分类为负面（一共 5 句）
if no one singles out any of these performances as award-worthy , it 's only because we would expect nothing less from this bunch .

the movie has an **avalanche** of eye-popping visual effects .

it never **fails** to engage us .

if this movie were a book , it would be a page-turner , you ca n't wait to see what happens next .

it represents better-than-average movie-making that does n't demand a **dumb** , **distracted** audience .

真实情况是负面，被分类为负面（随机抽取 10 句）

but as a movie , it 's a **humorless** , **disjointed** mess .

it does n't help that the director and cinematographer stephen kazmierski shoot on grungy video , giving the whole thing a **dirty** , **tasteless** feel .

schmaltzy and **unfunny** , adam sandler 's cartoon about hanukkah is numbingly **bad** , little nicky **bad** , 10 **worst** list bad .

the premise is in extremely **bad** taste , and the film 's supposed insights are so **poorly** thought-out and substance-free that even a high school senior taking his or her first psychology class could **dismiss** them .

unfortunately , the picture **failed** to capture me .

godawful **boring** slug of a movie .

this **miserable** excuse of a movie runs on **empty** , believing flatbush machismo will get it through .

desperately **unfunny** when it tries to makes us laugh and desperately **unsuspenseful** when it tries to make us jump out of our seats .

so **stupid** , so **ill-conceived** , so **badly** drawn , it created whole new levels of ugly .

i **hated** every minute of it .

真实情况是负面，被分为偏负面（随机抽取 10 句）

as with too many studio pics , plot mechanics get in the way of what should be the lighter-than-air adventure .

the film would work much **better** as a video installation in a museum , where viewers would be free to **leave** .

how on earth , or anywhere else , did director ron underwood manage to blow \$ 100 million on this ?

this is n't a `` friday " worth waiting for .

if this is the resurrection of the halloween franchise , it would have been **better off dead** .

obstacles are too easily overcome and there is n't much in the way of character development in the script .

a **waste of good** performances .

let 's cut to the consumer-advice bottom line : stay home .

looks like a high school film project completed the day before it was due .

it 's an awfully derivative story .

从上面的结果可以看出，不论是被错分为正面的负面句子还是被错分为负面的正面句子，它们大多表面上包含具有与它们所表达的感情色彩相反的词（已经用加粗标出），这类句子通常是观影者的间接表达或讽刺。相反地，被分类正确的负面句子（正面句子相似）大多都含有比较直接的、明显的含有感情色彩的形容词；而被误分为偏负面的负面句子则很少含有较为明显的形容词。

浅层的卷积神经网络提取句子结构信息的能力也比较有限，含有 not 和 never 等否定词的句子更容易分错，部分原因可能是该神经网络无法获取这些否定词所修饰的对象，而这些表示否定的词又相对均匀地分布在各个类别的句子中，基于局部相关的浅层网络不能很好地发现这些词所包含的信息。此外，句子中的一些结构和固定搭配的组成单词在空间上相距较远，这也加深了浅层卷积网络的分类困难。

对于反馈卷积神经网络模型，我们选取和普通卷积神经网络结构差距最远的迭代次数为 5 且同样不带有 Dropout 的模型（4.4.1 节中的 RCNN5），它的测试结果如表 4.9 所示，横向表示句子的真实属性，纵向表示句子的分类结果。

表格 4.9 反馈卷积神经网络分类结果

真实\预测	负面	偏负面	中性	偏正面	正面
负面	49	203	13	14	0
偏负面	30	491	37	68	6
中性	2	200	54	126	8
偏正面	0	111	17	326	54
正面	1	26	7	213	152

从上述结果可以看出，与普通的卷积神经网络相比，反馈卷积神经网络把句子属性完全分反的情况要少很多（表格的左下角和右上角），这说明后者在学习结构化特征的能力要强于前者。同时，我们也可以看到，反馈卷积神经网络把负面和偏负面、正面和偏正面搞混的情况依然没有得到好转，这可能是因为反馈卷积神经沿时间展开后本质将各阶特征相加，与普通深度网络将各阶特征分开处理相比，更容易造成混乱。

4.6 实验总结与分析

浅层神经网络句子模型有先天缺陷，它没有把卷积网络在图像识别上的优势发挥出来：它没有通过卷积层的叠加来抽取具有层次性的特征，它只能通过一层卷积来抽取非常基本的统计特征，而这些特征对于我们做语义分类而言是远远不够的。

本文中提出的深度卷积神经网络在语义分类问题上表现不佳的原因，本质是我们目前通过 Skip-Gram 或者 CBOW 所训练出来的词向量和卷积运算本身的不兼容。一方面，通过 4.3 节的实验，我们发现非监督语料和由此训练出来的高质量词向量的必要性和重要性，随机初始化的词向量很难取得比较好的结果；另一方面，通过 Skip-Gram 和 CBOW 算法的研究我们可以发现，通过这两种算法训练出的词向量各个维度是相互独立的，由这种词向量所得到的句子矩阵在词向量维度上不满足局部相关性。由于词向量的约束，我们在第一层只能使用相当于词向量维度宽的卷积核，所以在这种结构下，从神经网络的第二层开始，我们都是通过一维向量来刻画句子特征的，从此向上的卷积都退化为一维卷积。这样，这个网络就会产生一种“头重脚轻”的问题，第一层和第二层神经元个数的巨大差异（如果使用 300 维的 GoogleNews，则这种差异将会是 300 倍）会导致句子信息传递到第二层时的大量丢失，这不利于网络高层提取更高阶的特征。

反馈卷积神经网络是利用深层神经网络的优势来改造浅层神经网络，它的本质是一个权重共享的、带有短路连接的深层神经网络（正如图 2.1 左上角的展开图所示）。由上一段的讨论，我们发现深层神经网络并没有训练出比较高质量的高阶特征，在这种情况下，反馈卷积神经网络的自连接优势没有体现出来，实验中它的性能和一般的卷积神经网络相似。

第5章 总结与展望

本文针对用于语义分析的浅层神经网络模型存在的缺陷，结合深层神经网络和反馈卷积神经网络在图像分类问题上得优良表现，将这两个模型作为原型引入到语义分类问题中。然而，与图像问题不同的是，我们通过已有的算法学习到的词向量表示并不能很好地适应深层卷积网络，深层神经网络没有发挥出在图像上提取高阶特征的优势，在语义分类问题上效果表现不佳。由于反馈卷积神经网络可以视为带有短路的深层神经网络，它融合了低阶特征和高阶特征，在高阶特征没有有效提取出来的情况下，它的性能与普通浅层神经网络类似。

