基于深度学习的文本分类：综合回顾

摘要。基于深度学习的模型在各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，包括情感分析、新闻分类、问答和自然语言推理。在本文中，我们全面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、相似之处和优势。我们还提供了 40 多个广泛用于文本分类的流行数据集的摘要。最后，我们对不同深度学习模型在流行基准上的性能进行了定量分析，并讨论了未来的研究方向。

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 简介

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，旨在为句子、查询、段落和文档等文本单元分配标签或标签。它具有广泛的应用，包括问答、垃圾邮件检测、情感分析、新闻分类、用户意图分类、内容审核等。文本数据可以来自不同的来源，包括网络数据、电子邮件、聊天、社交媒体、票证、保险索赔、用户评论以及客户服务的问答等等。文本是极其丰富的信息来源。但是，由于其非结构化的性质，从文本中提取见解可能具有挑战性且耗时。

文本分类可以通过人工标注或自动标注来进行。随着工业应用中文本数据规模的不断扩大，自动文本分类变得越来越重要。自动文本分类的方法可以分为两类：

- 基于规则的方法
- 基于机器学习（数据驱动）的方法

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，并且需要深厚的领域知识。另一方面，基于机器学习的方法学习根据对数据的观察对文本进行分类。使用预先标记的示例作为训练数据，机器学习算法可以学习文本与其标签之间的内在关联。

机器学习模型近年来引起了很多关注。大多数经典的基于机器学习的模型都遵循两步过程。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。在第二步中，这些特征被馈送到分类器以进行预测。流行的手工制作功能包括词袋 (BoW) 及其扩展。分类算法的流行选择包括朴素贝叶斯、支持向量机 (SVM)、隐马尔可夫模型 (HMM)、梯度提升树和随机森林。两步法有几个限制。例如，对手工特征的依赖需要繁琐的特征工程和分析才能获得良好的性能。此外，设计特征对领域知识的强烈依赖使得该方法难以推广到新任务。最后，这些模型无法充分利用大量训练数据，因为特征（或特征模板）是预先定义的。

已经探索了神经方法来解决由于使用手工特征而造成的限制。这些方法的核心组件是机器学习的嵌入模型，它将文本映射到低维连续特征向量，因此不需要手工制作的特征。最早的嵌入模型之一是

Dumais 等人开发的潜在语义分析 (LSA)。[1] 于 1989 年。LSA 是一个参数少于 100 万个的线性模型，训练了 200K 个单词。2001 年，Bengio 等人。[2] 提出了第一个基于前馈神经网络的神经语言模型，该网络训练了 1400 万个单词。然而，这些早期的嵌入模型不如使用手工特征的经典模型，因此没有被广泛采用。当使用大量训练数据开发更大的嵌入模型时，范式转变就开始了。2013 年，Google 开发了一系列 word2vec 模型 [3]，这些模型在 60 亿个单词上进行了训练，并立即在许多 NLP 任务中流行起来。2017 年，来自 AI2 和华盛顿大学的团队开发了一个基于 3 层双向 LSTM 的上下文嵌入模型，该模型具有 93M 参数，训练有 1B 个单词。该模型称为 ELMo [4]，
比 word2vec 工作得更好，因为它们捕获上下文信息。2018 年，OpenAI 开始使用 Transformer [5] 构建嵌入模型，这是一种由 Google 开发的新 NN 架构。Transformer 完全基于注意力，大大提高了 TPU 上大规模模型训练的效率。他们的第一个模型称为 GPT [6]，现在广泛用于文本生成任务。同年，Google 开发了基于双向 Transformer 的 BERT [7]。BERT 由 3.4 亿个参数组成，训练了 33 亿个单词，是当前最先进的嵌入模型。使用更大模型和更多训练数据的趋势仍在继续。到本文发表时，OpenAI 最新的 GPT-3 模型 [8] 包含 1700 亿个参数，Google 的 GShard [9] 包含 6000 亿个参数。

尽管这些巨大的模型在各种 NLP 任务上表现出令人印象深刻的性能，但一些研究人员认为，它们并不真正理解语言，并且对于许多关键任务领域来说不够健壮[10-14] 。最近，人们越来越有兴趣探索神经符号混合模型（例如，[15-18] ），以解决神经模型的一些基本限制，例如缺乏基础、无法执行符号推理、不可解释。这些工作虽然很重要，但超出了本文的范围。

虽然有很多关于文本分类方法和一般应用的优秀评论和教科书，例如 [19 – 21]，但这项调查的独特之处在于它全面回顾了为各种文本开发的 150 多种深度学习 (DL) 模型过去六年的分类任务，包括情感分析、新闻分类、主题分类、问答 (QA) 和自然语言推理 (NLI)。特别是，我们根据它们的神经网络架构将这些作品分为几类，包括循环神经网络 (RNN)、卷积神经网络 (CNN)、注意力、变形金刚、胶囊网络等。本文的贡献可以总结如下：

- 我们详细介绍了为文本分类提出的 150 多个 DL 模型。
- 我们审查了 40 多个流行的文本分类数据集。
- 我们对一组选定的深度学习模型在 16 个流行基准上的性能进行定量分析。

- 我们讨论剩余的挑战和未来的方向。

1.1 文本分类任务

文本分类 (TC) 是将文本（例如，推文、新闻文章、客户评论）分类为有组织的组的过程。典型的 TC 任务包括情感分析、新闻分类和主题分类。最近，研究人员表明，通过允许基于 DL 的文本分类器将一对文本作为输入（例如，[7、22、23] ）。_ _ 本节介绍本文讨论的五个 TC 任务，包括三个典型的 TC 任务和两个 NLU 任务，这些任务在最近的许多 DL 研究中通常被称为 TC。

*情绪分析。*这是分析人们在文本数据（例如，产品评论、电影评论或推文）中的意见，并提取他们的极性和观点的任务。该任务可以转换为二元或多类问题。二元情感分析将文本分为正面和负面类别，而多类别情感分析将文本分类为细粒度标签或多级强度。

新闻分类。 新闻内容是最重要的信息来源之一。新闻分类系统通过识别新兴新闻话题或根据用户兴趣推荐相关新闻等方式，帮助用户实时获取感兴趣的信息。

*主题分析。*该任务也称为*主题分类*，旨在识别文本的主题或主题（例如，产品评论是关于“客户支持”还是“易用性”）。

*问答（QA）。*有两种类型的 QA 任务：抽取式和生成式。抽取式 QA 是 TC 任务：给定一个问题和一组候选答案（例如，SQuAD [24] 中文档中的文本跨度），系统将每个候选答案分类为正确与否。生成式 QA 是一项文本生成任务，因为它需要即时生成答案。本文仅讨论抽取式 QA。

自然语言推理（NLI）。 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。NLI 系统需要为一对文本单元分配一个标签，例如蕴含、矛盾和中性 [25]。释义是 NLI 的一种广义形式，也称为*文本对比较*，其任务是测量句子对的语义相似性，表明一个句子是另一个句子的释义的可能性。

1.2 论文结构

本文的其余部分结构如下：第 2 节全面回顾了 150 多个基于 DL 的文本分类模型。第 3 节介绍了使用 DL 模型构建文本分类器的方法。第 4 节回顾了一些最流行的 TC 数据集。第 5 节介绍了在 16 个基准上对一

组选定的 DL 模型进行的定量性能分析。第 6 节讨论了基于 DL 的 TC 方法的主要挑战和未来方向。第 7 节总结了本文。

2 文本分类的深度学习模型

本节回顾了为各种 TC 任务提出的 150 多个 DL 模型。为了澄清起见，我们根据它们的模型架构将这些模型分为几类[1][1]：[2]

- 前馈网络将文本视为一个词袋（第 2.1 节）。
- 基于 RNN 的模型将文本视为单词序列，旨在捕获单词依赖关系和文本结构（第 2.2 节）。
- 训练基于 CNN 的模型以识别文本中的模式，例如 TC 的关键短语（第 2.3 节）。
- 胶囊网络解决了 CNN 的池化操作所遭受的信息丢失问题，最近已应用于 TC（第 2.4 节）。
- 注意机制可以有效地识别文本中的相关词，并已成为开发 DL 模型的有用工具（第 2.5 节）。
- 记忆增强网络将神经网络与一种外部记忆相结合，模型可以读取和写入（第 2.6 节）。
- 图神经网络旨在捕捉自然语言的内部图结构，例如句法和语义分析树（第 2.7 节）。
- 连体神经网络专为文本匹配而设计，是 TC 的一个特例（第 2.8 节）。
- 混合模型结合注意力、RNN、CNN 等来捕捉句子和文档的局部和全局特征（第 2.9 节）。
- Transformer 允许比 RNN 更多的并行化，从而可以使用 GPU 高效（预）训练非常大的语言模型（第 2.10 节）。
- 最后，在第 2.11 节中，我们回顾了监督学习之外的建模技术，包括使用自动编码器和对抗训练的无监督学习，以及强化学习。

读者应该相当熟悉基本的 DL 模型才能理解本节的内容。读者可以参考 Goodfellow 等人的 DL 教科书。[26] 了解更多详情。

2.1 前馈神经网络

前馈网络是用于文本表示的最简单的 DL 模型之一。然而，它们在许多 TC 基准测试中实现了高精度。这些模型将文本视为一袋单词。对于每个单词，他们使用诸如 word2vec [27] 或 Glove [28] 之类的嵌入模型来学习向量表示，将嵌入的向量和或平均值作为文本的表示，将其传递给一个或多个前馈层，称为多层感知器 (MLP)，然后使用逻辑回归、朴素贝叶斯或 SVM [29] 等分类器对最后一层的表示进行分类。这些模型的一个例子是深度平均网络（DAN）[29]，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合性。例如，DAN 在具有高句法差异的数据集上优于句法模型。乔林等人。[30] 提出了一种简单高效的文本分类器，称为 fastText。与 DAN 一样，fastText 将文本视为一个词袋。与 DAN 不同，fastText 使用一袋 n-gram 作为附加特征来捕获本地词序信息。事实证明，这在实践中非常有效，取得了与明确使用词序的方法相当的结果 [31]。

Le 和 Mikolov [32] 提出了 doc2vec，它使用无监督算法来学习可变长度文本片段的固定长度特征表示，例如句子、段落和文档。如图 2 所示，doc2vec 的架构类似于连续词袋 (CBOW) 模型 [3 , 27]。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，

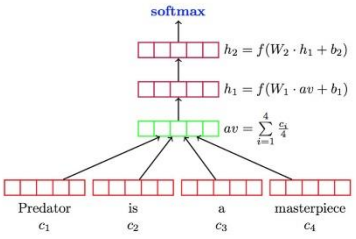

图 1. 深度平均网络 (DAN) [29] 的架构。

该向量与三个词的上下文的连接或平均值用于预测第四个词。段落向量表示当前上下文中缺失的信息，可以作为段落主题的记忆。在被训练之后，段落向量被用作段落的特征（例如，代替或补充 BoW），并馈送到分类器进行预测。Doc2vec 在发布时在多个 TC 任务上取得了最新的最新成果。

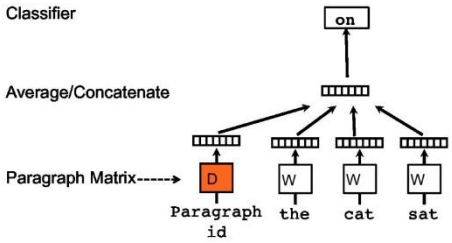

图 2. doc2vec 模型 [32]。

2.2 基于 RNN 的模型

基于 RNN 的模型将文本视为单词序列，旨在为 TC 捕获单词依赖性和文本结构。然而，vanilla RNN 模型表现不佳，并且通常表现不如前馈神经网络。在 RNN 的众多变体中，长短期记忆 (LSTM) 是最流行的架构，

旨在更好地捕获长期依赖关系。LSTM 通过引入一个记忆单元来记住任意时间间隔内的值，以及三个门（输入门、输出门、遗忘门）来调节信息流入和流出细胞。已经有一些工作通过捕获更丰富的信息来改进 TC 的 RNN 和 LSTM 模型，例如自然语言的树结构、文本中的长跨度单词关系、文档主题等。

泰等人。[33] 开发 Tree-LSTM 模型，将 LSTM 推广到树结构网络类型，以学习丰富的语义表示。作者认为，对于 NLP 任务，Tree-LSTM 是比链式 LSTM 更好的模型，因为自然语言表现出可以自然地将单词与短语结合起来的句法特性。他们验证了 Tree-LSTM 在两个任务上的有效性：情感分类和预测两个句子的语义相关性。这些模型的架构如图 3 所示。朱等人。[34]还将链结构的 LSTM 扩展到树结构，使用一个记忆单元来存储递归过程中多个子单元或多个后代单元的历史。他们认为，新模式提供了一个有原则的考虑层次结构上的长距离交互的方法，例如语言或图像解析结构。

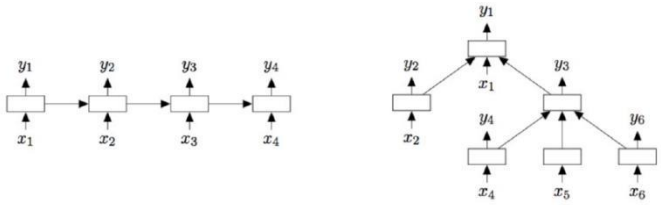

图 3.（左）链结构 LSTM 网络和（右）具有任意分支因子的树结构 LSTM 网络 [33]。这里x和 y表示每个单元的输入和输出。

为了模拟机器阅读的长跨度单词关系，Cheng 等人。[35] 用记忆网络代替单个记忆单元来增强 LSTM 架构。这使得在具有神经注意力的复发期间能够自适应内存使用，提供了一种弱诱导令牌之间关系的方法。该模型在语言建模、情感分析和 NLI 方面取得了可喜的成果。

多时间尺度 LSTM (MT-LSTM) 神经网络 [36] 还旨在通过捕获具有不同时间尺度的有价值信息来对长文本（例如句子和文档）进行建模。MT-LSTM 将标准 LSTM 模型的隐藏状态分为几组。每个组在不同的时间段被激活和更新。因此，MT-LSTM 可以对很长的文档进行建模。据报道，MT-LSTM 在 TC 上优于一组基线，包括基于 LSTM 和 RNN 的模型。

RNN 擅长捕捉词序列的局部结构，但难以记住长期依赖关系。相比之下，潜在主题模型能够捕捉文档的全局语义结构，但不考虑词序。迪恩等人。[37] 提出了一个 TopicRNN 模型来整合 RNN 和潜在主题模

型的优点。它使用 RNN 捕获本地（句法）依赖关系，并使用潜在主题捕获全局（语义）依赖关系。据报道，TopicRNN 在情感分析方面优于 RNN 基线。

还有其他有趣的基于 RNN 的模型。刘等人。[38] 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 [39] 探索了一种使用 LSTM 的文本区域嵌入方法。周等人。[40] 将双向 LSTM (Bi-LSTM) 模型与二维最大池化集成以捕获文本特征。王等人。[41]提出了“匹配-聚合”框架下的双边多视角匹配模型。万等人。[42] 使用由双向 LSMT 模型生成的多个位置句子表示来探索语义匹配。

值得注意的是，RNN 属于广泛的 DNN 类别，称为*递归神经网络*。递归神经网络在结构输入上递归地应用相同的权重集，以在可变大小输入上产生结构化预测或向量表示。RNN 是具有线性链结构输入的递归神经网络，而递归神经网络则在层次结构上运行，例如自然语言句子的解析树 [43]，将子表示组合成父表示。RNN 是 TC 中最流行的递归神经网络，因为它们有效且易于使用——它们将文本视为一系列标记，而不需要额外的结构标签，例如解析树。

2.3 基于 CNN 的模型

RNN 被训练为跨时间识别模式，而 CNN 学习跨空间识别模式 [44]。RNN 非常适用于 NLP 任务，例如需要理解远程语义的 POS 标记或 QA，而 CNN 在检测局部和位置不变模式很重要的情况下非常适用。这些模式可能是表达特定情绪的关键短语，如“我喜欢”或“濒危物种”等主题。

因此，CNN 已成为 TC 最流行的模型架构之一。

Kalchbrenner 等人提出了第一个基于 CNN 的 TC 模型。[45]。该模型使用动态 k -max-pooling，称为动态 CNN（DCNN）。如图 4 所示，DCNN 的第一层使用句子中每个单词的嵌入构造了一个句子矩阵。*然后，使用由动态*k -max-pooling 给出的将宽卷积层与动态池化层交替使用的卷积架构在句子上生成特征图，该特征图能够显式地捕获单词和短语的短期和长期关系。池化参数 k可以根据句子大小和卷积层次结构中的级别动态选择。

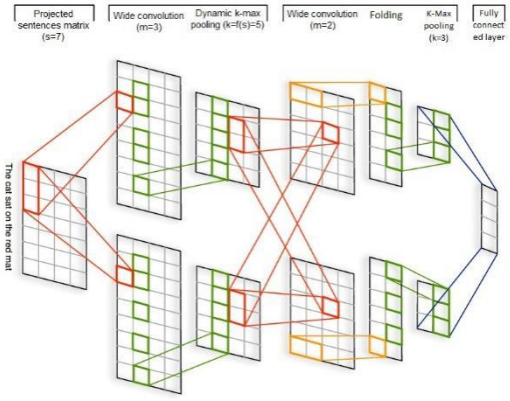

图 4 DCNN 模型的架构 [45]。

后来，Kim [46] 提出了一个比 DCNN 更简单的基于 CNN 的 TC 模型。如图 5，Kim 的模型仅在从无监督神经语言模型（即 word2vec）获得的词向量之上使用了一层卷积。Kim 还比较了四种不同的词嵌入学习方法：（1）CNN-rand，其中所有词嵌入随机初始化，然后在训练期间进行修改；(2) CNN-static，其中使用预训练的 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期间都经过微调；(4) CNN-multi-channel，其中使用了两组词嵌入向量，均使用 word2vec 进行初始化，其中一组在模型训练期间更新，另一组固定。据报道，这些基于 CNN 的模型改进了情感分析和问题分类方面的最新技术。

人们一直在努力改进 [45 , 46] 的基于 CNN 的模型的架构。刘等人。[47] 提出了一种新的基于 CNN 的模型，该模型对 Kim-CNN [46]的架构进行了两次修改。首先，采用动态最大池化方案从文档的不同区域捕获更细粒度的特征。其次，在池化层和输出层之间插入一个隐藏的瓶颈层，以学习紧凑的文档表示，以减小模型大小并提高模型性能。在 [48 , 49]，作者没有使用预训练的低维词向量作为 CNN 的输入，而是直接将 CNN 应用于高维文本数据，以学习小文本区域的嵌入以进行分类。

