

文章编号: 2095-2163(2020)04-0133-05

中图分类号: TP181

文献标志码: A

基于深度学习的自然语言处理研究综述

罗 泉

(浙江农林大学 信息工程学院, 浙江 杭州 311300)

摘要: 本文对深度神经网络模型做了简单介绍; 阐述了自然语言处理研究领域中所使用的深度学习方法及研究进展和成果; 重点介绍当前最新的预训练语言模型。并对深度学习在自然语言处理领域的发展趋势和有待深入研究的难点进行了总结及展望。

关键词: 深度学习; 自然语言处理; 深度神经网络; 语言模型

A survey of natural language processing based on Deep learning

LUO Xiao

(School of Information Engineering, Zhejiang A&F University, Lin'an Zhejiang 311300, China)

[Abstract] In this paper, the deep neural network model is briefly introduced. The deep learning methods used in the field of natural language processing and the research progress and achievements are described. Focus on the latest pre-training language models. The development trend of deep learning in the field of natural language processing and the difficulties to be further studied are summarized and prospected.

[Key words] Deep learning; Natural language processing; Deep neural network; Language model

0 引言

目前, 人工智能领域中最热的研究方向当属深度学习。深度学习的迅速发展受到了学术界和工业界的广泛关注, 由于其拥有优秀的特征选择和提取能力, 对包括机器翻译、目标识别、图像分割等在内的诸多任务中产生了越来越重要的影响。同时在自然语言处理(Natural Language Processing, NLP)、计算机视觉(Computer Vision, CV)、语音识别(Speech Recognition, SR)领域得到广泛应用。

自然语言处理被称为人工智能皇冠上的明珠, 因此如何使用深度学习技术推动 NLP 中各个任务的发展是当前研究热点和难点。语言是人类所特有的一种能力, 而如何用自然语言与计算机进行通信, 是人们长期以来追求的。自然语言处理就是实现人机间通过自然语言交流。但自然语言是高度抽象的符号化系统, 文本间存在数据离散、稀疏, 同时还存在多义词、一词多义等问题。而深度学习方法具有强大的特征提取和学习能力, 可以更好地处理高维度稀疏数据, 在 NLP 领域诸多任务中都取得了长足发展。因此, 本文将当前深度学习在 NLP 领域的发展展开综述性讨论, 详细阐述目前 NLP 的研究进展和最新的技术方法。

1 深度学习概述

深度学习的概念最早是由 Hinton^[1] 在 2006 年

提出的, 是研究如何从数据中自动提取多层特征表示。其核心思想是通过数据驱动的方式, 采用一系列的非线性变换, 从原始数据中提取由低层到高层、由具体到抽象的特征。不同于传统的浅层学习, 深度学习强调模型结构的深度, 通过增加模型深度来获取深层次含义。其次, 深度学习明确特征学习的重要性, 通过逐层特征变换, 将样本在原空间的特征表示变换到一个新特征空间, 从而使分类或预测更容易。最经典的深度学习网络包括卷积神经网络(Convolutional Neural Network, CNN)和递归神经网络(Recurrent Neural Network, RNN)。

1.1 卷积神经网络

CNN 是一种前馈神经网络, 区别于其它神经网络模型, 卷积运算操作赋予了 CNN 处理复杂图像和自然语言的特殊能力^[2]。CNN 神经元之间采用局部连接和权值共享的连接方式。其中, 局部连接是指每个神经元只需对图像或者文本中的部分元素进行感知, 最后的神经元对感知到的局部信息进行整合, 最后得到图像或文本的综合表示信息。权值共享使得模型在训练时, 可以使用较少的参数, 以此来降低深度神经网络模型的复杂性, 加快模型训练速度, 从而使深度神经网络模型可以被应用到实际生产。

