

MLDS Homework 1 Report

B05902127 資工二 劉俊緯
B05902013 資工二 吳宗翰

1 Deep vs Shallow

在作業 1-1 中，我們將比較 deep 和 shallow 兩個 network 架構在參數量相同的時候的表現及其收斂過程。以下我們將做兩個實驗：實驗一是試著去 fit 一個函數、實驗二則是去做真實生活中的 MNIST 和 CIFAR-10。

1.1 Simulate a function

在這個實驗中，我們選用兩個函數，其中一個是 $f(x) = \sin(x) + \cos(x^2)$ ；另外一個則是著名的碎形函數 Weierstrass function，以下我們將分別討論其實驗過程與結果。

1.1.1 訓練模型與參數

1. Optimizer: Adam, learning rate: 0.01, loss function: Mean square error
2. batch: 10000, epoch: 200000
3. 訓練範圍：0.0 到 10.0，間隔 0.001

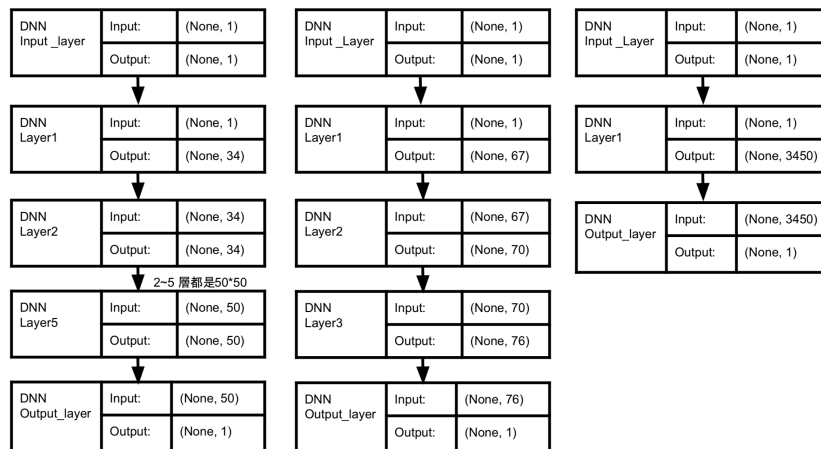


Figure 1: 本次實驗分別實作了 1, 3, 5 層的 DNN 網路，共 10367 個參數

1.1.2 Function-1 $f(x) = \sin(x) + \cos(x^2)$

1. Training Loss

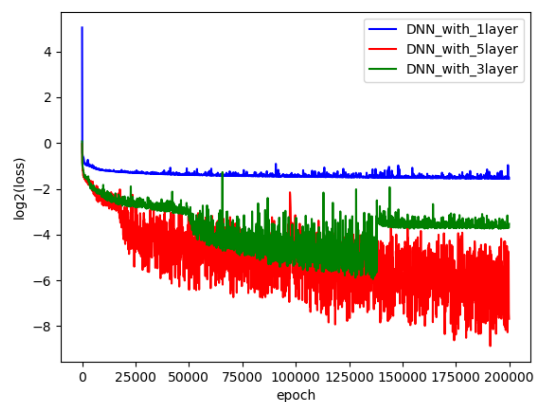


Figure 2: 上圖可以看到 deep 的表現比較好

2. Prediction Result

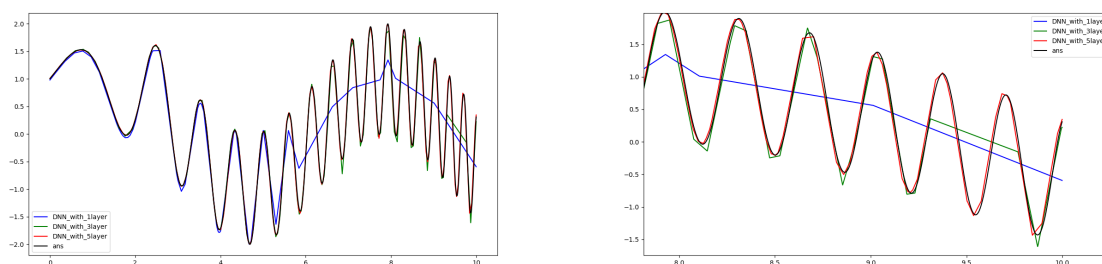


