杨帆-20170926-读书笔记

Mask R-CNN

- 作者：Kaiming He, Georgia Gkioxari, Piotr Doll´ar, Ross Girshick, Facebook AI Research (FAIR).
- 继Faster-RCNN后，将pixel-wise segmentation引入这一框架。
- 创新点：
  - 1、Segmentation和classification、bounding box regression同步进行，之前的框架利用segmentation结果辅助classification。
  - 2、提出ROI Align方法，因为ROI Pooling方法计算bin大小时取整操作会引入偏差，导致pixel wise segmentation精度不够。
    - > Our proposed change (ROI Align) is simple: we avoid any quantization of the RoI boundaries or bins (i.e., we use $x/ = 16$ instead of $[x/ = 16]$). We use bilinear interpolation to compute the exact values of the input features at four regularly sampled locations in each RoI bin, and aggregate the result (using max or average).
    - > RoIAlign has a large impact: it improves mask accuracy by relative 10% to 50%.

- Segmentation输出结果不是$(C+1)$类的One Hot Vector，而是Binary Mask；将classification与segmentation解耦。
  - > We found it essential to decouple mask and class prediction: we predict a binary mask for each class independently, without competition among classes, and rely on the network's RoI classification branch to predict the category.

Spatial Transformer Networks

- 作者：Max Jaderberg, Karen Simonyan, Andrew Zisserman, Koray Kavukcuoglu, Google DeepMind.
- 通过网络自主学习对输入数据（中间层特征）的空间变换，使得网络能够适应平移、缩放、旋转等变换。

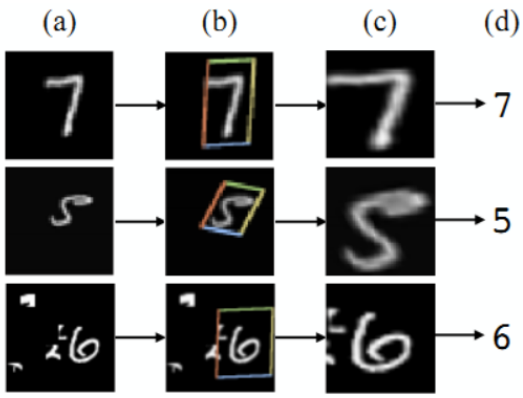


Figure 1: The result of using a spatial transformer as the first layer of a fully-connected network trained for distorted MNIST digit classification. (a) The input to the spatial transformer network is an image of an MNIST digit that is distorted with random translation, scale, rotation, and clutter. (b) The localisation network of the spatial transformer predicts a transformation to apply to the input image. (c) The output of the spatial transformer, after applying the transformation. (d) The classification prediction produced by the subsequent fully-connected network on the output of the spatial transformer. The spatial transformer network (a CNN including a spatial transformer module) is trained end-to-end with only class labels – no knowledge of the groundtruth transformations is given to the system.

- 上图中，没有Ground Truth，STN能够定位到数字(b)，并且进行纠正(c)。

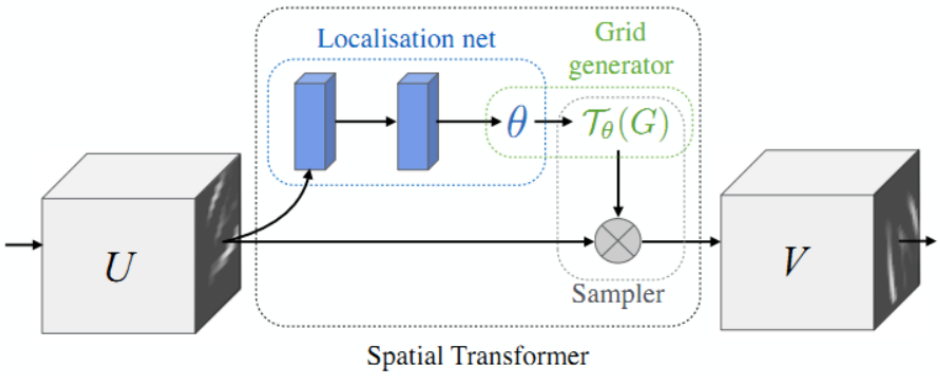


Figure 2: The architecture of a spatial transformer module. The input feature map U is passed to a localisation network which regresses the transformation parameters θ. The regular spatial grid G over V is transformed to the sampling grid $\mathcal{T}_\theta(G)$, which is applied to U as described in Sect. 3.3, producing the warped output feature map V. The combination of the localisation network and sampling mechanism defines a spatial transformer.

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = \mathcal{T}_\theta(G_i) = \mathtt{A}_\theta \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix}$$

- 如上图所示，通过一个Localization net分支，学习到一组变换参数θ，经过如上计算，得到对应的新坐标，通过Sampling（最邻近插值、双线性插值等），得到变换后的图像（特征）。
- 整个过程可以反向传播；对于输入的不同channel，操作相同。

Training RNNs as Fast as CNNs

- 作者：Tao Lei, Yu Zhang.
- RNN难以并行化的主要原因是当前的状态要依靠上一时刻的状态进行计算，本文将状态计算简化，不再依赖于前一时刻的状态，使得RNN计算可以并行化，达到近似CNN的速度，比经过cudnn优化过的lstm快5~10倍。
- Simple Recurrent Unit (SRU) 内部状态c_t的计算仍然依赖于c_{t-1},而h_t则不依赖于h_{t-1}，由c_t和x_t计算得到。

We propose to completely drop the connection (between $\mathbf{h}_{t-1}$ and the neural gates of step t). The associated equations of SRU are given below,

$$\tilde{\mathbf{x}}_t = \mathbf{W}\mathbf{x}_t \tag{3}$$
$$\mathbf{f}_t = \sigma(\mathbf{W}_f\mathbf{x}_t + \mathbf{b}_f) \tag{4}$$
$$\mathbf{r}_t = \sigma(\mathbf{W}_r\mathbf{x}_t + \mathbf{b}_r) \tag{5}$$
$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + (1 - \mathbf{f}_t) \odot \tilde{\mathbf{x}}_t \tag{6}$$
$$\mathbf{h}_t = \mathbf{r}_t \odot g(\mathbf{c}_t) + (1 - \mathbf{r}_t) \odot \mathbf{x}_t \tag{7}$$

Given a sequence of input vectors $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$, $\{\tilde{\mathbf{x}}_t, \mathbf{f}_t, \mathbf{r}_t\}$ for different $t = 1 \cdots n$ are independent and hence all these vectors can be computed in parallel. Our formulation is similar to the recently

- 两个创新点：
  - 不同layer之间添加skip connections，利于加深网络；
  - 使用Variational Dropout，在不同time step下共享dropout mask。

Deformable Convolutional Networks

- 作者：Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, Yichen Wei, Microsoft.
- 一般的卷积操作以矩形框的方式进行特征提取，作者提出可形变的卷积操作（Deformable CNN），使得提取特征时学习到特征点的偏置信息；同时提出了可形变的ROI Pooling（Deformable ROI Pooling）。

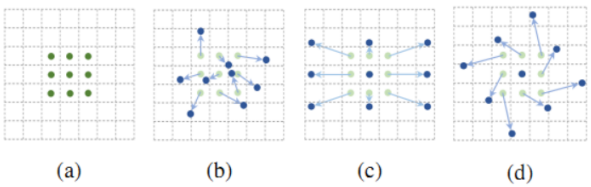


Figure 1: Illustration of the sampling locations in 3×3 standard and deformable convolutions. (a) regular sampling grid (green points) of standard convolution. (b) deformed sampling locations (dark blue points) with augmented offsets (light blue arrows) in deformable convolution. (c)(d) are special cases of (b), showing that the deformable convolution generalizes various transformations for scale, (anisotropic) aspect ratio and rotation.

