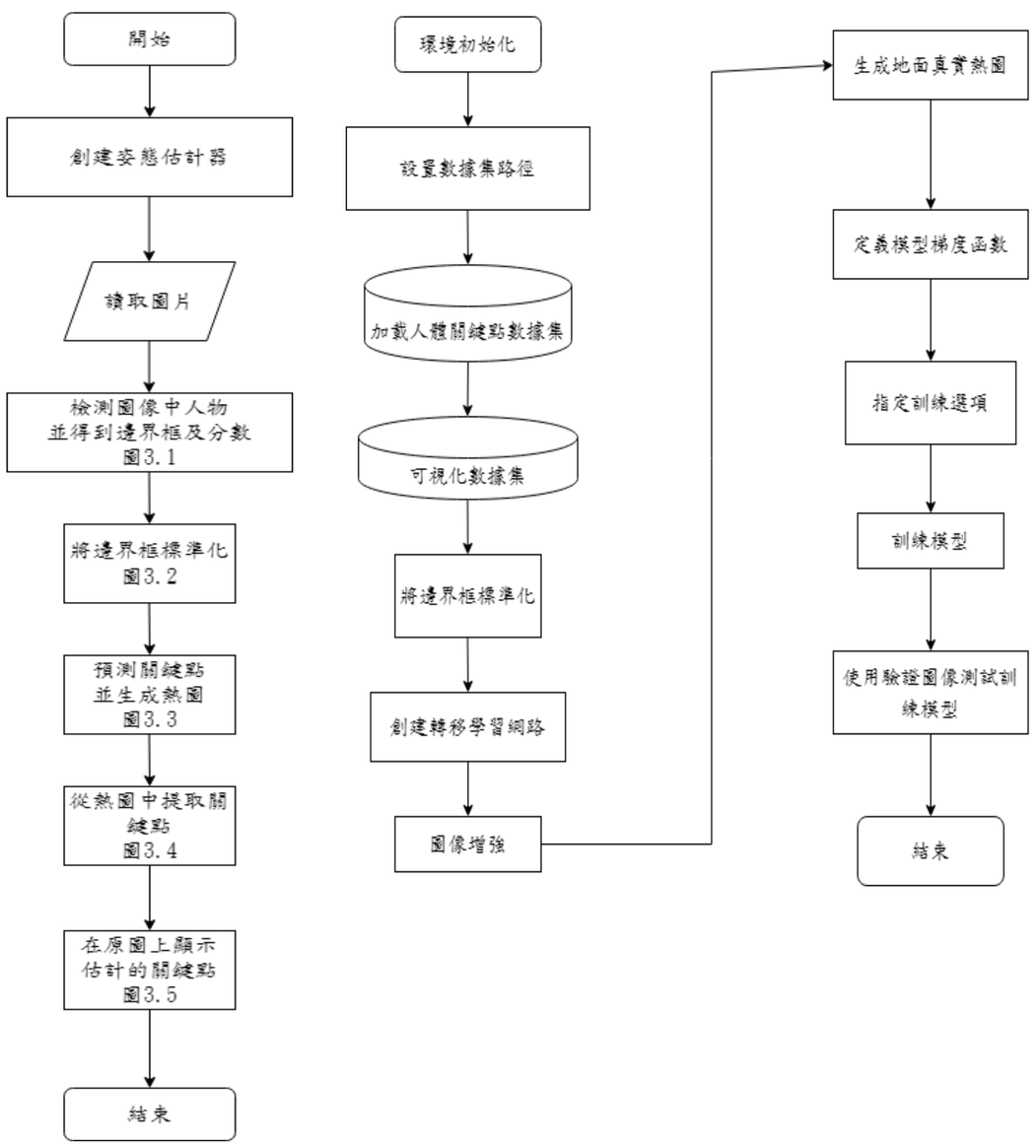
利用深度學習實現人體姿勢之估計研究
Human Pose Estimation with Deep Learning

指導教授:余松年 教授　　專題生:張庭瑋 江佳峰

摘要

這個專案的主要目標是透過深度學習技術，實現對照片中人物姿勢骨架的精確估計。我們首先從照片中精確地剪切出人物的圖像，再透過ResNet模型識別出人物的關鍵節點，並將這些節點連線，生成一個完整的人物姿勢估計圖像。我們首先定義了一些參數，例如COCO數據集的路徑，並讀取和分析COCO數據集的JSON註釋文件，提取訓練和驗證圖像的路徑，以及對應的關鍵點和邊界框信息。接著創建了一個基於ResNet18的深度學習網絡，並對其進行了一些修改，以適應人體姿勢估計的任務。在進行數據增強操作後，我們使用ADAM優化器和均方誤差損失函數來訓練網絡。訓練完成後，我們使用驗證數據集來評估模型的性能，並將預測結果視覺化出來。最後，訓練好的模型和骨架連接圖會被保存到一個MAT文件中，以便後續使用。

