PVANET: Deep but Lightweight Neural Networks for Real-time Object Detection

Kye-Hyeon Kim; Sanghoon Hong; Byungseok Roh

Intel Imaging and Camera Technology

{*kye-hyeon.kim, sanghoon.hong, peter.roh*}*@intel.com*

Report By: *Cao Wenlong*

2017 年 3 月 7 日

Overview

1. Introduction
   - Abstract
   - Introduction
   - Review Related Work
2. Nerual Network Design
   - C.ReLU
   - Inception
3. Experimental Results
4. Summary

Abstract

- 使用"Feature Extraction+Region Proposal+RoI Classification”的结构，主要对Feature Extraction进行重新设计。因为，Region Proposal部分计算量不太大而且classification部分可以使用通用的技术(例如：Truncated SVD)进行有效的压缩。
- 设计原则：Less channels with more layers 和采用一些Building blocks （包括：串级的ReLU、Inception和HyperNet)

结果

VOC2007－83.8%mAP；VOC2012－82.5%mAP，46ms/image在NVIDIA Titan X GPU；计算量是ResNet-101的12.3% (理论上)

Introduction

准确率很高的检测算法有往往需要很大的计算量。现在压缩和量化技术的发展对减小网络的计算量很重要。这篇文章展示了我们用于目标检测的一个轻量级的特征提取的网络结构－*PVANET*

- 串级的ReLU[1](C.ReLU－Concatenated rectified linear unit)被用在我们的CNNs 的初期阶段来减少一半的计算数量而不损失精度。
- Inception[3]被用在剩下的生成feature的子网络中。一个Inception module产生不同大小的感受野（receptive fields）的输出激活值，所以增加前一层感受野大小的变化。我们观察叠加的Inception modules可以比线性链式的CNNs更有效的捕捉大范围的大小变化的目标。
- 采用multi-scale representation的思想[2], 结合多个中间的输出，所以，这使得可以同时考虑多个level的细节和非线性。我们展示设计的网络deep and thin，在batch normalization、residual connections和基于plateau detection的learning rate的调整的帮助下进行有效地训练