TC [50 , 51]也探索了字符级 CNN 。Zhang 等人提出了最早的此类模型之一。[50]。如图 6 所示，该模型将固定大小、编码为 one-hot 向量的字符作为输入，将它们传递给一个深度 CNN 模型，该模型由六个具有池化操作的卷积层和三个全连接层组成。普鲁萨等人。[52] 提出了一种使用 CNN 对文本进行编码的方法

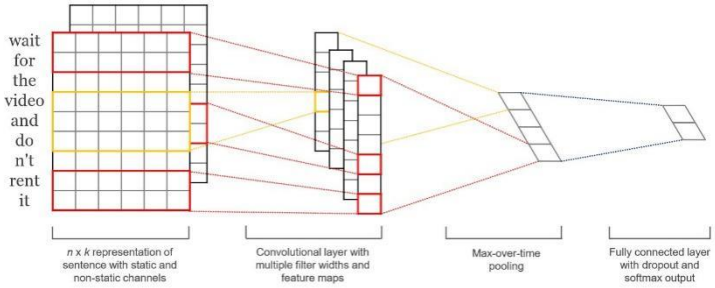

图 5. 用于文本分类的示例 CNN 模型的架构。由 Yoon Kim [46] 提供。

这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大小，允许从原始文本中保留更多信息以提高分类性能。

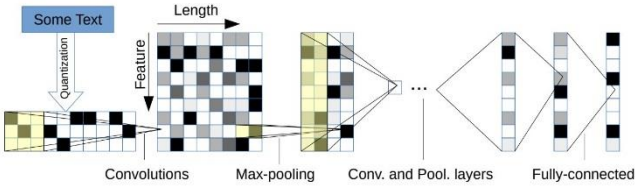

图 6. 字符级 CNN 模型的架构 [50]。

有研究调查词嵌入和 CNN 架构对模型性能的影响。受 VGG [53] 和 ResNets [54] 的启发，Conneau 等人。[55] 提出了一种用于文本处理的非常深的 CNN（VDCNN）模型。它直接在字符级别进行操作，并且仅使用小的卷积和池化操作。这项研究表明，VDCNN 的性能随着深度的增加而提高。杜克等人。[56] 修改 VDCNN 的结构以适应移动平台的约束，而不会造成太大的性能下降。他们能够将模型大小压缩 10 倍到 20 倍，精度损失在 0.4% 到 1.3% 之间。乐等人。[57] 表明当文本输入表示为字符序列时，深度模型确实优于浅层模型。然而，一个简单的浅而宽的网络在单词输入方面优于 DenseNet[58] 等深度模型。郭等人。[59] 研究词嵌入的影响，并建议通过多通道 CNN 模型使用加权词嵌入。张等人。[60] 研究了不同词嵌入方法和池化机制的影响，发现使用非静态 word2vec 和 GloVe 优于 one-hot 向量，并且最大池化始终优于其他池化方法。

还有其他有趣的基于 CNN 的模型。牟等人。[61] 提出了一个基于树的 CNN 来捕获句子级语义。庞等人。[62] 将文本匹配作为图像识别任务，并使用多层 CNN 来识别显着的 n-gram 模式。王等人。[63] 提出了一种基于 CNN 的模型，该模型结合了 TC 短文本的显式和隐式表示。将 CNN 应用于生物医学文本分类的兴趣也越来越大[64-67] 。

2.4 胶囊神经网络

CNN 通过使用连续的卷积层和池化对图像或文本进行分类。尽管池化操作可以识别显着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，并且可能会根据实体的方向或比例对实体进行错误分类。

为了解决池化问题，Hinton 等人提出了一种新方法，称为胶囊网络（CapsNets）[68、69] 。胶囊是一组神经元，其活动向量表示特定类型实体（例如对象或对象部分）的不同属性。向量的长度代表实体存在的概率，向量的方向代表实体的属性。与 CNN 的最大池化不同，它选择一些信息并丢弃其余信息，胶囊将下层中的每个胶囊“路由”到上层中的最佳父胶囊，使用网络中所有可用的信息直到最后一层进行分类。路由可以使用不同的算法来实现，例如动态路由协议 [69] 或 EM 算法 [70]。

最近，胶囊网络已应用于 TC，其中胶囊适用于将句子或文档表示为向量。[71 – 73] 提出了一个基于 CapsNets 变体的 TC 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，以及 (4) 全连接胶囊层。作者实验了三种策略来稳定动态路由过程，以减轻包含背景信息（例如停用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架构，Capsule-A 和 Capsule-B，如图 7 所示。Capsule-A 类似于 [69] 中的 CapsNet]。Capsule-B 在 n-gram 卷积层中使用三个具有不同窗口大小的过滤器的并行网络来学习更全面的文本表示。CapsNet-B 在实验中表现更好。

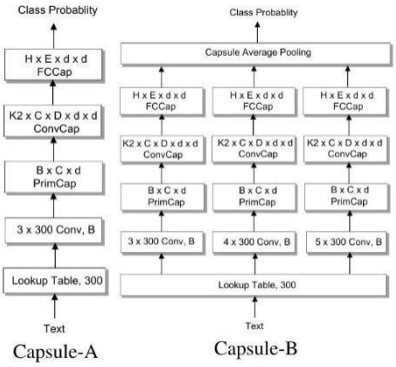

图 7. CapsNet A 和 B 用于文本分类 [71]。

Kim 等人提出的基于 CapsNet 的模型。[74] 使用类似的架构。该模型由 (1) 一个输入层组成，该输入层将文档作为词嵌入序列；(2) 卷积层，生成特征图并使用门控线性单元保留空间信息；(3) 卷积胶囊层通过聚合卷积层检测到的局部特征来形成全局特征；(4) 一个文本胶囊层来预测类标签。作者观察到对象可以在文本中比在图像中更自由地组合。例如，即使某些句子的顺序发生变化，文档的语义也可以保持不变，这与人脸上眼睛和鼻子的位置不同。因此，他们使用静态路由模式，该模式始终优于动态路由 [69] 为 TC。阿里等人。[75] 建议使用 CapsNets 进行分层多标签分类 (HMC)，认为 CapsNet 编码子父关系的能力使其更好

与 HMC 任务的传统方法相比，在传统方法中，文档被分配一个或多个按层次结构组织的类标签。他们的模型架构类似于 [71 , 72 , 74] 中的架构。

任等人。[76] 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，词嵌入是使用码本中的所有码字向量形成的。*然后低层胶囊捕获的特征通过k* - means 路由聚合到高层胶囊中。

2.5 带有注意力机制的模型

注意力的动机是我们如何将视觉注意力集中在图像的不同区域或在一个句子中关联单词。在开发用于 NLP 的 DL 模型时，注意力成为一个越来越流行的概念和有用的工具 [77 , 78]。简而言之，语言模型中的注意力可以解释为重要性权重的向量。为了预测句子中的一个词，我们使用注意力向量估计它与其他词的相关性或“关注度”有多强，并将注意力向量加权的它们的值的总和作为目标的近似值。

本节回顾了一些最突出的注意力模型，这些模型在 TC 任务上创造了新的艺术状态，当它们发布时。

杨等人。[79]提出了一种用于文本分类的分层注意网络。该模型具有两个显着特征：（1）反映文档层次结构的层次结构，以及（2）应用于单词和句子级别的两级注意机制，使其能够以不同的方式关注越来越重要的内容在构建文档表示时。该模型在六个 TC 任务上的性能大大优于以前的方法。周等人。[80] 将分层注意模型扩展到跨语言情感分类。在每种语言中，都使用 LSTM 网络对文档进行建模。然后，通过使用分层注意机制实现分类，其中句子级注意模型学习文档中的哪些句子对于确定整体情绪更重要。而
词级注意力模型学习每个句子中的哪些词是决定性的。

沉等人。[81] 提出了一种用于 RNN/CNN-free 语言理解的定向自注意力网络，其中来自输入序列的元素之间的注意力是定向的和多维的。一个轻量级的神经网络用于学习句子嵌入，仅基于提出的注意力，没有任何 RNN/CNN 结构。刘等人。[82] 提出了一个用于 NLI 的具有内部注意力的 LSTM 模型。该模型使用两阶段过程对句子进行编码。首先，在词级 Bi-LSTM 上使用平均池化来生成第一阶段的句子表示。其次，使用注意力机制来代替同一个句子上的平均池化以获得更好的表示。句子的第一阶段表示用于关注本身出现的单词。

注意力模型也广泛应用于成对排序或文本匹配任务。桑托斯等人。[83] 提出了一种双向注意力机制，称为注意力池 (AP)，用于成对排序。AP 使池化层能够了解当前输入对（例如，问答对），从而使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之外，AP 还联合学习了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力向量以指导池化。AP 是一个独立于底层表示学习的通用框架，可以应用于 CNN 和 RNN，如图 8 所示（一种）。王等人。[84] 将 TC 视为一个标签-词匹配问

题：每个标签与词向量嵌入在同一空间中。作者介绍了一个注意框架，该框架通过余弦相似度测量文本序列和标签之间嵌入的兼容性，如图 8 (b) 所示。

金等人。[85] 提出了一种语义句子匹配方法，该方法使用密集连接的循环和覆盖网络。与 DenseNet [58] 类似，该模型的每一层都使用注意力特征的连接信息以及所有前面循环层的隐藏特征。它可以保留原始和

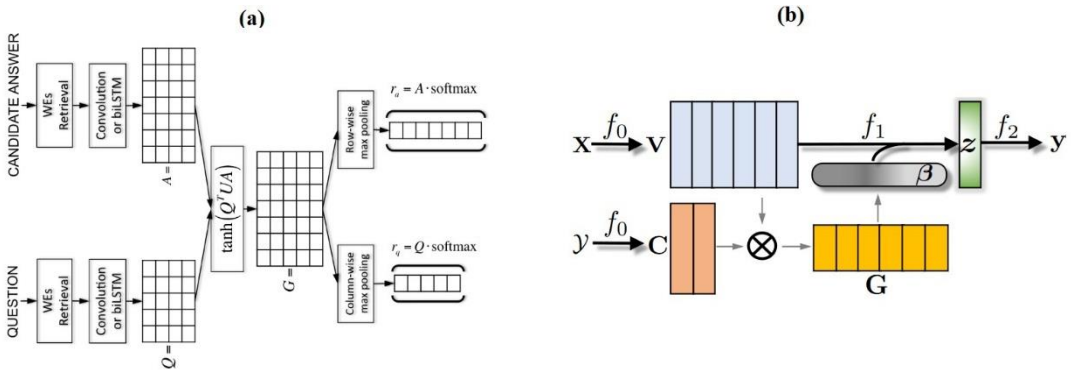

图 8. (a) 注意力池网络的架构 [83]。(b) 标签文本匹配模型的架构[84]。

从最底层的词嵌入层到最上层的循环层的共同注意特征信息。尹等人。[86] 提出了另一种基于注意力的 CNN 模型，用于句子对匹配。他们研究了三种将句子之间的相互影响整合到 CNN 中的注意力方案，以便每个句子的表示都考虑到它的配对句子。这些相互依赖的句子对表示比孤立的句子表示更强大，这在包括答案选择、释义识别和文本蕴涵在内的多个分类任务中得到了验证。谭等人。[87]在匹配聚合框架下使用多个注意力函数来匹配句子对。杨等人。[88] 介绍了一种基于注意力的神经匹配模型，用于对简短答案文本进行排序。他们采用价值共享加权方案而不是位置共享加权方案来组合不同的匹配信号，并使用问题注意力网络结合问题术语重要性学习。该模型在 TREC QA 数据集上取得了可喜的结果。

还有其他有趣的注意力模型。林等人。[89] 使用自我注意来提取可解释的句子嵌入。王等人。[90] 提出了一种具有多尺度特征注意力的密集连接 CNN，以产生可变的 n-gram 特征。Yamada 和 Shindo [91] 使用神经注意力实体袋模型来使用知识库中的实体执行 TC。帕里克等人。[92] 使用注意力将问题分解为可以单独解决的子问题。陈等人。[93]探索了增强句子嵌入的广义池化方法，并提出了一种基于向量的多头注意力模型。巴西里等人。[94] 提出了一种基于注意力的双向 CNN-RNN 深度模型进行情感分析。

2.6 记忆增强网络

虽然注意力模型在编码过程中存储的隐藏向量可以被视为模型*内部记忆*的条目，但记忆增强网络将神经网络与*外部记忆*结合在一起，模型可以读取和写入外部记忆。

Munkhdalai 和 Yu [95] 提出了一种用于 TC 和 QA 的记忆增强神经网络，称为神经语义编码器 (NSE)。NSE 配备了一个可变大小的编码内存，它随着时间的推移而演变，并通过读、写和写操作保持对输入序列的理解，如图 9 所示。

韦斯顿等人。[96] 为合成 QA 任务设计一个记忆网络，其中向模型提供一系列语句（记忆条目）作为问题的支持事实。该模型根据问题和先前检索到的记忆，学习一次从记忆中检索一个条目。苏赫巴托尔等人。[97] 扩展了这项工作并提出了端到端的记忆网络，其中记忆条目以软方式检索

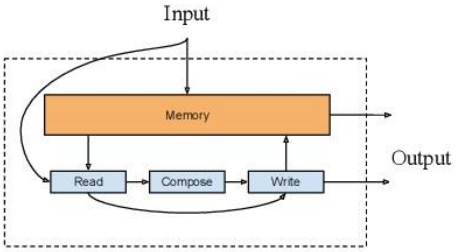

图 9. NSE [95] 的架构。

注意力机制，从而实现端到端的训练。他们表明，通过多轮（跳跃），该模型能够检索和推理几个支持事实来回答特定问题。

库马尔等人。[98] 提出了一种动态记忆网络 (DMN)，它处理输入序列和问题，形成情景记忆，并生成相关答案。问题触发了一个迭代注意过程，该过程允许模型将其注意力集中在输入和先前迭代的结果上。然后在分层循环序列模型中对这些结果进行推理以生成答案。DMN 进行端到端训练，并在 QA 和 POS 标记方面获得最先进的结果。熊等人。[99] 对 DMN 进行了详细分析，并改进了它的记忆和输入模块。

2.7 图神经网络

尽管自然语言文本表现出顺序，但它们也包含内部图结构，例如句法和语义分析树，它们定义了句子中单词之间的句法和语义关系。

为 NLP 开发的最早的基于图的模型之一是 TextRank [100]。作者建议将自然语言文本表示为图G（V，E），其中V表示一组节点，E表示节点之间的一组边。根据手头的应用，节点可以表示各种类型的文本单

元，例如单词、搭配、整个句子等。类似地，边可以用来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

现代图神经网络 (GNN) 是通过扩展图数据的 DL 方法开发的，例如 TextRank 使用的文本图。深度神经网络，如 CNN、RNN 和自动编码器，在过去几年中已被推广用于处理图数据的复杂性 [101]。例如，用于图像处理的 CNN 的 2D 卷积被泛化为通过取节点邻域信息的加权平均值来执行图卷积。在各种类型的 GNN 中，卷积 GNN，例如图卷积网络 (GCN) [102] 及其变体是最流行的，因为它们可以有效且方便地与其他神经网络组合，并在许多应用中取得了最先进的结果。GCN 是图上 CNN 的一种有效变体。GCN 堆叠学习的一阶光谱滤波器层，然后是非线性激活函数来学习图形表示。

GNN 在 NLP 中的一个典型应用是 TC。GNN 利用文档或单词的相互关系来推断文档标签 [102 – 104]。在下文中，我们回顾了为 TC 开发的 GCN 的一些变体。

彭等人。[105]提出了一种基于 graph-CNN 的 DL 模型，首先将文本转换为词图，然后使用图卷积操作对词图进行卷积，如图 10 所示。他们通过实验表明，文本的词图表示具有捕获非连续和长距离语义的优势，而 CNN 模型具有学习不同级别语义的优势。

在[106]中，彭等人。提出了一种基于分层分类感知和注意力图胶囊 CNN 的 TC 模型。该模型的一个独特功能是使用类标签之间的层次关系，即

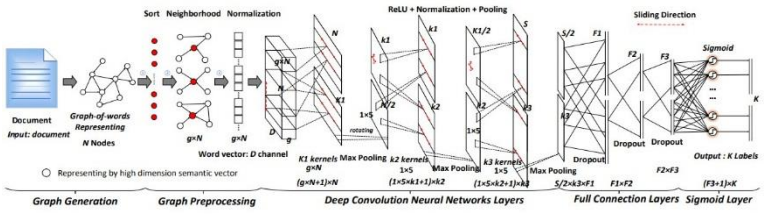

图 10. Peng 等人使用的 GNN 架构。[105]。

在以前的方法中被认为是独立的。具体来说，为了利用这种关系，作者开发了一种分层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性来定义一种新的加权边距损失。

姚等人。[107] 对 TC 使用类似的 Graph CNN (GCNN) 模型。他们基于单词共现和文档单词关系为语料库构建单个文本图，然后为语料库学习文本图卷积网络（Text GCN），如图 11 所示。Text GCN 使用单词和文档的 one-hot 表示进行初始化，然后在已知文档类标签的监督下联合学习单词和文档的嵌入。

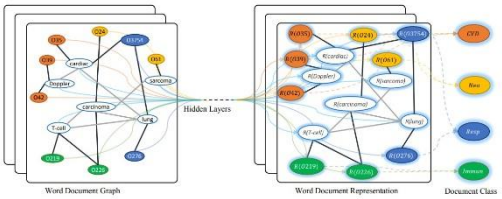

图 11. GCNN [107] 的架构。

为大规模文本语料库构建 GNN 的成本很高。通过降低模型复杂性或改变模型训练策略来降低建模成本已经有很多工作。前者的一个例子是 [108] 中提出的简单图卷积 (SGC) 模型，其中通过重复去除连续层之间的非线性并将结果函数（权重矩阵）折叠成单个线性来简化深度卷积 GNN 转型。后者的一个例子是文本级 GNN [109]。文本级 GNN 不是为整个文本语料库构建图，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，以减少训练期间的内存消耗。其他一些有前途的基于 GNN 的作品包括 GraphSage [103] 和上下文化的非局部神经网络 [110]。

2.8 连体神经网络

连体神经网络 (S2Nets) [111 , 112] 及其 DNN 变体，称为深度结构化语义模型（DSSM）[113 , 114]，专为文本匹配而设计。该任务是许多 NLP 应用程序的基础，例如抽取式 QA 中的查询文档排名和答案选择。这些任务可以看作是 TC 的特例。例如，在问题文档排名中，我们希望将文档分类为与给定查询相关或不相关。

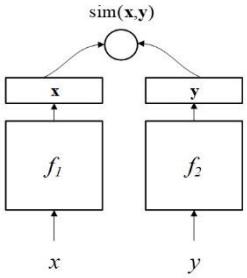

图 12. DSSM 的架构，如 [115]所示

如图 12 所示，DSSM（或 S2Net）由一对 DNN，f_1和f_2组成，它们将输入x和y *映射*到公共低维语义空间中的相应向量 [115]。然后通过两个向量的余弦距离来衡量x和y*的相似度*。虽然 S2Nets 假设f_1和f

