# 基於元學習之深度卷積孿生網路做圖像識別之設計與實作

**Design and Implementation of Deep Convolutional Siamese Network for Image Recognition Based on One-Shot-Learning**


## 摘要

深度學習（Deep Learning, DL）已發展成為這波人工智慧（Artificial Intelligence, AI)熱潮的主流技術，應用於各個領域，如人臉辨識，其中孿生網路（Siamese Network）扮演重要角色。然而，對於某些應用中資料不足的情況，元學習（Meta Learning）提供一個有效且辨識結果不錯的學習模式。但過少的資料會導致過擬合（overfitting）使網路表現不佳。

在本研究中，將透過設計一個孿生網路架構，並基於幾個常用的損失函數之下嘗試提議用於孿生網路上更有效的損失函數，研究結果顯示，使用調適後的基本網路和對比損失函數（Contrastive Loss）、Adam 優化器和 ReLU 激活函數，在 MNIST、LFW、AT&T 資料集上，以小資料量（基於 One-Shot 學習架構）和較少的網路參數量，分別在測試資料上，達到 97.27%、76%與 75.63%的識別準確率，且訓練收斂速度較快。



關鍵詞：元學習、影像辨識、孿生網路、One-Shot Learning、對比損失



## Abstract

Deep Learning has become the mainstream technology in current AI development and facial recognition one of DL application, where Siamese Networks play an important role. However, for task with insufficient data, Meta Learning provides an effective learning model with good recognition results.

In this study, a Siamese network architecture is designed and several commonly used loss functions are proposed to be used on Siamese networks for better performance. Results show that using a modified Siamese network and Contrastive Loss, Adam optimizer, and ReLU activation function, the proposed model achieves 97.27%, 76%, and 72.72% accuracy on MNIST, LFW, and AT&T datasets, respectively, with less data and fewer network parameters and faster convergence speed.



**Keywords**：Meta Learning; Image Recognition; Siamese Network; One-Shot Learning; Contrastive Loss

## (一)前言

在人臉辨識並分類的任務中，元學習固然是一個有效的學習模式，但是在辨識準確率中，依然很難達到如同擁有海量學習資料集的網路訓練出來的準確率，以及目前在實務中，依然使用大資料量做訓練，只在測試任務中實作元學習。因此，本研究計畫先由經典的基於元學習架構下的孿生網路應用於圖像辨識的文獻入手，完全依照元學習架構，在訓練與測試任務中使用 One-Shot 學習，並加以調整後的孿生網路架構與損失函數，觀察是否可以提高準確率和模型泛化性。

## (二) 研究材料

### 3.1 元學習（Meta Learning）

一個應用在小資料集的有效學習模式，也被稱為"**學習如何學習**"（Learn to Learn），它與常見的深度學習模式不同。不是學習適當權重分佈，而是學習適當權重分佈的學習模式。這意味著學習者通過推論方法獲取知識，例如通過學習貓、豹和老虎的外貌，推論出猞猁可能屬於貓科，這種融會貫通的能力更接近人類學習方式。

元學習的具體概念是讓網絡訓練多種任務，藉由訓練網絡辨識三角形與圓形、三角形與長方形以及圓形與長方形等任務，找到在各種任務中表現最好的模型，將其應用於蘋果和香蕉的辨識中。元學習學習架構根據使用資料的數量($N$)和資料類別數($K$)來命名，稱為 $K$-way $N$-shot Learning。例如，5-way One-shot Learning 的支持集有五個類別，每個類別只有一個資料。

因為元學習的訓練單位是任務，而非單張資料，每個訓練、驗證和測試子任務都包含訓練和測試資料，為方便閱讀，下文將子任務中的訓練資料稱為支援集（Support set），測試資料稱為查詢集（Query set）。

