Final project of Machine Learning

111652001 吳文生

1. AI 的未來能力
   請具體描述一件你認為目前 AI 無法做到，但 20 年後有可能做到的重要事情。

我認為 AI 有可能在 20 年後實現《名偵探柯南：黑鐵的魚影》中的「全年齡認證」（All-Age Recognition）系統，它被安裝在位於太平洋上用以連接全球各地警備監控數據的「太平洋浮標」（Pacific Buoy）設施內。一般的臉部辨識系統，就像你在手機上解鎖一樣，只能認出你「現在」的樣子，但這個「全年齡認證」系統厲害的地方在於：

1. 它看過全世界所有人的照片：它連接了全球各地的監視器和身份資料，等於它手上有全世界幾十億人的「從小到大」的各種照片。

2. 它會「猜」你長大或縮小的樣子：無論是一個人從小孩長大變老，或者像柯南、小哀一樣，吃了藥身體縮小，臉變回小孩子的樣子，這個系統都能靠著那些大量的照片，準確地比對出來，它能從一個小孩子的臉，推測他成年時的樣子；也能從一個人成年的樣子，倒推回他幼年時期的外貌。

在電影中，苦艾酒透過偽裝成相同面孔的人，便可讓系統判斷為是同一人。但走路姿勢也是識別一個人的一種方式，就算苦艾酒偽裝技術再厲害，我想也無法完美複製另一人的走路姿勢，因此我認為在 20 年後的「全年齡認證」除了能從影像中識別、預測人臉，甚至具備處理動態影片的功能。

透過「全年齡認證」，警方在追蹤可疑人士或失蹤兒童時，可以輸入一張舊照片，讓系統分析或預測目標人物經過指定時間後的可能樣貌，這在找尋失蹤多年的人口時有很大幫助。但現實生活中，若政府或大型機構部署類似的「全年齡認證」系統，它可以將任何時期的影像都歸檔到你的個人檔案下，形同無處不在、跨越時間的監控，這會嚴重侵害個人的隱私權和自由。

2. 所需的成分與資源（Ingredients）
   若要實現問題一中描述的能力，請具體說明你認為所需的「成分（ingredients）」有哪些。

系統的資料來源是來自全球各地的影像序列，包括靜態照片和動態步態影片，目標訊號除了個體的唯一身份 ID 外，還包括特定年齡的真實樣貌影像（用於訓練預測生成模型）。由於此

系統較不需要與物理環境進行即時互動，因此其學習回饋主要透過損失函數 (Loss Function) 實現，例如將識別結果與真實標籤進行比對的分類損失，或將預測影像與真實影像進行比對的生成損失。

3. 涉及的機器學習類型
   根據你對問題一的構想，判斷該能力的實現主要涉及哪一類機器學習方法。

要實現「全年齡認證」系統，我們會用到監督式學習 (Supervised Learning) 與非監督式學習 (Unsupervised Learning) 。

1. 監督式學習 (Supervised Learning)：主要負責最終的識別與預測功能，系統需要大量的「標籤化數據」，即同一個人的多張不同年齡、不同外貌狀態（輸入）與其唯一的身份 ID（目標訊號）進行配對訓練。

2. 非監督式學習 (Unsupervised Learning)：主要負責特徵提取與表徵學習，人臉和動態步態（Gait）影像數據維度高且複雜，年齡變化是一個巨大的干擾因素，非監督式方法，如自編碼器（Autoencoders），能從大量未標記的影像中學習並提取出穩定的潛在表徵 (Latent Representation)，例如步頻、步長比例、關節轉動的角度、雙腳著地時的身體晃動幅度，這些潛在表徵可以做為跨年齡識別的基礎。

4. 第一步的「可實作模型問題 (Solvable Model Problem)
   假設要讓 AI 在 20 年後達到問題一的能力，請設計並實際解出一個目前可行的「簡化模型問題」，作為邁向該能力的第一步。

如果要在二十年後實現此系統，我想第一步是處理如何從有限的數據中學習到穩定的潛在表徵，因此，我們設計的簡化模型是透過「跨年齡小樣本身份識別」，來邁向最終的大規模、跨時間的識別 (Image$(t) \to$ Identity ID)。

而簡化問題則將其縮減為一個二元分類問題：判斷兩張不同年齡的影像 Image(t_1) 和 Image(t_2) 是否為同一個人 ($\to$ Yes/No)，這個簡化模型的核心任務不僅僅是輸出 Yes/No，而是要構建一個與年齡無關的特徵空間。