作者简介: 罗 泉(1993-), 男, 硕士研究生, 主要研究方向: 深度学习、自然语言处理。

收稿日期: 2019-12-04

卷积神经网络通常由输入层、卷积层、池化层、全连接层和输出层5部分组成,CNN的网络结构如图1所示。

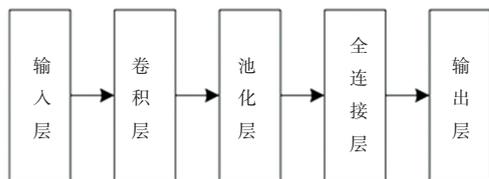


图1 卷积神经网络结构图

Fig. 1 The structure of Convolutional Neural Network

在自然语言处理领域,CNN的输入通常是单句或者句子表示成向量矩阵。卷积层是CNN中的重要组成部分,卷积层中的每一个节点输入是上一神经网络层的一部分,其目的是提取输入图片或者文本的不同特征。卷积层在处理文本序列问题时,通常使用不同大小的滤波器提取文本序列中不同特征。池化层是为了降低网络模型的输入维度,从而降低网络模型复杂度,减少整个模型参数,使神经网络模型具有更高的鲁棒性,同时在一定程度上能有效防止模型过拟合问题。其中最常见池化方式为最大池化(Max-Pooling)和平均池化(Average-Pooling)。CNN一般会在卷积层和池化层之后加上全连接层,该层可以把高维度转换成低维度,同时把有用的信息保留下来。通常将卷积层、池化层的组成部分视为自动提取特征的过程,在特征提取完成之后,需要使用输出层来完成分类或者预测任务。一般将学习到的高维度特征表示馈送到输出层,通过Softmax函数可以计算出当前样本属于不同类别的概率。

1.2 递归神经网络

递归神经网络具有树状层结构,网络节点按其连接顺序对输入信息进行递归的人工神经网络^[3]。RNN具有可变的拓扑结构且权重共享,多被用于包含结构关系的机器学习任务,在NLP领域受到研究者的重点关注。

RNN的基本结构包括输入层、隐藏层和输出层。与传统神经网络最大的区别在于RNN每次计算都会将前一词的输出结果送入下一词的隐藏层中一起训练,最后仅仅输出最后一个词的计算结果。RNN的缺点:

- (1)对短期的记忆影响比较大,但对长期的记忆影响很小,无法处理很长的输入序列。
- (2)训练RNN需要极大的成本投入。
- (3)RNN在反向传播时求底层的参数梯度会涉

及到梯度连乘,容易出现梯度消失或者梯度爆炸^[4]。而长短时记忆网络(Long Short-Term Memory, LSTM)和门控循环单元(Gated Recurrent Unit, GRU)在一定程度上可以解决该问题。

2 NLP应用研究进展

NLP领域主要研究任务包括语言建模、机器翻译、问答系统、情感分析、文本分类、阅读理解、中文分词、词性标注及命名实体等。国内外学者使用深度学习在NLP各个任务都取得了前所未有的发展。

Conneau等人^[5]将残差网络和CNN相结合,提出一种最多包含29个卷积层的深度神经网络架构VDCNN用于文本分类。Chen^[6]将条件随机场(Conditional Random Field, CRF)引入情感分析中,利用BiLSTM和CRF组合模型捕获句子中不同目标,然后利用一维CNN进行分类。Yang^[7]提出结合注意力机制的层次结构模型,该模型通过使用门控循环单元(Gated Recurrent Unit, GRU)分别对词和句子进行建模,然后在GRU之后加入注意力机制,并成功将该模型应用在篇章级别文本分类任务中。Choi等人^[8]提出一种细粒度注意力用于机器翻译任务中,其中上下文向量的每个维度都将收到单独的注意力分数。Wu等人^[9]提出一种轻量级的机器翻译模型,将动态卷积与自注意力机制相结合在英语—德语翻译任务中取得了优异的效果。Tan等人^[10]提出一种在匹配聚合框架下,利用多个注意力函数匹配句子对的多向注意力网络,并成功将该网络应用在语义相似计算任务中。Tay等人^[11]提出一种以紧密连接方式的神经网络结构用于阅读理解任务,该网络以紧密方式连接网络的所有层,建立了跨层次的通道和查询之间的关系,此外该网络的密集连接器是通过注意力机制学习,而不是标准的跳跃连接器。