要在自然语言处理问题上发挥深层卷积网络的优势，一方面，我们可以对句子的输入进行变换，而这种变换必须满足变换后的输入矩阵能够在两个维度上均有局部相关性并且损失的信息足够小；另一方面，我们可以研究一种新的训练词向量的方法，使得在这个词向量设定下，句子矩阵在两个维度都具有很好的局部相关性。

插图索引

图 1.1 卷积神经网络	3
图 1.2 反馈神经网络	4
图 1.3 时间展开后的反馈神经网络	4
图 1.4 Skip-Gram 和 CBOW	5
图 2.1 反馈卷积神经网络	9
图 2.2 用于语义分类的浅层卷积神经网络	10
图 2.3 用于语义分类的深层卷积神经网络	11
图 2.4 用于比较句子相似度的卷积神经网络	12
图 3.1 深层卷积神经网络	13
图 3.2 反馈卷积神经网络	14
图 3.3 相似矩阵输入	15
图 3.4 线性变换输入	16
图 4.1 卷积神经网络在 MR 数据集上的训练情况	21
图 4.2 循环次数为 1 的反馈卷积神经网络在 MR 数据集上的训练情况	22
图 4.3 数据集中某一个句子的相似矩阵	24

表格索引

表格 4.1 浅层卷积和反馈卷积神经网络参数规模对比	20
表格 4.2 不同配置的卷积和反馈卷积神经网络在 MR 数据集上的表现 ...	20
表格 4.3 不同配置的卷积和反馈卷积神经网络在 SST 数据集上的表现 ..	21
表格 4.4 不同配置的深度卷积神经网络在 MR 数据集上得表现	23
表格 4.5 不同配置的深度卷积神经网络在 SST 数据集上的表现	23
表格 4.6 非监督语料对语义分类模型性能的影响	25
表格 4.7 浅层卷积神经网络分类结果	26
表格 4.8 一些分类结果的句子举例	26
表格 4.9 反馈卷积神经网络分类结果	28

参考文献

- [1] Erk K, Pado S. A structured vector space model for word meaning in context[C]. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2008: 897-906.
- [2] Mitchell J, Lapata M. Vector-based Models of Semantic Composition[C]. ACL. 2008: 236-244.
- [3] Mitchell J, Lapata M. Composition in distributional models of semantics[J]. Cognitive science, 2010, 34(8): 1388-1429.
- [4] Turney P D. Domain and function: A dual-space model of semantic relations and compositions[J]. Journal of Artificial Intelligence Research, 2012: 533-585.
- [5] Erk K. Vector space models of word meaning and phrase meaning: A survey[J]. Language and Linguistics Compass, 2012, 6(10): 635-653.
- [6] Clarke D. A context-theoretic framework for compositionality in distributional semantics[J]. Computational Linguistics, 2012, 38(1): 41-71.
- [7] Zettlemoyer L S, Collins M. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars[J]. arXiv preprint arXiv:1207.1420, 2012.
- [8] Küchler A, Goller C. Inductive learning in symbolic domains using structure-driven recurrent neural networks[M]. KI-96: Advances in Artificial Intelligence. Springer Berlin Heidelberg, 1996: 183-197.
- [9] Socher R, Pennington J, Huang E H, et al. Semi-supervised recursive autoencoders for predicting sentiment distributions[C]. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011: 151-161.
- [10] Hermann K M, Blunsom P. The Role of Syntax in Vector Space Models of Compositional Semantics[C]. ACL (1). 2013: 894-904.
- [11] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]. Advances in neural information processing systems. 2012: 1097-1105.
- [12] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
- [13] Funahashi K, Nakamura Y. Approximation of dynamical systems by continuous time recurrent neural networks[J]. Neural networks, 1993, 6(6): 801-806.
- [14] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]. Advances in neural information processing systems. 2013: 3111-3119.

- [15] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.
- [16] Liang M, Hu X. Recurrent Convolutional Neural Network for Object Recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3367-3375.
- [17] Kim Y. Convolutional neural networks for sentence classification[J]. arXiv preprint arXiv:1408.5882, 2014.
- [18] Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences[J]. arXiv preprint arXiv:1404.2188, 2014.
- [19] Hu B, Lu Z, Li H, et al. Convolutional neural network architectures for matching natural language sentences[C]. Advances in Neural Information Processing Systems. 2014: 2042-2050.
- [20] 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: 115-124.
- [21] Socher R, Perelygin A, Wu J Y, et al. Recursive deep models for semantic compositionality over a sentiment treebank[C]. Proceedings of the conference on empirical methods in natural language processing (EMNLP). 2013, 1631: 1642.
- [22] Zeiler M D. ADADELTA: An adaptive learning rate method[J]. arXiv preprint arXiv: 1212.5701, 2012.

致谢

在综合论文训练期间，衷心感谢胡晓林老师对我的悉心指导。从开题到中期再到最终成稿，胡老师每一周都会专门抽出时间跟我讨论研究进展与想法，他严谨治学的态度使我在综合论文训练期间受益匪浅，特别是在神经网络方面获得了很多最新知识。同时，我也感谢实验室的梁鸣师兄和钱桥同学，你们在我有问题不太清楚的时候随时答疑解惑，包括实验的各个细节，在此由衷地表示敬意。

声明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签 名：_____ 日 期：_____

附录 A 外文资料的调研阅读报告或书面翻译

用卷积神经网络来刻画句子

Nal Kalchbrenner, Edward Grefenstette, Phil Blunsom

Department of Computer Science

University of Oxford

摘要

精确地表示句子的能力是语义理解技术的核心。我们将介绍一个被称作动态卷积神经网络的模型（Dynamic Convolutional Neural Network, DCNN）并利用这个模型来刻画句子的语义特征。这个神经网络在句子上使用一种称为动态 K-Max Pooling 的运算。它能够处理不同长度的句子并且引入一个句子矩阵使我们能够显式捕捉句子中存在的长短不一的关系。此外，该模型不依赖语法树并且能够被应用到各种语言上去。我们将在四个实验中来测试这个模型：小规模的句子二分类问题、小规模的句子多分类问题、六种类型的问题分类任务和 Twitter 语义预测。这个网络模型在前三个问题中都取得了非常好的结果并且在最后一个问题上的错误率比之前最好的方法低 25%。

1. 前言

句子模型的目的在于分析或者表示句子的语义特征以便于进行句子分类或句子生成操作。很多涉及到自然语言理解的问题的核心就是构造合适的句子模型。这些问题包括语义分析、短语检测、含义识别、摘要生成、对话分析、机器翻译和基础的语言学习以及图像检索。然而单独的句子却很少或者压根没有真正被关注过，我们必须通过一个句子中经常连续出现的一个或者几个词来提取句子的特征。提取句子特征的核心问题是找到一个定义了从句子中连续出现的一个或者几个词映射到句子特征的函数。

