2022 Computer Vision Final Project Report
Team 5

李采蓉 111062537 / 陳怡汝 111062511 / 黃寶萱 111052543
Equal Contribution

Abstract

近年來人臉辨識技術被廣泛使用在不同領域，例如身份辨識、人員追蹤及社交軟體等等的應用。目前許多深度學習的模型，如 VGGFace, ResNet, FaceNet 等，皆在人臉辨識上達到了可觀的準確率。在這次的期末專題中，我們參考了經典的 ArcFace 模型，並使用 MobileFaceNet 作為我們的模型骨架，同時結合 arc margin 方法來學習圖片中的人臉特徵。此外，我們也使用 Siamese Network 架構針對人臉的跨年齡特徵進行進一步的 fine-tune，使我們的模型能夠學習到跨年齡的特徵資訊，以達到更好的模型辨識結果及準確率。

1. Introduction

近年來，隨著人臉辨識技術的提升與進步，人臉辨識模型被廣泛使用在不同領域，例如身份辨識、人員追蹤及社交軟體等等的應用。目前常用的人臉辨識方法多使用深度學習網路，例如卷積神經網路(CNN)，先利用模型對輸入的人臉圖像進行特徵提取，並將提取到的特徵與目前模型看過的訓練資料進行比較，找出與輸入圖像最相似的人臉。

雖然在人臉辨識技術上，使用現有的深度學習的模型，如 VGGFace, ResNet, FaceNet 等，普遍皆可以達到不錯的辨識結果，但是在處理「跨年齡 (Cross-age)」的人臉辨識問題則有較多困難的地方。其困難點主要有以下三點：

1. 人臉特徵隨年齡改變：隨著人們年齡的增長，人臉特徵會隨之改變，進而造成人臉辨識難以準確匹配不同年齡的臉部特徵。
2. 訓練資料的取得困難：目前常用的人臉辨識資料集，如 LFW 、YTF、MegaFace 等，皆不包含年齡變化的資料，針對年齡變化搜集的資料集相對較少，而針對不同年齡長期追蹤長相也容易造成隱私問題，進一步提升資料搜集的難度。
3. 真實場景拍攝影像品質不穩定：在真實世界拍攝的影像包含許多變量，包括真實世界光照環境的變化，人臉姿勢及背景，不同的相機參數及位置等，都會影響到模型辨識的準確度。

在我們的方法上，我們使用經典的 ArcFace 的 arc margin 搭配 MobileFaceNet 作為模型骨架，並且使用 Siamese Network 針對跨年齡特徵進行 fine-tune，讓模型能夠學到跨年齡的資訊，進而達到更好的人臉辨識結果。

2. Method

我們的模型主要分成兩階段的訓練：第一階段使用 MobileFaceNet 作為模型骨架，提取臉部特徵，搭配 ArcFace margin，對不同的人臉進行分類。在第二階段，我們對第一階段學到的特徵提取器進行 fine-tune，在這個階段我們主要使用了 Siamese Network 搭配 contrastive loss 對人臉跨年齡的 feature similarity 進行 supervision，迫使特徵提取器學到跨年齡資訊，以達更好的模型辨識結果。

2-1. Stage-1 MobileFaceNet-Based ArcFace Feature Extraction

ArcFace 是一種著名的人臉辨識模型，在人臉辨識上有極高的正確率，其作法主要針對輸入影像進行特徵提取，並透過 ArcFace Margin 來提高人臉辨識的準確度。ArcFace Margin 透過對特徵向量和權重歸一化，對角度空間 θ 加上角度間隔 m，使得模型可以更好地辨別出相似的人臉。

我們採用 MobileFaceNet 作為特徵提取器，對輸入的圖片進行特徵提取，並且利用 ArcFace margin 對特徵向量進行分群，得到最終預測的 label。針對此 label，我們與 ground truth 計算 cross-entropy loss，對模型進行迭代訓練。