### 3.2 孿生網路（Siamese Networks）

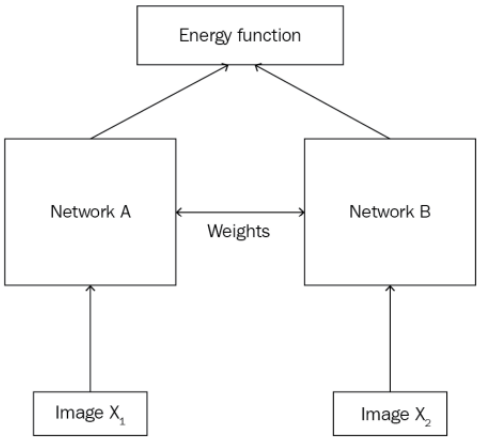


圖 1：基本孿生網路架構[3]

網路架構包含兩個共享權重且構造相同的神經網路，同時輸入來自不同類別的訓練資料，生成特徵向量後比較差異。孿生網路的目標為拉近相似圖片的距離或讓它們在特徵空間中的位置相近，以提高辨識準確率。基本結構包含 energy function、loss function、Sigmoid 正規化或 Softmax 分類層及全連接層。

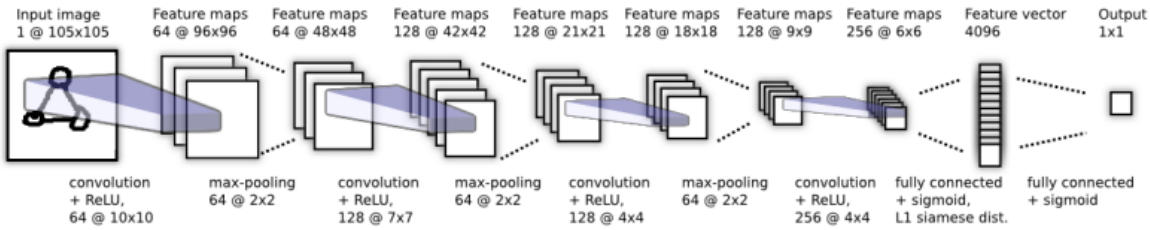


圖 2 文獻[4]使用之孿生網路基本模型

## (三)相關文獻

文獻[4]為基於元學習架構下的孿生網路應用於圖像辨識十分代表性的研究，展現了孿生網路在少資料學習下依然可以達到良好的準確率。該篇論文實驗用資料集為 Omniglot，由 50 個來自不同地區的字母表組成。在實作時僅在各類別中抽取一個數據，稱為 One-Shot 學習。提取特徵使用的基本網路如圖 2 所示。損失函數為交叉熵損失(Cross Entropy Loss)，為了增加資料的多樣性，作者加入仿射轉換(Affine distortions)增強訓練集，最後在 Omniglot 的辨識準確度為 92%。將相同網路架構應用於 MNIST 資料集時，準確度也有 70.3%。

在人臉辨識任務中，文獻[5]為最早將孿生網路應用於小量資料集的真假人臉辨識任務中，使用的人臉資料集為 AT&T，每個類別都有十張圖像，可做為 Few-Shot 學習架構。基本網路架構使用 LeNet-5，並用 $L_1$ 距離度量輸入對的特徵向量間的距離，且提出了新的損失函數：對比損失（Contrastive Loss），可以驅動模型正確的辨識人臉的相似與不相似，最終在 AT&T 數據集中達到不錯的準確率。

