Recent Advances in Convolution Neural Networks

Tianyuan Zhang
Department of Computer Science
Peking University
1600012888@pku.edu.cn

Abstract

Since 2012, deep convolution networks have outperformed the state of the art in many visual recognition tasks, and many lager and deeper networks has been desinged. In this paper, we give a brief review over some of the most popular models and insights behind them.

1 Tensor Decomposition

Traditional convolution layer have to learn a filter in 3D space, with two spatial dimension and one channel dimension; thus one single convolution operation is tasked with mapping cross-channel correlations and spatial correlations simultaneously.

The idea behind many works is to make this process mush simpler and more computational efficient by decompose convolution into a series of operations.

1.1 Inception Module

In figure 1, we show one canonical form of an Inception module found in the Inception V4 architecture[1]. Typical Inception modules first deals with cross-channel correlations via 1x1 convolutions(or point-wise convolution), mapping the input data into several smaller separate 3D spaces, then perform convolution with larger kernels in each branch(pooling can been seen as a form of convolution) to pay more attention to spatial correlations. Alternating 7x1 and 1x7 convolutions is to independently look at width-wise correlations and height-wise correlations.

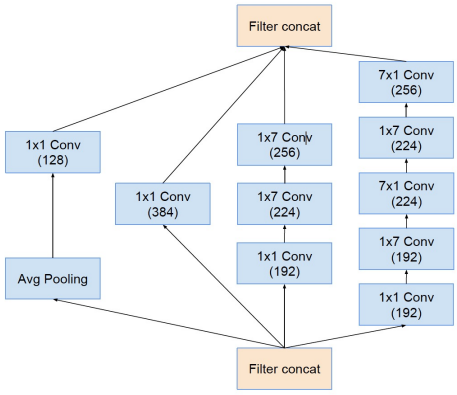

Figure 1: A canonical form of an Inception module(Inception V4).

Course Project.

1.2 Depthwise Separable Convolution

We can see, the hypothesis behind Inception modules is that to some extent cross-channel correlations and spatial correlations can decoupled. Xception[2], which stands for 'Extreme Inception', made a stronger hypothesis: these two correlations can be entirely decoupled.

Xception uesed *depthwise separable convolution* which has been used as early as 2014[3] to factorize traditional convolutions. And briefly, Xception architecture is a linear stack or depthwise separable convlotion with residual connections.

Due to it's great gain in computational efficiency without loosing too much representational power, depthwise separable convolution has been widely used in lighweight convolution networks[4],[5],[6]

1.3 Bottleneck and Inverted Bottleneck

Inspired by the intuition that the manifold of interest lies in a low-dimensional subspace of the higher-dimensional activation space, bottleneck design was proposed and successfully used in[7],[8] to keep the computational budget, it's main idea is to perform a dimensionality reduction via 1x1 convolution before convolutions with larger kernel size. This design was very popular in modern architectures.

Sandler M, et al.[5] reported gain of expressiveness when increasing the dimensionality before applying the 3x3 depthwise separable convolution. Surprsingly under the same computational budget, with much thinner feature maps, mobilenet-v2 achieves comparable results with Shufflenet[6] and NasNet-A[21].This may give rise to further exploration how bottleneck structures equipped with different activation functions and different expansion ratio work.

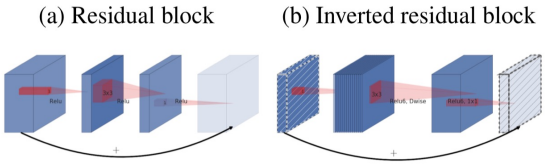

Figure 2: Demonstration of the difference between Bottleneck and Inverted Bottleneck structure

1.4 Group Convolution

The idea of group convolution was first introduced in *AlexNet*[9] to distributed the model over two GPUs, depthwise separable convolution used in Xception generalized this idea further.

Roughly speaking, group convolution on a feature map of g evenly distributed groups reduces the computational complexity by a factor of g, but suffers from poor feature fusion.

Combining the bottleneck architecture, ResNeXt[10] adds 1x1 convolution both before and after 3x3 group convolution, thanks to group convolution ResNeXt can be equipped with wider feature maps compared to its ResNet[8] counterpart under the same computational budget, thus has a much improved representational capability.This architecture is shown in figure 3.