$_2$共享相同的架构甚至相同的参数，但在 DSSM 中，f_1和f_2可以具有不同的架构，具体取决于x和y。例如，为了计算图像-文本对的相似度，f_1可以是深度 CNN，f_2可以是 RNN 或 MLP。

根据(x,y)的定义，这些模型可以应用于广泛的 NLP 任务。例如，(x,y)可以是查询-文档排名的查询-文档对 [114 , 116]，或 QA 中的问答对 [117 , 118]。

模型参数θ通常使用成对排名损失进行优化。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，而y^-不相关。令 $\text{sim}_\theta(x,y)$ 为x和y在θ参数化的语义空间中的余弦相似度。训练目标是将基于边际的损失最小化为

$$L(\theta) = \gamma + \text{sim}_\theta(x,y^-) - \text{sim}_\theta(x,y^+)_+, \tag{1}$$

其中 $[x]_+ := \max(0,x)$ 和γ是边距超参数。

由于文本呈现顺序，因此很自然地使用 RNN 或 LSTM 来实现f_1和f_2来测量文本之间的语义相似度。图 13 显示了 [119] 中提出的孪生模型的架构，其中两个网络使用相同的 LSTM 模型。Neculoiu 等人。[120] 提出了一个类似的模型，该模型使用字符级 Bi-LSTM 用于f_1和f_2以及余弦函数来计算相似度。刘等人。[121] 模拟句子对与两个耦合 LSTM 的交互。除了 RNNs，BOW 模型和 CNNs 也被用于 S2Nets 来表示句子。例如，他等人。[122] 提出了一个使用 CNN 对多视角句子相似度建模的 S2Net。承租人等。[123] 提出了一种 Siamese 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 成为最先进的句子嵌入模型，已经有人尝试构建基于 BERT 的 S2Net，例如 SBERT [124] 和 TwinBERT [125]。

S2Nets 和 DSSM 已广泛用于 QA。达斯等人。[117] 提出了一个用于 QA（SCQA）的连体 CNN 来测量问题与其（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 的参数以最大化问题与其相关答案之间的语义相似性，如公式 1 所示，其中x是问题，y是候选答案。谭等人。[118] 提出了一系列用于答案选择的孪生神经网络。如图 14，这些是使用卷积、循环和注意力神经网络处理文本的混合模型。为 QA 开发的其他孪生神经网络包括基于 LSTM 的非事实答案选择模型 [126]、双曲线表示学习 [127] 和使用深度相似性神经网络的 QA [128]。

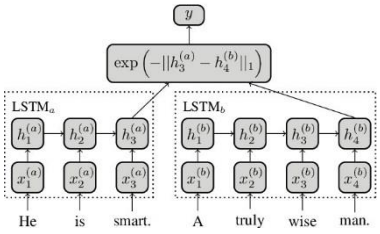

图 13. Mueller 等人提出的 Siamese 模型的架构。[119]。

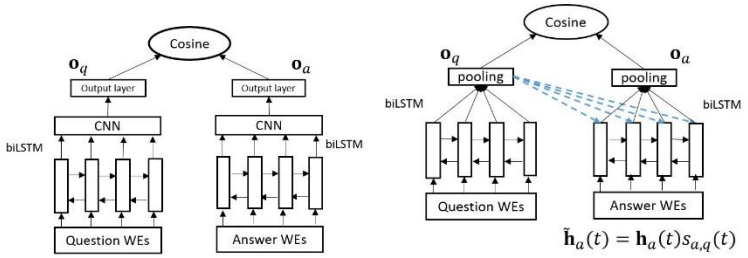

图 14. [118] 中研究的连体模型的架构。

2.9 混合模型

已经开发了许多混合模型来结合 LSTM 和 CNN 架构来捕获句子和文档的局部和全局特征。朱等人。[129] 提出了一个卷积 LSTM (C-LSTM) 网络。如图 15 (a) 所示，C-LSTM 利用 CNN 提取一系列高级短语（n-gram）表示，将其馈送到 LSTM 网络以获得句子表示。同样，张等人。[130]提出了一种用于文档建模的依赖敏感 CNN（DSCNN）。如图 15 (b) 所示，DSCNN 是一个分层模型，其中 LSTM 学习句子向量，这些向量被馈送到卷积层和最大池化层以生成文档表示。

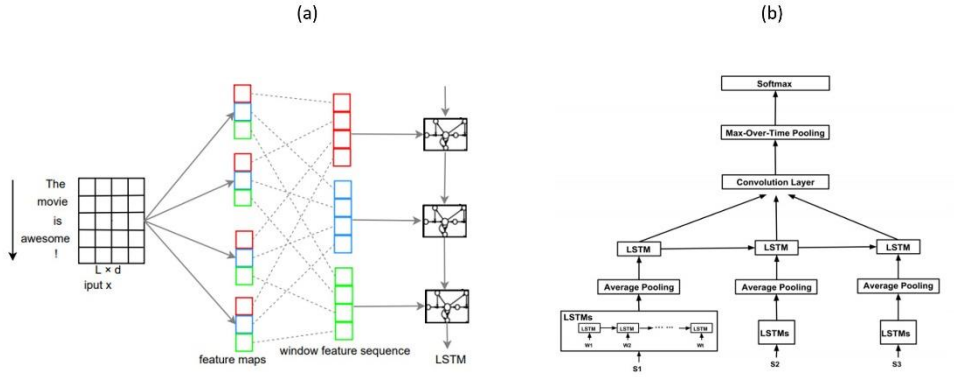

图 15. (a) C-LSTM [129] 的架构。(b) 用于文档建模的 DSCNN 架构 [130]。

陈等人。[131] 通过 CNN-RNN 模型执行多标签 TC，该模型能够捕获全局和局部文本语义，因此，在具有易于处理的计算复杂性的同时对高阶标签相关性进行建模。唐等人。[132] 使用 CNN 来学习句子表示，并使用门控 RNN 来学习编码句子之间内在关系的文档表示。肖等人。[133]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循环层进行文档编码。与单词级模型相比，该模型以更少的

参数实现了可比的性能。循环 CNN [134] 应用循环结构来捕获用于学习单词表示的长期上下文相关性。为了减少噪声，使用最大池来自动选择对文本分类任务至关重要的显着词。

陈等人。[135] 提出了一种通过句子类型分类进行情感分析的分而治之的方法，其动机是观察到不同类型的句子以非常不同的方式表达情感。作者首先应用 Bi-LSTM 模型将自以为是的句子分为三种类型。然后将每组句子分别馈送到一维 CNN 进行情感分类。

在 [136] 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 采用混合 DL 模型架构堆栈，包括 MLP、RNN 和 CNN，以在文档层次结构的每个级别提供专门的理解。

Liu [137] 提出了一种鲁棒的随机答案网络 (SAN)，用于机器阅读理解中的多步推理。SAN 结合了不同类型的神经网络，包括记忆网络、Transforms、Bi-LSTM、注意力和 CNN。Bi-LSTM 组件获取问题和段落的上下文表示。它的注意力机制派生了一个问题感知的段落表示。然后，另一个 LSTM 用于为该段落生成工作记忆。最后，基于门控循环单元的答案模块输出预测。

一些研究集中在将高速公路网络与 RNN 和 CNN 相结合。在典型的多层神经网络中，信息逐层流动。随着深度的增加，基于梯度的 DNN 训练变得更加困难。高速公路网络 [138] 旨在简化非常深的神经网络的训练。*它们允许信息高速公路*上的多个层畅通无阻的信息流，类似于 ResNet [139] 中的快捷连接。金等人。[140] 在字符上使用带有 CNN 和 LSTM 的高速公路网络进行语言建模。如图 16 所示，第一层执行字符嵌入的查找，然后应用卷积和最大池操作以获得单词的固定维度表示，将其提供给高速网络。高速公路网络的输出用作多层 LSTM 的输入。最后，将仿射变换和 softmax 应用于 LSTM 的隐藏表示，以获得下一个单词的分布。其他基于高速公路的混合模型包括循环高速公路网络 [141] 和带有高速公路的 RNN [142]。

2.10 Transformer 和预训练的语言模型

RNN 遇到的计算瓶颈之一是文本的顺序处理。尽管 CNN 的顺序性不如 RNN，但捕获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，类似于 RNN。Transformers [5] 通过应用自注意力来并行计算句子中的每个单词或记录一个“注意力分数”来模拟每个单词对另一个单词的影响，从而克服了这一限制[3][2]。由于这个特性，Transformers 允许比 CNN 和 RNN 更多的并行化，这使得在 GPU 上的大量数据上有效地训练非常大的模型成为可能。[4]

[3]严格来说，Transformer 是混合模型（2.9）的一个实例，因为每个 Transformer 层都是一个由前馈层和多头注意力层组成的复合结构。

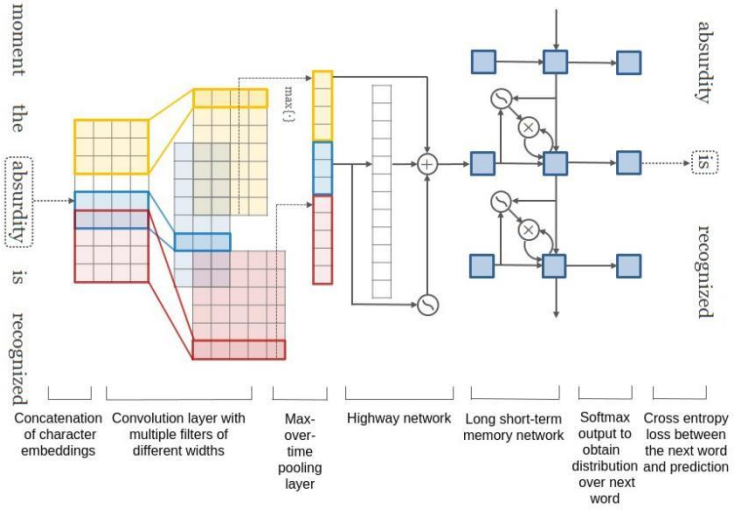

图 16. 使用 CNN 和 LSTM [140] 的高速公路网络架构。

自 2018 年以来，我们看到了一组基于 Transformer 的大规模预训练语言模型 (PLM) 的兴起。与早期基于 CNN [143] 或 LSTM [4] 的上下文化嵌入模型相比，基于 Transformer 的 PLM 使用更深的网络架构（例如，48 层 Transformer [144]），并且 在大量文本上进行了*预训练*语料库通过根据上下文预测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*，并在包括 TC 在内的许多下游 NLP 任务中创造了新的技术水平。虽然预训练是无监督的（或自我监督的），但微调是有监督的学习。邱等人最近的一项调查。[145] 按其表示类型、模型架构、预训练任务和下游任务对流行的 PLM 进行分类。

PLM 可以分为两类，自回归 PLM 和自编码 PLM。最早的自回归 PLM 之一是 OpenGPT [6 , 144]，这是一种单向模型，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先前的预测。图 17 显示了 OpenGPT 的架构。它由 12 层 Transformer 块组成，每层都包含一个带掩码的多头注意力模块，然后是一个层归一化和一个位置前馈层。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 TC 等下游任务。

使用最广泛的自动编码 PLM 之一是 BERT [7]。与 OpenGPT 根据之前的预测预测单词不同，BERT 使用掩蔽语言建模 (MLM) 任务进行训练，该任务随机掩蔽文本序列中的一些标记，然后通过调节双向 Transformer 获得的编码向量来独立恢复被掩蔽的标记. 有很多关于改进 BERT 的工作。RoBERTa [146] 比 BERT 更健壮，并且使用更多的训练数据进行训练。ALBERT [147] 降低了内存消耗，提高了 BERT 的训练速度。DistillBERT [148] 在预训练期间利用知识蒸馏将 BERT 的大小减少 40%，同时保留 99% 的原始能力，并使推理速度提高 60%。SpanBERT [149] 扩展了 BERT 以更好地表示和预测文本跨度。Electra [150] 使用比

MLM 更有效的样本预训练任务，称为替换标记检测。它不是屏蔽输入，而是通过用从小型生成器网络中采样的合理替代方案替换一些令牌来破坏它。ERNIE [151 , 152] 结合了来自外部知识库的领域知识，例如命名实体，用于

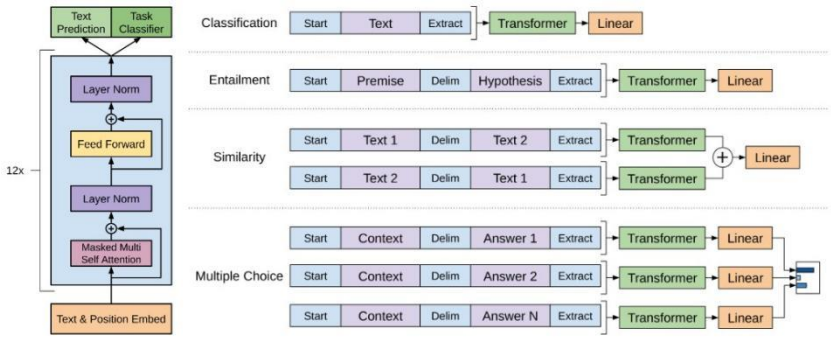

图 17. OpenGPT-1 的架构 [144]

模型预训练。ALUM [14] 为模型预训练引入了对抗性损失，提高了模型对新任务的泛化能力和对抗性攻击的鲁棒性。BERT 及其变体已针对各种 NLP 任务进行了微调，包括 QA [153]、TC [154] 和 NLI [23、155] 。

已经尝试将自回归和自编码 PLM 的优势结合起来。XLNet [156] 集成了 OpenGPT 等自回归模型的思想和 BERT 的双向上下文建模。XLNet 利用*置换操作*在预训练期间，允许上下文包含左右两个标记，使其成为通用的顺序感知自回归语言模型。排列是通过在 Transformers 中使用特殊的注意掩码来实现的。XLNet 还引入了双流自我注意模式，以允许位置感知词预测。这是由于观察到单词分布根据单词位置而有很大差异。例如，句子的开头与句子中的其他位置有很大不同的分布。如图 18 所示，为了预测排列 3-2-4-1 中位置 1 的单词标记，通过包含所有先前单词 (3, 2, 4) 的位置嵌入和标记嵌入形成内容流，然后查询流是由内容流和要预测的单词（位置为 1 的单词）的位置嵌入组成，最后模型根据查询流中的信息进行预测。

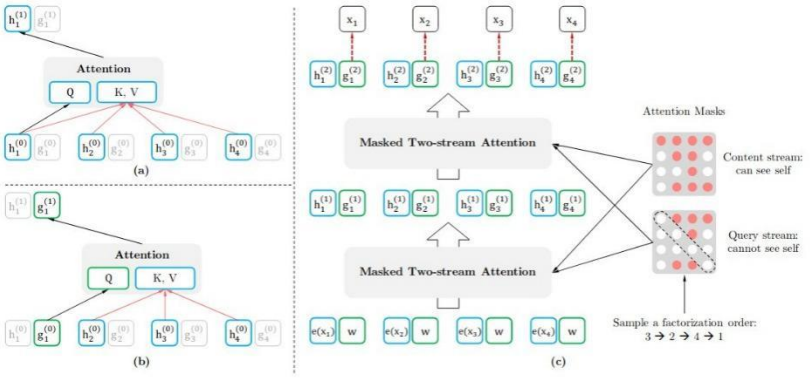

图 18. XLNet [156] 的架构：a) 内容流注意力，b) 查询流注意力，c) 具有双流注意力的置换语言建模训练概述。

如前所述，OpenGPT 使用从左到右的 Transformer 来学习用于自然语言生成的文本表示，而 BERT 使用双向 Transformer 来进行自然语言理解。统一语言模型 (UniLM) [157] 旨在解决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。统一建模是通过使用共享的 Transformer 网络并利用特定的自注意力掩码来控制预测条件的上下文来实现的，如图 19 所示。UniLM 第二版[158] 据报道，在广泛的自然语言理解和生成任务上实现了最新的技术水平，显着优于以前的 PLM，包括 OpenGPT-2、XLNet、BERT 及其变体。

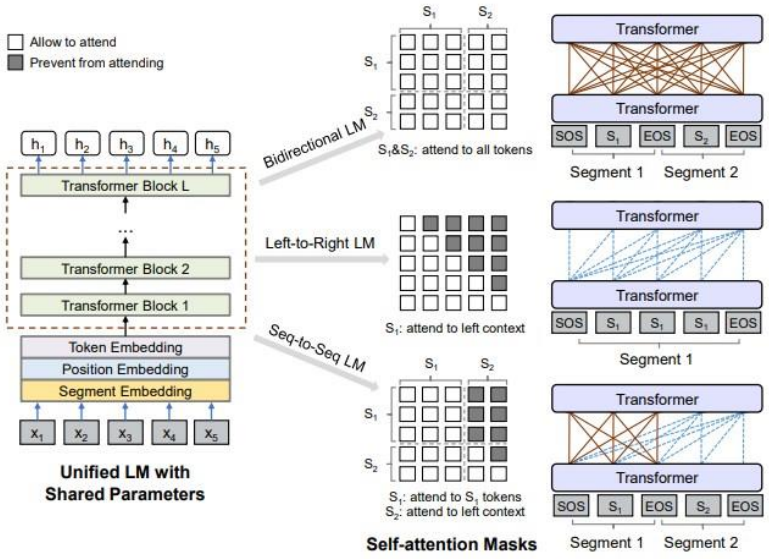

图 19. UniLM 预训练概述 [157]。模型参数在语言建模目标之间共享，即双向、单向和序列到序列的语言建模。使用不同的自注意力掩码来控制每个词标记对上下文的访问。

拉菲尔等人。[159] 提出了一个统一的基于 Transformer 的框架，它将许多 NLP 问题转换为文本到文本的格式。他们还进行系统研究，比较数十种语言理解任务的预训练目标、架构、未标记数据集、微调方法和其他因素。

2.11 超越监督学习

*使用自动编码器的无监督学习。*与词嵌入类似，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 [160]。这种无监督学习的结果是句子编码器，它可以将具有相似语义和句法属性的句子映射到相似的固定大小向量表示。第 2.10 节中描述的基于 Transformer 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。