令 $\mathcal{X} \subset \mathbb{R}^n$ 為輸入影像空間，我們的目標是學習一個參數化的非線性函數 $f_\theta : \mathcal{X} \to \mathbb{R}^D$，其中 θ 為神經網絡參數，D 為特徵向量的維度，對於任意兩張影像 x_i, x_j，若它們屬於同一身份（無論年齡差距多大），則其歐氏距離 $d(f_\theta(x_i), f_\theta(x_j))$ 應小於閾值 τ；若屬於不同身份，則距離應大於 τ。

如果 f_θ 能成功訓練，則向量 $v = f_\theta(x)$ 即為「身份特徵碼」。在未來的「全年齡認證」中，當警方輸入一張舊照片 x_{old}，系統只需計算 $v_{old} = f_\theta(x_{old})$，然後在數據庫中尋找最近鄰 v_{target}，即可鎖定目標。

在模型與方法上，我們採用 Siamese/Triplet 架構，使用 ResNet-50 作為基礎，Residual Connections 能學習到更細微的臉部骨骼與五官幾何特徵，這些特徵比皮膚紋理更抗老化。在 Triplet Network 中，處理 Anchor, Positive, Negative 的三個 CNN 完全共享同一組參數 θ，這保證了特徵空間的唯一性。我們還需要將其結果轉化為一個精簡的「身份特徵碼」(特徵向量)，也就是將特徵圖壓縮成 D 維向量，去除多餘資訊，只保留最關鍵的身份特徵。另外，我們將輸出向量縮放為單位長度，這使得所有特徵向量都分佈在一個 D 維的超球面上，防止影像亮度或對比度影響距離計算，並使 Triplet Loss 的收斂更穩定。在 Loss function 上，我們採用 Triplet Loss，樣本是三個一組的 $\mathcal{T} = (A, P, N)$：

- A (Anchor)：基準樣本
- P (Positive)：與 A 同一身份的正樣本（跨年齡）
- N (Negative)：與 A 不同身份的負樣本

Triplet Loss 公式為：

$$\mathcal{L} = \sum_i \max(0, ||f(x_i^A) - f(x_i^P)||_2^2 - ||f(x_i^A) - f(x_i^N)||_2^2 + \tau)$$

其中 τ 是閾值。

隨著訓練進行，隨機選取的 Triplet 大部分會很容易滿足條件 (即 $d(A, P) + \tau < d(A, N)$)，導致 Loss 為 0，模型停止學習。為了讓模型具備「小樣本」泛化能力，我們採用 Hard Negative Mining 策略：在每一個訓練 Batch 中，針對每一個 Anchor A，我們不隨機選 N，而是刻意挑選那些「長得最像 A 的不同人」(Hard Negative)，強迫模型去區分那些細微的差異（例如區分兩個長得很像的老人)。

在數據的輸入上，我們隨機選取 P 個不同的身份 (Identity)，每個身份隨機選取 K 張不同年齡的照片，即 Batch Size $= P \times K$。假設 $P = 8, K = 4$：這個 Batch 裡會有 8 個人，每人有 4 張不同年齡的照片。

- 對於 Anchor (A)：模型可以在這 4 張照片中，任選一張當基準。
- 對於 Positive (P)：模型可以在該人剩下的 3 張照片中，選一張當正樣本（例如 A 是 5 歲，P 是 40 歲)。

- 對於 Negative (N)：模型可以從 Batch 裡剩下的其他 7 個人（28 張照片）中，挑選一張當負樣本。
- Hard Negative Mining：系統在這 28 張負樣本中，找出「長得最像 A 的那個人」來計算 Loss，強迫模型學得更精準。

模型訓練完成後，我們透過以下步驟評估其在「未見過的測試集」上的效能：

1. 距離矩陣計算：計算測試集中所有影像兩兩之間的歐氏距離。
2. ROC 曲線繪製：
    - 設定一系列變動的距離閾值 τ。
    - 真陽性率 (TAR, True Accept Rate): 同一人被正確判定為同一人的比例。
    - 假陽性率 (FAR, False Accept Rate): 不同人被錯誤判定為同一人的比例。
3. 等錯誤率 (EER): 找到一個閾值 τ，使得 $FAR(\tau) = 1 - TAR(\tau)$（即錯誤接受率等於錯誤拒絕率），EER 越低，代表系統在嚴格的安全性和便利性之間取得了更好的平衡。

