Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks

- **Introduction**

  這篇是 MS COCO 2015 Detection 競賽第三名，同時也發表在 CVPR 2016。Inside-Outside 是相對 ROI 而言。過去 RCNN, Fast-RCNN 及 Faster-RCNN 一系列的做法，經過 ROI-pooling 後，只會利用到此 ROI 範圍內的 feature。而這篇 paper 希望除了可以使用 Inside ROI 的資訊，同時也將 Outside ROI 的資訊加入。

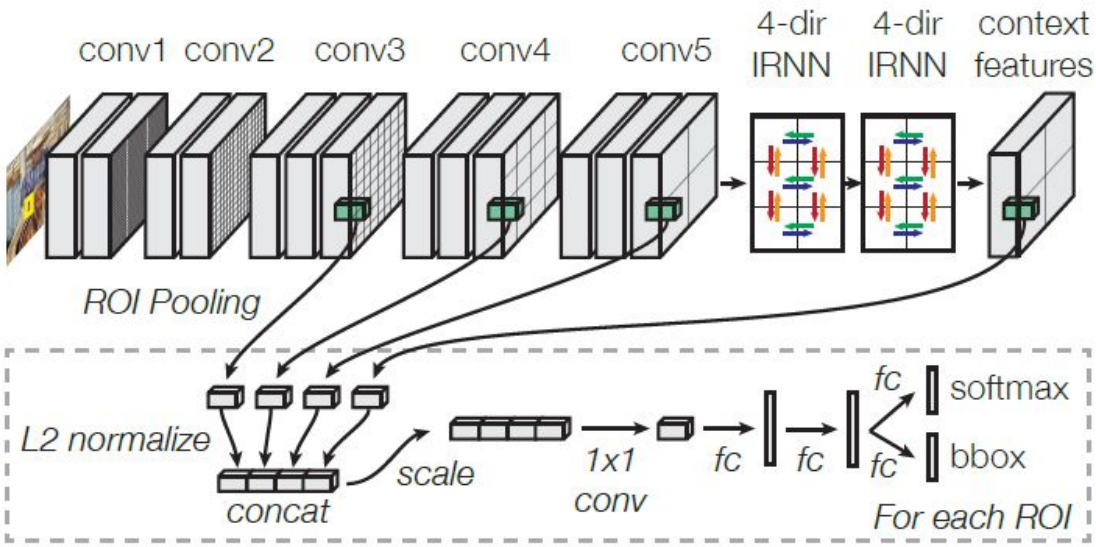

Fig1 :ION架構圖

  - ❖ Inside ROI
    inside ROI最主要是對每個region proposals取不同layer的convolution資訊，也就是multi-scale的feature來偵測各種不同大小的物體，承襲 VGG-16 的架構，取 conv3、conv4、conv5 的 feature 作為 Inside ROI 的資訊，因為這樣可以留下較多圖片中局部的資訊，較能分辨小物體