基罗斯等人。[161] 提出了用于无监督学习通用句子编码器的 Skip-Thought 模型。训练编码器-解码器模型以重建编码句子的周围句子。戴和

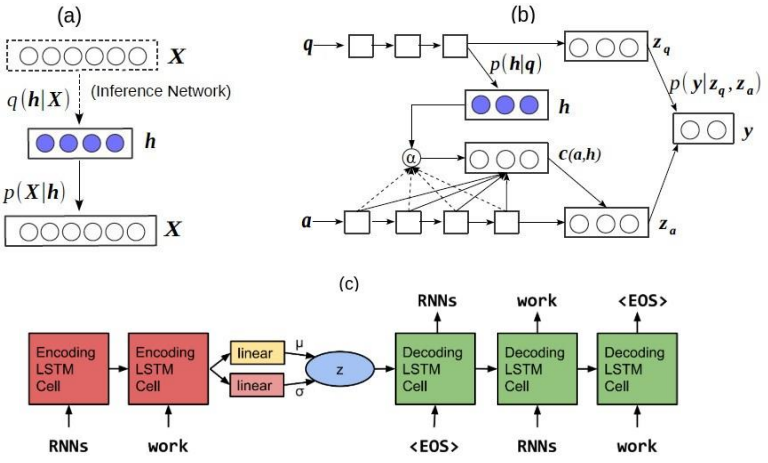

图 20. (a) 用于文档建模的神经变分文档模型 [166]。(b) QA [166]的神经答案选择模型。(c) 基于 RNN 的变分自编码器语言模型 [167]。

Le [162] 研究了序列自动编码器的使用，它将输入序列读入向量并再次预测输入，用于句子编码。他们表明，在大型无监督语料库上预训练句子编码器比仅预训练词嵌入产生更好的准确性。张等人。[163]提出了一种 mean-max 注意力自动编码器，它使用多头自注意力机制来重构输入序列。在编码中使用均值-最大值策略，其中对隐藏向量应用均值和最大池化操作来捕获输入的不同信息。

自编码器学习输入的压缩表示，而变分自编码器 (VAE) [164，165] 学习表示数据的分布，并且可以被视为自编码器 [26] 的正则化版本。由于 VAE 学习对数据建模，我们可以轻松地从分布中采样以生成新

样本（例如，新句子）。苗等人。[166] 将 VAE 框架扩展到文本，并提出了用于文档建模的神经变分文档模型 (NVDM) 和用于 QA 的神经答案选择模型 (NASM)。如图 20 (a) 所示，NVDM 使用 MLP 编码器将文档映射到连续语义表示。如图 20 (b)，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它们的相关性。注意力模型专注于与问题语义密切相关的答案短语，并由潜在分布建模，从而使模型能够处理任务中固有的歧义。鲍曼等人。[167]提出了一种基于 RNN 的 VAE 语言模型，如图 20 （c）所示。该模型结合了整个句子的分布式潜在表示，允许显式地建模句子的整体属性，例如风格、主题和高级句法特征。古鲁兰甘等人。[168] 将文档模型预训练为域内未标记数据的 VAE，并将其内部状态用作文本分类的特征。一般来说，使用 VAE 或其他模型的数据增强 [169 , 170] 广泛用于半监督或弱监督 TC。

*对抗训练。*对抗性训练 [171] 是一种用于改进分类器泛化的正则化方法。它通过提高模型对对抗性示例的鲁棒性来做到这一点，这些对抗性示例是通过对输入进行小扰动而创建的。对抗训练需要使用标签，并应用于监督学习。虚拟对抗训练 [172] 将对抗训练扩展到半监督学习。这是通过正则化模型来完成的，以便给定一个示例，该模型产生的输出分布与它在该示例的对抗性扰动下产生的输出分布相同。宫藤等人。[173] 通过将扰动应用于 RNN 中的词嵌入而不是原始输入本身，将对抗性和虚拟对抗性训练扩展到监督和半监督 TC 任务。萨切尔等人。[174] 研究半监督 TC 的 LSTM 模型。他们发现，使用混合目标函数结合了标记和未标记数据的交叉熵、对抗性和虚拟对抗性损失，可以显着改进监督学习方法。刘等人。[175] 将对抗性训练扩展到 TC [36]的多任务学习框架，旨在减轻任务无关（共享）和任务相关（私有）潜在特征空间的相互干扰。

*强化学习。*强化学习 (RL) [176] 是一种训练代理根据策略执行离散动作的方法，该策略被训练以最大化奖励。沉等人。[177] 使用硬注意力模型为 TC 选择输入序列的关键词标记的子集。硬注意力模型可以看作是一个代理，它采取是否选择令牌的动作。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。刘等人。[178] 提出了一种将 TC 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，并在它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，由使用 RL 训练的策略控制。沉等人。[179] 提出了一个用于机器阅读理解的多步推理网络 (ReasoNet)。ReasoNets 执行多个步骤来推理查询、文档和答案之间的关系。ReasoNets 不是在推理过程中使用固定数量的步骤，而是引入了终止状态来放松对推理步骤的限制。通过使用 RL，ReasoNets 可以动态地确定是在消化中间结果后继续理解过程，还是在得出结论现有信息足以产生答案时终止阅读。李等人。[180] 结合 RL、GAN 和 RNN 构建一个新模型，称为类别句子生成对抗网络 (CS-GAN)，该模型能够生成扩大原始数据集的类别句子并在监督训练期间提高其泛化能力。张等人。[181] 提出了一种基于 RL 的方法来学习用于文本分类的结构化表示。他们提出了两个基于 LSTM 的模型。第一个只选择输入文本中重要的、与任务相关的单词。另一个发现句子的短语结构。使用这两个模型的

结构发现被制定为由策略网络引导的顺序决策过程，策略网络在每个步骤中决定使用哪个模型，如图 21 所示。使用策略梯度优化策略网络。

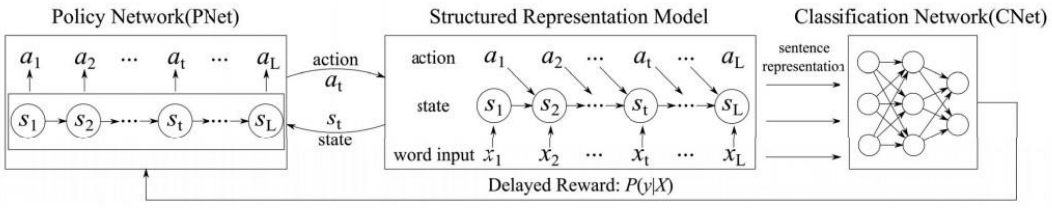

图 21. 基于 RL 的文本分类结构化表示学习方法 [181]。策略网络对每个状态的动作进行采样。结构化表示模型更新状态，并在剧集结束时将最终的句子表示输出到分类网络。文本分类损失用作训练策略的（负）奖励。

作为本节的总结，图 22 展示了自 2013 年以来一些最流行的基于 DL 的 TC 模型的时间线。

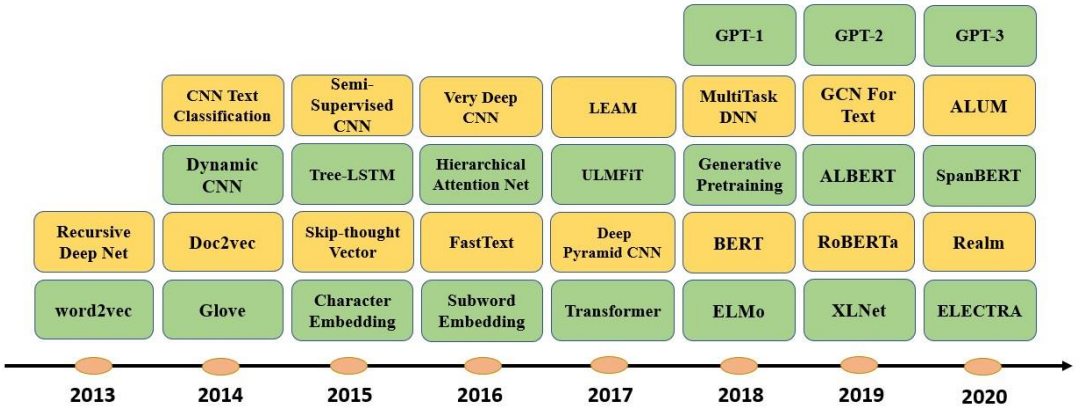

图 22. 2013 年至 2020 年发布的一些最突出的文本嵌入和分类深度学习模型。

3　如何为我的任务选择最佳的神经网络模型

回答“什么是 TC 的最佳神经网络架构？”取决于目标任务和域的性质、域内标签的可用性、应用程序的延迟和容量限制等，差异很大。尽管毫无疑问，开发文本分类器是一个反复试验的过程，但通过分析公共基准（例如，GLUE [22]）的最新结果，我们提出了以下方法来简化该过程。配方包括五个步骤：

(1) PLM 选择。如第 5 节所示，使用 PLM 可以显着改进所有流行的文本分类任务，并且自动编码 PLM（例如，BERT 或 RoBERTa）通常比自回归 PLM（例如，OpenAI GPT）工作得更好。Hugging Face [5][3] 维护着为各种任务和设置开发的丰富的 PLM 存储库。[6]

(2) 领域适应。大多数 PLM 都在通用域文本语料库（例如，Web）上进行训练。如果目标域与一般域显着不同，我们可能会考虑通过持续预训练选定的一般域 PLM 来使用域内数据来调整 PLM。对于具有大量未标记文本的领域，例如生物医学，从头开始预训练语言模型也可能是一个不错的选择 [182]。

(3) 特定任务模型设计。给定输入文本，PLM 在上下文表示中生成向量序列。然后，在顶部添加一个或多个特定于任务的层，以生成目标任务的最终输出。任务特定层架构的选择取决于任务的性质，例如，需要捕获文本的语言结构。如第 2 节所述，前馈神经网络将文本视为一个词袋，RNN 可以捕获词序，CNN 擅长识别关键短语等模式，注意力机制可以有效识别文本中的相关词，Siamese NN 是用于文本匹配任务，如果自然语言的图结构（例如解析树）对目标任务有用，GNN 可能是一个不错的选择。

(4) 特定任务微调。根据域内标签的可用性，任务特定层可以单独使用固定 PLM 进行训练，也可以与 PLM 一起训练。如果需要构建多个相似的文本分类器（例如，不同域的新闻分类器），多任务微调 [23] 是利用相似域的标记数据的好选择。

(5) 模型压缩。PLM 的服务成本很高。它们通常需要通过例如知识蒸馏[183、184]进行压缩， 以满足实际应用程序中的延迟和容量限制。

4 文本分类数据集

本节介绍了广泛用于 TC 研究的数据集。我们根据它们的主要目标应用程序将这些数据集分组为情感分析、新闻分类、主题分类、QA 和 NLI 等类别。

4.1 情感分析数据集

喊叫。 Yelp [185] 数据集包含两个情绪分类任务的数据。一种是检测细粒度的情感标签，称为 Yelp-5。另一个预测负面和正面情绪，被称为 Yelp Review Polarity 或 Yelp-2。Yelp-5 每个类有 650,000 个训练样本和 50,000 个测试样本，Yelp-2 包括 560,000 个训练样本和 38,000 个负类和正类的测试样本。

[5] https://huggingface.co/

IMDB。IMDB 数据集 [186] 是为电影评论的二元情感分类任务而开发的。IMDB 包含相同数量的正面和负面评论。它在训练集和测试集之间平均分配，每个都有 25,000 条评论。

电影评论。电影评论 (MR) 数据集 [187] 是为检测与特定评论相关的情绪并确定其是负面还是正面的任务而开发的电影评论集合。它包括 10,662 个带有偶数个负样本和正样本的句子。带有随机分割的 10 倍交叉验证通常用于在此数据集上进行测试。

不锈钢。斯坦福情绪树库 (SST) 数据集 [43] 是 MR 的扩展版本。有两个版本可用，一个带有细粒度标签（五类），另一个带有二进制标签，分别称为 SST-1 和 SST2。SST-1 由 11,855 条电影评论组成，分为 8,544 个训练样本、1,101 个开发样本和 2,210 个测试样本。SST-2 被分成三个大小分别为 6,920、872 和 1,821 的集合作为训练集、开发集和测试集。

MPQA。多视角问答 (MPQA) 数据集 [188] 是具有两个类别标签的意见语料库。MPQA 包含从与各种新闻来源相关的新闻文章中提取的 10,606 个句子。这是一个不平衡的数据集，包含 3,311 个正面文档和 7,293 个负面文档。

亚马逊。这是从亚马逊网站 [189]收集的一个流行的产品评论语料库。它包含二进制分类和多类（5 类）分类的标签。亚马逊二元分类数据集分别包含 3,600,000 和 400,000 条用于训练和测试的评论。Amazon 5 类分类数据集 (Amazon-5) 分别包含 3,000,000 和 650,000 条用于训练和测试的评论。

4.2 新闻分类数据集

AG 新闻。AG 新闻数据集 [50] 是学术新闻搜索引擎 ComeToMyHead 从 2,000 多个新闻来源收集的新闻文章集合。该数据集包括 120,000 个训练样本和 7,600 个测试样本。每个样本都是一个带有四类标签的短文本。

20 个新闻组。20 个新闻组数据集 [190] 是发布在 20 个不同主题上的新闻组文档的集合。该数据集的各种版本用于文本分类、文本聚类等。最受欢迎的版本之一包含 18,821 个文档，这些文档均匀地分类在所有主题中。

搜狗新闻。搜狗新闻数据集 [154] 是 SogouCA 和 SogouCS 新闻语料库的混合体。新闻的分类标签由其在 URL 中的域名决定。例如，URL 为 http://sports.sohu.com 的新闻被归类为体育类。

路透社消息。Reuters-21578 数据集 [191] 是最广泛使用的文本分类数据集之一，于 1987 年从路透社金融新闻专线服务收集。ApteMod 是 Reuters-21578 的多类版本，包含 10,788 个文档。它有 90 个课程，

7,769 个培训文档和 3,019 个测试文档。从路透社数据集的子集派生的其他数据集包括 R8、R52、RCV1 和 RCV1-v2。

为新闻分类开发的其他数据集包括：必应新闻 [192]、BBC [193]、谷歌新闻 [194]。

4.3 主题分类数据集

*数据库百科。*DBpedia 数据集 [195] 是一个大规模的多语言知识库，由维基百科中最常用的信息框创建。DBpedia 每月发布一次，每个版本中都会添加或删除一些类和属性。最受欢迎的 DBpedia 版本包含 560,000 个训练样本和 70,000 个测试样本，每个都有 14 类标签。

*哦苏梅德。*Ohsumed 集合 [196] 是 MEDLINE 数据库的一个子集。Ohsumed 包含 7,400 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。

*欧元法。*EUR-Lex 数据集 [197] 包括不同类型的文档，这些文档根据几个正交分类方案进行索引，以允许使用多种搜索工具。该数据集最受欢迎的版本基于欧盟法律的不同方面，包含 19,314 个文档和 3,956 个类别。

哇。 Web Of Science (WOS) 数据集 [136] 是 Web of Science 中已发表论文的数据和元数据的集合，Web of Science 是世界上最受信任的独立于出版商的全球引文数据库。WOS 已经发布了三个版本：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。

*考研。*PubMed [198] 是美国国家医学图书馆为医学和生物科学论文开发的搜索引擎，其中包含一个文档集。每个文档都标有 MeSH 集的类别，MeSH 集是 PubMed 中使用的标签集。摘要中的每个句子都使用以下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 [199]、Irony（由来自社交新闻网站 reddit 的注释评论、用于推文主题分类的 Twitter 数据集、arXiv 集合）[200] 等等。

4.4 质量保证数据集

队。 斯坦福问答数据集 (SQuAD) [24] 是来自维基百科文章的问答对集合。在 SQuAD 中，问题的正确答案可以是给定文本中的任何标记序列。由于问题和答案是由人类通过众包产生的，因此它比其他一些问答数据集更加多样化。SQuAD 1.1 包含 536 篇文章的 107,785 个问答对。最新版本 SQuAD2.0 结合了

SQuAD1.1 中的 100,000 个问题和超过 50,000 个由众包工作者以与可回答问题相似的形式以对抗方式编写的无法回答的问题 [201]。

*马可女士。*该数据集由 Microsoft [202] 发布。与 SQuAD 不同，所有问题都是由编辑产生的；在 MS MARCO 中，所有问题都是使用 Bing 搜索引擎从用户查询和真实 Web 文档的段落中抽取的。MS MARCO 中的一些答案是生成性的。因此，该数据集可用于开发生成 QA 系统。

*TREC-QA。*TREC-QA [203] 是 QA 研究中最受欢迎和研究最多的数据集之一。该数据集有两个版本，称为 TREC-6 和 TREC-50。TREC-6 由 6 个类别的问题组成，而 TREC-50 由 50 个类别组成。对于这两个版本，训练和测试数据集分别包含 5,452 和 500 个问题。

*维基问答。*WikiQA 数据集 [204] 由一组问答对组成，为开放域 QA 研究收集和注释。该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。

*知乎。*Quora 数据集 [205] 是为释义识别（检测重复问题）而开发的。为此，作者提供了 Quora 数据的一个子集，其中包含超过 400,000 个问题对。为每个问题对分配一个二进制值，指示这两个问题是否相同。

用于 QA 的其他数据集包括 Situations With Adversarial Generations (SWAG) [206]、WikiQA [204]、SelQA [207]。

4.5 NLI 数据集

SNLI。 斯坦福自然语言推理 (SNLI) 数据集 [208] 广泛用于 NLI。该数据集由 550,152、10,000 和 10,000 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之一进行注释：中性、蕴涵、矛盾。

多 NLI。 Multi-Genre Natural Language Inference (MNLI) 数据集 [209] 是 433k 句对的集合，并带有文本蕴涵标签。语料库是 SNLI 的扩展，涵盖了更广泛的口语和书面文本流派，并支持独特的跨流派泛化评估。

生病的。 涉及组合知识的句子 (SICK) 数据集 [25] 由大约 10,000 个英语句子对组成，这些句子对使用三个标签进行注释：蕴含、矛盾和中性。

建议零售价。 Microsoft Research Paraphrase (MSRP) 数据集 [210] 通常用于文本相似性任务。MSRP 包含 4,076 个用于训练的样本和 1,725 个用于测试的样本。每个样本是一个句子对，用一个二进制标签进行注释，指示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) [211]、RTE [212]、SciTail [213] 等等。

5 实验性能分析

在本节中，我们首先描述一组通常用于评估 TC 模型性能的指标，然后对一组基于 DL 的 TC 模型在流行基准上的性能进行定量分析。

5.1 文本分类的流行指标

*准确性和错误率。*这些是评估分类模型质量的主要指标。让 TP，FP、TN、FN 分别表示真阳性、假阳性、真阴性和假阴性。分类准确性和错误率在方程式中定义。2

$$\text{准确率} = \frac{(TP + TN)}{N}, \text{错误率} = \frac{(FP + FN)}{N} \quad (2)$$

其中N是样本总数。显然，我们的错误率 = 1 - 准确度。

*精度/召回/F1 分数。*这些也是主要指标，并且比准确性或错误率更常用于不平衡测试集，例如，大多数测试样本具有一个类别标签。二元分类的精度和召回率定义为方程式。3 . F1 分数是准确率和召回率的调和平均值，如等式所示。3 . F1 分数在 1 处达到其最佳值（完美的精度和召回率），在 0 处达到最差。

$$\text{Precision} = \frac{TP}{TP + FP}, \text{Recall} = \frac{TP}{TP + FN}, \text{F1-score} = \frac{2\,\text{Prec}\,\text{Rec}}{\text{Prec} + \text{Rec}} \quad (3)$$