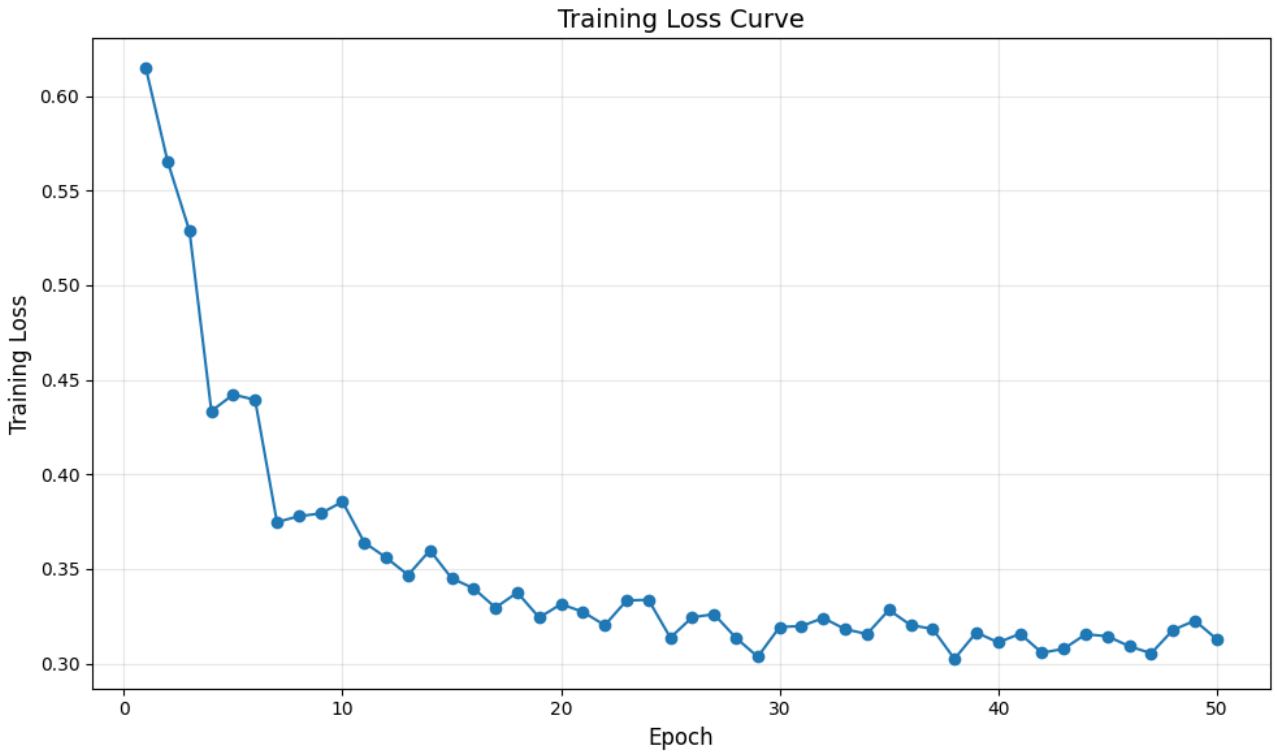

Figure 1: Training Loss Curve

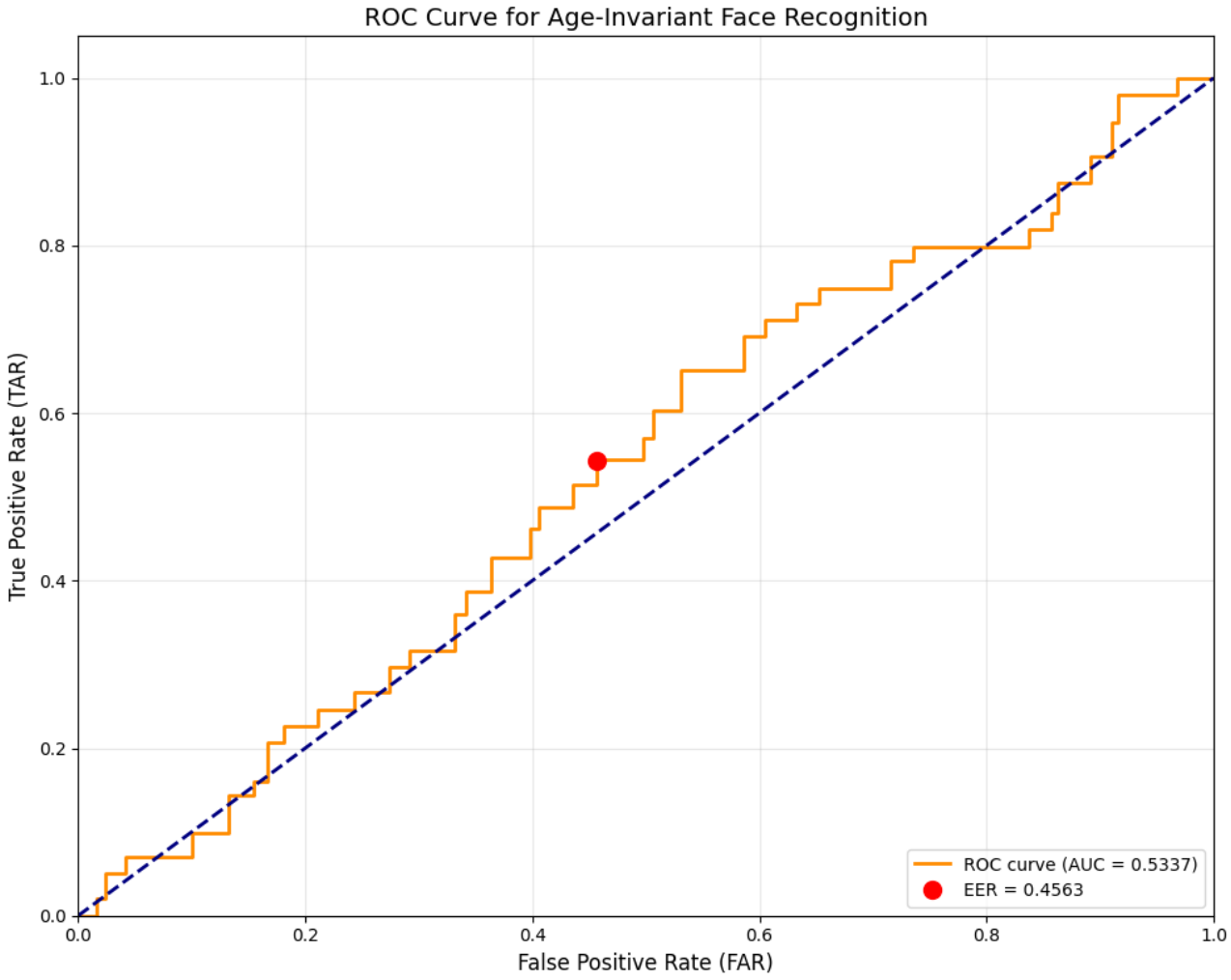

