**Deep Learning Based Text Classification A Comprehensive Review**

摘要。

基于深度学习的文本分类模型在情感分析、新闻分类、问答和自然语言推理等多种文本分类任务中已经超越了经典的基于机器学习的方法。在这项工作中，我们详细回顾了近年来开发的150多个基于深度学习的文本分类模型，并讨论了它们的技术贡献、相似性和优势。我们还总结了40多个广泛用于文本分类的流行数据集。最后，我们对不同深度学习模式在流行基准上的表现进行量化分析，并探讨未来的研究方向。

附加关键词和短语：文本分类、情感分析、问答、新闻分类、深度学习、自然语言推理、主题分类。

**1.引言**

文本分类，也称为文本分类，是自然语言处理（NLP）中的一个经典问题，其目的是为句子、查询、段落和文档等文本单位指定标签或标记。它有着广泛的应用，包括问答、垃圾邮件检测、情感分析、新闻分类、用户意图分类、内容调节等。文本数据可以来自不同的来源，例如web数据、电子邮件、聊天、社交媒体、门票、保险索赔、用户评论、客户服务的问题和答案，等等。文本是一种极其丰富的信息源，但由于其非结构化的特性，从文本中提取洞察信息可能具有挑战性和耗时性。

文本分类可以通过手动注释或自动标记来执行。随着文本数据在工业应用中的规模不断扩大，文本自动分类变得越来越重要。自动文本分类的方法可以分为三类：

•基于规则的方法

•基于机器学习（数据驱动）的方法

•混合方法

基于规则的方法使用一组预定义的规则将文本分类为不同的类别。例如，任何带有“football”、“basketball”或“basketball”字样的文档都被指定为“sport”标签。这些方法需要对领域有深入的了解，而且系统很难维护。另一方面，基于机器学习的方法学习根据过去对数据的观察进行分类。一种机器学习算法利用预先标记的例子作为训练数据，学习文本片段与其标签之间的内在联系。因此，基于机器学习的方法可以检测数据中隐藏的模式，具有更高的可扩展性，并且可以应用于各种任务。这与基于规则的方法不同，后者需要为不同的任务设置不同的规则集。顾名思义，混合方法使用基于规则和机器学习方法的组合来进行预测。

近年来，机器学习模型引起了人们的广泛关注。大多数经典的基于机器学习的模型遵循流行的两步过程，第一步从文档（或任何其他文本单元）中提取一些手工制作的特征，第二步将这些特征输入分类器进行预测。一些流行的手工制作功能包括单词包（BoW）及其扩展。常用的分类算法有朴素贝叶斯、支持向量机、隐马尔可夫模型、梯度提升树和随机森林。两步方法有几个局限性。例如，依赖手工制作的特性需要繁琐的特性工程和分析才能获得良好的性能。此外，特征设计对领域知识的依赖性强，使得该方法很难推广到新的任务中。最后，这些模型不能充分利用大量的训练数据，因为特征（或特征模板）是预先定义的。

2012年开始了一个范式转变，当时一个基于深度学习的模型AlexNet[1]以很大的优势赢得了ImageNet的竞争。自那时起，深度学习模型被应用于计算机视觉和自然语言处理领域的一系列任务，提高了最新水平[2-5]。这些模型试图学习特征表示并以端到端的方式执行分类（或回归）。它们不仅能够发现数据中隐藏的模式，而且可以从一个应用程序转移到另一个应用程序。毫不奇怪，这些模型正成为近年来各种文本分类任务的主流框架。

在这项调查中，我们回顾了150多个针对不同文本分类任务开发的深度学习模型，包括情感分析、新闻分类、主题分类、问答和自然语言推理，在过去的六年里几年了。我们根据它们的神经网络结构，将它们分为几个类别，例如基于递归神经网络（RNNs）、卷积神经网络（CNNs）、注意力、变压器、胶囊网络等的模型。本文的贡献概括如下：

•我们详细介绍了150多种针对文本分类的深度学习模型。

•我们回顾了40多个流行的文本分类数据集。

•我们在16个流行基准上对一组选定的深度学习模型的绩效进行定量分析。

•我们讨论剩余的挑战和未来的方向。

**1.1文本分类任务**

本节简要介绍了本文讨论的不同文本分类任务：情感分析、新闻分类、主题分类、问答和自然语言推理。

情绪分析。情感分析是文本分类的一个热门分支，它旨在分析人们在文本数据（如产品评论、电影评论和推特）中的观点，并提取他们的极性和观点。情感分类可以是一个二值问题，也可以是一个多类问题。二元情感分析是将文本分为正类和负类，而多类情感分析则侧重于将数据分为细粒度的标签或多层次的强度。

新闻分类。新闻内容是影响人们生活的重要信息来源之一。新闻分类系统可以帮助用户实时获取感兴趣的信息。基于用户兴趣的新闻主题识别和相关新闻推荐是新闻分类的两个主要应用。

主题分析。主题分析试图通过识别文本的主题从文本中自动获得意义。主题分类是主题分析的重要组成技术之一。主题分类的目标是为每个文档分配一个或多个主题，以便于分析。

问答（QA）。有两种类型的QA系统：抽取式和生成式。抽取式QA可以看作是文本分类的一个特例。给定一个问题和一组候选答案（例如班[6]中给定文档的文本跨度），我们需要将每个候选答案分类为正确与否。生成型QA学习从零开始生成答案（例如使用序列到序列模型）。除非另有说明，否则本文讨论的QA任务是抽取式QA。

自然语言推理。NLI，也称为识别文本蕴涵（RTE），预测一个文本的含义是否可以从另一个文本中推断出来。特别是，一个系统需要给每对文本单元分配一个标签，比如蕴涵、矛盾和中立[7]。释义是NLI的一种普遍形式，也称为文本对比较。任务是测量句子对的语义相似度，以确定一个句子是否是另一个句子的释义。

1.2论文结构

论文的其余部分结构如下：第2节全面介绍了150多种基于深度学习的文本分类模型。第3节回顾了一些最流行的文本分类数据集。第4节对16个基准上的一组选定的深度学习模型进行了定量绩效分析。第五部分讨论了基于深度学习的文本分类方法面临的主要挑战和未来的发展方向。第六节对论文进行总结。附录A概述了一些常用于文本分类的神经网络模型体系结构。

**2.文本分类的深度学习模型**

•基于前馈网络的模型，将文本视为一袋文字（第2.1节）。

•基于RNN的模型，将文本视为一个单词序列，旨在捕获单词相关性和文本结构（第2.2节）。

•基于CNN的模型，训练识别文本中的模式，如关键短语，以便分类（第2.3节）。

•胶囊网络，它解决了CNN汇集操作所带来的信息丢失问题，最近被应用于文本分类（第2.4节）。

•注意机制，有效识别文本中的相关词汇，并已成为开发深度学习模型的有用工具（第2.5节）。

•记忆增强网络，它将神经网络与外部记忆结合在一起，模型可以读取和写入（第2.6节）。

•变压器，它允许比RNN更多的并行化，使得使用GPU集群高效（预）训练非常大的语言模型成为可能（第2.7节）。

•图形神经网络，用于捕捉自然语言的内部图形结构，如句法和语义分析树（第2.8节）。

•孪生神经网络，专为文本匹配而设计，是文本分类的特例（第2.9节）。

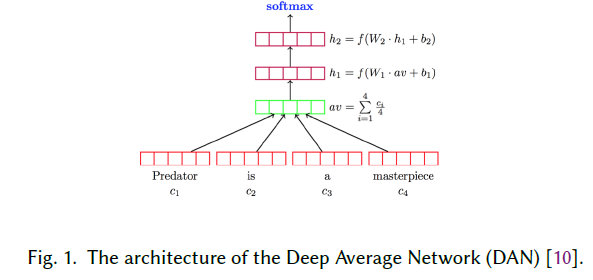
•混合模型，将注意力、RNN、CNN等结合起来，捕捉句子和文档的本地和全局特征（第2.10节）。

•最后，在第2.11节中，我们回顾了超出监督学习的建模技术，包括使用自动编码器和对抗性训练的无监督学习，以及强化学习。

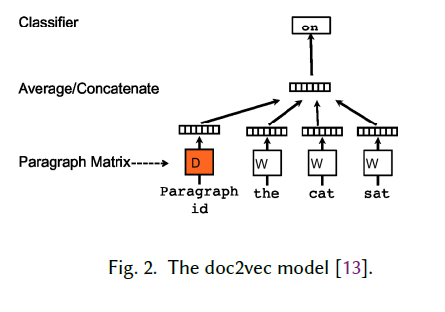
读者应合理熟悉基本的深度学习模式，以理解本节内容。关于基本的深度学习架构和模型的更多细节，我们请读者参考Goodfellow等人的深度学习教科书。[140]或本文附录。

**2.1前馈神经网络**

前馈网络是最简单的文本表示深度学习模型之一。然而，它们在许多文本分类基准上都达到了很高的精度。这些模型将文本视为一个单词包。对于每个单词，他们使用嵌入模型（如word2vec[8]或Glove[9]）学习向量表示，将嵌入的向量和或平均值作为文本的表示，将其通过一个或多个前馈层（称为多层感知器（MLPs））传递，然后使用logistic回归、朴素贝叶斯或支持向量机等分类器对最终层的表示进行分类[10]。这些模型的一个例子是深平均网络（DAN）[10]，其架构如图1所示。尽管它很简单，但DAN优于其他更复杂的模型，这些模型旨在明确地学习文本的组成性。例如，DAN在句法方差较高的数据集上优于句法模型。Joulin等人。[11] 提出了一种简单有效的文本分类器fastText。与DAN一样，fastText将文本视为一个单词包。与DAN不同，fastText使用一包n-grams作为附加功能来捕获本地单词顺序信息。这在实践中证明是非常有效的，同时实现了与显式使用词序的方法类似的结果[12]。