对于多类分类问题，我们总是可以计算每个类标签的准确率和召回率，并分析类标签上的各个性能或平均这些值以获得整体准确率和召回率。

*精确匹配 (EM)。*精确匹配指标是问答系统的一种流行指标，它衡量与任何一个基本事实答案完全匹配的预测的百分比。EM 是用于 SQuAD 的主要指标之一。

*平均倒数秩 (MRR)。*MRR 通常用于评估排序算法在 NLP 任务中的性能，例如查询文档排序和 QA。MRR 在方程式中定义。4，其中Q是所有可能答案的集合，$rank_i$是真实答案的排名位置。

$$MR = \frac{1}{|Q|}\sum_{i=1}^{Q}\frac{1}{rank_i} \quad (4)$$

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

5.2 定量结果

我们将前面讨论的几种算法在流行的 TC 基准上的性能制成表格。在每个表格中，除了一组具有代表性的 DL 模型的结果外，我们还使用非深度学习模型展示了结果，这些模型要么是先前的技术水平，要么在 DL 时代之前被广泛用作基线。我们可以看到，在所有这些任务中，使用 DL 模型带来了显着的改进。

表 1 总结了第 2 节中描述的模型在几个情感分析数据集上的结果，包括 Yelp、IMDB、SST 和 Amazon。我们可以看到，自从引入第一个基于 DL 的情感分析模型以来，准确度得到了显着提高，例如，分类错误相对减少了大约 78%（在 SST-2 上）。

表 2 报告了三个新闻分类数据集（即 AG News、20-NEWS、搜狗新闻）和两个主题分类数据集（即 DBpedia 和 Ohsummed）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 DL 模型在 SQuAD 和 WikiQA 上的性能。值得注意的是，在这两个数据集上，显着的性能提升都归功于 BERT 的使用。

表 5 展示了两个 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稳步提升。

6 挑战与机遇

在 DL 模型的帮助下，TC 在过去几年中取得了长足的进步。已经提出了几个新颖的想法（例如神经嵌入、注意力机制、自我注意力、Transformer、BERT 和 XLNet），这些想法导致了过去十年的快速发展。尽管取得了进展，但仍有挑战需要解决。本节介绍了其中的一些挑战，并讨论了有助于推动该领域发展的研究方向。

*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 TC 任务收集了许多大规模数据集，但对于更具挑战性的 TC 任务，例如具有多步推理的 QA、多语言文档的文本分类和 TC，仍然需要新的数据集。对于极长的文档。

表 1. 基于深度学习的文本分类模型在情感分析数据集上的准确度（在分类准确度方面），在 IMDB、SST、Yelp 和 Amazon 数据集上进行了评估。斜体表示非深度学习模型。

方法	数据库	SST-2	亚马逊-2	亚马逊-5	Yelp-2	Yelp-5
朴素贝叶斯[43]	-	81.80	-	-	-	-
LDA [214]	67.40	-	-	-	-	-
弓+支持向量机[31]	87.80	-	-	-	-	-
tf。Δ idf [215]	88.10	-	-	-	-	-

字符级 CNN [50]	-	-	94.49	59.46	95.12	62.05
深度金字塔 CNN [49]	-	84.46	96.68	65.82	97.36	69.40
ULMFiT [216]	95.40	-	-	-	97.84	70.02
BLSTM-2DCNN [40]	-	89.50	-	-	-	-
神经语义编码器 [95]	-	89.70	-	-	-	-
BCN+Char+CoVe [217]	91.80	90.30	-	-	-	-
胶水 ELMo 基线 [22]	-	90.40	-	-	-	-
BERT ELMo 基线 [7]	-	90.40	-	-	-	-
CCCapsNet [76]	-	-	94.96	60.95	96.48	65.85
虚拟对抗训练 [173]	94.10	-	-	-	-	-
块稀疏 LSTM [218]	94.99	93.20	-	-	96.73	
基于 BERT 的[7、154]	95.63	93.50	96.04	61.60	98.08	70.58
BERT-大 [7 , 154]	95.79	94.9	96.07	62.20	98.19	71.38
阿尔伯特 [147]	-	95.20	-	-	-	-
多任务 DNN [23]	83.20	95.60	-	-	-	-
通气管金属 [219]	-	96.20	-	-	-	-
BERT 微调 + UDA [220]	95.80		96.50	62.88	97.95	62.92
RoBERTa（+附加数据）[146]	-	96.40	-	-	-	-
XLNet-Large（合奏）[156]	96.21	96.80	97.60	67.74	98.45	72.20

建模常识知识。 将常识知识融入 DL 模型有可能显着提高模型性能，这与人类利用常识知识执行不同任务的方式几乎相同。例如，配备常识知识库的 QA 系统可以回答有关现实世界的问题。在信息不完整的情况下，常识性知识也有助于解决问题。使用对日常对象或概念的广泛持有的信念，人工智能系统可以以与人类类似的方式基于对未知数的“默认”假设进行推理。虽然这个想法已经被研究用于情感分类 [?]，需要更多的研究来探索在 DL 模型中有效地建模和使用常识知识。

可解释的 DL 模型。 虽然 DL 模型在具有挑战性的基准测试中取得了可喜的性能，但这些模型中的大多数是不可解释的。例如，为什么一个模型在一个数据集上的表现优于另一个模型，但在其他数据集上表现不佳？DL 模型究竟学到了什么？什么是可以在给定数据集上达到一定精度的最小神经网络架构？尽管注意力和自注意力机制为回答这些问题提供了一些见解，但仍然缺乏对这些模型的潜在行为和动态的详细研究。更好地理解这些模型的理论方面可以帮助开发针对各种文本分析场景的更好的模型。

表 2. 分类模型对新闻分类和主题分类任务的准确性。斜体表示非深度学习模型。

	新闻分类			话题分类	
方法	AG 新闻	20 新闻	搜狗新闻	数据库百科	大须
分层对数双线性模型[221]	-	-	-	-	52
文本 GCN [107]	67.61	86.34	-	-	68.36
简化的 GCN [108]	-	88.50	-	-	68.50
字符级 CNN [50]	90.49	-	95.12	98.45	-
CCCapsNet [76]	92.39	-	97.25	98.72	-
学习 [84]	92.45	81.91	-	99.02	58.58
快速文本 [30]	92.50	-	96.80	98.60	55.70
胶囊网 B [71]	92.60	-	-	-	-
深度金字塔 CNN [49]	93.13	-	98.16	99.12	-
ULMFiT [216]	94.99	-	-	99.20	-
L 混装 [174]	95.05	-	-	99.30	-
BERT-大 [220]	-	-	-	99.32	-
XLNet [156]	95.51	-	-	99.38	-

表 3. SQuAD 问答数据集上分类模型的性能。在这里，F1 分数衡量了预测和真实答案之间的平均重叠。斜体表示非深度学习模型。

	小队 1.1		小队 2.0	
方法	电磁场	F1 分数	电磁场	F1 分数
滑动窗口+距离。 [222]	13.00	20.00	-	-
手工特征+逻辑回归[24]	40.40	51.00	-	-
BiDAF + 自注意力 + ELMo [4]	78.58	85.83	63.37	66.25
SAN（单一型号）[137]	76.82	84.39	68.65	71.43
FusionNet++（合奏）[223]	78.97	86.01	70.30	72.48
SAN（合奏）[137]	79.60	86.49	71.31	73.70
BERT（单一模型）[7]	85.08	91.83	80.00	83.06

BERT-large（集合）[7]	87.43	93.16	80.45	83.51
BERT + 多卷积神经网络 [137]	-	-	84.20	86.76
XL-Net [156]	89.90	95.08	84.64	88.00
斯潘伯特 [149]	88.83	94.63	71.31	73.70
罗伯塔 [146]	-	-	86.82	89.79
ALBERT（单一模型）[147]	-	-	88.10	90.90
阿尔伯特（合奏）[147]	-	-	89.73	92.21
阿尔伯特的复古阅读器	-	-	90.11	92.58
ELECTRA+ALBERT+EntitySpanFocus	-	-	90.42	92.79

*内存高效模型。*大多数现代神经语言模型都需要大量内存来进行训练和推理。为了满足边缘应用程序的计算和存储限制，必须压缩这些模型。这可以通过使用知识蒸馏构建学生模型来完成，

表 4. WikiQA 数据集上分类模型的性能。

方法	地图	MRR
段落向量 [32]	0.511	0.516
神经变分推理 [166]	0.655	0.674
细心的汇集网络 [83]	0.688	0.695
超级质量保证 [127]	0.712	0.727
BERT（单一模型）[7]	0.813	0.828
坦达-罗伯塔 [153]	0.920	0.933

表 5. 自然语言推理数据集上分类模型的性能。对于 Multi-NLI，Matched 和 Mismatched 分别指的是匹配和不匹配的测试精度。斜体表示非深度学习模型。

	SNLI	多 NLI	
方法	准确性	匹配	不匹配
Unigrams 特征[208]	71.6	-	-
词汇化[208]	78.2	-	-
LSTM 编码器 (100D) [208]	77.6	-	-
基于树的 CNN [61]	82.1	-	-
biLSTM 编码器 [209]	81.5	67.5	67.1
神经语义编码器 (300D) [95]	84.6	-	-

基于 RNN 的句子编码器 [224]	85.5	73.2	73.6
迪桑 (300D) [81]	85.6	-	-
可分解注意力模型 [92]	86.3	-	-
强化自注意力（300D）[177]	86.3	-	-
广义池化 (600D) [93]	86.6	73.8	74.0
双边多视角匹配[41]	87.5	-	-
多路注意力网络 [87]	88.3	78.5	77.7
ESIM + ELMo [4]	88.7	72.9	73.4
带有强化学习的 DMAN [225]	88.8	88.8	78.9
BiLSTM + ELMo + Attn [22]	-	74.1	74.5
微调 LM 预训练变压器 [6]	89.9	82.1	81.4
多任务 DNN [23]	91.6	86.7	86.0
森伯特 [155]	91.9	84.4	84.0
罗伯塔 [146]	92.6	90.8	90.2
XLNet [156]	-	90.2	89.8

或通过使用模型压缩技术。开发与任务无关的模型压缩方法是一个活跃的研究课题 [226]。

*少样本和零样本学习。*大多数 DL 模型都是需要大量域标签的监督模型。实际上，为每个新域收集此类标签的成本很高。与从头开始训练模型相比，将 PLM（例如，BERT 和 OpenGPT）微调到特定任务所需的域标签要少得多，从而为开发基于 PLM 的新的零样本或少样本学习方法提供了机会。

7 结论

在本文中，我们调查了 150 多个 DL 模型，这些模型是在过去六年中开发的，并且显着提高了各种 TC 任务的最新技术水平。我们还概述了 40 多个流行的 TC 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。

致谢

作者要感谢 Richard Socher、Kristina Toutanova、Brooke Cowan 和所有匿名审稿人审阅了这项工作并提供了非常有见地的评论。

参考

[1] S. Deerwester、ST Dumais、GW Furnas、TK Landauer 和 R. Harshman，“通过潜在语义分析进行索引”，*美国信息科学学会杂志*，第一卷。41，没有。6，第 391-407 页，1990 年。

[2] Y. Bengio、R. Ducharme、P. Vincent 和 C. Jauvin，“神经概率语言模型”，*机器学习研究杂志*，第一卷。3，没有。2003 年 2 月，第 1137-1155 页。

[3] T. Mikolov、I. Sutskever、K. Chen、GS Corrado 和 J. Dean，“单词和短语的分布式表示及其组合性”，《*神经信息处理系统进展*》，2013 年，第 3111-3119 页.

[4] ME Peters、M. Neumann、M. Iyyer、M. Gardner、C. Clark、K. Lee 和 L. Zettlemoyer，“深度上下文化的词表示”，*arXiv 预印本 arXiv:1802.05365，2018*。

[5] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、AN Gomez、Ł。Kaiser 和 I. Polosukhin，“Attention is all you need”，*神经信息处理系统进展*，2017 年，第 5998-6008 页。

[6] A. Radford、K. Narasimhan、T. Salimans 和 I. Sutskever，“通过生成式预训练提高语言理解能力”，*网址 https://s3-us-west-2。亚马逊。com/openai-assets/researchcovers/languageunsupervised/语言理解论文。pdf*，2018 年。 [7] J. Devlin，M.-W。Chang、K. Lee 和 K. Toutanova，“Bert：用于语言理解的深度双向转换器的预训练”，*arXiv 预印本 arXiv:1810.04805，2018* 年。

[8] TB Brown、B. Mann、N. Ryder、M. Subbiah、J. Kaplan、P. Dhariwal、A. Neelakantan、P. Shyam、G. Sastry、A. Askell *等人。*，“语言模型是少数人的学习者”，*arXiv 预印本 arXiv:2005.14165，2020*。

[9] D. Lepikhin、H. Lee、Y. Xu、D. Chen、O. Firat、Y. Huang、M. Krikun、N. Shazeer 和 Z. Chen，“Gshard：使用条件计算和自动分片，” *arXiv 预印本 arXiv:2006.16668* , 2020。

[10] G. Marcus 和 E. Davis，*重启 AI：构建我们可以信任的人工智能。* 万神殿，2019 年。

[11] G. Marcus，“人工智能的下一个十年：迈向强大人工智能的四个步骤”，*arXiv 预印本 arXiv:2002.06177，2020* 年。

[12] Y. Nie、A. Williams、E. Dinan、M. Bansal、J. Weston 和 D. Kiela，“对抗性 nli：自然语言理解的新基准”，*arXiv 预印本 arXiv:1910.14599，2019*。 13] D. Jin、Z. Jin、JT Zhou 和 P. Szolovits，“bert 真的很健壮吗？对文本分类和蕴涵的自然语言攻击，” *arXiv 预印本 arXiv:1907.11932*，第一卷。2019 年 2 月。

[14] X. Liu、H. Cheng、P. He、W. Chen、Y. Wang、H. Poon 和 J. Gao，“大型神经语言模型的对抗性训练”，*arXiv 预印本 arXiv:2004.08994，2020*。

[15] J. Andreas、M. Rohrbach、T. Darrell 和 D. Klein，“学习构建用于问答的神经网络”，*arXiv 预印本 arXiv:1601.01705，2016* 年。

[16] M. Iyyer，W.-t。Yih 和 M.-W。Chang，“用于顺序问答的基于搜索的神经结构学习”，*在第 55 届计算语言学协会年会论文集（第 1 卷：长篇论文）*，2017 年，第 1821-1831 页。

[17] I. Schlag、P. Smolensky、R. Fernandez、N. Jojic、J. Schmidhuber 和 J. Gao，“通过显式关系编码增强变压器以解决数学问题”，*arXiv 预印本 arXiv:1910.06611，2019* 年。

[18] J. Gao、B. Peng、C. Li、J. Li、S. Shayandeh、L. Liden 和 H.-Y。Shum，“具有扎根文本生成的强大对话式人工智能”，*arXiv 预印本 arXiv:2009.03457，2020* 年。

[19] K. Kowsari、K. Jafari Meimandi、M. Heidarysafa、S. Mendu、L. Barnes 和 D. Brown，“文本分类算法：调查”，*信息*，卷。10，没有。第 4 页 150，2019。

[20] CD Manning、H. Schütze 和 P. Raghavan，*信息检索简介*。 剑桥大学出版社，2008 年。

[21] D. Jurasky 和 JH Martin，“语音和语言处理：自然语言处理简介”，*计算语言学和语音识别。新泽西州普伦蒂斯霍尔*，2008 年。

[22] A. Wang、A. Singh、J. Michael、F. Hill、O. Levy 和 SR Bowman，“Glue：用于自然语言理解的多任务基准和分析平台”， *arXiv 预印本 arXiv:1804.07461*， 2018 年。

[23] X. Liu、P. He、W. Chen 和 J. Gao，“用于自然语言理解的多任务深度神经网络”，*arXiv:1901.11504，2019*。 [24] P. Rajpurkar，J. Zhang， K. Lopyrev 和 P. Liang，“Squad：机器理解文本的 100,000+ 个问题”，*arXiv 预印本 arXiv:1606.05250，2016* 年。

[25] M. Marelli、L. Bentivogli、M. Baroni、R. Bernardi、S. Menini 和 R. Zamparelli，“Semeval-2014 任务 1：通过语义相关性和文本蕴涵评估完整句子的组合分布语义模型,”在*第 8 届语义评估国际研讨会论文集 (SemEval 2014)*中，2014 年，第 1-8 页。

[26] I. Goodfellow、Y. Bengio 和 A. Courville，*深度学习*。 麻省理工学院出版社，2016 年。

[27] T. Mikolov、K. Chen、G. Corrado 和 J. Dean，“向量空间中单词表示的有效估计”，*arXiv 预印本 arXiv:1301.3781*，2013 年。

[28] J. Pennington、R. Socher 和 C. Manning，“Glove: Global vectors for word representation”，*2014 年自然语言处理经验方法会议论文集 (EMNLP)*，2014 年，第 1532-1543 页.

