National Tsing Hua University
Fall 2023 11210IPT 553000
Deep Learning in Biomedical Optical Imaging
Report

AUTHOR ONE[1]

謝瑋哲 清華大學動機系 碩士一年級 新竹 台灣

Student ID: 112033645

1. 介紹

本報告是基於清華大學光電所陳鴻文教授開設的深度學習於生醫光學影像之應用課程，這次的訓練資料是六種癌症組織細胞的分類，總資料近 4000 張的彩色醫學影像，我這次的報告主要比對的使用了在課程資料中的 Vit 模型，為了評估此模型的優劣，也另外對比了一個簡單的 CNN 模型以及更多轉移學習的有名模型。

2. 實驗

2.1. 資料

資料主要有六類:腫瘤、基質、複合體、淋巴、碎片、黏膜，共 3750 張照片，分為:訓練 2550 張、驗證 600 張、測試 600 張。

2.2. 模型

模型分為兩個部分:自建模型與轉移學習的有名模型，自建模型皆來自之前課堂的部分，而轉移學習模型則是來自 pytorch 模組的預設模型，初始權重也都是使用各模型的預設權重，總共訓練 30 個 epoch，而比較模型淺層模型三個以及深層模型三個最為對比。

2.2.1. Vit(Vision Transformer)

模型架構: 總共有 47 層，ViT 模型包含 9 個 Transformer 層，而 Transformer 由 Multi-Head Self-Attention 模組 Feed Forward Network 模組組合而成，第一層為一個補丁輸入層，最後再接上一個全連接層。

2.2.2. ConvModel

模型架構: 總共有 5 層，由 3 個卷積層接上 2 個全連接層組成。

2.2.3. resnet18

模型架構: 總共有 18 層，其中是 17 個卷積與 1 層的全連接層，架構為兩個卷積層構成 Residual Block，兩個基本塊構成的殘差層 Resnet 塊。順序為卷積層、四個連續的 Resnet 塊，最後再接上一個全連接層。

2.2.4. AlexNet

模型架構: 總共 8 層，5 個卷積層和 3 個全連接層。前兩個卷積層後接最大池化層，第五個卷積層後接平均池化層，最後三個全連接層中的前兩層後接 Dropout 層。AlexNet 使用 ReLU 作為激活函數，並在第一層和第二層卷積層後使用局部響應歸一化（Local Response Normalization）。

2.2.5. googlenet

模型架構: GoogleNet 總共約有 22 層，每個 Inception 模組包括多個並行的捲積層和池化層。

2.2.6. ResNet50

模型架構:總共 50 層，由更多的殘差層基本塊構成。這些基本塊包含 3 個卷積層（不同於 ResNet18 的 2 個卷積層基本塊）。最後是全局平均池化層和全連接層。

2.2.7. EfficientNetB4

模型架構:總共大概 50 層，架構使用了深度可分卷積和倒置殘差結構。包括多個倒置殘差塊，每個塊都具有一系列的卷積、激活和批量正規化操作。最後次全局平均池化層和全連接層。

2.2.8. DenseNet121

模型架構:共有 121 層，由多個密集塊（Dense Block）和過渡層（Transition Layer）組成。每個密集塊內的層都與之前所有層直接相連。最後以全球平均池化層和一個全連接層結束。

3. 實驗結果與討論

3.1. 訓練結果與分析

3.1.1. 總體表現

在表 1 的所有訓練結果中，訓練效果最好的是 DenseNet121，最差的為 ConvModel，雖然 ConvModel 毫不意外的是最差的，但他的模型是極其簡單 5 層，也就是說在本組影像分類問題之上，隨意挑選模型即可達到一定程度上的分類，因為隨意亂猜僅會有 16.7%的準確率，所有結果皆在 70%之上。

3.1.2. Vit 模型不準結果討論

首先討論自建模型 Vit，對比 Vit 的訓練效果，而無論是何者轉移學習比較都比 Vit 的表現良好，甚至僅只比簡單設置的 ConvModel 表現好上一點點而已，我推測可能原因有三個:模型補丁個數、預訓練差別以及圖片性質。

1. 模型補丁個數: 在我所使用的 Vit 中為了能夠整除影像大小(150*150)，我將補丁設定為 10 與 15 作為對比，但是 10 和 15 除了一點準確度的差別僅只有訓練速度的差異，所以或許需要使用原先設定的 16 或是常用的 32，而如需改為使用 16 或 32，需要直接動到訓練的圖片，將大小變為 16 或 32 的倍數。
2. 預訓練差別: 在轉移訓練模型中，我皆是使用了 IMAGENET1K_V1 作為初始的預訓練，或許 IMAGENET1K_V1 的權重與本資料及的相容性佳，導致所有使用了 IMAGENET1K_V1 的模型最差也有 80%
3. 圖片性質: 在圖片集中，我們所有訓練資料皆是挑選六種分類的照片，而每一種照片大多都是完整表明清楚，這或許也使 Vit 的 attention 部分沒有特別的表現空間，因為資料及已經對圖片進行了一次人工的 attention 了。