Later, Zhang X et al.[6] noticed for each redisual unit in ResNeXt, the pointwise convolutions occupy 93.4% multiplication adds(cardinality as 32) . So they proposed using 1x1 group convolution and a novel channel shuffle operation to replace expensive pointwise convolutions, and achieved state-of-the-art in computational-efficient *CNN* architecture. Several shufflenet[6] units are shown in figure 4.

2 Spatial Operator

A main concern in visual recognition is how to accommodate multiple geometric transformations. This was usually addressed in two approaches. One is to build massive datasets with sufficient variations. The other is to design transformation-invariant feature extractors or algorithms.

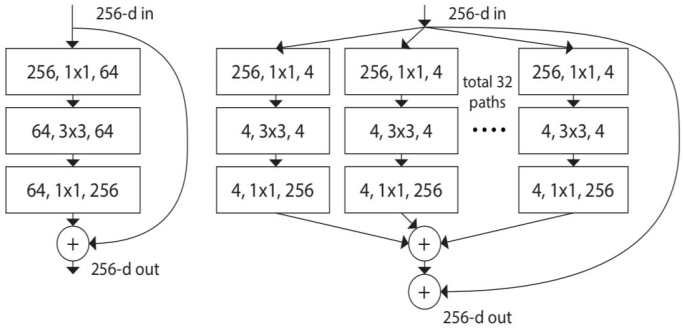

Figure 3: Illustration of the design of ResNeXt. **Left:** bottleneck architecture used in ResNet[8]. **Right:** bottleneck architecture combined with group convolution in ResNeXt with roughly the same complexity

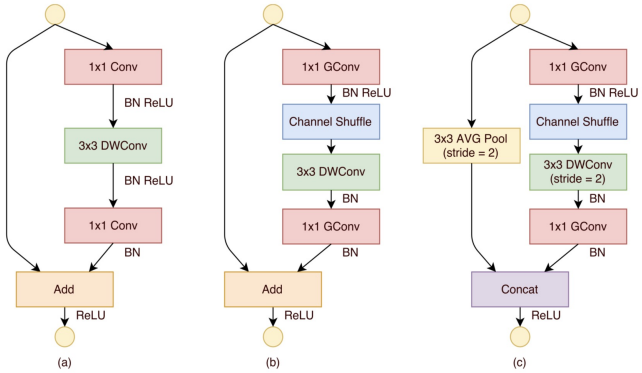

Figure 4: Demonstration of different types of shufflenet's unit.

Though,convolution neural networks have achieved great success in past few years, the fixed geometric structure of *CNN* modules limit it to model large transformations in object scale, viewpoint, pose, etc. So lots of work has been done in designing new spatial operators or modules to address this problem.

2.1 STN and Deformable Convolution

Jaderberg M et al.[11] introduced *Spatial Transformer* module, a dynamic mechanism that learns to actively perform spatial transformations on single image(or a feature map), which can be trained end-to-end. The transformation it learns can include scaling, cropping, rotations and non-grid deformations.

Deformable convolution[3] made this process simpler and more efficient by adaptively adding 2D offsets to the regular grid sampling locations in traditional convolution. As shown in figure 5, the offsets are learned via additional convolutions conditioned on the preceding feature maps, combined with bilinear interpolation this module can be trained in a end-to-end manner.

2.2 Receptive Filed and Dilated Convolution

We can see that successive stride 1 convolution have only linear increasing receptive field size, while most popular image classification networks achieves multi-scale contextual reasoning via successive pooling or downsampling layers which doubles the receptive filed and reduces the resolution. But dense prediction tasks such as semantic segmentation calls for large receptive filed in combination with high-resolution output.

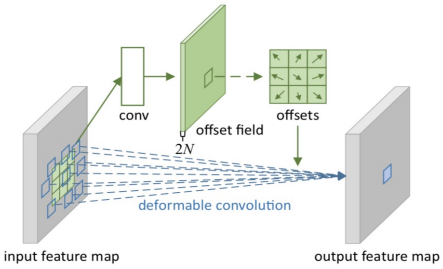

Figure 5: Illustration of 3x3 deformable convolution[11].

