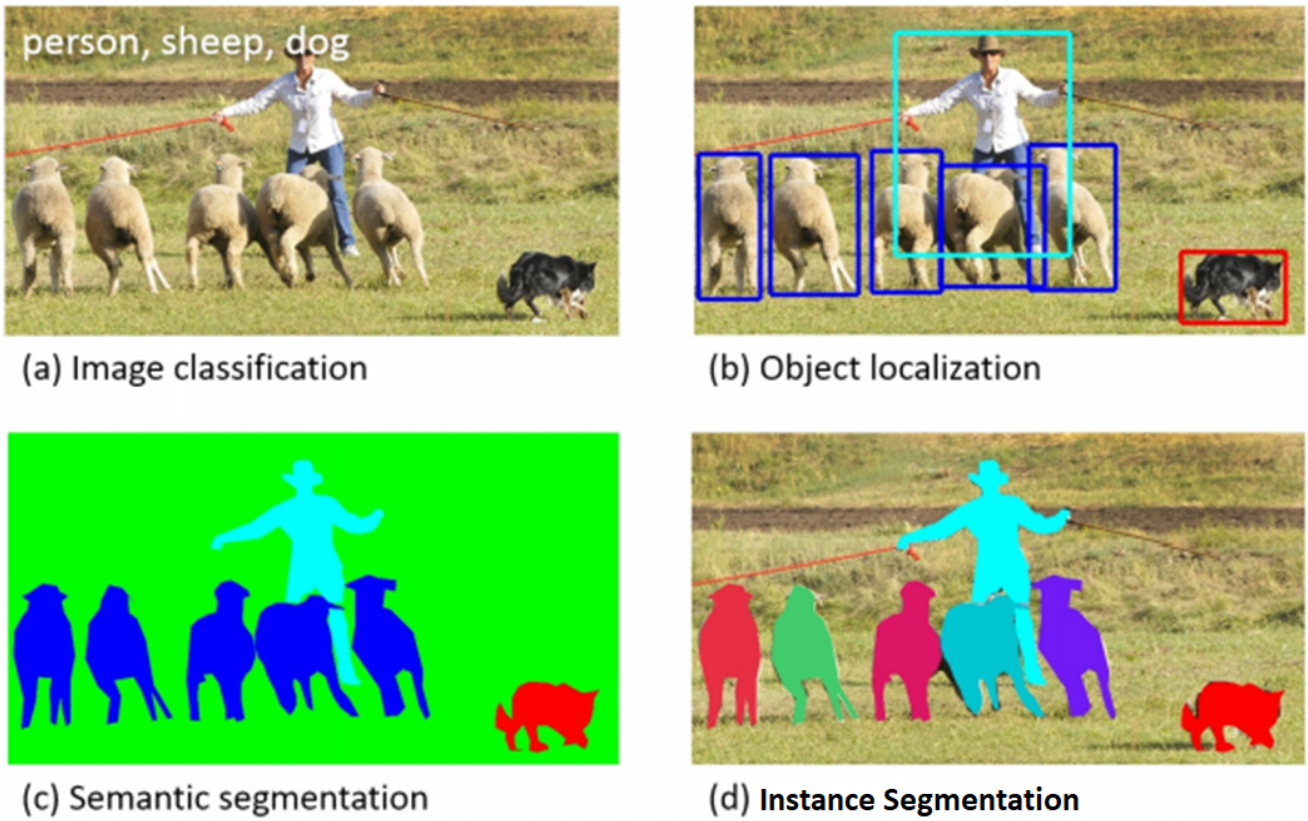
影像分割 Image Segmentation — 語義分割 Semantic Segmentation(1)