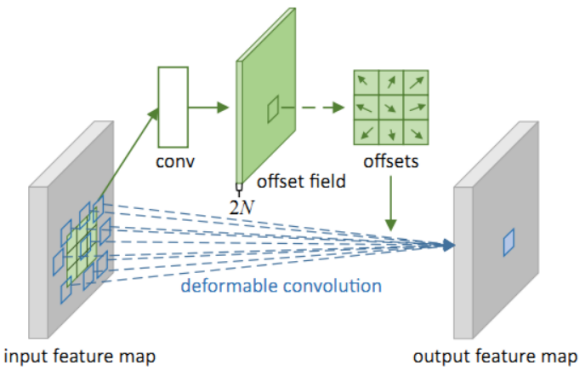


Figure 2: Illustration of 3×3 deformable convolution.

- Deformable CNN：在一般CNN基础上，加一路分支（如上右图）使得feature map上每一个特征值学习到一组$k^2 * (x, y)$偏置量（k^2为卷积核大小），偏置(x, y)使卷积操作时，以该特征值相对偏置处（小数部分使用

双线性插值）的特征值进行卷积。所以，Deformable CNN可以提取缩放、旋转特征（如上左图）。

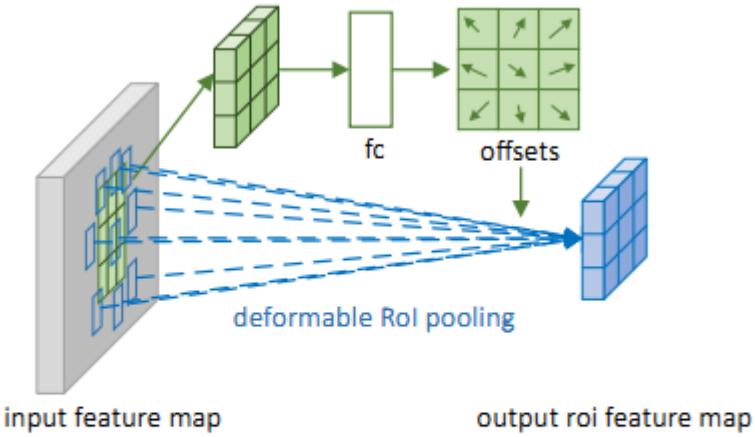


Figure 3: Illustration of 3 × 3 deformable RoI pooling.

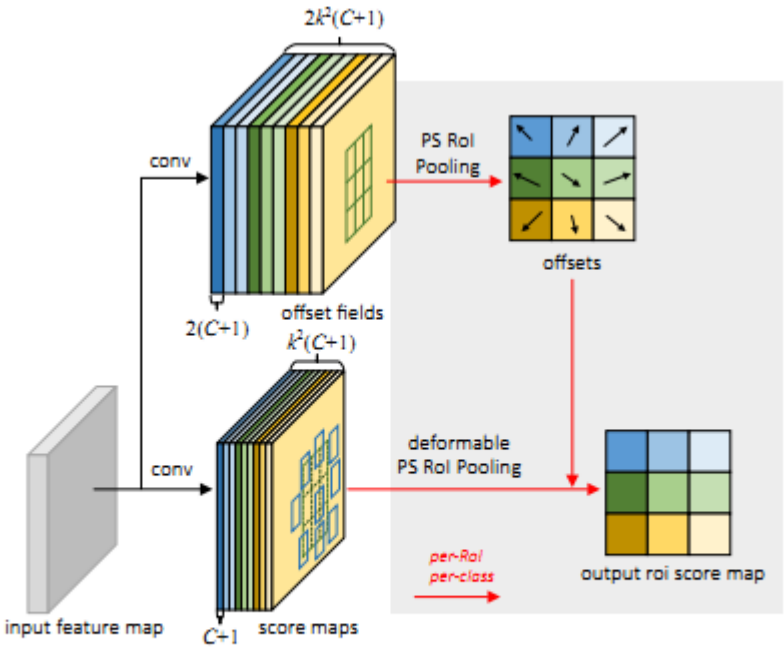


Figure 4: Illustration of 3 × 3 deformable PS RoI pooling.

- Deformable ROI Pooling：类似于D-CNN，采用一个分支学习每个Bin的偏置。

Focal Loss for Dense Object Detection

- 作者：Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, Piotr Dollar. [Facebook AI Research (FAIR)]
- 本文先分析了当前流行的检测框架，主要分为两类：
  - Faster RCNN（two stage）:先提取Proposals，再进行分类和回归；检测效果好，实时性较差；
  - Yolo、SSD（one stage）：直接出结果，不用Proposals；速度快，效果较two stage差。
- 分析如下：

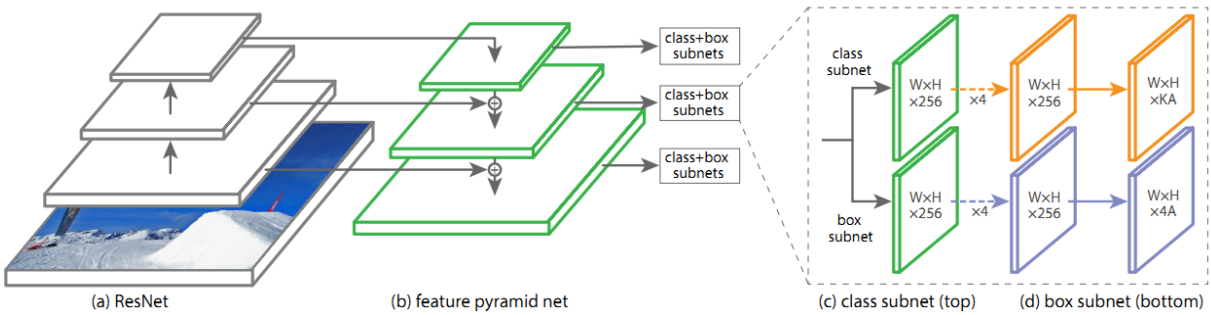
- Two stage方法速度慢主要是因为提取proposals后，每个ROI都需要进行一次ROI pooling+分类+回归，速度较慢；而One stage方法将proposals提取固定在预先设定的网格中，直接一次前向对proposals进行分类、回归，所以速度较快；
- One stage方法虽简单粗暴，但是会造成proposals正负样本比差距较大，因为训练时没有对负样本进行抑制；而Two stage方法训练时，对于提取到的proposals会进行一定比例的筛选抑制，保持一定的正负样本比例，使得训练较为稳定，效果较好；

- 文章提出了新的损失函数：Focal loss，可用于缓解正负样本比例差距较大的问题；提出一个新的网络结构：Retina Net，借助Focal loss，Retina Net能够兼顾One stage方法的速度和Two stage方法的效果。
- Focal loss：

$$FL(p_t) = -(1-p_t)^{\gamma} log(p_t)$$

  - p_t是softmax输出的置信度，当置信度越高，该sample对应的loss就越低，使网络偏向于拟合困难样本；
  - 对于简单样本，p_t较大loss较小；对于困难、易错样本，p_t较小loss较大（指数级的差距），使得网络更加focus困难样本；不过，对于网络中已识别错误但p_t较大的样本，纠错能力有限；
  - γ是可调参数，当$\gamma = 0$，Focal loss就成为交叉熵，随着γ增大，调节效果越明显，文章中取2效果最好。