Figure 3: 右圖是局部放大圖，可以明顯看出 shallow 表現得很差

1.1.3 Function-2 $f(x) = \sum_{n=0}^{10} \left(\frac{1}{2}\right)^n \cos(2^n \pi x)$

1. Training Loss

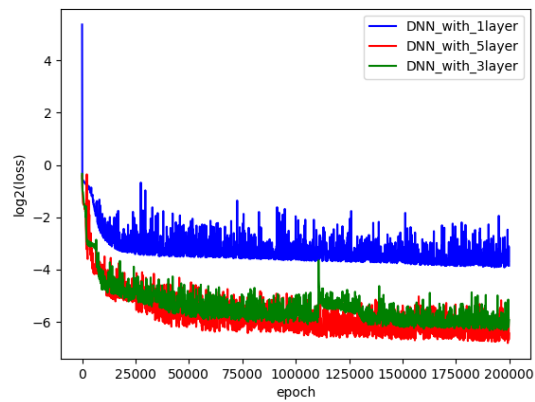


Figure 4: 上圖可以看到 deep 的表現比較好

2. Prediction Result

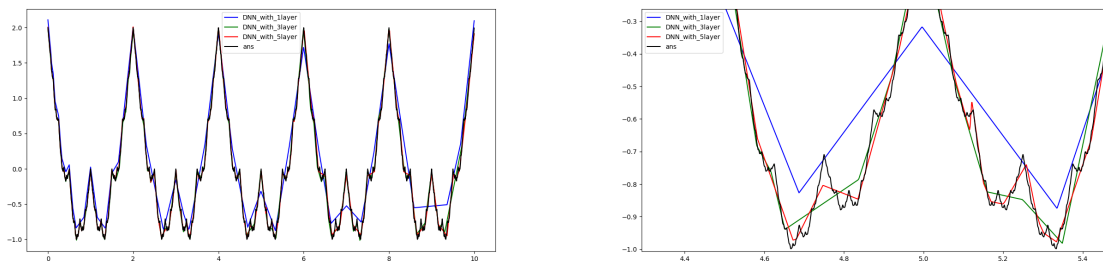


Figure 5: 右圖是局部放大圖，可以明顯看出 deep 的 piece 比較多也比較細緻

1.1.4 綜合討論

1. 實驗結果的 GIF 演進圖在：<https://imgur.com/gallery/MdENL>
2. 在 GIF 上面可以看到 shallow 在收斂過程中是沿著 x 軸一塊一塊的去 fit，然而在 deep 的架構下則有發現模組化的現象。在前幾個 epoch 就在整個 x 軸上掌握了大部分的特徵，在接下來的 epoch 只是在細部的地方微調而已。
3. 在兩個函數大圖 result 上面可以看到 deep 可以 fit 比較好是因為他產生比較多的 pieces，在折角處都表現得較好，這個也是符合理論。
4. 從 predict result 中可以看到越深的網路架構在收斂的過程中震盪比較大，這是因為在反向傳播的時候 deep 前面的參數改了一下會造成比較大的影響
5. 有點可惜的是照理說 shallow 一開始 loss 會掉得比較快，但是在這次的圖中不是很明顯
6. 在 $f(x) = \sin(x) + \cos(x^2)$ 的 case 中可以蠻明顯的看到 shallow 的比較難走到 global minima