李謦伊 Follow
May 27 · 8 min read

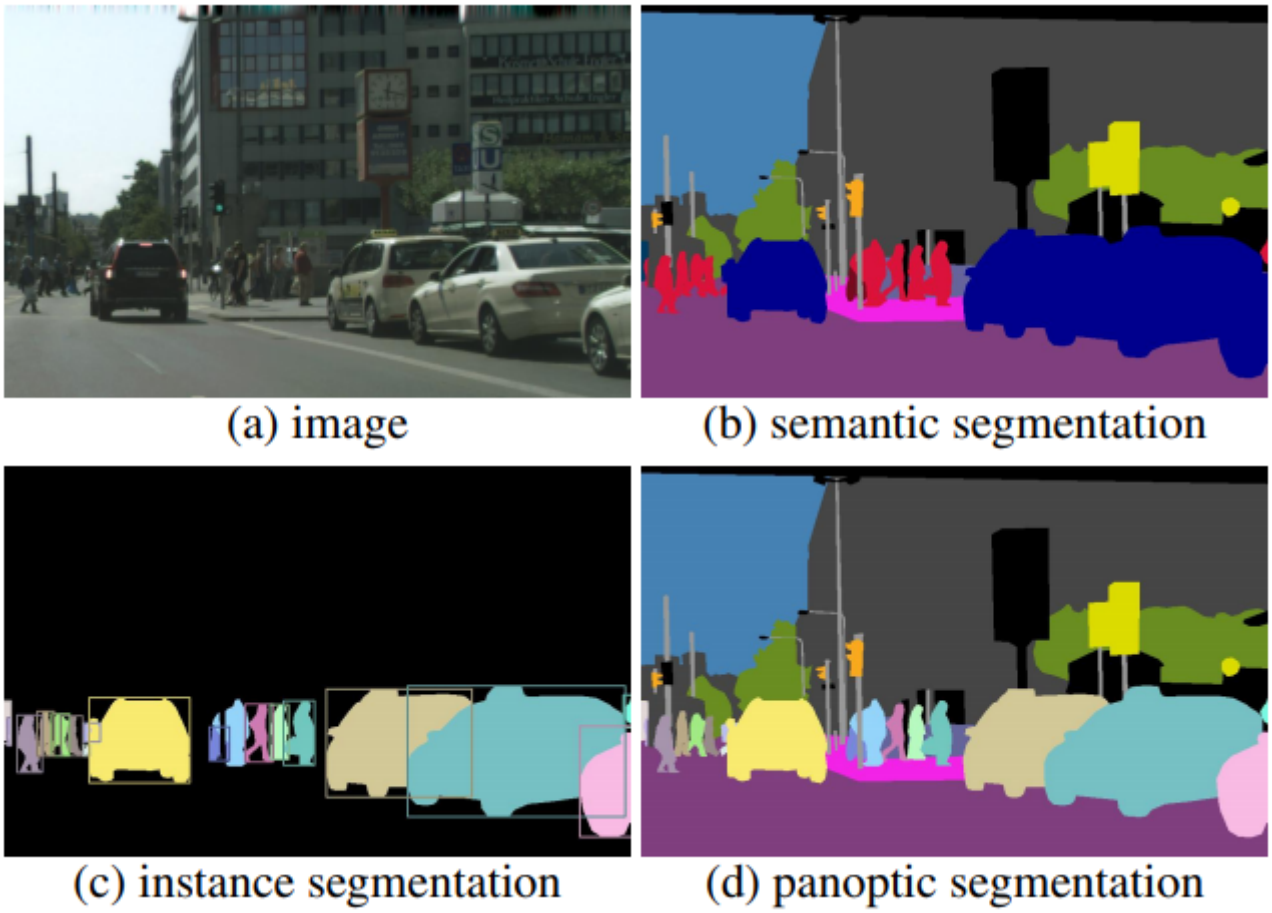
深度學習在 Computer Vision (CV) 領域的幾項重要任務應用分別有 Image Classification (影像分類)、Object Detection (物件偵測)、Image Segmentation (圖像分割)，其中 Image Segmentation 是針對像素進行檢測分類，能應用於人臉辨識、自動駕駛、醫學圖像分析等任務上。



source

Image Segmentation 包括 Semantic Segmentation (語義分割)、Instance Segmentation (實例分割)、Panoramic Segmentation (全景分割)。