## (四)研究方法

### 5.1 模型設計

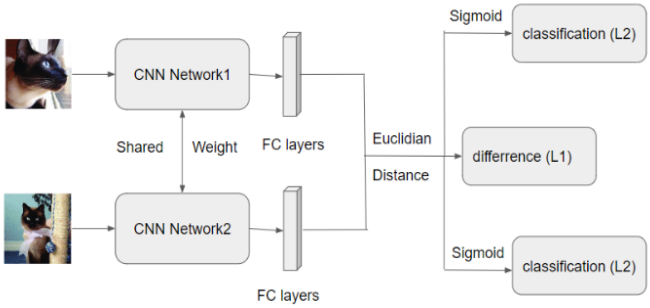


圖 3：本計畫所使用之孿生網路架

文獻[7][8]中作者混合不同損失函數作孿生網路的聯合訓練，參考這兩篇文獻，此研究使用的孿生網路架構如圖 5，權重更新同時考慮分類損失函數與相似度損失函數。

### 5.2 損失函數設計[1][26]

基於元學習架構訓練，資料很少（一個類別只有少量資料，甚或只有一筆資料），若訓練中控制優化方向的損失函數對訓練資料量大小不太敏感，可以避免梯度消失的問題，因此選擇對比損失（Contrastive Loss）作為新損失函數構成之一。此外，在人

臉多元分類的任務中，損失函數的優化方向要可以明確地分辨各類別，將同類別資料在特徵空間中的距離縮小，增大不同類別資料的特徵向量距離，因選擇 $L$-Softmax Loss 作為新損失函數的構成之二，並加入 Focal Loss 防止訓練網路時發生過擬合。綜上所述，本研究中採用的損失函數將考慮結合兩類損失函數：$Loss = L_1(\textit{相似度損失函數}) + L_2(\textit{多元分類損失函數})$，其中 $L_1$ 為對比損失（Contrastive Loss），而 $L_2$ 為 $L$-Softmax Loss 和 Focal Loss 的加總。接下來將更深入的解釋各項損失函數。

**5.2.1 $L_1$-相似度損失函數**[17]

對比損失（contrastive loss）基於預測和真實輸出之間相似性度量，函數表示如下：

$$L_1 = L(W, Y, X_1, X_2) = (1-Y)\frac{1}{2}(D_W)^2 + \frac{1}{2}Y\{\max(0, m - D_W)\}^2 \quad D_W = ||X_1 - X_2||^2$$

$W$ 是網路參數權重，$Y$ 是成對標籤，如果 $X_1$、$X_2$ 這對樣本屬於同一類，則 $Y$=0；$Y$=1 代表樣本屬於不同類。$D$ 是樣本對的 $L_1$ 距離。當 $Y$=0，調整參數，最小化 $X_1$、$X_2$ 之間的距離。當 $Y$=1，且 $X_1$、$X_2$ 之間的距離大於 $m$，則不做優化（省時省力）；當 $X_1$、$X_2$ 之間的距離小於 $m$，則增大兩者距離到 $m$。

**5.2.2 $L_2$-多元分類損失函數-1**[29]

$L$-Softmax（Large-Margin Softmax Loss），在原 Softmax Loss 下使用餘弦距離代替原 $L_1$ 距離作為距離度量的函數，表示如下：

$$L_s = \left(-log\frac{e^{\|W_{yi}\|\|X_i\|\varphi(\theta_{yi})}}{e^{\|W_{yi}\|\|X_i\|\varphi(\theta_{yi})} + \sum_{j\neq yi}(e^{\|W_j\|\|X_i\|\varphi(\theta_j)})}\right) \quad \varphi(\theta) = \begin{cases}\cos(m\theta) & 0 \le \theta \le \frac{\pi}{m} \\ D(\theta) & \frac{\pi}{m} \le \theta \le \pi\end{cases}$$

由於使用餘弦距離作為度量工具，可以減少類內距離並增大類外距離，加強分類的準確度，還可以學習更多的判別特徵。

**5.2.3 $L_2$-多元分類損失函數-2**[30]

Focal Loss 與交叉熵損失函數相似，解決“簡單”樣本主導訓練過程並關注學習對“困難”樣本（難以分類或屬於少數類別的樣本）進行分類的問題，函數表示如下：

$$L_f = (-\alpha_t(1-p_t)^{\gamma}\log(p_t)) \quad p_t = \begin{cases}p & y = 1 \\ 1-p & else\end{cases}$$

