Table 2 Overall compression performance

表 2 总体压缩效果

参考文献	模型	Top-1 准确率(%)	参数(byte)	压缩率(%)	运行效率提升
Ref.[24]	VGG-16 基准	93.25	1.50E+07	–	使用 FLOP 评价指标,对不同深度模型,分别获得 34.20%,13.70%,38.60%, 15.5%的加速
	VGG-16 压缩	93.4	5.40E+06	64.00	
	ResNet-56 基准	93.04	8.60E+05	–	
	ResNet-56 压缩	93.06	7.30E+05	15.12	
	ResNet-110 基准	93.53	1.72E+06	–	
	ResNet-110 压缩	93.3	1.16E+06	32.56	
	ResNet-34 基准	73.23	2.16E+07	–	
	ResNet-34 压缩	72.56	1.99E+07	7.87	
Ref.[20]	LeNet 基准	99.06	1.96E+07	–	由于全连接层对网络运行时间影响不大,所以对其剪枝不能显著提升运行效率
	基于幅度剪枝	96.5	1.65E+07	16.01	
	随机剪枝	91.37	1.65E+07	16.01	
	数据无关剪枝	98.35	1.65E+07	16.01	
	AlexNet 基准	57.84	6.09E+07	–	
	对 FC6 数据无关剪枝	56.08	4.23E+07	30.57	
	对 FC7 数据无关剪枝	56	5.37E+07	11.80	
	对 FC6 和 FC7 数据无关剪枝	55.6	3.97E+07	34.89	
Ref.[25]	基于 CASIA 数据的基准模型	86.04	1.30E+04	–	使用实验测得平均前向传播所需时间作为评价指标,对于不同深度模型,分别获得 59.51%, 25.93%,37.11%,57.81%, 62.26%的加速
	Fitnet	82.08	8.64E+03	33.33	
	低秩分解	84.79	1.17E+04	9.91	
	Inbound Prune	84.74	1.16E+04	10.71	
	RR Prune	85.08	6.14E+03	52.61	
	Hyb. Prunne	85.32	6.06E+03	53.27	
Ref.[22]	LeNet-5 基准	80	4.31E+05	–	–
	剪枝+量化	77	3.60E+04	91.67	
	AlexNet 基准	57.22	6.10E+07	–	
	剪枝+量化	57.22	6.70E+06	88.89	
	VGG-16 基准	31.5	1.38E+08	–	
	剪枝+量化	31.34	1.03E+07	92.31	
Ref.[21]	AlexNet	57.22	2.40E+08	–	引入了非结构化的稀疏性,需要专门的软件计算库或者未来的硬件获得运行效率的提升
	Fastfood-32-AD	58.07	1.31E+02	50.00	
	Fastfood-16-AD	57.1	6.40E+07	72.97	
	Collins&Kohli	55.6	6.10E+07	75.00	
	SVD	55.98	4.78E+07	80.00	
	剪枝	57.22	8.90E+06	88.89	
	剪枝+量化	57.22	6.90E+06	96.30	
	剪枝+量化+编码	57.22	6.90E+06	98.97	
	VGG-16 基准	68.5	5.52E+08	–	
	VGG-16 压缩	68.83	1.13E+07	97.95	
Ref.[37]	Teacher	90.18	9.00E+06	–	使用 MULT 评价指标,对 FitNet1,FitNet2,FitNet3, FitNet4, 分别获得了 92.51%,78.45%,27.01%, 34.21%的加速(FitNet1~FitNet4 采用不同的人工设计的网络架构)
	FitNet1	89.01	2.50E+05	97	
	FitNet2	91.06	8.62E+05	90.42	
	FitNet3	91.1	1.60E+06	82.22	
	FitNet4	91.61	2.50E+06	72.22	
	Mimic single	84.6	5.40E+07	–	
	Mimic ensemble	85.8	7.00E+07	–	

5.3 分层压缩效用评价

在网络分解的研究中,针对深度网络中的卷积层和全连接层通常采用不同的方法.因而,逐层分析可以帮助我们获得一些经验性的规律.逐层分析实验包含两种方法:(1) 逐层对照实验:固定其余所有层,每次对一层进行压缩.从而可以对压缩率、效率提升进行全面的评价,见表 3;(2) 只进行一次实验,统计压缩前和压缩后的各层参数数量.从而获得压缩率指标,见表 4.