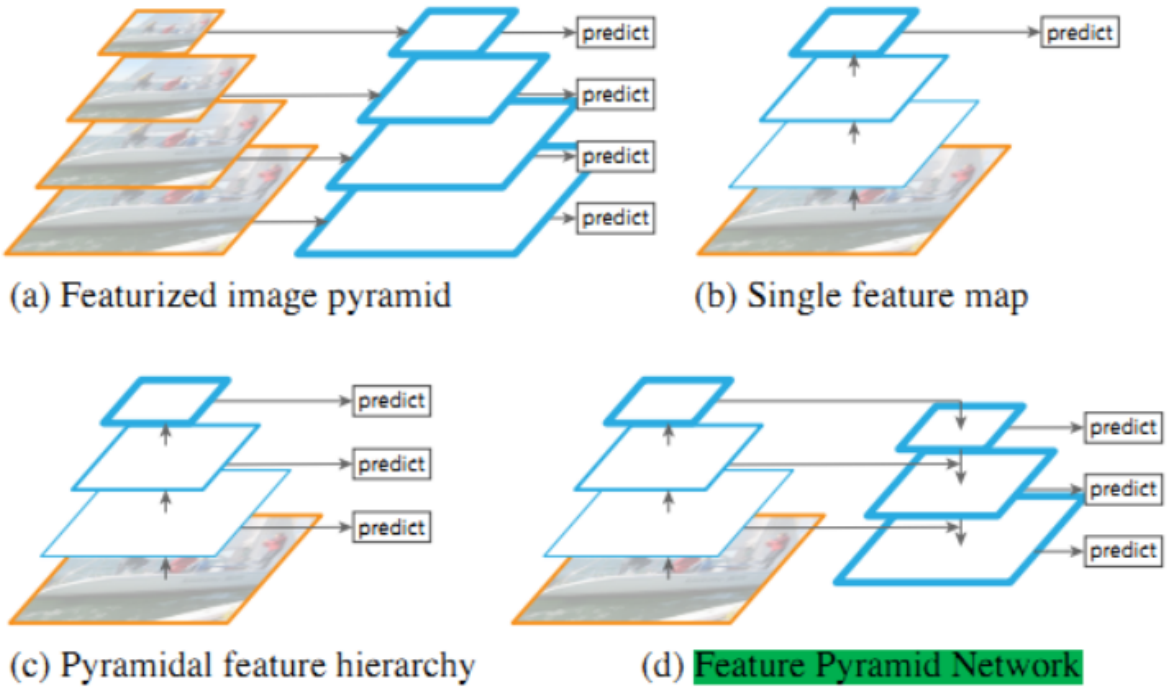
- Retina Net：



  - 使用“ResNet” + “Feature Pyramid Network”【Feature pyramid networks for object detection. In CVPR, 2017.】作为特征提取网络；
  - “FPN”可以提取多尺度特征，对于不同尺度的特征设计不同大小、比例的anchor；
  - 不通过ROI pooling，直接对多尺度特征进行分类+回归，网络同“Faster RCNN”。

Feature Pyramid Networks for Object Detection

- 作者：Tsung-Yi Lin, Piotr Doll, Ross Girshick1, Kaiming He, Bharath Hariharan, Serge Belongie2.



(a) Featurized image pyramid (b) Single feature map

(c) Pyramidal feature hierarchy (d) Feature Pyramid Network

Figure 1. (a) Using an image pyramid to build a feature pyramid. Features are computed on each of the image scales independently, which is slow. (b) Recent detection systems have opted to use only single scale features for faster detection. (c) An alternative is to reuse the pyramidal feature hierarchy computed by a ConvNet as if it were a featurized image pyramid. (d) Our proposed Feature Pyramid Network (FPN) is fast like (b) and (c), but more accurate. In this figure, feature maps are indicate by blue outlines and thicker outlines denote semantically stronger features.

- 如上图(d)所示，不仅利用各个中间层的输出，同时自上而下结合高层、低层特征（结合方式见下图）。

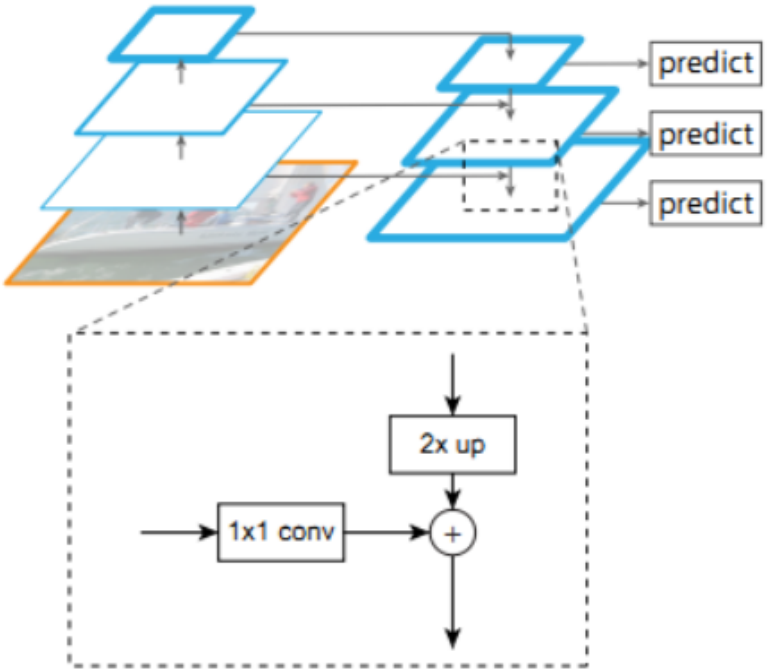


Figure 3. A building block illustrating the lateral connection and the top-down pathway, merged by addition.

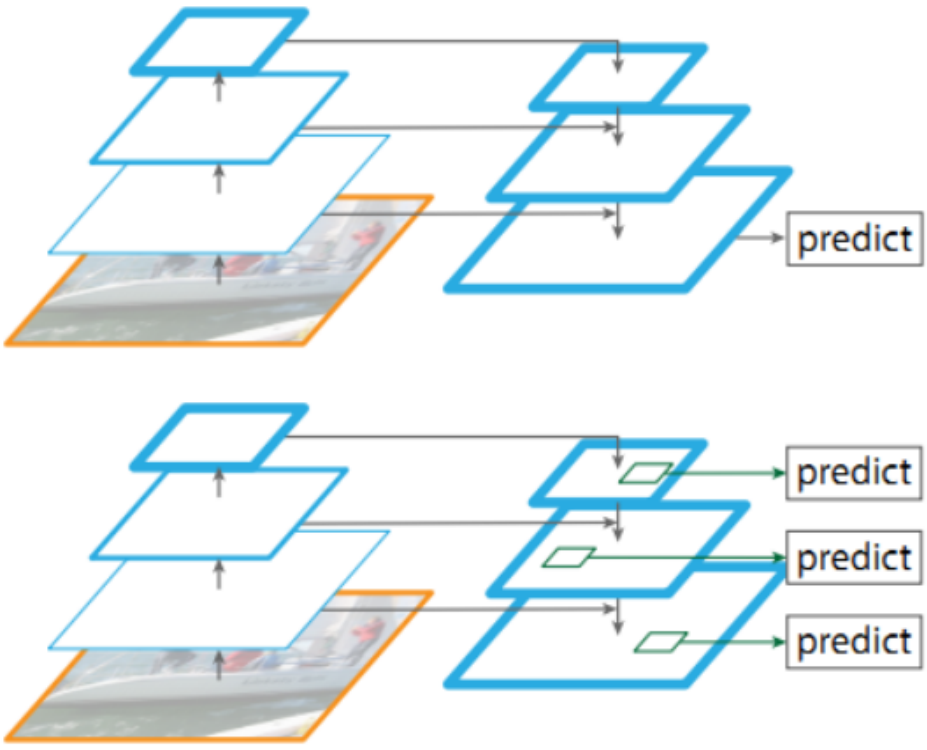


Figure 2. Top: a top-down architecture with skip connections, where predictions are made on the finest level (*e.g.*, [28]). Bottom: our model that has a similar structure but leverages it as a *feature pyramid*, with predictions made independently at all levels.