[29] M. Iyyer、V. Manjunatha、J. Boyd-Graber 和 H. Daumé III，“深度无序组合与文本分类的句法方法相媲美”，*第 53 届计算语言学协会年会论文集和第七届自然语言处理国际联合会议（第 1 卷：长篇论文）*，2015，第 1681-1691 页。

[30] A. Joulin、E. Grave、P. Bojanowski、M. Douze、H. Jégou 和 T. Mikolov，“Fasttext。zip：压缩文本分类模型，” *arXiv 预印本 arXiv:1612.03651，2016* 年。

[31] S. Wang 和 CD Manning，“基线和二元组：简单、良好的情感和主题分类”，*计算语言学协会第 50 届年会论文集：短篇论文第 2 卷*。计算语言学协会，2012，第。90-94。

[32] Q. Le 和 T. Mikolov，“句子和文档的分布式表示”， *机器学习国际会议*，2014 年，第 1188-1196 页。

[33] KS Tai、R. Socher 和 CD Manning，“改进的树结构长短期记忆网络的语义表示”，*arXiv 预印本 arXiv:1503.00075，2015* 年。

[34] X. Zhu、P. Sobihani 和 H. Guo，“递归结构上的长短期记忆”，*国际机器学习会议*，2015，第 1604-1612 页。

[35] J. Cheng、L. Dong 和 M. Lapata，“机器阅读的长短期记忆网络”，*arXiv 预印本 arXiv:1601.06733，2016* 年。

[36] P. Liu、X. Qiu、X. Chen、S. Wu 和 X.-J。Huang，“用于建模句子和文档的多时间尺度长短期记忆神经网络”，*2015 年自然语言处理经验方法会议论文集*，2015 年，第 2326-2335 页。

[37] AB Dieng、C. Wang、J. Gao 和 J. Paisley，“Topicrnn：具有远程语义依赖的循环神经网络”，*arXiv 预印本 arXiv:1611.01702，2016* 年。

[38] P. Liu、X. Qiu 和 X. Huang，“用于多任务学习的文本分类的循环神经网络”，*arXiv 预印本 arXiv:1605.05101*，2016 年。

[39] R. Johnson 和 T. Zhang，“使用 lstm 进行区域嵌入的监督和半监督文本分类”，*arXiv 预印本 arXiv:1602.02373，2016* 年。

[40] P. Zhou、Z. Qi、S. Zheng、J. Xu、H. Bao 和 B. Xu，“通过将双向 lstm 与二维最大池化相结合来改进文本分类”，*arXiv 预印本 arXiv:1611.06639*， 2016 年。

[41] Z. Wang、W. Hamza 和 R. Florian，“自然语言句子的双边多视角匹配”，*arXiv 预印本 arXiv:1702.03814*，2017 年。

[42] S. Wan, Y. Lan, J. Guo, J. Xu, L. Pang, and X. Cheng，“A deep architecture for semantic matching with multiple positional sentence representations”，*第三十届 AAAI 人工智能会议*， 2016 年。

[43] R. Socher、A. Perelygin、J. Wu、J. Chuang、CD Manning、AY Ng 和 C. Potts，“情感树库上语义组合性的递归深度模型”，*2013 年会议论文集自然语言处理中的经验方法*，2013 年，第 1631-1642 页。

[44] Y. LeCun、L. Bottou、Y. Bengio 和 P. Haffner，“基于梯度的学习应用于文档识别”，*IEEE 会议记录*，第一卷。86，没有。11，第 2278-2324 页，1998 年。

[45] N. Kalchbrenner、E. Grefenstette 和 P. Blunsom，“用于建模句子的卷积神经网络”，*计算语言学协会第 52 届年会，ACL 2014 - 会议论文集*，2014 年。

[46] Y. Kim，“用于句子分类的卷积神经网络”，*EMNLP 2014 - 2014 自然语言处理经验方法会议，会议论文集*，2014。

[47] J. Liu、WC Chang、Y. Wu 和 Y. Yang，“深度学习用于极端多标签文本分类”，*SIGIR 2017 - 第 40 届国际 ACM SIGIR 信息检索研究与开发会议论文集*, 2017 年。

[48] R. Johnson 和 T. Zhang，“使用卷积神经网络有效地使用词序进行文本分类”，*NAACL HLT 2015 - 2015 年计算语言学协会北美分会会议：人类语言技术，论文集会议*，2015 年。

[49] ——，“用于文本分类的深度金字塔卷积神经网络”，*载于第 55 届计算语言学协会年会论文集（第 1 卷：长篇论文)*，2017 年，第 562-570 页。

[50] X. Zhang、J. Zhao 和 Y. LeCun，“用于文本分类的字符级卷积网络”，《*神经信息处理系统进展*》，2015 年，第 649-657 页。

[51] Y. Kim、Y. Jernite、D. Sontag 和 AM Rush，“字符感知神经语言模型”，*第 30 届 AAAI 人工智能会议*，2016 年。

[52] JD Prusa 和 TM Khoshgoftaar，“为深度神经网络和文本分类设计更好的数据表示”，载于 *Proceedings - 2016 IEEE 第 17 届信息重用和集成国际会议，IRI 2016，2016*。

[53] K. Simonyan 和 A. Zisserman，“用于大规模图像识别的非常深的卷积网络”，*第三届国际学习表示会议，ICLR 2015 - Conference Track Proceedings*，2015 年。

[54] K. He、X. Zhang、S. Ren 和 J. Sun，“用于图像识别的深度残差学习”，*IEEE 计算机学会计算机视觉和模式识别会议论文集*，2016 年。

[55] A. Conneau、H. Schwenk、L. Barrault 和 Y. Lecun，“用于文本分类的非常深的卷积网络”，*arXiv 预印本 arXiv:1606.01781，2016* 年。

[56] AB Duque、LLJ Santos、D. Macêdo 和 C. Zanchettin，“用于文本分类的压缩非常深的卷积神经网络”，*计算机科学讲义（包括人工智能子系列讲义和生物信息学讲义）*，2019 年。

[57] HT Le、C. Cerisara 和 A. Denis，“卷积网络是否需要深度才能进行文本分类？”在 2018 年*第三十二届 AAAI 人工智能会议的研讨会上。*

[58] G. Huang、Z. Liu、L. Van Der Maaten 和 KQ Weinberger，“密集连接的卷积网络”，在 *Proceedings - 第 30 届 IEEE 计算机视觉和模式识别会议上，CVPR 2017，2017*。

[59] B. Guo、C. Zhang、J. Liu 和 X. Ma，“通过多通道 TextCNN 模型使用加权词嵌入改进文本分类”，*神经计算*，2019 年。

[60] Y. Zhang 和 B. Wallace，“用于句子分类的卷积神经网络（和从业者指南）的敏感性分析”，*arXiv 预印本 arXiv：1510.03820，2015* 年。 [61] L. Mou，R. Men， G. Li、Y. Xu、L. Zhang、R. Yan 和 Z. Jin，“基于树的卷积和启发式匹配的自然语言推理”，*arXiv 预印本 arXiv:1512.08422，2015* 年。

[62] L. Pang, Y. Lan, J. Guo, J. Xu, S. Wan, and X. Cheng，“文本匹配作为图像识别”，*第 30 届 AAAI 人工智能会议，AAAI 2016，2016*。

[63] J. Wang、Z. Wang、D. Zhang 和 J. Yan，“结合知识与深度卷积神经网络进行短文本分类”，*IJCAI 国际人工智能联合会议*，2017。

[64] S. Karimi、X. Dai、H. Hassanzadeh 和 A. Nguyen，“放射学报告的自动诊断编码：深度学习和传统分类方法的比较”，2017 年。

[65] S. Peng、R. You、H. Wang、C. Zhai、H. Mamitsuka 和 S. Zhu，“DeepMeSH：用于改进大规模 MeSH 索引的深度语义表示”，*生物信息学*，2016 年。

[66] A. Rios 和 R. Kavuluru，“用于生物医学文本分类的卷积神经网络：在索引生物医学文章中的应用”，*BCB 2015 - 第 6 届 ACM 生物信息学、计算生物学和健康信息学会议*，2015 年。

[67] M. Hughes、I. Li、S. Kotoulas 和 T. Suzumura，“使用卷积神经网络的医学文本分类”，*健康技术和信息学研究*，2017 年。

[68] GE Hinton、A. Krizhevsky 和 SD Wang，“转换自动编码器”，*人工神经网络国际会议*。
施普林格，2011，第 44-51 页。

[69] S. Sabour、N. Frost 和 GE Hinton，“胶囊之间的动态路由”，*神经信息处理系统进展*，2017 年，第 3856-3866 页。

[70] S. Sabour、N. Frosst 和 G. Hinton，“采用 em 路由的矩阵胶囊”，*第 6 届国际学习表示会议，ICLR*，
2018 年，第 1-15 页。

[71] W. Zhao、J. Ye、M. Yang、Z. Lei、S. Zhang 和 Z. Zhao，“使用动态路由研究用于文本分类的胶囊网络”，*arXiv 预印本 arXiv:1804.00538，2018* 年。

[72] M. Yang、W. Zhao、L. Chen、Q. Qu、Z. Zhao 和 Y. Shen，“研究用于文本分类的胶囊网络的传输能力”，*神经网络*，第一卷。第 118 页，第 247-261 页，2019 年。

[73] W. Zhao、H. Peng、S. Eger、E. Cambria 和 M. Yang，“面向具有挑战性的 NLP 应用的可扩展且可靠的胶囊网络”，*ACL*，2019 年，第 1549-1559 页。

[74] J. Kim、S. Jang、E. Park 和 S. Choi，“使用胶囊的文本分类”，*神经计算*，卷。376，第 214-221 页，2020 年。

[75] R. Aly、S. Remus 和 C. Biemann，“使用胶囊网络对文本进行分层多标签分类”，*第 57 届计算语言学协会年会论文集：学生研究研讨会*，2019 年，pp . 323–330。

[76] H. Ren 和 H. Lu，“用于文本分类的带有 k-means 路由的组合编码胶囊网络”，*arXiv 预印本 arXiv:1810.09177*，
2018 年。

[77] D. Bahdanau、K. Cho 和 Y. Bengio，“联合学习对齐和翻译的神经机器翻译”，*arXiv 预印本 arXiv:1409.0473，2014* 年。

[78] M.-T. Luong, H. Pham 和 CD Manning，“基于注意力的神经机器翻译的有效方法”，*arXiv 预印本 arXiv：1508.04025，2015* 年。
[79] Z. Yang，D. Yang，C. Dyer，X. He，A . Smola 和 E. Hovy，“文档分类的分层注意力网络”，在*计算语言学协会北美分会 2016 年会议论文集：人类语言技术*，2016 年，第 1480-1489 页。

[80] X. Zhou, X. Wan, and J. Xiao，“Attention-based lstm network for cross-lingual motion classification”，*2016 年自然语言处理经验方法会议论文集*，2016 年，第 247- 256.

[81] T. Shen、T. Zhou、G. Long、J. Jiang、S. Pan 和 C. Zhang，“Disan：用于 rnn/cnn-free 语言理解的定向自注意力网络”，*第 32 期 AAAI 人工智能会议*，2018。

[82] Y. Liu, C. Sun, L. Lin, and X. Wang, “Learning natural language inference using bidirectional lstm model and inner-attention,” *arXiv preprint arXiv:1605.09090* , 2016.

[83] C. d. Santos、M. Tan、B. Xiang 和 B. Zhou，“注意力池网络”，*arXiv 预印本 arXiv:1602.03609，2016* 年。

[84] G. Wang、C. Li、W. Wang、Y. Zhang、D. Shen、X. Zhang、R. Henao 和 L. Carin，“用于文本分类的词和标签的联合嵌入”，*arXiv 预印本 arXiv：1805.04174，2018*。

[85] S. Kim、I. Kang 和 N. Kwak，“语义句子匹配与密集连接的循环和共同注意信息”，*AAAI 人工智能会议论文集*，卷。33，2019，第 6586-6593 页。[86] W. Yin、H. Schütze、B. Xiang 和 B. Zhou，“Abcnn：用于对句子对建模的基于注意力的卷积神经网络”，*计算语言学协会会刊*，第一卷。4，第 259-272 页，2016 年。

[87] C. Tan、F. Wei、W. Wang、W. Lv 和 M. Zhou，“用于建模句子对的多路注意力网络”，*IJCAI*，2018 年，第 4411-4417 页。[88] L. Yang、Q. Ai、J. Guo 和 WB Croft，“anmm：使用基于注意力的神经匹配模型对简答文本进行排名”，*第 25 届 ACM 国际信息与知识管理会议论文集*，2016 年，第 287-296 页。

[89] Z. Lin, M. Feng, CN d. Santos、M. Yu、B. Xiang、B. Zhou 和 Y. Bengio，“一种结构化的自注意力句子嵌入”，*arXiv 预印本 arXiv:1703.03130，2017* 年。

[90] S. Wang、M. Huang 和 Z. Deng，“具有多尺度特征注意的密集连接 cnn 用于文本分类。”在 *IJCAI*，2018 年，第 4468-4474 页。

[91] I. Yamada 和 H. Shindo，“用于文本分类的神经注意力袋实体模型”，*arXiv 预印本 arXiv:1909.01259，2019* 年。 [92] AP Parikh、O. Tackstrom、D. Das 和 J. Uszkoreit，“自然语言推理的可分解注意力模型”，*arXiv 预印本 arXiv:1606.01933，2016* 年。

[93] Q. Chen, Z.-H. Ling 和 X. Zhu，“使用广义池化增强句子嵌入”，*arXiv 预印本 arXiv:1806.09828，2018* 年。

[94] ME Basiri、S. Nemati、M. Abdar、E. Cambria 和 UR Acharya，“Abcdm：用于情绪分析的基于注意力的双向 cnn-rnn 深度模型”，*未来一代计算机系统*，第一卷。115，第 279-294 页，2020 年。

[95] T. Munkhdalai 和 H. Yu，“神经语义编码器”，*会议记录。计算语言学协会。会议*，卷。1. NIH 公共访问，2017 年，p。397.

[96] J. Weston、S. Chopra 和 A. Bordes，“记忆网络”，*第 3 届学习表征国际会议，ICLR 2015 会议跟踪记录*，2015 年。

[97] S. Sukhbaatar、J. Weston、R. Fergus *等人。*，“端到端记忆网络”，*神经信息处理系统进展*，2015 年，第 2440-2448 页。[98] A. Kumar、O. Irsoy、P. Ondruska、M. Iyyer、J. Bradbury、I. Gulrajani、V. Zhong、R. Paulus 和 R. Socher，“问我任何问题：自然的动态记忆网络语言处理”，*第 33 届机器学习国际会议，ICML 2016，2016*。

[99] C. Xiong、S. Merity 和 R. Socher，“用于视觉和文本问答的动态记忆网络”，*第 33 届机器学习国际会议，ICML 2016 年*，2016 年。

[100] R. Mihalcea 和 P. Tarau，“Textrank：为文本带来秩序”，*2004 年自然语言处理经验方法会议论文集*，2004 年，第 404-411 页。

[101] Z. Wu、S. Pan、F. Chen、G. Long、C. Zhang 和 PS Yu，“图神经网络综合调查”，*arXiv 预印本 arXiv:1901.00596，2019* 年。

[102] TN Kipf 和 M. Welling，“图卷积网络的半监督分类”，*arXiv 预印本 arXiv:1609.02907，2016*。

[103] W. Hamilton、Z. Ying 和 J. Leskovec，“大图上的归纳表示学习”，*神经信息处理系统进展*，2017 年，第 1024-1034 页。

[104] P. Veličković、G. Cucurull、A. Casanova、A. Romero、P. Lio 和 Y. Bengio，“图注意力网络”，*arXiv 预印本 arXiv:1710.10903*，2017 年。

[105] H. Peng, J. Li, Y. He, Y. Liu, M. Bao, L. Wang, Y. Song 和 Q. Yang，“具有递归正则化深度图-cnn 的大规模分层文本分类，”在 *2018 年万维网会议记录中*。国际万维网会议指导委员会，2018 年，第 1063-1072 页。

[106] H. Peng, J. Li, Q. Gong, S. Wang, L. He, B. Li, L. Wang, 和 PS Yu，“用于大规模多的分层分类感知和注意图胶囊 rcnns -label 文本分类，” *arXiv 预印本 arXiv:1906.04898，2019* 年。

[107] L. Yao、C. Mao 和 Y. Luo，“用于文本分类的图卷积网络”，*在 AAAI 人工智能会议论文集*，卷。33，2019，第 7370-7377 页。

[108] F. Wu, T. Zhang, AH d. Souza Jr、C. Fifty、T. Yu 和 KQ Weinberger，“简化图卷积网络”，*arXiv 预印本 arXiv:1902.07153，2019* 年。

[109] L. Huang, D. Ma, S. Li, X. Zhang, and H. WANG, “用于文本分类的文本级图神经网络”，*arXiv:1910.02356，2019* . [110] P. Liu, S. Chang、X. Huang、J. Tang 和 JCK Cheung，“用于序列学习的上下文化非局部神经网络”，*在 AAAI 人工智能会议论文集*，卷。33，2019，第 6762-6769 页。

[111] J. BROMLEY、JW BENTZ、L. BOTTOU、I. GUYON、Y. LECUN、C. MOORE、E. SÄCKINGER 和 R. SHAH，“使用连体时间延迟神经网络进行签名验证”，*国际期刊模式识别与人工智能*，1993 年。

[112] W. tau Yih、K. Toutanova、JC Platt 和 C. Meek，“Learning discriminative projections for textsimilarity measure”，在 *CoNLL 2011 年第十五届计算自然语言学习会议上，会议论文集*，2011 年。

[113] P.-S。Huang, X. He, J. Gao, L. Deng, A. Acero 和 L. Heck，“使用点击数据学习用于网络搜索的深度结构化语义模型”，*第 22 届 ACM 信息与知识管理国际会议论文集*，2013 年，第 2333-2338 页。

[114] Y. Shen、X. He、J. Gao、L. Deng 和 G. Mesnil，“用于信息检索的具有卷积池结构的潜在语义模型”，*ACM 信息与知识管理国际会议*. ACM，2014 年，第 101-110 页。

[115] J. Gao、M. Galley 和 L. Li，“对话式人工智能的神经方法”，*信息检索中的基础和趋势®*，卷。13，没有。2-3，第 127-298 页，2019 年。

[116] A. Severyn 和 A. Moschittiy，“Learning to rank short text pairs with convolutional deep neural networks”，*2015 年 SIGIR - 第 38 届国际 ACM SIGIR 信息检索研究与开发会议论文集*，2015 年。

[117] A. Das、H. Yenala、M. Chinnakotla 和 M. Shrivastava，“我们站在一起：类似问题检索的连体网络”，*计算语言学协会第 54 届年会，ACL 2016 - 长篇论文*， 2016 年。

[118] M. Tan、CD Santos、B. Xiang 和 B. Zhou，“用于问答匹配的改进表示学习”，*第 54 届计算语言学协会年会，ACL 2016 - 长论文*，2016。

[119] J. Mueller 和 A. Thyagarajan，“用于学习句子相似性的连体循环架构”，*第 30 届 AAAI 人工智能会议，AAAI 2016，2016*。

[120] P. Neculoiu、M. Versteegh 和 M. Rotaru，“使用连体循环网络学习文本相似性”，2016 年。

[121] P. Liu、X. Qiu 和 X. Huang，“使用耦合 lstms 建模句子对的交互”，*arXiv 预印本 arXiv:1605.05573，2016*。 [122] H. He、K. Gimpel 和 J. Lin，“使用卷积神经网络进行多视角句子相似性建模”，*会议论文集 - EMNLP 2015：自然语言处理经验方法会议*，2015 年。

[123] T. Renter、A. Borisov 和 M. De Rijke，“Siamese CBOW：优化句子表示的词嵌入”，*计算语言学协会第 54 届年会，ACL 2016 - 长论文*，2016。

[124] N. Reimers 和 I. Gurevych，“Sentence-BERT：使用 Siamese BERT-Networks 的句子嵌入”，2019 年。

[125] W. Lu、J. Jiao 和 R. Zhang，“Twinbert：将知识提炼成双结构 bert 模型以实现高效检索”，*arXiv 预印本 arXiv：2002.06275，2020* 年。

[126] M. Tan, C. d. Santos、B. Xiang 和 B. Zhou，“用于非事实答案选择的基于 Lstm 的深度学习模型”，*arXiv 预印本 arXiv:1511.04108，2015* 年。

[127] Y. Tay、LA Tuan 和 SC Hui，“双曲线表示学习用于快速高效的神经问题回答”，*第 11 届 ACM 网络搜索和数据挖掘国际会议论文集*，2018 年，第 583-591 页。

[128] S. Minaee 和 Z. Liu，“使用深度相似性神经网络的自动问答”，*2017 年 IEEE 全球信号和信息处理会议 (GlobalSIP)*。 IEEE，2017，第 923-927 页。

[129] C. Zhou、C. Sun、Z. Liu 和 F. Lau，“用于文本分类的 c-lstm 神经网络”，*arXiv 预印本 arXiv：1511.08630，2015*。[130] R. Zhang，H. Lee 和 D. Radev，“用于建模句子和文档的依赖敏感卷积神经网络”，*计算语言学协会北美分会 2016 年会议：人类语言技术，NAACL HLT 2016 - 会议论文集*，2016 年。

[131] G. Chen、D. Ye、E. Cambria、J. Chen 和 Z. Xing，“卷积和循环神经网络在多标签文本分类中的集成应用”，*IJCNN*，2017 年，第 2377- 2383.