前人提出过很多基于此的模型。基于组合的方法从词语共现中得到词向量表示词义，以便于得到更长短语的向量表示。在一些情况下，这种组合就是词义向量通过代数运算得到句子向量的过程（Erk and Pado, 2008; Mitchell and Lapata, 2008; Mitchell and Lapata, 2010; Turney, 2012; Erk, 2012; Clarke, 2012）；在其它情况下，我们学习一个组合函数并且把它用在特定的语义关联（Guevara, 2010; Zanzotto et al., 2010）或者一些特定类型的词语上（Baroni and Zamparelli, 2010;

Coecke et al., 2010; Grefenstette and Sadrzadeh, 2011; Kartsaklis and Sadrzadeh, 2013; Grefenstette, 2013)。此外，还有一些研究者利用自动提取的逻辑关系来得到句子的表示 (Zettlemoyer and Collins, 2005)。

一类很主流的模型是基于神经网络的，它们的范畴从最基础的基于词袋模型的神经网络到更加结构化的递归神经网络再到带有时间延迟的基于卷积的神经网络 (Collobert and Weston, 2008; Socher et al., 2011; Kalchbrenner and Blunsom, 2013b)。神经网络模型有很多的优势。通过预测训练诸如上下文单词或者短语等信息，它们能够得到更加通用的词向量和短语向量。通过这些有监督的学习，神经网络句子模型能够学习到针对特定问题的词向量。除了将强大的分类器整合到模型中，神经网络模型还能学习一个语言模型，使之能够一个词一个词地生成一句话 (Schwenk, 2012; Mikolov and Zweig, 2012; Kalchbrenner and Blunsom, 2013a)。

我们定义一个卷积神经网络模型并且把它用到刻画句子语义特征的问题上来。这个网络处理长短不一的输入句子，然后交错使用一维的卷积运算和动态 K-Max Pooling 运算。动态 K-Max Pooling 是 Max Pooling 运算的一个泛化，后者是一个非线性采样函数，该函数返回输入值中最大的那个 (LeCun et al., 1998)。我们从两个方面对 Max Pooling 进行泛化：首先，K-Max Pool 是返回前 K 个最大值而不是单独一个最大值；然后，K 也可以是一个变量，我们通过网络的输入规模来动态决定 K 的值。

卷积层在句子矩阵的每一个特征方向上使用一维卷积核，在句子的不同位置使用同样地一组卷积核，这样能够保证提取出来的特征不受词语在句子中的位置影响。每一个卷积层之后是一个动态 Pooling 层，经过非线性的 Pooling 之后，我们便得到一个特征映射空间。像卷积神经网络应用到物体识别中一样 (LeCun et al., 1998)，我们通过多个不同的卷积核计算多个特征映射空间来丰富对输入句子的特征表示。后续的卷积层都有很多不同的特征映射空间，它们是由下层的特征空间通过卷积运算得到的，卷积核的权重构成一个四维的张量。这样的一个模型，我们称之为动态卷积神经网络。

多层卷积核加上动态 Pooling 运算能够提取出输入句子的结构化特征，图 A-1 阐述了这样的结构。高层的小卷积核能够捕捉到输入句子中距离很远短语的句法或语义关系。这种层次性的特征分布某种程度上跟句法树很类似，但是这种结构并不是建立在纯粹的句法关系上，这是神经网络的内在性质。

我们在四种配置上测试这个网络。前两个实验包括对电影评论的语义进行预测 (Socher et al., 2013b)。在二分类和多分类问题上，这个网络的性能超过了其

它任何应用。第三个实验包括在 TREC 数据集上 (Li and Roth, 2002) 对六种类型的问题进行分类。这个网络的准确率能够与其它最好的方法不相上下，而后者是建立在工程化提取的大量特征和一些手工的知识来源上的。第四个实验包括利用 distant supervision 方法 (Go et al., 2009) 来预测发送的 Twitter 语义。这个网络通过 160 万根据情绪自动标好标签的 Twitter 来进行训练，在人工标定的测试集上，该网络的错误率比最强大的基于 unigram 和 bigram 的 Baseline (Go et al. 2009) 低超过 25%。

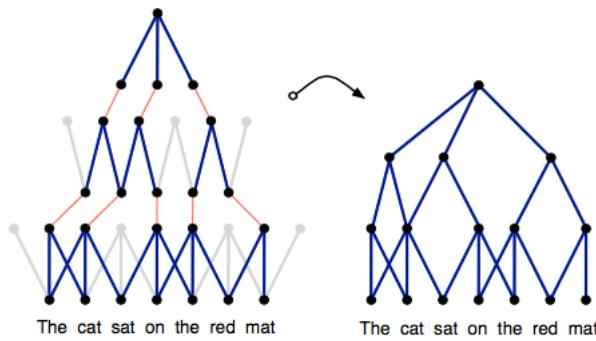


图 A-1 DCNN 所生成的特征图的子图，整个图的有很多边互不交叉的这种子图，这些子图会在不同的层次连接。上图左半部分强调了 Pooling 运算。左右半图卷积核的大小分别为 3 和 2. 通过动态 Pooling，高层的一个窄的卷积核能够联系到输入句子中相对较远的短语

下面是这篇论文的概要：第二部分阐述 DCNN 模型的背景，包括最核心的概念和相关的神经网络句子模型；第三部分定义了相关的运算符和网络的层次；第四部分说明了网络模型的特征分布和其它特性；第五部分分析实验结果和网络学到的特征。

2. 背景

DCNN 的各层是由一个卷积运算和一个 Pooling 运算相互连结而成。我们首先回顾相关的神经网络句子模型，然后我们介绍一维卷积运算和经典的延时神经网络 (Time-Delay Neural Network, TDNN) (Hinton, 1989; Waibel et al., 1990)。通过添加一个 Max-Pooling 层，TDNN 也可以成为一个句子模型 (Collobert and Weston, 2008)。

2.1 相关的句子模型

我们已经介绍了很多神经网络句子模型。一种基本类型的句子模型是词袋模型 (Neural Bag-of-Words models, NBoW models)。这些模型一般都会有一个投影层，它把词、词的组成部分或者连续几个词映射到一个高维的表示空间上，这些

