Dimension Reduction Homework

Basic Dataset

先針對 `iris`, `breast_cancer`, `digits` 這三個 dataset 進行降維，然後再用降維後的數據來訓練模型，看模型的 performance 有沒有變高。

Models

```
models = {
    'SVM': SVC(),
    'Logistic Regression': LogisticRegression(random_state=42, max_iter=5000),
    'Decision Tree': DecisionTreeClassifier(random_state=42),
    'Random Forest': RandomForestClassifier(n_estimators=100, random_state=42),
    'KNN': KNeighborsClassifier(n_neighbors=5),
    'Neural Network': MLPClassifier(hidden_layer_sizes=(100,), max_iter=1000, random_state=42)
}
```

使用了 SVM, Logistic Regression, Decision Tree, Random Forest, KNN, Neural Network 等方法來訓練。

Analysis

使用 PCA, Kernel PCA, t-SNE, UMAP 來進行降維分析。

```
analysis_methods = {
    'Original': None,
    'PCA': PCA(n_components=2),
    'KernelPCA (poly)': KernelPCA(n_components=2, kernel='poly'),
    'KernelPCA (rbf)': KernelPCA(n_components=2, kernel='rbf'),
    'TSNE': TSNE(n_components=2, random_state=42),
    'UMAP': UMAP(n_components=2, random_state=42)
}
```

Iris

Original Analysis

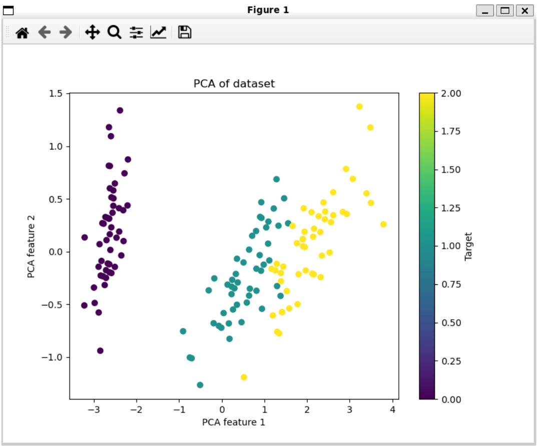
不經過處理然後把資料丟給模型訓練

result:

```
Original Analysis
Current Model is SVM, Average accuracy: 0.96666667
Current Model is Logistic Regression, Average accuracy: 0.97333333
Current Model is Decision Tree, Average accuracy: 0.95333333
Current Model is Random Forest, Average accuracy: 0.96000000
Current Model is KNN, Average accuracy: 0.97333333
Current Model is Neural Network, Average accuracy: 0.97333333
```

可以看到其實模型表現都已經很好了

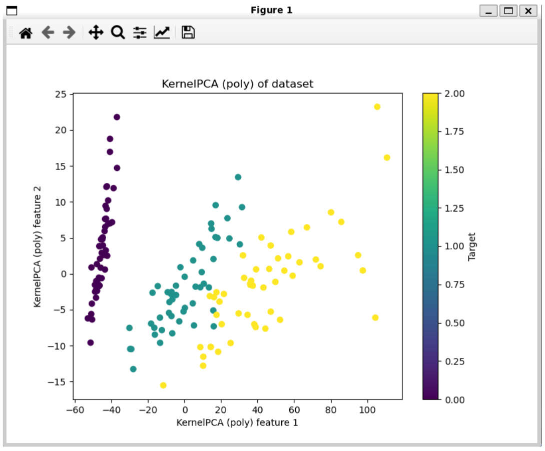
PCA Analysis



```
PCA Analysis
Explained variance ratio: [0.92461872 0.05306648]
Current Model is SVM, Average accuracy: 0.96666667
Current Model is Logistic Regression, Average accuracy: 0.96666667
Current Model is Decision Tree, Average accuracy: 0.96000000
Current Model is Random Forest, Average accuracy: 0.95333333
Current Model is KNN, Average accuracy: 0.97333333
Current Model is Neural Network, Average accuracy: 0.96000000
```

Kernel PCA (poly) Analysis

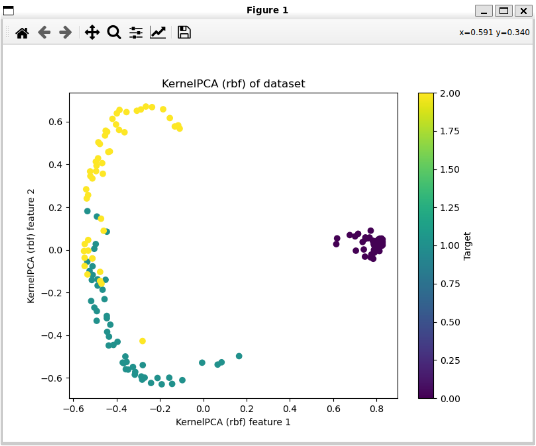
使用 `poly` 作為 kernel function。



result:

```
KernelPCA (poly) Analysis
Current Model is SVM, Average accuracy: 0.95333333
Current Model is Logistic Regression, Average accuracy: 0.96666667
Current Model is Decision Tree, Average accuracy: 0.95333333
Current Model is Random Forest, Average accuracy: 0.96666667
Current Model is KNN, Average accuracy: 0.97333333
Current Model is Neural Network, Average accuracy: 0.95333333
```