[132] D. Tang、B. Qin 和 T. Liu，“使用门控循环神经网络进行情感分类的文档建模”，*2015 年自然语言处理经验方法会议论文集*，2015 年，第 1422-1432 页.

[133] Y. Xiao 和 K. Cho，“通过结合卷积层和循环层实现高效的字符级文档分类”，*arXiv 预印本 arXiv:1602.00367，2016*。

[134] S. Lai、L. Xu、K. Liu 和 J. Zhao，“用于文本分类的循环卷积神经网络”，*第 29 届 AAAI 人工智能会议*，2015 年。

[135] T. Chen、R. Xu、Y. He 和 X. Wang，“使用 bilstm-crf 和 cnn 通过句子类型分类改进情感分析”，*应用专家系统*，卷。72，第 221 – 230 页，2017 年。[在线]。可用：http ://www.sciencedirect.com/science/article/pii/S0957417416305929

[136] K. Kowsari、DE Brown、M. Heidarysafa、KJ Meimandi、MS Gerber 和 LE Barnes，“Hdltex：用于文本分类的分层深度学习”，*2017 年第 16 届 IEEE 机器学习与应用国际会议 (ICMLA)*。 IEEE，2017，第 364-371 页。

[137] X. Liu、Y. Shen、K. Duh 和 J. Gao，“机器阅读理解的随机答案网络”，*arXiv:1712.03556，2017*。

[138] R. Srivastava、K. Greff 和 J. Schmidhuber，“训练非常深的网络”，*神经信息处理系统进展*，2015 年。

[139] K. He、X. Zhang、S. Ren 和 J. Sun，“用于图像识别的深度残差学习”，*IEEE 计算机视觉和模式识别会议论文集*，2016 年，第 770-778 页。

[140] Y. Kim、Y. Jernite、D. Sontag 和 AM Rush，“字符感知神经语言模型”，*第 30 届 AAAI 人工智能会议，AAAI 2016 年*，2016 年。

[141] JG Zilly、RK Srivastava、J. Koutnik 和 J. Schmidhuber，“循环高速公路网络”，*第 34 届机器学习国际会议，ICML 2017 年*，2017 年。

[142] Y. Wen、W. Zhang、R. Luo 和 J. Wang，“使用带有高速公路层的循环卷积神经网络学习文本表示”，*arXiv 预印本 arXiv:1606.06905，2016* 年。

[143] R. Collobert、J. Weston、L. Bottou、M. Karlen、K. Kavukcuoglu 和 P. Kuksa，“自然语言处理（几乎）从零开始”，*机器学习研究杂志*，第一卷。12，没有。2011 年 8 月，第 2493-2537 页。

[144] A. Radford、J. Wu、R. Child、D. Luan、D. Amodei 和 I. Sutskever，“语言模型是无监督的多任务学习者”，*OpenAI 博客*，第一卷。1，没有。第 8 页 9, 2019. [145] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai 和 X. Huang，“自然语言处理的预训练模型：调查”，*arXiv 预印本 arXiv： 2003.08271* , 2020.

[146] Y. Liu、M. Ott、N. Goyal、J. Du、M. Joshi、D. Chen、O. Levy、M. Lewis、L. Zettlemoyer 和 V. Stoyanov，“罗伯塔：稳健优化 bert 预训练方法，” *arXiv 预印本 arXiv:1907.11692，2019* 年。

[147] Z. Lan、M. Chen、S. Goodman、K. Gimpel、P. Sharma 和 R. Soricut，“Albert: A lite bert for self-supervised learning of language representations”，*arXiv 预印本 arXiv:1909.11942*， 2019 年。

[148] V. Sanh、L. Debut、J. Chaumond 和 T. Wolf，“Distilbert，bert 的精炼版：更小、更快、更便宜、更轻”，*arXiv 预印本 arXiv:1910.01108，2019* 年。

[149] M. Joshi、D. Chen、Y. Liu、DS Weld、L. Zettlemoyer 和 O. Levy，“Spanbert：通过表示和预测跨度来改进预训练”，*arXiv 预印本 arXiv:1907.10529，2019* 年。

[150] K.克拉克，M.-T。Luong、QV Le 和 CD Manning，“Electra：将文本编码器预训练为鉴别器而不是生成器”，*arXiv 预印本 arXiv:2003.10555，2020*。

[151] Y. Sun、S. Wang、Y. Li、S. Feng、X. Chen、H. Zhang、X. Tian、D. Zhu、H. Tian 和 H. Wu，“Ernie：通过知识整合”，*arXiv 预印本 arXiv:1904.09223，2019* 年。

[152] Y. Sun, S. Wang, Y.-K. Li、S. Feng、H. Tian、H. Wu 和 H. Wang，“Ernie 2.0：语言理解的持续预训练框架”。在 *AAAI*，2020 年，第 8968-8975 页。

[153] S. Garg、T. Vu 和 A. Moschitti，“Tanda：转移和调整预训练的变压器模型以进行答案选择”，*arXiv 预印本 arXiv:1911.04118，2019* 年。

[154] C. Sun、X. Qiu、Y. Xu 和 X. Huang，“如何微调 bert 以进行文本分类？”在*中国计算语言学全国学术会议上*。施普林格，2019 年，第 194-206 页。

[155] Z. Zhang、Y. Wu、H. Zhao、Z. Li、S. Zhang、X. Zhou 和 X. Zhou，“语义感知 bert 用于语言理解”，*arXiv 预印本 arXiv:1909.02209，2019*。

[156] Z. Yang、Z. Dai、Y. Yang、J. Carbonell、RR Salakhutdinov 和 QV Le，“Xlnet：用于语言理解的广义自回归预训练”，*神经信息处理系统进展*，2019 年，第 5754 页–5764。

[157] L. Dong、N. Yang、W. Wang、F. Wei、X. Liu、Y. Wang、J. Gao、M. Zhou 和 H.-W。Hon，“用于自然语言理解和生成的统一语言模型预训练”，《*神经信息处理系统进展*》，2019 年，第 13 042–13 054 页。

[158] H. Bao, L. Dong, F. Wei, W. Wang, N. Yang, X. Liu, Y. Wang, S. Piao, J. Gao, M. Zhou *等。*，“Unilmv2：用于统一语言模型预训练的伪掩码语言模型”，*arXiv 预印本 arXiv:2002.12804，2020*。

[159] C. Raffel、N. Shazeer、A. Roberts、K. Lee、S. Narang、M. Matena、Y. Zhou、W. Li 和 PJ Liu，“用统一文本探索迁移学习的极限文本转换器”，*arXiv 预印本 arXiv:1910.10683，2019* 年。

[160] DE Rumelhart、GE Hinton 和 RJ Williams，“通过错误传播学习内部表征”，加州大学圣地亚哥分校拉霍亚认知科学学院，科技。众议员，1985 年。

[161] R. Kiros、Y. Zhu、RR Salakhutdinov、R. Zemel、R. Urtasun、A. Torralba 和 S. Fidler，“Skip-thought vectors”，《*神经信息处理系统进展*》，2015 年，第。 3294-3302。

[162] AM Dai 和 QV Le，“半监督序列学习”，*神经信息处理系统进展*，2015 年。 [163] M. Zhang、Y. Wu、W. Li 和 W. Li，“学习通用使用 Mean-Max Attention Autoencoder 的句子表示”，2019 年。

[164] DP Kingma 和 M. Welling，“自动编码变分贝叶斯”，*第二届国际学习表示会议，ICLR 2014 - Conference Track Proceedings*，2014 年。

[165] DJ Rezende、S. Mohamed 和 D. Wierstra，“深度生成模型中的随机反向传播和近似推理”，*ICML*，2014 年。

[166] Y. Miao、L. Yu 和 P. Blunsom，“文本处理的神经变分推理”，*机器学习国际会议*，2016 年。 [167] SR Bowman, L. Vilnis, O. Vinyals, AM Dai 、R. Jozefowicz 和 S. Bengio，“从连续空间生成句子”，*2016 年 CoNLL - 第 20 届 SIGNLL 计算自然语言学习会议，论文集*，2016 年。

[168] S. Gururangan、T. Dang、D. Card 和 NA Smith，“半监督文本分类的变分预训练”，*arXiv 预印本 arXiv:1906.02242，2019* 年。

[169] Y. Meng、J. Shen、C. Zhang 和 J. Han，“弱监督神经文本分类”，*CIKM*，2018 年。

[170] J. Chen、Z. Yang 和 D. Yang，“混合文本：用于半监督文本分类的隐藏空间的语言知情插值”，*ACL*，2020 年。

[171] IJ Goodfellow、J. Shlens 和 C. Szegedy，“解释和利用对抗性示例”，*arXiv 预印本 arXiv:1412.6572，2014* 年。

[172] T. Miyato，S.-i。Maeda、M. Koyama、K. Nakae 和 S. Ishii，“虚拟对抗训练的分布式平滑”，*ICLR*，2016 年。 [173] T. Miyato、AM Dai 和 I. Goodfellow，“对抗性训练方法监督文本分类”，*arXiv 预印本 arXiv:1605.07725，2016* 年。

[174] DS Sachan、M. Zaheer 和 R. Salakhutdinov，“通过混合目标函数重新审视用于半监督文本分类的 lstm 网络”，*在 AAAI 人工智能会议论文集*，卷。33，2019，第 6940-6948 页。

[175] P. Liu、X. Qiu 和 X. Huang，“用于文本分类的对抗性多任务学习”，*arXiv 预印本 arXiv:1704.05742，2017*。 [176] RS Sutton 和 AG Barto，*强化学习：简介*。 麻省理工学院出版社，2018 年。

[177] T. Shen、T. Zhou、G. Long、J. Jiang、S. Wang 和 C. Zhang，“增强的自注意力网络：用于序列建模的硬注意力和软注意力的混合体”，*arXiv 预印本 arXiv：1801.10296，2018*。

[178] X. Liu, L. Mou, H. Cui, Z. Lu, and S. Song，“寻找文本分类中的决策跳跃”，*神经计算*，卷。371，第 177-187 页，2020 年。 [179] Y. Shen，P.-S。Huang、J. Gao 和 W. Chen，“Reasonet：在机器理解中学习停止阅读”，*第 23 届 ACM SIGKDD 知识发现和数据挖掘国际会议论文集*，2017 年，第 1047-1055 页。

[180] Y. Li、Q. Pan、S. Wang、T. Yang 和 E. Cambria，“类别文本生成的生成模型”，*信息科学*，卷。450，页。301–315，2018 年。

[181] T. Zhang、M. Huang 和 L. Zhao，“通过强化学习学习文本分类的结构化表示”，*第三十二届 AAAI 人工智能会议*，2018 年。

[182] Y. Gu、R. Tinn、H. Cheng、M. Lucas、N. Usuyama、X. Liu、T. Naumann、J. Gao 和 H. Poon，“生物医学自然领域的特定领域语言模型预训练语言处理”，*arXiv 预印本 arXiv：2007.15779，2020* 年。 [183] S. Mukherjee 和 AH Awadallah，“Xtremedistil：大规模多语言模型的多阶段蒸馏”，2020 年*计算语言学协会第 58 届年会论文集*，第 2221-2234 页。

[184] R. Tang、Y. Lu、L. Liu、L. Mou、O. Vechtomova 和 J. Lin，“将特定任务的知识从 bert 提取到简单的神经网络中”，*arXiv 预印本 arXiv:1903.12136，2019* 年。

[185] https://www.kaggle.com/yelp-dataset/yelp-dataset 。

[186] https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews 。

[187] B. Pang、L. Lee 和 S. Vaithyanathan，“竖起大拇指？：使用机器学习技术进行情感分类”，*ACL 会议论文集自然语言处理中的经验方法*，2002 年，第 79-86 页.

[188] L. Deng 和 J. Wiebe，“Mpqa 3.0：实体/事件级情感语料库”，*计算语言学协会北美分会 2015 年会议论文集：人类语言技术*，2015 年，第. 1323–1328。

[189] https://www.kaggle.com/datafiniti/consumer-reviews-of-amazon-products 。

[190] http://qwone.com/~jason/20Newsgroups/ 。

[191] https://martin-thoma.com/nlp-reuters 。

[192] F. Wang、Z. Wang、Z. Li 和 J.-R。温，“基于概念的短文本分类和排名”，*第 23 届 ACM 信息与知识管理国际会议论文集*。ACM，2014 年，第 1069-1078 页。

[193] D. Greene 和 P. Cunningham，“内核文档聚类中对角优势问题的实用解决方案”，*Proc。第 23 届机器学习国际会议（ICML'06）*。ACM 出版社，2006 年，第 377-384 页。

[194] AS Das、M. Datar、A. Garg 和 S. Rajaram，“Google 新闻个性化：可扩展的在线协同过滤”，*第 16 届万维网国际会议论文集*。ACM，2007，第 271-280 页。

[195] J. Lehmann、R. Isele、M. Jakob、A. Jentzsch、D. Kontokostas、PN Mendes、S. Hellmann、M. Morsey、P. Van Kleef、S. Auer *等人。*，“Dbpedia——从维基百科中提取的大规模、多语言知识库”，*语义网*，卷。6，没有。2，第 167-195 页，2015 年。

[196] http://davis.wpi.edu/xmdv/datasets/ohsumed.html 。

[197] EL Mencia 和 J. Fürnkranz，“法律领域大规模问题的高效成对多标签分类”，*欧洲机器学习和数据库知识发现联合会议*。施普林格，2008 年，第 50-65 页。

[198] Z. Lu，“Pubmed 及其他：用于搜索生物医学文献的网络工具的调查”，*数据库*，卷。2011 年，2011 年。

[199] F. Dernoncourt 和 JY Lee，“Pubmed 200k rct：医学摘要中顺序句子分类的数据集”，*arXiv 预印本 arXiv:1710.06071，2017*。

[200] BC Wallace、L. Kertz、E. Charniak *等人。*，“人类需要上下文来推断反讽意图（因此计算机也可能这样做）”，*第 52 届计算语言学协会年会论文集（第 2 卷：短论文）*，2014 年，第 512-516 页。

[201] P. Rajpurkar、R. Jia 和 P. Liang，“知道你不知道的事情：团队无法回答的问题”，*arXiv 预印本：1806.03822，2018* 年。

[202] T. Nguyen、M. Rosenberg、X. Song、J. Gao、S. Tiwary、R. Majumder 和 L. Deng，“Ms marco：人类生成的机器阅读理解数据集”，2016 年。

[203] https://cogcomp.seas.upenn.edu/Data/QA/QC/ 。

[204] Y.杨，W.-t。Yih 和 C. Meek，“Wikiqa：开放域问答的挑战数据集”，*2015 年自然语言处理经验方法会议论文集*，2015 年，2013-2018 年。

[205] https://data.quora.com/First-Quora-Dataset-Release-QuestionPairs 。

[206] R. Zellers、Y. Bisk、R. Schwartz 和 Y. Choi，“Swag：基于常识推理的大规模对抗性数据集”，*arXiv 预印本 arXiv:1808.05326，2018* 年。

[207] T. Jurczyk、M. Zhai 和 JD Choi，“Selqa：基于选择的问答的新基准”，*2016 年 IEEE 第 28 届人工智能工具国际会议 (ICTAI)*。IEEE，2016，第 820-827 页。

[208] SR Bowman、G. Angeli、C. Potts 和 CD Manning，“用于学习自然语言推理的大型注释语料库”，*arXiv 预印本 arXiv:1508.05326，2015* 年。

[209] A. Williams、N. Nangia 和 SR Bowman，“通过推理实现句子理解的广泛覆盖挑战语料库”，*arXiv 预印本 arXiv:1704.05426，2017* 年。

[210] B. Dolan、C. Quirk 和 C. Brockett，“大型释义语料库的无监督构建：利用大规模并行新闻来源”，*第 20 届计算语言学国际会议论文集*。ACL，2004 年，第 350。

[211] D. Cer、M. Diab、E. Agirre、I. Lopez-Gazpio 和 L. Specia，“Semeval-2017 任务 1：语义文本相似性 - 多语言和跨语言重点评估”，*arXiv 预印本 arXiv： 1708.00055* , 2017.