- 如上图所示，另一个创新点是检测时各层检测是独立进行（anchor不共享）的，最后再进行合并和抑制。

Training Neural Networks with Very Little Data-A Draft

- 作者：Hojjat Salehinejad, Joseph Barfett, Shahrokh Valaee, Senior Member, IEEE, and Timothy Dowdell.

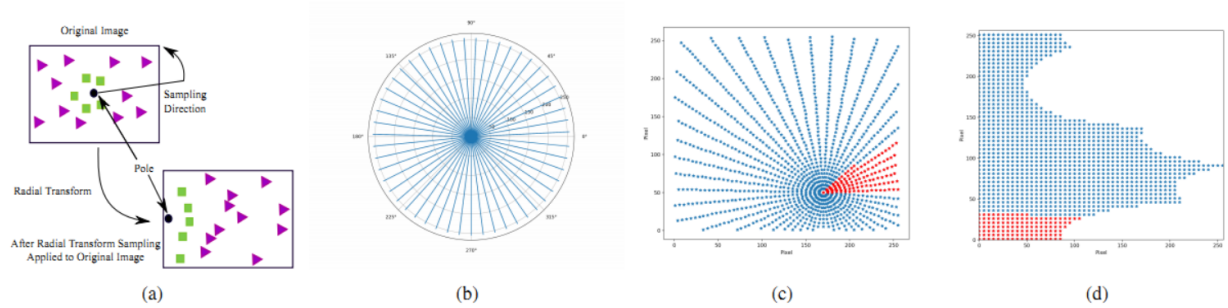


Fig. 1: Radial transform sampling. a) Mapping samples from Cartesian coordinate system (left) into a polar coordinate system (right) using the radial transform. b) Radial transform in polar coordinate system. c) Selected discrete samples of a 256 × 256 image (2D plane) using radial transform. The arbitrary selected pole is at pixel (170,50). d) Mapping of selected samples in (c) from polar coordinate system to Cartesian coordinate system. The red samples show the direction of mapping the samples in (c) into (d).

- 如上图所示，将一张输入图像由欧式空间表示转为极坐标表示，再输入网络进行训练。

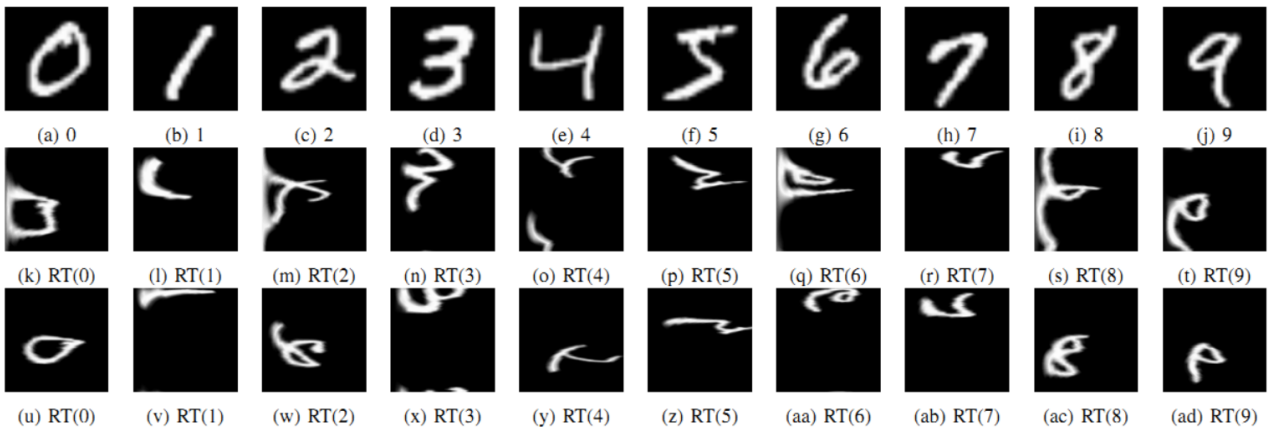
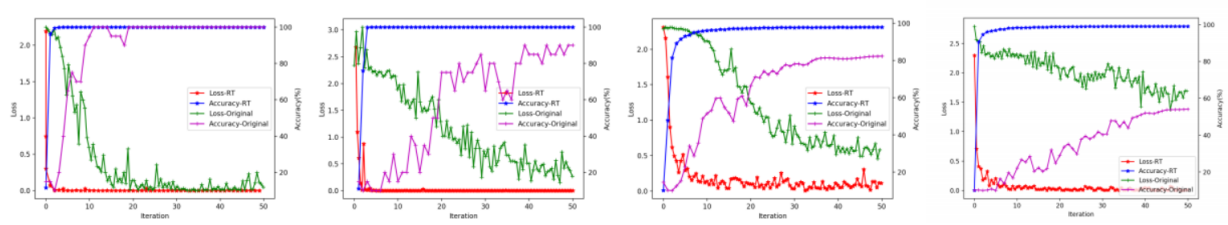


Fig. 2: Samples from the MNIST dataset and corresponding representations using radial transform RT(·) in polar coordinate system.

- 由坐标原点选取的不同，可以产生很多不同的极坐标图像（如上图所示），大大增加了数据量。
- 训练效果：
  - > AlexNet and GoogLeNet are trained once with a small dataset (i.e., 20 images per class) and once with 100 radial transformed images generated from each image (i.e., 2,000 radial images per class) of the same dataset. The number of images per class is selected such that it represents a very small dataset. For example, since the MNIST dataset has 10 classes, the original truncated dataset has 200 images and the radial transformed dataset has 20,000 images.
  - > The results clearly show that the models trained with radial transformed data have higher accuracy and greater confidence value. At competitive accuracy, the confidence of models trained with radial transformation is greater.



(a) AlexNet-Multimodal (b) GoogLeNet-Multimodal (c) AlexNet-MNIST (d) GoogLeNet-MNIST