语言单元会通过某种方式混合，混合后的向量会通过一个或多个全连接层接入到一个分类器上。

有一种神经网络模型，它采用了由外部句法树提供的结构，这种模型称作递归神经网络模型（Recursive Neural Network, RecNN）（Pollack, 1990; Kuchler and Goller, 1996; Socher et al., 2011; Hermann and Blunsom, 2013）。这个树中的每一个节点都是它的左儿子和右儿子通过某种层连接合并得来，而这种层连接中的权值是全局共享的，该树计算出来的根节点作为它所对应的一个句子的表示。反馈神经网络（Recurrent Neural Network, RNN）可以看成 RecNN 的一种特殊形式，前者可以看成后者所有节点分布在一个线性链上的情况（Gers and Schmidhuber, 2001; Mikolov et al., 2011）。RNN 主要被当成一个语言模型，但是也可以看成一个具有线性结构的句子模型，最后一个词输入完后计算所得的状态就是整个句子的表示。

最后，一种更深一步的神经网络模型是建立在卷积运算和 TDNN 结构上的（Collobert and Weston, 2008; Kalchbrenner and Blunsom, 2013b）。这些模型中所使用的概念也是 DCNN 的核心概念，我们下面会阐述它们。

2.2 卷积

一维卷积是在一个 M 维权重向量 m 和一个 S 维输入向量 s 之间的运算。向量 m 被称为卷积运算的卷积核，向量 s 可以视为输入序列。在本问题中， s_i 是句子中第 i 个词在某一个特征维度上的值。一维卷积的思想是将向量 m 和句子（可以看成词的序列） s 中得每个 M -gram 进行点积得到一个新的序列 c 。

$$c_j = m^T s_{j-m+1:j}$$

根据下标 j 的取值，上式可以解释两种类型的卷积运算。窄型卷积需要 $S \geq M$ 并且得到的是一个 $(S-M+1)$ 维序列，此时 j 的取值范围是从 M 到 S ；宽型卷积没有对 S 或者 M 的限制并且得到的是一个 $(S+M-1)$ 维序列，此时 j 的取值范围是从 1 到 $S+M-1$ ，对于向量 s 而言，所有下标超出范围的值均视为 0。窄型卷积的结果是宽型卷积所得序列的一个子序列，这两种类型的卷积运算会在图 A-2 中阐明。

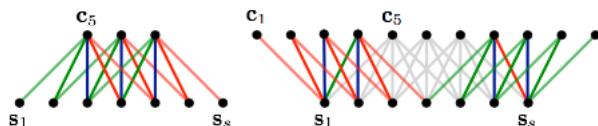


图 A-2 窄型和宽型卷积，卷积核的长度为 5。

一个训练好权重的卷积核 m 对应发现一个语言特征，这个特征能够识别一类连续出现的 N 个词，其中 $N \leq M$ 且 M 为 m 的维度。将向量 m 运用到宽型卷积比

运用到窄型卷积要更有优势。宽型卷积能够保证卷积核中得所有权重都能够到达整个句子，包括那些边缘的单词，这个性质在 M 比较大（例如 8 到 10）的时候特别重要。另外，不管 M 和 S 的值如何，宽型卷积能够保证卷积核 m 和输入句子 s 的运算能够始终得到一个有效的、非空的结果 c 。我们接下来将介绍 TDNN 模型中的经典卷积层。

2.3 延时神经网络 TDNN

TDNN 用一些卷积核 m 对输入的序列 s 进行卷积运算。像被用于语言识别的 TDNN (Waibel et al., 1990)，序列 s 被看做有一个时间维度而卷积就是被用在时间维度上的。每一个 s_j 通常不只是一个值，而是一个 D 维向量，所以 s 是一个 $D \times S$ 的矩阵。同样地， m 也是一个 $D \times M$ 的权值矩阵， m 的每一行都会和 s 的对应行进行卷积运算并且这时的卷积往往是窄卷积。通过把输出序列 c 作为下一层的输入序列，我们可以将若干个卷积层叠起来。

Max-TDNN 句子模型是建立在 TDNN 模型 (Collobert and Weston, 2008) 之上的。在这个模型中，卷积层对句子矩阵 s 使用的是窄型卷积。如下所示句子矩阵的每一列对应它的一个词的特征向量 w_i :

$$s = [w_1, w_2 \dots w_s]$$

为了解决句子长短不一的问题，Max-TDNN 取了矩阵 c 中每一行的最大值组成了一个 D 维的向量，如下所示:

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

这样做的目的是提取出最显著的特征，比如我们会从矩阵 c 中得每一行挑选最大的值。这样，大小恒定的向量 c_{max} 就作为输入可以通过一个全连接层来进行分类了。

Max-TDNN 模型有很多很好的性质。它对句子中词语的排列顺序敏感而且它不需要诸如语法树等外部的跟语种相关的特征。除了在进行窄型卷积的时候句子两侧的词语考虑的次数会少些，它给予了句子中每个词相同的权重。但是，该模型也有一些缺点。特征检测向量能够检测的特征范围被限制为 M ，增加 M 的值或者将若干个窄型卷积层相叠可以添加检测特征的范围，然而这也同时加重了对边缘单词的忽略以及对输入句子的最小长度提出了更高的要求。正因为这个原因，这个模型不能整合进高阶的、长距离的特征。此外，Max-Pooling 方法也有一些弊端：它不能分辨出 c 矩阵一行中的特征出现一次还是多次并且它会丢失这些特

征出现的顺序等信息。更加概括地说，Pooling 运算将矩阵 c 中每一行的 $S-M+1$ 个值压缩为一个值，对于 c 矩阵中等的值来说，这种方法损失的信息量很大。下面一节就旨在保留 Pooling 运算的优势的同时解决这些问题。

3. 带有 K-Max Pooling 的卷积神经网络

我们使用一种新的卷积神经网络来刻画句子，这个网络交替使用宽型卷积和动态 K-Max Pooling。在这个网络中，中间层 feature map 的宽度取决于输入句子的长度。这样的模型我们称为动态卷积神经网络模型，如图 A-3 所示。接下来，我们将详细描述这个模型。