- Semantic Segmentation 是指將圖像中的所有像素點進行分類。
- Instance Segmentation 是物件偵測和語義分割的結合，任務相對較難。其方法是針對感興趣的像素點進行分類，並且將各個物件定位，即使是相同類別也會分割成不同物件。由上圖可以很清楚地看到 Semantic Segmentation 及 Instance Segmentation 的差異。
- Panoramic Segmentation 則是更進一步結合了語義分割和實例分割，顧名思義就是要對各像素進行檢測與分割，同時也將背景考慮進去。



source

本文要來介紹 Semantic segmentation 的代表演算法、paper、code，由於內容篇幅過多，會拆分成 (1)、(2)。

Semantic segmentation

代表演算法有分為以下部分，將介紹其中幾種較著名的模型

- 基於卷積神經網路: FCN、DeconvNet、U-Net、SegNet、DeepLab、RefineNet、PSPNet、GSCNN 等
- 基於 Recurrent Neural Network (RNN): ReNet、ReSeg 等

- 基於 Generative adversarial network (GAN): pix2pix、Probalistic Unet 等
- 基於 Transformer: HRNet、OCRNet、HRNet-OCR、Point Transformer、SETR

Fully Convolutional Networks (FCN，2014)

FCN 是圖像分割的開山始祖，對於後續的語義分割任務奠定了很重要的基礎。詳細可參考我之前寫的文章: Fully Convolutional Networks 論文閱讀

DeconvNet (2015)

Github: https://github.com/HyeonwooNoh/DeconvNet

DeconvNet 基於 FCN 做改進，網路架構也是由卷積、反卷積所組成 (也可稱為 Encoder-Decoder 結構)，卷積網路包括卷積層與 Maxpooling、反卷積則是有 Unpooling 與反卷積層，其中中間層使用 1x1 卷積層連接。

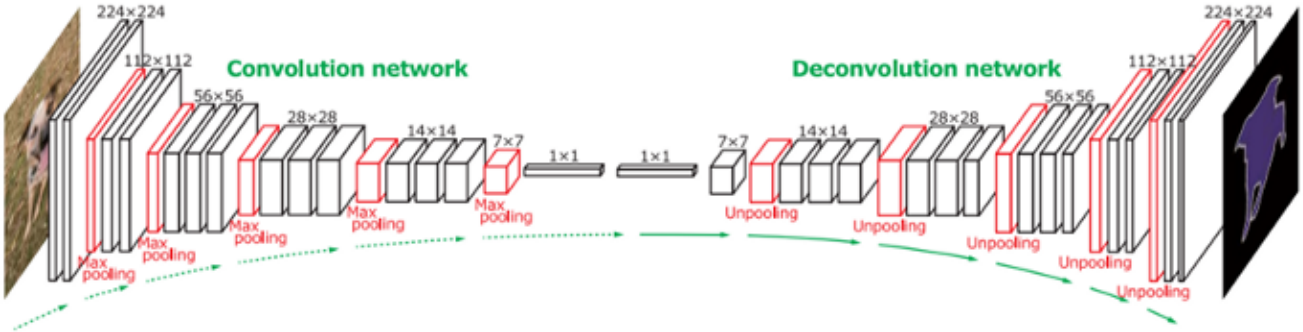


Figure 2. Overall architecture of the proposed network. On top of the convolution network based on VGG 16-layer net, we put a multi-layer deconvolution network to generate the accurate segmentation map of an input proposal. Given a feature representation obtained from the convolution network, dense pixel-wise class prediction map is constructed through multiple series of unpooling, deconvolution and rectification operations.

為了改善 FCN 對於細節處理不夠好的問題，在進行 maxpooling 時，會先保存 max 值的位置，之後 unpooling 就可以將其對應回原來的位置，其餘非 max 值則補 0，然後再藉由反卷積進行上採樣 (upsampling)