3.1.3. 轉移訓練模型表現

在轉移模型中，我以 30 層為分隔點把分為兩組: 低層組與深層組，我們可以發現，在低層組中，層數與訓練時間是等比例的，層數越深，訓練得越久，而在高層組中，層數越深就不完全與時間成等比關係了，這可能是因為在某些深層模型中，透過參數共享和重複使用技術，即使層數增加，模型的參數總數和計算複雜度可能不會線性成長。這有助於在保持較高層數的同時，控制訓練時間和計算需求。，比較兩組結果，我們可以得出如果隨意選擇越深層的模型，更有可能訓練的比低層數模型訓練的好。

而如果加上所花上的時間而言，高層數的模型並未取得飛躍性的增長，甚至在選擇 Resnet 的模型架構之上，更深層的 Resnet50 效果竟不比 Resnet18 好，或許是因為更深的 Resnet 模型需要更多的資料來充分訓練。

3.2. 結論

總的來說，在本次實驗之中，比起自己建立模型，如果目標僅是常見的影像分類，或是對於訓練資料集有特別的理解之外，直接選擇有名的模型做轉移學習，可能會比自己從頭建起還要優秀，更何況許多模型還能夠使用預訓練的權重，而在表一中，模型選擇並非皆是越深表現越好，而在圖一中，我們可以由訓練時間與準確率看到個模型的效率，由此可以看出性價比最高的模型是 ConvModel、AlexNet 與 ResNet18，而其中，ResNet18 是可以用較低的資源得到模型中，準確率最高的。

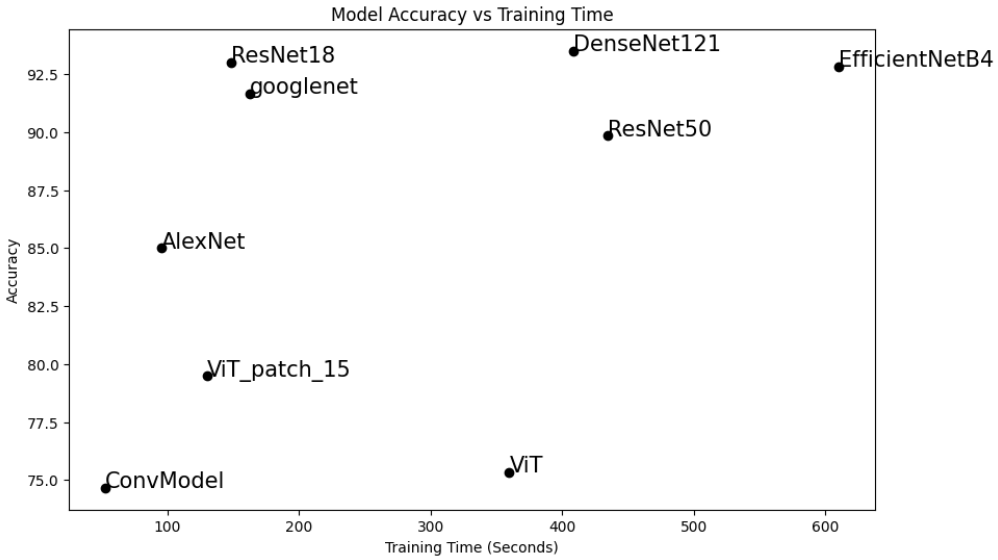
4. 實驗圖表

4.1. 結果表

Name	*test accuracy*	*test_avg_loss*	*Time cost*	*accuracy_per_second*	*loss_per_second*	*layer*
ViT_patch10	75.333%	0.683	359.7 秒	0.20943 %	0.0019	47
ViT_patch15	79.5%	0.633	129.9 秒	0.61201 %	0.00487	47
ConvModel	74.667	0.709	52.8 秒	1.41414 %	0.01342	5
ResNet18	93.0%	0.307	148.5 秒	0.62626 %	0.00207	18
AlexNet	85.0%	0.543	95.5 秒	0.89005 %	0.00568	8
googlenet	91.667%	0.276	162.2 秒	0.56515 %	0.0017	22
DenseNet121	93.5%	0.289	408.5 秒	0.22889 %	0.00071	121
ResNet50	89.833%	0.276	434.3 秒	0.20685 %	0.00063	50
efficientnet_b4	92.833%	0.265	609.8 秒	0.15224 %	0.00043	50

表 1. 實驗結果

4.2. 結果圖



圖一、各模型訓練時間與準確率