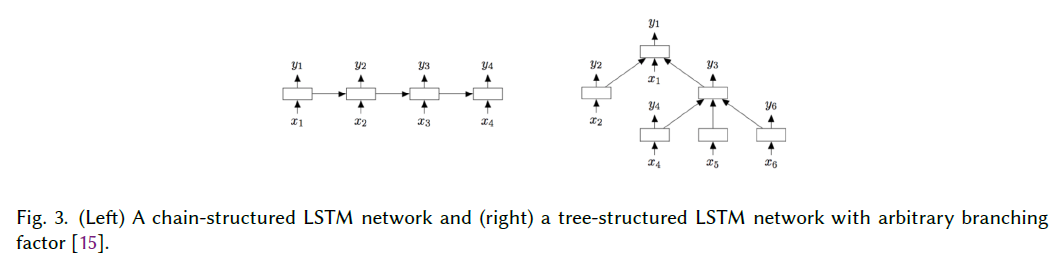
Le和Mikolov[13]提出了doc2vec，它使用无监督算法来学习可变长度文本片段（如句子、段落和文档）的定长特征表示。如图2所示，doc2vec的架构类似于连续字袋（CBOW）模型的架构[8，14]。唯一的区别是通过矩阵D映射到段落向量的附加段落标记。在doc2vec中，该向量与三个单词的上下文的连接或平均值用于预测第四个单词。段落向量表示当前上下文中缺少的信息，可以作为段落主题的记忆。经过训练后，段落向量被用作段落的特征，然后输入分类器进行预测。Doc2vec发表时，在多个文本分类和情感分析任务上取得了最新的研究成果。



**2.2 RNN-Based Models**

基于RNN的模型将文本视为一个单词序列，旨在捕获用于文本分类的单词依赖项和文本结构。然而，普通的RNN模型不能很好地工作，并且常常表现得不如前馈神经网络。在RNN的众多变体中，长-短期内存（LSTM）是最流行的体系结构，其设计目的是更好地捕获长期依赖关系。LSTM通过引入一个存储单元来记忆任意时间间隔内的值，并引入三个门（输入门、输出门、遗忘门）来调节进出单元的信息流，解决了普通RNNs所面临的梯度消失或爆炸问题。通过获取更丰富的信息，如自然语言的树状结构、文本中的大跨度词关系、文档主题等，对RNNs和LSTM模型进行了改进。

Tai等人。[15] 开发了树型LSTM模型，将LSTM推广到树型结构的网络类型，学习丰富的语义表示。作者认为，对于NLP任务，树型LSTM比链式LSTM是一个更好的模型，因为自然语言具有自然地将单词组合成短语的句法特性。验证了Tree-LSTM在情感分类和语义相关性预测两个任务中的有效性。这些模型的架构如图3所示。朱等人[16] 还将链式结构的LSTM扩展到树结构，使用内存单元在递归过程中存储多个子单元或多个子单元的历史。他们认为，新模型提供了一种考虑层次结构（如语言或图像解析结构）上的远程交互的原则方法。



为机器阅读建立大跨度单词关系模型，Cheng等人。[17] 用一个内存网络代替单个内存单元来扩充LSTM体系结构。这使得在重复期间能够自适应地使用内存带有神经注意，提供一种弱诱导标记之间关系的方法。该模型在语言建模、情感分析、自然语言识别等方面取得了很好的效果。

多时间尺度LSTM（MT-LSTM）神经网络[18]也被设计用于通过捕获不同时间尺度的有价值信息来对长文本（如句子和文档）进行建模。MT-LSTM将标准LSTM模型的隐藏状态划分为若干组。每个组在不同的时间段被激活和更新。因此，MT-LSTM可以对非常长的文档建模。据报道，在文本分类方面，MT-LSTM优于一组基线，包括基于LSTM和RNN的模型。

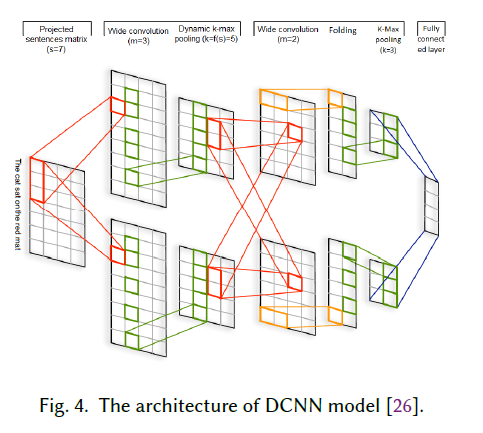
RNNs擅长于捕获单词序列的局部结构，但在记住长距离依赖关系时遇到困难。相比之下，潜在主题模型能够捕获文档的全局语义结构，但不考虑单词排序。Bieng等人。[19] 提出了一种TopicRNN模型，综合了RNNs和潜在主题模型的优点。它使用RNNs捕获本地（语法）依赖，使用潜在主题捕获全局（语义）依赖。据报道，TopicRNN在情绪分析方面优于RNN基线。

还有其他有趣的基于RNN的模型。Liu等人。[20] 使用多任务学习来训练RNN，以利用来自多个相关任务的标记训练数据。Johnson和Rie[21]探索了一种基于LSTM的文本区域嵌入方法。Zhou等人。[22]将双向LSTM（bilstm）模型与二维max池集成，以捕获文本特征。Wang等人。[23]提出了“匹配聚合”框架下的双边多视角匹配模型。Wan等人。[24]使用双向LSMT模型生成的多个位置句子表示来探索语义匹配。

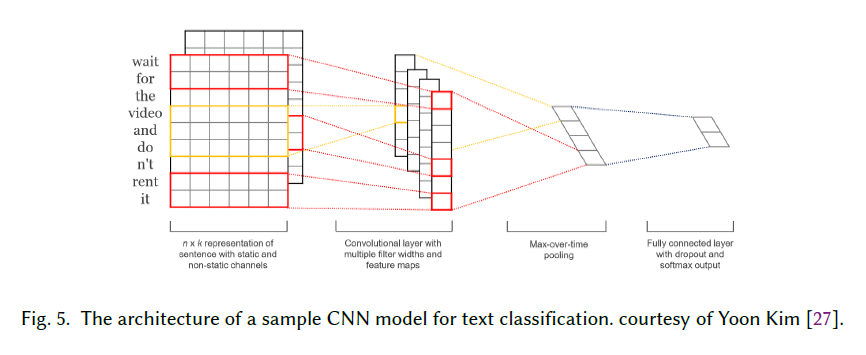
**2.3 CNN-Based Models**

RNN被训练成跨时间识别模式，而CNN则学习跨空间识别模式[25]。RNNs在需要理解远程语义的NLP任务（如POS标记或QA）中工作良好，而CNNs在检测局部和位置不变模式非常重要的情况下工作良好。这些模式可能是表达“我喜欢”或“濒危物种”等特定情感的关键短语。因此，CNNs已经成为最流行的文本分类模型体系结构之一。

Kalchbrenner等人提出了一种基于CNN的文本分类模型。[26]。该模型采用动态k-max池，称为动态CNN（DCNN）。如图4所示，DCNN的第一层使用对句子中的每个单词的嵌入来构造句子矩阵。然后，利用动态k-max池给出的动态池层与宽卷积层交替的卷积结构，在句子上生成一个能够显式捕获单词和短语的短距离和长距离关系的特征映射。池参数k可以根据句子大小和卷积层次中的级别动态选择。

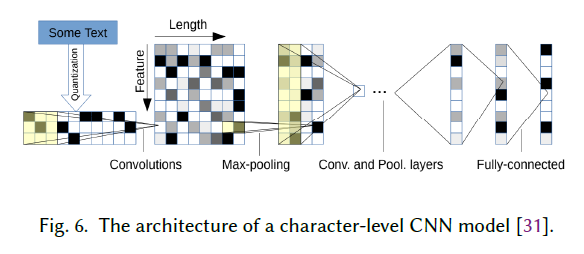


后来，Kim[27]提出了一个比DCNN更简单的基于CNN的文本分类模型。如图5所示，Kim的模型仅在从无监督神经语言模型即word2vec获得的词向量上使用一层卷积。Kim还比较了四种不同的单词嵌入学习方法：（1）CNN-rand，其中所有的单词嵌入都是随机初始化的，然后在训练期间进行修改；（2）CNN-static，其中使用预先训练好的word2vec嵌入，并在模型训练期间保持固定；（3）CNN-non-static，其中word2vec嵌入在每个任务的训练；和（4）CNNmultichannel，其中使用了两组单词嵌入向量，它们都使用word2vec初始化，其中一个在模型训练期间更新，而另一个是固定的。据报道，这些基于CNN的模型改进了情感分析和问题分类的最新技术。



我们一直在努力改进基于CNN的模型的体系结构[26，27]。Liu等人。[28]提出了一个新的基于CNN的模型，对Kim-CNN的架构进行了两次修改[27]。首先，采用动态最大池方案从文档的不同区域捕获更细粒度的特性。其次，在池和输出层之间插入一个隐藏的瓶颈层来学习压缩文档用于减小模型大小和提高模型性能的表示。在文献[29,30]中，作者没有使用预先训练好的低维词向量作为CNNs的输入，而是直接将CNNs应用到高维文本数据中，学习小文本区域的嵌入以进行分类。

字符级cnn也被用于文本分类[31，32]。第一个这样的模型是由Zhang等人提出的。[31]。如图6所示，该模型以固定大小、编码为一个热向量的字符作为输入，将其通过由具有池操作的六个卷积层和三个完全连接层组成的深CNN模型。Prusa等人。[33]提出了一种使用CNNs对文本进行编码的方法，大大减少了学习字符级文本表示所需的内存消耗和训练时间。这种方法可以很好地适应字母表的大小，允许从原始文本中保留更多的信息以提高分类性能。



