# 機器學習第二次報告
# Music genre classification

陳奕傑
大數據產學研發班
國立中興大學

## I. Music genre classification using machine learning techniques

### A. 簡介

此篇論文來自 BAHULEYAN, Hareesh(2018) 利用了卷積神經網路 (Convolutional neural network, CNN) 等方法去將歌曲做分類，因 CNN 應用於圖像辨識的技術已經相當成熟。音樂的聲波可以表示為頻譜圖，而頻譜圖又可以視為圖像。故在這理利用 CNN 去處理曲風的分類也應該是合適的。在這裡 CNN 的架構為 VGG-16，擁有 13 個卷積層，3 個全連接層並搭配遷移學習 (Transfer learning) 與微調 (Fine tuning) 去應用於曲風的分類。

### B. 使用的方法:*VGG-16*

VGG 是由 Simonyan 和 Zisserman 在 2014 文獻《Very Deep Convolutional Networks for Large Scale Image Recognition》中提出的卷積神經網絡模型，其名稱來源於作者所在的牛津大學視覺幾何組 (Visual Geometry Group) 的縮寫。從 Fig.1 可知，VGG-16 可以由捲積層與池化層劃分為不同的塊（Block），從前到後依次編號為 Block1 至 Block5。每一個 Block 內包含若干卷積層和一個池化層。

例如：

Block2 中包含 2 個卷積層，1 個池化層 (maxpool)，每個卷積層用 conv3-128 表示, 即卷積核為：3×3，通道數都是 128，激活函數為 ReLU。

Block3 中包含 3 個卷積層，1 個池化層 (maxpool)，每個卷積層用 conv3-256 表示, 即卷積核為：3×3，通道數都是 256，激活函數為 ReLU。

比起 AlexNet VGG-16 採用連續幾個 3x3 的卷積核代替 AlexNet 較大的卷積核 (11x11,7x7,5x5) 如 Fig.2., 在同樣的感受野中，採用堆積的小卷積核會優於採用較大的卷積核。因為採用堆積的小卷積核會比起較大的卷積核有更深的網路深度，並且能減少參數。

例如: $C$ 為通道數，$A$ 及 $B$ 各為不同的參數數量

$$\begin{aligned} A &= 2 \times (3 \times 3 \times C^2) = 18C^2 \\ B &= (5 \times 5 \times C^2) = 25C^2 \\ A &< B \end{aligned} \tag{1}$$

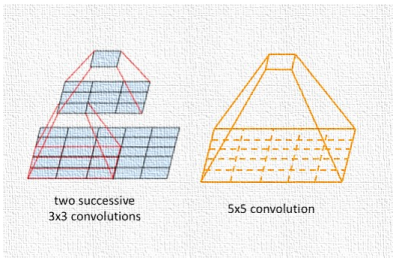



Fig. 2. 3*3 convolutions and 5*5 convolutions

在解決問題的時候，不用從零開始訓練一個新模型。可以從在類似問題中訓練過的模型入手，我們利用了遷移學習與微調。在應用兩種方法之前, 其中兩者區別為在遷移學習中, 我們僅優化了以添加的新分類層的權重, 而保留原始 VGG-16 模型的權重。在微調中，我們優化了以添加的新分類層的權重，也優化了 VGG-16 模型中的部分或全部層的權重。

如 Fig.3.，在另一個數據集上使用 VGG-16 模型時，我們可能必須替換所有 dense layers。為了避免 overfiting 我們添加了另一個 dense layers 和一個 dropout-layer 。並且在最後一層的 softmax 激活函數設定為七種分類各輸出的機率。