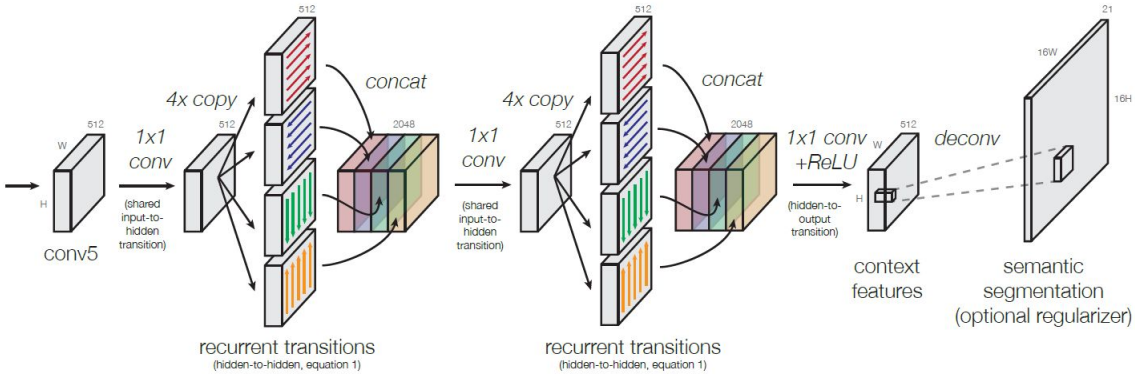

Fig2 :Four-directional IRNN架構

- ❖ Outside ROI
  inside ROIoutside ROI是利用作者新提出的4-directional IRNN擷取整體或局部的context feature，把 ROI 範圍外的資訊一起納入，作者利用由Le等人提出的IRNN，IRNN最大的好處是解決原本的RNN很容易會因為層數過多而導致vanishing gradients的問題，Le以ReLU代替logistic 或tanh組成的RNN架構，並利用identity matrix 初始化recurrent matrix，而IRNN 和 LSTM、GRU 的表現接近，但是速度更快，作者採用這樣的架構不只因為速度，也因為IRNN比較容易實作及平行化。
  本篇作者提出的 4-directional 的概念，由上、下、左、右，四個獨立的IRNN組成，而這四個 IRNN 是共享相同的 Input，也就是將 VGG conv5 的 feature 經過 1x1 convolution 即為 IRNN 的 input。利用 RNN 會隱含過去資訊的特性，把 ROI 周圍四個方向的資訊包含進來。

最後將 conv3、conv4、conv5 的 feature，以及 stacked 4-directional IRNN 輸出的 context feature 送入 ROI pooling 的流程，為了從不同的layer取出feature，每層的feature都必須要經過 feature normalize、concat、scale 以及利用 1x1 convolutional 降維產生統一大小的 feature descriptor，再送進 fully-connected layers 預測 class 及 bounding box 修正量。

- **Unique properties and where it can be applied to take advantage of the properties**

  - ❖ multi-layer ROI pooling
    本篇在MSCOCO以及VOC2007中，對於小物體的偵測進步特別大，除了RNN以外，也要歸功於只用了多層convolution的資訊，以Faster-RCNN為例，若只有利用conv5的資訊，則因為reveptive field過大，因此較能偵測大物體，但若加上了conv3 、conv4的資訊，則可以對於小物體有更高解析度的資訊，進而提升了小物體的偵測。

- 利用normalize、concat、scale、convolutional連接資訊
  本篇的方法連接方式較複雜，不只結合許多層的資訊，還有不同方式(convolution、IRNN)產生的資訊，作者善用1 x 1的convolution，讓資訊可以整合卻又可以保留各層的feature。

- 4-directional IRNN
  此方法最特別之處在於 4 個方向的 IRNN 。過去處理的 sequential 的資料，通常是和時間有關，例如連續的語音資料、文字、或是影片中連續的 frame，所以有 Bidirectional RNN 來利用前後文資訊。單張影像的預測則較少使用RNN來處理，通常透過 receptive field 的變化，可以在越後面的 layer 隱含 global 的資訊。而作者將 RNN 直接使用在單張影像上，同時配合 image 空間的特性，將圖片中各個pixels當作連續的資訊，將相鄰pixel視為連續的資訊，創造了這樣 4 個方向的 RNN。 CNN 能在後面的 layer 把整張圖的資訊整合，但是對於影像中的特定區塊(ROI)，整張影像其他部分，哪些區域對此 ROI 是更重要的，或許是個 learnable 的東西。所以當需要 focus 的區域在整張圖片中占很小部分時，我們不可能使用整張圖片的 feature，會使 focus 區域的 resolution 變很低，但是圖片中其他部分可能又對判斷這個區域是有幫助的，或許就能使用4-directional IRNN 來擷取影像中剩餘部分可能的重要 feature。

- **分工**
  周佳蓉:找資料、撰寫報告
  魏妤安:找資料、撰寫報告