在网络分解的逐层对照实验中,对卷积层的压缩率在不断的改进中不断提升,但是总体上小于对全连接层的 SVD 分解.稀疏卷积操作采用固定卷积核矩阵的方法,对训练好的卷积核矩阵的稀疏性具有一定的依赖性,

但总体上更好地利用了卷积核的稀疏性,提升了运行效率.

Table 3 Performance evaluations of compressing single convolutional layer or fully-connected layer

表 3 卷积层和全连接层逐层压缩对照实验

参考文献	压缩层类别	压缩方法	压缩率(%)	准确率下降比(%)	运行加速比(%)
Ref.[45]	Conv	Monochromatic, *C*=4	87.90	1.90	66.33
	Conv	Monochromatic, *C*=6	84.52	0.43	66.10
	Conv	Monochromatic, *C*=8	81.13	0.20	65.99
	Conv	Monochromatic, *C*=12	74.36	0	65.64
	Conv	双聚类+外积分解	92.54	0.68	23.07
	Conv	双聚类+SVD,	71.42	0.90	37.50
	FC	SVD, *K*=250	92.54	0.84	20.19
	FC	SVD, *K*=950	71.42	0.09	17.31
Ref.[46]	Conv	Sparse CNN, kernel size=11	92.70	0.62	61.69
	Conv	Sparse CNN, kernel size=5	95.00	1.43	85.99
	Conv	Low rank, kernel size=5	89.00	0.61	60.00

在权重数量上,Han 进行了系统的实验,从中可以总结出一些经验性的结论.全连接层参数密集,不仅参数数量比卷积层高出一个数量级,而且对最终压缩率的贡献也远超过卷积层.同时,文献[22]中的方法进一步改进后,采用第 4 节中提到的量化编码方式,最高可以达到 97.95%的压缩率.

Table 4 Compression rate of convolutional layers and fully-connected layers

表 4 卷积层和全连接层的压缩率

参考文献	压缩层类别	压缩方法	压缩率(%)	总体压缩率(%)
Ref.[21]	AlexNet,Conv	剪枝+量化	67.40	88.89
	AlexNet,Conv	剪枝+量化+编码	79.47	
	AlexNet,FC	剪枝+量化	97.00	
	AlexNet,FC	剪枝+量化+编码	97.61	
	VGGNet,Conv	剪枝+量化	60.00	92.31
	VGGNet,Conv	剪枝+量化+编码	70.03	
	VGGNet,FC	剪枝+量化	98.40	
	VGGNet,FC	剪枝+量化+编码	98.90	
Ref.[22]	VGGNet,Conv	剪枝	76.20	98.97
	VGGNet,FC	剪枝	89.80	97.95

6 未来研究方向

网络剪枝、网络精馏和网络分解都能在一定程度上实现网络压缩的目的.回归到深度网络压缩的本质目的上,即提取网络中的有用信息,以下是一些值得研究和探寻的方向.

(1) 权重参数对结果的影响度量.深度网络的最终结果是由全部的权重参数共同作用形成的,目前,关于单个卷积核/卷积核权重的重要性的度量仍然是比较简单的方式,尽管文献[14]中给出了更为细节的分析,但是由于计算难度大,并不实用.因此,如何通过更有效的方式来近似度量单个参数对模型的影响,具有重要意义.

(2) 学生网络结构的构造.学生网络的结构构造目前仍然是由人工指定的,然而,不同的学生网络结构的训练难度不同,最终能够达到的效果也有差异.因此,如何根据教师网络结构设计合理的网络结构在精简模型的条件下获取较高的模型性能,是未来的一个研究重点.

(3) 参数重建的硬件架构支持.通过分解网络可以无损地获取压缩模型,在一些对性能要求高的场景中是非常重要的.然而,参数的重建步骤会拖累预测阶段的时间开销,如何通过硬件的支持加速这一重建过程,将是未来的一个研究方向.

(4) 任务或使用场景层面的压缩.大型网络通常是在量级较大的数据集上训练完成的,比如,在 ImageNet 上训练的模型具备对 1 000 类物体的分类,但在一些具体场景的应用中,可能仅需要一个能识别其中

几类的小型模型.因此,如何从一个全功能的网络压缩得到部分功能的子网络,能够适应很多实际应用场景的需求.

(5) 网络压缩效用的评价.目前,对各类深度网络压缩算法的评价是比较零碎的,侧重于和被压缩的大型网络在参数量和运行时间上的比较.未来的研究可以从提出更加泛化的压缩评价标准出发,一方面平衡运行速度和模型大小在不同应用场景下的影响;另一方面,可以从模型本身的结构性出发,对压缩后的模型进行评价.