This dilemma was usually addressed in Three popular ways. One approach is to use U-Net[13] like Encoder-Decoder structure with skip connections. Another approach involves multiple branches to deal with features in different scales([14]).

Yu F, et al.[15] proposed a module using successive dilated convolutions to support exponential expansion of the receptive file without loss of resolution.This is demonstrated in figure 6

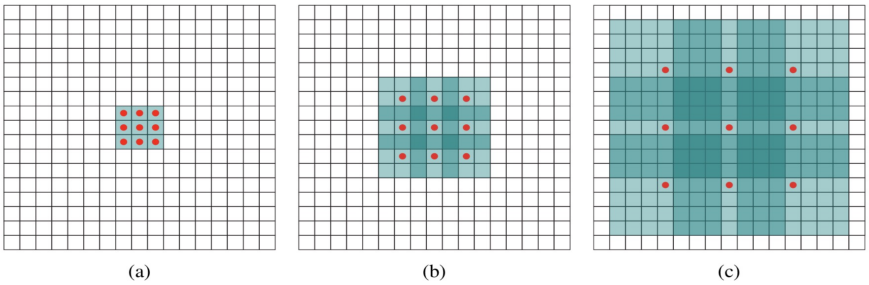

Figure 6: Systematic successive dilation convolutions. (a) 1-dilated convolution. (b) 2-dilated convolution following behind. (c) 4-dilated convolutions behind.

2.3 Deconvolution and Checkerboard Artifacts

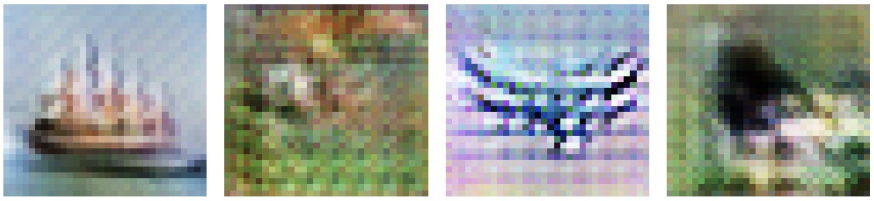

Figure 7: Deconvolution suffers from checkerboard artifacts

Deconvolution(or transpose convolution) suffers from checkerboard artifacts as shown in figure 7, and many approaches have been proposed to alleviate this problem.

Odena et al.[16] highlight three sources of checkerboard artifacts: deconvolution overlap, random initialization and loss functions, and proposed separating the feature aggregation stage and upsampling to higher-resolution process by first resizing the feature map(using nearest-neighbor interpolation or bilinear interpolation) and then performing a regular convolution.

Shi W, et al.[17] proposed a sub-pixel convolution layer to alleviate this problem, where a single reshape operation was utilized to upscale the feature maps.This is showed in figure 8

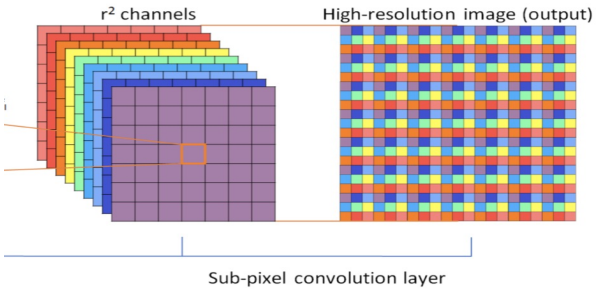

Figure 8: Illustration of sub-pixel convolution layer

3 Skip Connection and Feature Reuse

He K, et al.[8] first successfully trained networks with more than 100 layers by utilizing identity mappings to ease the optimization. After that, most modern architecture equip theirselves with skip connection.

DenseNet[18] connects each layer to every other layers via concatenation to encourages feature reuse and alleviate the vanishing-gradient problem, exploiting this idea to extreme.