當$\alpha_t = 0.5, \gamma = 0$，此函數變為交叉熵損失。當$\gamma$增加時，調製係數$1-p_t$也增加。專注參數$\gamma$平滑地調節了易分類樣本調低權值的比例。調製係數目的是減少易分類樣本的權重，使模型在訓練時更專注於難分類的樣本。$p$表示預測樣本屬於 1 的機率（範圍為 0~1）。

## （五）訓練設置

**6.1 實驗用資料集**

此研究使用以下幾種資料集進行實驗，分別有 MNIST 手寫阿拉伯數字資料集、LFW 野外標記的面孔資料集、mini-ImageNet、AT&T（Our Database of Faces），其資料組成如表 1 所列。

表 1 資料集分析

| DataSet | | MNIST | LFW | mini-ImageNet | AT&T |
|---|---|---|---|---|---|
| Train | Class | 10 | 4599 | 64 | 40 |
| | Total Images | 6000 | 4599 | 38,400 | 320 |
| Valadian | Class | Same as Train | Same as Train | 20 | Same as Train |
| | Total Images | Same as Train | Same as Train | 12000 | Same as Train |
| Test | Class | 10 | 1150 | 16 | 40 |
| | Total Images | 1000 | 1150 | 9600 | 80 |

## 6.2 資料生成器（Data Generater）

本研究中的資料生成器，分為訓練對（Training Pairs）和 10-Way One-Shot 支持集（Support Set）。訓練對若取自同類，則標籤設為 0，反之，取自不同類則設為 1；10-Way One-Shot 支持集取自 10 類，每類選取一張圖像。