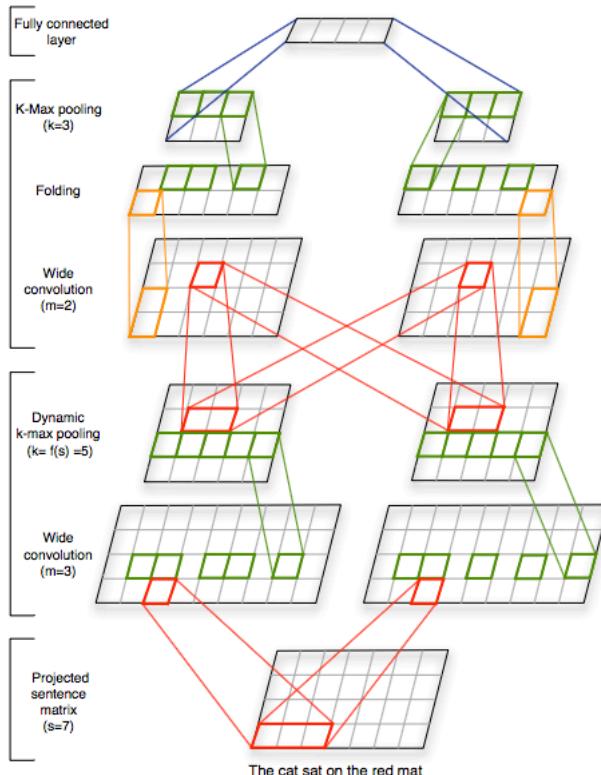


图 A-3 这是一个输入 7 个单词组成句子的 DCNN，词向量的维度为 4。这个神经网络有两个卷积层并且每个卷积层有两个特征空间，这两个卷积层的卷积核长度分别为 3 和 2。（动态）K-Max Pooling 层 K 的值分别为 5 和 3。

3.1 宽型卷积

给定一个输入句子，我们利用句子中每个词的词向量 w_i 来组成一个句子矩阵作为第一层神经元的输入。这些词的词向量的值是在训练的过程中不断调整优化的。网络中的卷积层将一个 $D \times M$ 维的权重矩阵与输入神经元的值进行卷积计算，例如第二层神经元就是通过将句子矩阵和卷积核作卷积计算得到。这里的卷积运算是指 2.2 节中所称的宽型卷积，所以得到矩阵的维度是 $D \times (S + M - 1)$ 。

3.2 K-Max Pooling

我们接下来将介绍一种 Pooling 运算，它不同于 Max-TDNN 模型中在时间维度上的 Max Pooling 和在物体识别应用中的 Max Pooling 运算(LeCun et al., 1998)。给定一个值 K 和一个长度为 P 的序列 p ($K \leq P$)，K-Max Pooling 就是选取 p 中值最大的 K 个值组成一个子串 p_{\max}^K ， p_{\max}^K 中元素的排列顺序与 p 一致。

K-Max Pooling 运算使从 p 中提取 K 个最活跃的特征成为可能，而这 K 个特征也许是不连续的。这个运算保证了这 K 个特征的顺序，但是并不关心它们原来的位置，它更加关注活跃特征出现的次数和它们在 p 中的变化。K-Max Pooling 被应用在卷积层的最顶端，这保证了全连接层的输入跟句子的长短没有关系。但是，我们下面会指出，在中间的卷积层后的 Pooling， K 的值是不固定的，它会动态选择以便能够平滑提取高阶的和更加宽范围的特征。

3.3 动态 K-Max Pooling

当我们让 K 成为输入句子的长度和网络深度的函数时，K-Max Pooling 就成为动态 K-Max Pooling。尽管可以构造很多函数，这里我们构造了如下一个相对简单地模型。

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

其中 l 代表当前使用 Pooling 运算的卷积层深度（从下往上从 1 开始）， L 是整个网络中卷积层的总数， k_{top} 是 3.2 节中最顶层卷积层的固定 Pooling 参数。例如，对于一个 3 层的卷积神经网络并且 $k_{top} = 3$ ，如果输入一个长度为 18 的句子，第一层的 Pooling 参数 $K_1 = 12$ ，第二层的 Pooling 参数 $K_2 = 6$ ，而第三层的 Pooling 参数是一个定值 $K_3 = k_{top} = 3$ 。上面的式子给定了在一定输入句子长度情况下，各层循序渐进捕捉特征的情况。例如，在语义预测实验中，一阶特征是诸如在句子中出现次数最多的正面词语，而二阶特征是出现次数最多的否定短语或者从句。

3.4 非线性特征函数

在对卷积之后的结果进行（动态）K-Max Pooling 之后，一个偏置 b 和一个非线性函数会逐元素地用在经过 Pooling 之后的矩阵上。在该矩阵的每一行以后都有一个偏置值 b 。

我们暂时先忽略 Pooling 层，我们讨论如何计算经过卷积和非线性变换后矩阵 a 中的每一个 D 维列向量。我们定义 m 是一个对角矩阵。

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

上式中的 m 是用于宽型卷积的 d 个卷积核的权值。在经过第一个卷积核非线性变换后，矩阵 a 中第 j 列是这样得到的。

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

这里的 a 可以认为是一阶特征的一列。二阶和三阶特征能够通过类似的方法得到，不过这里输入换成了矩阵 a 并且 M 矩阵也是不同的。除了 Pooling 外，上式是特征提取的核心函数并且有一个更加一般化的形式，这个我们下面会提到。加上 Pooling 之后，这个函数能够做到位置不敏感和提取高阶特征的作用。

3.5 多重特征空间

到现在为止，我们介绍了怎么通过宽型卷积、(动态) K-Max Pooling 和非线性函数来从输入句子矩阵得到一阶特征映射空间 (feature map)。这三种操作可以反复应用来得到更高阶的空间和一个更深的网络，我们将 i 阶的特征记为 F^i 。正如被用于物体识别的卷积神经网络一样，我们可以增加学习到的特征映射空间的数量，而计算这些在同一层中特征 $F_1^i, F_2^i \dots F_n^i$ 是可以并行的。通过将下一层的低阶特征 F_k^{i-1} 与本层的卷积核 $m_{j,k}^i$ 进行卷积计算，我们能够得到本层的特征 F_j^i 。

$$F_j^i = \sum_{k=1}^n m_{j,k}^i \otimes F_k^{i-1}$$

上式中的卷积都是宽型卷积，所有权值 $m_{j,k}^i$ 都是一个四阶张量。通过这样的一个宽型卷积之后，动态 K-Max Pooling 和非线性函数将被运用到每一个特征上去。

3.6 折叠

从目前所描述模型看，被用于句子矩阵某一行的用于抽取特征的卷积核可以在该行中抽取高阶特征，也可以抽取多个特征。然而，在顶部的全连接层之前，不同行的卷积核都是互相独立的。通过将 3.4 节中的稀疏的、对角的 M 矩阵改为稠密的矩阵，我们能够引入不同行卷积核之间的依赖关系。这里我们提出一种叫折叠的方法能够避免引入太多的额外参数：在卷积层之后和(动态)K-Max Pooling 层之前，我们只需要将特征矩阵中相邻两行对应的元素相加即可。对于一个有 D 行的特征矩阵，折叠后其行数变为 $D/2$ ，其规模减半。有折叠层之后，上层的特征矩阵的一行会依赖于下层的两行，这就是完整地 DCNN 模型。