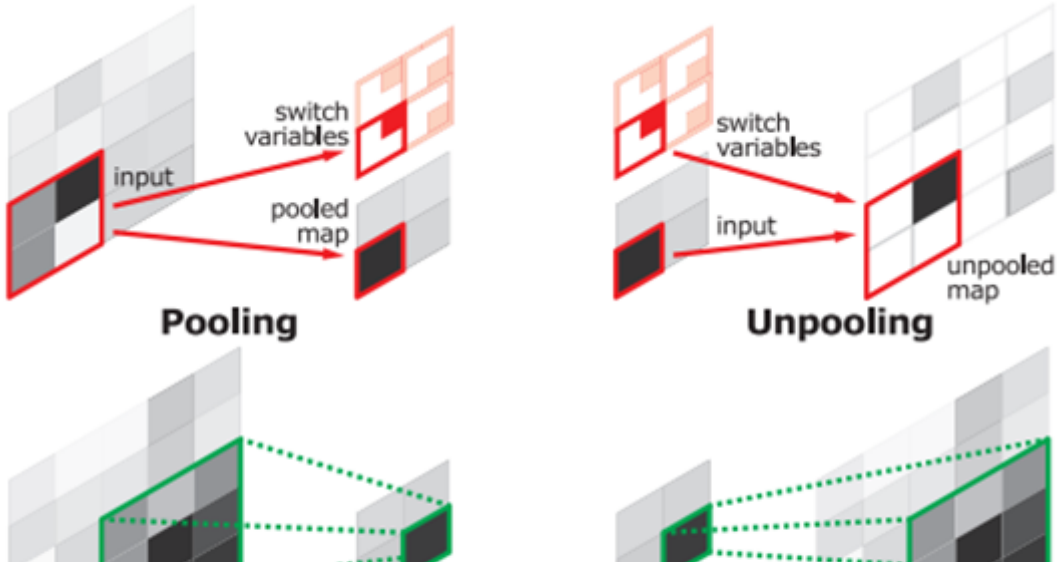


Figure 3. Illustration of deconvolution and unpooling operations.

U-Net (2015)

Keras github: https://github.com/zhixuhao/unet

U-Net 基於 Encoder-Decoder 結構，主要應用於醫學影像分割。Encoder (U-Net 稱為 contracting path) 負責提取特徵、Decoder (U-Net 稱為 expansive path) 則是用於 upsampling，網路架構形似 U。此外，經過上採樣後會與 contracting path 進行 concat，不過 contracting path 特徵圖的尺寸較大需要經過 crop 裁剪。

