$\frac{\partial e}{\partial \boldsymbol{h}_t}$, 由于输出节点同时作用于下一状态的多个门和输入节点,输出梯度 $\frac{\partial e}{\partial \boldsymbol{h}_t}$ 受多方面因素影响,所以第 1 项乘积结果具有较强的不确定性.第 2 项为上一时刻的记忆单元梯度值与忘记门的乘积,影响因素较少,结果具有稳定性.结合两者的特性,在简单的迭代计算中加入具有动态化的子项,既满足梯度迭代对差异性的需求,又保证了整体的稳定性.

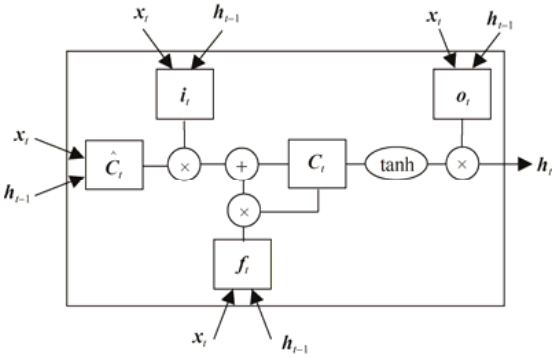

Fig.10 Architecture of long short-term memory unit

图 10 长短期记忆单元结构

LSTM 存在众多变体,Gers 等人为了更精确地利用历史信息,将记忆单元加入到门单元的输入中,提出了窥视孔连接(peephole connections)的方法[38],在丰富门限控制因素的同时增加了梯度传播路径,进一步提高了梯度的稳定性.针对 LSTM 参数过多的问题,Sak 等人[39]采用了投影变换的方法对输出进行线性降维,而 Cho 等人将 LSTM 进行简化,提出 GRU(gated recurrent unit)[40],只设置了重置门(reset gate)和更新门(update gate),实验结果表明,GRU 具有与 LSTM 相同的长期依赖能力.

受到 LSTM 的启发,Srivastava 等人将门限策略迁移到深度前馈神经网络中,提出了 highway network 模型[41,42],该模型利用门限将每层节点的输出分为两部分:其一是前一层传到本层的输入向量 $\boldsymbol{x}_{l-1}$,不进行任何处理直接通过,如同“高速公路”一样;其二是由传统隐层结构拟合的非线性函数 $H(\boldsymbol{x}_{l-1})$的计算结果,具体内容可形式化表述为

$$\boldsymbol{x}_l = \boldsymbol{t}_l \times H(\boldsymbol{x}_{t-1}) + \boldsymbol{c}_l \times \boldsymbol{x}_{l-1} \tag{60}$$

$$\boldsymbol{t}_l = \sigma(\boldsymbol{W}_{tl}\boldsymbol{x}_{l-1} + \boldsymbol{b}_{tl}) \tag{61}$$

$$\boldsymbol{c}_l = \sigma(\boldsymbol{W}_{cl}\boldsymbol{x}_{l-1} + \boldsymbol{b}_{cl}) \tag{62}$$

其中,$\boldsymbol{t}_l$ 和 $\boldsymbol{c}_l$ 作为门限的控制器,由饱和非线性函数产生.与之对应的反向传播表达式为

$$\frac{\partial e}{\partial \boldsymbol{x}_{l-1}} = \frac{\partial \boldsymbol{t}_l}{\partial \boldsymbol{x}_{l-1}}\frac{\partial e}{\partial \boldsymbol{t}_l} + \frac{\partial \boldsymbol{c}_l}{\partial \boldsymbol{x}_{l-1}}\frac{\partial e}{\partial \boldsymbol{c}_l} + \frac{\partial \boldsymbol{x}_l}{\partial \boldsymbol{x}_{l-1}}\frac{\partial e}{\partial \boldsymbol{x}_l} \tag{63}$$

由当前层梯度计算下一层梯度的表达式与 LSTM 类似,由多个项相加组成,在一定程度上降低了梯度不稳定现象发生的可能性.

