Midterm project

STuser19 賴昱凱

1. 模型架構
2. 模型架構概述

本期中專題我參考“Accurate Hand Keypoint Localization on Mobile Devices”（後簡稱甲論文）中的模型架構，不過該篇論文也是基於“VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera”（後簡稱乙論文）之模型，並以創建可於行動裝置上執行之模型為目的，大幅減少乙論文計算量及參數量所創照之模型架構。而我基於甲論文之模型做更進一步的縮減規模，同時也參考乙論文之transposed convolution架構（於甲論文中無詳細說明），做出在模型大小及表現上有優異平衡的模型。

乙論文之模型架構於ResNet50後接上一連串的convolution layers做後續處理（圖一），但其模型規模龐大，無法於行動裝置做即時影像處理，因此甲論文將原Resnet50利用MobileNetV2代替，並將mobilenetv2的第十四層stride由（2, 2）改為（1, 1），使特徵輸出由原先的1080\*7\*7改為1080\*14\*14，以避免過小的資料量造成特徵得提取困難，同時為了在後續可以繼續對特徵做處理，甲論文不使用mobilenetv2之linear layers，僅使用其feature maps的輸出1080 \* 14 \* 14，並將乙論文原先後續處理的convolution layers更換使用depth-wise convolution進一步減少參數及計算量（圖二）。

一張含有 文字, 圖表, 字型, 方案 的圖片

自動產生的描述

圖一、”VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera” 模型架構圖

一張含有 螢幕擷取畫面, 行, 圖表, 繪圖 的圖片

自動產生的描述

圖二、”Accurate Hand Keypoint Localization on Mobile Devices" 模型架構圖

1. MobileNet
   1. MobileNetV2

MobileNetV2之中最重要的概念就是inverted residual block。在大部分常見Efficient CNN中，都會使用Depthwise Separable Convolution來減少參數量，包括乙論文所使用的ResNet50。然而在甲論文中所使用的MobileNetV2中，其利用了inverted residual blocks技巧來達成更少的參數及計算量。

簡單來說，residual blocks是藉由減少計算時的channel數量來達到計算量及參數量的減少，因此利用1x1的convolution layers在兩端做channel數量的調整，兩端連接的是channel數較多的expansion layers做資料傳遞。然而inverted residual blocks在做完convolution layers後的資料傳遞是使用linear bottleneck layers，是將channel數減少而不是增加，且因為“manifold of interest”在降至低維度時不會有大多的資料損失，只會在經過non-linear function時才會，因此只要不使用relu等activation function就可以完整保留特徵資訊同時減少參數量及計算量，convolution layers前的inear bottleneck layers利用1x1 convolution後也是增加channel而非減少。

由於前面提到linear bottleneck layers並不會遺失資訊，因此在MobileNetV2中是對非expansion layers做shortcut，與一般residual blocks相反，因此稱為inverted residual blocks。

一張含有 螢幕擷取畫面, 文字, 圖表, 字型 的圖片

自動產生的描述

圖三、Residual / Inverted residual block架構圖

* 1. MobileNetV3

而在MobileNetV3中，使用了SENet (Squeeze and Excitation)架構，透過 Global Average Pooling計算每個 feature map 的權重，用來強化重要的 feature map 並減弱不重要的 feature map 。且在SE block中使用 Hard-sigmoid 代替 sigmoid 以實現更高效率的計算。

SE block首先先利用Squeeze function（Fsq）對每個channel（C）分別作全域平均（Global Average Pooling），得到一個1\*1\*C的資料，再利用Excitation function（Fex）做兩層fully connection layers，分別使用Relu及Hard-sigmoid做activation function，此時的輸出即為該channel的權重，並與對應的channel相乘即可達到強化重要的 feature map 並減弱不重要的 feature map 的效果。

一張含有 圖表, 行, 方案, 繪圖 的圖片

自動產生的描述

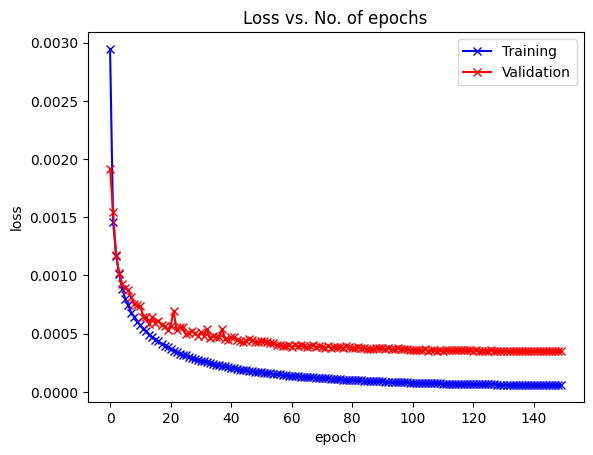
圖四、Squeeze and Excitation block架構圖

1. 專題使用之模型架構

在研讀甲論文後，因為專題評分方式會將模型規模納入考慮，因此我決定以甲論文之模型為雛形，創建一個計算量及參數量皆小的模型。在多次實驗後，我得到以下結論。

* 1. 雖然模型channel數量與準確度有正相關，但考慮到ranking formula後，將channel數量大幅減少有利於分數計算，且損失的準確度並無影響Kaggle grade太多。
  2. 使用MobileNetV2的準確度比MobileNetV3\_small高，但在同樣的前後期處理上，MobileNetV3\_small的規模（計算量\*參數量）為MobileNetV2的0.27倍，MobileNetV2需在準確度上有約1.3 pixel difference的領先才有優勢，實驗後發現V2並無如此優勢，因此使用V3在最終模型中，實驗結果如下(非最終結果)。

一張含有 文字, 螢幕擷取畫面, 繪圖, 行 的圖片

自動產生的描述 

FLOPS: 1.031351216 G

Params: 4.152448 M

avg pixeldiff: 5.108583596827928

avg loss: 0.0007516901241615415

ranking score: 708.5009779051727

圖五、使用MobileNetV2之loss曲線 圖六、使用MobileNetV3之loss曲線

FLOPS: 0.474115488 G

Params: 2.585248 M

avg pixeldiff: 5.965919247358571

avg loss: 0.0011419530492275953

ranking score: 477.91665120700634

最終模型如下：

FLOPS: 0.302601376 G

Params: 2.007968 M.