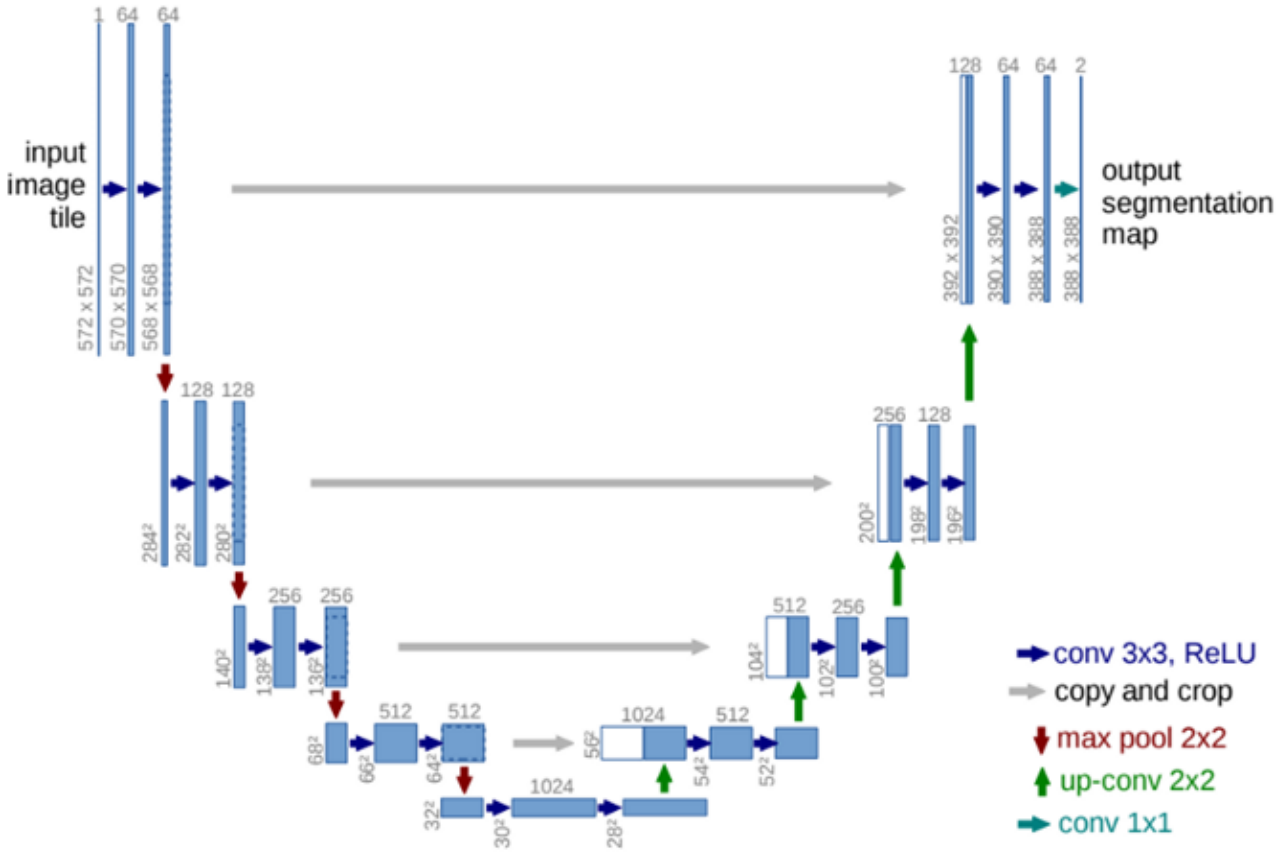


Fig. 1. U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations.

SegNet (2015)

Github: https://github.com/alexgkendall/SegNet-Tutorial

SegNet 結構與 DeconvNet 類似，不同的是去除了中間的 1x1 卷積層，為了降低記憶體使用及提升推理速度，主要應用於場景理解。由下圖可以看到 SegNet 的網路架構，Encoder 用於提取特徵、Decoder 則是將特徵圖做 upsampling。

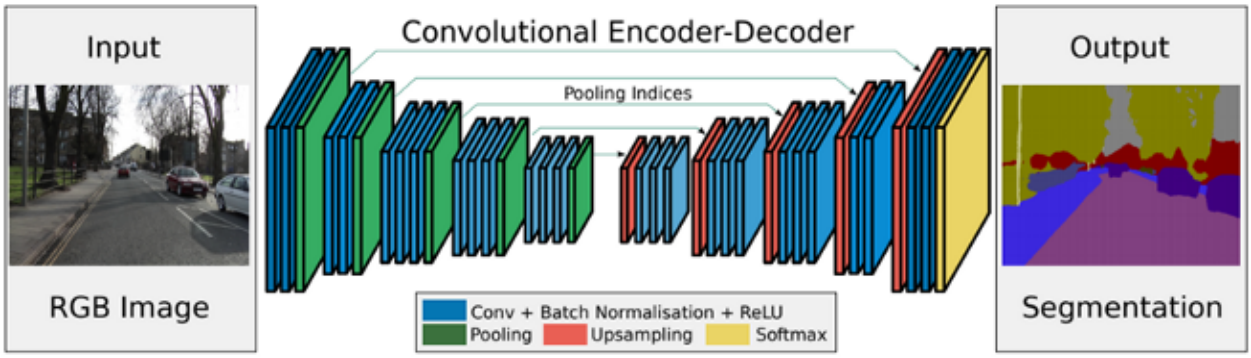


Fig. 2. An illustration of the SegNet architecture. There are no fully connected layers and hence it is only convolutional. A decoder upsamples its input using the transferred pool indices from its encoder to produce a sparse feature map(s). It then performs convolution with a trainable filter bank to densify the feature map. The final decoder output feature maps are fed to a soft-max classifier for pixel-wise classification.

SegNet 與 U-Net 差異?

SegNet 的 upsampling 只採用反卷積，而 U-Net 的 upsampling 是採用反卷積後與 contracting path 進行 concat，再經過兩個 3x3 卷積層。

SegNet 與 FCN 差異?