4. 句子模型的性质

我们接下来介绍基于 DCNN 的句子模型的一些性质。通过对一个句子进行卷积和 Pooling 运算，我们将阐述特征图的概念。我们会简要地将这些性质并且和其它神经网络句子模型的性质联系起来。

4.1 单词和 n-Gram

该模型的一个基本的性质是对句子中词语顺序的敏感。在大多数应用中，为了学到更加精细的特征，能够判别一句话中有没有连续出现 n 个特定的单词（称为 n-Gram，下同）对模型提升性能是十分有好处的。同样地，能够区分出相关的几个 n-Gram 的相对位置也是一个模型很好的特征。我们的模型就是为了捕捉这两个方面的信息而设计的。在第一层的宽型卷积计算中， M 维的卷积核能够捕捉到跨度小于 M 的 n-Gram；我们将在实验中看到，第一层的 M 通常设定为一个较大的值例如 10。通过 Pooling 运算提取出来的 n-Gram 序列虽然忽略了绝对位置信息，但是还是保持了它们的顺序和相对位置。

关于其它的神经网络句子问题，NBoW 类模型肯定是忽略词语之间的顺序的。基于反馈神经网络的句子模型对词语顺序敏感，但是它对最后输入的单词有偏袒 (Mikolov et al., 2011)，这使得反馈神经网络在语言模型上表现优秀，但是它在记忆距离结尾很远处的 n-Gram 上很有劣势。类似地，递归神经网络对词语顺序也敏感但是它更加偏袒位于树顶端的节点，比较浅的树结构能从一定程度上缓解这种问题 (Socher et al., 2013a)。2.3 节也指出，Max-TDNN 模型也对词语顺序敏感，但是 Max Pooling 层只从输入的句子矩阵中抽取一个 n-Gram 特征。

4.2 生成的特征图

一些句子模型使用内部的或者外部的结构来计算句子的表示。在 DCNN 中，卷积层和 Pooling 层能够生成关于输入的内部特征图：如果下层的神经元参与到上层某个神经元的计算中，那么这两个神经元被视作相连的；没有被 Pooling 运算选中的神经元将从图中丢弃；在最后一个 Pooling 层之后，剩下的神经元将会连接到一个顶层根节点中。这样生成的是一个由根节点和相互连接的无循环的边构成的图，图 A1.1 提供了两种等效的表示。在没有折叠层的 DCNN 中，输入句子矩阵中的每一行都是一个子图，这些子图只在根节点相互连接。每一个子图都可能有不同的形状，这些形状反映了这个子图提取的特征类型。折叠层的作用在于让这些子图能够在根节点之前连结起来。

用于物体识别的卷积神经网络也同样在输入图像的基础上生成了一个特征图。DCNN 的特征图之所以奇特，是因为它采用了一个全局的 Pooling 运算。(动态) 的 K-Max Pooling 运算符能够刻画分布在句子不连续位置的特征，高阶的特征所

涉及的范围可以短小且集中也可以长至整个句子的长度。同样地，DCNN 生成特征图子图的边缘也可以反映这种范围的可变性，某一子图既可以聚焦在句子的某一个或者几个部分，也可以扩展到整个句子中。这种结构是神经网络的内在特性并且通过输入沿着前向网络的传递得以定义。

在其它的句子模型中，NBoW 模型是浅层的而 RNN 模型具有一个线性链的结构。Max-TDNN 所生成出得特征图子图通过 Max Pooling 只能得到一个固定范围的特征。递归神经网络采用了一个外部语法树的结构，不同范围的特征是通过一个或多个树的子节点合成父节点的过程计算出来的。跟 DCNN 学习出一个明显的特征层次结构不同的是，在递归神经网络中，诸如单个单词的低阶特征可能和和诸如整个短语的高阶特征合并。在结构方面，DCNN 对 RecNN 进行了泛化，3.4 节中得公式比 RecNN 的合并函数更加通用，后者相当于前者等式中 $m=2$ 的情况。DCNN 所提取出来的特征图也比句法树更加通用，它不仅仅再局限在句法上的短语，它还能捕捉到一些句法树无法对应到的或长或短的语义关系。DCNN 具有内在的、只与输入相关的结构，它不依赖外部提供的句法树。这些使得 DCNN 能够直接应用到很难提取句法树的句子（如 Twitter）或者其它语言上。

5. 实验

我们从四个不同的实验来测试这个网络模型。我们首先明确实现的各个方面和模型的训练，然后我们展示实验结果和分析学到的特征。

5.1 训练

在所有的实验中，最顶层的网络都是一个全连接层后接一个 Softmax 非线性分类器，它预测给定输入句子的情况下，该句子属于某一类的概率分布。网络训练旨在减小预测分布和真实分布的 cross-entropy，目标还包括一个针对所有系数的 L2 正则化。网络的系数包括所有词的词向量、卷积核的权重和全连接层的权重。网络采用 mini-batch 误差回传的方式训练，更新采用基于梯度优化的 Adagrad 更新法则 (Duchi et al., 2011)。运用熟知的卷积理论，我们可以利用快速傅里叶变换快速计算输入矩阵每一行的一维卷积。为了充分利用平行计算的优势，我们在 GPU 上训练这个网络。一个 Matlab 代码能够一个小时在 GPU 上处理几百万个句子，这主要取决于神经网络的层数。

5.2 Movie Review 上的语义预测

前面两个实验都涉及对 Stanford Sentiment Treebank (Scocher et al., 2013b) 上的影评进行语义分类。其中一个实验的输出只有正负二元而另一个实验有五种可能的输出：负面、偏负面、中性、偏正面和正面。在二分类实验中，我们使用 6920

个训练样例、872 个验证样例和 1821 个测试样例，在精细分类中，我们使用数据所给定的 8544/1101/2210 划分。在训练句子中被标记的短语被作为一个独立的训练样例训练。数据集词汇总量为 15448。

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

表 A-1 在影评数据集上语义分类的准确性。最前面的四个结果是 Socher 等人 2013 年所报告的。NB 和 BiNB 模型是朴素贝叶斯分类器分别采用 unigram 特征、unigram 加上 bigram 特征得到的。SVM 是采用 unigram 和 bigram 特征的支持向量机。RecNN 是用基于张量函数作为非线性的递归神经网络，它依赖从外部句法树得到的特征。