1.2 An actual task

1.2.1 訓練模型與參數

1. Optimizer: Adam, learning rate: 0.0001
2. loss function: Cross Entropy
3. batch: 100, epoch: 30000

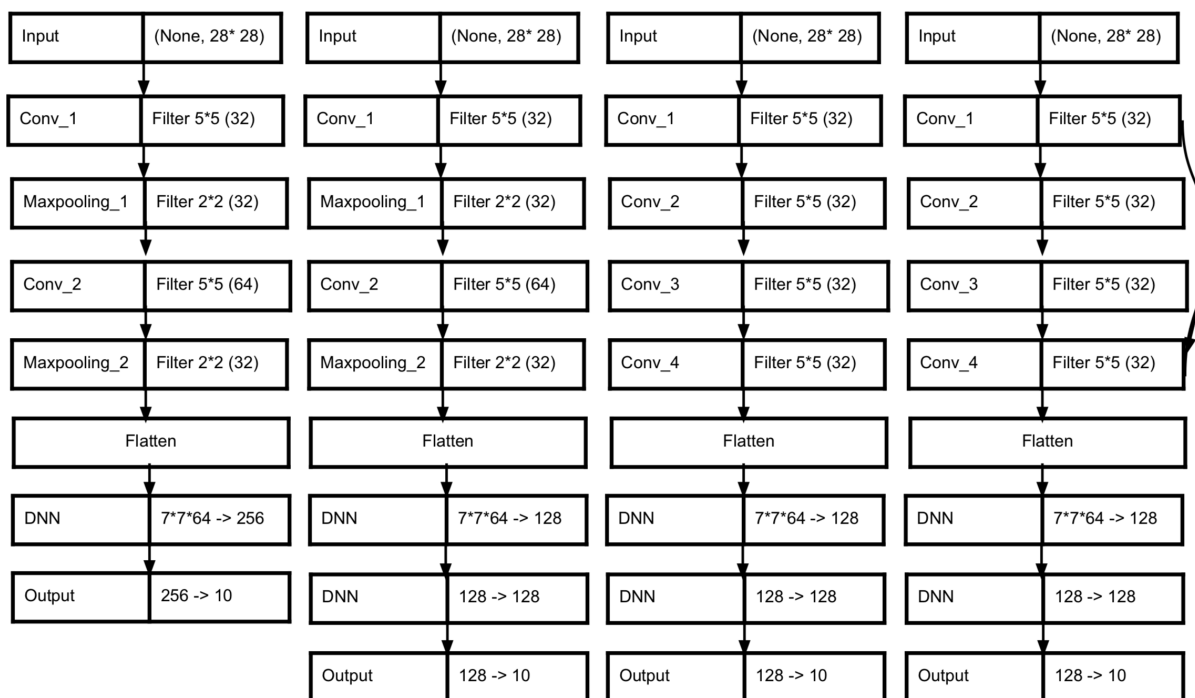


Figure 6: 本次實驗分別試了 DNN 深淺、有無 pooling 以及有無 residual

1.2.2 Task1 – MNIST

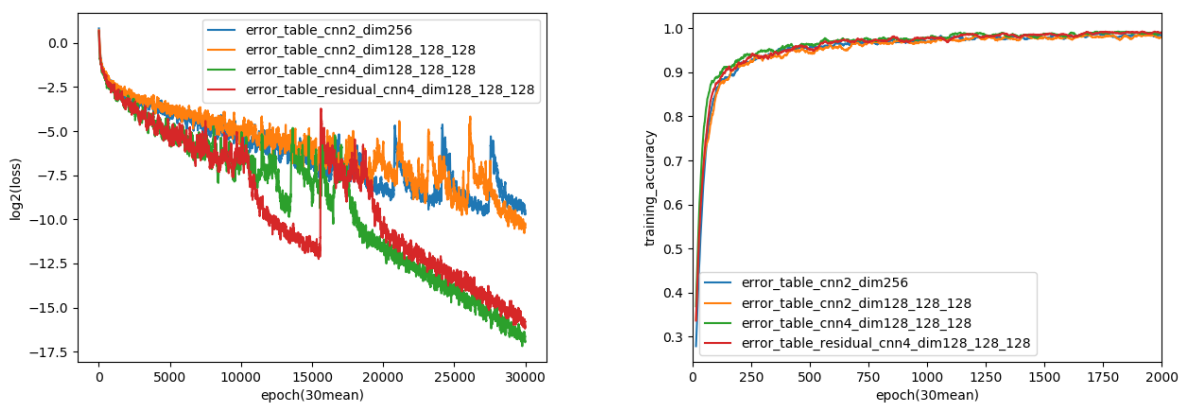


Figure 7: 左圖是 training loss，右圖是 Accuracy

1.2.3 Task2 – CIFAR-10

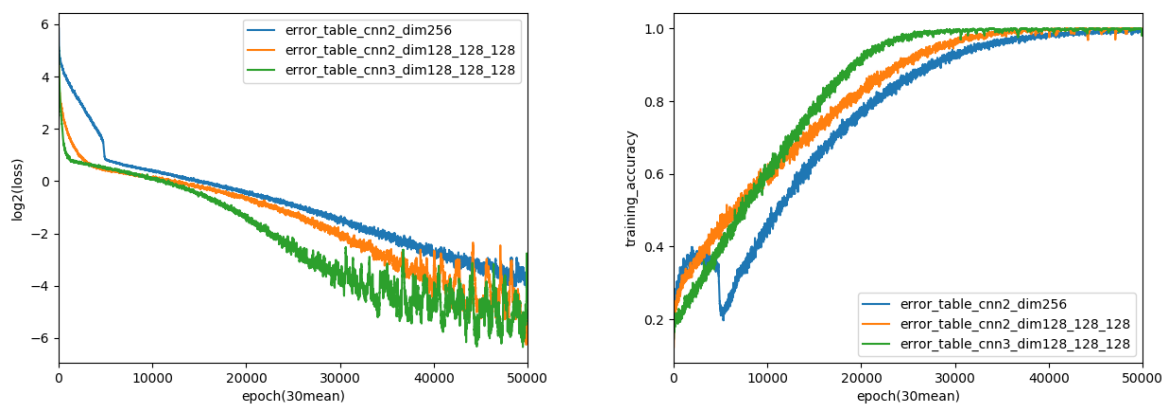


Figure 8: 左圖是 training loss，右圖是 Accuracy

1.2.4 綜合討論

1. 在紅色和綠色的比較中，有用 residual 的 case 確實收斂的比較快，只是表現沒有比較好。我們想要實驗 residual 的動機只是好奇這個得獎的做法有沒有比較好然後順便比較一下。
2. 在橘色與綠色的比較中，發現了沒有 max-pooling 的表現比較差
3. 在藍色與橘色兩條線的比較中，deep 的表現還是比 shallow 好，不過在上面沒有很明顯的原因可能是因為給的 neuron 比需要的多太多了
4. 因為收斂的圖震盪很大，因此我們都取每 30 個的平均一次讓圖變得比較易讀
5. 補充說明，因為 CIFAR 的參數比較難調，所以卷積跟池化就只做了三層，也沒有做 residual，另外 CIFAR 的 CNN filter 是 5×5