研究了word嵌入和CNN结构对模型性能的影响。受到VGG[34]和ResNets[35]的启发，Conneau等人。[36]提出了一个非常深入的文本处理CNN（VDCNN）模型。它直接在字符级操作，只使用小卷积和池操作。研究表明，VDCNN的性能随着深度的增加而提高。Duque等人。[37]修改VDCNN的结构以适应移动平台的限制并保持性能。他们能够将模型尺寸压缩10到20倍，精度损失在0.4%到1.3%之间。Le等人。[38]表明，当文本输入被表示为字符序列时，深度模型确实优于浅层模型。然而，一个简单的浅层和宽层网络在单词输入方面优于deep模型，如DenseNet[39]。郭等人。[40]研究了单词嵌入的影响，并提出通过多通道CNN模型使用加权单词嵌入。Zhang等人。[41]研究了不同的单词嵌入方法和池机制的影响，发现使用非静态word2vec和GloVe优于一个热向量，而max pooling始终优于其他池方法。

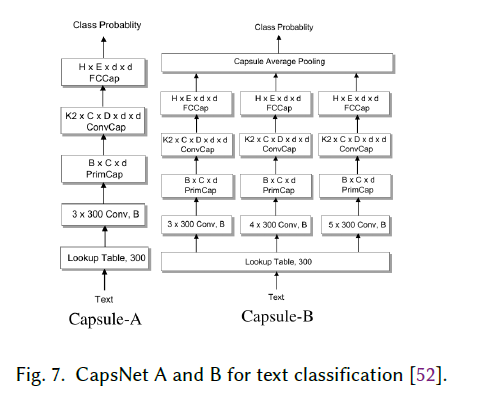
还有其他有趣的基于CNN的模型。Mou等人。[42]提出了一个基于树的CNN来捕获句子级语义。庞等人。[43]将文本匹配作为图像识别任务，利用多层CNNs来识别显著的n-gram模式。Wang等人。[44]提出了一个基于CNN的模型，该模型将短文本的显式和隐式表示结合起来进行分类。将CNNs应用于生物医学文本分类的兴趣也越来越大[45-48]。

**2.4 Capsule Neural Networks**

CNNs通过使用连续的卷积层和池对图像或文本进行分类。尽管池操作识别显著特征并降低卷积操作的计算复杂度，但它们会丢失有关空间关系的信息，并且可能基于实体的方向或比例对实体进行错误分类。

为了解决池的问题，杰弗里辛顿提出了一种新的方法，称为胶囊网络（CapsNets）[49，50]。胶囊是一组神经元，其活动向量代表特定类型实体（如对象或对象部分）的不同属性。向量的长度表示实体存在的概率，向量的方向表示实体的属性。与CNN的最大池（选择一些信息并丢弃其余信息）不同，胶囊“路由”下层的每个胶囊到上层的最佳父胶囊，使用网络中的所有可用信息到最后一层进行分类。路由可以使用不同的算法来实现，例如通过协议的动态路由[50]或EM算法[51]。

最近，胶囊网络被应用于文本分类，其中胶囊被用来表示一个句子或文档作为一个矢量。[52-54]提出了一个基于CapsNets变体的文本分类模型。该模型由四层组成：（1）n-gram卷积层，（2）胶囊层，（3）卷积胶囊层，（4）全连通胶囊层。作者尝试了三种策略来稳定动态路由过程，以减轻包含背景信息（如停止词或与任何文档类别无关的词）的噪声胶囊的干扰。他们还探索了两种胶囊结构，如图7所示，分别表示为胶囊-A和胶囊-B。胶囊-A类似于[50]中的CapsNet。Capsule-B在n-gram卷积层使用三个带不同窗口大小滤波器的并行网络来学习更全面的文本表示。CapsNet-B在实验中表现较好。



Kim等人提出的基于CapsNet的模型。[55]使用了类似的架构。该模型包括：（1）以文档为嵌入序列的输入层；（2）生成特征映射并使用选通线性单元保留空间信息的卷积层；（3）卷积胶囊层，通过对卷积层检测到的局部特征进行聚合形成全局特征；以及（4）文本胶囊层预测类标签。作者观察到，物体在文本中的组装比在图像中的组装更为自由。例如，文档的语义可以保持不变，即使某些句子的顺序发生了变化，这与眼睛和鼻子在人脸上的位置不同。因此，它们使用静态路由模式，在文本分类方面始终优于动态路由[50]。Aly等人。[56]建议使用CapsNet进行分层多标签分类（HMC），认为CapsNet编码子-父关系的能力使得它比传统方法更好地解决HMC任务，在HMC任务中，文档被分配一个或多个分层结构中组织的类标签。他们的模型结构与[52，53，55]中的相似。

Ren等人。[57]提出了CapsNets的另一种变体，使用胶囊之间的组合编码机制和基于k-均值聚类的新路由算法。首先，单词嵌入是使用码本中的所有码字向量形成的。然后通过k-means路由将低级胶囊捕获的特征聚合到高级胶囊中。

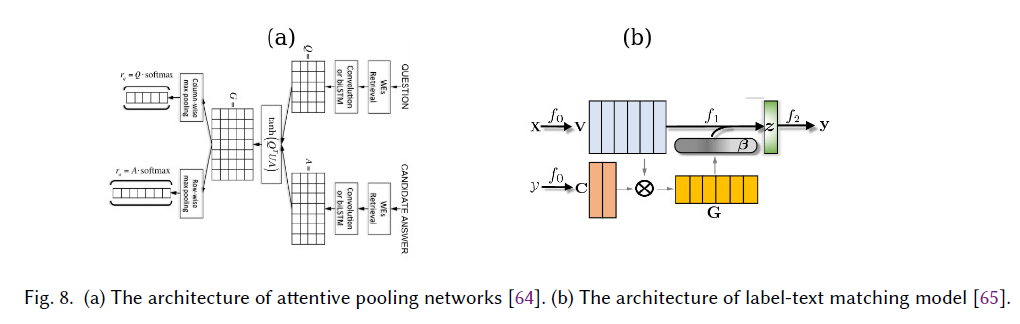
**2.5 Models with Attention Mechanism**

注意力是由我们如何将视觉注意力集中在一个图像的不同区域或一个句子中的相关词引起的。注意力成为一个越来越流行的概念和有用的工具，为NLP开发深度学习模型[58，59]。简言之，语言模型中的注意力可以解释为一个重要权重向量。为了预测句子中的一个词，我们使用注意向量来估计它与其他词或“注意”有多强的相关性，并将其加权值之和作为目标的近似值。

Yang等人。[60]提出了一种用于文本分类的分层注意网络。该模型有两个显著的特点：（1）反映文档层次结构的层次结构；（2）在单词和句子层次上应用两个层次的注意机制，使其能够在构建文档表示时区别地关注更多和更少的重要内容。该模型在6个文本分类任务上都有显著的优势。Zhou等人。[61]将分层注意模型扩展到跨语言情感分类。在每种语言中，LSTM网络被用来对文档进行建模。然后，利用层次注意机制进行分类，其中句子级注意模型学习文档中哪些句子对于确定整体情感更为重要。而单词级注意模型则学习每个句子中哪些单词是决定性的。

Shen等人。[62]提出了一个用于RNN/CNN自由语言理解的定向自注意网络，其中输入序列中元素之间的注意是定向的和多维的。使用一个轻量级的神经网络来学习句子嵌入，仅仅基于所提出的注意，而不需要任何RNN/CNN结构。Liu等人。[63]提出了一个对NLI有内在关注的LSTM模型。这个模型使用两个阶段的过程来编码一个句子。首先，在单词级Bi LSTM上使用平均池来生成第一阶段句子表示。其次，采用注意机制代替同一句子的平均集合，以获得更好的表达。句子的第一阶段表示法是用来处理出现在句子中的词。

注意模型也广泛应用于成对排序或匹配任务。Santos等人。[64]提出了一种双向注意机制，称为注意池（AP），用于成对排序。AP使池层能够意识到当前的输入对（例如，问答对），这样来自两个输入项的信息可以直接影响彼此表示的计算。除了学习输入对的表示外，AP还共同学习对投影片段的相似性度量，然后为每个输入导出相应的注意向量，以指导池。AP是独立于底层表示学习的通用框架，并且可以应用于CNNs和RNNs，如图8（a）所示。Wang等人。[65]将文本分类看作是一个标签词匹配问题：每个标签都嵌入到与词向量相同的空间中。作者介绍了一个注意框架，该框架通过余弦相似性度量文本序列和标签之间嵌入的兼容性，如图8（b）所示。



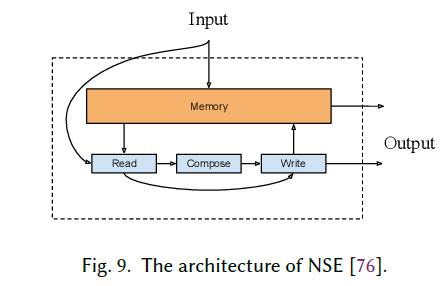
Kim等人。[66]提出了一种使用紧密连接的递归和共注意网络的语义句子匹配方法。与DenseNet[39]类似，该模型的每一层都使用注意特征的串联信息以及所有先前递归层的隐藏特征。它能够保持从最底层的单词嵌入层到最上层的递归层的原始和共同注意的特征信息。尹等人。[67]提出了另一种基于注意的句子对匹配CNN模型。他们研究了三种注意模式，将句子之间的相互影响整合到cnn中，以便每个句子的表示都考虑到它的成对句子。这些相互依赖的句子对表示比孤立的句子表示更强大，这在多个分类任务（包括答案选择、释义识别和文本蕴涵）中得到了验证。Tan等人。[68]在匹配聚合框架下，使用多个注意函数来匹配句子对。Yang等人。[69]提出了一种基于注意的神经匹配模型，用于对短答案文本进行排序。他们采用值共享加权方案代替位置共享加权方案来组合不同的匹配信号，并结合问题注意网络进行问题项重要度学习。该模型在TREC QA数据集上取得了很好的效果。

还有其他有趣的注意力模式。Lin等人。[70]利用自我注意提取可解释的句子嵌入。Wang等人。[71]提出了一种具有多尺度特征注意的密集连接CNN，以产生可变的n-gram特征。Yamada和Shindo[72]使用神经注意包实体模型，使用知识库中的实体执行文本分类。Parikh等人。[73]使用注意将问题分解为可以单独解决的子问题。Chen等人。[74]探索了增强句子嵌入的广义池方法，提出了一种基于向量的多头部注意模型。Liu和Lane[75]提出了一种基于注意的RNN模型，用于联合意图检测和时隙填充。