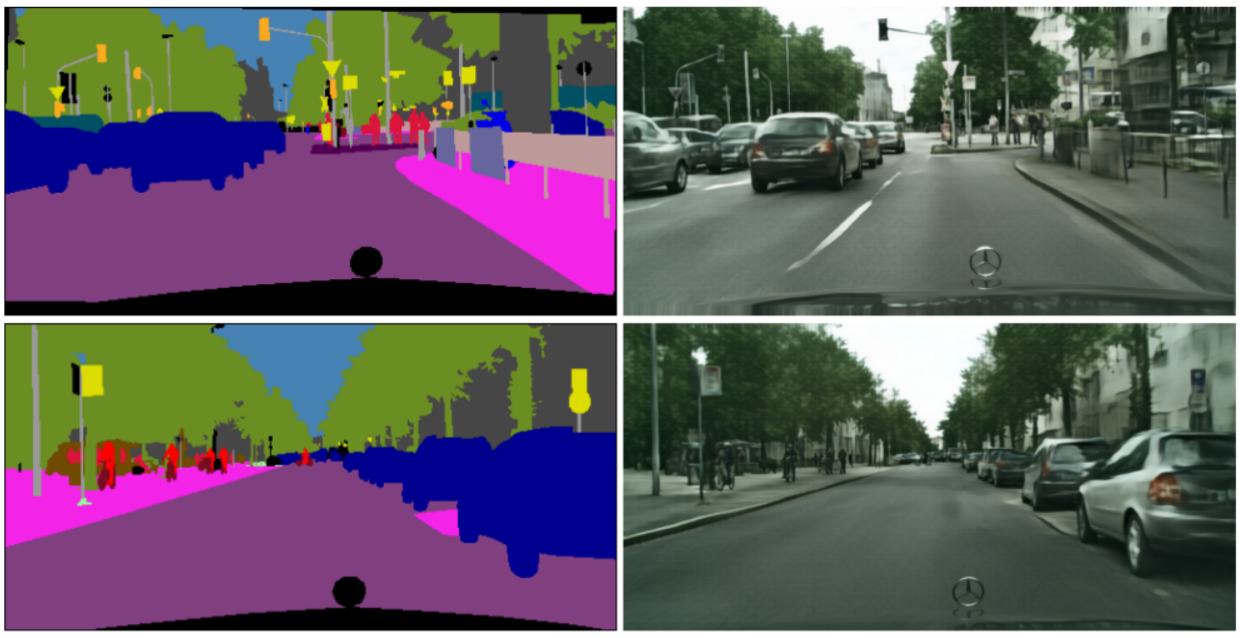
- 上图中，绿色、粉色为原始loss、accuracy曲线，红色、蓝色为极坐标系下loss、accuracy曲线。可以看到，使用极坐标变换后，极大的增加了数据量，同时loss下降、accuracy上升更快，效果更好。

Robust Scene Text Recognition with Automatic Rectification

- 作者：Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, Xiang Bai.
- STN+CNN+BLSTM+Attention-GRU.

Photographic Image Synthesis with Cascaded Refinement Networks

- 作者：Qifeng Chen, Vladlen Koltun, Intel Labs, Stanford University.
- 通过输入semantic layout信息，输出生成图像；不采用对抗式的方法（GAN），可以生成高分辨率的图像。



(a) Input semantic layouts (b) Synthesized images

Figure 1. Given a pixelwise semantic layout, the presented model synthesizes an image that conforms to this layout. (a) Semantic layouts from the Cityscapes dataset of urban scenes; semantic classes are coded by color. (b) Images synthesized by our model for these layouts. The layouts shown here and throughout the paper are from the validation set and depict scenes from new cities that were never seen during training. Best viewed on the screen.

- 输入是Pixel Wise的semantic layout信息（One Hot），网络会自动学习在对应layout中生成对应图像。

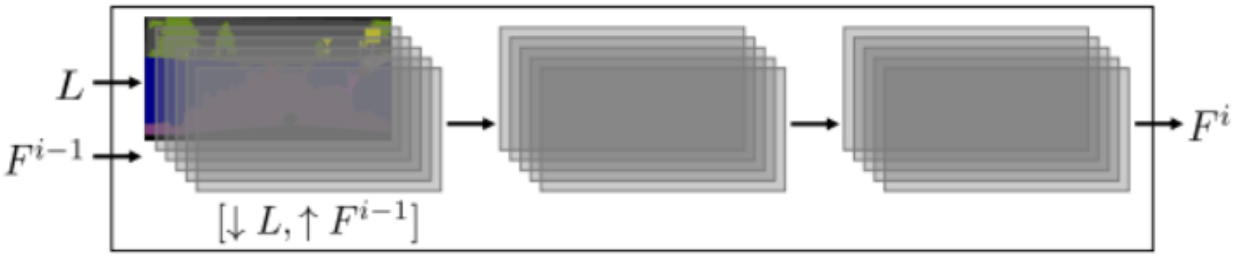
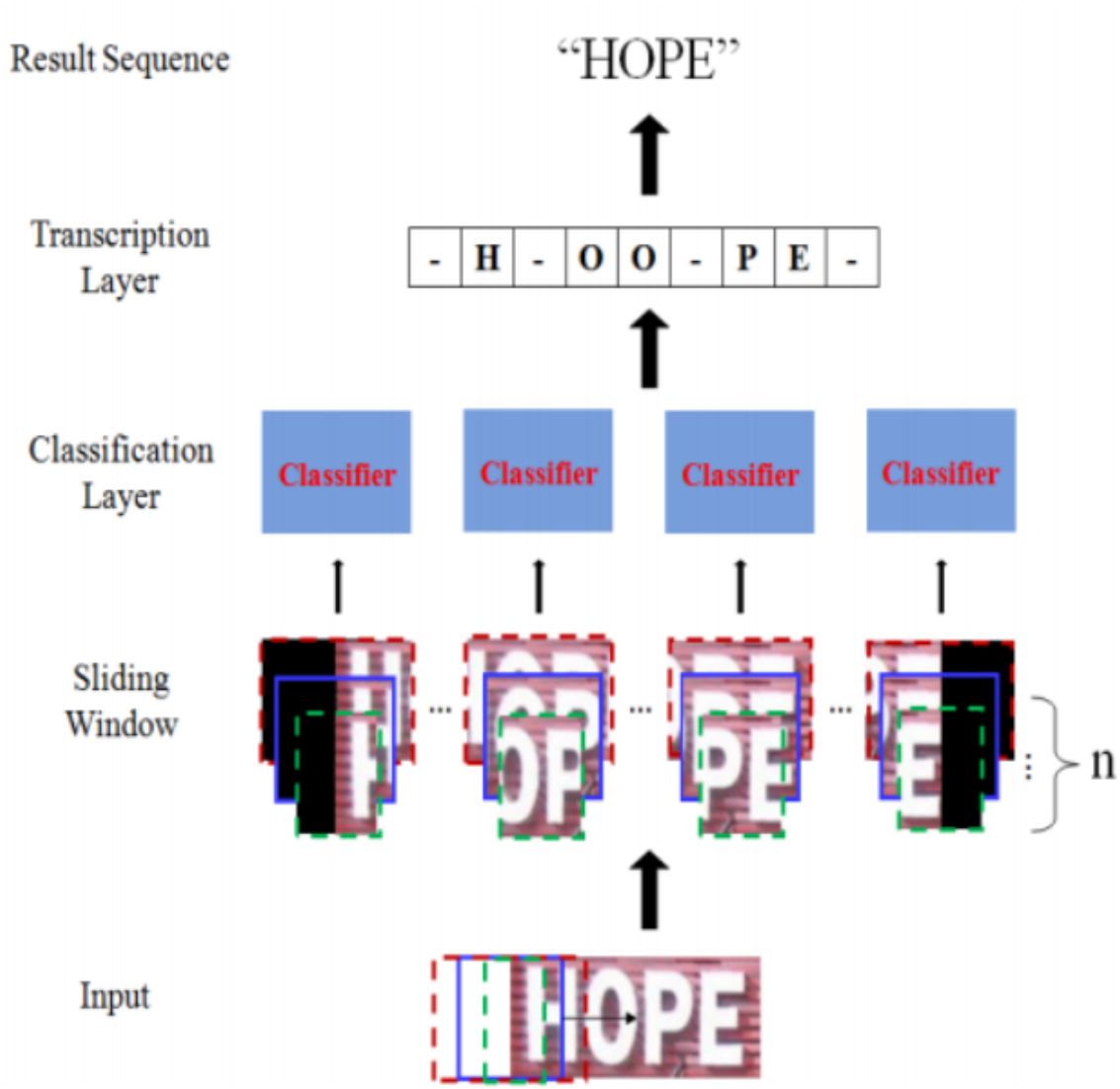


Figure 3. A single refinement module.