门限策略虽然在 LSTM 模型和 highway network 模型中发挥着作用,提高了模型的训练效果,但仍有以下几点需要进一步考虑.

(1) 梯度的波动性.门限策略通过改变传统梯度迭代公式,引入更多的动态化因子,每次梯度迭代结果相比于原来梯度,可能更大也可能更小,虽然避免了由于梯度连续增加或减小而导致的梯度不稳定现象,但是梯度传播呈波动形式,这对模型的梯度下降训练具有怎样的影响仍需进一步研究.

(2) 参数的繁重性.每增加一个门,就需要分配相应的权重和偏置,这导致隐层参数成倍增加.特别是在 highway network 模型中,过多的参数将加重内存负担.门限计算也将消耗更多的计算资源,影响训练效率.简化模型、提高效率始终是门限策略待解决的关键问题.

(3) 解释的真实性.门限策略的设计初衷是根据需求对特征进行筛选过滤.但实际上只是对特征的每个维

度上增加比例系数,而比例系数只是由含参线性变换和激活单元产生.神经网络的特征往往隐含在数据内部,特别是在底层网络(前期时刻)中,重构特征往往需要复杂运算,所以门限策略简单的运算除了提高梯度稳定性之外,是否真正能够筛选过滤特征仍需进一步验证.

2.3.2 捷径连接

门限的方法引入了大量的参数,增加了模型的复杂度,并且不能完全避免梯度不稳定现象.在 highway network 中,如果门限值接近于 0,将失去使用门限策略的意义.He 等人采用捷径连接(shortcut connection)的思想,提出了残差网络(residual network,简称 ResNet)[43,44].ResNet 利用非线性网络可以拟合任意函数的特性,使用包含少数隐层的浅层网络拟合自定义的残差函数 $F(\boldsymbol{x})=H(\boldsymbol{x})-\boldsymbol{x}$,残差函数再与恒等映射 $\boldsymbol{x}$ 相加,构成基本的残差单元,实现期望的特征映射关系 $H(\boldsymbol{x})$,如图 11(a)所示.将上述残差单元逐个连接构成深层网络,其前馈传播过程为

$$\boldsymbol{y}_l = F(\boldsymbol{x}_{l-1}) + \boldsymbol{x}_{l-1} \tag{64}$$

$$\boldsymbol{x}_l = \sigma(\boldsymbol{y}_l) \tag{65}$$

在反向传播中,梯度迭代计算表达式为

$$\frac{\partial e}{\partial \boldsymbol{x}_{l-1}} = (1 + F'(\boldsymbol{x}_{l-1})) \times \left[\sigma'(\boldsymbol{y}_l) \times \frac{\partial e}{\partial \boldsymbol{x}_l} \right] \tag{66}$$

(a) 正常残差单元 (b) 预激活残差单元

Fig.11 Architecture of residual unit

图 11 残差单元结构

其中,σ表示 ReLU 激活函数,恒等映射的导数为 1,与传统前馈神经网络的反向梯度迭代运算(见公式(3)和公式(10))相比,$1+F'(\boldsymbol{x}_{l-1})$稳定在“1”附近,降低了权重或卷积核的不确定性导致的梯度不稳定现象,而且在反向传播过程中,恒等映射一直存在,极大程度地缓解了梯度不稳定现象,解决了深层网络比浅层网络更难训练的问题.该团队在 2015 ImageNet 图像识别比赛中利用 152 层的 ResNet 取得最高的图像分类准确率.通过后续研究发现[44],恒等映射相比于定值缩放变换、含参缩放变换、1×1 卷积变换以及 Dropout 更适合作为层间捷径.ReLU 激活函数、BN 算法在数据流中的位置对模型效果有直接影响,将激活函数置于浅层网络中,构成预激活(pre-activation)残差单元,如图 11(b)所示.由该单元组成的残差网络模型训练后的准确度更高.