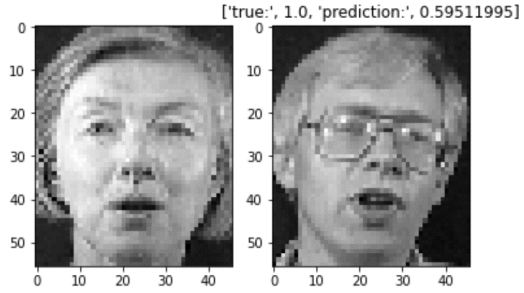


圖 4 取自不同類的訓練對

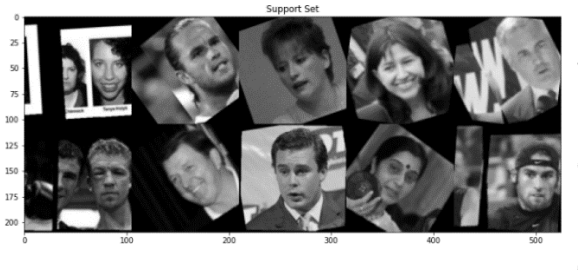


圖 5 10-Way One-Shot 支持集

### 6.2.1 孿生網路中的網路模型

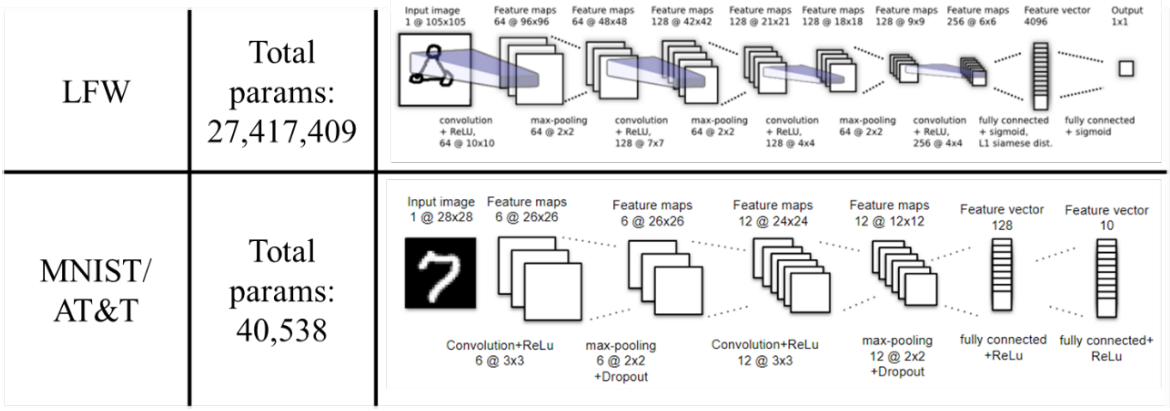


圖 6 不同訓練集中使用的孿生網路基本模型與總參數量

### 6.2.2 訓練設置

- **優化器(Optimizer)**：使用 Adam 最為優化器，學習率設為 5e-4，權重更新根據下

列函數：$w \leftarrow w - \eta \frac{\widehat{m_t}}{\sqrt{\widehat{v_t}}+\epsilon}$ ； $\widehat{m_t} = \frac{m_t}{1-\beta_1{}^t}$ ； $\widehat{v_t} = \frac{v_t}{1-\beta_2{}^t}$。

- **初始化**：所有層的初始化來自於正態分佈的卷積層中，權重為零均值(zero-mean)和$10^{-2}$的標準偏差（bias)；偏差為 0.5 均值和標準偏差$10^{-2}$。在全連接層，也有權重與偏差的初始化，但權重是從具有零均值和標準差$2 \times 10^{-1}$更廣泛的正態分佈。
- **過擬和**：卷積層中加入權重正則化(Regularize)，對損失函數使用 L2 正則化(懲罰項)，減少特徵權重差異，讓某些特徵的權重不要太突出；也使用 Dropout 層代替 BatchNormalization 層。
- **學習規劃**：不同損失函數有不同資料輸入型態，Focal Loss 使用 10-Way One-Shot 支持集；Contrastive Loss 使用 200 對訓練對，標籤為 1(取自不同類別)與 0(取自同一類別)的圖像對各占一半；在測試任務中，準確率為自 10-Way One-Shot 支持集中預測到與真實圖像相同類別的圖片的機率。
- **資料增強(data augmentation)**：LFW 資料集每類只取一張圖像，孿生網路所需輸入的訓練對之另一張圖像使用資料增強補全，仿射轉換（Affine transformation）或正、逆時鐘旋轉方法擇一，生成另一張圖像。

## (六) 結果與討論

### 6. 1 Mnist 資料集

表 2 MNIST 資料集辨識準確率

| MNIST test accuracy | |
|---|---|
| Model | Accuracy |
| Humans | 95.5 |
| Hierarchical Bayesian Program Learning | 95.2 |
| Affine model | 81.8 |
| Hierarchical Deep | 65.2 |
| Deep Boltzmann Machine | 62.0 |
| Siamese Neural Net | 58.3 |
| **Siamese-Focal Loss** | **91.0** |
| **Siamese-Contrastive Loss** | **95.63** |
| **Siamese-L-softmax Loss** | **100.0** |
| **Siamese-Mixed Loss** | **44.54** |

