要有减少卷积核的数量、减少特征的通道数以及设计更高效的卷积操作等关键技术,但是非常依赖设计者的经验.如何有效地将针对特定问题的先验知识加入到模型构建过程中,是未来研究的重点方向.通过网络剪枝、权重压缩和低秩分解是对已有的网络进行压缩,但是压缩算法需要设计者探索较大的设计空间以及在模型大小、速度和准确率之间权衡.为了减少人为因素的干扰,自动机器学习技术是未来研究的热点,联合优化深度神经网络流程的所有模型参数.神经网络架构搜索的研究主要集中在深度神经网络上,许多搜索架构都源自 NASNet[6]搜索空间,通过各种搜索算法在定义的搜索空间内自动生成的,广泛应用于解决图像识别、图像分割和语言建模等任务[6,7,98,99],但是只能针对某一特定或同一类型的数据集.如何使用跨不同数据集的知识来加速优化过程,是未来研究的热点.其他的挑战是联合优化深度神经网络流程的所有模型参数.到目前为止,深度神经网络的通用自动化仍处于起步阶段,许多问题尚未得到解决.然而,这仍然是一个令人兴奋的领域,并且未来的工作的方向需要强调其突出的实用性.

轻量级模型的发展使得神经网络更加高效,从而能够广泛地应用到各种场景任务中.一方面,轻量级神经网络有更小的体积和计算量,降低了对设备存储能力和计算能力的需求,既可以装配到传统家电中使其更加智能化,也可以将深度学习系统应用在虚拟现实、增强现实、智能安防和智能可穿戴设备等新兴技术中;另一方面,轻量级神经网络具有更快的运行速度和更短的延时,能够对任务进行实时处理,对于在线学习、增量学习和分布式学习有重大意义;另外,实时处理的神经网络能够满足自动驾驶技术的需求,提高自动驾驶的安全性.轻量级神经网络将对于人工智能技术的普及、建立智能化城市起不可或缺的作用.

References:

[1] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proc. of the 3rd Int'l Conf. on Learning Representations. 2015.

[2] 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. 2016. 770−778.

[3] Huang G, Liu Z, Maaten LVD, Weinberger K. Densely connected convolutional networks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 4700−4708.

[4] Han S, Mao H, Dally WJ. Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding. In: Proc. of the 4th Int'l Conf. on Learning Representations. 2016.

[5] He Y, Lin J, Liu Z, Wang H, Li L, Han S. AMC: Automl for model compression and acceleration on mobile devices. In: Proc. of the European Conf. on Computer Vision. 2018. 784−800.

[6] Zoph B, Vasudevan V, Shlens JV, Le Q. Learning transferable architectures for scalable image recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 8697−8710.

[7] Tan M, Chen B, Pang R, Vasudevan V, Sandler M, Howard AV, Le Q. Mnasnet: Platform-aware neural architecture search for mobile. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2019. 2820−2828.

[8] Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 432−445.

[9] Zhang X, Zhou X, Lin M, Sun J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 6848−6856.

[10] Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L. MobileNetV2: Inverted residuals and linear bottlenecks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 4510−4520.

[11] Qin Z, Li Z, Zhang Z, Bao Y, Yu G, Peng Y, Sun J. ThunderNet: Towards real-time generic object detection on mobile devices. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2019. 6718−6727.

[12] Ma N, Zhang X, Zheng H, Sun J. ShuffleNet V2: Practical guidelines for efficient cnn architecture design. In: Proc. of the European Conf. on Computer Vision. 2018. 116−131.

[13] Landola FN, Han S, Moskewicz MW, Ashraf K, Han S, Dally WJ, Keutzer K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. In: Proc. of the 5th Int'l Conf. on Learning Representations. 2017.

[14] Bergstra JS, Bardenet R, Bengio Y, Kégl B. Algorithms for hyper-parameter optimization. In: Proc. of the Advances in Neural Information Processing Systems. 2011. 2546−2554.

[15] Yoon J, Kim T, Dia O, Kim S, Bengio Y, Ahn S. Bayesian model-agnostic meta-learning. In: Proc. of the Advances in Neural Information Processing Systems. 2018. 7332−7342.