2-2. Stage-2 Cross-age Feature Supervision

由於僅對圖片使用 ArcFace margin 進行特徵分群無法使模型學習到跨年齡的人臉特徵，因此，除了第一階段的訓練之外，我們針對跨年齡特徵進行第二階段的學習，迫使特徵提取器在跨年齡的資料上亦可學出相似的特徵。

我們參考了 Siamese Network 的做法，讓兩張輸入影像通過與第一階段相同的 shared weights 特徵提取器並且計算兩特徵的相似度。當兩輸入影像為同一人時，其相似度應較高，反之則較低。我們對此相似度進行 supervision，使用 contrastive loss 針對 positive pair 及 negative pair 分別計算 loss value，並且加入第一階段使用的 cross-entropy loss，以確保特徵提取器在學習年齡資訊的同時亦不會丟失掉人臉辨識的能力。

在第二階段的訓練中，我們的模型使用了第一階段的訓練結果作為第二階段特徵提取器的初始值，並對其進行 fine-tune，因為我們在訓練過程中發現，如果我們一開始就使用 Siamese Network 的方法，只對年齡特徵進行學習，會因為缺少良好的初始特徵權重，而導致模型無法收斂。而如果對第二階段訓練只採用 contrastive loss 而不加上第一階段的 cross-entropy loss，也會使模型因為缺少對人臉辨識特徵的限制，而無法有更好的訓練結果。

2-3. Loss Function for Stage-2

在第二階段的訓練中，為了使我們的模型在學習年齡資訊的同時仍保有提取人臉特徵的能力，因此我們的 Loss Function 使用第一階段的 Cross-Entropy Loss ，並結合 Contrastive Loss 使我們的模型能有較好的訓練結果。Cross-Entropy Loss $(L1)$ 的部分，我們分別對輸入的成對圖像進行人臉特徵提取，計算其與 ground truth 的 cross-entropy loss，並取兩者的平均值。Contrastive Loss (式 1) 的部分，我們計算兩圖片特徵向量的 Euclidean distance (D_W)，並設定 margin (m)為 2.5，ground truth similarity (Y)在 positive pair 為 1，

negative pair 為 0，得到最終的 Total Loss (式 2)，即為我們 Stage-2 的 Loss Function。

$$Contrastive\ Loss\ (L2) = Y\frac{1}{2}(D_W) + (1-Y)\frac{1}{2}\{max(0,\ m - D_W)\}^2 \quad \text{- 式 1}$$

$$Total\ Loss = \frac{1}{2}(L1_i + L1_j) + L2 \quad \text{- 式 2}$$

3. Experiment

Training details

我們使用 CALFW 資料集進行訓練，其中包含 12174 張圖片，9739 張訓練圖片及 2435 測試圖片，針對訓練圖片，我們做了 Random Horizontal Flip 作為 data augmentation，並將圖片以 mean 為 0.5、standard deviation 為 0.5 做 normalization。在訓練參數的設定上，第一階段我們設定初始 learning rate 為 0.1，並使用 SGD optimizer，設定 learning rate step scheduler 每 10 個 epochs 降低 0.95，batch size 為 16；第二階段訓練我們設定 learning rate 為 0.1，一樣使用 SGD optimizer，設定 learning rate step scheduler 每 10 個 epochs 降低 0.95，batch size 為 16。

在訓練設備上，我們使用兩張 GTX 1080 Ti GPUs 進行訓練，第一階段訓練 100 個 epoch，共需 5 小時，在第二階段訓練共訓練 100 個 epochs，共需 8 小時。

Data preprocess

我們使用 CALFW aligned image 作為 Training data，訓練圖片大小為 128x128。第一階段直接使用訓練圖片，並將每個不同人的身份以數字編號成從 0 到 4025 進行分類。第二階段我們將所有訓練資料進行配對，產生 11169 組 positive pairs 及 11169 組 negative pairs，並將 positive pair 的 ground truth similarity 設為 1，negative pairs 的 ground truth similarity 設為 0 進行訓練。