SegNet 在進行 maxpooling 時，會先保存 max 值的位置，之後 upsampling 就可以將其對應回原來的位置，其餘非 max 值則補 0。而 FCN 的 upsampling 是採用反卷積後跟另一個相同尺寸的特徵圖相加，再進行一次上採樣。

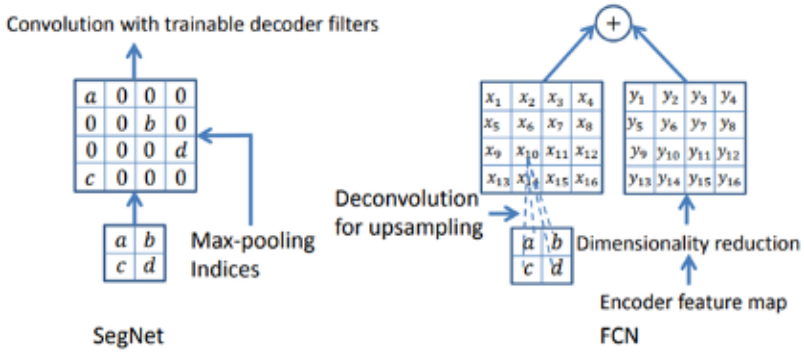


Fig. 3. An illustration of SegNet and FCN [2] decoders. a, b, c, d correspond to values in a feature map. SegNet uses the max pooling indices to upsample (without learning) the feature map(s) and convolves with a trainable decoder filter bank. FCN upsamples by learning to deconvolve the input feature map and adds the corresponding encoder feature map to produce the decoder output. This feature map is the output of the max-pooling layer (includes sub-sampling) in the corresponding encoder. Note that there are no trainable decoder filters in FCN.

DeepLab v1~v3、DeepLab v3+

Github: https://github.com/tensorflow/models/tree/master/research/deeplab

可參考: Deeplab 系列介紹

RefineNet (CVPR 2017)

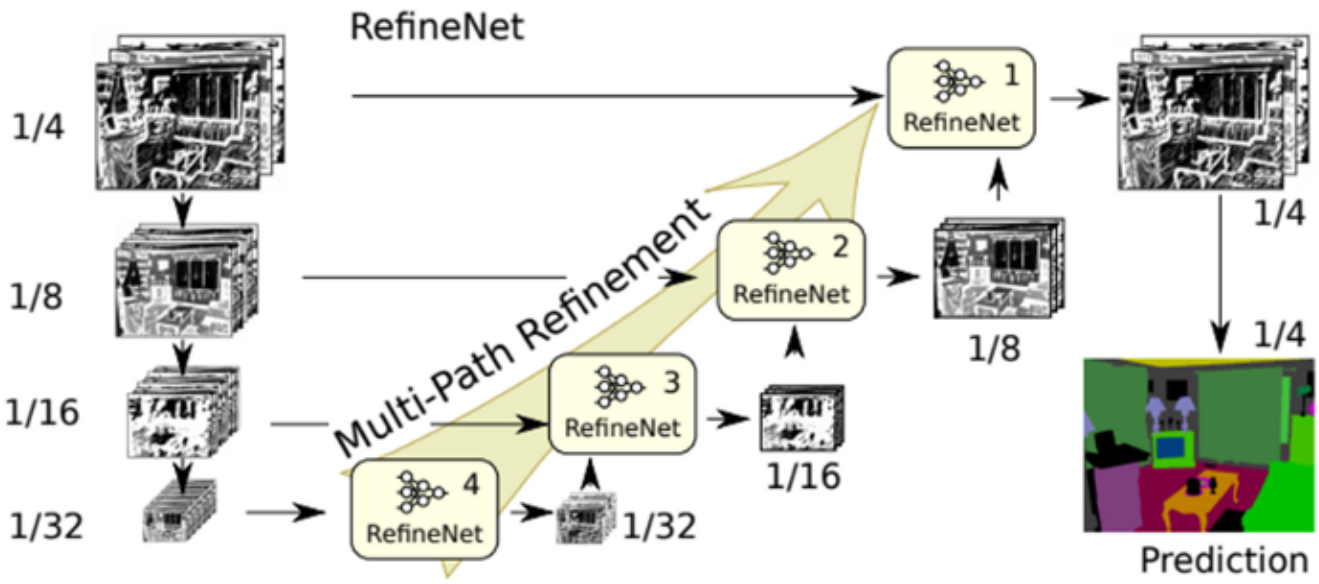
Github: https://github.com/guosheng/refinenet

