基于注意力的权重分配机制

张亚飞
(中国石油大学(华东) 计算机科学与技术学院,山东 青岛 266580)

摘 要:目前,基于神经网络的深度学习技术得到了飞速的发展,已经广泛应用于日常生活,如行人检测、车牌识别、人脸识别等。理论上可以通过不断扩大神经网络规模来提高算法准确度,然而这种方法并不可行。原因在于单纯扩大网络规模会导致过拟合。为了解决这个问题,通过人的先验知识来指导神经网络结构的设计以及明确神经网络每一个模块需要学习的目标,进而通过明确的模块分工来提升神经网络性能。受注意力机制和正则化方法的启发,提出了一个基于注意力机制的自适应权重分配算法,通过对神经网络各模块进行合理的权重分配,强调或者弱化某些输入数据对于下一步处理的贡献并以可微分的方式进行设计,完成一个端对端的神经网络。实验结果显示相比于其他方法,该算法达到了更好的效果。

关键词:深度学习;神经网络;网络优化;注意力机制;先验知识
中图分类号:TP302.1 **文献标识码**:A **文章编号**:1673-629X(2020)09-0049-05
doi:10.3969/j.issn.1673-629X.2020.09.009

Attention-based Weight Allocation Mechanism

ZHANG Ya-fei
(School of Computer Science and Technology, China University of Petroleum, Qingdao 266580, China)

Abstract: Recently, the deep learning technology based on neural network has been developed rapidly and widely used in daily life, such as pedestrian detection, license plate recognition, face recognition and so on. Theoretically, the neural networks have the ability to approximate any complicated functions abstracting from real world. However, in practice, it is impossible to reach this target. The reason is that enlarging networks would cause over-fitting. To tackle this question, some researches design the neural networks structure guided by person's priori knowledge to narrow down the function of every neural network module for increasing its performance. Inspired by attention mechanism and regularization methods, we propose an attention-based weight allocation mechanism to optimize the network structure. This method emphasizes or reduces contributions of some input data by distributing weights into different neural networks modules, which is designed as a differentiable end-to-end neural network. In a number of experiments on citation networks and on some public datasets, we demonstrate that the proposed algorithm has a better quality and outperforms other methods.

Key words: deep learning; neural networks; network optimization; attention mechanism; priori knowledge

0 引 言

近年来,神经网络在图像分类和目标识别领域取得了巨大的成功[1-3]。然而,研究人员对于提升准确度的追求没有改变。因此针对神经网络的各种优化方法层出不穷,然而已有的算法大多针对具体的问题进行调参,对于通用框架的改进则相对较少。第一个原因是普适性的解决框架难以找到;第二个原因是在实际问题中往往面对的是具体问题,需要针对特定问题进行偏置归纳以便使网络更符合真实数据集。

在深度学习框架设计中有一个基本原则是进行稀疏学习,使用较少的参数来表征数据特征,从而达到良好的抽象效果和泛化效果。例如针对权重的 L1[4] 和 L2[5] 正则化,L1 正则化针对权重绝对值之和进行约束,使其尽可能小,L2 正则化针对权重的平方之和的平方根进行约束,使其权重值更小,这也就限制了多项式中某些分量的影响,相当于减少参数个数。然而 L1 和 L2 正则化仅仅是针对其所约束的权重矩阵的,然而针对更高一层级的同层之间的神经元,以及更大范围的神经层则没有相应的稀疏约束来实现网络结构的稀疏化。

因此,设计针对网络的稀疏性约束对于提升网络的泛化能力具有很大的作用,然而普通的数值型约束

收稿日期:2019-10-24 **修回日期**:2020-03-05
基金项目:国家自然科学基金(61402533);山东省自然科学基金(ZR2019MF049)
作者简介:张亚飞(1994-),男,硕士,研究方向为深度学习、机器学习、数据挖掘。