Kernel PCA (rbf) Analysis

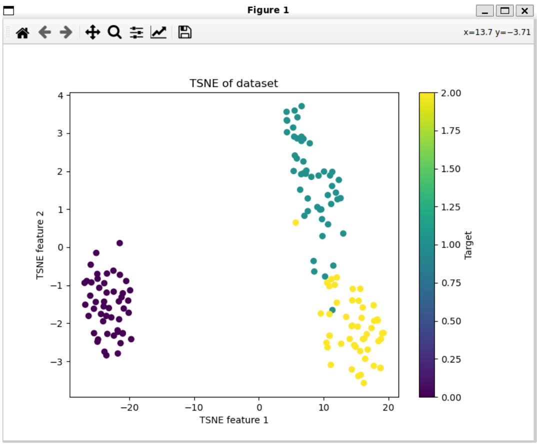


result:

```
KernelPCA (rbf) Analysis
Current Model is SVM, Average accuracy: 0.90000000
Current Model is Logistic Regression, Average accuracy: 0.92000000
Current Model is Decision Tree, Average accuracy: 0.90000000
Current Model is Random Forest, Average accuracy: 0.92666667
Current Model is KNN, Average accuracy: 0.91333333
Current Model is Neural Network, Average accuracy: 0.92000000
```

使用 RBF (Gaussian Kernel) 作為 kernel function 效果反而沒那麼好。

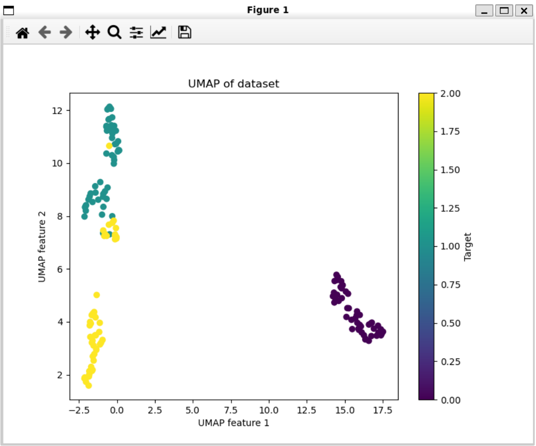
t-SNE Analysis



result:

```
TSNE Analysis
Current Model is SVM, Average accuracy: 0.96666667
Current Model is Logistic Regression, Average accuracy: 0.96666667
Current Model is Decision Tree, Average accuracy: 0.97333333
Current Model is Random Forest, Average accuracy: 0.97333333
Current Model is KNN, Average accuracy: 0.97333333
Current Model is Neural Network, Average accuracy: 0.96666667
```

UMAP Analysis



result:

```
UMAP Analysis
Current Model is SVM, Average accuracy: 0.90000000
Current Model is Logistic Regression, Average accuracy: 0.97333333
Current Model is Decision Tree, Average accuracy: 0.96000000
Current Model is Random Forest, Average accuracy: 0.97333333
Current Model is KNN, Average accuracy: 0.97333333
Current Model is Neural Network, Average accuracy: 0.96000000
```

結論

在這個 dataset 中，由於 original 的表現就很好，所以降維的幫助並不大。

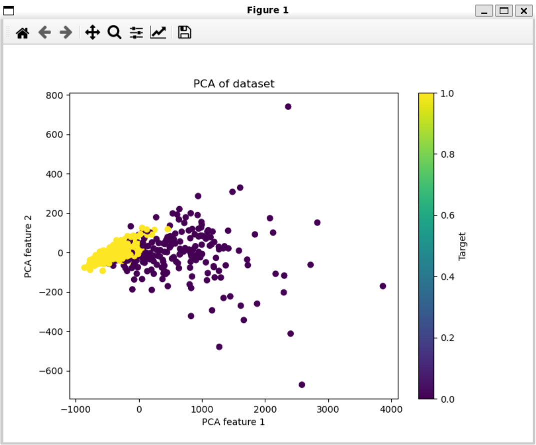
Breast Cancer

Original Analysis

result:

```
Original Analysis
Current Model is SVM, Average accuracy: 0.91731098
Current Model is Logistic Regression, Average accuracy: 0.95073746
Current Model is Decision Tree, Average accuracy: 0.93322465
Current Model is Random Forest, Average accuracy: 0.95783263
Current Model is KNN, Average accuracy: 0.93668685
Current Model is Neural Network, Average accuracy: 0.94195001
```