在當前的語義分割問題主要採用反卷積和 Atrous Convolution，但這兩種方法有一些缺點:

- 反卷積不能恢復 low-level 的特徵，對於細節的預測不佳
- Atrous Convolution 需要消耗大量的 GPU 和計算量

作者認為所有 level 的特徵都是有作用的，因而提出一種多階段的網路架構 RefineNet，目的就是把各個 level 的特徵融合以生成 high resolution 預測。

其網路架構以 ResNet 為基礎，根據 image resolution 分成四部分：原始圖像的 1/4, 1/8, 1/16, 1/32。接著分別向右輸入至對應的 RefineNet block 進行融合及 refine，其中 RefineNet-4 只有一個輸入，而其他的 block 都有兩個輸入。另外，這四個 RefineNet block 的參數不共享。



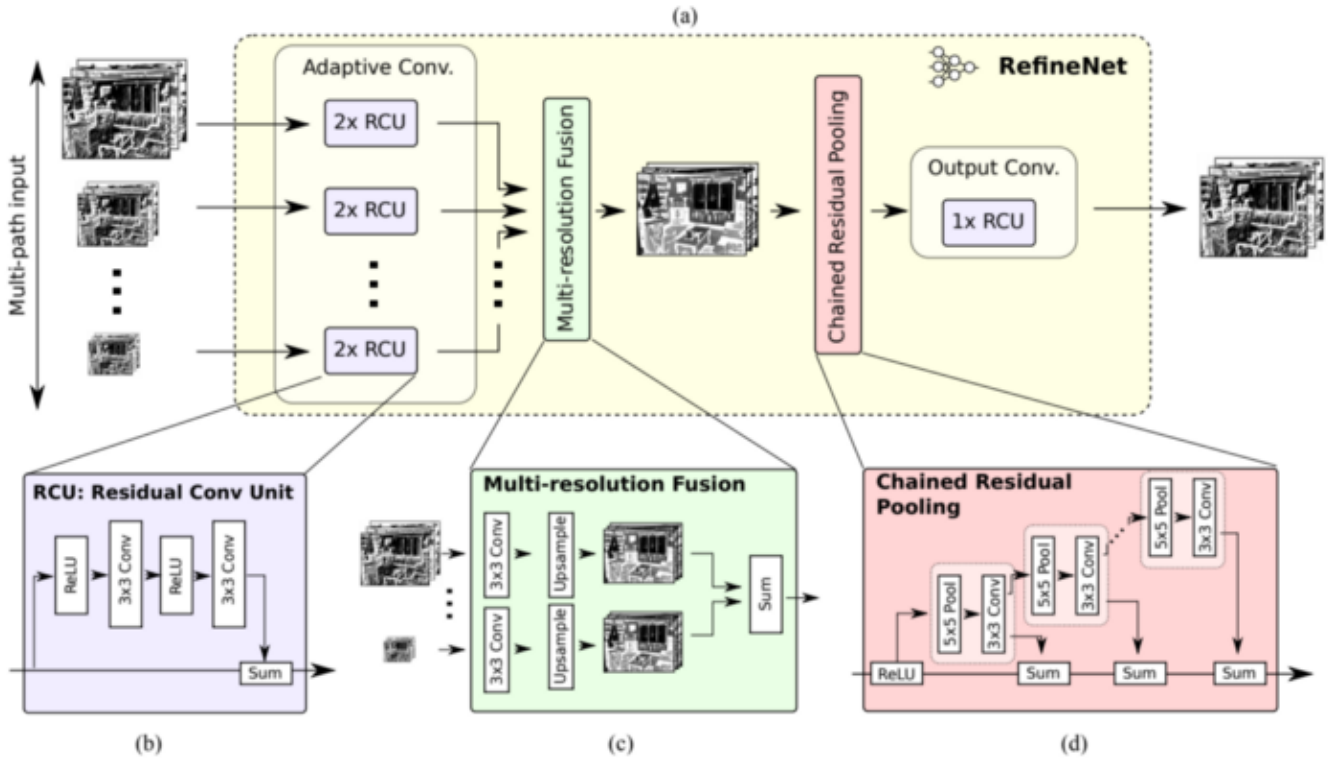
RefineNet

RefineNet block 是由 Residual convolution unit (RCU)、Multi-resolution fusion、Chained residual pooling、Output convolutions 所組成

- Residual convolution unit (RCU): 就是簡化版的 ResNet block，用於提取特徵。
- Multi-resolution fusion: 除了 RefineNet-4 之外的 block 都有兩個輸入，因此在這部分要將兩個不同 resolution 的特徵融合。作法是先進行 3x3 卷積後，把較小的特徵圖上採樣至較大的特徵圖大小，再將兩者相加。

- Chained residual pooling: 這部分是要從大範圍的區域中獲取背景內容，作法是先通過 ReLU 後分為兩路，上半部進行多次池化和卷積，再與下半部相加。

- Output convolutions: 在輸出前會再經過一次 RCU。



RefineNet block

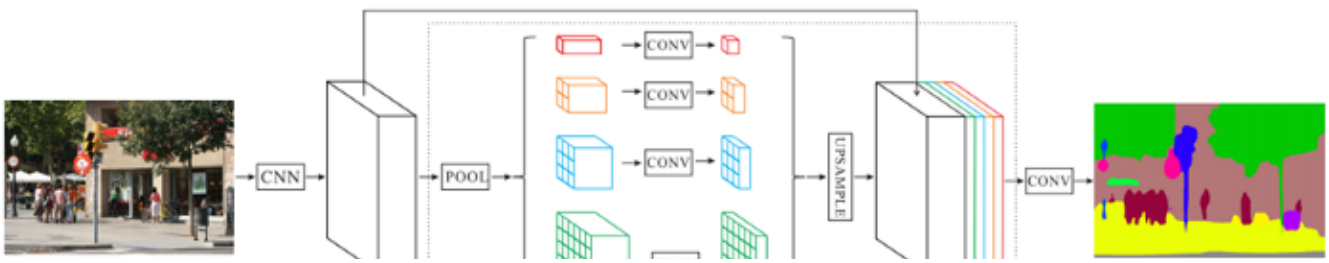
PSPNet (CVPR 2017)

Github: https://github.com/hszhao/PSPNet