**2.6 Memory-Augmented Networks**

虽然注意模型在编码过程中所存储的隐藏向量可以看作是模型内部存储器的条目，但是记忆增强网络将神经网络与外部存储器结合起来，模型可以读取和写入外部存储器。

Munkhdalai和Yu[76]提出了一种记忆增强神经网络，称为神经语义编码器（NSE），用于文本分类和质量保证。NSE配备有一个可变大小的编码存储器，该存储器随时间演变，并通过读取、合成和写入操作保持对输入序列的理解，如图9所示。



韦斯顿等人。[77]为一个综合的QA任务设计了一个内存网络，其中一系列的语句（内存条目）作为问题的支持事实提供给模型。该模型根据问题和以前检索到的内存，一次从内存中检索一个条目。Sukhbatar等人。[78]扩展了这项工作，提出了端到端的记忆网络，其中的记忆条目通过注意机制以软方式检索，从而实现端到端的训练。他们表明，通过多轮（hops），该模型能够检索和推理多个支持事实来回答特定问题。

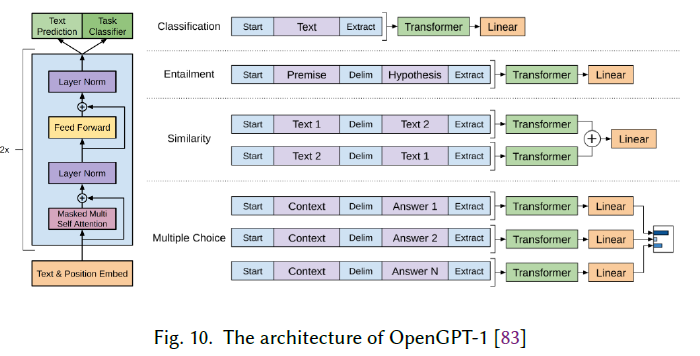
Kumar等人。[79]提出了一种动态记忆方法（DMN），它处理输入序列和问题，形成情景记忆，并生成相关答案。问题触发一个迭代注意过程，它允许模型将其注意条件设置为前一次迭代的输入和结果。然后，这些结果在层次递归序列模型中进行推理以生成答案。DMN是端到端训练，并获得最新的QA和POS标记结果。熊等。[80]对DMN进行了详细的分析，改进了DMN的存储和输入模块。

**2.7 Transformers**

RNNs所面临的计算瓶颈之一是文本的顺序处理。尽管CNNs比RNNs的顺序性差，但是捕捉句子中单词之间关系的计算成本也随着句子长度的增加而增加，这与RNNs类似。变形金刚[2]克服了这一局限性，将自我注意应用于句子中每个单词的并行计算，或记录一个“注意分数”来模拟每个词对另一个词的影响。由于这个特性，变压器允许比CNNs和RNNs更多的并行化，这使得在GPU集群上有效地训练大量数据的非常大的模型成为可能。

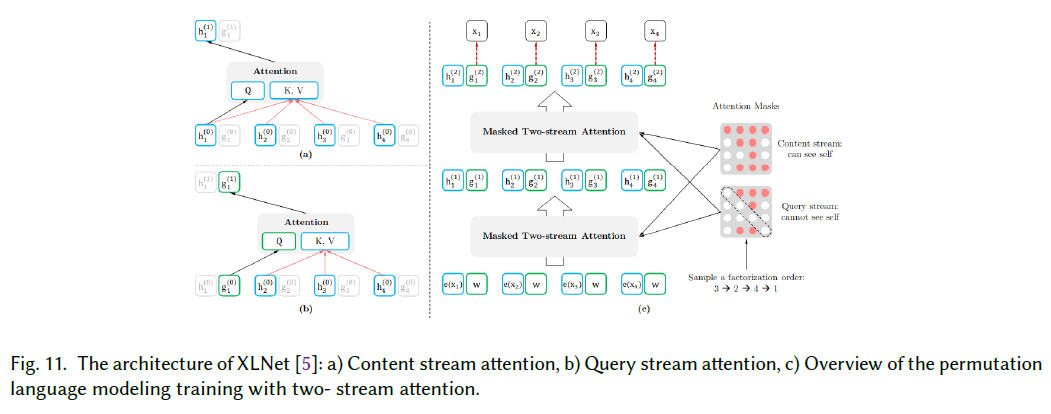
自2018年以来，我们看到了一组基于变压器的大规模预训练语言模型（plm）的兴起。与早期基于CNNs[81]或LSTMs[82]的上下文嵌入模型相比，基于Transformer的plm使用更深层次的网络架构（例如，48层Transformers[83]），并在更大量的文本语料库上进行预训练，通过预测基于上下文的单词来学习上下文文本表示。这些plm使用特定于任务的标签进行了微调，并在许多下游NLP任务（包括文本分类）中创建了新的最新技术。尽管预训练是无监督的，但微调是有监督的学习。

plm可分为两类：自回归plm和自编码plm。最早的自回归plm之一是OpenGPT[83，84]，这是一个单向模型，它从左到右（或从右到左）逐字预测文本序列，每个单词的预测取决于以前的预测。图10示出了OpenGPT的体系结构。它由12层变压器块组成，每个变压器块由一个屏蔽的多头注意模块、一个层归一化和一个位置前馈层组成。OpenGPT可以通过添加特定于任务的线性分类器和使用特定于任务的标签进行微调来适应下游任务，例如文本分类。

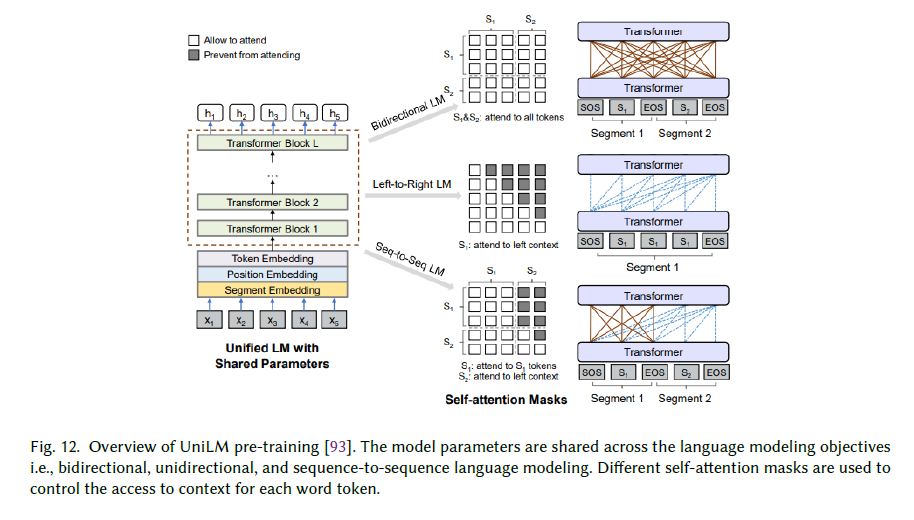


最广泛使用的自动编码plm之一是BERT[4]。与OpenGPT基于先前的预测预测单词不同，BERT使用屏蔽语言建模任务进行训练，该任务随机屏蔽文本序列中的一些标记，然后通过对双向转换器获得的编码向量进行调节，独立地恢复屏蔽标记。关于改进BERT的工作有很多。RoBERTa[85]比BERT更健壮，并且使用更多的训练数据进行训练。阿尔伯特[86]降低了记忆消耗，提高了伯特的训练速度。distributbert[87]在预训练期间利用知识蒸馏将BERT的大小减少40%，同时保留其99%的原始能力，并使推理速度加快60%。SpanBERT[88]扩展了BERT以更好地表示和预测文本跨度。BERT及其变体已经针对各种NLP任务进行了微调，包括QA[89]、文本分类[90]和NLI[91，92]。

有人试图将自回归和自编码plm的优点结合起来。XLNet[5]集成了OpenGPT等自回归模型和BERT的双向上下文建模思想。XLNet在预训练期间使用了一个置换操作，它允许上下文包括来自左和右的令牌，使其成为一个广义的顺序感知自回归语言模型。这种排列是通过在变压器中使用一个特殊的注意遮罩来实现的。XLNet还引入了一个双流自关注模式允许位置感知单词预测。这是由于观察到单词分布因单词位置的不同而变化很大。例如，句子的开头与句子中其他位置的分布有很大的不同。如图11所示，为了在置换3-2-4-1中预测位置1中的字令牌，通过包括所有先前字（3、2、4）的位置嵌入和令牌嵌入来形成内容流，然后通过包括要预测的字（位置1中的字）的内容流和位置嵌入来形成查询流，最后根据查询流中的信息进行预测。



如前所述，OpenGPT使用从左到右的转换器来学习自然语言生成的文本表示，而BERT使用双向转换器来理解自然语言。统一语言模型（UniLM）[93]旨在处理自然语言理解和生成任务。UniLM使用三种类型的语言建模任务进行预训练：单向、双向和序列到序列预测。如图12所示，通过使用共享变压器网络和利用特定的自我注意遮罩来控制预测条件所处的上下文来实现统一建模。据报道，UniLM的第二个版本[94]在广泛的自然语言理解和生成任务上达到了最新的水平，显著优于以前的plm，包括OpenGPT-2、XLNet、BERT及其变体。



Raffel等人。[95]提出了一个统一的基于转换器的框架，将许多NLP问题转换为文本到文本格式。他们还进行了一项系统研究，比较了数十项语言理解任务的训练前目标、体系结构、未标记数据集、微调方法和其他因素。