PCA Analysis

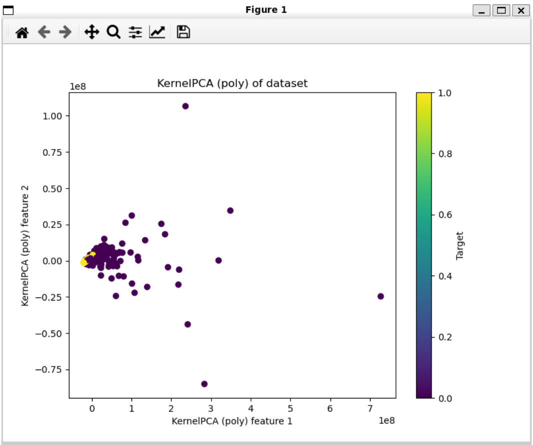


result:

```
PCA Analysis
Explained variance ratio: [0.98204467 0.01617649]
Current Model is SVM, Average accuracy: 0.91557212
Current Model is Logistic Regression, Average accuracy: 0.92612948
Current Model is Decision Tree, Average accuracy: 0.90335352
Current Model is Random Forest, Average accuracy: 0.92793045
Current Model is KNN, Average accuracy: 0.93142369
Current Model is Neural Network, Average accuracy: 0.92437510
```

整體的準確度都降低了一些... 但還是跟原本的差不多。

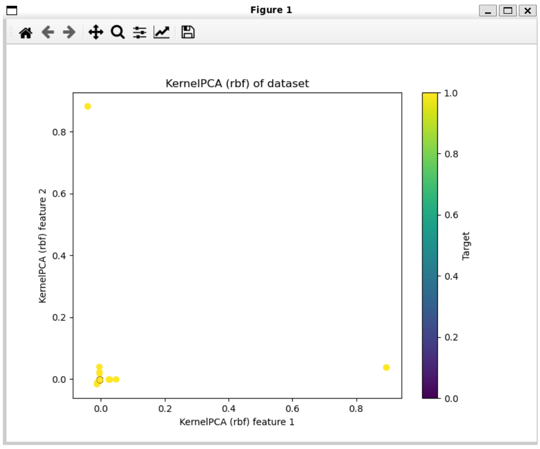
Kernel PCA (poly) Analysis



```
KernelPCA (poly) Analysis
Current Model is SVM, Average accuracy: 0.89094861
Current Model is Logistic Regression, Average accuracy: 0.83823940
Current Model is Decision Tree, Average accuracy: 0.91564974
Current Model is Random Forest, Average accuracy: 0.92263624
Current Model is KNN, Average accuracy: 0.91912746
Current Model is Neural Network, Average accuracy: 0.87165036
```

準確度降低不少，看來Kernel PCA不太適合這個dataset。

Kernel PCA (rbf) Analysis



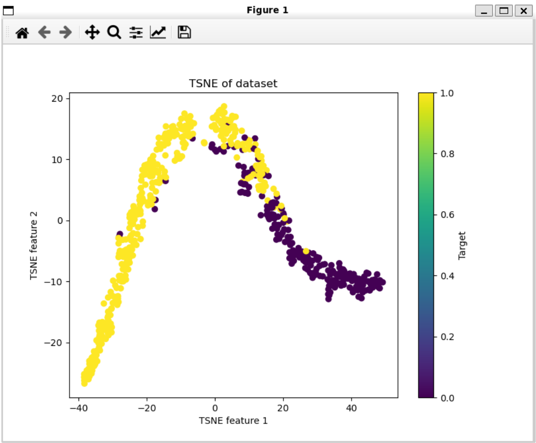
這甚麼鬼

result:

```
KernelPCA (rbf) Analysis
Current Model is SVM, Average accuracy: 0.62735600
Current Model is Logistic Regression, Average accuracy: 0.62735600
Current Model is Decision Tree, Average accuracy: 0.81547896
Current Model is Random Forest, Average accuracy: 0.83129949
Current Model is KNN, Average accuracy: 0.87700668
Current Model is Neural Network, Average accuracy: 0.62735600
```

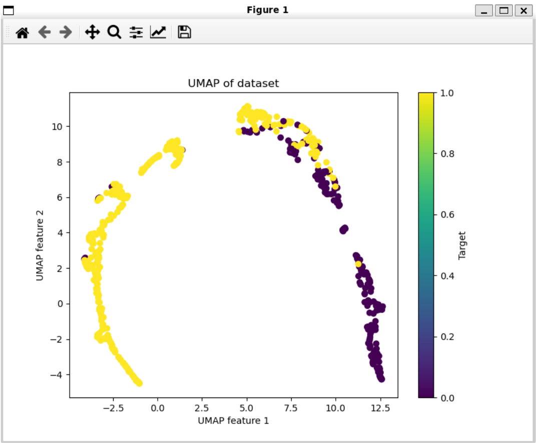
準確度降低很多。

t-SNE Analysis



```
TSNE Analysis
Current Model is SVM, Average accuracy: 0.91029343
Current Model is Logistic Regression, Average accuracy: 0.88745536
Current Model is Decision Tree, Average accuracy: 0.89979817
Current Model is Random Forest, Average accuracy: 0.92788387
Current Model is KNN, Average accuracy: 0.92439062
Current Model is Neural Network, Average accuracy: 0.91027791
```