Evaluation Metrics

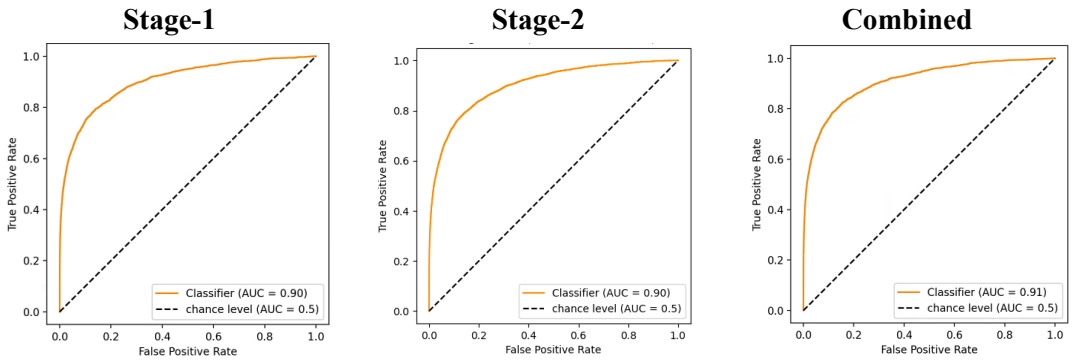
我們使用兩個常用的 evaluation metric 進行結果的評估：

1. Rank-1 accuracy: 我們對所有測試資料進行窮舉，計算所有 pair 的 similarity，如果第一高的 similarity 所代表的圖片與目前圖片為同一個人，則 accuracy 加一，最後計算 accuracy 的總和並與所有 pair 求平均，得到 Rank-1 accuracy。
2. ROC/AUC: 如上所述，我們對所有測試資料進行窮舉並計算所有 pair 的 similarity，並且記錄所有 pair 是否為同一個人。若為同一個人，則以 similarity 為 1 作為 ground truth，反之以 0 為 ground truth，並使用 scikit-learn 套件計算 ROC 及 AUC。

Comparisons

下圖(圖一)及下表(表一)分別表示 Stage-1 (MobileFaceNet+ArcFace margin)、Stage-2 (Siamese Network)、Combined (Stage-1 + Stage-2)的訓練結果。Fig1 為針對 Validation set 的 ROC 曲線表示，藉由 ROC 曲線可以看出我們的方法可以正確預測成對圖像的相似程度，並且相較於使用 Stage-1 或 Stage-2 的訓練有更好的結果。Table1 則表示不同設定下測試在 validation 及 testing set 的 Rank-1 accuracy 及 AUC 數值比較，藉由 Table1 也可以發現我們的方法無論在 Rank-1 accuracy 亦或是 AUC 的數值上相較於 Stage-1 或 Stage-2

皆有顯著的提升。因此可知，我們使用兩階段訓練模型的方法，不僅可以讓模型學到跨年齡的臉部資訊，同時也保有區分不同人臉臉部特徵的能力，相較於 ArcFace 僅考慮不同人的臉部特徵，以及 Stage-2 缺乏良好初始化模型權重的設定，我們的方法在處理跨年齡人臉辨識問題上能有較好的表現。



圖一

	Stage-1	Stage-2	Combined
Rank-1 acc. (Validation)	0.33	0.31	**0.35**
AUC (Validation)	0.90	0.90	**0.91**
Rank-1 acc. (Testing)	0.62	0.58	**0.70**
AUC (Testing)	0.77	0.76	**0.78**

表一

Bonus

在 Bonus 的處理上，我們利用兩種分類依據對我們的 combined model 做測試。

1. 以臉部特徵作依據：我們使用 K-means clustering 以及 hierarchical clustering 以臉部特徵作為 grouping 的方法，藉由 scikit-learn 套件來完成分群的任務，分別可以得到 72 以及 80 個 grouping 結果。
2. 以 feature similarity 作依據：藉由計算單一輸入圖像與其他圖像的 feature similarity，取出與之相近的前五張圖像分類作為一組，一組 6 張圖像，直至分出 20 組，可以得到 48 個 grouping 結果。