针对网络结构并没有很好的约束效果,而且如果基于人为设计进行权重分配,那么就会因为需要设计的超参数太多而导致学习效果不佳,因此更好的分配方式是采用自动化权重分配,即权重自学习的方式。目前最好的自学习方式是使用神经网络。同样,可以利用神经网络的这个特点来学习权重分配函数。

因此,文中提出了一种可学习的权重分配机制,与注意力机制相似,该机制为神经元与神经元之间分配权重,为并行化的神经层与神经层之间分配权重。具体的做法是使用一个多层神经网络对权重分配函数进行学习。而学习方式有别于普通的神经网络训练方式,首先只是单纯训练一个目标网络,训练完成后在网络中添加权重分配网络,进而固定目标网络的参数,针对权重分配网络进行训练,迭代训练目标网络和权重分配网络直至效果最优。

创新点:针对神经元在设计过程中采用未对同层神经元之间进行权重区分的问题,采用注意力机制对其进行同层权重分配,通过强调或者弱化神经元学习到的特征的方式来提高神经网络的精度;提出一种新的针对于注意力机制的训练方法,即循环迭代训练,首先训练常规神经网络关系层,然后训练注意力层,迭代循环,直至目标函数收敛。

1 相关工作

文中的主要工作基于注意力机制和图卷积神经网络,因此接下来对这两个领域的工作进行介绍。

1.1 注意力机制

注意力机制受人的视觉认知启发,人的视觉在处理图像信息的时候并不总是关注全局信息,而是根据任务目标来重点关注某个具体的区域获取最有用的信息。基于图像上不同区域的信息来建立内部联系[6]指导注意力的转移和决策。作为注意力机制的基础,人眼的注意力已经从神经学和认知学上得到了充分的研究,图像中最低层级的信息在视觉注意力中起着重要的作用[7],同时人眼关注的图像的区域与具体的任务目标具有很强的相关性[8-9]。基于此,Volodymyr 将注意力应用于视觉图像处理来缩减网络规模进而降低计算资源消耗,这项工作首次将注意力机制引入深度学习框架中[10]。在注意力机制展现其卓越的性能之后,越来越多的学者将注意力机制纳入其研究领域进行创新和发展,作为注意力机制的一个方向,自注意力机制已经被成功应用于阅读理解、文本摘要等任务中[11-12]。

1.2 图卷积神经网络

在图卷积神经网络中,受限于图上边的存在依赖于具体问题的特性,因此一般情况下在进行图卷积操作时会输入一个全局的邻接矩阵来表征图上节点与节点的连接关系。在消息传递算法[13]中,每个节点状态的更新是基于该节点的邻居节点的状态,然而如何知道其邻居是谁,这种情况下就需要用到邻接矩阵来获取该节点的邻居信息以及存在的边信息。而邻接矩阵通常是作为全局信息而存在的。那么受此启发是否可以将这种针对数据之间的邻接矩阵约束,或者也可以称之为稀疏约束,因为其相对于全连接形式来说,在信息与信息的关联性上具有很大的稀疏性,反映到神经网络结构上就是,神经元与神经元之间的连接并不需要全连接的形式,更好的方法是只选取其中重要信息的方法,就可以达到一种优化的效果。类似于在权重矩阵中,有 L1 正则化和 L2 正则化对权重矩阵进行稀疏化约束。

然而针对更高层级,神经元与神经元的稀疏激活约束,神经层与神经层之间的稀疏约束还没有方法涉及到相关方面。DropOut 和 DropConnect[14],可近似地看作神经元之间的稀疏约束,通过屏蔽部分神经元使其不工作来实现,但是其具有很大的随机性,无法获取到一个全局有效的信息对神经元的激活性或者稀疏性进行约束,而稀疏性约束对于神经网络结构来说同样是重要的[15]。