架構介紹



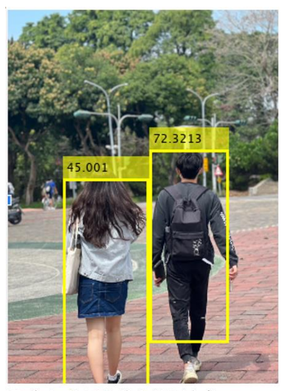


圖 3.1 檢測人像和檢測分數

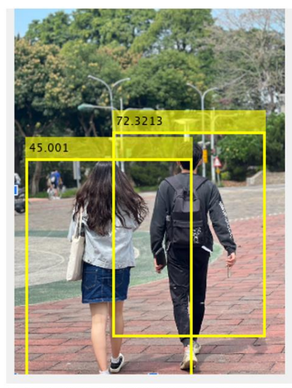


圖 3.2 將邊界框調整大小，使其在網絡輸入中具有相同的長寬比。

圖3.3 關鍵熱圖

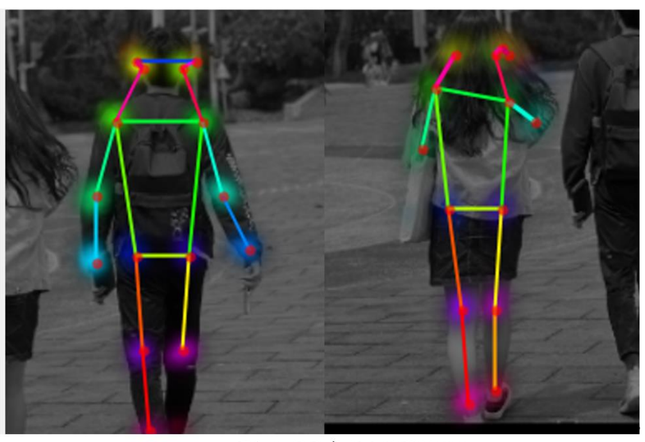


圖3.4 提取每個人的關鍵點

圖3.5 在原始圖像的坐標中估計的關鍵點

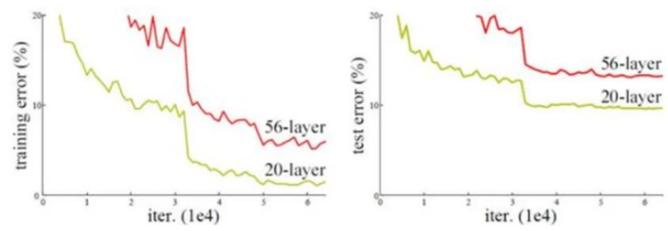
深度學習網路
ResNet 18 (Residual Network)

➢ 什麼是殘差(residual)?

簡單來說就是誤差的觀測值，例如，想找一個x，使得使得 f(x)=b，我們估計x=x’ 殘差（residual）就是b-f(x’)。誤差就是x-x’ 。

➢ 什麼是退化?

理論上Resnet深度神經網路越多層越可以提取更好的圖片特徵，但事實卻相反



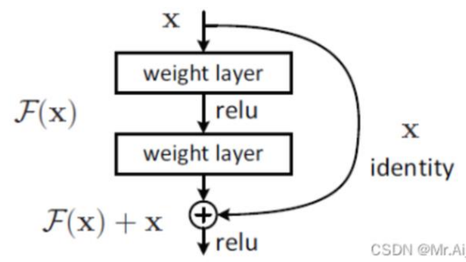
在CIFAR-10數據集上，使用20層和56層一般網路進行訓練的錯誤率（左）和測試的錯誤率（右）。

從圖片可以看到當我們直接增加層數時，錯誤率反而更高，這並不是過度擬和的問題，這個現象被稱作為退化(Degradation)。ResNet團隊把退化現象歸因為深度神經網路難以實現「恆等變換（y=x）」隨著層數的增加，引入的激活函數也越來越多，數據被映射到更加離散的空間，此時難以實現恆等變換(讓數據回到原點)，於是有了在網路中增加線性轉換分支的方法。

➢ 捷徑連接（shortcut connections）

ResNet的核心實是通過添加額外的連接來解決深度神經網絡訓練中的梯度消失和梯度爆炸等問題，從而允許建構非常深的神經網絡。ResNet通過引入所謂的"捷徑連接"，允許某一層的輸出直接跳過一個或多個層，連接到後續層的輸入。這樣做的好處是，即使某些層不做任何有意義的變換，它們仍然可以傳遞之前層的信息，而不會對梯度產生過多的損失。這可以用一個公式來表示：

$$H(x) = F(x) + x$$



結論

本專題利用coco data資料集來自訓練深度神經網路，如用原來的預訓練模型，在一些比較昏暗的場景，很容易偵測錯誤，以及過度敏感。此次專題所使用的網路(Simple Baseline)為改變過後的ResNet網路，結構較ResNet簡單，所輸出的feature map的分辨率也高。

在一開始使用預訓練模型時，當框選人物時可能會框選到背景中的樹以及背景中模糊的人影(圖3.6)，這顯然不符合我們的需求。為了改善這種情況，我們決定訓練自己的模型，以提高人體姿勢估計的精確度經過訓練後的模型，我們再次使用同一張圖片進行檢測，結果顯示更為精確且符合我們的期望(圖3.7)。

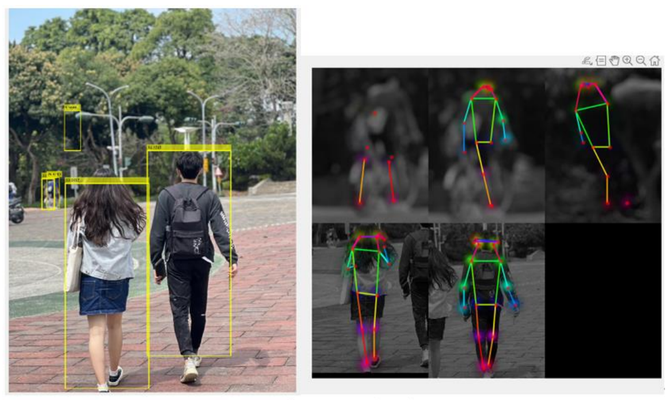


圖3.6 預訓練網路模型所產生的結果圖

圖3.7 自訓練深度神經網路模型所產生的結果圖