黄改娟等人^[12]提出一种双重注意力模型,在模型训练过程中使用微博数据集,数据中不仅包含文本信息还包括情感符号。通过注意力机制和情感符号的结合,模型增加了对微博数据中情感知识的获取能力,进而将分类的准确率进行了提升。金志刚等人^[13]通过对BiLSTM和Bagging算法的改进,提出一种新的情感分析Bi-LSTM-B模型,该模型的优点在于结合了深度学习模型可提取抽象特征的优势和集成学习多分类器共同决策的思想,相比于其它模型,该模型提高了情感分析的准确率。张新路^[14]针对维吾尔语到汉语这种低资源语料库使用神经机器翻译容易出现局部最优解问题,提出利用

集成策略整合多个模型预测概率分布,将多个模型作为一个整体,训练选取出得分最高的候选作为翻译结果输出。陈瑛等人^[15]以食品安全为研究对象,采用 Lucene 全文检索架构和 LSTM 构建了食品安全自动问答系统。李佳媛等人^[16]提出一种基于语义词典和语料库相结合的词语语义相似度计算模型,在 Word Similarity-353 数据集上取得了优异结果,显著提高了中文词语语义相关度计算效果。

3 预训练语言模型

预训练思想的本质是模型参数不再随机初始化,而是通过语言模型进行训练。目前 NLP 各项任务的解决思路是预训练加微调。预训练对于 NLP 任务有着巨大的提升帮助,而预训练语言模型也越来越多,从最初的 Word2vec^[17]、Glove^[18]到通用语言文本分类模型 ULMFiT^[19]以及 EMLo^[20]等。而当前最优秀的预训练语言模型是基于 Transformer 模型构建。该模型是由 Vaswani 等人^[21]提出的,其是一种完全基于 Self-Attention 所构建的,是目前 NLP 领域最优秀的特征提取器,不但可以并行运算而且可以捕获长距离特征依赖。

3.1 BERT

当前影响最大的预训练语言模型是基于 Transformer 的双向深度语言模型—BERT^[22]。其网络结构如图 2 所示。

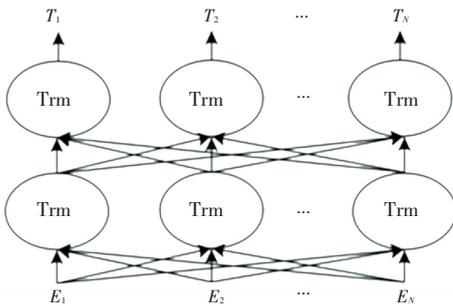


图2 BERT模型

Fig. 2 BERT model

BERT 是由多层双向 Transformer 解码器构成,主要包括 2 个不同大小的版本:基础版本有 12 层 Transformer,每个 Transformer 中的多头注意力层是 12 个,隐藏层大小为 768;加强版有 24 层 Transformer,每个 Transformer 中的多头注意力层是 24 个,隐藏层大小为 1 024。由此可见深而窄的模型效果要优于浅而宽的模型。目前 BERT 在机器翻译、文本分类、文本相似性、阅读理解等多个任务中都有优异的表现。BERT 模型的训练方式包括 2 种:

(1)采用遮盖单词的方式。将训练语料中的 80%的单词用 [MASK] 替换,如 my dog is hairy→ my dog is [MASK]。还有 10%的单词进行随机替换,如 my dog is hairy→ my dog is banana。剩下 10%则保持句子内容不变。