在常规网络中并没有针对神经网络结构所做的稀疏性约束或者称之为神经元权重分配机制,针对于特征选择器所选择出来的特征没有施加权重系数,即其对于最终结果的贡献度,因此,平等对待并不能达到最优的效果。为了解决这个问题,提出新的权重分配机制来奖励重要的特征,减少低贡献的特征权重。基于注意力机制,通过神经网络来自动化拟合权重分配函数,分别对同层的神经元和不同的神经层之间进行权重分配。其最终目的是将具有并行关系的神经元或者神经层看作具有竞争注意力关系的目标,对其进行自适应权重分配。

2 权重分配模型

针对注意力机制可以针对数据进行合理分配权重的特性,其核心原理是从大量信息中找到目标信息。按照已有的研究总结如下:加权可以作用在原图上;加权可以作用在空间尺度上,给不同空间区域加权[16];加权可以作用在 channel 尺度上,给不同的通道特征加权[17];加权可以作用在不同时刻的历史特征上,结合循环结构添加权重[18]。上述研究证明了注意力机制在神经网络里边的广泛应用,然而,更本质上来说,它们针对的数据处理结构都有一个共同的特点,即上一步处理出来的数据对于下一步数据处理模块的重要程度是不同的。即忽略了特征与特征之间的关系。因此

文中提出的是一个通用的模型优化方法,即基于注意力机制对神经网络结构的稀疏性进行约束。

在神经网络结构设计上,权重分配主要体现在对下一步处理具有平行关系的输入上,因此,在从整个神经网络处理流程来看,其所包含的是一大的平行模块里边包含着一个一个小的并行模块。因为最终是通过神经网络计算出来一个损失值,优化目标也是一个。最终其中的各种大的小的并行模块必然要汇聚在一起,然而传统上,它们单纯以一种简单线性相加的方式来汇聚,可能在某些网络设计的时候考虑多个优化目标是对其设置一个经验参数用来平衡不同优化目标之间的比例关系,但更具体,更深入网络结构内部的权重分配关系却并没有得到充分研究。在此将借助注意力机制来对此进行权重分配和研究。

要在实际计算中准确进行注意力分配,首先要了解神经网络的具体设计流程,或者也可以称之为数据处理流程,然后进行逐层解析,在神经网络中最基本的处理元素是权重矩阵,而针对于权重矩阵的稀疏性约束已经有了 L1 和 L2 正则化项可以选择,接下来要面对全连接模块权重分配。

由于全连接[19]是关注信息过多,对于输入数据的处理并不能有效地区分重要的和非重要的数据,因此在处理过程中针对非重要的数据和重要的数据以同等程度对待会引出一个问题,即非重要的数据会对重要数据造成干扰以致神经网络无法进一步提高拟合精度。

而在针对性解决问题时,知道全连接层中前一层的每一个神经元对于下一层的每一个神经元所起的作用是不一样的,而在当前的处理中,它们是以一种均等的方式输入下一层,而在此要做的是基于注意力机制对其进行自学习形式的权重分配,通过外接神经网络的形式来进行权重分配函数的学习。即实现函数:

$$W = f1(v_1, v_2, \cdots, v_n) \tag{1}$$

$$\text{N1\_input} = V * W \tag{2}$$

其中,W 表示全连接层中所有连接的权重向量,f 表示输入与权重之间的映射函数,即注意力分配函数,在这里,由于神经网络对于函数的拟合性较好,在此采用外接神经网络的形式来逼近该注意力分配函数。