为了充分发挥捷径连接的优势,有学者尝试将网络中的每一层都连接起来,构成稠密网络(DenseNet)[45],图 12 展示了 4 层稠密卷积网络结构.假设存在 m 层稠密网络,第 l 层有 l 个输入,第 l 层输出连接到后面共 $m-l$ 层,整个网络共有 $\frac{m(m+1)}{2}$ 条连接.层与层全连接的网络使数据信息流途径最大化,丰富了特征提取的内容,可以对之前所有节点的输出进行特征提取和重构.反向传播过程的梯度信息流也更加丰富,缓解了梯度不稳定问题.有学者尝试构建了百层以上的深层稠密网络,并应用在图像识别任务中[45].

捷径连接利用残差单元达到了稳定梯度的目的,但是残差单元提取特征的能力仍需研究.在对传统模型隐层特征函数 $H(x)$没有充分认识的前提下,拟合残差函数 $F(x)$无疑是具有挑战性的.残差单元内部参数和结构设计若仍采用传统方法,能否拟合残差函数,能否有效提取特征,仍需考证.目前有学者对此进行了研究,将残差网络根据连接关系展开,解析展开视图(unraveled view),如图 13 所示.将残差网络分解为不同长度的网络路径,并且证明真正有价值的是长度较短的路径,这与残差网络层数越深效果越好的特点相矛盾[46].这说明,部分残差

单元内部的网络结构作用很小,更多的是依靠捷径连接的恒等映射传递来自底层的特征以及来自高层的梯度,从而实现信息的稳定流动.这违背了残差网络的设计初衷,如何真正拟合残差函数,依靠残差单元提取特征,这不仅需要充分认识特征提取函数,而且需要了解构建网络的函数拟合能力,这些内容均需进一步研究.

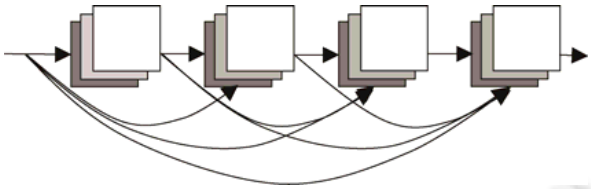

Fig.12 Architecture of DenseNet

图 12 稠密卷积网络结构

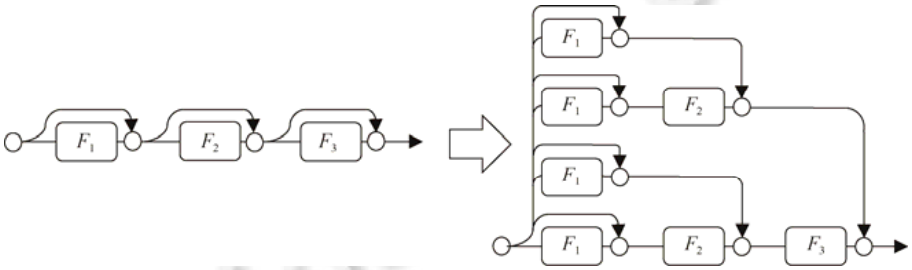

Fig.13 Unraveled view of residual network

图 13 残差网络展开视图

3 梯度不稳定现象未来研究方向

上文依据 3 种策略总结了缓解梯度不稳定现象的研究工作,但梯度不稳定现象依旧存在,导致训练深度神经网络仍然面临着巨大的挑战.根据目前的研究分析,梯度不稳定现象来源于深度神经网络的深层结构,并作用于深度神经网络的梯度下降训练算法,所以未来的研究工作应该从模型设计和训练算法两大方向上展开.