UMAP Analysis



result:

```
UMAP Analysis
Current Model is SVM, Average accuracy: 0.90675361
Current Model is Logistic Regression, Average accuracy: 0.87692905
Current Model is Decision Tree, Average accuracy: 0.89628940
Current Model is Random Forest, Average accuracy: 0.92439062
Current Model is KNN, Average accuracy: 0.92437510
Current Model is Neural Network, Average accuracy: 0.88920975
```

結論

整體而言，t-SNE 和 PCA 比較適合這個 dataset 的降維處理。

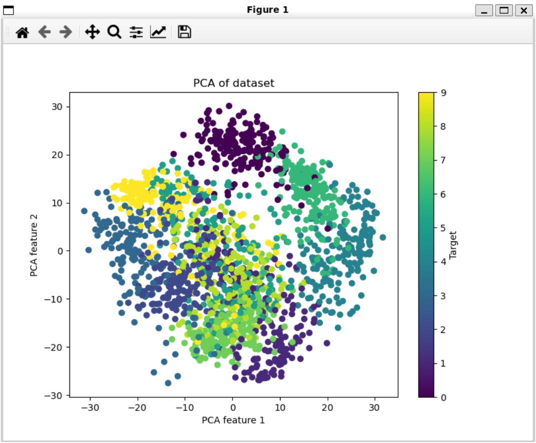
Digits

Original Analysis

```
Original Analysis
Current Model is SVM, Average accuracy: 0.98775766
Current Model is Logistic Regression, Average accuracy: 0.96159393
Current Model is Decision Tree, Average accuracy: 0.85698236
Current Model is Random Forest, Average accuracy: 0.97551223
Current Model is KNN, Average accuracy: 0.98608326
Current Model is Neural Network, Average accuracy: 0.96939338
```

除了 Decision Tree 以外，其他模型都有很不錯的結果。

PCA Analysis



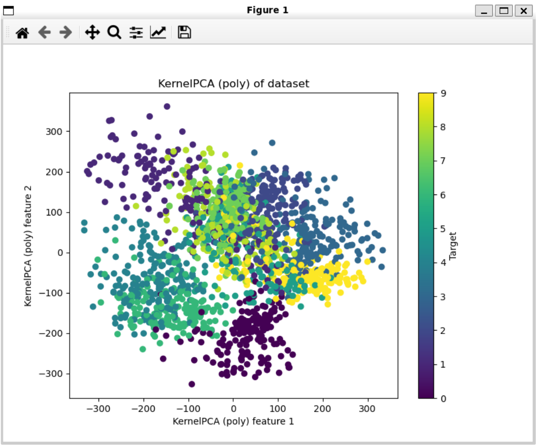
漆彈大作戰(?

result:

```
PCA Analysis
Explained variance ratio: [0.14890594 0.13618771]
Current Model is SVM, Average accuracy: 0.65832714
Current Model is Logistic Regression, Average accuracy: 0.60266171
Current Model is Decision Tree, Average accuracy: 0.58933148
Current Model is Random Forest, Average accuracy: 0.61881151
Current Model is KNN, Average accuracy: 0.63104302
Current Model is Neural Network, Average accuracy: 0.64885949
```

準確度降低不少。

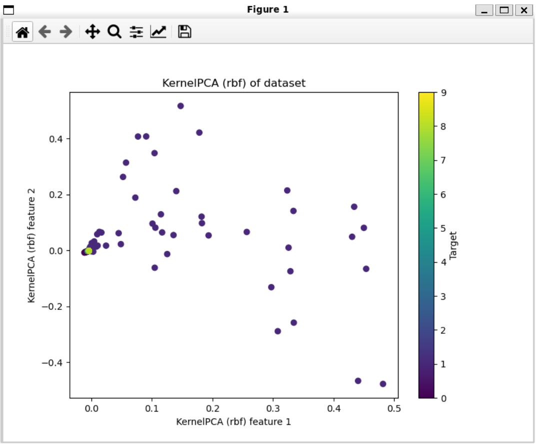
Kernel PCA (poly) Analysis



result:

```
KernelPCA (poly) Analysis
Current Model is SVM, Average accuracy: 0.60822965
Current Model is Logistic Regression, Average accuracy: 0.57761374
Current Model is Decision Tree, Average accuracy: 0.54533890
Current Model is Random Forest, Average accuracy: 0.59208140
Current Model is KNN, Average accuracy: 0.57205819
Current Model is Neural Network, Average accuracy: 0.55983906
```