Skip connection was also used in encoder-decoder like structures to strengthen feature propagation and gain more precise results in dense prediction tasks.([13])

But how ResNet works still remains mysterious. Two school of thoughts have tried to explain it. One is the ensemble view — ResNets attempts to learn an exponential ensemble of shallow networks. Another is the unrolled iterative estimation view — ResNets layers are thought to iteratively refine representaions.

4 Attention Mechanism

Attention mechanism, a heuristically inspired tool, has been used across a range of tasks. In *CNN*s, this mechanism was usually utilized to help modelling either interdependencies between channels or spatial correlations.

4.1 Channel Relationships

Hu J, et al.[19] introduced *Sequeeze-and-Excitation Blocks*, a computational efficient and flexible module to adaptively recalibrates channel-wise features, this architecture generalize extremely well in many challenging tasks. They also introduced SE-ResNet module by integrating SE blocks and ResNet, this is depicted in figure 9.

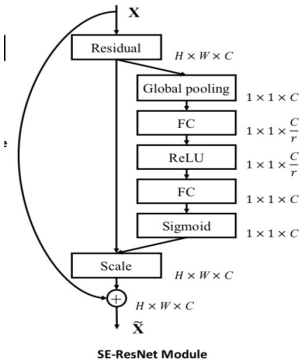

Figure 9: The schema of SE-ResNet module.

4.2 Spatial Attention

*CNN*s inherently produce smooth feature maps which are harmful for per-pixel tasks where rapid change in responses are required.

Harley A W, et al.[20] proposed a novel structure, which can generates accurate foreground-background spatial masks for convolution to sharpen the feature map.

5 Acknowledgments

Thanks Xiangyu Zhang for insightful instructing.

6 Conclusions

Though in this paper, we focused mainly on novel or obviously different architectures, but there is some subtle difference that matters: where or weather to use BN[22] or activation functions.This was noted in many papers especially in[2],[5],[8],[19].

References

[1] Szegedy, Christian, et al. "Inception-v4, inception-resnet and the impact of residual connections on learning." AAAI. Vol. 4. 2017.

[2] Carreira J, Madeira H, Silva J G. Xception: A technique for the experimental evaluation of dependability in modern computers[J]. IEEE Transactions on Software Engineering, 1998, 24(2): 125-136.

[3] Sifre L, Mallat S. Rigid-motion scattering for texture classification[J]. arXiv preprint arXiv:1403.1687, 2014.

[4] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.

[5] Sandler M, Howard A, Zhu M, et al. Inverted Residuals and Linear Bottlenecks: Mobile Networks for Classification, Detection and Segmentation[J]. arXiv preprint arXiv:1801.04381, 2018.

[6] Zhang X, Zhou X, Lin M, et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices[J]. arXiv preprint arXiv:1707.01083, 2017.

[7] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]. Cvpr, 2015.

[8] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[9] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.

[10] Xie S, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks[C]//Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. IEEE, 2017: 5987-5995.

[11] Jaderberg M, Simonyan K, Zisserman A. Spatial transformer networks[C]//Advances in neural information processing systems. 2015: 2017-2025.

[12] Dai J, Qi H, Xiong Y, et al. Deformable convolutional networks[J]. CoRR, abs/1703.06211, 2017, 1(2): 3.

[13] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015: 234-241.

[14] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//CVPR. 2017, 1(2): 4.

[15] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions[J]. arXiv preprint arXiv:1511.07122, 2015.

[16] Odena, Augustus, Vincent Dumoulin, and Chris Olah. "Deconvolution and checkerboard artifacts." Distill 1, no. 10 (2016): e3.

[17] Shi W, Caballero J, Huszár F, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1874-1883.

[18] Huang G, Liu Z, Weinberger K Q, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, 1(2): 3.

[19] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[J]. arXiv preprint arXiv:1709.01507, 2017.

[20] Harley A W, Derpanis K G, Kokkinos I. Segmentation-aware convolutional networks using local attention masks[C]//IEEE International Conference on Computer Vision (ICCV). 2017, 2: 7.

[21] Zoph B, Vasudevan V, Shlens J, et al. Learning transferable architectures for scalable image recognition[J]. arXiv preprint arXiv:1707.07012, 2017.

[22] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[J]. arXiv preprint arXiv:1502.03167, 2015.