[16] Jaderberg M, Vedaldi A, Zisserman A. Speeding up convolutional neural networks with low rank expansions. In: Proc. of the British Machine Vision Conf. 2014.

[17] Peng C, Zhang X, Yu G, Luo G, Sun J. Large kernel matters—Improve semantic segmentation by global convolutional network. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 4353−4361.

[18] Rizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Proc. of the Advances in Neural Information Processing Systems. 2012. 1097−1105.

[19] Xie S, Girshick R, Dollár P, Tu Z, He K. Aggregated residual transformations for deep neural networks. In: Proc. of the IEEE Conf. on Computer Vision And Pattern Recognition. 2017. 1492−1500.

[20] Zhang T, Qi GJ, Xiao B, Wang J. Interleaved group convolutions. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 4373−4382.

[21] Xie G, Wang J, Zhang T, Lai J, Hong R, Qi GJ. Interleaved structured sparse convolutional neural networks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 8847−8856.

[22] Mehta S, Rastegari M, Caspi A, Shapiro L, Hajishirzi H. Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In: Proc. of the European Conf. on Computer Vision. 2018. 552−568.

[23] Mehta S, Rastegari M, Shapiro L, Hajishirzi H. Espnetv2: A light-weight, power efficient, and general purpose convolutional neural network. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2019. 9190−9200.

[24] Li H, Kadav A, Durdanovic I, Samet H, Graf HP. Pruning filters for efficient convnets. In: Proc. of the 5th Int'l Conf. on Learning Representations. 2017.

[25] Liu Z, Li J, Shen Z, Huang G, Yan S, Zhang C. Learning efficient convolutional networks through network slimming. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 2736−2744.

[26] Hu H, Peng R, Tai YW, Tang CK. Network trimming: A data-driven neuron pruning approach towards efficient deep architectures. In: Proc. of the 5th Int'l Conf. on Learning Representations. 2015.

[27] Tian Q, Arbel T, Clark JJ. Deep LDA-pruned nets for efficient facial gender classification. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops. 2017. 10−19.

[28] Molchanov P, Tyree S, Karras T, Aila T, Kautz J. Pruning convolutional neural networks for resource efficient inference. In: Proc. of the 5th Int'l Conf. on Learning Representations. 2018.

[29] Luo JH, Wu J, Lin W. Thinet: A filter level pruning method for deep neural network compression. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 5058−5066.

[30] He Y, Zhang X, Sun J. Channel pruning for accelerating very deep neural networks. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 1389−1397.

[31] Wen W, Wu C, Wang Y, Chen Y, Li H. Learning structured sparsity in deep neural networks. In: Proc. of the Advances in Neural Information Processing Systems. 2016. 2074−2082.

[32] Yu R, Li A, Chen CF, Lai JH, Morariu VI, Han X, Davis LS. Nisp: Pruning networks using neuron importance score propagation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 9194−9203.

[33] Gupta S, Agrawal A, Gopalakrishnan K, Barayanan P. Deep learning with limited numerical precision. In: Proc. of the Int'l Conf. on Machine Learning. 2015. 1737−1746.

[34] Dettmers T. 8-bit approximations for parallelism in deep learning. In: Proc. of the 4th Int'l Conf. on Learning Representations. 2016.

[35] 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. 2015. 3123−3131.

[36] Hubara I, Courbariaux M, Soudry D, Yaniv R, Bengio Y. Binarized neural networks. In: Proc. of the Advances in Neural Information Processing Systems, Vol.29. 2016. 4107−4115.

[37] Li F, Zhang B, Liu B. Ternary weight networks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops. 2016.

[38] Leng C, Dou Z, Li H, Zhu S, Jin R. Extremely low bit neural network: Squeeze the last bit out with ADMM. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. 2018.

[39] Hu Q, Wang P, Cheng J. From hashing to CNNs: Training binary weight networks via hashing. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. 2018.

[40] Wang P, Hu Q, Zhang Y, Zhang C, Liu Y, Cheng J. Two-Step quantization for low-bit neural networks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 4376−4384.