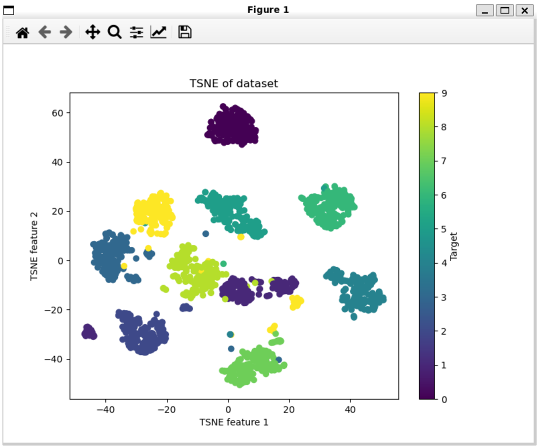
Kernel PCA (rbf) Analysis



```
KernelPCA (rbf) Analysis
Current Model is SVM, Average accuracy: 0.12075054
Current Model is Logistic Regression, Average accuracy: 0.11240173
Current Model is Decision Tree, Average accuracy: 0.26209842
Current Model is Random Forest, Average accuracy: 0.26712318
Current Model is KNN, Average accuracy: 0.25264779
Current Model is Neural Network, Average accuracy: 0.17808418
```

惨不忍睹

t-SNE Analysis

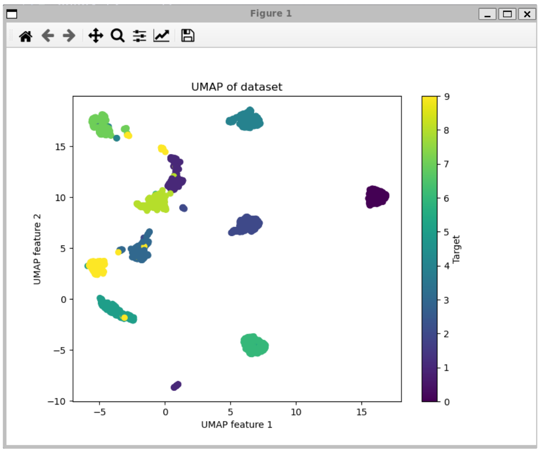


看起來還不錯?

```
TSNE Analysis
Current Model is SVM, Average accuracy: 0.97328381
Current Model is Logistic Regression, Average accuracy: 0.92208759
Current Model is Decision Tree, Average accuracy: 0.97829000
Current Model is Random Forest, Average accuracy: 0.98775611
Current Model is KNN, Average accuracy: 0.98775611
Current Model is Neural Network, Average accuracy: 0.98051842
```

準確度提高了不少，連 Decision Tree 都練起來了

UMAP Analysis



result:

```
UMAP Analysis
Current Model is SVM, Average accuracy: 0.96493810
Current Model is Logistic Regression, Average accuracy: 0.93043330
Current Model is Decision Tree, Average accuracy: 0.98051842
Current Model is Random Forest, Average accuracy: 0.98664036
Current Model is KNN, Average accuracy: 0.98719901
Current Model is Neural Network, Average accuracy: 0.98052461
```

跟 t-SNE 差不多，都得到了很好的表現。

結論

在這個 dataset 中，t-SNE 和 UMAP 表現都很好，這個資料集中，有非線性和多個局部特徵的這些特性，PCA這種線性降維就不太適合，Kernel PCA雖然可以處理非線性資料，但他不如 t-SNE 和 UMAP 能夠有效地保持局部結構。

CIFAR10

在之前我有用 AlexNet, ResNet-9, ResNet-152 等架構去訓練 CIFAR10 這個資料集，可以針對訓練好的模型提取出來的特徵進行降維分析。

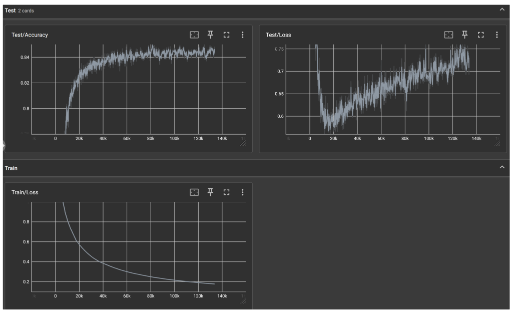
然後分類方法這裡就只用 pytorch 手刻的神經網路，因為 CIFAR10 這個資料集較大，sklearn 的分類器都是只能跑在 CPU 上的，訓練起來很慢而且效果也不太好。