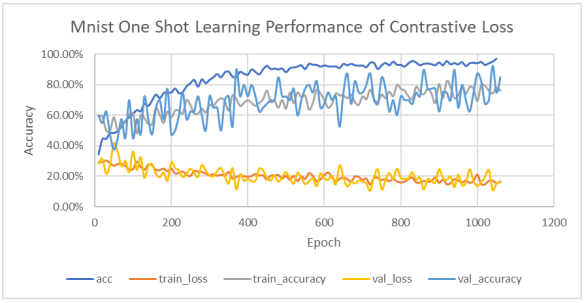


圖 7 Contrastive Loss 在 Mnist 資料集的準確率

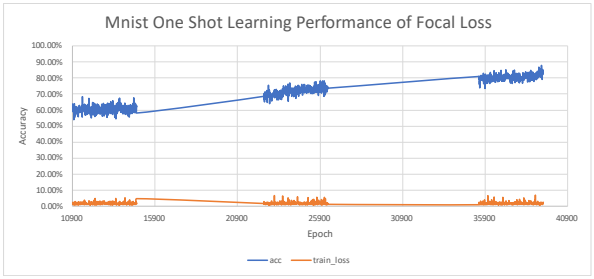


圖 8 Focal Loss 在 Mnist 資料集的準確率

- 辨識準確率（97.27%）較其他同為元學習的孿生網路架構高（70.3%）。
- 孿生網路使用 Contrastive Loss 收斂速度（1060 個 Epoch）較使用 Focal Loss 或其他網路架構快（90 萬個 Epoch）。

- 訓練使用的網路架構參數（40,538） 較其他網路架構（38,954,049）少很多。
- Focal Loss：將訓練輸入改成與測試輸入的數據型態（10-Way One-Shot 支持集）。雖然資料型態屬於 1（同類)的資料只有一個，其他都是 0（不同類）的資料，但 Focal loss 的特性正好可以解決資料分布不均的問題，因此可以表現良好。
- L-softmax Loss：非常敏感，會發生 1.梯度消失，2.過擬和的問題，調整網路參數與層數都無法改進，嘗試過 L-Softmax 論文裡的網路架構，還有先用 Softmax Loss 訓練再使用 L-softmax Loss，都造成過擬和，正嘗試修改 L-Softmax Loss 的函式。

### 6.2 LFW 資料集

表 3 LFW 資料集辨識準確率

| LFW test accuracy | |
|---|---|
| Model | Accuracy |
| OpenFace | 92.92 |
| DeepID | 97.05 |
| DeepFace | 98.37 |
| VGG-Face | 98.78 |
| CNN-3DMM estimation | 92.35 |
| FaceNet | 99.63 |
| **Siamese-Focal Loss** | **75.64** |
| **Siamese-Contrastive Loss** | **76.0** |
| **Siamese-L-softmax Loss** | **100.0** |
| **Siamese-Mixed Loss** | **50.0** |

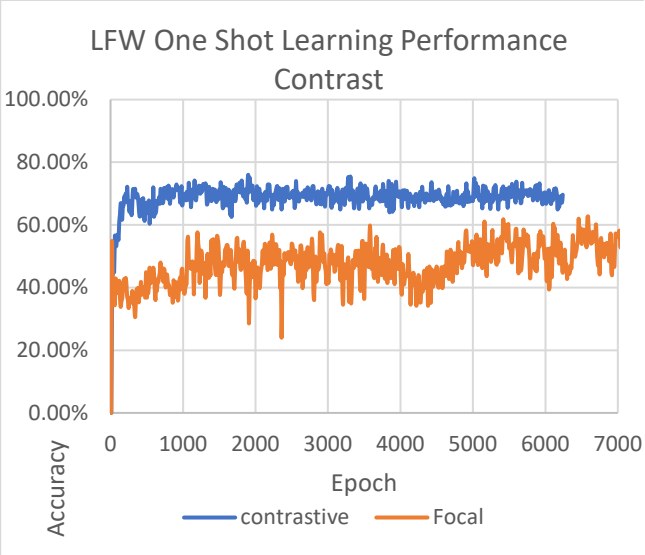


圖 9 Focal Loss 與 Contrastive Loss 的準確率比較

- 使用之資料量（原始資料 5749 張，加上資料增強 11,498 張）較其他訓練方式（3 萬-1 百萬張）少，訓練時實作 One-Shot 學習架構。
- 使用 Contrastive Loss 收斂速度（6240 個 Epoch）較使用 Focal Loss 或其他網路架構快（50 萬個 Epoch）。
- 訓練使用的網路架構參數（27,417,409）較其他網路架構（38,954,049）少。
- Mixed Loss：訓練之 Loss 已下降到很低，準確率卻沒有升高（維持在 50%），可以推斷模型梯度更新的方向並非最優解，而是陷入局部最優解中。模型於測試集上的預測值：