(2)采用预测句子下一句的方式。将语料中的语句分为 A 和 B,B 中的 50%的句子是 A 中的下一句,另外的 50%则是随机的句子。通过上述 2 种方式训练得到通用语言模型,然后利用微调的方法进行下游任务,如文本分类、机器翻译等任务。较比以前的预训练模型,BERT 可以捕获真正意义上的双向上下文语义。但 BERT 也有一定的缺点,既在训练模型时,使用大量的 [MASK] 会影响模型效果,而且每个批次只有 15%的标记被预测,因此 BERT 在训练时的收敛速度较慢。此外由于在预训练过程和生成过程不一致,导致在自然语言生成任务表现不佳,而且 BERT 无法完成文档级别的 NLP 任务,只适合于句子和段落级别的任务。

3.2 XLNet

XLNet^[23]是一种广义自回归的语言模型,是基于 Transformer-XL^[24]而构建的。Transformer 的缺点:

(1)字符之间的最大依赖距离受输入长度的限制。

(2)对于输入文本长度超过 512 个字符时,每个段都是从头开始单独训练,因此使训练效率下降,影响模型性能。针对以上 2 个缺点,Transformer-XL 引入了 2 个解决方法:分割循环机制 (Division Recurrence Mechanism) 和相对位置编码 (Relative Positional Encoding)。Transformer-XL 的测试速度更快,可以捕获更长的上下文长度。

无监督表征学习在 NLP 领域取得了巨大成功,在这种理念下,很多研究者探索了不同的无监督预训练目标,而自回归语言建模和自编码语言是 2 个最成功的预训练目标。而 XLNet 是一种集合了自回归和自编码 2 种方式的泛化自回归方法。XLNet 不使用传统自回归模型中的固定前向或后向因式分解顺序,而使用一种随机排列自然语言预测某个位置可能出现的词,这种方式不仅可以使句子中的每个位置都能学习来自所有位置的语境信息,而且还可以构建双向语义,更好地获取上下文语义。由于 XLNet 采用的是 Transformer-XL,因此模型性能更优,尤其在包含长文本序列的任务中。通过 XLNet 训练得到语言模型后,可以用于下游相关任务,如阅

读理解,基于XLNet得到的结果已经远超人类水平,在文本分类、机器翻译等任务中取得了优异的效果。

3.3 ERNIE

无论是BERT还是XLNet语言模型,在英文语料中表现都很优异,但在中文语料中效果一般,ERNIE^[25]则是以中文语料训练得出一种语言模型。ERNIE是一种知识增强语义表示模型,其在语言推断、语义相似度、命名实体识别、文本分类等多个NLP中文任务上都有优异表现。ERNIE在处理中文语料时,通过对预测汉字进行建模,可以学习到更大语义单元的完整语义表示。ERNIE模型内部核心是由Transformer所构成,其模型结构如图3所示。模型结构主要包括2个模块,下层模块的文本编码器(T-Encoder)主要负责捕获来自输入标记的基本词汇和句法信息,上层模块的知识编码器(K-Encoder)负责从下层获取的知识信息集成到文本信息中,以便能够将标记和实体的异构信息表示成一个统一的特征空间中。

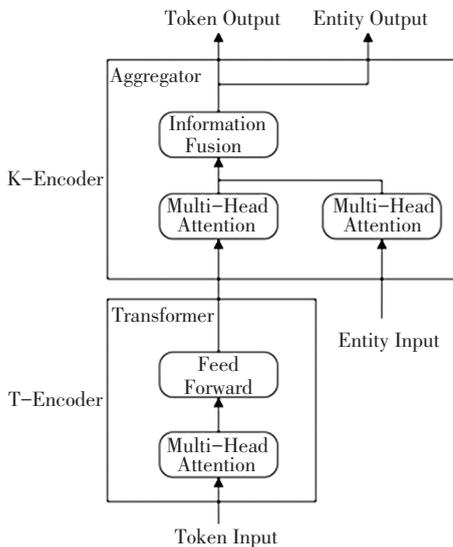


图3 ERNIE模型

Fig. 3 ERNIE model

ERNIE模型通过建立海量数据中的实体概念等先验语义知识,学习完整概念的语义表示,即在训练模型时采用遮盖单词的方式通过对词和实体概念等语义单词进行遮盖,使得模型对语义知识单元的表达更贴近真实世界。此外,ERNIE模型引入多源语料训练,其中包括百科类、新闻资讯类、论坛对话等数据。总体来说,ERNIE模型通过对实体概念知识的学习来学习真实世界的完整概念语义表示,使得模型对实体概念的学习和推理能力更胜一筹,其次通过对训练语料的扩充,尤其是引入了对话语料