3.1 针对模型设计的研究

(1) 将具有良好梯度稳定特性的模型设计优化思想迁移到更多模型.目前的模型设计优化方法,无论是节点运算或是网络结构,主要思路均是针对某一类模型设计缓解梯度不稳定现象的优化方法,但是这样的方法往往无法直接移植到其他模型,或者移植后无法发挥作用.解决这个问题的思路是:深入理解优化方法的核心思想以及对特定模型有效的原因,再根据其他模型的特点进行调整,从而将优化方法应用到更多模型中,这样做不仅能够检验优化方法的正确性与通用性,还有可能在迁移的过程中产生新的方法或思路,更好地缓解其他模型中的梯度不稳定现象.例如,有学者借鉴前馈神经网络 highway network 和残差网络模型的优秀特性,将其迁移到 LSTM 等循环神经网络结构中,缓解其中的梯度不稳定现象[47−50].所以,在深入理解现有优化方法的基础上,如何将优化思想迁移到更多模型并推陈出新,是未来重要的研究方向之一.

(2) 基于新的梯度稳定性约束条件的模型设计.在参数随机初始化的条件下,训练中各参数的梯度也是不确定的.在本文第 1 节中,论证了在梯度呈现随反向传播逐层递增或递减的不稳定现象时,不利于模型训练.那么梯度应该满足怎样的条件才是稳定且有利于训练的,是否可以通过模型设计使梯度服从这样的约束条件,从而缓解梯度不稳定现象.目前,针对该问题的主要思路是:以分布规律作为梯度稳定性的约束条件,即保证各层梯度的均值和方差是稳定的,并据此设计了相应的参数初始化算法、激活函数等节点运算,目前这方面的研究已经相当完善.最近有学者另辟蹊径,从梯度的 L2 范数入手,结合梯度迭代公式,提出基于酉矩阵的参数结构和激活函数,使得各层梯度的 L2 范数不会随反向传播递增,在一定程度上避免了梯度爆炸现象[51,52].这说明,其他形

式的约束条件也可以指导模型的梯度稳定性设计.因此,从梯度稳定性的约束条件入手,据此设计网络模型,是未来解决梯度不稳定现象的一个重要研究方向.

3.2 针对训练算法的研究

(1) 面向梯度不稳定现象研究新的梯度下降法.当采用梯度下降算法训练深度神经网络时,受梯度不稳定现象的影响,无法有效地更新参数,致使网络训练效果差,说明现有的梯度下降法不适合训练深度神经网络.除了对模型设计优化外,是否可以从训练算法入手,研究新的梯度下降算法,使得在梯度不稳定的条件下,有效训练模型,而且随着不断训练,使各层梯度趋于稳定.目前已有学者提出对梯度采用归一化的方法,强调以梯度方向训练参数,而忽略梯度本身的大小,以定长梯度训练可以避免梯度不稳定现象导致的收敛速度缓慢问题[53,54].而梯度下降算法的关键因素包括梯度和学习率,是否可以通过优化学习率,摆脱梯度不稳定现象的影响.这方面优化方法的提出与证明,不仅需要充分的数学解释,还需要完备的实验验证,这也是一个具有重要价值的研究方向.

(2) 探索无需反向梯度传播的模型训练新算法.深度神经网络训练中发生梯度不稳定现象的原因在于其采用梯度下降法训练,需要以反向传播的方式计算梯度.所以,对训练算法的研究除了改进梯度下降算法之外,尝试颠覆这种传统训练方法,探索其他的训练算法与相应模型,在不需要反向梯度传播的条件下实现模型的训练.深度置信网络的无监督预训练过程即是采用了上述思想[15],在对当前层充分训练后,再开始下一层的训练,不需要从高到低地反向传播梯度,也就避免了梯度不稳定现象的发生.深度置信网络首次挖掘了深度神经网络的巨大潜力,但是由于无监督训练的代价过高,以及针对梯度下降法的不断优化,导致无监督训练逐渐被抛弃.但是该方法为设计无需反向梯度传播的训练算法提供了一个思路,至于更多的无需反向梯度传播的训练算法设计仍需继续研究,包括理论上的收敛性证明以及应用中的模型训练效率.目前,无论是对算法可行性的分析,还是对算法不可行性的证明,相关的结论和研究都比较匮乏,所以这是一个极具挑战性的未来研究方向.