分類器的模型結構很簡單，只有兩層隱藏層 (1024, 1024)。

而降維分析也只有使用 t-SNE 和 UMAP，因為 PCA 和 Kernel PCA 顯然不適合這種 dataset，在前面的分析中就可以看出。

AlexNet

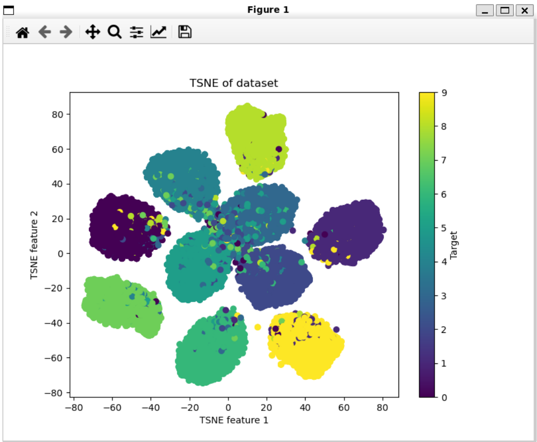
預訓練模型，大約收斂到 84.5% 的正確率



Original Analysis

```
Original Analysis
Epoch 1: [    0 / 50000] (0 %)  Loss: 2.312197
Epoch 1: [25600 / 50000] (51 %)  Loss: 1.156103
Epoch 2: [    0 / 50000] (0 %)  Loss: 0.771158
Epoch 2: [25600 / 50000] (51 %)  Loss: 0.578371
...
Epoch 19: [    0 / 50000] (0 %)  Loss: 0.062606
Epoch 19: [25600 / 50000] (51 %)  Loss: 0.060959
Epoch 20: [    0 / 50000] (0 %)  Loss: 0.059396
Epoch 20: [25600 / 50000] (51 %)  Loss: 0.057911
Current Model is Neural Network, Average accuracy: 0.84300000
```

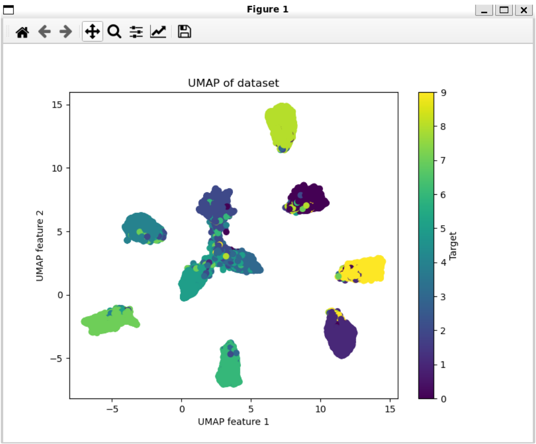
t-SNE Analysis



result:

```
TSNE Analysis
Epoch 1: [    0 / 50000] (0 %)  Loss: 4.835803
Epoch 1: [25600 / 50000] (51 %)  Loss: 2.639431
Epoch 2: [    0 / 50000] (0 %)  Loss: 1.819552
Epoch 2: [25600 / 50000] (51 %)  Loss: 1.386053
...
Epoch 19: [    0 / 50000] (0 %)  Loss: 0.180113
Epoch 19: [25600 / 50000] (51 %)  Loss: 0.175481
Epoch 20: [    0 / 50000] (0 %)  Loss: 0.171708
Epoch 20: [25600 / 50000] (51 %)  Loss: 0.167430
Current Model is Neural Network, Average accuracy: 0.83560000
```

UMAP Analysis

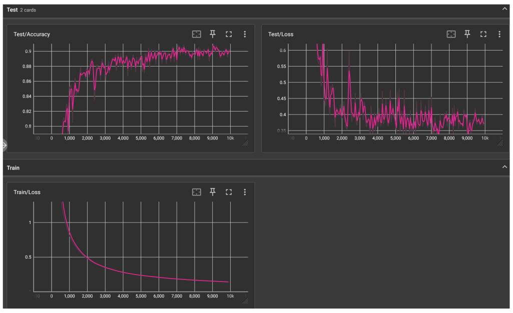


result:

```
UMAP Analysis
Epoch 1: [    0 / 50000] (0 %)  Loss: 2.467787
Epoch 1: [25600 / 50000] (51 %)  Loss: 1.237651
Epoch 2: [    0 / 50000] (0 %)  Loss: 0.825718
Epoch 2: [25600 / 50000] (51 %)  Loss: 0.619543
...
Epoch 19: [    0 / 50000] (0 %)  Loss: 0.076735
Epoch 19: [25600 / 50000] (51 %)  Loss: 0.075148
Epoch 20: [    0 / 50000] (0 %)  Loss: 0.073367
Epoch 20: [25600 / 50000] (51 %)  Loss: 0.071550
Current Model is Neural Network, Average accuracy: 0.83300000
```