**2.8 Graph Neural Networks**

尽管自然语言文本具有顺序性，但它们也包含内部的图形结构，如句法和语义分析树，它们定义了句子中单词之间的句法/语义关系。

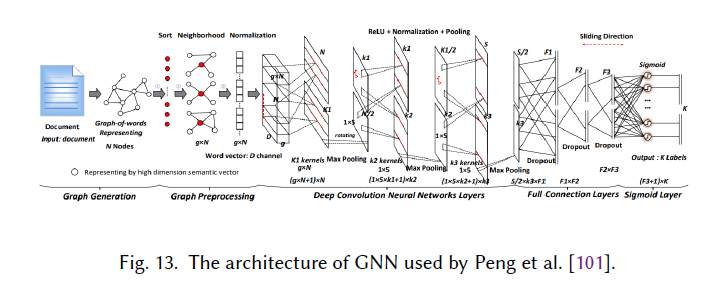
为NLP开发的最早的基于图的模型之一是TextRank[96]。作者提出将自然语言文本表示为图G（V，E），其中V表示一组节点，E表示节点间的一组边。

根据手头的应用程序，节点可以表示各种类型的文本单位，例如单词、搭配、整句话等。同样，边缘可以表示任何节点之间的不同类型的关系，例如词汇或语义关系、上下文重叠等。

现代图形神经网络（GNNs）是通过扩展对图形数据的深度学习方法而发展起来的，例如TextRank使用的文本图。深层神经网络，如CNN，RNN和自动编码，在过去的几年里，为了处理图形数据的复杂性，已经进行了推广[97]。例如，将用于图像处理的CNNs的2D卷积推广到通过取节点邻域信息的加权平均来执行图卷积。在各种类型的gnn中，卷积gnn，如图卷积网络（GCNs）[98]及其变体，是最受欢迎的gnn，因为它们与其他神经网络组合有效且方便，并且在许多应用中取得了最新的结果。GCNs是图上CNNs的一个有效变种。GCNs堆叠学习的一阶谱滤波器层，然后是一个非线性激活函数来学习图形表示。

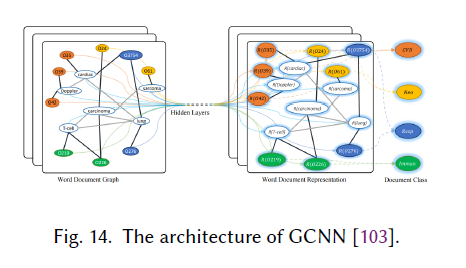
GNNs在NLP中的一个典型应用是文本分类。GNNs利用文档或单词之间的相互关系来推断文档标签[98–100]。在下文中，我们将回顾一些为文本分类而开发的gcn变体。

Peng等人。[101]提出了一种基于图CNN的深度学习模型，首先将文本转换成词的图，然后使用图卷积运算卷积词的图，如图13所示。实验表明，文本的词表示图具有捕获非连续语义和远距离语义的优势，CNN模型具有学习不同层次语义的优势。



在[102]中，Peng等人。提出了一种基于层次分类感知和注意图胶囊CNNs的文本分类模型。该模型的一个独特之处是使用了类标签之间的层次关系，在以前的方法中，这些关系被认为是独立的。具体来说，为了利用这种关系，作者开发了一种层次分类嵌入方法来学习它们的表示，并通过结合标签表示相似度定义了一种新的加权边缘损失。

Yao等人。[103]使用相似图CNN（GCNN）模型进行文本分类。他们基于单词共现和文档单词关系为一个语料库构建了一个文本图，然后学习了一个文本图卷积网络（text GCN），如图14所示。文本GCN用一个热表示初始化对于word和document，然后共同学习word和document的嵌入，由已知的文档类标签监督。

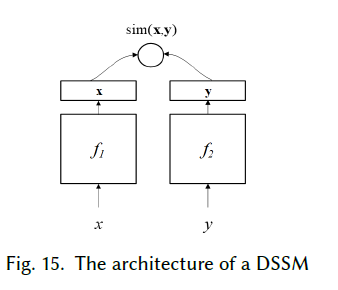


为大规模文本语料库构建GNNs代价高昂。通过降低模型复杂度或改变模型训练策略来降低建模成本的研究已经开始。前者的一个例子是[104]中提出的简单图卷积（SGC）模型，其中通过反复消除连续层之间的非线性并将得到的函数（权重矩阵）折叠成一个线性变换来简化深卷积GNN。后者的一个例子是文本级GNN[105]。文本级GNN不是为整个文本语料库构建一个图，而是为文本语料库上的滑动窗口定义的每个文本块生成一个图，从而减少训练过程中的内存消耗。同样的想法推动了GraphSage[99]的发展，GraphSage[99]是一种用于卷积GNNs的批训练算法。

**2.9 Siamese Neural Networks**

暹罗神经网络（S2Nets）[106，107]及其DNN变体，称为深结构语义模型（DSSMs）[108]，是为文本匹配而设计的。该任务是许多NLP应用程序的基础，例如QA中的querydocument排序和答案选择。这些任务可以看作是文本分类的特例。

例如，在有问题的文档排序中，我们希望将文档分类为与给定查询相关或无关的文档。如图15所示，DSSM（或S2Net）由DNNs、f1和f2对组成，DNNs、f1和f2将输入x和y映射到公共低维语义空间中的对应向量中。然后用两个向量的余弦距离来度量x和y的相似性。当S2Nets假设f1和f2共享相同的架构甚至相同的参数时，在DSSMs中，f1和f2可以根据x和y具有不同的架构。例如，为了计算图像-文本对的相似性，f1可以是深CNN，f2可以是RNN或MLP。根据（x，y）的定义，这些模型可以应用于各种NLP任务。例如，（x，y）可以是用于查询文档排名[108，109]的查询文档对，也可以是QA[110，111]中的问答对，等等。

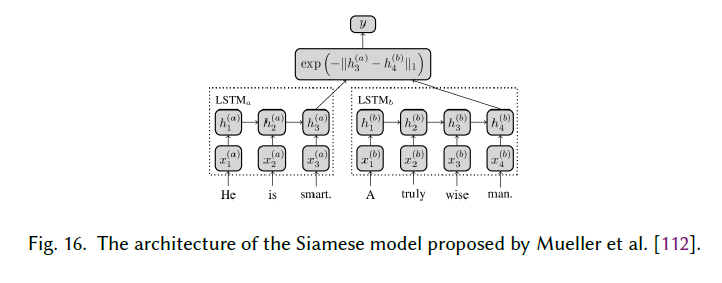


模型参数θ通常使用成对秩损失进行优化。以文档排名为例。考虑一个查询x和两个候选文档y+和y-，其中y+与x相关，y-与x无关。设simθ（x，y）为θ参数化语义空间中x和y的余弦相似性。培训目标是将基于利润的损失最小化为

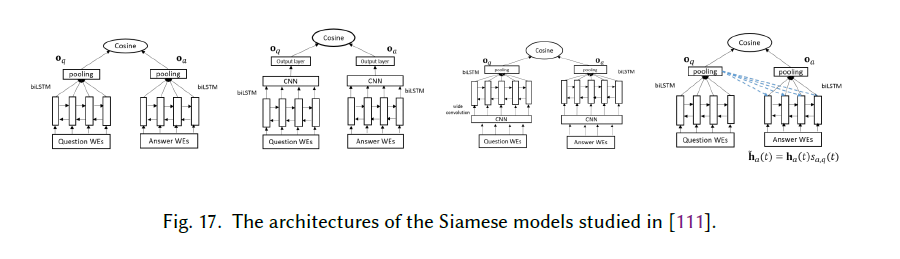


其中[x]+：=max（0，x）和γ是margin超参数

由于文本具有顺序性，因此使用RNNs或LSTMs来度量文本之间的语义相似性是很自然的。图16示出了Mueller等人提出的暹罗模型的架构。[112]，其中两个网络使用相同的LSTM模型。Neculoiu等人。[113]提出了一个相似的模型，该模型对f1和f2使用字符级Bi LSTMs，并使用余弦函数计算相似性。除RNNs外，s2net中还使用BOW模型和CNNs来表示句子。例如，他等。[114]提出了一种利用CNNs对多视角句子相似度进行建模的S2Net。Renter等人。[115]提出了一个连体CBOW模型，该模型通过平均句子的词嵌入来形成句子向量表示，并将句子相似度计算为句子向量之间的余弦相似度。随着BERT成为新的最先进的句子嵌入模型，有人试图构建基于BERT的s2net，如SBERT[116]和TwinBERT[117]。

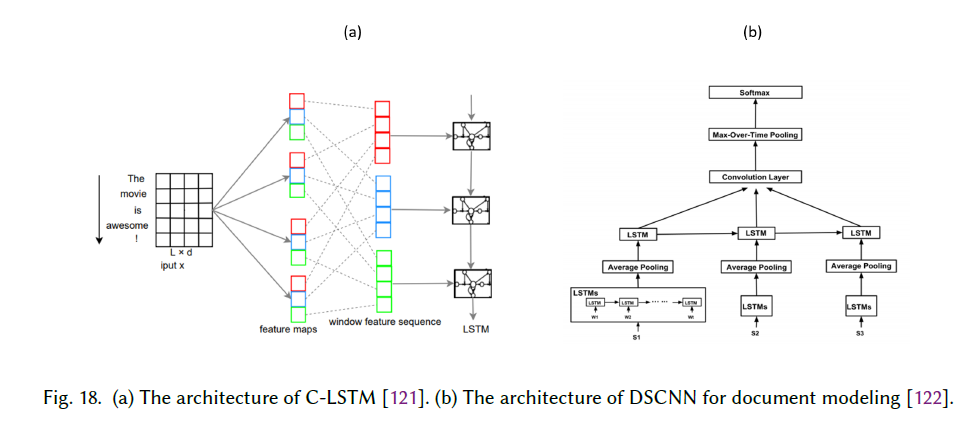


S2Nets和DSSMs在质量保证中得到了广泛的应用。Das等人。[110]提出了一个连体问题回答CNN（SCQA）来衡量问题及其（候选）答案之间的语义相似性。为了降低计算复杂度，SCQA使用问答对的字符级表示。训练SCQA的参数，使问题与其相关答案之间的语义相似度最大，如方程1，其中x是问题，y是候选答案。Tan等人。[111]提出了一系列连体神经网络用于答案选择。如图17所示，这些是混合模型，它们使用卷积、递归，以及注意神经网络。其他为QA开发的暹罗神经网络包括基于LSTM的非因子答案选择模型[118]、双曲表示学习[119]和使用深度相似神经网络的问题回答[120]。