- 整个网络由多个上图模块组成。$\boldsymbol{L}$是Layout信息，reshape到该模块设定大小作为输入；$\boldsymbol{F^{i-1}}$是上一模块输出的feature map，通过线性插值放大（2倍），与$\boldsymbol{L}$进行通道拼接作为该模块的输入；每个模块的输出是输入大小的2倍。
- 训练时，不是简单的比较生成图像与原图的差距（距离），而是通过利用VGG-19，将生成图像（中间层特征）、原图（reshape到特征图大小）输入VGG-19，提取特征后，计算$\boldsymbol{L1}$距离作为Loss Function。
- 进一步改进，对于每一类图像，仅计算生成效果最好的图像与原图在VGG-19下特征差距，效果更好。

Scene Text Recognition with Sliding Convolutional Character Models

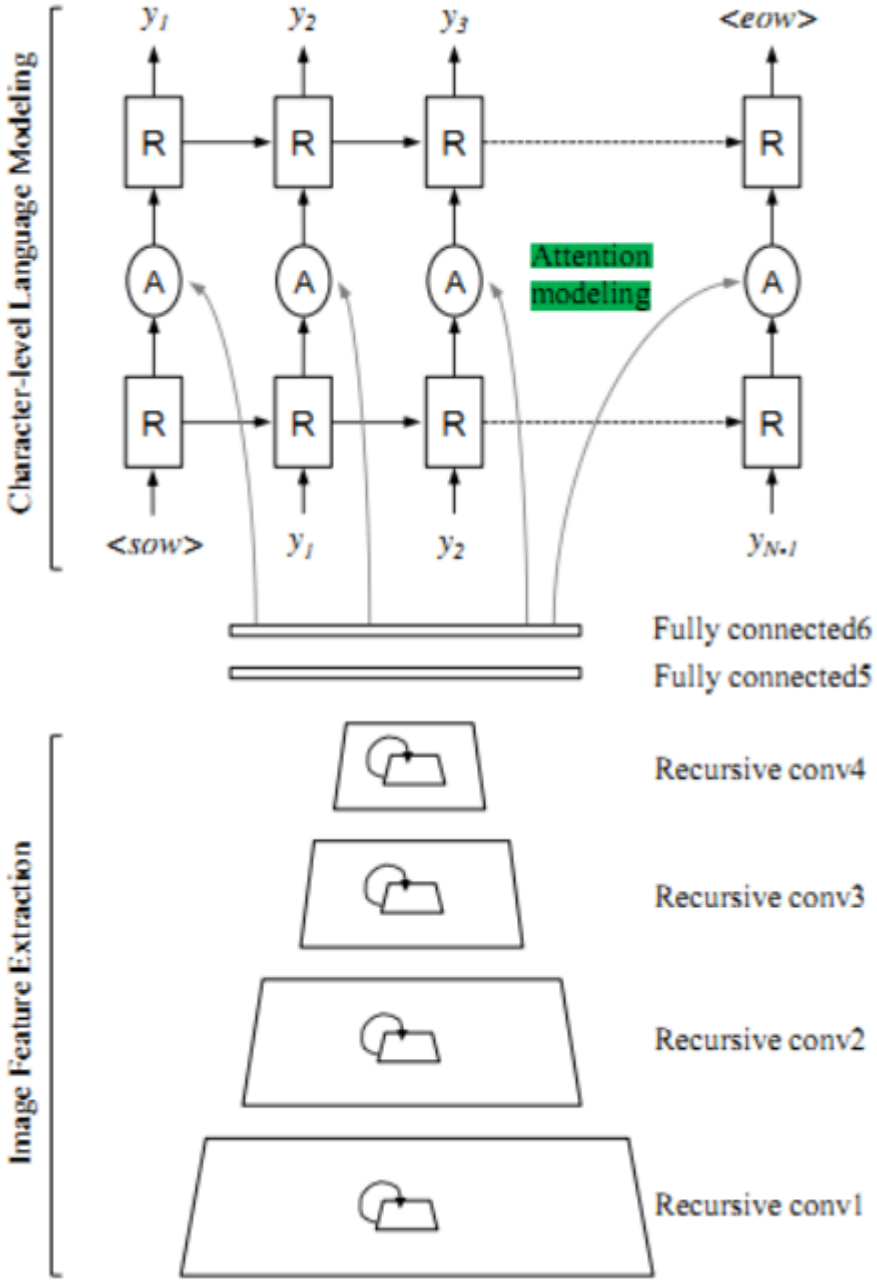
- 作者：Fei Yin, Yi-Chao Wu, Xu-Yao Zhang, Cheng-Lin Liu.



- 对比于CNN+CTC方法，将对原图直接卷积的方法改为滑窗，滑窗单字识别结果用CTC解码训练。

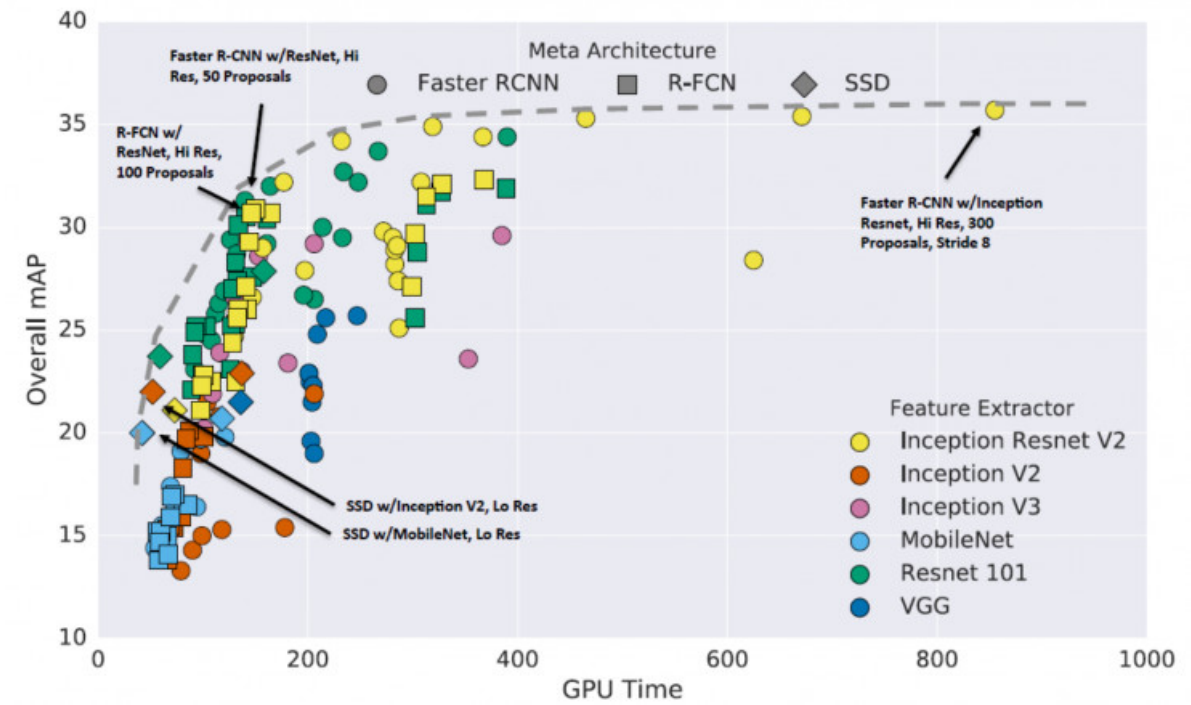
Recursive Recurrent Nets with Attention Modeling for OCR in the Wild

- 作者：Chen-Yu Lee [UC San Diego], Simon Osindero [Yahoo Inc].
- Recursive Recurrent Neural Networks with Attention Modeling (R^2AM)：
  - 使用Recursive CNN，减少参数数量；使用RNN进行序列建模；使用Soft Attention。