7 结束语

网络压缩旨在减少模型参数、降低存储空间和减少运算开销,在深度网络的实际应用中发挥着越来越重要的作用.本文总结了网络剪枝、网络精馏和网络分解这 3 个方向的压缩方法,并在压缩性能上提供了相应指标评价.其中,网络剪枝关注于去掉模型中影响较小的卷积结构;网络精馏侧重于训练一个较小的学生网络结构去模拟较大的教师网络的性能;网络分解强调从数据存储方式和结构以及运算优化的角度来降低开销.希望通过以上介绍,读者能够对深度网络压缩有一个较为全面的了解,并在相关基于深度模型的实际任务中加以利用.

References:

[1] Le Cun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015,521(7553):436−444. [doi: 10.1038/nature14539]

[2] Plamondon R, Srihari SN. Online and off-line handwriting recognition: A comprehensive survey. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2000,22(1):63−84. [doi: 10.1109/34.824821]

[3] Wan J, Wang D, Hoi SCH, Wu P, Zhu J, Zhang Y, Li J. Deep learning for content-based image retrieval: A comprehensive study. In: Proc. of the 22nd ACM Int'l Conf. on Multimedia (MM). Orlando: ACM Press, 2014. 157−166. [doi: 10.1145/2647868.2654948]

[4] Girshick R. Fast *r*-cnn. In: Proc. of the IEEE Int'l Conf. on Computer Vision (ICCV). Santiago: IEEE, 2015. 1440−1448. [doi: 10.1109/iccv.2015.169]

[5] Wang N, Yeung DY. Learning a deep compact image representation for visual tracking. In: Proc. of the Advances in Neural Information Processing Systems (NIPS). Tahoe: IEEE, 2013. 809−817.

[6] Severyn A, Moschitti A. Learning to rank short text pairs with convolutional deep neural networks. In: Proc. of the 38th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Santiago: ACM Press, 2015. 373−382. [doi: 10.1145/2766462.2767738]

[7] Ngiam J, Coates A, Lahiri A, Prochnow B, Ng AY. On optimization methods for deep learning. In: Proc. of the 28th Int'l Conf. on Machine Learning (ICML). Bellevue: ACM Press, 2011. 265−272.

[8] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Proc. of the Advances in Neural Information Processing Systems (NIPS). Tahoe: IEEE, 2012. 1097−1105.

[9] Sercu T, Puhrsch C, Kingsbury B, Le Cun Y. Very deep multilingual convolutional neural networks for LVCSR. In: Proc. of the Acoustics, Speech and Signal Processing (ICASSP). Shanghai: IEEE, 2016. 4955−4959. [doi: 10.1109/icassp.2016.7472620]

[10] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 770−778. [doi: 10.1109/cvpr.2016.90]

[11] Setiono R, Liu H. Neural-Network feature selector. IEEE Trans. on Neural Networks, 1997,8(3):654−662. [doi: 10.1109/72.572104]

[12] Hanson SJ, Pratt LY. Comparing biases for minimal network construction with back-propagation. In: Proc. of the Advances in Neural Information Processing Systems (NIPS). Denver: IEEE, 1989. 177−185.

[13] Whitley D, Starkweather T, Bogart C. Genetic algorithms and neural networks: Optimizing connections and connectivity. Parallel Computing, 1990,14(3):347−361. [doi: 10.1016/0167-8191(90)90086-o]

[14] Oberman SF, Flynn MJ. Design issues in division and other floating-point operations. IEEE Trans. on Computers, 1997,46(2):154−161. [doi: 10.1109/12.565590]

[15] Anwar S, Sung WY. Coarse pruning of convolutional neural networks with random masks. In: Proc. of the Int'l Conf. on Learning and Representation (ICLR). IEEE, 2017. 134−145.

[16] Le Cun Y, Denker JS, Solla SA. Optimal brain damage. In: Proc. of the Advances in Neural Information Processing Systems (NIPS). Denver: IEEE, 1989. 598−605.

[17] Rosenblueth E. Point estimates for probability moments. Proc. of the National Academy of Sciences, 1975,72(10):3812−3814. [doi: 10.1073/pnas.72.10.3812]

[18] Hassibi B, Stork DG, Wolff GJ. Optimal brain surgeon and general network pruning. In: Proc. of the Int'l Conf. on Neural Networks (ICNN). San Francisco: IEEE, 1993. 293−299. [doi: 10.1109/icnn.1993.298572]

[19] Hassibi B, Stork DG. Second order derivatives for network pruning: Optimal brain surgeon. In: Proc. of the Advances in Neural Information Processing Systems (NIPS). Denver: IEEE, 1993. 164−171.