[41] Tung F, Mori G. CLIP-Q: Deep network compression learning by in-parallel pruning-quantization. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 7873−7882.

[42] Denton EL, Zaremba W, Bruna J, Cun YL, Fergus R. Exploiting linear structure within convolutional networks for efficient evaluation. In: Proc. of the Advances in Neural Information Processing Systems. 2014. 1269−1277.

[43] Zhang X, Zou J, He K, Sun J. Accelerating very deep convolutional networks for classification and detection. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2015,38(10):1943−1955.

[44] Lebedev V, Ganin Y, Rakhuba M, Oseledets I, Lempitsky V. Speeding-Up convolutional neural networks using fine-tuned cp-decomposition. In: Proc. of the 3rd Int'l Conf. on Learning Representations. 2015.

[45] Kim YD, Park E, Yoo S, Choi T, Yang L, Shin D. Compression of deep convolutional neural networks for fast and low power mobile applications. In: Proc. of the 4th Int'l Conf. on Learning Representations. 2016.

[46] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. In: Proc. of the Advances in Neural Information Processing Systems Workshop, Vol.27. 2014.

[47] Romero A, Ballas N, Kahou SE, Chassang A, Gatta C, Bengio Y. Fitnets: Hints for thin deep nets. In: Proc. of the 3rd Int'l Conf. on Learning Representations. 2015.

[48] Zagoruyko S, Komodakis N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: Proc. of the 5th Int'l Conf. on Learning Representations. 2017.

[49] Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y, Wierstra D. Continuous control with deep reinforcement learning. In: Proc. of the 4th Int'l Conf. on Learning Representations. 2016.

[50] Ashok A, Rhinehart N, Beainy F, Kitani KM. N2N learning: Network to network compression via policy gradient reinforcement learning. In: Proc. of the 6th Int'l Conf. on Learning Representations. 2018.

[51] Wong C, Houlsby N, Lu Y, Gesmundo A. Transfer learning with neural AutoML. In: Proc. of the Advances in Neural Information Processing Systems, Vol.31. 2018. 8366−8375.

[52] Lin J, Rao Y, Lu J, Zhou J. Runtime neural pruning. In: Proc. of the Advances in Neural Information Processing Systems. 2017. 2181−2191.

[53] Wang H, Zhang Q, Wang Y, Hu H. Structured probabilistic pruning for convolutional neural network acceleration. In: Proc. of the British Machine Vision Conf. 2018.

[54] Real E, Aggarwal A, Huang Y, Le QV. Regularized evolution for image classifier architecture search. In: Proc. of the AAAI Conf. on Artificial Intelligence. 2019.

[55] Chen LC, Collins M, Zhu Y, Papandreou G, Zoph B, Schroff F, Adam H, Shlens J. Searching for efficient multi-scale architectures for dense image prediction. In: Proc. of the Advances in Neural Information Processing Systems. 2018. 8699−8710.

[56] Chollet F. Xception: Deep learning with depth-wise separable convolutions. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 1251−1258.

[57] Yu F, Koltun V. Multi-Scale context aggregation by dilated convolutions. In: Proc. of the 4th Int'l Conf. on Learning Representations. 2016.

[58] Baker B, Gupta O, Naik N, Raskar R. Designing neural network architectures using reinforcement learning. In: Proc. of the 5th Int'l Conf. on Learning Representations. 2017.

[59] Suganuma M, Shirakawa S, Nagao T. A genetic programming approach to designing convolutional neural network architectures. In: Proc. of the Genetic and Evolutionary Computation Conf. 2017. 497−504.

[60] Cai H, Chen T, Zhang W, Yu Y, Wang J. Efficient architecture search by network transformation. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. 2018.

[61] Mendoza H, Klein A, Feurer M, Springenberg J, Hutter F. Towards automatically-tuned neural networks. In: Proc. of the Workshop on Automatic Machine Learning. 2016. 58−65.

[62] Zoph B, Le QV. Neural architecture search with reinforcement learning. In: Proc. of the 5th Int'l Conf. on Learning Representations. 2017.