最終我們選用以 hierarchical clustering 的方式對於臉部特徵作分類的方法完成 Bonus 部分，可以得到 80 個 grouping 結果。

4. Other Experiments

4-1 Model backbone

Resnet: 除了 MobileFaceNet 之外，我們另外嘗試了著名的 Resnet18 及 Resnet50 作為 Stage-1 訓練使用的 model backbone，但是結果卻無法收斂，推測原因可能是 Resnet 為了有更深的模型結構所使用的 shortcut 反而限制了模型學習特徵的能力，因此在我們的 task 上並不適用。

4-2 Loss function

Focal loss: 我們在 Stage-1 除了使用常見的 cross-entropy 之外，也嘗試使用 focal loss 進行訓練，但模型訓練結果在 validation set 的測試上結果不如 cross-entropy loss，因此在我們在後續實驗中皆使用 cross-entropy 進行 loss 運算。

4-3 Data augmentation

Adjust image brightness: 我們在 Stage-2 fine-tune 訓練時嘗試先對我們的資料進行圖像亮度及色調的微調以增加圖片之間的相異度，試圖增加資料量使讓我們的模型可以利用在 Stage-1 未看過的影像進行訓練，達到避免發生 overfitting 的效果，但從經過測試後發現模型並未有更好的表現，我們推測經過調整後的影像可能喪失原本影像的特徵，導致原本在 Stage-1 得到的特徵提取器無法有效的進行人臉特徵提取，因此無法提升最後模型的分辨能力。

	Brightness	**Grayscale**	**Original**
Rank-1 acc. (Testing)	0.56	0.66	**0.70**
AUC (Testing)	0.77	0.76	**0.78**

Random rotate: 我們在訓練時嘗試對資料進行 random rotate 以增加資料量及避免 overfitting 的發生，然而我們發現這樣的訓練反而會因為擾動了已對齊的臉部 landmarks 而造成訓練上的困難，因此在後續的實驗中我們皆僅使用 random horizontal flip 作為 data augmentation。

5. Conclusion

在我們的方法上，我們參考了經典的 ArcFace 模型，使用 MobileFaceNet 作為模型骨架，結合 arc margin 學習人臉特徵，並使用 Siamese Network 針對人臉跨年齡特徵進行近一步的 fine-tune，讓模型能夠學到跨年齡特徵資訊，以達到更好的模型辨識結果。從我們實驗的結果可以看出，加入第二階段的訓練能夠有效幫助我們的模型提升跨年齡臉部辨識的能力，並且達到 0.70 的 rank-1 accuracy 以及 AUC 0.78，相比於單純使用 ArcFace 訓練的結果更加準確。在未來，我們認為我們的模型還有繼續改善的空間，我們可以透過 Heatmaps 分析並了解我們的模型在訓練過程中參考了圖片上的哪些特徵，進一步針對特定特徵改善特徵提取器的能力，此外，在第二階段訓練的過程中，我們也可以進一步探討如何選取對提升模型更有幫助的成對圖片，進而讓模型有更好的收斂結果。

6. Reference

1. Jiankang Deng, Jia Guo, Niannan Xue, Stefanos Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4690-4699
2. Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, Piotr Dollar. Focal Loss for Dense Object Detection. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980-2988
3. Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Sackinger and Roopak Shah. Signature verification using a "Siamese" time delay neural network. Proceedings of the 6th International Conference on Neural Information Processing Systems (NIPS), 1993, pp. 737–744
4. Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, Wei Liu. CosFace: Large Margin Cosine Loss for Deep Face Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5265-5274
5. Florian Schroff, Dmitry Kalenichenko, James Philbin. FaceNet: A Unified Embedding for Face Recognition and Clustering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 815-823
6. Chen, S., Liu, Y., Gao, X., Han, Z. MobileFaceNets: Efficient CNNs for Accurate Real-Time Face Verification on Mobile Devices. Biometric Recognition (CCBR), 2018, Lecture Notes in Computer Science, vol 10996. Springer, Cham.