Machine Learning Project-2

Kuan-Lin Chen
Department of Applied Mathematics
National Chung Hsing University
Taichung, Taiwan
04131045@gm.scu.edu.tw

I. Introduction

從第一次報告中，我們先篩選了7篇較有興趣的paper，而在這次的報告中，我們會再從這裡面挑選兩篇出來實作。這次挑選的兩篇主要跟表情辨識相關，其中都會使用到VGG16這個非常深度的model當作pretrain的model，這裡兩篇paper使用的數據集分別是：CK+,CUB-200-2011，以下會介紹兩篇論文的實作細節。

II. Related Work

A. FaceNet2ExpNet: Regularizing a Deep Face Recognition Net for Expression Recognition -2017 CVPR

這篇論文的目標是，利用訓練好的人臉識別網路去訓練表情網路，最後做表情分類，分別是生氣、失望、害怕、傷心、開心、驚訝、鄙視、其他，在這裡我們會分兩個階段進行(如圖一)，我們先介紹大概要做甚麼下一章節我們在細講每個步驟應該做甚麼，在這我們的input image是1535 張大小為(224*224*3)，label 分別是我們預測出來(output label)的跟data 輸入的(input label)，output label主要是像(0.1,0.2,0.1,0.2,0,0,0.3,0.1) 的八維向量，而input label 主要是(0,0,0,0,0,0,1,0) 這樣的one hot vector，我們會用cross entropy去計算loss，因此在第一階段我們會將影像丟到VGG16 跟我們自己要訓練的emotion model，在這個階段我們將兩個conv layer的output 結果相減再取norm，目的是希望兩個conv.出來的結果要相似，接著第二階段，我們會在剛剛訓練好的emotion net後面接上一個1*1 conv去降低表情網路跟人臉網路之間的差距，接著為了避免overfitting，我們只接了一層的fully connect layer維度為256維，最後在softmax到8維向量。

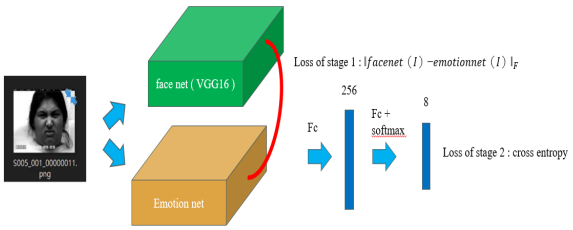

Fig. 1. First paper model

B. Learning a Discriminative Filter Bank within a CNN for Fine-grained Recognition -2018 ECCV

這篇論文想要解決的問題是希望能夠不需要額外end-to-end的cnn架構，而是透過他們設計的RA-CNN(細粒度圖像識別技術)，由較為細節的地方去做CNN進而將分類問題做得更為精確，在這我們使用的數據集是CUB-200-2011，且輸入的train image 與test image 分別有5994張與5794 張，大小則是448*448*3，label 則有200 種，這裡就不做額外說明，total epoch 是十萬次，所有的loss 則是cross entropy。這裡我們把圖片丟到VGG16裡面的前十層，將這個output 分別丟到兩個conv. layer 去，圖二中綠色區塊我們稱為P-Stream，我們會接一層conv. 然後fc + softmax 到200 維，圖二中紫色區塊我們稱為G-Stream，在這我們也會接一層conv.然後去計算這層的global maximum pooling(GMP)到pool6，這層的輸出我們會分別做兩件事，一、我們會直接fc+softmax 到200 維，二、我們會在這做cross channel pooling(也就是圖二中紅色區塊，我們稱為Side Branch)，最後softmax 到200 維，這三個會分別算一個loss，然後將這三個loss 乘上自己的權重由下而上分別是1.0,1.0,0.1 整合成一個total loss，在一次全部更新。

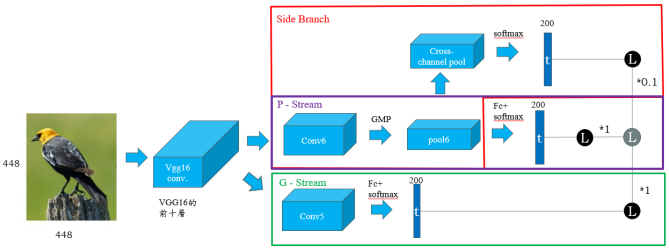

Fig. 2. second paper model

III. Detail for Model

在這裡我們會分別介紹兩篇論文裡面的實作細節，包含參數的設定使用的環境...等等。

A. FaceNet2ExpNet: Regularizing a Deep Face Recognition Net for Expression Recognition -2017 CVPR

前文有提到，這篇會有兩個階段，但是原VGG16這個model是沒有看過人臉的，原VGG16在訓練過程中丟的影像都是日常生活比較常見的物體，因此如果用這個模型直接訓練，會有不好的效果，所以在開始訓練之前我們

要先fine tune VGG16 這個pre-train model，如圖三所示，我們會將我們的資料丟到VGG16裡面然後接三層的fc 最後softmax 到8 維向量。