[63] Szegedy C, Liu W, Jia Y, Sermanet P, Reed RE, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 1−9.

[64] Liu C, Zoph B, Neumann M, Shlens J, Hua W, Li LJ, Fei L, Yuille AL, Huang J, Murphy K. Progressive neural architecture search. In: Proc. of the European Conf. on Computer Vision. 2018. 19−34.

[65] Pham H, Guan MY, Zoph B, Le QV, Dean J. Efficient neural architecture search via parameter sharing. In: Proc. of the 35th Int'l Conf. on Machine Learning. 2018. 4092−4101.

[66] Elsken T, Metzen JH, Hutter F. Efficient multi-objective neural architecture search via lamarckian evolution. In: Proc. of the 7th Int'l Conf. on Learning Representations. 2019.

[67] Cai H, Yang J, Zhang W, Han S, Yu Y. Path-Level network transformation for efficient architecture search. In: Proc. of the 35th Int'l Conf. on Machine Learning. 2018. 677−686.

[68] Liu H, Simonyan K, Yang Y. DARTS: Differentiable architecture search. In: Proc. of the 7th Int'l Conf. on Learning Representations. 2019.

[69] Zhong Z, Yan J, Wu W, Shao J, Liu CL. Practical block-wise neural network architecture generation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 2423−2432.

[70] Dong JD, Cheng AC, Juan DC, Wei W, Sun M. Dpp-Net: Device-aware progressive search for pareto-optimal neural architectures. In: Proc. of the European Conf. on Computer Vision. 2018. 517−531.

[71] Zhong Z, Yan J, Wu W, Shao J, Liu CL. Practical block-wise neural network architecture generation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 2423−2432.

[72] Chen T, Goodfellow I, Shlens J. Net2net: Accelerating learning via knowledge transfer. In: Proc. of the 4th Int'l Conf. on Learning Representations. 2016.

[73] Goldberg DE, Deb K. A comparative analysis of selection schemes used in genetic algorithms. In: Proc. of the Foundations of Genetic Algorithms. Elsevier. 1991. 69−93.

[74] Cubuk ED, Zoph B, Schoenholz SS, Le QV. Intriguing properties of adversarial examples. In: Proc. of the 6th Int'l Conf. on Learning Representations Workshop. 2018.

[75] Liu H, Simonyan K, Vinyals O, Fernando C, Kavukcuoglu K. Hierarchical representations for efficient architecture search. In: Proc. of the 6th Int'l Conf. on Learning Representations. 2018.

[76] Elsken T, Metzen JH, Hutter F. Simple and efficient architecture search for convolutional neural networks. In: Proc. of the 6th Int'l Conf. on Learning Representations Workshop. 2018.

[77] Wistuba M. Deep learning architecture search by neuro-cell-based evolution with function-preserving mutations. In: Proc. of the Joint European Conf. on Machine Learning and Knowledge Discovery in Databases. 2018. 243−258.

[78] Real E, Moore S, Selle A, Saxena S, Suematsu YL, Tan J, Le QV, Kurakin A. Large-scale evolution of image classifiers. In: Proc. of the 34th Int'l Conf. on Machine Learning, Vol.70. 2017. 2902−2911.

[79] Xie L, Yuille A. Genetic CNN. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 1379−1388.

[80] Kandasamy K, Neiswanger W, Schneider J, Póczos B, Xing EP. Neural architecture search with bayesian optimisation and optimal transport. In: Proc. of the Advances in Neural Information Processing Systems. 2018. 2016−2025.

[81] Luo R, Tian F, Qin T, Chen E, Liu TY. Neural architecture optimization. In: Proc. of the Advances in Neural Information Processing Systems. 2018. 7816−7827.

[82] Bender G, Kindermans PJ, Zoph B, Vasudenvan V, Le QV. Understanding and simplifying one-shot architecture search. In: Proc. of the Int'l Conf. on Machine Learning. 2018. 549−558.

[83] Brock A, Lim T, Ritchie JM, Weston N. SMASH: One-shot model architecture search through HyperNetworks. In: Proc. of the 6th Int'l Conf. on Learning Representations. 2018.

