轻量化网络发展现状及其在遥感图像处理领域研究成果综述

徐浩*

2020 年 10 月

目录

*电子邮件: xu_hao_98@163.com

1 引言

本次报告从研究目的，轻量化网络的国内外研究现状以及在遥感图像领域的应用现状，同时详细介绍当前采取的轻量化网络的主要策略。

1.1 研究目的

卷积神经网络在图像的分类、分割以及检测等领域都获得了广泛的应用，并且有着十分优异的检测性能，但随之而来的就是越来越深，越来越复杂的网络，这些网络模型虽然可以获得很好的检测效果，但过于复杂的模型以及过于多的参数和计算量，使得模型的部署成为了一个很大的问题，就是说一个好的深度学习模型无法部署在性能一般的硬件上，从 AlexNet [1] 提出到现在已经八年了，中间出现了 VGGNet [2],ResNet [3],DenseNet [4] 等性能优异的网络结构，总的发展趋势就是网络的层数越来越深，一般情况下，网络越深，特征提取的能力越强，但是随之而来的是，中间会造成大量的特征冗余以及参数冗余，这就大大限制了网络的工业应用范围，因此，对神经网络进行轻量化是当前需要解决的问题，故对轻量化网络的发展进行一个简单调研。

1.2 轻量化网络国内外研究现状

通过对普通卷积神经网络的调查研究发现，这些算法总是把提高网络的检测准确率放在第一位，但是在相同准确率的情况下，一个轻量化网络，即参数更少，运算速度更快的网络可以更容易的进行模型的迁移和部署，比如从服务器部署到自动驾驶系统，并且可以将模型部署到算力相对较弱的平台中，比如 FGPA。因此，在 2016 年，Berkeley 和 Stanford 的团队提出了 SqueezeNet [5], 该模型提出了一个全新的卷积结构，在该模型下获得和 AlexNet 相同精度下减少了 50 倍参数量。2017 年，Google 的团队提出了 MobileNets [6], 在该论文中，应用深度可分离卷积作为主要的卷积方式代替传统的全卷积，通过这种方式，可以大大减少网络的参数量，提高网络的运算速度，采取和 VGG16 [2] 一样的网络结构，将其中的部分卷积换为深度可分离卷积，可以在减少 30 倍参数量下，在 ImageNet 数据上仅损失 0.9% 的精度。2017 年，Facebook 团队提出 ShuffleNet [7]，该模型利用分组卷积以及通道打乱的方式减少模型的运算开销同时维持检测精度，分组卷积的方式会造成通道间信息的不流通，因此通过通道打乱这种方式来解决通道不流通的带来的问题，在参数量更少的情况下，与 MobileNets [6] 相比，在 ImageNet 数据集上提高了 3.1% 的精度。2018 年，Sandler 等人提出了 MobileNetV2 [8], 在这篇论文中指出该模型是基于倒残差设计的网络结构，同时还指出使用 RELU 非线性激活函数可能会丢失部分信息，因此，在 MobileV2 倒残差结构中 1x1 卷积过后移除了 RELU 非线性激活函数，在 ImageNet 数据集上实验表明，与 MobileNets [6] 相比参数量减少 1/4, 同时检测准确率提高 1.4%, 且在同一块 CPU 上前向推理速度由 113ms 提升至 75ms，与

ShuffleNet [7] 相比，在参数量一样的情况下，准确率提升 0.5%。2018 年，Ma [9] 等人提出了 ShuffleNet V2, 在论文中指出，参数量往往不能代表一个模型运算的速度，而受到内存使用量 (memory access cost,MAC) 的限制，同时模型的并行程度也会影响处理速度，理论上来说，并行度高的模型速度相对较快，而且采用不同的平台运算速度也是有差异的，比如：GPU 和 ARM。2019 年，Google 团队提出 MobileNetV3 [10], 在论文中提出了 MobileNetV3-Large 和 MobileNetV3-Small 两种结构，其中 MobileNetV3-Large 在 ImageNet 上的检测结果与 MobileNetV2 相比提高了 3.2% 的准确度。同时，在 MobileNetV2 结构的基础上，加入了注意力模块 SE 去改善网络性能，并且提出利用 h-swish 作为网络中一部分层的激活函数以代替 RELU。

以上，均是通过将全卷积网络替换为深度可分离卷积网络，并在此基础上进行改进和发展，另外，进行模型轻量化的设计方法，还可以通过对模型进行压缩，去除特征冗余的层进行，由于时间仓促，暂时还未对这个方法进行研究，因为当前模型轻量化采取的主流方法是上述所示的一些方法，当然，在利用深度可分离卷积进行网络设计和改进时，可以考虑特征冗余的情况，并进行改进。

1.3 轻量化网络在遥感图像处理领域的应用现状

2 当前采取的轻量化网络的策略

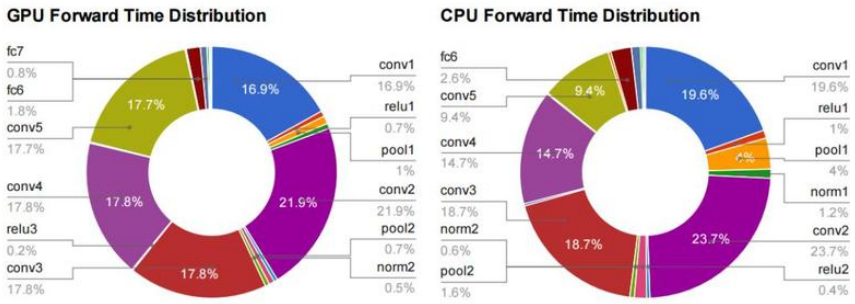

图 1: GPU-CPU 运算时间分布

如图 1 所示，是 AlexNet 网络中不同的层在 GPU 和 CPU 中的运算时间消耗，可以明显看到，无论是在 GPU 还是 CPU 上最耗时的层是卷积层，因此，若要明显改善网络的运算性能，就要提高卷积层的计算效率。

2.1 模型压缩策略

2.2 改变卷积方式策略

3 参考文献

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[2] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[4] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[5] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size, 2016.

[6] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017.

[7] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[8] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[9] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[10] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.