  [[0.6279017 ][0.64900494][0.63365936]...[0.690137 ][0.68854505][0.6770405 ]]；標籤為 1 與 0 各半，可以看出輸出預測值大多被判斷為 1，所以準確率在 50%，可能因為特徵向量間的距離位於 0~2 之間，在 Sigmoid 函數轉換後使梯度更新陷於 0.5 附近，優化困難。

圖 10 LFW 辨識結果

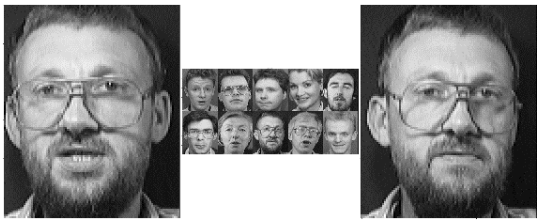

圖 11 AT&T 辨識結果

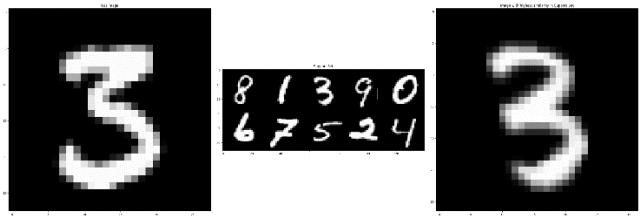

圖 12 MNIST 辨識結果

### 6.3 AT&T 資料集

表 4 AT&T 資料集

| AT&T test accuracy | |
|---|---|
| Model | Accuracy |
| self-organizing map (SOM)[11] | 85.5 |
| Fusion of global and local matching[12] | 95.83 |
| EDLGP[13] | 90.3 |
| SVM | 90.0 |
| CESR[14] | 88.5 |
| Siamese NeuralNetwork | 75.0 |
| **Siamese-Focal Loss** | **70.36** |
| **Siamese-Contrastive Loss** | **75.63** |

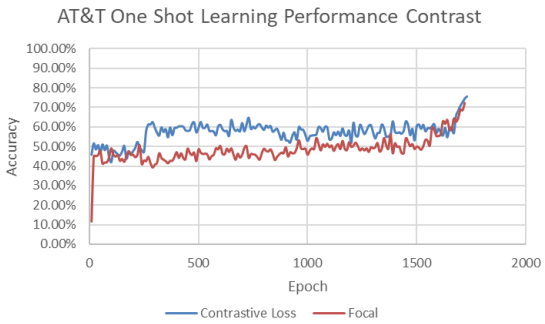


圖 13 Focal Loss 與 Contrastive Loss 的準確率比較

- 辨識準確率（75.63%）較其他同使用孿生網路架構高（75.0%），並實作 One-Shot 學習架構。
- 訓練使用的網路架構參數（40,538） 較其他網路架構（4,018,650）少很多。

### 6.4 mini-ImageNet 資料集

表 5 Mini-ImageNet 資料集

| Mini-ImageNet test accuracy | | | | |
|---|---|---|---|---|
| Loss function | | Focal Loss | Contrastive Loss | |
| **Method** | **Optimizer** | Accuracy (%) | | **Active Function** |
| 文獻[26] | Adam | 12 | 100 | ReLU |
| | SGD | 14.72 | 14.9 | ReLU |
| | RMSprop | 16 | 12.54 | ReLU |
| +Dropout | Adam | 8 | 100 | ReLU |
| 2-CNN | Adam | 15.09 | 14.36 | ReLU |
| | Adam | 12.72 | 21.63 | sigmoid |

| | | | | |
|---|---|---|---|---|
| | Adam | 12.54 | 12.54 | softmax |
| 3-CNN | Adam | 14.18 | 13.81 | ReLU |