使得模型的语义表示能力更强。

4 结束语

本文主要对深度学习中的卷积神经网络和递归神经网络做了简单介绍,阐述了目前NLP领域各个任务的研究进展。当前NLP的研究重点是预训练语言模型,因此详细介绍BERT、XLNet和ERNIR3种模型。尽管深度学习在NLP各个任务中取得了巨大成功,但若大规模投入使用,仍然有许多研究难点需要克服。深度神经网络模型越大,使得模型训练时间延长,如何减小模型体积但同时保持模型性能不变是未来研究的一个方向。此外深度神经网络模型可解释性较差,在自然语言生成任务研究进展不大。但是,本文认为随着深度学习的不断研究深入,在不久的将来,NLP领域将会取得更多研究成果和发展。

参考文献

- [1] HINTON G E, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets[J]. *Neural Computation*, 2006, 18(7): 1527-1554.
- [2] 周飞燕,金林鹏,董军. 卷积神经网络研究综述[J]. *计算机学报*, 2017, 40(6): 1229-1251.
- [3] GOODFELLOW I, BENGIO Y, COURVILLE A. Deep learning[M]. Cambridge: MIT Press. 2016(1): 394-396.
- [4] 奚雪峰,周国栋. 面向自然语言处理的深度学习研究[J]. *自动化学报*, 2016, 42(10): 1445-1465.
- [5] CONNEAU A, SCHWENK H, BARRAULT L, et al. Very Deep Convolutional Networks for Text Classification[C]// In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. 2017, 1: 1107-1116.
- [6] TAO C, XU R F, HE Y L, et al. Improving Sentiment Analysis via Sentence Type Classification Using BiLSTM-CRF and CNN[J]. *Expert Systems With Applications*. 2017(72): 221-230.
- [7] YANG Z C, YANG D Y, DYER C, et al. Hierarchical Attention Networks for Document Classification[C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016: 1480-1489.
- [8] CHOI H Y, CHO K, BENGIO Y H. Fine-Grained Attention Mechanism for Neural Machine Translation[J]. *Neurocomputing*. 2018, 284: 171-176.
- [9] WU F, FAN A, BAEVSKI A, et al. Pay Less Attention with Lightweight and Dynamic Convolutions[C]// ICLR 2019: 7th International Conference on Learning Representations. 2019.
- [10] TAN C Q, WEI F, WANG W H, et al. Multiway Attention Networks for Modeling Sentence Pairs[C]// IJCAI 2018: 27th International Joint Conference on Artificial Intelligence. 2018: 4411-4417.
- [11] TAY Y, LUU A T, HUI S C, et al. Densely Connected Attention Propagation for Reading Comprehension[C]// NIPS 2018: The 32nd Annual Conference on Neural Information Processing Systems, 2018: 4906-4917.
- [12] 张仰森,郑佳,黄改娟,等. 基于双重注意力模型的微博情感分