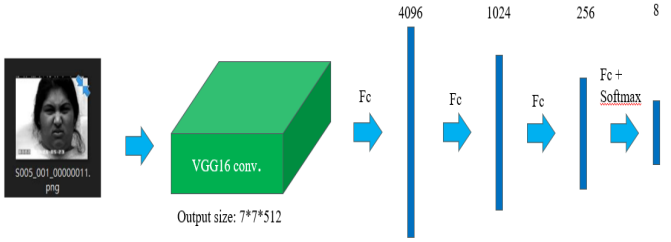

Fig. 3. FaceNet2ExpNet - stage1

接下來會將fine tune 好的VGG16 的這個模型當作我們訓練emotion net 的label，我們訓練的網路會有五層其channel分別是64,128,256,512,512每層的kernel size皆為3*3，stride為1，激勵函數為ReLu，loss function則是將這兩個conv. layer 相減取norm，total epoch為300，oprimizer為SGD，learning rate為1e-7在100 epoch後開始遞減，每次遞減1e-10。

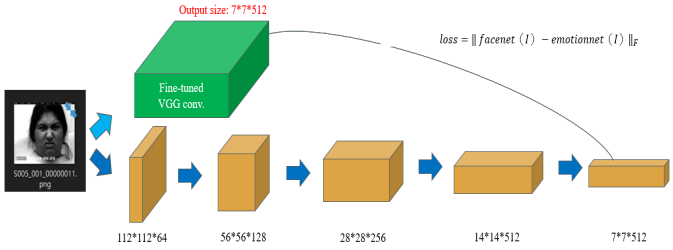

Fig. 4. FaceNet2ExpNet - stage2

上述部份訓練完後，我們會多接一層的1*1conv.，這是為了降低人臉神經網路與表情神經網路的差距，然後在接上一層fc維度為256最後fc+softmax到8維，loss function則用cross entropy，optimizer則是SGD，learning rate是0.0001，total epoch為50。

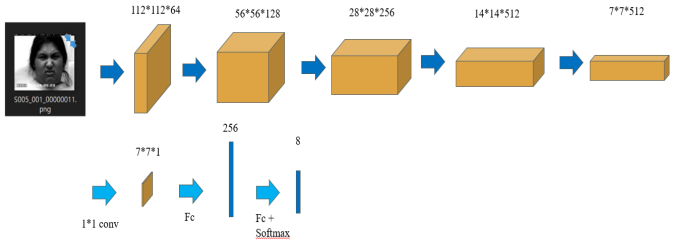

Fig. 5. FaceNet2ExpNet - stage3

B. Learning a Discriminative Filter Bank within a CNN for Fine-grained Recognition -2018 ECCV

接續第二章節說明P-Stream，在前十層的output of VGG16維度為56*56*256，在這裡我們接上一層的conv layer維度為56*56*200，然後接上一層fc最後softmax到200維，optimizer為adam，lerning rate為1e-4，激勵函數為ReLu，loss function為cross entropy。

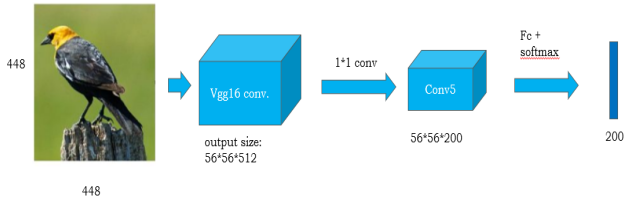

Fig. 6. Fine-grained model P-Stream

接著我們說明G-Stream，這裡我們接上一層的conv layer維度為56*56*(K*200)，然後做GMP到K*200(也就是pool6那個區塊的維度)，這裡的K 是因為要在後面多做cross channel pooling 才會把維度拉到K*200(稍後說明K為何)，然後接上一層fc最後softmax到200維，optimizer 為adam，lerning rate為1e-4，激勵函數為ReLu，loss function為cross entropy。

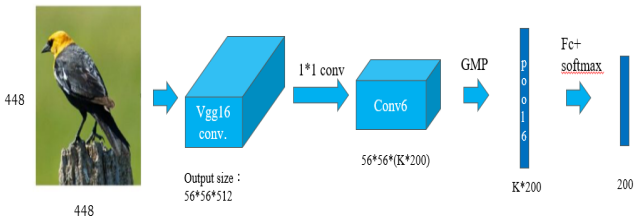

Fig. 7. Fine-grained model G-stream

最後我們說明Side Branch，由output of pool6做以K為一組的average pooling(也就是我們一直提到cross channel pooling)，我們假設K=10，當我們做完cross channek pooling後(此時維度應為200)，最後接上softmax層，optimizer為adam，lerning rate為1e-4，激勵函數為ReLu，loss function為cross entropy。

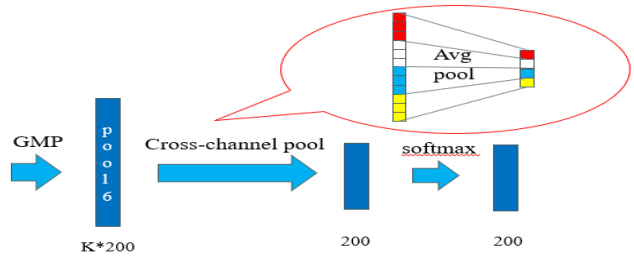

Fig. 8. Fine-grained model Side Branch