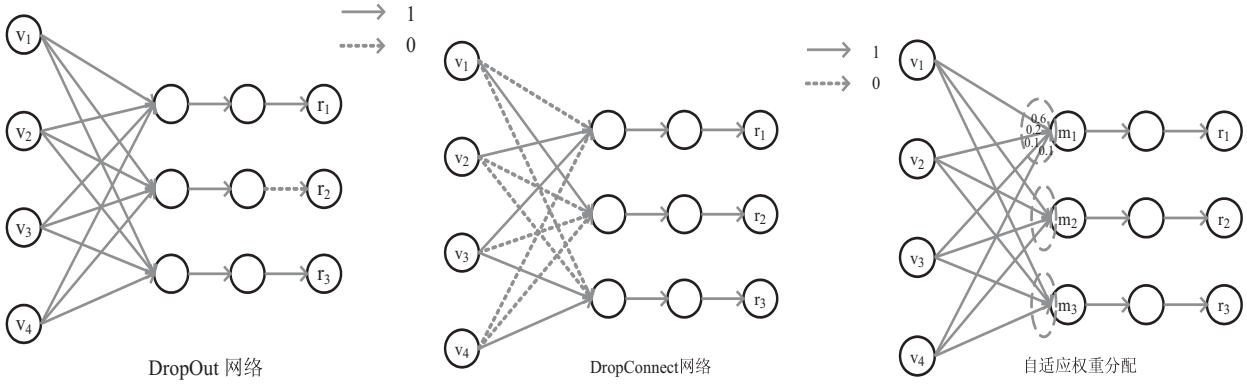

图 1　DropOut,DropConnect 和自适应权重分配

图 1 分别表示 DropOut,DropConnect 和文中方法针对网络连接中的权重调整,DropOut 和 DropConnect 仅仅是针对不同的连接随机进行屏蔽,截断相应的数据流。而文中的自适应权重分配方法则针对神经网络的各个连接设置权重来强化或者抑制某些数据流信息的影响。这种动态的调整是根据数据流内容而不是随机进行,更有助于提升网络的泛化性能。

在训练阶段,由于文中的注意力处理模块时外接神经网络而存在,在训练阶段将采用迭代训练的方式。首先将神经网络结构中的注意力网络全部屏蔽,单独对于原始网络进行充分训练,并优化相应超参数以达到最优的目标效果。第二步将注意力网络添加进入主网络中,固定主网络的所有参数,单独针对于注意力网络进行训练直到效果最优,然后固定注意力网络的参数优化主网络的参数。然后迭代训练直到结果的精度不变为止,最后将主网络和注意力网络参数同时设置为可学习状态进而进行最后的微调。

3　实验过程

在实验中,使用 TensorFlow 来实现文中方法,使用开源代码以及预训练好的模型,在模型中相应的模块上添加设计的模块,进而进行迭代训练。而在对比实验中,以原有实验精度为基础,发现所提出的模型对原有模型的精度具有明显的提升作用。

表 1 展示了针对各个不同数据集所设计的深度学习模型不同层的神经元数量以及施加在其上的注意力矩阵大小,其中结构上第一个数字表示输入的数据特征数,中间的数字是隐层神经元数目,后边的数字是输出向量维数,用于和向量化的标签进行比较。

针对于这种单隐层神经网络,采用全连接作为注意力分配层,使用 sigmoid 激活函数生成注意力分配矩阵,然后将注意力分配矩阵与隐层输出做内积,在输

出层输出结果。

表 1 针对不同数据集的网络参数配置

数据集	结构	注意力层	初始化	最大迭代次数
Balance	5-18-3	[18,3]	glorot_uniform	15 000
Ecoli	8-20-3	[20,3]	glorot_uniform	30 000
Glass	11-22-3	[22,3]	glorot_uniform	40 000
Liver	7-22-2	[22,2]	glorot_uniform	20 000
Sonar	61-20-2	[20,2]	glorot_uniform	15 000
Vehicle	19-30-2	[30,2]	glorot_uniform	40 000

对比了 WDBP(weight decay back propagation)、WEBP(weight elimination back propagation)和 SGLBP(smoothing group lasso BP)方法。它们作为对比算法将和文中算法一起进行对比实验,同时采用相同的数据集和模型配置,不同的地方在于每种算法各自采用的稀疏化方法不一样。从表 2 中可以看出,在大部分情况下,文中方法在该分类任务针对该数据集具有更好的分类效果,但是在耗时上却比较多,原因在于网络设计中添加了基于注意力的权重分配层,相比于原来的网络多了一些需要进行训练的参数,因此,计算量相比于其他网络要大得多,最终导致比其他网络耗时更长。