Speed/accuracy trade-offs for modern convolutional object detectors

- 作者：Google Research
- 这篇论文基于Tensor Flow实现了Faster-RCNN、R-FCN、SSD，并做了Speed/Accuracy trade-offs测试。
- 结论：

- 

  - （如上图）R-FCN、SSD速度较Faster-RCNN快，然而Faster-RCNN可以通过减少Proposal减少训练时间，而且精度影响较小。
  - SSD受feature extractor的影响较Faster RCNN和R-FCN小。
  - 参数相同情况下，精度：Faster RCNN > R-FCN > SSD。
Model summary	minival mAP	test-dev mAP
(Fastest) SSD w/MobileNet (Low Resolution)	19.3	18.8
(Fastest) SSD w/Inception V2 (Low Resolution)	22	21.6
(Sweet Spot) Faster R-CNN w/Resnet 101, 100 Proposals	32	31.9
(Sweet Spot) R-FCN w/Resnet 101, 300 Proposals	30.4	30.3
(Most Accurate) Faster R-CNN w/Inception Resnet V2, 300 Proposals	35.7	35.6

    Table 3: Test-dev performance of the "critical" points along our optimality frontier.
  - （如上图）作者在保证相同条件下，各种模型方法达到的最好效果。

R-FCN: Object Detection via Region-based Fully Convolutional Networks

- 作者：Jifeng Dai, Yi Li, Kaiming He, Jian Sun, Microsoft Research.
- Faster-RCNN速度较慢因为ROI Pooling后的计算（分类、回归）并不共享；R-FCN（本文）使用region based selective pooling，RPN之后直接selective pooling+分类/回归，大大降低了计算量，提升了速度。
- Seletive Pooling：

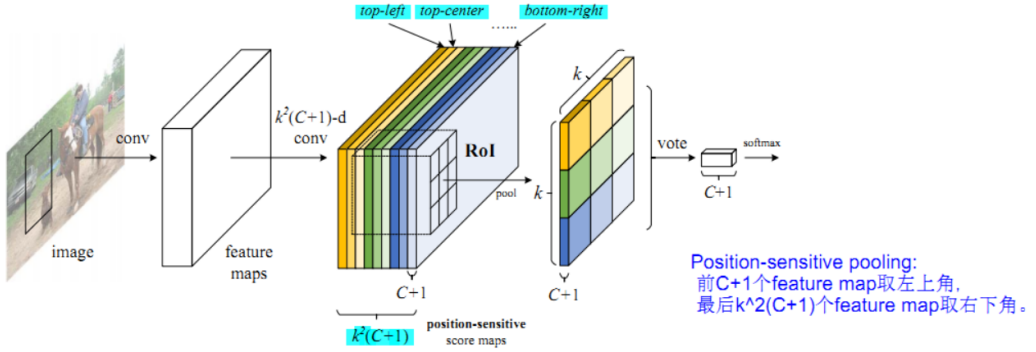
- 

  Figure 1: Key idea of **R-FCN** for object detection. In this illustration, there are $k \times k = 3 \times 3$ position-sensitive score maps generated by a fully convolutional network. For each of the $k \times k$ bins in an RoI, pooling is only performed on one of the k^2 maps (marked by different colors).

  - RPN产生$k^2(C+1)$个channel的feature map，对每个ROI进行Pooling的时候，将ROI分为k^2个区域，每$(C+1)$个channel的feature map取其中一个区域进行Pooling，使得学习到的feature map响应于ROI/原图中特定位置；得到的$(C+1)$个channel做average pooling，得到$(C+1)$维的Vector，直接作为softmax的输入进行分类。回归分支的做法类似，将$k^2(C+1)$换成$4k^2$即可。

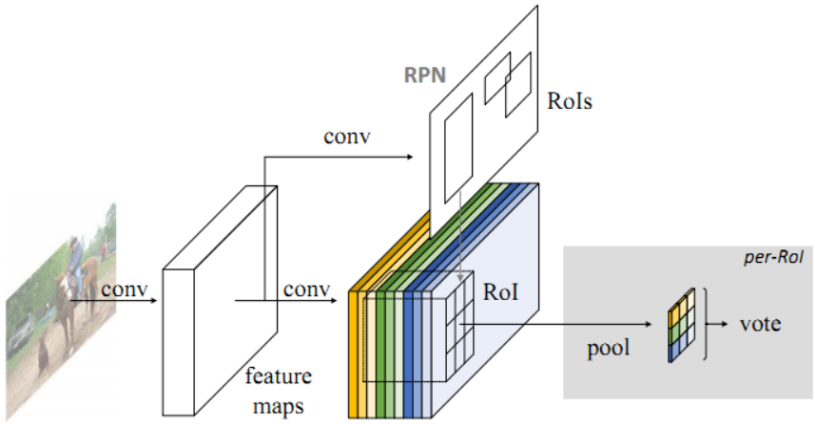
- 

  Figure 2: Overall architecture of R-FCN. A Region Proposal Network (RPN) [19] proposes candidate RoIs, which are then applied on the score maps. All learnable weight layers are convolutional and are computed on the entire image; the per-RoI computational cost is negligible.

  - 整体架构如上图，RPN产生ROI后，直接Selective Pooling，pooling输出直接用于分类、回归；Pooling后不再有参数需要学习，同时对每个ROI的selective pooling操作可以合并，大大降低了计算量。

Understanding Deep Architectures using a Recursive Convolutional Network

- 作者：David Eigen，Jason Rolfe，Rob Fergus，Yann LeCun。
- 这是14年的文章，提出**Recursive CNN** 结构，讨论网络层数、feature map数量和参数对识别效果的影响。
- Recursive CNN:
  - 简单概括：所有层（除了第一层）的feature map数相同，参数共享，层数可调。
  - 可以看成类似递归结构，故名Recursive CNN。
  - > The final hidden layer is subject to pixel-wise L2 normalization and passed into a logistic classifier to produce a prediction Y.

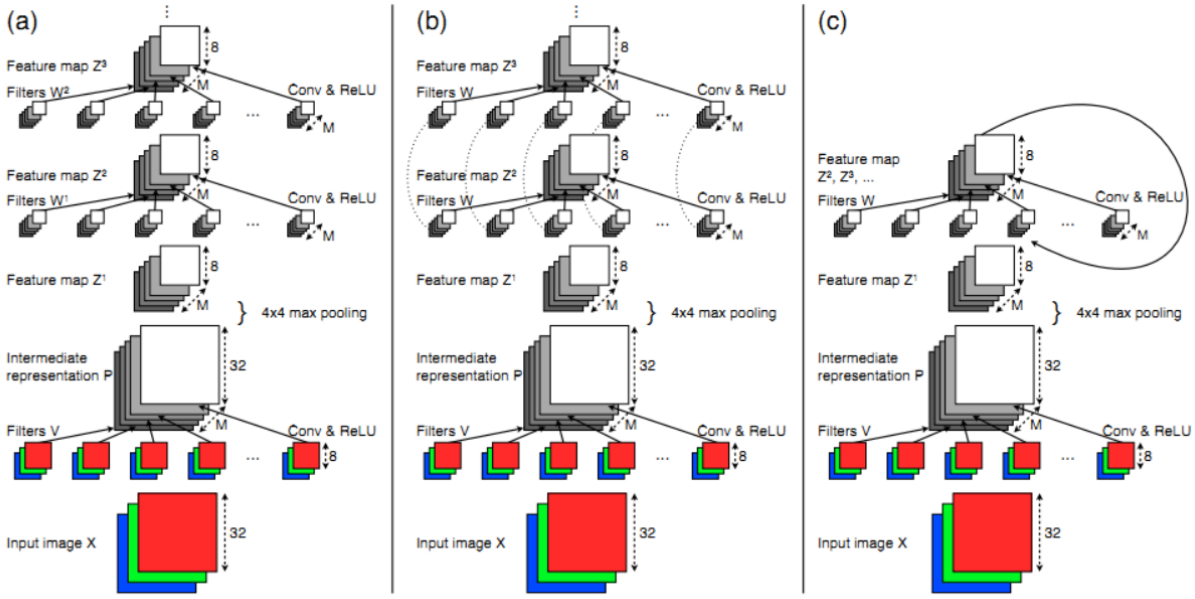


Figure 1: Our model architecture prior to the classification layer, as applied to CIFAR and SVHN datasets. (a): Version with un-tied weights in the upper layers. (b): Version with tied weights. Kernels connected by dotted lines are constrained to be identical. (c): The network with tied weights from (b) can be represented as a recursive network.