Review:HyperNet

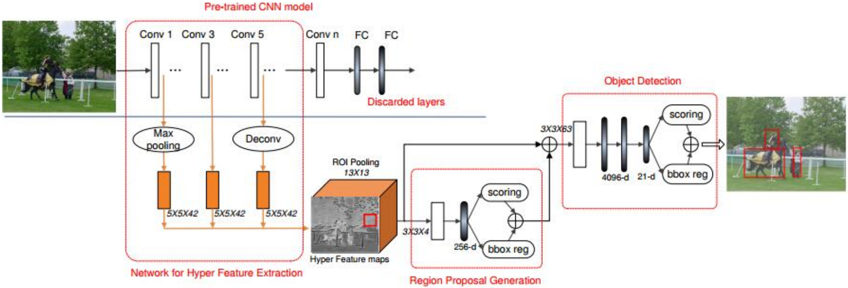


Figure 1: HyperNet的网络结构示意图

Review:Inception

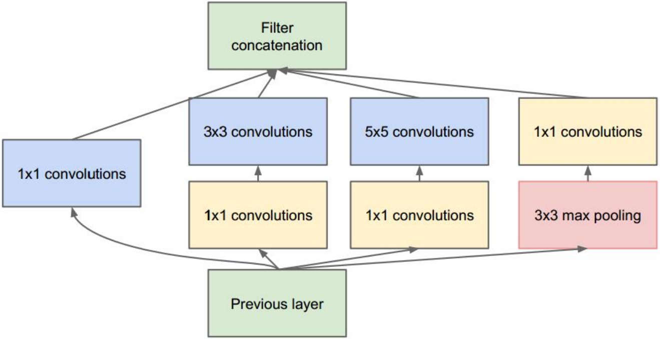


Figure 2: Inception的网络结构示意图

C.ReLU来源于CNN中间激活模式引发的。观察发现，输出节点倾向于是"配对的"，一个节点激活是另一个节点的相反面。

求同

C.ReLU减少一半输出通数量，通过简单的连接相同的输出和negation 使其变成双倍，这使得2倍的速度提升而没有损失精度

存异

同时，增加了scaling and shifting在concatenation之后，这允许每个channel 的斜率和激活阈值与其相反的channel不同。

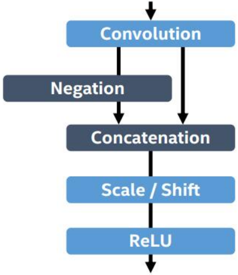


Figure 3: C.ReLU的设计结构

Inception是捕获图像中小目标和大目标的最具有成效的Building Blocks之一;
为了学习捕获大目标的视觉模式，CNN特征应该对应于足够大的感受野，这可以很容易的通过叠加3x3或者更大的核卷积实现;
为了捕获小尺寸的物体，输出特征应该对应于足够小的感受野来精确定位小的感兴趣区域。

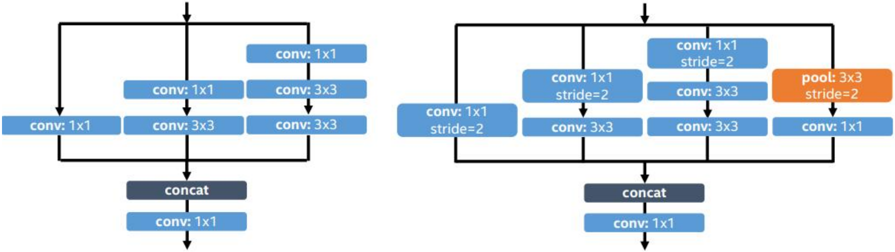


Figure 4: (Left) Our Inception building block. 5x5 convolution is replaced with two 3x3 convolutional layers for efficiency. (Right) Inception for reducing feature map size by half

1x1的conv扮演了关键的角色，保留上一层的感受野。只是增加输入模式的非线性，它减慢了一些输出特征的感受野的增长，使得可以精确地捕获小尺寸的目标。

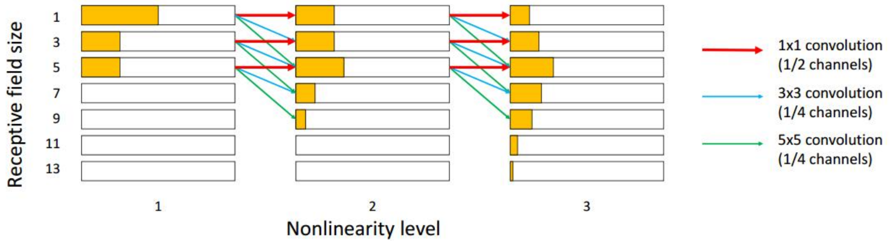
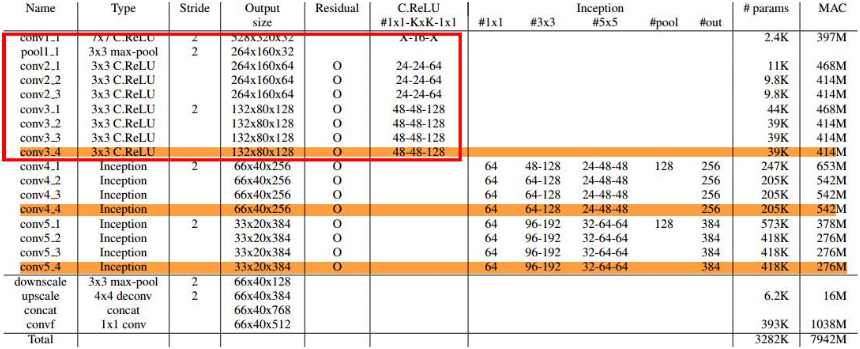

Receptive field size
1
3
5
7
9
11
13
1
2
3
Nonlinearity level
1x1 convolution (1/2 channels)
3x3 convolution (1/4 channels)
5x5 convolution (1/4 channels)

Figure 2: Example of a distribution of (expected) receptive field sizes of intermediate outputs in a chain of 3 Inception modules. Each module concatenates 3 convolutional layers of different kernel sizes, 1x1, 3x3 and 5x5, respectively. The number of output channels in each module is set to $\{1/2, 1/4, 1/4\}$ of the number of channels from the previous module, respectively. A latter Inception module can learn visual patterns of wider range of sizes, as well as having higher level of nonlinearity.