- 使用文獻[4]（圖二）的網路，訓練正確率差，加入 Dropout 層，容易在很小的 Epoch 內過擬和。
- 激活函數（Active Function）使用 ReLu、Sigmoid 較好。
- 辨識效果差可能原因 1.照片的主體不明，對於孿生網路，可能把背景都考慮進相似度內。2.使用彩色圖片，孿生網路可能受顏色相似度影響，顏色相似度可能比形狀相似度高，讓網路誤將兩顏色相近的類別分在同一類。

## (七)結論與展望

目前成功實作 One-Shot 學習架構，並減少訓練成本（降低網路參數量、使用少量訓練資料）和提高收斂速度，在手寫字資料集上達到 97.27%的準確率，在更複雜的人臉資料集上也達到 76%的準確率。下一步希望應用其他最佳化方法優化準確率，並提升模型在其他資料集上的泛化能力。以及改進 L-Softmax Loss 會過擬合的問題。

- **最佳化方法**

  1.過擬和：增加網路層數與神經參數。

  2.使用分層學習率/動量、貝葉斯超參數優化學習最佳參數。

  3.修改各超參數，例如學習率、初始化、Bath Size……等。

  4.更改正則化層，或更改最後激活函數（Active Fuction）。

  5.PCA 與 DWT 等影像處理方法，對資料進行預處理。

## (八)參考文獻


[1] M. Shorfuzzaman and M. S. Hossain, "MetaCOVID: A Siamese neural network framework with contrastive loss for n-shot diagnosis of COVID-19 patients," *Pattern recognition,* 113: 107700, 2021.

[2] E. Khvedchenya and T. Gabruseva, "Fully convolutional Siamese neural networks for buildings damage assessment from satellite images," *arXiv:2111.00508*, 2021.

[3] Ravichandiran and Sudharsan, *Hands-On Meta Learning with Python*, Packt Publishing, 2018.

[4] G. Koch, R. Zemel and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," *Deep Learning Workshop on ICML*, vol. 2, 2015.

[5] S. Chopra, R. Hadsell and Y. LeCun. "Learning a similarity metric discriminatively, with application to face verification," *CVPR*, 2005.

[6] C. Feng, *et al*., "MTCSNN: Multi-task clinical siamese neural network for diabetic retinopathy severity prediction," *arXiv:2208.06917*, 2022.

[7] I. A. Lungu, *et al*., "Siamese networks for few-shot learning on edge embedded devices," *IEEE*

*Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 10, no. 4, pp. 488-497, 2020.

[8] I. A. Lungu, Y. Hu and S.-C. Liu, "Multi-resolution siamese networks for one-shot learning," *2nd IEEE Int. Conf. on Artificial Intelligence Circuits and Systems (AICAS)*, 2020.

[9] H. Wang, *et al*., "Cosface: Large margin cosine loss for deep face recognition," *CVPR*, 2018.

[10] Y. Zhao, *et al*., "Paralleled attention modules and adaptive focal loss for Siamese visual tracking," *IET Image Processing*, vol. 15, no. 6, pp. 1345-1358, 2021.

[11] Santaji Ghorpade, *et al*., "Pattern Recognition Using Neural Networks" *International Journal of Computer Science and Information Technology (IJCSIT)*, 2010.

[12] Kisku, D. R., Tistarelli, M., Sing, J. K., & Gupta, P. (2009, June). " *Face recognition by fusion of local and global matching scores using DS theory: An evaluation with uni-classifier and multi-classifier paradigm*". In 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (pp. 60-65). IEEE.

[13] Bi, Ying, Bing Xue, and Mengjie Zhang. "*Genetic Programming-Based Evolutionary Deep Learning for Data-Efficient Image Classification.*" IEEE Transactions on Evolutionary Computation (2022).

[14] He, Ran, Wei-Shi Zheng, and Bao-Gang Hu. "*Maximum correntropy criterion for robust face recognition.*" IEEE Transactions on Pattern Analysis and Machine Intelligence 33.8 (2010): 1561-1576.