**CNN-RNN: A Unified Framework for Multi-label Image Classification**

1. Introduction

這篇paper是Baidu Research和Facebook Speech等單位在CVPR2016發表的paper.

* 1. Problem

這篇paper要處理的問題就是題目提到的multi-label的問題. 在過去幾年因為CNN的成功,single-label image classification獲得了大大的成功,但是現實的世界裡並非一張image裡只包含一種類別,如Figure 1所示.



* 1. Challenges

處理這類問題有兩大挑戰,第一是label correlation,第二是semantic redundancy, label correlation像是一張圖中,如有船,那船跟海就有比較高的相關性,相對的和雲、天空的相關性就比較低,過去處理這類問題的方式是用Graphical Model,但是在high-order correlation的表現並不好.

Semantic redundancy是,例如cat跟kitty和kitten指的是同一個類別,過去處理這類問題是利用joint image/label embedding,但是此方法label co-occurrence的特性卻會大大的被忽略.

* 1. Framework

為了解決舊方法的限制,作者提出了一個架構,CNN-RNN A Unified Framework,這是一個end-to-end的架構,作者利用RNN去model label dependencies有high-order label relationship的問題,關於CNN的部份,image feature是利用RNN前一個預測的結果,encode attention models到CNN-RNN這個架構裡.Figure 2



因此此架構的RNN部份是去學到joint low-dimensional image-label embedding用來處理images和labels之間的semantic relevance,在CNN的部份,主要是去學image embedding vectors,因為每個label有自己的label embedding vector,所以此架構結合了joint image/label embedding跟label co-occurrence models的優點.

1. Proposed Method
   1. LSTM

RNN的部份就是LSTM,以下是LSTM的公式



f是forget gate, i是input gate, o是output gate, x是candidate gate, r(t)是candidate, 0(t)是output,其中δ(.)是activation function,⊙是product with gate value, U是要學的參數(weight).

* 1. Model

詳細架構圖請見Figure 4



先將Current label表示成one-hot vector的形式,再用一個label embedding matrix 將vector embed到一個較低的維度上,這邊有點將字詞embed到一個字典空間去得到latent語義的作用,像是word to vector(一種神經網路)也可以做到,得到公式(2).



所以與相關的值的變數會變成公式(3)



其中(.)跟(.)是公式(1)裡的non-linear functions.

再把recurrent layer和image representation project到相同low-dimension space,就像是剛剛公式(2)的label embedding,如公式(4)



其中跟分別是recurrent layer和image representation的projection matrix. 這邊有一個重點,即跟的column數要跟一樣多,原因如下.

最後要計算跟each label embedding之間的距離,公式如下



然後再利用softmax去計算scores.

* 1. Inference

RNN model預測multiple labels的path即是最大化priori probability,

如公式(6)



機率並不具備Markov property,這也是RNN(LSTM)的特性,即t的label未必會跟t+1有關係,但是卻可以決定要不要將它往下傳.

因此作者利用Beam search的方式來找top-ranked prediction path,如公式(7)



Beam search會在t找到top-N個最有可能的預測路徑,然後在t+1時利用top-N去預測t+1的路徑.

Beam search的優點是,它一次估計N個最好的路徑,僅從這N個路徑出發向下搜索,即每一層只有滿意的節點會被保留,其它節點則永久拋棄,能大大節省時間.

1. Experiments

CNN是使用16 layers的VGG network, pre-trained在ImageNet 2012上,實驗的dataset是NUS-WIDE, Microsoft COCO和VOC PASCAL 2007.

* 1. Evaluation Metric

C-P(per-class precision), O-P(overall precision), P-R(per-class recall scores), O-R(overall recall score), F1(average of the precision and recall scores) and MAP(mean average precision)

* 1. NUS-WIDE

有269,648 images(from Flickr),在移除noisy之後有1000 tags,之後又人工又更進一步去annotate到只有81 concepts.

Table 1跟Tabel 2分別呈現81 concepts和1000 tags的結果





* 1. Microsoft COCO

123 thousand images,利用object annotation當label.

結果請見Table 3.



在NUS-WIDE和Microsoft COCO這兩個dataset上,都可以發現P-R都不是很高,這是因為output可能低於k個labels,且較小的objects一般都不會去predict,所以和其它large objects的co-occurrence就比較小,而本篇就是利用RNN(LSTM)去解決co-occurrence的問題,如原本就沒被predict到,當然無法利用到RNN(LSTM)的特性.

Figure 7是MS-COCO dataset的per-class precision跟recall,可以發現到,很小的物體幾乎沒有被偵測出來,像是toaster和hair drier.



* 1. PASCAL VOC 2007

Table 4可以發現大部分的物體的結果都比之前的方法好



1. Conclusion

作者利用RNN(LSTM)學到前後文相關聯性的特性,在joint image/label embedding space中去model label co-occurrence dependency.