表 A-1 展示了实验结果的细节。在三个神经网络句子模型——Max TDNN, NBoW 和 DCNN——中，词向量都是作为模型的参数并且随机初始化的，它们的维度为 48。Max-TDNN 第一层使用窄型卷积的卷积核的宽度为 6，比较短的短语都使用零向量进行补全。卷积层后面接的是一个非线性的 Max-Pooling 层和一个 Softmax 分类器。NBoW 模型是将词向量相加后接一个非线性函数和 Softmax 分类器，我们采用的非线性函数是 tanh 函数。DCNN 的超参数设定如下。对于二分类问题，模型是一个宽型卷积层后接一个折叠层，之后是一个 K-Max Pooling 层和非线性映射；它的第二个宽型卷积层后面是一个折叠层，之后是一个 K-Max Pooling 层和非线性映射。卷积层的宽度分别为 7 和 5，最顶层 K-Max Pooling 层的 K 值为 4。第一个卷积层的特征映射空间数为 6，第二个卷积层的特征映射空间数为 14. 模型的最顶端是一个 Softmax 分类器。用于五分类问题的 DCNN 有着相同的结构，但是卷积核的大小分别为 10 和 7，最顶层 K-Max Pooling 层的 K 值是 5。两个卷积层的特征数分别为 6 和 12。整个模型使用 tanh 作为非线性函数。在训练的过程中，我们使用 Dropout 在倒数第二层的非线性函数之后(Hinton et al., 2012)。

我们可以看到 DCNN 显著地比其它神经网络或者非神经网络模型都要表现地好。NBoW 模型跟非神经网络的基于 n-Gram 的分类器表现类似。Max-TDNN 比 NBoW 模型表现差，这可能是 Max Pooling 运算的过度压缩所致，它丢弃了输

入句子中绝大部分语义特征。除了 RecNN 需要使用一个外部的 parser 来生成模型的结构特征，其它模型都是使用基于 n-Gram 或者神经元特征的，而后者是不需要外部资源和额外标记的。在下面的实验中，我们将把 DCNN 对比那些使用了工程化色彩很浓的资源的方法。

5.3 问题种类分类

作为问题回答的辅助，一个问题需要被分为属于某一问题的类别。TREC 数据集包括了 6 中不同类别的问题，例如关于地点的，关于人物的，或者一些关于数字信息的 (Li and Roth, 2002)。训练数据集包括 5452 条有标签的问题，测试数据集包括 500 个问题。

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

表 A-2 这是在 TREC 数据集上的六种问题分类准确率。第二列是在各种方法中所用到的外部特征。前四个结果的来源分别是：Li and Roth (2002), Blunsom et al. (2006), Huang et al. (2008) 和 Silva et al. (2011)。

结果展示在表格 A-2 中。非神经网络的方法使用了基于大量人手工提取特征和手工标注资源的分类器。例如 Blunsom 等 (2006) 提出了一个最大熵模型，它依赖于 26 种句法的和语义的特征，包括 unigram, bigrams, trigrams, POS 标签，专有名词标签，CCG parse 的结构化信息和 WordNet 同义词集合。我们在这个数据集上使用的三个神经网络模型采用和 5.2 节中二分类问题有相同的超参数配置。因为数据集的规模很小，我们使用维度较低的 32 维词向量，它们是通过一种预测上下文共现的非监督方法训练得来 (Turian et al., 2010)。DCNN 使用一个卷积核长度为 8、特征数为 5 的卷积层。DCNN 和那些在表格 A2 中表现比较好的方

法差异不大（准确率差异小于 0.09）。在给定的用来训练的带标签信息只是训练集本身的情况下，很明显，神经网络模型的性能能够与那些依赖大量手工特征规则和手工资源的方法相媲美。

5.4 使用远距离监督对 Twitter 语义进行预测

在我们最后一个实验中，我们用大量的 Twitter 状态来训练模型。根据它里面的情感，每一条状态都能够被自动标记为正面或者负面。训练集数据包括 160 万条状态和它们的语义标签，测试集包括 400 条手工标定的状态。我们根据 Go 等人 (2009) 的方法最小化地预处理数据，此外，我们还把所有大写字母变成小写，这产生了 76643 个不同的词语。DCNN 和其它神经网络模型的结构都采用和 5.2 节中二分类问题一样的配置，随机初始化的词向量的维数增加到 60 维。表格 A-3 报告了这个实验的结果。我们可以看到 DCNN 跟基于 n-Gram 的非人工神经网络分类器相比性能有明显的提升，而在大规模训练数据的情况下后者是很强大的 Baseline。我们看到把 DCNN 应用到自动提取情感特征的大规模语料中进行情感分类任务能够得到很好的结果。DCNN 和 NBoW 模型的性能差异更加能够说明 DCNN 提取长范围 n-Gram 特征和将这些特征层次化组合的能力，这些能力对于处理分类问题是十分有利的。

Classifier	Accuracy (%)
SVM	81.6
BiNB	82.7
MAXENT	83.0
MAX-TDNN	78.8
NBoW	80.9
DCNN	87.4

表 A-3 在 Twitter 语义数据集上得准确率。三个非神经网络模型都是基于 unigram 和 bigram 特征的，它们的结果是 Go 等人在 2009 年报告的。

5.5 特征抽取可视化

DCNN 中得每一个卷积核都跟一个特征抽取符或者神经元密切相关，它在训练的过程中学习使得只要输入一个特定的词语序列就能够激发。在第一层，这个序列是一个连续的出现的 n 个单词 (n-Gram)；在更高的层次里，这个序列可由若干个分离的 n-Gram 组成。我们将 5.2 节中用于语义二分类问题的网络第一层特征描述符可视化。因为每个卷积核的宽度为 7，所以对于这 288 个特征描述符中每一个，我们将在验证集和测试集中出现的每一个连续 7 个词根据他们对卷积核

的响应大小进行排序。图 A-4 展示了 4 个特征描述符的前 5 名，除了我们预料之中的表示正面或者负面语义的特征描述符，我们也发现了诸如 ‘not’ 那样否定语义的描述符和诸如 ‘too’ 那样增强语义的描述符。我们还发现了像 ‘all’、‘or’、‘with ... that’、‘as ... as’ 那样具有明显结构的描述符。这些特征描述符不仅仅能够识别出单个 n-Gram，而且能识别出 n-Gram 之间句法的、语义的或者结构的信息。