4 结束语

近年来,深度神经网络在机器学习领域展现了优秀的特性,受到了广泛关注.但在其训练过程中发生的梯度不稳定现象,严重影响了模型的学习能力,已经成为制约深度神经网络发展的关键问题.本文针对梯度不稳定现象进行综述:首先,通过分析全连接神经网络、卷积神经网络以及循环神经网络的梯度计算,认为发生梯度不稳定现象的根本原因在于训练中参数的不确定性,并总结了导致梯度不稳定现象的主要因素;然后,通过理论分析与模拟实验,论证了梯度不稳定现象会导致前馈神经网络的多隐层结构失效和收敛速度缓慢,以及循环神经网络的长期依赖问题;之后,以改进训练算法、优化节点运算和调整网络结构这 3 种不同策略,归纳整理了缓解梯度不稳定现象发生的重要方法.结合数学解释,论述了各种方法针对梯度不稳定现象的优化思想;最后,从模型设计与训练算法的角度展望了未来对梯度不稳定现象的研究方向.希望本综述可以让更多的人关注到梯度不稳定现象,并期望与有研究意向的学者共同探索新的研究方向.

References:

[1] Judd S. On the complexity of loading shallow neural networks. Journal of Complexity, 1988,4(3):177–192.

[2] Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. IEEE Trans. on Neural Networks, 1994,5(2):157–166.

[3] Hochreiter S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. Int'l Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 1998,6(2):107–116.

[4] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature, 1986,323(6088):533–536.

[5] Liu SG, Zheng CX, Liu MY. Back propagation algorithm in feedforward neural network and its improvement: Progress and prospect. Computer Science, 1996,23(1):76–79 (in Chinese with English abstract).

[6] Lecun Y, Bottou L, Orr GB, Muller KR. Efficient BackProp. Neural Networks Tricks of the Trade, 1998,1524(1):9–50.

[7] Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-Based learning applied to document recognition. Proc. of the IEEE, 1998,86(11): 2278–2324.

[8] Li YD, Hao ZB, Lei H. Survey of convolutional neural network. Journal of Computer Applications, 2016,36(9):2508–2515 (in Chinese with English abstract).

[9] Elman JL. Finding structure in time. Cognitive Science, 1990,14(2):179–211.

[10] Jordan MI. Serial order: A parallel distributed processing approach. Advances in Psychology, 1997,121:471–495.

[11] Williams RJ, Zipser D. Gradient-Based learning algorithms for recurrent networks and their computational complexity. Backpropagation: Theory, Architectures, and Applications, 1995,1:433–486.

[12] Wang XG, Guo YB, Qi ZQ. The effect of activation function on the performance of BP network and its simulation research. Techniques of Automation & Applications, 2002,21(4):15–17 (in Chinese with English abstract).

[13] Huang Y, Duan XS, Sun SY, Lang W. A study of training algorithm in deep neural networks based on sigmoid activation function. Computer Measurement & Control, 2017,25(2):126–129 (in Chinese with English abstract).

[14] Yuan ZR. Artificial Neural Network and Its Application. 5th ed., Beijing: Tsinghua University Press, 1999 (in Chinese).

[15] Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. Neural Computation, 2006,18(7):1527–1554.

[16] Bengio Y, Lamblin P, Popovici D, Larochelle H. Greedy layer-wise training of deep networks. In: Scholkopf B, ed. Proc. of the Int'l Conf. on Neural Information Processing Systems. MIT Press, 2006. 153–160.

[17] Bengio Y. Practical recommendations for gradient-based training of deep architectures. In: Neural Networks: Tricks of the Trade. 2nd ed., Berlin: Springer-Verlag, 2012. 437–478.

[18] Smolensky P. Information processing in dynamical systems: Foundations of harmony theory. In: Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol.1. MIT Press, 1986. 194–281.

[19] Zhang CX, Ji NN, Wang GW. Restricted Boltzmann machines. Chinese Journal of Engineering Mathematics, 2015,32(2):159–173 (in Chinese with English abstract).