一張含有 圖表, 方案, 螢幕擷取畫面, 設計 的圖片

自動產生的描述

圖七、專題模型架構圖

1. Training skill
2. 資料前處理

訓練資料數量與準確率有正相關，我將原先30000筆的training data依靠旋轉、鏡像及縮放增加資料量。但實驗後發現縮放對結果為負面影響，因此只使用了隨機旋轉以及隨機水平鏡像處理，同時也發現過多使用前處理新增的資料也會對結果有負面影響，因此最終使用7倍的訓練資料。

一張含有 螢幕擷取畫面, 鮮豔, 藝術, 設計 的圖片

自動產生的描述

圖八、資料前處理結果示意圖

1. Optimizer

於甲論文中所使用的Optimizer是Adadelta，然而我在嘗試過後表現並沒有比Adam好，因此最終我使用的是Adam，且參數皆為預設。

1. Scheduler

我一開始使用warn up的技巧來訓練模型，但後來發現並沒有太大的幫助，因此之後使用“CosineAnnealingLR”，由初始learning rate：0.0005至最小的0.00005，T\_max為（config["rot\_num"] \* (len(train\_dl.dataset) / config["batch\_size"]) \* config["n\_epochs"]），就是半個完整的cos波形。

1. 訓練結果
2. Loss

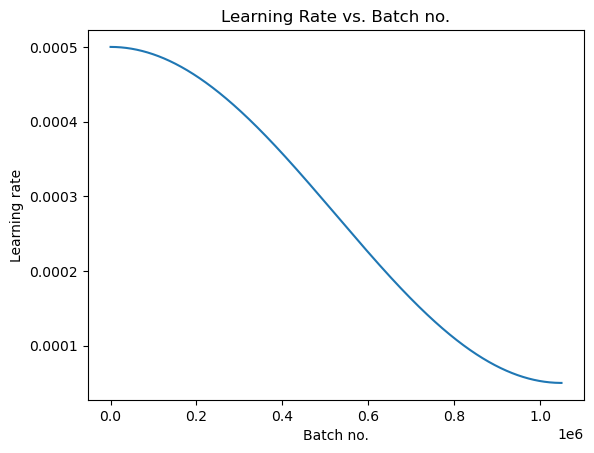
一張含有 文字, 螢幕擷取畫面, 繪圖, 行 的圖片

自動產生的描述

圖九、loss vs. epochs

可以明顯的發現training loss與validation loss是有相同的變化趨勢，然而training loss皆會穩定地低於validation loss，這是因為model的參數調整是基於training data與該model的差距做gradient descent，然而validation data僅用來判斷這次model調整的好壞，並不會直接與model計算有關。因此對於model來說，計算上本身就是為了迎合training data，有這樣的結果也十分合理。基於同個理由，testing loss也會比validation loss更高，因為model是選擇使用對於validation data有最好表現的參數，與testing data毫無相關。

1. Scheduler



圖十、learning rate vs.batch number

1. Result

FLOPS: 0.302601376 G

Params: 2.007968 M.

avg pixeldiff: 6.152803876982536

avg loss: 0.00122355786152184

ranking score: 285.59885234255654

1. Reference
2. Filippos Gouidis et al. "Accurate Hand Keypoint Localization on Mobile Devices." Proceedings of the 16th International Conference on Machine Vision Applications (MVA), National Olympics Memorial Youth Center, Tokyo, Japan, May 27-31, 2019.
3. Dushyant Mehta et al. "VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera." ACM Transactions on Graphics (TOG), SIGGRAPH 2017.
4. 速讀論文 Accurate Hand Keypoint Localization on Mobile Devices

<https://allen108108.github.io/blog/2020/01/11/%5B%E8%AB%96%E6%96%87%5D%20%E9%80%9F%E8%AE%80%E8%AB%96%E6%96%87%20Accurate%20Hand%20Keypoint%20Localization%20on%20Mobile%20Devices/>

1. 手部關鍵點 (Hand Keypoints) 的預測及其應用

<https://allen108108.github.io/blog/2020/06/06/%E6%89%8B%E9%83%A8%E9%97%9C%E9%8D%B5%E9%BB%9E%20(Hand%20Keypoints)%20%E7%9A%84%E9%A0%90%E6%B8%AC%E5%8F%8A%E5%85%B6%E6%87%89%E7%94%A8/>

1. Efficient CNN 介紹(二)：MobilenetV2

<https://medium.com/ai-academy-taiwan/efficient-cnn-%E4%BB%8B%E7%B4%B9-%E4%BA%8C-mobilenetv2-7809721f0bc8>

1. [論文筆記] MobileNet演變史-從MobileNetV1到MobileNetV3

<https://chihangchen.medium.com/%E8%AB%96%E6%96%87%E7%AD%86%E8%A8%98-mobilenetv3%E6%BC%94%E8%AE%8A%E5%8F%B2-f5de728725bc>