[20] Srinivas S, Babu RV. Data-Free parameter pruning for deep neural networks. In: Proc. of the 26th British Machine Vision Conf. (BMVC). Swansea: IEEE, 2015. 120−129. [doi: 10.5244/c.29.31]

[21] Han S, Mao H, Dally WJ. Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. In: Proc. of the Int'l Conf. on Learning and Representation (ICLR). San Juan: IEEE, 2016. 233−242.

[22] Han S, Pool J, Tran J, Dally WJ. Learning both weights and connections for efficient neural network. In: Proc. of the Advances in Neural Information Processing Systems. Montreal: IEEE, 2015. 1135−1143.

[23] Anwar S, Hwang K, Sung W. Structured pruning of deep convolutional neural networks. ACM Journal on Emerging Technologies in Computing Systems (JETC), 2017,13(3):Article No.32. [doi: 10.1145/3005348]

[24] Li H, Kadav A, Durdanovic I, Samet H, Graf HP. Pruning filters for efficient ConvNets. In: Proc. of the Int'l Conf. on Learning and Representation (ICLR). IEEE, 2017. 34−42.

[25] Polyak A, Wolf L. Channel-Level acceleration of deep face representations. IEEE Access, 2015,3:2163−2175. [doi: 10.1109/access.2015.2494536]

[26] Figurnov M, Ibraimova A, Vetrov DP, Kohli P. PerforatedCNNs: Acceleration through elimination of redundant convolutions. In: Proc. of the Advances in Neural Information Processing Systems (NIPS). Barcelona: IEEE, 2016. 947−955.

[27] Hu H, Peng R, Tai YW, Tang CK. Network trimming: A data-driven neuron pruning approach towards efficient deep architectures. In: Proc. of the Int'l Conf. on Learning and Representation (ICLR). IEEE, 2017. 214−222.

[28] Molchanov P, Tyree S, Karras T, Aila T, Kautz J. Pruning convolutional neural networks for resource efficient transfer learning. In: Proc. of the Int'l Conf. on Learning and Representation (ICLR). IEEE, 2017. 324−332.

[29] Rueda FM, Grzeszick R, Fink GA. Neuron pruning for compressing deep networks using maxout architectures. In: Proc. of the German Conf. on Pattern Recognition (GCPR). Saarbrücken: Springer-Verlag, 2017. 110−120. [doi: 10.1007/978-3-319-66709-6_15]

[30] Zhou ZH. Rule extraction: Using neural networks or for neural networks? Journal of Computer Science and Technology, 2004, 19(2):249−253. [doi: 10.1007/BF02944803]

[31] Zhou ZH, Jiang Y. NeC4.5: Neural ensemble based C4.5. IEEE Trans. on Knowledge and Data Engineering, 2004,16(6):770−773.

[32] Buciluǎ C, Caruana R, Niculescu-Mizil A. Model compression. In: Proc. of the 12th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Philadelphia: ACM Press, 2006. 535−541. [doi: 10.1145/1150402.1150464]

[33] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. In: Proc. of the Advances in Neural Information Processing Systems (NIPS). Montrea: IEEE, 2014. 2644−2652.

[34] Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans. on Knowledge and Data Engineering, 2010,22(10):1345−1359. [doi: 10.1109/TKDE.2009.191]

[35] Lowd D, Domingos P. Naive Bayes models for probability estimation. In: Proc. of the 22nd Int'l Conf. on Machine Learning (ICML). Bonn: ACM Press, 2005. 529−536. [doi: 10.1145/1102351.1102418]

[36] Ba J, Caruana R. Do deep nets really need to be deep? In: Proc. of the Advances in Neural Information Processing Systems (NIPS). Montrea: IEEE, 2014. 2654−2662.

[37] Romero A, Ballas N, Kahou SE, Chassang A, Gatta C, Bengio Y. Fitnets: Hints for thin deep nets. In: Proc. of the Int'l Conf. on Learning and Representation (ICLR). IEEE, 2017. 124−133.

[38] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015. 1−9. [doi: 10.1109/cvpr.2015.7298594]

[39] Chen T, Goodfellow I, Shlens J. Net2net: Accelerating learning via knowledge transfer. In: Proc. of the Int'l Conf. on Learning and Representation (ICLR). San Juan: IEEE, 2016. 27−35.