ResNet-9

預訓練模型，大約收斂到 90.5% 的正確率。

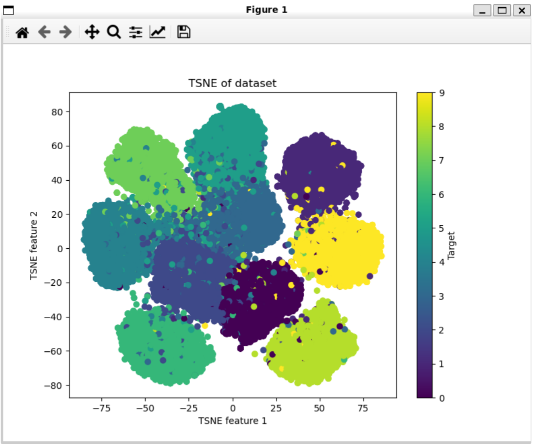


Original Analysis

```
Original Analysis
Epoch 1: [    0 / 50000] (0 %)  Loss: 2.362766
Epoch 1: [25600 / 50000] (51 %)  Loss: 1.202956
Epoch 2: [    0 / 50000] (0 %)  Loss: 0.804938
Epoch 2: [25600 / 50000] (51 %)  Loss: 0.605108
...
Epoch 19: [    0 / 50000] (0 %)  Loss: 0.069009
Epoch 19: [25600 / 50000] (51 %)  Loss: 0.067193
Epoch 20: [    0 / 50000] (0 %)  Loss: 0.065471
Epoch 20: [25600 / 50000] (51 %)  Loss: 0.063834
Current Model is Neural Network, Average accuracy: 0.91660000
```

換了個分類器讓準確度變高了?!

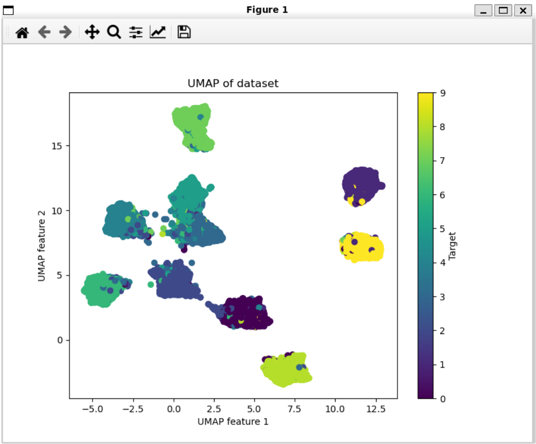
t-SNE Analysis



result:

```
TSNE Analysis
Epoch 1: [    0 / 50000] (0 %)  Loss: 6.549839
Epoch 1: [25600 / 50000] (51 %)  Loss: 3.511696
Epoch 2: [    0 / 50000] (0 %)  Loss: 2.468609
Epoch 2: [25600 / 50000] (51 %)  Loss: 1.905890
...
Epoch 19: [    0 / 50000] (0 %)  Loss: 0.278508
Epoch 19: [25600 / 50000] (51 %)  Loss: 0.272290
Epoch 20: [    0 / 50000] (0 %)  Loss: 0.266938
Epoch 20: [25600 / 50000] (51 %)  Loss: 0.261526
Current Model is Neural Network, Average accuracy: 0.91250000
```

UMAP Analysis



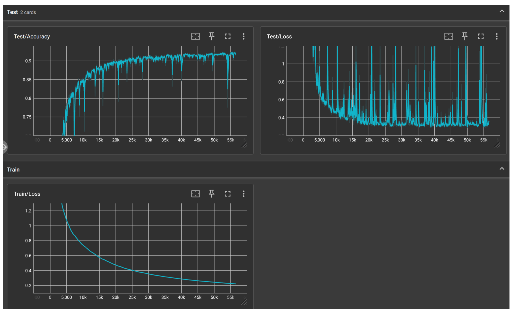
result:

```
UMAP Analysis
Epoch 1: [    0 / 50000] (0 %)  Loss: 2.517041
Epoch 1: [25600 / 50000] (51 %)  Loss: 1.345312
Epoch 2: [    0 / 50000] (0 %)  Loss: 0.917512
Epoch 2: [25600 / 50000] (51 %)  Loss: 0.694256
...
Epoch 19: [    0 / 50000] (0 %)  Loss: 0.131593
Epoch 19: [25600 / 50000] (51 %)  Loss: 0.132036
Epoch 20: [    0 / 50000] (0 %)  Loss: 0.129874
Epoch 20: [25600 / 50000] (51 %)  Loss: 0.127358
Current Model is Neural Network, Average accuracy: 0.91170000
```

看起來模型的bottleneck在於4 ~ 6這幾個類別不容易分清楚，查了一下是有關貓狗類的。

ResNet-152

預訓練模型，大約收斂到 92% 的正確率。

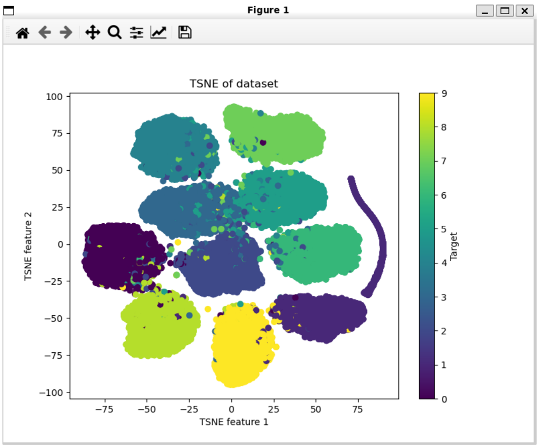


Original Analysis

result:

```
Original Analysis
Epoch 1: [    0 / 50000] (0 %)  Loss: 2.313861
Epoch 1: [25600 / 50000] (51 %)  Loss: 1.163437
Epoch 2: [    0 / 50000] (0 %)  Loss: 0.776662
Epoch 2: [25600 / 50000] (51 %)  Loss: 0.582697
...
Epoch 19: [    0 / 50000] (0 %)  Loss: 0.064706
Epoch 19: [25600 / 50000] (51 %)  Loss: 0.063003
Epoch 20: [    0 / 50000] (0 %)  Loss: 0.061388
Epoch 20: [25600 / 50000] (51 %)  Loss: 0.059854
Current Model is Neural Network, Average accuracy: 0.92270000
```