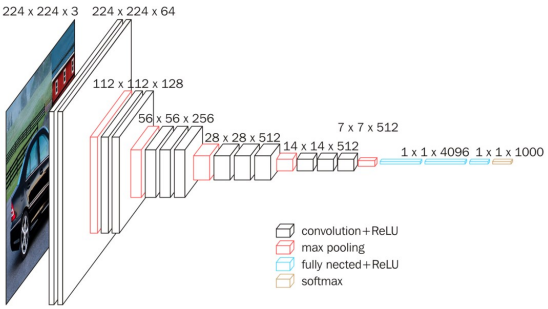



Fig. 1. VGG-16

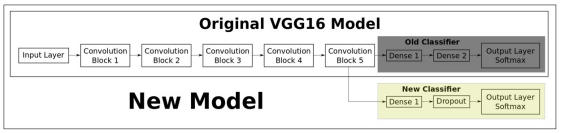



Fig. 3. Convolutional neural network architecture

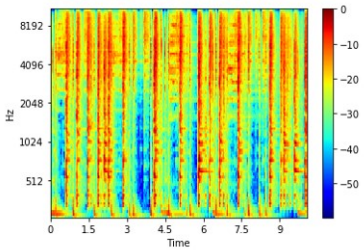



Fig. 4. MEL spectrogram

### C. *Dataset* 與實作

第一步我們利用了 AudioSet:Gemmeke.ate 2017 網站中提供的 csv 檔,此資料檔包含了他們標籤好的 youtube 影片 id 以及其標籤的類型。我們從中選了七種音樂類型，分別為 Pop Music、Rock Music、Hip Hop Music、Techno、Rhythm Blues、Vocal 以及 Reggae Music。第二步透過 youtube-dl(Gonzalez,2006) 與 AudioSet 的 youtube Id 去下載所需的 mp4 並透過 ffmpeg(Tomar,2006) 將 mp4 檔轉成較小的 wav 檔。最後透過短時距傅立葉變換將 wav 轉成 MEL spectrogram, 如 Fig.4.。

轉成 MEL spectrogram 後就可以帶入 VGG-16，首先將圖片轉成為 216*216，但原 VGG-16 是使用 224*224。對於此篇論文作者利用 216*216 可能需要在實作中進行驗證。將資料切割為 train (90%),validation (5%) and test (5%) sets 後，將資料導入 VGG-16 模型，並採用已建立好的 ImageNet 模型參數。根據 transfer learning 以及 fine tuning 兩種不同的方式設置 VGG-16 模型。每個模型都做 10 次的 epochs，每次的 epochs 中其 batch size 為 32，並採用 Adam Optimizer 作為優化器。

## II. Music Genre classification using a hierarchical Long Short TermMemory (LSTM) model

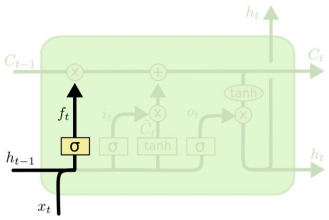
### A. 簡介

此篇論文來自 TANG, Chun Pui, et al.(2018) 本文探討了長短期記憶網路 (Long Short Term Memory Network, LSTM) 模型如何應用在音樂曲風分類中，在一開始從音樂裡提取其音頻特徵，即為 Mel-frequency cepstral coefficients(MFCC)。利用兩種方法進行分類比較, 第一種只利用 LSTM 進行分類，第二種利用分層分類加上 LSTM 對於曲風做分類，第一層為音樂的強弱分成兩類，下一層再將其分層兩類，最後一層則是按照不同的曲風分類，最後再比較哪種方法達到的效果會比較好。

### B. 使用的方法:*LSTM*

Long Short-Term Memory，LSTM 是一種時間遞歸神經網絡（RNN），論文首次發表由 S Hochreiter (1997)。由於獨特的設計結構，LSTM 適合於處理和預測時間序列中間隔和延遲非常長的重要事件。

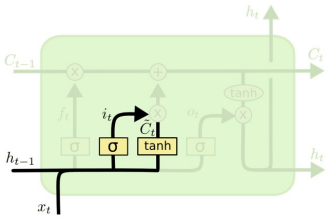
第一步為忘記層將 sigmoid 函式的輸出值直接決定了狀態資訊保留。將先前隱藏的訊息以及當前輸入的信息同時丟入 Sigmoid，輸出值處於 0 和 1 之間，越接近 0 意味著越應該忘記，越接近 1 意味著越應該保留。