- 实验：

> 1. Control for M and P, vary L: Using the tied model (constant M and P), we evaluate performance for different numbers of layers L.
> 2. Control for M and L, vary P: Compare pairs of tied and untied models with the same numbers of feature maps M and layers L. The number of parameters P increases when going from tied to untied model for each pair.
> 3. Control for P and L, vary M: Compare pairs of untied and tied models with the same number of parameters P and layers L. The number of feature maps M increases when going from the untied to tied model for each pair.

  - 一定范围内，层数越多、参数越多，训练、泛化效果越好（符合常识）。
  - > We find that despite the different numbers of feature maps, the tied and untied models perform about the same in each case. Thus, performance is determined by the number of parameters.

- 作者建议：
  - > Allocating a fixed number of parameters across multiple layers tends to increase performance compared to putting them in few layers, even though this comes at the cost of decreasing the feature map dimension.

Fully Convolutional Recurrent Network for Handwritten Chinese Text Recognition

- Path Signature + FCNN + BLSTM + CTC / Language Model.

MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications

- 作者：Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, Hartwig Adam, Google Inc.
- 使用 Depth-wise Separable Convolution（逐通道卷积）减少参数数量，使复杂 网络也可以部署在手机上。

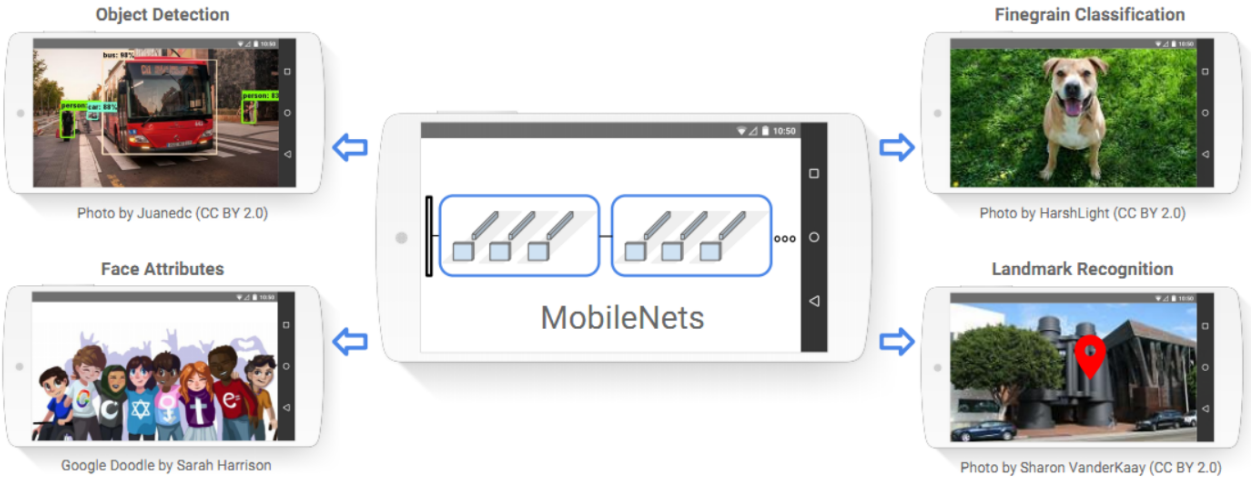


Figure 1. MobileNet models can be applied to various recognition tasks for efficient on device intelligence.

- Depth-wise Separable Convolution（逐通道卷积）中每个kernel仅对输入feature map中一个channel进行卷积，设输入为C_i通道，输出为C_o通道，卷积核大小为$k \times k$，则一般卷积参数为$C_i \times C_o \times k \times k$，DSC参数为$C_i \times k \times k$，当$C_i \neq C_o$时，采用$1 \times 1$的一般卷积进行通道融合。
- 每层结构如下右图所示：

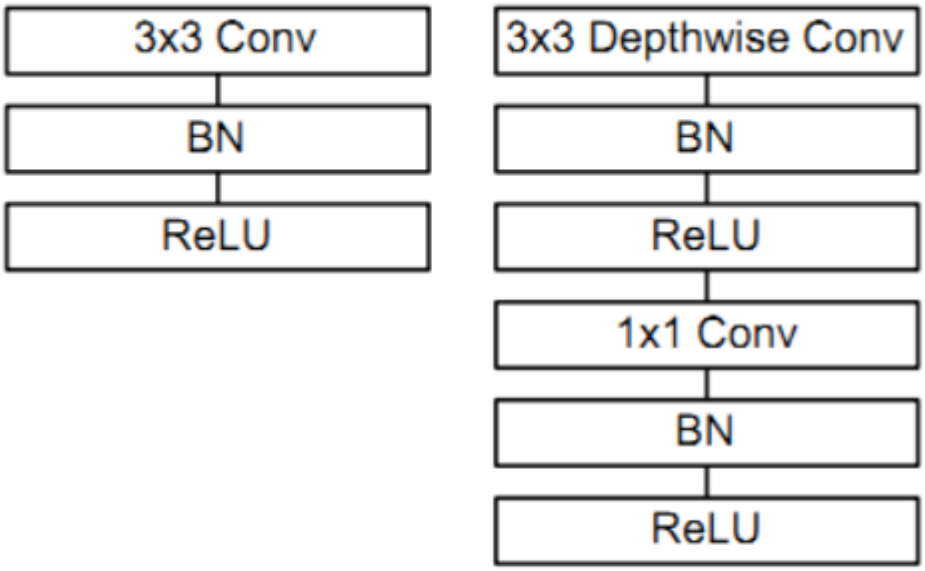


Figure 3. Left: Standard convolutional layer with batchnorm and ReLU. Right: Depthwise Separable convolutions with Depthwise and Pointwise layers followed by batchnorm and ReLU.

Design of a Very Compact CNN Classifier for Online Handwritten Chinese Character Recognition Using DropWeight and Global Pooling

- 通过动态调整阈值，逐步实现模型的压缩。

DropSample: [A New Training Method to Enhance Deep Convolutional Neural Networks for Large-Scale Unconstrained Handwritten Chinese Character Recognition]

- 以Softmax输出结果为置信度，针对mislabeled sample、noisy sample、hard sample进行优化。
    - 置信度高的、识别较好的：逐步减少出现几率；
    - 置信度较低、较难以识别的样本：逐步增加训练几率；
    - 误标注样本、置信度较低的样本：逐步除去该样本。

Toward high-performance online HCCR: a CNN approach with DropDistortion, path signature and spatial stochastic max-pooling

- 通过逐步降低数据形变程度，使得网络在训练后期能更好的拟合数据原始分布；引入空间随机Pooling。

An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition

- 作者：Baoguang Shi, Xiang Bai and Cong Yao.
- CNN+BLSTM+CTC