t-SNE Analysis

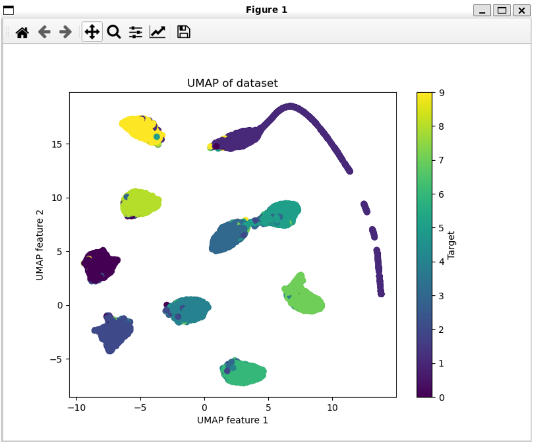


那條毛毛蟲是怎麼回事

result:

```
TSNE Analysis
Epoch 1: [    0 / 50000] (0 %)  Loss: 7.100447
Epoch 1: [25600 / 50000] (51 %)  Loss: 3.793626
Epoch 2: [    0 / 50000] (0 %)  Loss: 2.638284
Epoch 2: [25600 / 50000] (51 %)  Loss: 2.044024
...
Epoch 19: [    0 / 50000] (0 %)  Loss: 0.282281
Epoch 19: [25600 / 50000] (51 %)  Loss: 0.275001
Epoch 20: [    0 / 50000] (0 %)  Loss: 0.268637
Epoch 20: [25600 / 50000] (51 %)  Loss: 0.262047
Current Model is Neural Network, Average accuracy: 0.86670000
```

UMAP Analysis



result:

```
UMAP Analysis
Epoch 1: [    0 / 50000] (0 %)  Loss: 2.694722
Epoch 1: [25600 / 50000] (51 %)  Loss: 1.361425
Epoch 2: [    0 / 50000] (0 %)  Loss: 0.910756
Epoch 2: [25600 / 50000] (51 %)  Loss: 0.692358
...
Epoch 19: [    0 / 50000] (0 %)  Loss: 0.079821
Epoch 19: [25600 / 50000] (51 %)  Loss: 0.077747
Epoch 20: [    0 / 50000] (0 %)  Loss: 0.075764
Epoch 20: [25600 / 50000] (51 %)  Loss: 0.073986
Current Model is Neural Network, Average accuracy: 0.92290000
```

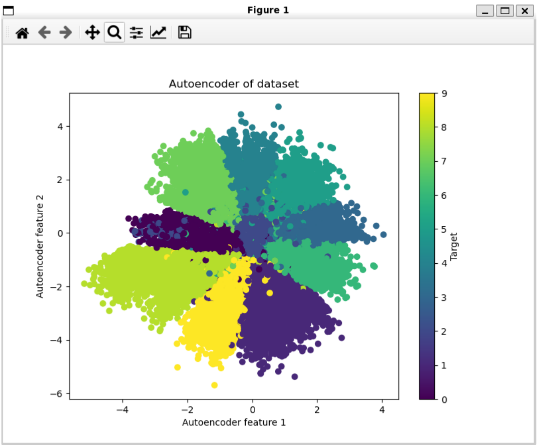
結論

雖然兩種降維分析後都可以維持模型的性能，但是UMAP更能看出模型在提取特徵時有哪些類別的特徵是容易搞混的。

UMAP 表現通常都比 t-SNE 好。

Autoencoder

用 Autoencoder 來進行降維分析，針對 ResNet-9 所提取的特徵進行降維



result:

```
Autoencoder Analysis
Epoch 1: [    0 / 50000] (0 %)  Loss: 2.300393
Epoch 1: [25600 / 50000] (51 %)  Loss: 1.194649
Epoch 2: [    0 / 50000] (0 %)  Loss: 0.810597
Epoch 2: [25600 / 50000] (51 %)  Loss: 0.617195
...
Epoch 19: [    0 / 50000] (0 %)  Loss: 0.108679
Epoch 19: [25600 / 50000] (51 %)  Loss: 0.106233
Epoch 20: [    0 / 50000] (0 %)  Loss: 0.104363
Epoch 20: [25600 / 50000] (51 %)  Loss: 0.102601
Current Model is Neural Network, Average accuracy: 0.8993000
```

just for fun :D