而且从表 2 中可以看出,在部分情况下该网络的泛化能力更好,在训练阶段的准确率相比于测试阶段并没有高出特别多,同时测试阶段的准确率达到了更好的效果,即训练准确率和测试准确率相差较小,同时训练准确率已经达到了一个比较高的水平,说明网络并没有欠拟合。

从表 2 中可以看出,WEBP、SGLBP 和文中方法具有相似的训练准确度,同时都比 WDBP 高。然而,文中方法具有最好的测试准确度,表明该方法具有更好的泛化能力。

表 2 不同算法结果比较

数据集	算法	训练时间/s	训练准确度	测试准确度	数据集	算法	训练时间/s	训练准确度	测试准确度
Balance	WDBP	6.542 7	0.967 2	0.898 2	Liver	WDBP	6.955 7	0.592 1	0.504 3
	WEBP	6.628 1	0.993 6	0.897 3		WEBP	7.024 2	0.892 6	0.622 0
	SGL-Ⅱ	8.708 4	0.994 7	0.910 5		SGLBP	10.181 2	0.937 0	0.638 4
	our	19.694 2	0.996 0	0.984 0		Our	17.705 4	0.759 3	0.740 4
Ecoli	WDBP	5.381 6	0.790 5	0.698 7	Sonar	WDBP	3.426 6	0.966 1	0.775 8
	WEBP	6.035 1	0.956 0	0.743 1		WEBP	3.565 1	0.984 0	0.760 6
	SGLBP	10.335 7	0.970 5	0.776 2		SGLBP	5.273 6	0.999 8	0.782 9
	our	21.559 3	0.791 5	0.762 4		our	11.045 5	1.000 0	0.808 0
Glass	WDBP	5.285 8	0.951 2	0.853 6	Vehicle	WDBP	11.633 9	0.777 4	0.657 6
	WEBP	5.317 8	0.988 9	0.898 3		WEBP	11.816 9	0.979 0	0.732 5
	SGLBP	8.259 7	1.000 0	0.906 6		SGLBP	14.848 7	0.975 3	0.738 3
	our	18.590 1	0.987 9	0.899 1		our	26.837 9	0.765 2	0.767 7

实验中仅仅是使用这些方法进行了对比,在实际应用中,完全可以将正则化方法中最好的 SGLBP 方法与文中方法进行结合从而更好地提高算法精度。因为相比于正则化方法,文中提出的方法针对的是神经元的权重分配问题,而正则化方法则是针对于稀疏化神经网络,针对神经元进行动态衰减,在其权重低于阈值之后进行裁剪,避免了 DropOut 的随机性。

上述实验为针对简单神经网络并行神经元的权重分配,基于此,可以发现注意力机制是一个比较好的权重分配方法,而在具体实现上拥有诸多变体可供选择,针对不同的任务可以选用不同的具体实现形式。而文中提出的权重分配方法正是基于此,由于注意力的集中性,可以针对每个神经元或者神经层的输入特征进行权重调节,因此,提出的更为泛化的自适应权重调节机制能够有效提升模型表现性能,将单纯针对神经网络输入输出的注意力分配拓展到了整个神经网络内部结构空间,是对神经网络正则化方法的有效补充。而且从实验结果来看,对于提升神经网络的泛化能力同