**2.10 Hybrid Models**

为了结合LSTM和CNN体系结构来捕捉句子和文档的局部和全局特征，已经开发了许多混合模型。朱等人。[121]提出了卷积LSTM（C-LSTM）网络。如图18（a）所示，C-LSTM利用C n n来提取高级短语（n-gram）表示序列，该序列被馈送到LSTM网络以获得句子表示。同样，张等人。[122]提出了一种用于文档建模的依赖敏感CNN（DSCNN）。如图18（b）所示，DSCNN是一个层次模型，其中LSTM学习被馈送到卷积层和最大池层的句子向量以生成文档表示。

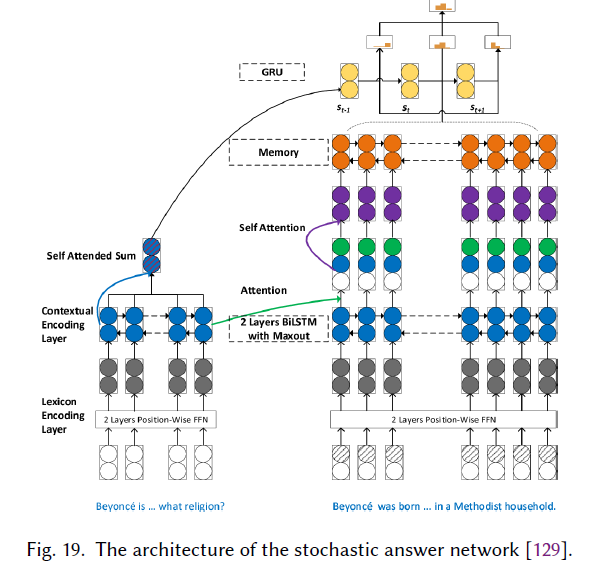


Chen等人。[123]通过CNN-RNN模型执行多标签文本分类，该模型能够捕获全局和局部文本语义，因此，在具有可处理的计算复杂性的同时，能够建模高阶标签相关性。Tang等人。[124]使用CNN学习句子表示，使用门控RNN学习编码句子之间内在关系的文档表示。肖等人。[125]将文档视为字符序列，而不是单词，并建议使用基于字符的卷积和递归层进行文档编码。与字级模型相比，该模型在参数较少的情况下取得了可比的性能。递归CNN[126]应用了递归结构获取学习单词表示的长期语境依赖。为了降低噪声，采用max-pooling算法自动选择对文本分类任务至关重要的显著词。

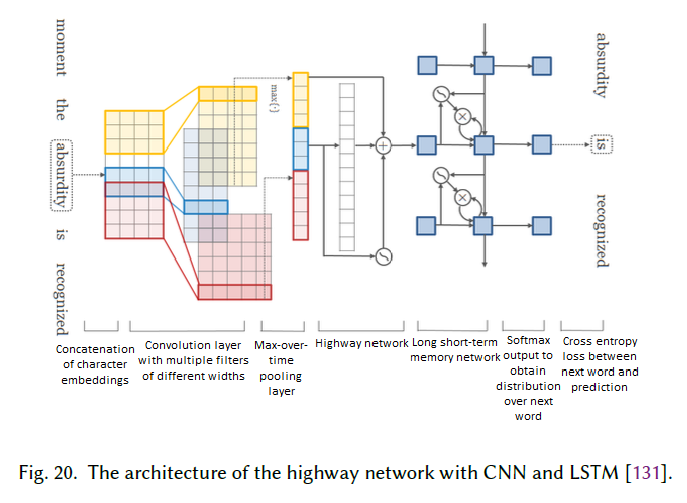
Chen等人。[127]通过观察不同类型的句子以不同的方式表达情感，提出了一种通过句子类型分类进行情感分析的分而治之方法。作者首先应用Bi LSTM模型将自以为是的句子分为三类。然后将每组句子分别输入一维CNN进行情感分类。在[128]中，Kowsari等人。提出了一种用于文本分类的分层深度学习方法（HDLTex）。

HDLTex采用了一系列混合的深度学习模型体系结构，包括MLP、RNN和CNN，以在文档层次结构的每个层次提供专门的理解。

Liu[129]提出了一种用于机器阅读理解中多步推理的鲁棒随机应答网络（SAN）。如图19所示，SAN结合了不同类型的神经网络，包括记忆网络、变换、Bi LSTM、注意和CNN。Bi LSTM组件获取问题和段落的上下文表示。它的注意机制导出了一个问题感知的段落表示。然后，使用另一个LSTM为文章生成一个工作内存。最后，基于门控递归单元（GRU）的应答模块输出预测。



一些研究集中在highway networks RNN和CNN的结合上。在典型的多层神经网络中，信息是逐层流动的。随着深度的增加，基于梯度的DNN训练变得越来越困难。highway networks [130]的设计是为了简化非常深的神经网络的训练。它们允许信息在信息高速公路上的多个层之间畅通无阻地流动，类似于ResNet中的快捷连接[3]。Kim等人。[131]使用带有CNN和LSTM over characters的高速公路网络进行语言建模。如图20所示，第一层执行字符嵌入的查找，然后应用卷积和最大池运算来获得给定给公路网的单词的固定维度表示。公路网的输出被用作多层LSTM的输入。最后，仿射变换后的softmax被应用于LSTM的隐藏表示，以获得下一个单词的分布。其他基于公路的混合模型包括重复公路网[132]和RNN与公路[133]。

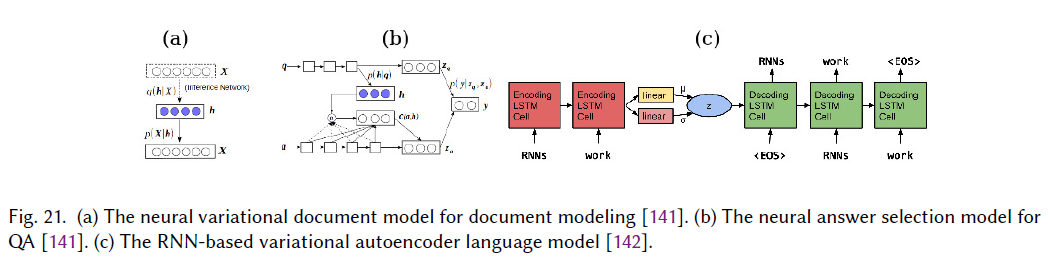


**2.11 Beyond Supervised Learning**

使用自动编码器的无监督学习。与单词嵌入类似，句子的分布式表示也可以以无监督的方式学习。通过优化一些辅助目标，例如自动编码器的重建损失[134]。这种无监督学习的结果是句子编码，它可以将具有相似语义和句法特性的句子映射到相似的固定大小向量表示。第2.7节中描述的基于变压器的plm也是可以用作句子编码器的无监督模型。本节讨论基于自动编码器及其变体的无监督模型。

Kiros等人。[135]提出了一种无监督学习通用句子编码器的跳跃思维模型。训练编码器-解码器模型来重建编码句子的周围句子。Dai和Le[136]研究了序列自动编码器（sequence autoencoder）在句子编码中的应用，该编码器将输入序列读入向量并再次预测输入。他们表明，在一个大的无监督语料库上进行预训练的句子编码比只进行预训练的单词嵌入产生更好的准确性。Zhang等人。[137]提出了一种平均最大注意自动编码器，该编码器利用多头自注意机制重构输入序列。在编码中使用mean-max策略，在编码过程中，对隐藏向量应用mean和max池操作来捕获输入的不同信息。

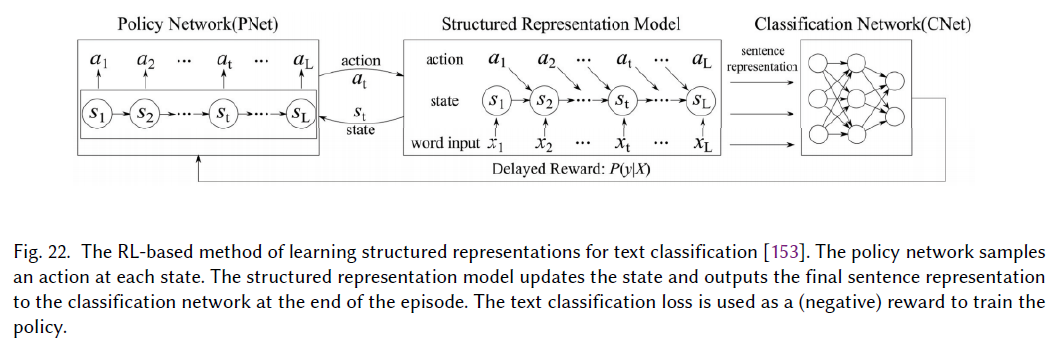
当自动编码器学习输入的压缩表示时，变分自动编码器（VAEs）[138，139]学习表示数据的分布，并且可以看作是自动编码器的正则化版本[140]。由于VAE学会了对数据建模，我们可以很容易地从分布中抽取样本来生成新的输入数据样本（例如，新的句子）。苗等人。[141]将VAE框架扩展到文本，提出了一种用于文档建模的神经变分文档模型（NVDM）和一种用于问答的神经应答选择模型（NASM）。如图21（a）所示，NVDM使用MLP编码器将文档映射到连续语义代表。如图21（b）所示，NASM使用LSTM和潜在随机注意机制来建模问答对的语义并预测其相关性。注意模型关注的是答案中与问题语义紧密相关的短语，并由潜在分布建模，从而允许模型处理任务中固有的歧义。Bowman等人。[142]提出了基于RNN的VAE语言模型，如图21（c）所示。该模型融合了整句的分布式潜在表示，允许显式地建模句子的整体属性，如风格、主题和高级句法特征。