Figure 2: ROC Curve for Age-Invariant Face Recognition

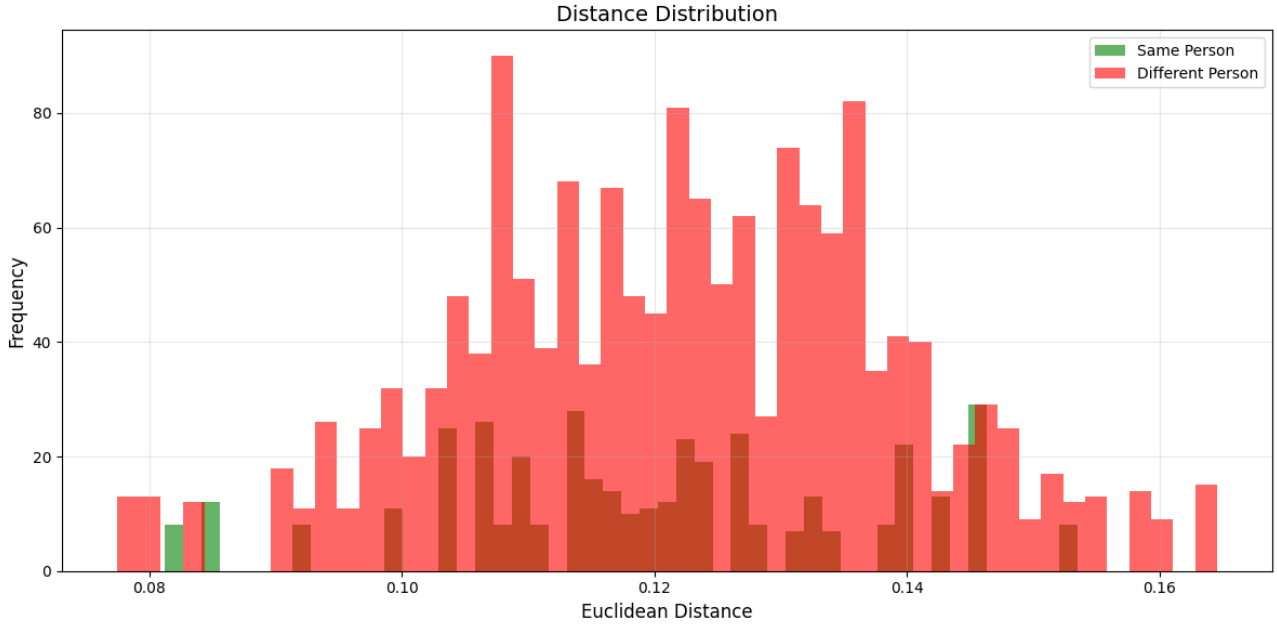

Figure 3: Distance Distribution