样具有很大的作用。同时,从实验结果也可以看出,该方法是以增加计算资源的消耗,提升网络模型的复杂性来提高神经网络的表现的,因此需要进一步提升算法效率。

4　结束语

提出了一种自适应的权重分配方法,并针对神经网络结构进行了详细分析和对比实验。实验结果显示,该方法对于不同任务均有相应的性能和准确度的提高。然而该方法还存在一些问题,在其计算权重分配时是通过神经网络模块来计算,增加了较多的计算开销,需要进一步提出更为有效率的权重分配方法。同时可以参考人类的视觉规律,基于信息熵的角度来进行自适应权重分配策略的研究。对于未来的工作,可以将这种自适应的权重分配用于模型的自适应稀疏化,因为表达的稀疏化有助于提升网络的泛化能力,同时,针对注意力机制进行进一步压缩处理,对不同权重进行差距放大处理,通过训练得出权重分布,然后基于一定阈值将网络进行剪枝以实现自动化的模型压缩,同时平衡模型大小与精确度之间的关系。同时,可以利用这种自适应权重分配对于未知的未知特征之间的关系进行挖掘,找出其潜在的函数映射关系。而且,由于整个训练网络可以表征为一个图,训练的神经网络可以表征为一个一个模块,网络对于数据的处理过程可以表征为模块与模块之间的连接,这种连接使用邻接矩阵来进行表征,进而可以使用该方法对于邻接矩阵进行优化,从而优化整个网络数据处理流程。

参考文献:

[1] 葛梦颖,于重重,周　兰,等. 基于协同半监督的深度学习图像分类算法[J]. 计算机仿真,2019,36(2):196-200.

[2] 张思雨. 智能机器人目标检测的深度学习算法研究[D]. 哈尔滨:哈尔滨工程大学,2018.

[3] 路　雪,刘　坤,程永翔. 一种深度学习的非机动车辆目标检测算法[J]. 计算机工程与应用,2019,55(8):182-188.

[4] KWAK N. Principal component analysis based on L1-norm maximization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2008,30(9):1672-1680.

[5] GOTO T,MIYAKURA J,UMEDA K,et al. A robust Spline filter on the basis of L2-norm[J]. Precision Engineering,2005,29(2):157-161.

[6] RENSINK R A. The dynamic representation of scenes[J]. Visual Cognition,2000,7(1-3):17-42.

[7] ITTI L,KOCH C,NIEBUR E. A model of saliency-based visual attention for rapid scene analysis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,1998,20(11):1254-1259.

[8] 李敏学. 基于注意力机制的图像显著区域提取算法分析与比较[D]. 北京:北京交通大学,2011.

[9] MATHE S, SMINCHISESCU C. Action from still image dataset and inverse optimal control to learn task specific visual scanpaths[C]//Proceedings of the 26th international conference on neural information processing systems-Volume 2. Sydney, Australia: Curran Associates Inc., 2013: 1923-1931.

[10] DENIL M,BAZZANI L,LAROCHELLE H,et al. Learning where to attend with deep architectures for image tracking[J]. Neural Computation,2012,24(8):2151-2184.

[11] 周　瑛,刘　越,蔡　俊. 基于注意力机制的微博情感分析[J]. 情报理论与实践,2018,41(3):89-94.

[12] 王　红,史金钏,张志伟. 基于注意力机制的 LSTM 的语义关系抽取[J]. 计算机应用研究,2018,35(5):1417-1420.

[13] GILMER J,SCHOENHOLZ S S,RILEY P F,et al. Neural message passing for quantum chemistry[C]//Proceedings of the 34th international conference on machine learning-volume 70. Sydney,Australia:JMLR,2017:1263-1272.

[14] WAN L,ZEILER M,ZHANG S,et al. Regularization of neural networks using DropConnect[C]//30th international conference on machine learning. GA, USA: JMLR, 2013: 2095-2103.

[15] 吕　伟. 基于稀疏表示和卷积神经网络的水果图像分类与实现[D]. 广州:华南农业大学,2016.

[16] WANG F,JIANG M,QIAN C,et al. Residual attention network for image classification[C]//2017 IEEE conference on computer vision and pattern recognition (CVPR). Hawaii, USA:IEEE,2017:6450-6458.

[17] WOO S,PARK J,LEE J,et al. CBAM:convolutional block attention module[C]//European conference on computer vision. Munich,Germany:[s. n.],2018:3-19.

[18] 王　路,张　璐,李寿山,等. 基于注意力机制的上下文相关的问答配对方法[J]. 中文信息学报,2019,33(1):125-132.

[19] ROSENBLATT F. Principles of neurodynamics perceptrons and the theory of brain mechanisms[J]. American Journal of Psychology,1963,76(4):245-248.