[20] Carreira-Perpinan MA, Hinton GE. On contrastive divergence learning. In: Cowell R, ed. Proc. of the 10th Int'l Workshop on Artificial Intelligence and Statistics. Barbados: The Society for Artificial Intelligence and Statistics, 2005. 33–40.

[21] Chen Y. Research on Chinese information extraction based on deep belief nets [Ph.D. Thesis]. Harbin: Harbin Institute of Technology, 2014 (in Chinese with English abstract).

[22] Sun JG, Jiang JY, Meng XF, Li XJ. Application of deep belief nets in spam filtering. Journal of Computer Applications, 2014,34(4): 1122–1125 (in Chinese with English abstract).

[23] Ranzato M, Boureau Y L, Lecun Y. Sparse feature learning for deep belief networks. In: Platt JC, ed. Proc. of the Advances in Neural Information Processing Systems. New York: Curran Associates Inc., 2007. 1185–1192.

[24] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. Journal of Machine Learning Research, 2010,9:249–256.

[25] Thimm G, Fiesler E. Neural network initialization. In: Mira J, ed. Proc. of the Int'l Workshop on Artificial Neural Networks. Berlin: Springer-Verlag, 1995. 535–542.

[26] He KM, Zhang XY, Ren SQ, Sun J. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In: Proc. of the IEEE Int'l Conf. on Computer Vision (ICCV). IEEE, 2015. 1026–1034.

[27] Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. In: Gordon G, ed. Proc. of the Int'l Conf. on Artificial Intelligence and Statistics. 2011. 315–323.

[28] Attwell D, Laughlin SB. An energy budget for signaling in the grey matter of the brain. Journal of Cerebral Blood Flow & Metabolism, 2001,21(10):1133–1145.

[29] Nair V, Hinton GE. Rectified linear units improve restricted Boltzmann machines. In: Proc. of the Int'l Conf. on Machine Learning. Wisconsin: Omnipress, 2010. 807–814.

[30] Hahnloser RHR, Sarpeshkar R, Mahowald MA, Douglas RS, Seung HS. Digital selection and analogue amplification coexist in acortex-inspired silicon circuit. Nature, 2000,405(6789):947–951.

[31] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Pereira F, ed. Proc. of the Advances in Neural Information Processing Systems. New York: Curran Associates, Inc., 2012. 1097–1105.

[32] Dugas C, Bengio Y, Bélisle F. Incorporating second-order functional knowledge for better option pricing. In: Leen TK, ed. Proc. of the Advances in Neural Information Processing Systems. MIT Press, 2001. 472–478.

[33] Maas AL, Hannun AY, Ng AY. Rectifier nonlinearities improve neural network acoustic models. In: Dasgupta S, ed. Proc. of the Int'l Conf. on Machine Learning. 2013.

[34] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Bach F, ed. Proc. of the 32nd Int'l Conf. on Machine Learning. 2015. 448–456.

[35] Laurent C, Pereyra G, Brakel P, Ying Z, Bengio Y. Batch normalized recurrent neural networks. In: Proc. of the IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. IEEE, 2016. 2657–2661.

[36] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997,9(8):1735–1780.

[37] Gers FA, Schmidhuber J, Cummins F. Learning to forget: Continual prediction with LSTM. Neural Computation, 2000,12(10): 2451–2471.

[38] Gers FA, Schmidhuber J. Recurrent nets that time and count. In: Proc. of the IEEE-INNS-ENNS Int'l Joint Conf. on Neural Networks. IEEE, 2000. 189–194.

[39] Sak H, Senior A, Beaufays F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In: Proc. of the 15th Annual Conf. of the Int'l Speech Communication Association. 2014. 338–342.

[40] Cho K, Van MB, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing. 2014. 1724–1734.

[41] Srivastava RK, Greff K, Schmidhuber J. Training very deep networks. In: Cortes C, ed. Proc. of the Advances in Neural Information Processing Systems. New York: Curran Associates, Inc., 2015. 2377–2385.

[42] Srivastava RK, Greff K, Schmidhuber J. Highway networks. arXiv: 1505.00387, 2015.