对抗性训练。对抗训练[143]是一种改进分类器泛化的正则化方法。它通过提高模型对对手示例的鲁棒性来实现，对手示例是通过对输入进行小的扰动而创建的。对抗训练要求使用标签，并应用于监督学习。虚拟对抗训练[144]将对抗训练扩展到半监督学习。这是通过正则化一个模型来实现的，这样在给定的一个例子中，该模型产生的输出分布与在该例子的敌对扰动下产生的输出分布相同。Miyato等人。[145]通过对RNN中嵌入的单词而不是原始输入本身应用扰动，将对抗性和虚拟对抗性训练扩展到有监督和半监督的文本分类任务。Sachel等人。[146]研究了用于半监督文本分类的LSTM模型。他们发现，对于有标签和无标签的数据，使用结合了交叉熵、对抗性和虚拟对抗性损失的混合目标函数，可以显著改善监督学习方法。Liu等人。[147]将对抗训练扩展到文本分类的多任务学习框架[18]，旨在减轻任务无关（共享）和任务依赖（私有）的潜在特征空间之间的相互干扰。

关键单词标记的子集，以进行文本分类。hard-attention模型可以看作是一个代理，它采取是否选择令牌的操作。在遍历整个文本序列后，它会收到一个分类损失，可以作为训练代理的奖励。Liu等人。[150]提出了一种将文本分类建模为顺序决策过程的神经代理。

强化学习。强化学习（RL）[148]是一种训练agent根据策略执行离散动作的方法，该策略被训练为使奖励最大化。Shen等人。[149]使用了一个硬注意模型来选择输入序列中关键单词标记的子集，以进行文本分类。hard-attention模型可以看作是一个代理，它采取是否选择令牌的操作。在遍历整个文本序列后，它会收到一个分类损失，可以作为训练代理的奖励。Liu等人。[150]提出了一种将文本分类建模为顺序决策过程的神经代理。受人类文本阅读认知过程的启发，智能体按顺序扫描一段文本，并在希望的时间做出分类决策。分类结果和何时进行分类都是决策过程的一部分，由经过RL训练的策略控制。Shen等人。[151]提出了一种用于机器阅读理解的多步骤推理网络（ReasoNet）。ReasoNets任务有多个步骤来对查询、文档和答案之间的关系进行推理。推理机在推理过程中不使用固定数量的步骤，而是引入终止状态来放松对推理步骤的约束。通过使用RL，推理机可以动态地确定在消化中间结果之后是继续理解过程，还是在它认为现有信息足以产生结果时终止阅读生成一个答案。Li等人。[152]将RL、GANs和RNNs相结合，建立了一种新的类别句子生成对抗网络模型（CS-GAN），该模型能够生成扩展原始数据集的类别句子，并在有监督的训练中提高其泛化能力。Zhang等人。[153]提出了一种基于RL的文本分类学习结构化表示方法。他们提出了两个基于LSTM的模型。第一个选项只选择输入文本中与任务相关的重要单词。另一个发现句子的短语结构。使用这两个模型的结构发现被表示为由策略网络引导的顺序决策过程，策略网络在每个步骤决定使用哪个模型，如图22所示。利用策略梯度优化策略网络。



**3 TEXT CLASSIFICATION DATASETS**

在本节中，我们将描述一些广泛用于文本分类的数据集研究。我们根据主要的目标应用程序，将这些数据集分为情绪分析、新闻分类、主题分类、QA和NLI等类别。

**3.1 Sentiment Analysis Datasets**

Yelp。Yelp[154]是最流行的情绪分类数据集之一。在此数据集上定义了两个分类任务。一种是检测细粒度的情感标签，称为Yelp-5。另一个预测消极和积极的情绪，被称为Yelp回顾极性或Yelp-2。Yelp-5每班有65万个训练样本和5万个测试样本，Yelp-2每班有56万个训练样本和3.8万个阴性和阳性测试样本。图23示出了该数据集的字云表示。



IMDB。IMDB数据集[155]是为电影评论的二元情感分类任务而开发的。IMDB由相同数量的正面和负面评论组成。它平均分为训练和测试集，每个测试集有25000条评论。

Movie Review.。电影评论（MR）数据集[156]是为检测与特定评论相关联的情绪并确定其是消极的还是积极的任务而开发的电影评论的集合。它包括10662个句子的偶数个否定和肯定的样本。在这个数据集上，通常使用随机分割的10倍交叉验证进行测试。

SST。斯坦福情感树库（SST）数据集[157]是MR的扩展版本，有两个版本，一个是细粒度标签（五类），另一个是二进制标签，分别称为SST-1和SST2。SST-1包含11855个电影评论，分为8544个训练样本、1101个开发样本和2210个测试样本。SST-2分为三组，分别为6920、872和1821个训练、开发和测试集。

MPQA。多视角问答（MPQA）数据集[158]是一个带有两个类标签的意见语料库。MPQA由10606个句子组成，这些句子摘自与各种新闻来源相关的新闻文章。这是一个包含3311个阳性文档和7293个阴性文档的不平衡数据集。

Amazon。这是一个流行的产品评论语料库，收集自亚马逊网站[159]。它包含二进制分类和多类（5类）分类的标签。Amazon二进制分类数据集分别包含3600000条和400000条用于培训和测试的评论。amazon5级分类数据集（Amazon-5）由3000000条和650000条分别用于培训和测试的评论组成。

基于方面的情感分析。除了上述数据集外，还提出了几个数据集用于方面级情绪分析任务[160]。一些最受欢迎的数据集包括SemEval-2014 Task 4[161]、Twitter[162]、SentiHood[163]等等。

**3.2 News Classification Datasets**

AG News。AG新闻数据集[31]是学术新闻搜索引擎cometmyhead从2000多个新闻源收集的新闻文章的集合。该数据集包括120000个训练样本和7600个测试样本。每个示例都是带有四个类标签的短文本。

20 Newsgroups。20个新闻组数据集[164]是发布在20个不同主题上的新闻组文档的集合。该数据集的不同版本用于文本分类、文本聚类等。其中一个最流行的版本包含18821个文档，这些文档在所有主题中被均匀地分类。



Sogou News.。Sogou新闻数据集[90]是SogouCA和SogouCS新闻语料库的混合体。新闻的分类标签由它们在URL中的域名决定。例如，带有URL的新闻http://sports.sohu.com是归类为运动类。

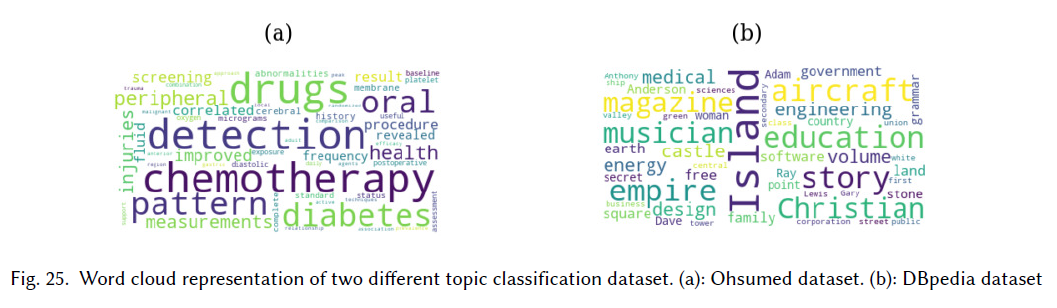
路透社新闻。Reuters-21578数据集[165]是文本分类研究中使用最广泛的数据集之一。它是1987年从路透社财经新闻通讯社收集的。ApteMod是Reuters-21578的多类版本，包含10788个文档。有90个班，7769份培训文件和3019份测试文件。还有许多其他数据集来自路透社数据集的不同子集，如R8、R52、RCV1和RCV1-v2。

为新闻分类开发的其他数据集包括：Bing news[166]、NYTimes[167]、BBC[168]、Google news[169]等等。

**3.3主题分类数据集**

DBpedia。DBpedia dataset[170]是一个大规模的多语言知识库，它是从Wikipedia中最常用的信息框创建的。DBpedia每月发布一次，在每个版本中添加或删除一些类和属性。最流行的DBpedia版本包含560000个训练样本和70000个测试样本，每个样本都有14个类标签。

Ohsumed。Ohsumed集合[171]是MEDLINE数据库的一个子集。Ohsumed包含7400个文档。每个文档都是一个医学摘要，由从23个心血管疾病类别中选择的一个或多个类别标记。图25提供了Ohsumed和DBpedia数据集的字云表示。



EUR-Lex。EUR-Lex数据集[172]包含不同类型的文档，这些文档根据几个正交分类方案进行索引，以允许多个搜索工具。这个数据集最流行的版本基于欧盟法律的不同方面，有19314个文件和3956个类别。

WOS。科学网（WOS）数据集[128]是从科学网获得的已发表论文的数据和元数据的集合，科学网是世界上最受信任的独立于出版商的全球引文数据库。WOS发布了三个版本：WOS-46985、WOS-11967和WOS-5736。WOS-46985是完整的数据集。WOS-11967和WOS-5736是WOS-46985的两个子集。

PubMed。PubMed[173]是国家医学图书馆为医学和生物科学论文开发的一个搜索引擎，包含一个文档集。每个文档都用网格集的类进行了标记，网格集是PubMed中使用的标签集。摘要中的每个句子都使用下列类别之一标记其在摘要中的角色：背景、目标、方法、结果或结论。

其他用于主题分类的数据集包括PubMed 200k RCT[174]、讽刺（由来自社交新闻网站reddit的注释评论、Twitter用于tweets主题分类的数据集、arXiv collection[175]组成）等等。

**3.4 QA Datasets**

SQuAD。斯坦福问答数据集（SQuAD）[6]是从维基百科文章中衍生出来的问答对的集合。在班里，问题的正确答案可以是给定文本中的任何标记序列。因为问答是由人类通过众包产生的，所以它比其他一些问答数据集更加多样化。SQuAD1.1包含536篇文章的107785个问答对。SQuAD2.0是最新版本，它将SQuAD1.1中的100000个问题与众包工人以类似于可回答问题的形式以对手的方式编写的50000多个不可回答的问题结合起来[176]。