'NOT'					
n't	have	any	huge	laughs	in
no	movement	,	no	,	not
n't	stop	me	from	enjoying	much
not	that	kung	pow	is	of
not	a	moment	that	is	funny
					false
'TOO'					
,	too	dull	and	pretentious	to
either	too	serious	or	too	be
too	slow	,	too	long	lighthearted
feels	too	formulaic	and	too	,
is	too	predictable	and	too	and
					too
POSITIVE					
lovely	comedic	moments	and	several	fine
good	script	,	good	dialogue	performances
sustains	throughout	is	daring	,	funny
well	written	,	nicely	acted	inventive
remarkably	solid	and	subtly	satirical	and
				tour	beautifully
NEGATIVE					
,	nonexistent	plot	and	pretentious	style
it	fails	the	most	basic	test
so	stupid	,	so	ill	as
,	too	dull	and	pretentious	conceived
hood	rats	butt	their	ugly	,
				heads	be
					in

表 A-4 在网络的第一层，4 个特征描述符所刻画的 7-Grams 的前五名。

6. 结论

我们介绍了一种将 K-Max Pooling 运算作为一种非线性采样函数的动态卷积神经网络。这个网络所生成的特征图能够识别不同远近大小的词语关系。在没有借助 parser 和其它资源所提供的外部信息的情况下，这个网络在问题和语义分类问题上取得了很好的性能。

鸣谢

我们感谢 Nando de Freitas 和 Yee Whye Teh 在这片论文上得讨论。这项工作被 Xerox Foundation Award 所资助，EPSRC 码为 EP/F042728/1，EPSRC 码为 EP/K036580/1。

参考文献

- [1] Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In EMNLP, pages 1183–1193. ACL.
- [2] Phil Blunsom, Krystle Kocik, and James R. Curran. 2006. Question classification with log-linear models. In SIGIR ’06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pages 615–616, New York, NY, USA. ACM.
- [3] Daoud Clarke. 2012. A context-theoretic framework for compositionality in distributional semantics. *Computational Linguistics*, 38(1):41–71.
- [4] Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. 2010. Mathematical Foundations for a Compositional Distributional Model of Meaning. March.
- [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] John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159, July.
- [7] Katrin Erk and Sebastian Pado. 2008. A structured vector space model for word meaning in context. Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP 08, (October): 897.
- [8] Katrin Erk. 2012. Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10): 635–653.
- [9] Felix A. Gers and Jrgen Schmidhuber. 2001. Lstm recurrent networks learn simple context-free and context-sensitive languages. *IEEE Transactions on Neural Networks*, 12(6): 1333–1340.
- [10] Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *Processing*, pages 1–6.
- [11] Edward Grefenstette and Mehrnoosh Sadrzadeh. 2011. Experimental support for a categorical compositional distributional model of meaning. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 1394–1404. Association for Computational Linguistics.

- [12] Edward Grefenstette. 2013. Category-theoretic quantitative compositional distributional models of natural language semantics. arXiv preprint arXiv:1311.1539.
- [13] Emiliano Guevara. 2010. Modelling Adjective-Noun Compositionality by Regression. ESSLLI'10 Workshop on Compositionality and Distributional Semantic Models.
- [14] Karl Moritz Hermann and Phil Blunsom. 2013. The Role of Syntax in Vector Space Models of Compositional Semantics. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Sofia, Bulgaria, August. Association for Computational Linguistics. Forthcoming.
- [15] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. CoRR, abs/1207.0580.
- [16] Geoffrey E. Hinton. 1989. Connectionist learning procedures. *Artif. Intell.*, 40(1-3):185–234.
- [17] Zhiheng Huang, Marcus Thint, and Zengchang Qin. 2008. Question classification using head words and their hypernyms. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '08, pages 927–936, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [18] Nal Kalchbrenner and Phil Blunsom. 2013a. Recurrent continuous translation models. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, October. Association for Computational Linguistics.
- [19] Nal Kalchbrenner and Phil Blunsom. 2013b. Recurrent Convolutional Neural Networks for Discourse Compositionality. In Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality, Sofia, Bulgaria, August. Association for Computational Linguistics.
- [20] Dimitri Kartsaklis and Mehrnoosh Sadrzadeh. 2013. Prior disambiguation of word tensors for constructing sentence vectors. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP), Seattle, USA, October.

- [21] Andreas Kuchler and Christoph Goller. 1996. Inductive learning in symbolic domains using structure-driven recurrent neural networks. In Gunther Gorz and Steffen Hölldobler, editors, KI, volume 1137 of Lecture Notes in Computer Science, pages 183–197. Springer.
- [22] Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11): 2278–2324, November.
- [23] Xin Li and Dan Roth. 2002. Learning question classifiers. In Proceedings of the 19th international conference on Computational linguistics-Volume 1, pages 1–7. Association for Computational Linguistics.
- [24] Tomas Mikolov and Geoffrey Zweig. 2012. Context dependent recurrent neural network language model. In SLT, pages 234–239.
- [25] Tomas Mikolov, Stefan Kombrink, Lukas Burget, Jan Cernocky , and Sanjeev Khudanpur. 2011. Extensions of recurrent neural network language model. In ICASSP, pages 5528–5531. IEEE.
- [26] Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In Proceedings of ACL, volume 8.
- [27] Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. Cognitive Science, 34(8): 1388–1429.
- [28] Jordan B. Pollack. 1990. Recursive distributed representations. Artificial Intelligence, 46:77–105.
- [29] Holger Schwenk. 2012. Continuous space translation models for phrase-based statistical machine translation. In COLING (Posters), pages 1071–1080.
- [30] Joo Silva, Lusa Coheur, AnaCristina Mendes, and An- dreas Wichert. 2011. From symbolic to sub- symbolic information in question classification. Artificial Intelligence Review, 35(2): 137–154.
- [31] Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. 2011. Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP).
- [32] Richard Socher, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. 2013a. Grounded Compositional Semantics for Finding and Describing Images with

Sentences. In Transactions of the Association for Computational Linguistics (TACL).

- [33] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013b. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1631–1642, Stroudsburg, PA, October. Association for Computational Linguistics.
- [34] Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 384–394. Association for Computational Linguistics.
- [35] Peter Turney. 2012. Domain and function: A dual- space model of semantic relations and compositions. *J. Artif. Intell. Res. (JAIR)*, 44:533–585.
- [36] 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.
- [37] Fabio Massimo Zanzotto, Ioannis Korkontzelos, Francesca Fallucchi, and Suresh Manandhar. 2010. Estimating linear models for compositional distributional semantics. In Proceedings of the 23rd International Conference on Computational Linguistics, pages 1263–1271. Association for Computational Linguistics.
- [38] Luke S. Zettlemoyer and Michael Collins. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorical grammars. In UAI, pages 658–666. AUAI Press.

原文索引：

Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences [J]. arXiv preprint arXiv:1404.2188, 2014.