[43] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. IEEE, 2016. 770–778.

[44] He KM, Zhang XY, Ren SQ, Sun J. Identity mappings in deep residual networks. In: Leibe B, ed. Proc. of the 14th European Conf. on Computer Vision. Berlin: Springer-Verlag, 2016. 630–645.

[45] Huang G, Liu Z, Weinberger KQ, *et al*. Densely connected convolutional networks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. IEEE, 2017. 2261–2269.

[46] Veit A, Wilber M, Belongie S. Residual networks behave like ensembles of relatively shallow networks. In: Lee DD, ed. Proc. of the Advances in Neural Information Processing Systems. New York: Curran Associates, Inc., 2016. 550–558.

[47] Yao K, Cohn T, Vylomova K, Duh K, Dyer C. Depth-Gated LSTM. arXiv: 1508.03790, 2015.

[48] Prakash A, Hasan SA, Lee K, Datla V, Qadir A, Liu J, Farri O. Neural paraphrase generation with stacked residual LSTM networks. In: Proc. of the 26th Int'l Conf. on Computational Linguistics. 2016. 2923–2934.

[49] Zhang Y, Chen GG, Yu D, Yaco K, Khudanpur S, Glass J. Highway long short-term memory RNNS for distant speech recognition. In: Proc. of the IEEE Conf. on Acoustics, Speech and Signal Processing. IEEE, 2016. 5755–5759.

[50] Kim J, El-Khamy M, Lee J. Residual LSTM: Design of a deep recurrent architecture for distant speech recognition. In: Proc. of the Conf. of the Int'l Speech Communication Association. 2017. 1591–1595.

[51] Arjovsky M, Shah A, Bengio Y. Unitary evolution recurrent neural networks. In: Balcan MF, ed. Proc. of the Int'l Conf. on Machine Learning. 2016. 1120–1128.

[52] Henaff M, Szlam A, Lecun Y. Recurrent orthogonal networks and long-memory tasks. In: Balcan MF, ed. Proc. of the Int'l Conf. on Machine Learning. 2016. 2034–2042.

[53] Hazan E, Levy KY, Shalevshwartz S. Beyond convexity: Stochastic quasi-convex optimization. In: Cortes C, ed. Proc. of the Advances in Neural Information Processing Systems. New York:Curran Associates, Inc., 2015. 1594–1602.

[54] Yu AW, Lin Q, Salakhutdinov R, Carbonell J. Normalized gradient with adaptive stepsize method for deep neural network training. arXiv: 1707.04822, 2017.

附中文参考文献:

[5] 刘曙光,郑崇勋,刘明远.前馈神经网络中的反向传播算法及其改进:进展与展望.计算机科学,1996,23(1):76–79.

[8] 李彦冬,郝宗波,雷航.卷积神经网络研究综述.计算机应用,2016,36(9):2508–2515.

[12] 王雪光,郭艳兵,齐占庆.激活函数对 BP 网络性能的影响及其仿真研究.自动化技术与应用,2002,21(4):15–17.

[13] 黄毅,段修生,孙世宇,郎巍.基于改进 sigmoid 激活函数的深度神经网络训练算法研究.计算机测量与控制,2017,25(2):126–129.

[14] 袁曾任.人工神经元网络及其应用.北京:清华大学出版社,1999.

[19] 张春霞,姬楠楠,王冠伟.受限波尔兹曼机.工程数学学报,2015,32(2):159–173.

[21] 陈宇.基于深度置信网络的中文信息抽取方法[博士学位论文].哈尔滨:哈尔滨工业大学,2014.

[22] 孙劲光,蒋金叶,孟祥福,李秀娟.深度置信网络在垃圾邮件过滤中的应用.计算机应用,2014,34(4):1122–1125.

陈建廷(1995－),男,吉林省吉林市人,硕士生,主要研究领域为数据挖掘.

向阳(1962－),男,博士,教授,博士生导师,CCF 专业会员,主要研究领域为数据挖掘.