PSPNet 提出基於融合全局上下文訊息的金字塔網路結構，在 2016 年 ImageNet 比賽中得到 scene parsing 的冠軍，且收錄於 CVPR 2017。

其架構如下圖所示，backbone 採用帶有 dilated 的 ResNet 以增大感受野，接著使用 Pyramid Pooling Module 提取特徵後，再經過一層卷積層得到最終的預測結果。

Pyramid Pooling Module 的作法為使用不同 size 的 Global average pooling 提取全局上下文訊息，分別為 1x1、2x2、3x3、6x6。再經過 1x1 卷積降維後，上採樣至與輸入特徵圖相同的大小，最後再將這些特徵進行 concat，融合全局特徵與細節特徵。



(a) Input Image (b) Feature Map (c) Pyramid Pooling Module (d) Final Prediction

Figure 3. Overview of our proposed PSPNet. Given an input image (a), we first use CNN to get the feature map of the last convolutional layer (b), then a pyramid parsing module is applied to harvest different sub-region representations, followed by upsampling and concatenation layers to form the final feature representation, which carries both local and global context information in (c). Finally, the representation is fed into a convolution layer to get the final per-pixel prediction (d).

相關文章

影像分割 Image Segmentation — 語義分割 Semantic Segmentation(2)

Deep Learning Segmentation Unet Segnet