[40] Li Z, Hoiem D. Learning without forgetting. In: Proc. of the European Conf. on Computer Vision (ECCV). Amsterdam: Springer Int'l Publishing, 2016. 614−629. [doi: 10.1007/978-3-319-46493-0_37]

[41] He ZF, Yang M, Liu HD. Joint learning of multi-label classification and label correlations. Ruan Jian Xue Bao/Journal of Software, 2014,25(9):1967−1981 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4634.htm [doi: 10.13328/j.cnki.jos.004634]

[42] Golub GH, Reinsch C. Singular value decomposition and least squares solutions. Numerische Mathematik, 1970,14(5):403−420. [doi: 10.1007/BF02163027]

[43] Zhang M, Ge WH. Overlap bicuster algorithm based on probability. Computer Engineering and Design, 2012,33(9):3579−3583 (in Chinese with English abstract). [doi: 10.16208/j.issn1000-7024.2012.09.046]

[44] Jaderberg M, Vedaldi A, Zisserman A. Speeding up convolutional neural networks with low rank expansions. In: Proc. of the 26th British Machine Vision Conf. (BMVC). Swansea: IEEE, 2015. 100−109. [doi: 10.5244/c.28.88]

[45] Denton EL, Zaremba W, Bruna J, Le Cun Y, Fergus R. Exploiting linear structure within convolutional networks for efficient evaluation. In: Proc. of the Advances in Neural Information Processing Systems (NIPS). Montrea: IEEE, 2014. 1269−1277.

[46] Liu B, Wang M, Foroosh H, Tappen M, Penksy M. Sparse convolutional neural networks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015. 806−814. [doi: 10.1109/cvpr.2015.7298681]

[47] Courbariaux M, Bengio Y, David JP. Binaryconnect: Training deep neural networks with binary weights during propagations. In: Proc. of the Advances in Neural Information Processing Systems (NIPS). Montreal: IEEE, 2015. 3123−3131.

[48] Gong Y, Liu L, Yang M, Bourdev L. Compressing deep convolutional networks using vector quantization. In: Proc. of the Int'l Conf. on Learning and Representation (ICLR). Toronto: IEEE, 2015. 102−110.

[49] Lee H, Battle A, Raina R, Ng AY. Efficient sparse coding algorithms. In: Proc. of the Advances in Neural Information Processing Systems (NIPS). IEEE, 2007. 789−801.

[50] Mairal J, Bach F, Ponce J, Sapiro G. Online dictionary learning for sparse coding. In: Proc. of the 26th Annual Int'l Conf. on Machine Learning (ICML). Montreal: ACM Press, 2009. 689−696. [doi: 10.1145/1553374.1553463]

[51] Zhou A, Yao A, Guo Y, Xu L, Chen Y. Incremental network quantization: Towards lossless cnns with low-precision weights. In: Proc. of the Int'l Conf. on Learning and Representation (ICLR). IEEE, 2017. 154−162.

[52] Monmasson E, Cirstea MN. FPGA design methodology for industrial control systems—A review. IEEE Trans. on Industrial Electronics, 2007,54(4):1824−1842. [doi: 10.1109/tie.2007.898281]

[53] Gupta S, Agrawal A, Gopalakrishnan K, Narayanan P. Deep learning with limited numerical precision. In: Proc. of the Int'l Conf. on Machine Learning (ICML). Lille: ACM Press, 2015. 1737−1746.

[54] Antipov G, Berrani SA, Dugelay JL. Minimalistic CNN-based ensemble model for gender prediction from face images. Pattern Recognition Letters, 2016,70:59−65. [doi: 10.1016/j.patrec.2015.11.011]

附中文参考文献:

[41] 何志芬,杨明,刘会东.多标记分类和标记相关性的联合学习.软件学报,2014,25(9):1967−1981. http://www.jos.org.cn/1000-9825/4634.htm [doi: 10.13328/j.cnki.jos.004634]

[43] 张敏,戈文航.基于概率计算的重叠双聚类算法.计算机工程与设计,2012,33(9):3579−3583. [doi: 10.16208/j.issn1000-7024.2012.09.046]

雷杰(1991－),男,湖北仙桃人,博士生,主要研究领域为计算机视觉,深度学习.

王兴路(1996－),男,本科生,主要研究领域为计算机视觉,深度学习.

高鑫(1992－),男,硕士生,主要研究领域为计算机视觉,深度学习.

宋明黎(1976－),男,博士,教授,博士生导师,CCF 专业会员,主要研究领域为计算机视觉,深度学习.

宋杰(1991－),男,博士生,主要研究领域为计算机视觉,深度学习.