MS MARCO.。此数据集由Microsoft[177]发布。与SQuAD不同，SQuAD中的所有问题都是通过编辑产生的；在MARCO女士中，所有问题都是通过Bing搜索引擎从用户查询和真实web文档中抽取的。马可女士的一些回答是有生成性的。因此，该数据集可用于生成性QA系统的开发。《马可女士》有多种版本，用于《摘录问答》、《文章排名》等。

TREC-QA。TREC-QA[178]是QA研究中最受欢迎和研究的数据集之一。这个数据集有两个版本，称为TREC-6和TREC-50。TREC-6包含6个类别的问题，TREC-50包含50个类别的问题。对于这两个版本，训练和测试数据集分别包含5452和500个问题。

WikiQA。WikiQA数据集[179]由一组问答对组成，收集并注释用于开放域QA研究。数据集还包括没有正确答案的问题，允许研究人员评估答案触发模型。

Quora。Quora数据集[180]用于释义识别（检测重复问题）。为此，作者提出了一个由40多万个问题对组成的Quora数据子集。为每个问题对分配一个二进制值，指示两个问题是否相同。

QAQA的其他数据集包括敌对世代（SWAG）[181]、WikiQA[179], SelQA[182],等。

**3.5 NLI Datasets**

SNLI。斯坦福自然语言推理（SNLI）数据集[183]被广泛用于自然语言推理。该数据集由550152、10000和10000个句子对组成，分别用于训练、开发和测试。每一对都用三个标签中的一个来标注：中性、蕴涵、矛盾。

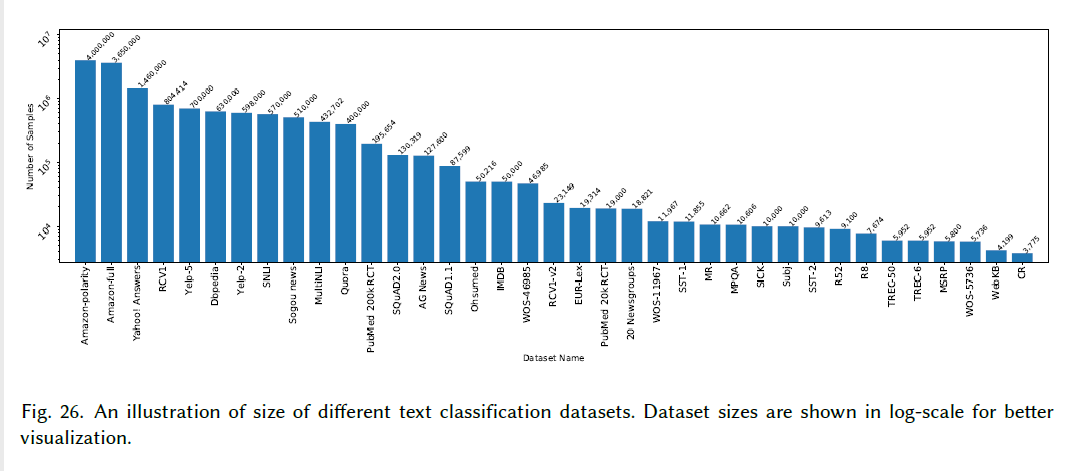
MNLI。多类型自然语言推理（MNLI）数据集[184]是433k个句子对的集合，这些句子对用文本蕴涵标签标注。语料库是SNLI的延伸，涵盖了更广泛的语篇和书面语篇体裁，支持独特的跨体裁概括评价。

SICK。包含合成知识（SICK）数据集的句子由大约10000个英语句子对组成，这些句子对用三个标签标注：蕴涵、矛盾和中性。

MSRP。微软研究释义（MSRP）数据集[185]通常用于文本相似性任务。MSRP包括4076个训练样本和1725个测试样本。每个样本都是一个句子对，用一个二进制标签标注，指示这两个句子是否是释义。

其他NLI数据集包括语义文本相似性（STS）[186]、RTE[187]、SciTail[188]，仅举几个例子。

图26示出了一组文本分类数据集，这些数据集根据其大小进行排序。

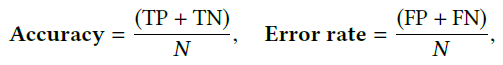


**4实验性能分析**

在这一部分中，我们首先描述了一组常用于评估文本分类模型性能的指标，然后对一组基于深度学习的文本分类模型在流行数据集上的性能进行了定量分析。

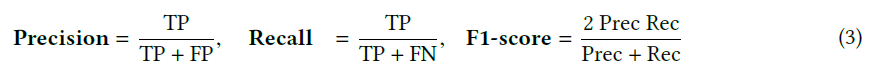
**4.1文本分类的常用度量**

准确度和错误率。这些是评估分类模型质量的主要指标。设TP、FP、TN、FN分别表示真阳性、假阳性、真阴性和假阴性。分类精度和错误率在公式2中定义



其中N是样本总数。很明显，我们有错误率=1-准确度

精确/召回/F1分数。这些也是主要的度量，并且比不平衡测试集的准确性或错误率更经常使用，例如，大多数测试样本都有一个类标签。二进制分类的精度和召回率定义为公式3。F1分数是精确性和召回率的调和平均值，如等式3所示。F1的分数在1（完美的精确性和召回率）时达到最佳值，在0时达到最差值。



对于多类分类问题，我们可以计算每个类标签的精确性和召回率，分析每个类标签上的个体性能或平均值，从而得到整体的精确性和召回率。

完全匹配（EM）。精确匹配度量是问答系统中一种流行的度量，它测量与任何一个基本真实答案完全匹配的预测的百分比。EM是SQuAD的主要指标之一。

平均倒数秩（MRR）。在查询文档排序和质量保证等NLP任务中，MRR经常被用来评估排序算法的性能。MRR在等式4中定义，其中Q是所有可能答案的集合，ranki是基本真理答案的排名位置。



其他广泛使用的度量包括平均精度（MAP）、曲线下面积（AUC）、错误发现率、错误遗漏率等等。

**4.2定量结果**

在本节中，我们将列出几个先前讨论的算法在流行的文本分类基准上的性能。在每一个表格中，除了一组有代表性的深度学习模型的结果外，我们还展示了使用非深度学习模型的结果，这些模型要么是最先进的，要么是在深度学习时代之前被广泛用作基线的。我们可以看到，在所有这些任务中，使用深度学习模型会带来显著的改进。

表1总结了第2节中描述的几个情绪分析数据集的模型结果，包括Yelp、IMDB、SST和Amazon。我们可以看到，自从引入第一个基于深度学习的情绪分析模型以来，在准确性方面已经取得了一些显著的提高，例如，分类误差相对减少了78%（在SST-2上）。

表2报告了三个新闻分类数据集（即AG news、20-news、Sogou news）和两个主题分类数据集（即DBpedia和ohsummated）的性能。观察到与情绪分析类似的趋势。

表3和表4分别展示了一些深度学习模型在SQuAD和WikiQA上的性能。

值得注意的是，在这两个数据集上，显著的性能提升归因于BERT的使用。

表5显示了两个NLI数据集（即SNLI和MNLI）的结果。在过去5年中，我们观察到这两个数据集的性能都在稳步提高。

**5 挑战与机遇**

近年来，文本分类在深度学习模型的帮助下取得了很大的进展。在过去的十年里，人们提出了一些新的思想（如神经嵌入、注意机制、自我注意、变换器、BERT和XLNet），这些思想导致了快速的发展。尽管取得了所有进展，但我们面前仍有若干挑战需要解决。本节介绍了其中一些挑战，并讨论了我们认为有助于推进该领域的研究方向。

用于更具挑战性任务的新数据集。尽管近年来已经为常见的文本分类任务收集了大量的大规模数据集，但是对于更具挑战性的任务，如具有多步骤推理的QA和多语言文档的文本分类，仍然需要新的数据集。为这些任务创建一个大规模的标记数据集有助于加快这些领域的进展。

建立常识知识模型。将常识性知识融入深度学习模型有可能显著提高模型性能，这与人类利用常识性知识执行不同任务的方式几乎相同。例如，一个配备常识知识库的QA系统可以回答关于真实世界的问题。常识知识也有助于在信息不完整的情况下解决问题。人工智能系统利用人们对日常事物或概念的普遍信念，以类似的方式，基于对未知事物的“默认”假设进行推理。尽管这一思想已经被用于情绪分类[202]，但是还需要更多的研究来探索如何在神经模型中有效地建模和使用常识知识。

可解释的深度学习模型。尽管深度学习模型在具有挑战性的基准测试中取得了令人满意的性能，但大多数模型都是不可解释的，并且仍然存在许多悬而未决的问题。例如，为什么一个模型在一个数据集上优于另一个模型，但在其他数据集上却表现不佳？深度学习模式到底学到了什么？什么是最小的神经网络结构，能够在给定的数据集上达到一定的精度？尽管注意和自我注意机制为回答这些问题提供了一些见解，但对这些模型的潜在行为和动力学的详细研究仍然缺乏。更好地理解这些模型的理论方面有助于开发针对各种文本分析场景的更好的模型。

内存效率模型。大多数现代神经语言模型需要大量的记忆来训练和推理。但是，为了满足移动设备的计算和存储限制，这些模型必须进行简化和压缩。这可以通过使用知识蒸馏来构建学生模型，或者使用模型压缩技术来实现。开发与任务无关的模型简化方法是一个活跃的研究课题[203]。

少射零射学习。大多数深度学习模型都是有监督的模型，需要大量的领域标签。实际上，为每个新域收集这样的标签是昂贵的。微调针对特定任务的预训练语言模型（PLM）如BERT和OpenGPT比从头开始训练一个模型需要更少的领域标签，从而为开发基于PLM的新的零炮或少炮学习方法提供了机会。

**6结论**

本文调查了150多个深度学习模型，这些模型是在过去的6-7年中发展起来的，在情感分析、新闻分类、主题分类、问答和自然语言学习等各种文本分类任务上都有了显著的进步。我们还概述了40多个流行的文本分类数据集，并对这些模型在几个公共基准上的性能进行了定量分析。最后，我们讨论了一些开放的挑战和未来的研究方向。