$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

Fig. 5. Step.1 LSTM

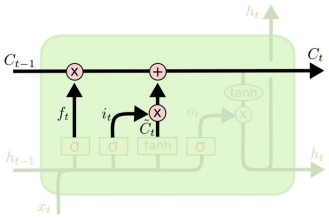
第二步將隱藏狀態和當前輸入傳輸給 Tanh 函數，並在-1 和 1 之間壓縮數值以調節網絡，說明細胞狀態在某些維度上需要加強，在某些維度上需要減弱。然後把 Tanh 輸出和 Sigmoid 輸出相乘，Sigmoid 輸出將決定在 Tanh 輸出中哪些信息是重要的且需要進行保留。




$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

Fig. 6. Step.2 LSTM

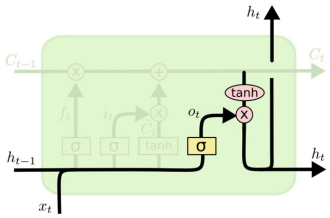
第三步把先前的單元狀態和遺忘向量逐點相乘，如果它乘以接近 0 的值，則意味在新的單元狀態中可能要丟棄這些值；然後把它和輸入門的輸出值逐點相加，把神經網絡發現的新信息更新到單元狀態中，這樣就得到了新的單元狀態。




$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Fig. 7. Step.3 LSTM

輸出門能決定下個隱藏狀態的值，隱藏狀態中包含了先前輸入的相關信息。首先把先前的隱藏狀態和當前輸入傳遞給 Sigmoid 函數；接著把新得到的單元狀態傳遞給 Tanh 函數；然後把 Tanh 輸出和 Sigmoid 輸出相乘，以確定隱藏狀態應攜帶的信息；最後把隱藏狀態作為當前單元輸出，把新的單元狀態和新的隱藏狀態傳輸給下個時間。




$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

Fig. 8. Step.4 LSTM

### C. *Dataset* 與實作

採用 Gtzan music dataset, 數據集包含 1000 個音軌, 每個音軌長 30 秒。包含 10 個流派, 每個流派由 100 首曲目代表。曲目皆為 22050Hz 單聲道 16-bit 音頻.wav 格式文件。利用 Librosa python Library 提取音頻特徵, 即 MFCC features。

實驗 1 將音樂分成六類,classic, hip-hop, jazz, metal, pop and reggae。將 audio tracks 分為 420 個訓練資料, 120 個驗證資料以及 60 個測試資料。

音框大小為 25 毫秒, 每個 30 秒的 soundtrack 有 1293 個音框以及 13 個 MFCC 特徵 $C_1, ..., C_{13}$。

在 Input Layer 將 13 個 MFCC 特徵輸入, 再接上兩個分別有 128 及 32 neurons 的 Hidden Layer。最後的 Output Layer 為六個輸出對應六個不同曲風的音樂, 在這裡取 5, 10, 20, 50, 100, 200, 400 Epochs 並取 Adam 作為優化器, 並且對於每個 case 運行 4 次。

在實驗 2 的部分, 如 Fig.9. 用人工的方式對於 input 至 LSTM 的資料作分層以 LSTM1 將音樂分成強音樂以及溫和音樂, LSTM2a 將音樂曲風分為 Sub-strong1 (hiphop, metal and rock) 以及 Sub-strong2 (pop and reggae) 。最後由 LSTM3a,LSTM3b,LSTM3c,LSTM3d 將 10 種不同曲風的音樂分類出來。其中每個 LSTM 參照的模式皆與第一種實驗一樣。




Fig. 9. Tree diagram of our approach

## References


[1] BAHULEYAN, Hareesh. Music genre classification using machine learning techniques. arXiv preprint arXiv:1804.01149, 2018.

[2] TANG, Chun Pui, et al. Music genre classification using a hierarchical long short term memory (LSTM) model. In: Third International Workshop on Pattern Recognition. International Society for Optics and Photonics, 2018. p. 108281B.

[3] SIMONYAN, Karen; ZISSERMAN, Andrew. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.