- 析方法[J]. 清华大学学报(自然科学版), 2018, 58(2): 122-130.
- [13] 金志刚, 韩玥, 朱琦. 一种结合深度学习和集成学习的情感分析模型[J]. 哈尔滨工业大学学报, 2018, 50(11): 32-39.
- [14] 张新路, 李晓, 杨雅婷, 等. 面向维汉神经机器翻译的双向重排序模型分析[J/OL]. 北京大学学报(自然科学版). 1-9[2019-10-15]. <https://doi.org/10.13209/j.0479-8023.2019.093>.
- [15] 陈瑛, 陈昂轩, 董玉博, 等. 基于 LSTM 的食品安全自动问答系统方法研究[J]. 农业机械学报, 2019, 50(S1): 380-384.
- [16] 张仰森, 郑佳, 李佳媛. 一种基于语义关系图的词语语义相关度计算模型[J]. 自动化学报, 2018, 44(1): 87-98.
- [17] MIKOLOV T, CHEN K, CORRADO G S, et al. Efficient Estimation of Word Representations in Vector Space [C]// In ICLR (Workshop Poster). 2013.
- [18] PENNINGTON J, SOCHER R, MANNING C D. Glove: Global Vectors for Word Representation [C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014: 1532-1543.
- [19] HOWARD J, RUDER S. Universal Language Model Fine-Tuning for Text Classification [C]// ACL 2018: 56th Annual Meeting of the Association for Computational Linguistics. 2018, 1: 328-339.
- [20] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations [C]// NAACL HLT 2018; 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2018, 1: 2227-2237.
- [21] VASWANI A, SHAZEER N, PARMAR N, et al. Attention Is All You Need [C]// NIPS '17 Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 6000-6010.
- [22] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding [C]// NAACL - HLT 2019: Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2019: 4171-4186.
- [23] YANG Z L, DAI Z H, YANG Y M, et al. XLNet: Generalized Autoregressive Pretraining for Language Understanding [C]// NeurIPS 2019: Thirty-Third Conference on Neural Information Processing Systems. 2019.
- [24] DAI Z H, YANG Z L, YANG Y M, et al. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context [C]// ACL 2019: The 57th Annual Meeting of the Association for Computational Linguistics. 2019: 2978 - 2888.
- [25] SUN Y, WANG S H, LI Y K, et al. ERNIE: Enhanced Representation through Knowledge Integration [DB]. ArXiv Preprint ArXiv:1904.09223. 2019.

(上接第 132 页)

根据上图可知,磨粒在线圈中获取的涡流密度与线圈内径成反比,且随着线圈内径的增大,涡流作用衰减速率缓慢减少。

4 结束语

本文根据等效电路原理,得到了影响电涡流传感器输出的主要因素为涡流检测理论和电涡流监测金属磨粒原理。应用 ANASYS Maxwell 软件进行仿真验证,从磨粒尺寸、磨粒材质、线圈激励频率以及线圈内径 4 个方面分别做了仿真分析,得出结论为:

(1) 电涡流密度随磨粒尺寸的增大而加快。因此,可以通过电涡流密度的变化识别出不同的磨粒尺寸。

(2) 不同的金属磨粒会产生差异性的涡流密度,而非金属磨粒无法引起电涡流作用。因此,可以通过涡流密度识别出不同的磨粒材质。

(3) 磨粒产生的涡流密度与激励频率变化呈正

比,但增长速率显示为下降趋势。

(4) 磨粒产生的涡流密度随线圈内径的增大而减小,且涡流作用衰减速率放缓。

参考文献

- [1] 郑晓虎. 基于平面线圈的微传感器及其制造技术研究[D]. 南京:南京航空航天大学,2007.
- [2] 于亚婷. 与被测材料无关的电涡流传感器基础理论与实现方法研究[D]. 成都:电子科技大学,2007.
- [3] 龚褪. 一般情况下两共轴圆线圈间互感系数的简便计算[J]. 大学物理,2011, 30(4): 46-48.
- [4] 于亚婷,杜平安,廖雅琴. 线圈形状及几何参数对电涡流传感器性能的影响[J]. 仪器仪表学报,2007,28(06): 1045-1050.
- [5] 赵博. Ansoft12 在工程电磁场中的应用[M]. 中国水利水电出版社,2010.
- [6] 刘国强,赵凌志,蒋继娅. Ansoft 工程电磁场有限元分析[M]. 电子工业出版社,2005.
- [7] 傅舰艇. 油路磨粒检测方法与电路研究[D]. 成都:电子科技大学,2012.
- [8] 冯丙华,杜永平. 电感式磨粒检测传感器参数的探讨[J]. 煤矿机械,2005(10): 52-54.