[84] Zhang C, Ren M, Urtasun R. Graph HyperNetworks for neural architecture search. In: Proc. of the 7th Int'l Conf. on Learning Representations. 2019.

[85] Real E, Aggarwal A, Huang Y, Le QV. Regularized evolution for image classifier architecture search. In: Proc. of the AAAI Conf. on Artificial Intelligence, Vol.33. 2019. 4780−4789.

[86] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein MS, Berg AC, Li FF. Imagenet large scale visual recognition challenge. Int'l Journal of Computer Vision, 2015,115(3):211−252.

[87] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft COCO: Common objects in context. In: Proc. of the European Conf. on Computer Vision. 2014. 740−755.

[88] Zela A, Klein A, Falkner S, Frank H. Towards automated deep learning: Efficient joint neural architecture and hyperparameter search. In: Proc. of the 21th Int'l Conf. on Artificial Intelligence and Statistics Workshop. 2018.

[89] Klein A, Falkner S, Bartels S, Hennig P, Hutter F. Fast Bayesian optimization of machine learning hyperparameters on large datasets. In: Proc. of the 20th Int'l Conf. on Artificial Intelligence and Statistics. 2017. 528−536.

[90] Domhan T, Springenberg JT, Hutter F. Speeding up automatic hyperparameter optimization of deep neural networks by extrapolation of learning curves. In: Proc. of the 24th Int'l Joint Conf. on Artificial Intelligence. 2015.

[91] Snoek J, Rippel O, Swersky K, Kiros R, Satish N, Sundaram N, Patwary M, Prabhat, Adams R. Scalable Bayesian optimization using deep neural networks. In: Proc. of the 32nd Int'l Conf. on Machine Learning. 2015. 2171–2180

[92] Klein A, Falkner S, Springenberg JT, Hutter F. Learning curve prediction with Bayesian neural networks. In: Proc. of the 5th Int'l Conf. on Learning Representations. 2017.

[93] Baker B, Gupta O, Raskar R, Naik N. Accelerating neural architecture search using performance prediction. In: Proc. of the 6th Int'l Conf. on Learning Representations. 2018.

[94] Wei T, Wang C, Rui Y, Chen C. Network morphism. In: Proc. of the 33nd Int'l Conf. on Machine Learning. 2016. 564–572.

[95] Runge F, Stoll D, Falkner S, Hutter F. Learning to design RNA. In: Proc. of the 7th Int'l Conf. on Learning Representations. 2019.

[96] Li L, Jamieson K, De Salvo G, Rostamizadeh A, Talwalkar A. Hyperband: A novel bandit-based approach to hyperparameter optimization. Journal of Machine Learning Research, 2017,18(185):1–52.

[97] Xie S, Zheng H, Liu C, Lin L. SNAS: Stochastic neural architecture search. In: Proc. of the 7th Int'l Conf. on Learning Representations. 2019.

[98] Weng Y, Zhou T, Li Y, Qiu X. NAS-Unet: Neural architecture search for medical image segmentation. IEEE Access, 2019,7: 44247–44257.

[99] Liu C, Chen LC, Schroff F, Adam H, Hua W, Yuille A, Li FF. Auto-Deeplab: Hierarchical neural architecture search for semantic image segmentation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2019. 82–92.

葛道辉(1994－),男,博士生,CCF 学生会员,主要研究领域为深度学习,计算机视觉,目标跟踪.

李洪升(1994－),男,博士生,主要研究领域为深度学习,视频分类,动作识别.

张亮(1981－),男,博士,副教授,博士生导师,主要研究领域为嵌入式多核系统,机器人语义 SLAM,三维场景语义分割,手势识别.

刘如意(1989－),女,博士,讲师,CCF 专业会员,主要研究领域为计算机视觉,机器学习,目标分割提取.

沈沛意(1969－),男,博士,教授,CCF 专业会员,主要研究领域为计算机视觉,DSPFPGA 理论及应用,数字图像处理,计算机网络.

苗启广(1972－),男,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为计算机视觉,机器学习,大数据分析.