[212] I. Dagan、O. Glickman 和 B. Magnini，“PASCAL 识别文本蕴涵挑战”，*计算机科学讲义（包括人工智能子系列讲义和生物信息学讲义）*，2006 年。

[213] T. Khot、A. Sabharwal 和 P. Clark，“Scitail：来自科学问答的文本蕴涵数据集”，*第 32 届 AAAI 人工智能会议，AAAI 2018 年*，2018 年。

[214] AL Maas、RE Daly、PT Pham、D. Huang、AY Ng 和 C. Potts，“学习词向量进行情感分析”，*计算语言学协会第 49 届年会论文集：人类语言技术-第 1 卷*，2011 年，第 142–150 页。

[215] JC Martineau 和 T. Finin，“Delta tfidf：情绪分析的改进特征空间”，*第三届国际 AAAI 网络博客和社交媒体会议*，2009 年。

[216] J. Howard 和 S. Ruder，“用于文本分类的通用语言模型微调”，*arXiv 预印本 arXiv:1801.06146，2018* 年。

[217] B. McCann、J. Bradbury、C. Xiong 和 R. Socher，“在翻译中学习：语境化词向量”，《*神经信息处理系统进展*》，2017 年，第 6294-6305 页。

[218] S. Gray、A. Radford 和 DP Kingma，“块稀疏权重的 Gpu 内核”，*arXiv 预印本 arXiv:1711.09224*，第一卷。2017 年 3 月 3 日。

[219] A. Ratner、B. Hancock、J. Dunnmon、F. Sala、S. Pandey 和 C. Ré，“使用多任务弱监督训练复杂模型”，*AAAI 会议论文集关于人工智能*，卷。33，2019，第 4763-4771 页。

[220] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong 和 QV Le，“无监督数据增强”，*arXiv 预印本 arXiv:1904.12848，2019*。

[221] M. Kusner、Y. Sun、N. Kolkin 和 K. Weinberger，“从词嵌入到文档距离”，*机器学习国际会议*，2015 年，第 957-966 页。

[222] M. Richardson、CJ Burges 和 E. Renshaw，“Mctest：用于文本的开放域机器理解的挑战数据集”，*2013 年自然语言处理经验方法会议论文集*，2013 年，pp。 193–203。

[223] H.-Y。Huang、C. Zhu、Y. Shen 和 W. Chen，“Fusionnet：通过完全感知注意力与机器理解应用进行融合”，*arXiv 预印本 arXiv:1711.07341，2017* 年。

[224] Q. Chen, X. Zhu, Z.-H. Ling、S. Wei、H. Jiang 和 D. Inkpen，“基于循环神经网络的句子编码器，具有门控注意的自然语言推理”，*arXiv 预印本 arXiv:1708.01353，2017* 年。

[225] B. Pan、Y. Yang、Z. Zhao、Y. Zhuang、D. Cai 和 X. He，“带有强化学习的话语标记增强网络用于自然语言推理”，*arXiv 预印本 arXiv:1907.09692，2019* 年。

[226] W. Wang、F. Wei、L. Dong、H. Bao、N. Yang 和 M. Zhou，“Minilm：用于预训练变压器的任务不可知压缩的深度自我注意蒸馏”，*arXiv 预印本 arXiv：2002.10957，2020*。

[227] D. Jurafsky 和 JH Martin，*语音和语言处理：自然语言处理、计算语言学和语音识别简介*，第 1 版。美国：Prentice Hall PTR，2000。

[228] R. Sennrich、B. Haddow 和 A. Birch，“带有子词单元的稀有词的神经机器翻译”，*arXiv 预印本：1508.07909，2015* 年。 [229] M. Schuster 和 K. Nakajima，“日语和韩语语音搜索”，*2012 年 IEEE 声学、语音和信号处理国际会议 (ICASSP)*。 IEEE，2012，第 5149-5152 页。

[230] T. Kudo，“子词正则化：改进具有多个候选子词的神经网络翻译模型”，*ACL 2018 - 计算语言学协会第 56 届年会，会议论文集（长篇论文）*，2018 年。

[231] DE Rumelhart、GE Hinton、RJ Williams *等人。*，“通过反向传播误差学习表示”，*认知建模*，第一卷。5，没有。3，第 1，1988 年。

[232] http://colah.github.io/posts/2015-08-Understanding-LSTMs/ 。

[233] K. Fukushima，“Neocognitron：一种不受位置变化影响的模式识别机制的自组织神经网络模型”，*生物控制论*，卷。36，没有。4，第 193-202 页，1980 年。

[234] A. Krizhevsky、I. Sutskever 和 GE Hinton，“使用深度卷积神经网络的 Imagenet 分类”，*神经信息处理系统进展*，2012 年，第 1097-1105 页。

[235] S. Minaee、Y. Boykov、F. Porikli、A. Plaza、N. Kehtarnavaz 和 D. Terzopoulos，“使用深度学习进行图像分割：一项调查”，*arXiv 预印本 arXiv:2001.05566，2020* 年。

[236] O. Abdel-Hamid，A.-r。Mohamed、H. Jiang、L. Deng、G. Penn 和 D. Yu，“用于语音识别的卷积神经网络”，*IEEE/ACM 音频、语音和语言处理*汇刊，卷。22，没有。10，第 1533-1545 页，2014 年。

[237] S. Minaee、A. Abdolrashidi、H. Su、M. Bennamoun 和 D. Zhang，“使用深度学习进行生物特征识别：一项调查”，*arXiv 预印本 arXiv:1912.00271，2019* 年。

[238] J. Gehring、M. Auli、D. Grangier、D. Yarats 和 YN Dauphin，“卷积序列到序列学习”，*第 34 届机器学习国际会议论文集 - 第 70 卷*。JMLR。组织，2017 年，第 1243-1252 页。

深度神经　　网络概述

本附录介绍了一些常用的 NLP 深度学习模型，包括 MLP、CNN、RNN、LSTM、编码器-解码器和 Transformer。感兴趣的读者可以参考[26]进行全面讨论。

A.1　　神经语言模型和词嵌入

语言建模采用数据驱动的方法来捕获自然语言中文本序列的显着统计特性，这些特性以后可用于预测序列中的未来单词，或在相关任务中执行槽填充。N-gram 模型是最简单的统计语言模型，它捕获连续标记之间的关系。然而，这些模型无法捕获通常编码语义关系的标记的长距离依赖性[227]。因此，人们在开发更丰富的语言模型方面付出了很多努力，其中最成功的一种是神经语言模型[2]。

神经语言模型学习以自我监督的方式将文本标记（例如单词）表示为密集向量，称为词嵌入。然后，这些学习到的表示可以用于各种 NLP 应用程序。一种流行的神经语言模型是 word2vec [27]，它学习将出现在相似上下文中的单词映射到相似的向量表示。学习到的 word2vec 表示还允许对向量空间中的词嵌入进行一些简单的代数运算，如方程式所示。 5 .

$$\text{“}king\text{”} - \text{“}man\text{”} + \text{“}woman\text{”} = \text{“}queen\text{”} \tag{5}$$

尽管 word2vec 受欢迎且语义丰富，但仍存在一些问题，例如词汇量不足 (OOV) 扩展，无法捕捉词的形态和词的上下文。有许多工作试图改进 word2vec 模型，根据他们处理的文本单元以及是否依赖于上下文，它们可以分为以下几类：

- 字级嵌入
- 子词嵌入
- 上下文嵌入

字级嵌入。词级嵌入模型的两个主要类别是基于预测和基于计数的模型。前一类中的模型经过训练以恢复令牌序列中丢失的令牌。Word2vec 是该类别的早期示例，它提出了两种用于词嵌入的架构，连续词袋（CBOW）和 Skip-Gram [3 , 27]，如图 23 所示。Skip-Gram 模型预测

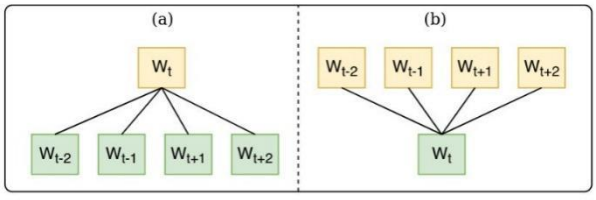

图 23. 两个 word2vec 模型 [27] (a) CBOW (b) Skip-Gram

每个上下文词都来自中心词，而 CBOW 模型根据其上下文词预测中心词。这些模型的训练目标是最大化正确单词的预测概率。

例如，CBOW 和 Skip-Gram 的训练目标如方程式所示。6 和等式。7 ，分别。

$$L_{CBOW} = -\frac{1}{|C|-C}\sum_{k=C+1}^{|C|-C}\log P(w_k \mid w_{k-C},\ldots,w_{k-1},w_{k+1},\ldots,w_{k+C}) \quad (6)$$

$$L_{Skip\text{-}Gram\sim i} = -[\log\sigma(v'^{\top}_{w} v_{w_I}) + \sum_{i=1,\, w\sim i\sim Q}^{N}\log\sigma(-v'^{\top}_{w} v_{w_I})] \quad (7)$$

GloVe [28] 是最广泛使用的基于计数的嵌入模型之一。它对单词的共现矩阵执行矩阵分解以学习嵌入。

子词和字符嵌入。词级嵌入模型存在 OOV 等问题。一种补救方法是将单词分割成子词或字符以进行嵌入。基于字符的嵌入模型不仅可以处理 OOV 词 [50 , 51]，而且可以减小嵌入模型的大小。子词方法找到最频繁的字符段（子词），然后学习这些段的嵌入。FastText [30] 是一种流行的子词嵌入模型，它将每个词表示为一个包含字符 n-gram 的包。这类似于 DSSM 中使用的字母三元组。其他流行的子词标记器包括字节对编码 [228]、WordPiece [229]、SentencePiece [230]， 等等。

上下文嵌入。一个词的意义取决于它的上下文。例如，“孩子在玩”句子中的“玩”这个词与“这出戏是莫扎特写的”中的“玩”有不同的含义。因此，词嵌入最好是上下文敏感的。Word2vec 和 Glove 都不是上下文敏感的。他们只是将一个词映射到同一个向量中，而不管它的上下文。另一方面，上下文化词嵌入模型可以根据上下文将一个词映射到不同的嵌入向量。ELMo [4] 是第一个大规模上下文相关嵌入模型，它在前向和后向使用两个 LSTM 来编码单词上下文。

A.2 循环神经网络 (RNN) 和长短期记忆 (LSTM)

RNN [231] 广泛用于处理序列数据，例如文本、语音、视频。vanilla RNN 模型的架构如图 24 （左）所示。模型从当前时间 X_i 获取输入，从上一步 h_{i-1} 获取隐藏状态，并生成隐藏状态和可选的输出。最后一个时间戳的隐藏状态（或所有隐藏状态的加权平均值）可以用作下游任务的输入序列的表示。

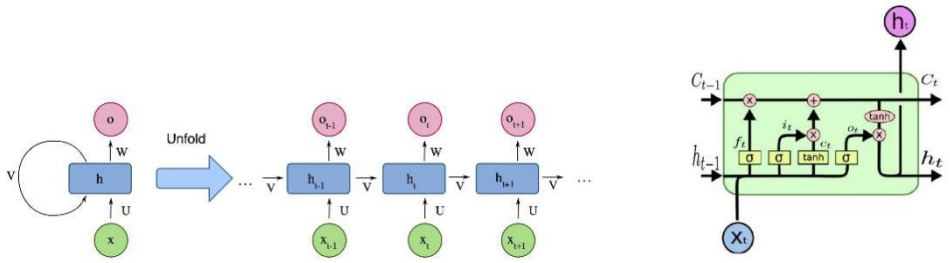

图 24.（左）RNN 的架构。（右）标准 LSTM 模块的架构 [232]。

由于梯度消失和爆炸问题，RNN 无法捕获在许多实际应用中出现的非常长序列的长期依赖关系。LSTM 是 RNN 的一种变体，旨在更好地捕捉长期依赖关系。如图 24 （右）和方程式所示。如图 8 所示，LSTM 层由一个记忆单元和三个门（输入门、输出门、遗忘门）和三个门（输入门、输出门、遗忘门）组成，该记忆单元在任意时间间隔内记住值。LSTM 的输入、隐藏状态和不同门之间的关系如公式 8 所示：

$$
\begin{aligned}
f_t &= \sigma(W^{(f)} x_t + U^{(f)} h_{t-1} + b^{(f)}), \\
i_t &= \sigma(W^{(i)} x_t + U^{(i)} h_{t-1} + b^{(i)}), \\
o_t &= \sigma(W^{(o)} x_t + U^{(o)} h_{t-1} + b^{(o)}), \\
c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(W^{(c)} x_t + U^{(c)} h_{t-1} + b^{(c)}), \\
h_t &= o_t \odot \tanh(c_t)
\end{aligned}
\tag{8}
$$

其中$x_t \in R^k$是在时间步长t输入的 kD 词嵌入，σ是逐元素 sigmoid 函数，⊙ 是逐元素积，W、U 和b是模型参数，c_t是记忆单元，遗忘门f_t决定是否重置记忆单元，输入门i_t和输出门o_t控制记忆单元的输入和输出存储单元，分别。

A.3 卷积神经网络 (CNN)

CNN 最初是为计算机视觉任务而开发的，但后来在各种 NLP 应用程序中出现了。CNN 最初是由 Fukushima 在他的开创性论文“Neocognitron”[233] 中提出的，基于 Hubel 和 Wiesel 提出的人类视觉系统模型。Yann LeCun 和他的同事通过开发一种基于反向传播来训练 CNN 的有效方法来普及 CNN [44]。LeCun 等人开发的 CNN 模型的架构。如图 25 所示。

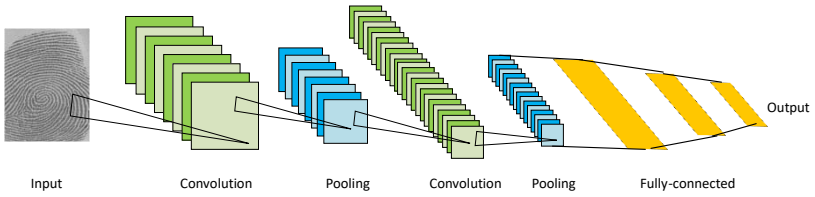

图 25. CNN 模型的架构，由 Yann LeCun [44] 提供。

CNN 由三种类型的层组成：（1）卷积层，将滑动核应用于图像（或文本段）的区域以提取局部特征；(2) 非线性层，将非线性激活函数应用于（局部）特征值；(3) 池化层，局部特征被聚合（通过最大池化或平均池化操作）以形成全局特征。CNN 的一个优势是由于使用了内核而实现的权重共享机制，这使得参数数量明显少于类似的全连接神经网络，从而使 CNN 更容易训练。CNN 已广泛用于计算机视觉、NLP 和语音识别问题 [45 , 139 , 234 – 238]。

A.4 编码器-解码器模型

编码器-解码器模型通过两阶段过程学习将输入映射到输出：（1）编码阶段，其中编码器f（.）将输入x压缩为潜在空间向量表示z，因为$z = f(x)$；(2) 解码阶段，解码器g(.)从z重构或预测输出y为 $y = g(z)$。潜在表示z预计将捕获输入的底层语义。这些模型广泛用于机器翻译等序列到序列的任务，如图 26 所示。

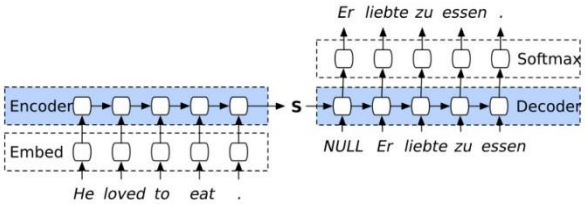

图 26. 用于机器翻译的简单编码器-解码器模型。输入是英语单词序列，输出是德语翻译版本。

自动编码器是编码器-解码器模型的特殊情况，其中输入和输出相同。自动编码器可以通过最小化重建损失以无监督的方式进行训练。

A.5 注意力机制

注意力的动机是我们如何将视觉注意力集中在图像的不同区域或在一个句子中关联单词。在开发 NLP 的深度学习模型时，注意力成为一个越来越流行的概念和有用的工具 [77 , 78]。简而言之，语言模型中的

注意力可以解释为重要性权重的向量。为了预测句子中的一个词，使用注意力向量，我们估计它与其他词的相关性或“关注度”有多强，并将注意力向量加权的它们的值的总和作为目标的近似值.

巴赫达瑙等人。[77]推测在 CNN 中使用固定长度的状态向量是提高编码器-解码器模型性能的瓶颈，并提出允许解码器在源语句中搜索与预测相关的部分。目标词，而不必将源语句压缩成状态向量。如图 27 （左）所示，输入词的隐藏向量 h 的线性组合，由注意力分数 α 加权，用于生成输出 y。从图 27 （右）可以看出，在目标句子中生成单词时，源句子中的不同单词具有不同的权重。

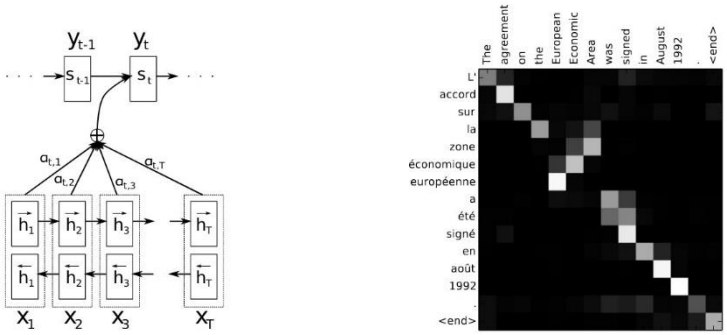

图 27.（左）[77]中提出的注意力机制。（右）法语到英语机器翻译中注意力机制的一个例子，它显示了法语中每个单词在翻译成英语时的影响，Brighter cell 有更大的影响。

自注意力是一种特殊的注意力机制，它允许学习同一句子中单词之间的相关性[35]。这在机器阅读、抽象摘要和图像字幕等 NLP 任务中非常有用。稍后将描述的 Transformer 也使用自注意力。

A.6 变压器

RNN 遇到的计算瓶颈之一是文本的顺序处理。尽管 CNN 的顺序性不如 RNN，但捕获句子中单词之间有意义关系的计算成本也随着句子长度的增加而增加，类似于 RNN。Transformers [5] 通过为句子或文档中的每个单词并行计算“注意力分数”来模拟每个单词对另一个单词的影响，从而克服了这一限制。由于这个特性，Transformer 允许比 CNN 和 RNN 更多的并行化，并且可以在 GPU 集群上的大量数据上有效地训练非常大的模型。

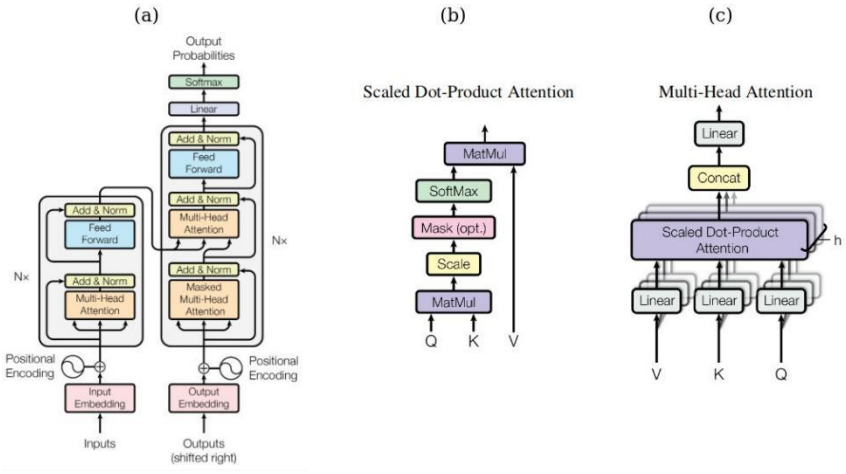

图 28. (a) Transformer 模型架构。(b) 缩放的点积注意力。(3) Multi-Head Attention 由多个并行运行的注意力层组成。[5]

.

如图 28 (a) 所示，Transformer 模型由编码器和解码器组件中的堆叠层组成。每层都有两个子层，包括一个多头注意力层（图 28 （c)），然后是一个位置前馈网络。对于每组查询Q、键K和值V，多头注意力模块使用缩放的点积注意力执行注意力 h 次，如图 28 所示（b)，其中 Mask（选项）是用于防止要预测的目标词信息在预测之前泄漏到解码器（在训练期间）的注意掩码。实验表明，多头注意力比单头注意力更有效。多个头的注意力可以解释为每个头在不同的位置处理不同的子空间。多个头的自我注意的可视化表明每个头处理语法和语义结构[5]。