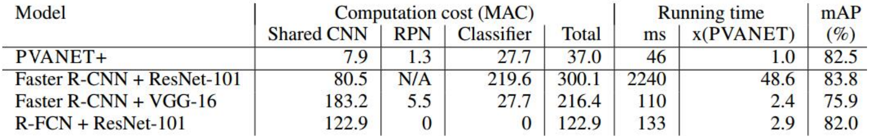
Figure 5: Inception中的感受野的直观表示



Name	Type	Stride	Output	Residual	C.ReLU	Inception					# params	MAC
			size		#1x1-KxK-1x1	#1x1	#3x3	#5x5	#pool	#out		
conv1_1	7x7 C.ReLU	2	528x320x32		X-16-X						2.4K	397M
pool1_1	3x3 max-pool	2	264x160x32									
conv2_1	3x3 C.ReLU		264x160x64	O	24-24-64						11K	468M
conv2_2	3x3 C.ReLU		264x160x64	O	24-24-64						9.8K	414M
conv2_3	3x3 C.ReLU		264x160x64	O	24-24-64						9.8K	414M
conv3_1	3x3 C.ReLU	2	132x80x128	O	48-48-128						44K	468M
conv3_2	3x3 C.ReLU		132x80x128	O	48-48-128						39K	414M
conv3_3	3x3 C.ReLU		132x80x128	O	48-48-128						39K	414M
conv3_4	3x3 C.ReLU		132x80x128	O	48-48-128						39K	414M
conv4_1	Inception	2	66x40x256	O		64	48-128	24-48-48	128	256	247K	653M
conv4_2	Inception		66x40x256	O		64	64-128	24-48-48		256	205K	542M
conv4_3	Inception		66x40x256	O		64	64-128	24-48-48		256	205K	542M
conv4_4	Inception		66x40x256	O		64	64-128	24-48-48		256	205K	542M
conv5_1	Inception	2	33x20x384	O		64	96-192	32-64-64	128	384	573K	378M
conv5_2	Inception		33x20x384	O		64	96-192	32-64-64		384	418K	276M
conv5_3	Inception		33x20x384	O		64	96-192	32-64-64		384	418K	276M
conv5_4	Inception		33x20x384	O		64	96-192	32-64-64		384	418K	276M
downscale	3x3 max-pool	2	66x40x128									
upscale	4x4 deconv	2	66x40x384								6.2K	16M
concat	concat		66x40x768									
convf	1x1 conv		66x40x512								393K	1038M
Total											3282K	7942M

Figure 6: The detailed structure of PVANET

Experimental Results



Model	Computation cost (MAC)				Running time		mAP
	Shared CNN	RPN	Classifier	Total	ms	x(PVANET)	(%)
PVANET+	7.9	1.3	27.7	37.0	46	1.0	82.5
Faster R-CNN + ResNet-101	80.5	N/A	219.6	300.1	2240	48.6	83.8
Faster R-CNN + VGG-16	183.2	5.5	27.7	216.4	110	2.4	75.9
R-FCN + ResNet-101	122.9	0	0	122.9	133	2.9	82.0

Figure 7: Comparisons between our network and some state-of-the-arts in the PASCAL VOC2012 leaderboard.

Summary

- C.ReLU减少训练过程中的网络大小
- Inception是网络设计中的用于压缩网络的技巧
- 1x1的使用相当于挖掘卷积结果中的冗余信息，从而减少channel个数

References

Sanghoon Hong, Byungseok Roh, Kye-Hyeon Kim, Yeongjae Cheon, and Minje Park.
PVANet: Lightweight deep neural networks for real-time object detection.
arXiv preprint arXiv:1611.08588, 2016.

Tao Kong, Anbang Yao, Yurong Chen, and Fuchun Sun.
Hypernet: Towards accurate region proposal generation and joint object detection.
CoRR, abs/1604.00600, 2016.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich.
Going deeper with convolutions.
CoRR, abs/1409.4842, 2014.